

CONTENTS;

	<u>Page</u>
Programme	3-5
Invitations	6
Committees	7-11
Main Topics	12
Keynote Lectures	13-18
Oral Abstracts	19-104
Poster Abstracts	105-107
Full Texts	108-472



www.irsysc2019.com



5th INTERNATIONAL RESEARCHERS, STATISTICIANS AND YOUNG STATISTICIANS CONGRESS

18-20 October 2019
Amara Sealight Elite Hotel, Kusadasi- TURKEY

TECHNICAL PROGRAM

MOTTO
www.motto.tc



5th INTERNATIONAL RESEARCHERS, STATISTICIANS AND YOUNG STATISTICIANS CONGRESS

18-20 October 2019
Amara Sealight Elite Hotel, Kusadasi- TURKEY



TECHNICAL PROGRAM

October 18, Friday

OPENING

13.00-13.30	Onur Köksoy (Congress Chair)
13.30-14.15	Claudiu Hertellu (Keynote Speaker) <i>Baby Conception between Religiousness and Pleasure in Romania. Comparing Eastern Orthodox and Non-Orthodox Populations through Very Long Daily Time Series (1905-2001) Analysis (Claudiu Hertellu, Bogdan Vasile Ileanu, Marcel Ausloos and Giulia Rotundo)</i>
14.30-15.15	Hakan Demirtaş (Keynote Speaker) <i>Hybrid data generation (Hakan Demirtaş)</i>
15.30-16.15	Luigi D'Ambra (Keynote Speaker) <i>Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation (Luigi D'Ambra, Pietro Amenta and Antonello D'Ambra)</i>

16:15-16:30 Coffee Break

Parallel Sessions - October 18, Friday • 16.30-17.45

Biostatistics - Hall 1 • Chair: Adnan Karaibrahimoğlu

- 16.30-16.45 Imputation of Missing Observations in Longitudinal Data via Neural Network - Marwa BenGhoul, Berna Yazıcı
16.45-17.00 Several Applications of New Generalized Entropy Optimization Methods in Survival Data Analysis - Aladdin Shamilov, Nihal Ince, Sevda Özdemir Çalikuşu
17.00-17.15 Performances of the distribution function estimators based on ranked set sampling using body fat data - Yusuf Can Sevil, Tuğba Yıldız
17.15-17.30 Anxiety and Attitudes towards Biostatistics and Scientific Research Methods Courses of Students of a Dental School - Adnan Karaibrahimoğlu, Nazan Karaoğlu, Said Karabekiroğlu
17.30-17.45 Robust Gene Co-Expression Network Analysis - Aylin Alın, Ayça Ölmez

Statistics and Probability - Hall 2 • Chair: Berna Yazıcı

- 16.30-16.45 A Simulation Study on the Unbalanced Design Properties of the Generalized P-value Based Tests - Mustafa Cavus, Berna Yazıcı
16.45-17.00 A New Lifetime Distribution Based on the Triangular Kernel - İsmet Birbiçer, Ali İhsan Genç
17.00-17.15 An Application of New Stochastic Model Using Generalized Entropy Optimization Methods - Aladdin Shamilov, Nihal Ince
17.15-17.30 Fitting One Life Expectancy at Birth Data to Stochastic Differential Equation - Aladdin Shamilov, Sevda Özdemir, Fevzi Erdogan
17.30-17.45 Convolutional Neural Network Architectures for Sentiment Analysis in Turkish - Aytuğ Onan

October 18, Friday • 18.00 -19.30

Industrial Statistics and Engineering Applications - Hall 1 • Chair: Murat Alper Basaran

- 18.00-18.15 A New Risk Assessment for the Right-Skewed Processes - *Melis Zeybek, Onur Köksoy*
 18.15-18.30 Combining Sales Forecasts at the Sku Level: An Application in the Food and Beverage Industry - *Erkan Işıklı*
 18.30-18.45 Support Vector Regression for Weather Forecasting - *Neslihan Çevik, Ahmet Sermet Anagün*
 18.45-19.00 Healthcare Tourism Demand and an Empirical Analysis for Istanbul - *Erkan Işıklı, Bilgesu Bayır*
 19.00-19.15 A Golden Ratio Control Chart for Monitoring the Process Mean - *Elif Kozan, Onur Köksoy*
 19.15-19.30 A Regression Analysis for Predicting the Amount of Yearly Greenhouse Gass Emissions in Turkey - *Merve Avcı, Banu Yetkin Ekren*

Statistics and Probability - Hall 2 • Chair: Selma Toker

- 18.00-18.15 Modelling and Analyzing Bivariate Survival Data Based on Copulas - *Ece Görceğiz, Burcu Hudaverdi Ucer*
 18.15-18.30 Ranking Error Models, Cost and Optimal Set Size in Ranked Set Sampling - *Sami Akdeniz, Tuğba Yıldız*
 18.30-18.45 A new parametric test for detecting linear trend in location - *Halil Tanil, Agah Kozan*
 18.45-19.00 Improving Two Stage Two Parameter Ridge Estimator under Linear Restrictions - *Selma Toker, Nimet Özbay*
 19.00-19.15 Defining Some Adaptive Optimal Estimators for the Distributed Lag Model - *Nimet Özbay, Selma Toker*
 19.15-19.30 Performance Comparison of Machine Learning Methods and Traditional Time Series Methods for Forecasting - *Ozancan Özdemir, Ceylan Yozgatlıgil*

October 19, Saturday

09.00-09.45 **Tatjana V. Šibalija (Keynote Speaker)**

Statistical vs. Metaheuristic Techniques in Parametric Optimisation of Industrial Processes (Tatjana V. Šibalija)

Parallel Sessions - October 19, Saturday • 10.00 - 11.15

Statistics and Probability - Hall 1 • Chair: Ali İhsan Genç

- 10.00-10.15 Multicomponent Stress-Strength Reliability Estimation based on the Standard Two-Sided Power Distribution - *Çağatay Çetinkaya, Ali İhsan GENÇ*
 10.15-10.30 Evaluations of the Mean Residual Lifetime Function of a Multi-state System - *Funda İscioğlu*
 10.30-10.45 Response Surface Approximation to Stress-Strength Reliability Under Dependency Structure - *Gözde Kuş, Sevcan DEMİR Atalay*
 10.45-11.00 Determining the types of missing data under supervised statistical models - *Vladimir Vasić*
 11.00-11.15 Classification of the Prices of Real Estate Using Machine Learning Methods - *Betül Kan Kılınc, Yonca Yazırlı*

Econometrics - Hall 2 • Chair: Fatih Cemrek

- 10.00-10.15 Investigating Chaotic Dependence between Economic Growth and Energy Consumption - *Aygül Anavatan*
 10.15-10.30 Analsis of Turkey Household Budget Survey Data with Quantile Regression - *Ismail Yenilmez, Yeliz Mert Kantar*
 10.30-10.45 The Relationship between Health, Education Expenditures and Economic Growth: The Case of NUTS-1 Regions in Turkey - *Aygül Anavatan, Zerife Yıldırım*
 10.45-11.00 Economic Analysis of the Effect of Opec Oil Policies on Economic Growth - *Fatih Cemrek, Hüseyin Naci Bayrak*
 11.00-11.15 Analysis of Internal Migration Movements in Izmir by Multinomial Logit Model - *Tuba İlhan, Şenay Üçdoğruk Birecikli*

11.15-11.30 **Coffee Break**

October 19, Saturday • 11.30 - 12.45

Biostatistics - Hall 1 • Chair: Hayal Boyacıoğlu

- 11.30-11.45 How is the performance of the Mc-Nemar test to determine cut-off comparing to the Youden Index and the minP methods for ordinal data? A simulation study - *Pervin Demir, Afra Alkan, Selcen Yüksel*
 11.45-12.00 The Performance Comparison of Feature Selection Methods in Health Datasets - *Mert Demiralp, Asli Suner*
 12.00-12.15 Evaluation of the Performance of Boosting and Bagging Classification Algorithms after Preprocessing in Health Data - *Yüksel Ozkan, Asli Suner*
 12.15-12.30 Estimating Species Diversity Components of Various Forest Stand Types in The Lake Districts using a Draft Software for Biodiversity Estimation (BYÇEP) - *Ahmet Mert, Kürşad Özkan*
 12.30-12.45 Estimating Community Rarity by Creating the Locked Matrix and Using the Tsallis Entropy - *Kürşad Özkan*

Econometrics - Hall 2 • Chair: İpek Deveci Kocakoç

- 11.30-11.45 Regression Analyses or Decision Trees? - *Burcu Kocarik Gacar, İpek Deveci Kocakoç*
 11.45-12.00 The Change of The Factors Determining Happiness in Turkey - *Eda Yalçın Kayacan*
 12.00-12.15 HEGY Seasonal Unit Root Test: An Application for Agricultural Products Producer Price Index - *Okan Küreş, Fatih Cemrek*
 12.15-12.30 R&D Expenditures for the NUTS-1 Regions in Turkey and a Bootstrap Panel Causality Analysis of the Relationship with GSYM - *Zerife Yıldırım*
 12.30-12.45 Contribution of Primary Production of Renewable Energy to Economic Growth and Labor Force: EU-28 Panel Data Analysis - *Selena Kantarmacı, Şenay Üçdoğruk Birecikli*
 12.45-13.00 Inflation Targeting and Taylor Rule Model for Developed And Developing Countries: A Panel Data Analysis - *Hande Erk, Hamdi Emec*

12.45-14.00 **LUNCH**

October 19, Saturday

14.00-14.45 **Rama Shanker (Keynote Speaker)**

Statistics and its uses in Biomedical Sciences

Parallel Sessions - October 19, Saturday • 15.00 - 16.15

Biostatistics - Hall 1 • Chair: İlker Ercan

- 15.00-15.15 A semi-parametric method to detect outbreaks in syndromic surveillance - *İmren Saygır Yılmaz, Eralp Doğu, Dursun Aydın*
 15.15-15.30 Diagnostic Meta-Analysis: An Application in Dentistry - *Merve Parmaksız, Hayal Boyacıoğlu, Pelin Güneri*
 15.30-15.45 "Investigation of Patients' Persistence Rate of Antibiotic Prescription by Sensitive Question Method; A Cross Sectional Study" - *Robab Ahmadian, İlker Ercan, Yesim Uncu, Ozlem Toluk, Fatma Ezgi Can*
 15.45-16.00 Combining Binary and Continuous Biomarkers - *Robab Ahmadian, İlker Ercan, Deniz Sığirli, Abdulmecit Yıldız*

Statistics and Probability - Hall 2 • Chair: Serpil Aktaş Altunay

- 15.00-15.15 The Effect of Multicollinearity on the Estimators of the Regression Coefficients - *Filiz Karadağ, Hakan Savaş Sazak*
 15.15-15.30 Artificial Neural Network Approach to Response Surface Model for Upper Limb Performance in Patients with Chronic Neck Pain - *Leyla Bakacak Karabelli, Serpil Aktaş Altunay*
 15.30-15.45 Quasi-Maximum Likelihood Estimator based on Moyal Distribution for Censored Data - *Ismail Yenilmez, İlhan Usta, Yeliz Mert Kantar*

15.45-16.00 On Using Structural Patterns In Data for Classification - *Güvenç Arslan, Bergen Karabulut, Halil Murat Ünver*

16.00-16.15 Ridge deviance residual charts for monitoring Poisson distributed data - *Ulduz Mammadova, M. Revan Özkale*

16.15-16.30 Coffee Break

October 19, Saturday • 16.30 - 17.45

Statistics and Probability - Hall 1 • Chair: Juergen Pilz

16:15-16.30M- Estimation Use Pearson Type IV Distribution Weight Function In Robust Regression - *Yasin BÜYÜKKÖR, Hatem ÇOBAN, Ali Kemal ŞEHİRLİOĞLU*

16.30-16.45 Quantification of verbal assessments using hesitant fuzzy sets: Computing with words - *Murat Alper Basaran*

16.45-17.00 Additive Gaussian Process Modeling and Novel Sparse Bayesian Regression with Applications In Business and Industry - *Juergen Pilz, Konstantin Posch, Maximilian Arbeiter*

17.00-17.15 Evaluation of The Reduction Algorithms Based on Rough Set Theory -An Application - *Yonca Yazırlı, Betül Kan Kılıç*

17.15-17.30 Parameter Estimation for the k-th Extreme Value Distribution - *Talha Arslan, Sukru Acitas, Birdal Senoglu*

17.30-17.45 Unit Lindley-Weibull Distribution and It's Some Properties - *Coşkun Kus, Kadir Karakaya, Buğra Saraçoğlu, İsmail Kinacı*

Industrial Statistics and Engineering Applications - Hall 2 • Chair: Hülya Çingir

16.30-16.45 Estimation of Cutting Tool Wear In Turning Operations via Fuzzy Logic - *Meryem Gamze Mutlu, Ceyda Arslan, Hakan Altunay*

16.45-17.00 Supply Chain Management in Dairy Industry: Mathematical Programming Approach - *Nursena Gülkaya, Ceyda Arslan, Hakan Altunay*

17.00-17.15 A Metaheuristic Algorithm For In-Plant Milk-Run System - *İslam Altın, Aydin Sipahioğlu*

17.15-17.30 A goal programming approach for the use of restricted data envelopment analysis as a tool in multi criteria decision analysis - *Esra Betül Kinacı, Harun Kinacı, Hasan Bal*

17.30-17.45 Investigation by Simple Correspondence Analysis of Regional Distribution of Hospitals and Beds in Turkey - *Ezgi Güler, Gülşen Akman, Zerrin Aladağ*

October 19, Saturday • 18.00 - 19.15

Statistics and Probability - Hall 1 • Chair: Güzin Yüksel

18.00-18.15 Comparison of Classification Algorithms on Different Data Sets - *Burcu Durmuş, Öznur İşçi Güneri, Nevin Güler Dincer*

18.15-18.30 Simpson's Paradox: Literature Review and a Dataset for the Treatment of Acne Rosacea Patients in the Muğla Region - *Burcu Durmuş, Öznur İşçi Güneri, Aslı Akın Belli*

18.30-18.45 Clustering with Genetic Algorithm: A Simulation Study - *Erkut Tekeli, Özlem Akay, Güzin Yüksel*

18.45-19.00 Reliability Analysis of Phased Mission System Using Markov Approach with Repairable Components - *Sibel Yılmaz, Özge Elmastaş Gültekin*

19.00-19.15 Predicting the Price of Real Estate Using Decision Tree Approach - *Simay Mirgen, Betül Kan Kılıç*

Statistics and Probability and Econometrics - Hall 2 • Chair: Ali Mert

18.00-18.15 A New Method Based on Interquartile Range to Feature Selection for Classification in Big Data - *Ahmet Kocatürk, Bülent Altunkaynak*

18.15-18.30 Generalized Gamma Parameters Estimation with Cuckoo Search Algorithm - *Ali Mert, Ridvan Temiz*

18.30-18.45 An Open Source Decision Tree Interface for MATLAB - *Ipek Deveci Kocakoç, Metin Öner*

18.45-19.00 Evaluation of Social Progress Performance Of European Union Countries and Turkey by Data Envelopment Analysis - *Esra Betül Kinacı, Hasan Bal, İhsan Alp*

19.00-19.15 The Impact on Student Achievement of School Volume in the PISA 2015 Turkey Sample - *Atalay Çağlar, Eda Yalçın Kayacan*

October 20, Sunday

Parallel Sessions - October 20, Sunday • 09.00 - 10.30

Statistics and Probability - Hall 1 • Chair: Hakan Demirtaş

09.00-09.15 A Fuzzy Approach for Parameters Estimation of Generalized Gamma Distribution - *Merve Dılmaç, Ridvan Temiz, Ali Mert*

09.15-09.30 Some tests and comparisons for homogeneity of variances - *Nilgün Nursu Öztürk, Hamza Gamgam, Bülent Altunkaynak*

09.30-09.45 The Performance of the VSI Tukey-Exponentially Weighted Moving Average Control Chart - *Selcem Adsız, Burcu Aytacıoğlu*

09.45-10.00 Forecasting the Number of Patients Arriving at a Hospital Emergency Department - *Aslı Kılıç, Murat Ersel*

10.00-10.15 Estimation of right-censored time-series with semi-parametric regression model - *Ersin Yılmaz, Dursun Aydın*

Statistics and Probability - Hall 2 • Chair: Güvenç Arslan

09.00-09.15 Bayesian Parameter Estimation of Akash Distribution - *İlhan Usta, Merve Akdede*

09.15-09.30 A new extended symmetry model for square contingency table - *Gökçen Altun*

09.30-09.45 The Modelling of Earthquake Events Based on Bivariate Extreme Value Theory - *Gamze Özel*

09.45-10.00 E-Bayesian Estimation for Topp-Leone Distribution - *İlhan Usta, Merve Akdede*

10.00-10.15 Evaluation of User Experience Effects on Ergonomic Behavior with using Rough-SWARA Method - *Sercan Madanlar, Şebnem Demirkol Akyol*

10.15-10.30 Time Series Analysis of Elce Prices using Box-Jenkins ARIMA Methodology in Hargelsa Somalia - *Abdışakur Ismeal Adam, Vedide Rezan Uslu*

Organization Secretariat

MOTTO
www.motto.tc

+90 232 446 06 10
info@motto.tc

Invitations

Dear Colleagues and Dear Students,

We are pleased to invite you to the 5th International Researchers, Statisticians and Young Statisticians Congress which will be held in Kuşadası Amara Sealight Elite Hotel in cooperation with Ege University Statistics Department on October 18-20, 2019.

Our aim is to bring together experts who are doing research in the field of Statistics during this two-day congress and to share our experiences in theoretical and practical fields with each other and with our students. In this context, the scientific committee of the congress is composed of scientists working in the sciences such as Statistics, Biostatistics, Industrial and Management Engineering, Econometrics, Actuarial and Data Mining, as well as experts in their fields. The language of the Congress is Turkish and English. I wish to meet you in Kuşadası, the beautiful resort of the Aegean.

Congress Chair

Prof. Dr. Onur Köksoy

Değerli Meslektaşlarımız ve Sevgili Öğrenciler,

Sizleri 18-20 Ekim 2019 tarihlerinde Kuşadası Amara Sealight Elite otelde ‘Ege Üniversitesi İstatistik Bölümü’ işbirliğinde düzenlenecek olan ‘5. Uluslararası Araştırmacılar, İstatistikçiler ve Genç İstatistikçiler’ Kongresine davet etmekten memnuniyet duyarız.

Hedefimiz bu iki günlük kongre süresince İstatistik alanında araştırma yapan uzman kişileri bir araya getirmek, teorik ve uygulama alanlarındaki tecrübelerimizi birbirimizle ve öğrencilerimizle paylaşmaktır. Bu bağlamda, kongre bilimsel komitesi İstatistik, Biyoistatistik, Endüstri ve İşletme Mühendisliği, Ekonometri, Aktüerya ve Veri Madenciliği gibi bilimlerde çalışan ve aynı zamanda alanlarında uzman bilim insanlarından teşkil edilmiştir. Kongre dili Türkçe ve İngilizce olarak belirlenmiştir. Ege’nin güzel beldesi Kuşadası’nda buluşmak dileğiyle saygılar sunarım.

Kongre Başkanı

Prof. Dr. Onur Köksoy

Organizing Committee

HONORARY COMMITTEE

Necdet Budak (Rector)

Canan Fisun Abay (Vice Rector)

İhsan Yaşa (Dean)

Onur Baskan (Founder of Statistics Department)

CHAIR

Onur Köksoy

VICE-CHAIRS

Halil Tanıl

Sevcan Demir Atalay

Melis Zeybek

ORGANIZING COMMITTEE

Onur Köksoy

Hayal Boyacıoğlu

Sevcan Demir Atalay

Ali Mert

Hakan Savaş Sazak

Halil Tanıl

Özge Elmastaş Gültekin

Funda İşçioğlu

Burcu Aytaçoğlu

Melis Zeybek

Emine Çetingöz

Aslı Kılıç

Agah Kozan

Gözde Kuş

Elif Kozan

Filiz Karadağ

Scientific Committee

INTERNATIONAL SCIENTIFIC COMMITTEE	
NAME/SURNAME	COMPANY
Ali Allahverdi	Kuwait University, KUWAIT
Anna Crisci	Università Federico II Napoli, ITALY
Antonello D'Ambra	University of Campania "Luigi Vanvitelli", ITALY
Atanu Bhattacharjee	Centre for Cancer Epidemiology Tata Memorial, INDIA
Biagio Simonetti	Università degli Studi del Sannio, ITALY
Claudiu Herteliu	Bucharest University of Economic Studies, ROMANIA
Eric Beh	The University of New Castle, AUSTRALIA
Fatmir Memaj	University of Tirana, ALBANIA
Hakan Demirtas	University of Illinois, Chicago, USA
Loon Ching Tang	National University of Singapore, SINGAPORE
Luigi D'Ambra	Università Federico II Napoli, ITALY
Michael Greenacre	Universitat Pompeu Fabra, SPAIN
Necip Doganaksoy	Siena College, USA
Nuno Costa	Instituto Politecnico de Setubal, PORTUGAL
Pietro Amenta	Università degli Studi del Sannio, ITALY
Rama Shanker	Devendra Mishra Institute of Statistics, INDIA
Rubén Ruiz	Universitat Politècnica de València, SPAIN
Sangmun Shin	Dong-A University, SOUTH KOREA
Shu-Kai S. Fan	National Taipei University of Technology, TAIWAN
Tatjana Sibalija	Belgrade Metropolitan University, Belgrade, SERBIA
Timothy J. Robinson	University of Wyoming, USA
Yongtao Cao	Indiana University of Pennsylvania, USA
Zhanpan Zhang	General Electric Global Research, USA

NATIONAL SCIENTIFIC COMMITTEE	
NAME/SURNAME	COMPANY
Adil Baykasođlu	Dokuz Eylöl University
Ali İhsan Genç	Çukurova University
Aşır Genç	Necmettin Erbakan University
Atalay Çađlar	Pamukkale University
Atila Göktaş	Muđla Sıtkı Koçman University
Aylin Alın	Dokuz Eylöl University
Bayram Şahin	Ege University
Berna Yazıcı	Eskişehir Technical University
Birdal Şenođlu	Ankara University
Buđra Saraçođlu	Selçuk University
Burcu Üçer	Dokuz Eylöl University
Cem Kadılar	Hacettepe University
Cemil Çolak	İnönü University
Çađdaş Hakan Aladađ	Hacettepe Üniversitesi
Coşkun Kuş	Selçuk University
Deniz Sıđırlı	Uludađ University
Dursun Aydın	Muđla Sıtkı Koçman University
Efendi Nasibov (Nasibođlu)	Dokuz Eylöl University
Ergun Karaađaođlu	Hacettepe University
Erol Eđriođlu	Giresun University
Esin Firuzan	Dokuz Eylöl University
Fatma Zehra Muluk	Başkent University
Femin Yalçın	İzmir Katip Çelebi University
Filiz Karaman	Yıldız Technical University
Gökhan Ocakođlu	Uludađ University

Gözde Yazgı Tütüncü	İzmir Ekonomi University
Güçkan Yapar	Dokuz Eylül University
Gülay Başarır	Mimar Sinan Güzel Sanatlar University
Gülhayat Gölbaşı Şimşek	Yıldız Technical University
Gülser Köksal	Odtü
Güvenç Arslan	Kırıkkale University
Güzin Yüksel	Çukurova University
Hakan Altunay	Süleyman Demirel University
Hale Köksoy	Selçuk University
H. Kıvanç Aksoy	Eskişehir Osmangazi University
Hamza Gamgam	Gazi University
Hasan Bal	Gazi University
Hülya Bayrak	Gazi University
Hülya Çıngı	Hacettepe University
Hüseyin Tatlıdil	Hacettepe University
İlker Ercan	Uludağ University
İnci Batmaz	ODTÜ
İpek Deveci Kocakoç	Dokuz Eylül University
İsmihan Bayramoğlu	İzmir Ekonomi University
M. Akif Bakır	Gazi University
Mahmude Revan Özkale Atıcıoğlu	Çukurova University
Mahmut Ali Gökçe	Yaşar University
Mehmet Aksaraylı	Dokuz Eylül University
Mehmet Mert	Akdeniz University
Mehmet N. Orman	Ege University
Meral Sucu	Hacettepe University
Murat Alper Başaran	Alanya Keykubat University
Murat Caner Testik	Hacettepe University
Müjgan Tez	Marmara University

Nimet Yapıcı Pehlivan	Selçuk University
Olca Arslan	Ankara University
Özge Akkuş	Muğla Sıtkı Koçman University
Özgür Yeniay	Hacettepe University
Özlem Ege Oruç	Dokuz Eylül University
Sadullah Sakallıoğlu	Çukurova University
Selahattin Kaçıranlar	Çukurova University
Selma Gürler	Dokuz Eylül University
Serdar Demir	Muğla Sıtkı Koçman University
Serkan Eryılmaz	Atılım University
Serpil Aktaş Altunay	Hacettepe University
Sevgi Yurt Öncel	Kırıkkale University
Şanslı Şenol	Ege University
Timur Köse	Ege University
Türkan Erbay Dalkılıç	Karadeniz Technical University
Veysel Yılmaz	Eskişehir Osmangazi University
Yeliz Mert Kantar	Eskişehir Technical University
Zafer Küçük	Karadeniz Technical University

*** In Alphabetical Order by Name**

Main Topics

BAYESIAN STATISTICS
BIOSTATISTICS
CATEGORICAL DATA ANALYSIS
CHEMOMETRICS
COMPUTATIONAL STATISTICS
DATA MINING
DECISION THEORY
DESIGN OF EXPERIMENTS
ECOLOGICAL & ENVIRONMENTAL STATISTICS
ECONOMETRICS
ESTIMATION
FINANCIAL MATHEMATICS
FUZZY THEORY AND APPLICATIONS
HEURISTIC ALGORITHMS
INDUSTRIAL STATISTICS
INFORMATION TECHNOLOGY
MULTIVARIATE STATISTICS
OFFICIAL STATISTICS
OPERATIONAL RESEARCH
OPTIMIZATION THEORY AND APPLICATIONS
ORDERED STATISTICAL DATA ANALYSIS
PROBABILITY THEORY
QUALITY ENGINEERING
QUEUEING THEORY
REGRESSION ANALYSIS
RELIABILITY THEORY AND ANALYSIS
RESPONSE SURFACE METHODOLOGY
RISK ANALYSIS
ROBUST PARAMETER DESIGN
ROBUST STATISTICS
SAMPLING METHODS
SIGNAL PROCESSING
SIMULATION METHODS
SPATIAL STATISTICS
STATISTICAL APPLICATIONS IN SOCIAL SCIENCES
STATISTICAL QUALITY CONTROL
STOCHASTIC PROCESSES
SURVIVAL ANALYSIS
TIME SERIES ANALYSIS
OTHERS

KEYNOTE LECTURES

Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation

Luigi D’Ambra^{1*}, Pietro Amenta² and Antonello D’Ambra³

¹*Department of Economics, management and institutions, Italy, dambra@unina.it*

²*Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy, amenta@unisannio.it*

³*Department of Economics, University of Campania “Luigi Vanvitelli”, Italy, antonello.dambra@unicampania.it*

(Corresponding author should have an asterisk sign () if possible.)*

Abstract – It is well known that the Pearson statistic χ^2 can perform poorly in studying the association between ordinal categorical variables. Taguchi’s and Hirotsu’s statistics have been introduced in the literature as simple alternatives to Pearson’s chi-squared test for contingency tables with ordered categorical variables. The aim of this paper is to shed new light on these statistics, stressing their interpretations and characteristics, providing in this way new and different interpretations of these statistics. Moreover, a theoretical scheme is developed showing the links between the different proposals and classes of cumulative chi-squared statistical tests, starting from a unifying index of heterogeneity, unlikability and variability measures. Users of statistics may find it attractive to understand well the different proposals. Some decompositions of both statistics are also highlighted. This paper presents a case study of optimizing the polysilicon deposition process in a very large-scale integrated circuit, to identify the optimal combination of factor levels. It is obtained by means of the information coming from a correspondence analysis based on Taguchi’s statistic and regression models for binary dependent variables. A new optimal combination of factor levels is obtained, different from many others proposed in the literature for this data.

Keywords – *Contingency table · Chi-squared statistic · Single and double cumulative chi-squared statistics · Likelihood ratio*

References

Agresti A (2013) *Categorical data analysis*, 3rd edn. Wiley, New York, US.

D’Ambra L., Beh E.J., Camminatiello I. (2014). “Cumulative correspondence analysis of two-way ordinal contingency tables”, *Communication Statistics-Theory and Methods* 43(6):1099–1113

D’Ambra L., Köksoy O., Simonetti B. (2009). “Cumulative correspondence analysis of ordered categorical data from industrial experiments”, *Journal of Applied Statistics* 36(12):1315–1328

Goodman L.A., Kruskal W.H. (1954). “Measures of association for cross-classifications”, *Journal of the American Statistical Association* 49:732–764

Hirotsu C. (1986). “Cumulative chi-squared statistic as a tool for testing goodness of fit”, *Biometrika* 73:165–173

Light R., Margolin B. (1971). “An analysis of variance for categorical data” *Journal of the American Statistical Association* 66(335):534–544

Nair V.N. (1986). “Testing in industrial experiments with ordered categorical data”, *Technometrics* 28(4):283–291

Satterthwaite F. (1946). “An approximate distribution of estimates of variance components”, *Biometrics Bulletin* 2:110–114

Taguchi G. (1974). “A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test”. *Saishin Igaku* 29:806–813

Baby Conception between Religiousness and Pleasure in Romania. Comparing Eastern Orthodox and Non-Orthodox Populations through Very Long Daily Time Series (1905-2001) Analysis

Claudiu Herteliu¹, Bogdan Vasile Ileanu¹, Marcel Ausloos^{1,2}, Giulia Rotundo³

¹Department of Statistics and Econometrics, Bucharest University of Economic Studies, Romania; Calea Dorobantilor 15-17, Bucharest, 010552 Sector 1, Romania

²School of Business, University of Leicester, Leicester; LE2 1RQ, United Kingdom

³Mathematical Methods of Economics, Finance and Actuarial Sciences, Sapienza University of Rome, Rome, Italy; Piazzale Aldo Moro, 5, 00185 Roma RM, Italy

E-mails: Claudiu Herteliu hertz@csie.ase.ro, Bogdan Vasile Ileanu ileanub@yahoo.com, Marcel Ausloos marcel.ausloos@uliege.be, Giulia Rotundo giulia.rotundo@live.com

Abstract

Eastern Orthodox tradition bans sexual intercourse during Nativity and Lent fasting periods. When one tests the effect of this interdiction during the 20th century, it seems that Lent is obeyed in a greater extent than the Nativity. Fertility (births and subsequently their conceptions) shows seasonality all around the world. There are geographical factors (latitude, weather, day-length), demographic, economic and socio-cultural characteristics (education, ethnicity, religion) that have been proven to affect babies' conception. Censuses (1992 and 2002) data for daily births for 97 years (35 429 points) is taken into consideration. Based on the reported birthday of each person, the estimated time of conception was computed using standard gestation duration. The population has been grouped into two categories (Eastern Orthodox and Non-Orthodox) based on religious affiliation. Data analysis was performed in a comparative manner for both groups. Religion affiliation acts as an important factor; the data analysis indicates smaller error bars on the parameters for the Eastern Orthodox group as compared to the Non-Orthodox one. The models are tested for validity using F test (ANOVA) while the regression coefficients are tested by the Student t-test. All models are statistically valid (p value less than 0.01) and all (except for rurality between 1990 and 2001 where the p value is less than 0.05) of the regression coefficients for the Eastern Orthodox group are valid to a large extent (p value less than 0.01).

Hybrid data generation

Hakan Demirtaş

University of Illinois at Chicago, School of Public Health, 1603 West Taylor Street, Room 950

Chicago, IL, 60612-4336, U.S.A., phone: 312-996-9841 fax:312-996-0064

email: demirtas@uic.edu

<http://demirtas.people.uic.edu/DEMIRTAS-CV.pdf>

<http://www.healthstats.org/members/hdemirtas.html>

Abstract: This talk is concerned with formulating and implementing a unified framework for generating data sets that include all four major types of variables (i.e., binary, ordinal, count, and continuous) when the marginal distributions and a feasible association structure are specified for simulation purposes. The final outcome is designed to be a public-domain, user-friendly software tool that can be employed in many different data-analytic contexts. Practitioners and methodologists across many disciplines in medical, managerial, social, biobehavioral, and physical sciences will be able to simulate multivariate data of hybrid types with relative ease. The proposed work can serve as a milestone for the development of more sophisticated simulation, computation, and data analysis techniques in the digital information, massive data era. As research efforts are unequivocally evolving from modeling the total and absolute truth to deciphering the empirical truth, from small data to big data, from mathematical perfection to reasonable approximation to reality, and from exact solutions to simulation-driven solutions in modern times, this work is a timely and needed step forward for substantially augmenting the range of problems that the statistical simulation paradigm can effectively address. Capability of jointly generating many variables of different distributional types, nature, and dependence structures may become an influential contributing factor for better comprehending the operational attributes of today's intensive data trends. Overall, this talk will provide the salient characteristics of a principled, practical, comprehensive, and broadly applicable set of computational tools for simulating data, whose generality and flexibility offer promising potential for building enhanced statistical computing infrastructure for research and education.

Keywords: Random number generation, stochastic simulation, mixed data, discretization, correlation mapping

Statistical vs. Metaheuristic Techniques in Parametric Optimisation of Industrial Processes

Tatjana V. Šibalija^{1, *}

¹ *Faculty of Management, Faculty of Information Technology, Belgrade Metropolitan University, Tadeuša Košćuška 63, 11 000, Belgrade, Serbia*

**Corresponding author:*

tsibalija@gmail.com; tatjana.sibalija@metropolitan.ac.rs

Abstract: The parametric optimisation, i.e. the process parameter design, is one of the essential issues in setting up and managing industrial processes nowadays, especially in terms of responding on rapidly changing circumstances in a globalised, dynamic industrial environment. The goal of process parameters design is to find such a setting of the process control factors that meets requirements for the mean value of process responses and minimise their variation simultaneously. The problem becomes more complex when process is characterised by multiple output characteristics, i.e. responses, which are typically correlated. Therefore, the methods used for process parameter optimisation are of utmost importance for improving quality of industrial processes today. Over the years, the optimisation methods have undergone substantial development and expansion to address rapid changes especially for new, emerging processes. This paper discusses the most frequently used conventional, i.e. statistical methods (e.g. Taguchi robust parameter design and its modifications; response surface methodology), and non-conventional methods based on metaheuristic search algorithms (e.g. genetic algorithm; simulated annealing algorithm; particle swarm optimisation; ant colony optimisation) for process parameter design. The implementation of both types of optimisation methods was critically appraised, in terms of their peculiarities, benefits, shortcomings and applicability for certain type of problems. Besides, the issues in modelling input-output process relationships are also addressed since they are inevitably linked with the process optimisation.

Key Words: process parameter optimisation; Taguchi method; response surface methodology (RMS); metaheuristic algorithms; genetic algorithm (GA); simulated annealing (SA); particle swarm optimisation (PSO); ant colony optimisation (ACO)

ORAL ABSTRACT

O-01 Evaluation of the Performance of Boosting and Bagging Classification Algorithms after Preprocessing in Health Data

Oral Abstract / Biostatistics

Yuksel Ozkan¹, Asli Suner¹,

¹Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, İzmir, Turkey,

Classification is a part of health studies that provides recognition and examination of diseases. Supervised machine learning methods using artificial intelligence are generally preferred in deciding how to break complex data into meaningful pieces and revealing medical information. Ensemble learning methods, called multiple classifier learning systems, provide more successful models by training multiple learners at the same time to solve the same problem. In this study, after the data pre-processing to solve possible problems such as missing value, class noise and class imbalance; we aimed to compare the performance of classification algorithms used for accurate diagnosis in health data. In this study, while random forest and weighted subspace random forest were used as bagging algorithms; additive logistic regression and gradient boosted machines algorithms were used as boosting algorithms. Eight disease diagnosis data sets with different data properties and sample sizes obtained from KEEL database, and were pre-processed. Accuracy, sensitivity, specificity, precision, Kappa statistic, Youden index, F - measure and ROC measurement metrics were used for performance comparison and, run times of algorithms were calculated. All analyzes were performed with R version 3.3.0. We observed that performance success of algorithms increased after data preprocessing. In general, performance of boosting algorithms yielded higher results than bagging algorithms. However, boosting algorithms had the longest running algorithms. As a result, data preprocessing process should not be ignored if high performance success is targeted by researchers.

KEYWORDS: Bagging, Boosting, Health Data, Preprocessing, Ensemble Learning Methods

O-02 Quantifying Opinion about a logistic regression Modelling for Fauna Habitat Distributions

Oral Abstract / Statistics & Probability

shafiqah Alawadhi¹,

¹Kuwait University,

An implementation of the method will be done through an example about Fauna Habitat Distributions. A method of quantifying opinion about a logistic regression model is proposed that can be used with large models. The method may also be used to elicit opinion about the regression coefficients of any generalized linear model (GLM). Subjective opinion is quantified through an interactive visualize computer program that asks the expert to perform various assessment tasks. The elicitation tasks are chosen so that a prior distribution for a logistic can be determined from the assessments. Then formula for using the elicited assessments from expert to estimate the parameters of the prior distribution is derived. When data becomes available, combining it with the prior distribution is intractable analytically but easily handled using a Markov chain Monte

KEYWORDS: Expert opinion, Assessment Task, Logistic learner model. Environmental experiment.

O-03 A Fuzzy Approach for Parameters Estimation of Generalized Gamma Distribution

Oral Abstract / Statistics & Probability

Merve Dilmaç¹, Rıdvan Temiz¹, Ali Mert¹,

¹Ege University, Department of Statistics,

The Generalized Gamma Distribution (GGD) is a very important distribution and occupies a great area in application fields. One of the reasons for the importance is that with respect to special values of the 3 parameters of the GGD, it can be transformed into various statistical distributions. Also, GGD is applicable to many real-life problems. The main purpose of the study is to estimate the parameter values of the GGD based on the data sets at hand. Maximum Likelihood Estimation Method was chosen. The motivation of the method is to maximize the Likelihood Function (LF) subject to parameters of the GGD. In the literature, there are various approaches in order to maximize the LF of GGD because of the complex structure of the function. Heuristic optimization algorithms are one of the approaches. The algorithms are derived from the behavior of physical or biological systems in nature. In this study, the Bat Algorithm which is inspired by the hunting behavior of bats is employed. Because of the complex structure of the LF, results of the Bat Algorithm sometimes are not satisfactory. At this point, we injected Fuzzy Logic in the Bat Algorithm in order to obtain more reasonable results. We described the parameters of the distribution as fuzzy coefficients of fuzzy LF of GGD. From now on, the maximization problem becomes a fuzzy maximization problem. In the study, we ran simulations in order to compare performances of Genetic, Cuckoo, Bat and Fuzzy Bat Algorithms with respect to LF of GGD.

KEYWORDS: Generalized Gamma Distribution, Maximum Likelihood, Bat Algorithm, Fuzzy Logic

O-04 Investigation by Simple Correspondence Analysis of Regional Distribution of Hospitals and Beds in Turkey

Oral Abstract / Industrial Statistics & Engineering Applications

Ezgi Güler¹, Gülşen Akman², Zerrin Aladağ²,

¹Bilecik Şeyh Edebali University, ²Kocaeli University,

Correspondence Analysis is a technique that allows interpretation for categorical variables, facilitating the interpretation of the relationships or similarities and differences between row and column variables in cross tables, and illustrating these changes in a graphical dimension. It is a widely used method with many different analyzes in areas where categorical data analysis is frequently used such as health sciences, economics and social sciences. The main objective of the method, which has two types; simple and multiple correspondence analysis; to show the relationship between variables and categories of variables graphically and to reduce the size of the cross-tables with simple factors to obtain this representation. Basic investments in the health sector require a relational analysis of the current situation. In this study, the number of hospitals and beds of all regions in 2017 was obtained from the most recently published TURKSTAT (TUİK) database and Simple Correspondence Analysis was performed on the data. In the quantitative sense, which hospital types are compatible with which regions and the reasons for this situation were interpreted. As a sub-investigation, a separate simple application analysis was conducted in order to examine the distribution of patient beds according to regions divided into hospital categories. SPSS software was used in all analyzes. The finding of the analysis were interpreted comparatively.

KEYWORDS: Correspondence Analysis, Categorical Data Analysis, Health

O-05 A Regression Analysis for Predicting the Amount of Yearly Greenhouse Gas Emissions in Turkey

Oral Abstract / Statistics & Probability

Merve Avcı¹, Banu Yetkin Ekren¹,

¹Yaşar University,

In this paper, we study a stepwise multiple linear regression analysis in order to predict the amount of yearly greenhouse gas (GHG) emissions based on weather related data in Turkey. Greenhouse gas is a compound of gaseous capable of absorbing infrared radiation, resulting with trapping and holding heat in the atmosphere. By the increased heat in the atmosphere, greenhouse gases cause greenhouse effect, which ultimately leads to global warming. By the presented study, we both investigate the statistically significant factors affecting the yearly GHG emission amount and the existence of a proper regression function that is able to predict the yearly GHG emission amount accurately based on those significant factors. Thus, while the output is considered to be the yearly GHG emission amount, the input factors are considered to be: annual average weather temperature (0C), relative humidity (%), hours of sunshine (h), average sea water temperature (0C), and rainfall (mm/m2). By those input factors, we also search whether or not there is relationship of yearly GHG emission amount on climate change. We use real data for the analysis which are obtained from the Turkish Statistical Institute's web site as well as Turkish state Meteorological Service for the years of 1990 and 2017. The results show that there is significant effect of GHG emission on climate change and there exist a good fit regression function estimating the yearly GHG emission amount based on those climate-based input factors.

KEYWORDS: Regression, greenhouse gas emission, stepwise regression, climate

O-06 Investigating Chaotic Dependence between Economic Growth and Energy Consumption

Oral Abstract / Econometrics

Aygül Anavatan¹,

¹Pamukkale University,

The relationship between energy consumption and economic growth releases information about the role of energy consumption in economic development. The aim of this study is to identify the nature of the dependence or causal relationship that exists between economic growth and energy consumption in Turkey. It was used the recent methods of linear cointegration, and non-linear Granger causality in this study. Additionally, it was constructed a noisy chaotic multivariate model, using the bivariate noisy Mackey–Glass process, in order to reveal the complex dynamics that cause the specific character of the observed causality and feedback.

KEYWORDS: Non-linear causality, Dynamic non-linearity, Economic growth, Energy consumption, Bivariate noisy Mackey–Glass model

O-07 The Impact of Primary Production of Renewable Energy on Labor Force: EU-28 Panel Data Analysis

Oral Abstract / Econometrics

Selena Kantarmacı¹, Şenay Üçdoğruk Birecikli²,

¹ Social Sciences Institute, ²Faculty of Economics and Administrative Sciences,

The increase in the use of non-renewable energy sources in the world leads to an increase in prices, an increase in emission problems and depletion of resources. As a result, countries have turned to the use of renewable energy sources. The effect of "fuel of the future" on economic growth or labor force or another development criterion have been investigated by most studies, but primary production has not been considered. This study seeks to examine relationships between primary production of renewable energy and labor force. We employed the Fully Modified Ordinary Least Square (FMOLS) regression model to sample of EU-28 countries for the period 2006-2016 by using panel data analysis. Our main empirical findings reveal that primary production of renewable energy is associated with a positive and statistically significant impact on labor force in EU-28 for the period 2006–2016. These findings are important for the development of energy policies and employment issues. Also reveal contribution of primary production of renewable energy to employment in the future.

KEYWORDS: Renewable Energy, Primary Production, EU-28, Panel Data Analysis, Labor Force

O-08 A Panel Data Analysis: Unemployment The Relationship Between Young Unemployment and Growth

Oral Abstract / Econometrics

merve altaylar¹,

¹Social Science İnstu,

A Panel Data Analysis: Unemployment The Relationship Between Young Unemployment and Growth
MERVE ALTAYLAR¹*, HAMDİ EMEÇ²* Social Sciences Institute, Dokuz Eylül University, Turkey
mervealtaylar37@gmail.com Faculty Of Economics And Administrative Sciences, Dokuz Eylül
University, Turkey hamdi.emec@deu.edu.tr
ABSTRACT In this paper examines youth unemployment and economic growth in unemployment between 24 and 20 OECD countries between 2000 and 2017. These variables were analyzed using panel time series techniques unemployment, youth unemployment and economic growth are examined with static and dynamic models. The causality between variables was also investigated. It does not examine the relationship between these variables just as Okun suggests and also explores how unemployment and youth unemployment affect economic growth. The traditional panel unit root test and panel cointegration tests as well as the structural breakage panel unit root and panel cointegration tests were used to investigate the diversity of these variables. as a result, youth unemployment is more sensitive to economic growth than unemployment,at the same time, improvements in youth unemployment do not affect economic growth as much as improvements in unemployment.
Keywords: structural break, dynamic, panel, causality

KEYWORDS: structural break, dynamic, panel, causality

O-09 Inflation Targeting and Taylor Rule Model for Developed And Developing Countries: A Panel Data Analysis

Oral Abstract / Econometrics

Hande Erk¹

¹Social Science Institute,

Since its introduction in 1990, the inflation targeting regime has become a Monetary policy strategy adopted by many developed and developing countries. Aiming the price stability and targeting nominal interest rates in this direction, this strategy is expected to have stronger effects on the expectation of Monetary policy when a rule based approach is followed. The purpose of this study is to test policy interest rates in the inflation targeting regime, based on the rule based approach with The Taylor's Rule and to provide a comparison to developed – developing countries. In this study, the macroeconomic variables which are effective in determining the policy interest rates are investigated in Taylor Rule for 10 countries of 2008Q1-2018Q4 period, 5 of which are developed and 5 of which are developing countries. Developed countries are modelled With original Taylor Rule and developing countries are modelled With expanded Taylor Rule. “ Panel Data Analysis” technique was used to determine the macroeconomic variables that are thought to have an impact on policy interest rate. According to the findings, While real interest rate, inflation rate and Output gap variables are effective in determining policy in developed countries, real interest rate, inflation gap and output gap variables are effective in determining policy in developing countries.

KEYWORDS: Panel Data Analysis, Taylor Rule, Inflation targeting Regime

O-10 The Impact on Student Achievement of School Volume in the PISA 2015 Turkey Sample

Oral Abstract / Econometrics

Atalay Çağlar¹, Eda Yalçın Kayacan¹,

¹Pamukkale University,

The Programme for International Student Assessment (PISA) is a three-year survey conducted by the Organization for Economic Cooperation and Development (OECD) to assess the knowledge and skills of 15-year-old students in the field of reading, science and mathematics. PISA, by providing the indicators of education, shows the participant countries what level they are in the field of education at international level and enables them to identify the deficiencies and issues that need to be improved in the educational systems of the countries. The question of whether school volume has a significant impact on student achievement has been the subject of many studies. Although this is the main question that is tried to be answered in the study; it is also aimed to answer the question of what is the optimal school volume which has an effect on student achievement. For the purposes mentioned, the impact on student achievement of school volume were analyzed by multilevel model, which allow to consider the hierarchical structure of the data being the students nested within schools, using PISA 2015 data for Turkey. Also, in the study; the students' achievement in reading, science and mathematics was evaluated separately and the effect of variables such as student characteristics, school type and regional differences which were thought to have an effect on achievement were also taken into consideration.

KEYWORDS: PISA 2015, School Size, Student Achievement, Multilevel Model

O-11 A new parametric test for detecting linear trend in location

Oral Abstract / Statistics & Probability

Halil Tanil¹, Agah Koza¹,

¹Ege University,

In this study, we present a new parametric test for detecting a linear trend in location under the assumption of normality. We use a similar approach of weighting observations while constructing a test statistic as in Brillinger (1989) and Balakrishnan and Tan (2016). We consider these tests while making empirical power comparison via Monte Carlo simulations for different sample sizes of $n=10,30,50$, and 100 at $\alpha=0.05$ level of significance. Also we discuss the real life applicability of the proposed test.

KEYWORDS: Parametric test, Linear trends in location, Empirical power

O-12 How is the performance of the Mc-Nemar test to determine cut-off comparing to the Youden index and the minP methods for ordinal data? A simulation study

Oral Abstract / Biostatistics

Pervin Demir¹, Afra Alkan¹, Selcen Yüksel¹,

¹Ankara Yıldırım Beyazıt Üniversitesi,

It is important to determine the optimal cut-off point that differentiates the patients and the healthy individuals for diagnostic tests with a continuous or an ordinal response. In this study, we proposed the Mc-Nemar test, which considers the association between two dependent categorical variables to estimate the optimal cut-off point for the ordinal response test with five point results. We evaluated the performance of this test statistics by a simulation design with considering the sample size and the balance of groups as simulation conditions and compared it to the Youden index method, which is the most applied method in the diagnostic studies, and the minimum P-value approach, which uses the chi-square test of independence to get optimal cut-off. The sample sizes were set 50, 100 and 200 per group in the balanced design and (50, 100), (50, 150) and (50, 200) for the diseased and non-diseased groups in the unbalanced design. For each scenario, 1000 MCMC repeats were generated. The range of bias was the largest in the Youden index method and the narrowest in the Mc-Nemar test in general and within all scenarios. The median bias was 0 for the Mc-Nemar test and 1 for the other methods. The proportion of unbiased estimation was 46.0%, 40.0% and 73.9% for the Youden index method, the minimum P-value approach, and the Mc-Nemar test, respectively. The proportion of unbiased estimation in the Mc-Nemar test was higher in the unbalanced design compared to the balanced design.

KEYWORDS: ordinal data, optimal cut-off, Youden index, minimum P-value, McNemar test

**O-13 Evaluation of Social Progress Performance Of European Union Countries and Turkey by
Data Envelopment Analysis**

Oral Abstract / Statistics & Probability

Esra Betül KINACI¹, Hasan BAL¹, İhsan ALP^{1, 2},

¹Gazi University,

States and societies always tend to increase the level of development of their country. In general, the growth in the economies of the countries constitutes the perception that there will be an increase in the level of development of the society. The economy will have an absolute impact on the level of development of society. However, it would not be right to make such a judgment considering only the economy so many other factors have to be considered. These factors include human development, the ability to meet the basic human needs of citizens, and the opportunities that enable citizens to improve and maintain their quality of life by creating space for themselves. In order to measure the development levels of societies, many variables that affect this structure evaluate together. Various institutions and organizations evaluate the level of development of countries using different methods. One of the studies to examine this complex structure is the Social Progress Index (SPI). This index is created taking into account measurable factors in many areas, including housing, nutrition, rights and education. Thus the index assess social progress performance levels of countries. The aim of this study is to evaluate Performance of Social Development the European Union countries and Turkey in a different method, taking into account variables used in SPI. Data envelopment analysis (DEA), which is a linear programming based method, was used in the study. The results obtained from the analysis were compared with SPI and the similarities were evaluated with statistical methods

KEYWORDS: Data Envelopment Analysis, Social Progress Index, Efficiency

O-14 The Performance Comparison of Feature Selection Methods in Health Datasets

Oral Abstract / Biostatistics

Mert Demirarslan¹, Ash Suner¹,

¹Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, Izmir, Turkey,

Nowadays, since data sets become very high-dimensional and specific with the data collected from different devices, attribute selection has an important pre-processing task in reducing data size in data mining. This study aims to improve classification performance by reducing the calculation time and cost by using attribute selection methods. Attribute selection methods are examined under three main headings; filter method, wrapper method, and embedded method. In the study, support vector machine, naive bayes and decision trees methods (J48) among the machine learning classification algorithms were used. Parkinson's disease, colorectal cancer, SCADI (Self-Care Activities Dataset based on ICF-CY), hepatocellular carcinoma (HCC) and breast cancer data sets were obtained from UCI and Kaggle databases. Accuracy values were calculated to compare the classification performances of the algorithms. WEKA version 3.8.3 and R3.3.0 programs were performed in all analyzes. After unnecessary features were extracted by using appropriate methods in the analysis; classification performances and run times of algorithms were calculated. Accuracy values increased to 87% for colorectal cancer, 85% for Parkinson's disease, 97% for SCADI, 100% for HCC, and 78% for breast cancer after attribute selection. The algorithm with the highest performance was found as a wrapper method with decision trees (J48). While the fastest algorithm was filter method, the longest-running algorithm was the wrapper method. According to results, the performance improvement was higher in feature sets with a large number of attributes after selecting feature. As a result, low dimensional data sets can provide higher classification accuracy with lower calculation costs.

KEYWORDS: Feature Selection Methods, Data Mining, Machine Learning, Health Dataset

O-15 Performances of the distribution function estimators based on ranked set sampling using body fat data

Oral Abstract / Biostatistics

Yusuf Can Sevil¹, Tuğba Yıldız¹,

¹Dokuz Eylül University,

In this study, we illustrate an application of empirical distribution function (EDF) estimators based on ranked set sampling (RSS) that are suggested by Yıldız & Sevil (2018, 2019) using real data set (body fat data). The body fat data for 252 men collected in 1985. In this application, we use three variables which are percentage of body fat (X), abdomen circumference (Y1) and age (Y2). Our target parameter is the distribution function of percentage of body fat. Age and abdomen circumference are separately used in ranking process as auxiliary variables which have correlation 0.813 (for perfect ranking) and 0.291 (for imperfect ranking) with the percentage of body fat, respectively. Ranked set samples are constructed by using three different sampling designs which are level-0 level-1 and level-2 (Deshpande et al., 2006). The effects of both perfect and imperfect ranking on the estimators of the sampling designs are investigated. Relative efficiencies of the EDF estimators are obtained by using their integrated mean squared errors (IMSE), numerically. For both perfect and imperfect ranking, these EDF estimators based on sampling designs have outperformance against EDF estimator based on SRS.

KEYWORDS: Ranked set sampling, sampling designs, empirical distribution function, integrated mean squared errors

O-16 Ranking Error Models, Cost and Optimal Set Size in Ranked Set Sampling

Oral Abstract / Statistics & Probability

Sami Akdeniz¹, Tuğba Yıldız¹,

¹Dokuz Eylül University,

Ranked Set Sampling (RSS) is a sampling method commonly used in recent years. RSS is developed by McIntyre (1952) as an alternative to Simple Random Sampling (SRS) in order to estimate population parameter more efficiently where the measurement of sampling units is difficult or costly but the units are easier to rank. There are several factors that make this method useful especially for studies in medicine, agriculture, forestry and ecology. The most important of these factors are the set size and the relative costs of some operations such as sampling, measurement and ranking. Ranking of the units in the set is made on the basis of the visual judgment of the researcher or a concomitant variable which has a strong correlation with the variable of interest. These ranking methods are defined as Ranking Error Models. In this study, the widely used cost and ranking error models in RSS literature are investigated. Besides, it is aimed to examine whether RSS is cost effective with respect to SRS in terms of mean squared error of the mean estimator considering ranking error models and the N-KPST cost model in infinite population. Then based on these results, the optimal set size for RSS is determined. A Monte Carlo simulation study is conducted for this purpose. Additionally, this study is supported by real life data.

KEYWORDS: Ranked set sampling, ranking error models, cost, optimal set size, abalone data set

O-17 Predicting the Price of Real Estate Using Decision Tree Approach

Oral Abstract / Statistics & Probability

Simay MİRGEN¹, Betul KAN-KILINC²,

¹Institute of Graduate Programs, Department of Statistics, Eskisehir Technical University, Turkey, ²Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey,

In this study, the relationship between the real estate and its properties are investigated by using a decision tree that is built from the training set including 70% of dataset. Using a public housing platform as a case study, the real estate for sale in Eskisehir have been collected. The existence of some real estate characteristics is recorded as 1 and 0 elsewhere. The dependent variable is the unit price of real estate for 280 observations that is classified in three categories, such as cheap, moderate and expensive. For model validation 5-fold cross validation is used and evaluation metrics are summarized for both train and test data. Addition, the built tree model identified the most significant characteristics of real estate in determining the unit price.

KEYWORDS: Splitting, real estate, tree algorithm

O-18 Anxiety and Attitudes towards Biostatistics and Scientific Research Methods Courses of Students of a Dental School

Oral Abstract / Biostatistics

Adnan Karaibrahimoğlu¹, Nazan Karaoğlu², Said Karabekiroğlu³,

¹Süleyman Demirel University, Faculty of Medicine, Biostatistics Dept , ²Necmettin Erbakan University, Faculty of Medicine, Family Practice Dept, ³Necmettin Erbakan University, Faculty of Dentistry, Restorative Dentistry Dept,

Attitude and anxiety are two important terms related with the psychology. However, they are often used in education. College students sometimes feel themselves anxious and develop an attitude for some courses because of many reasons. Statistics and scientific research methods courses can be two examples of such reasons. Unlike natural sciences and engineering, the students in health sciences like medicine, dental, pharmacy or veterinary schools are reported to have an attitude and anxiety towards statistics since they are not familiar with the mathematical ability. The aim of this study is to determine the anxiety and attitude level of students in a dental school during three semesters consecutively. After approval of the ethics committee, two separate scale which are Statistics Attitude Scale (SAS) and Scientific Research Methods Attitude Scale (SRMAS) with shown the reliability and validity were applied to the volunteer 152 first year students. The data collection process took three years to compare the different terms since the students take the biostatistics course only in the first year of their academic education. There was a significant difference in the anxiety level of both statistics and scientific research methods between the terms. The age, agender, high school type, economic level and the residence location were not affecting factors of the attitude and anxiety. However, the anxiety level of students having the willingness of education in dental school was significantly higher than the reluctant students. The above courses should be given to the students decreasing the level of anxiety.

KEYWORDS: Anxiety, Attitude, Dental Students, Biostatistics, Scientific Research

O-19 Analysis of Turkey Household Budget Survey Data with Quantile Regression

Oral Abstract / Econometrics

Ismail YENILMEZ¹, Yeliz MERT KANTAR¹,

¹Eskisehir Technical University,

The linear regression (LR) models the relationship between independent variable(s) and the conditional mean of a dependent variable. Ordinary least squares (OLS) estimation is the best estimation method for regression model under the certain assumptions such as homoscedasticity and normality. However, if these assumptions are not satisfied for LR and/or outliers are detected in data, LR is not suitable to model data. Such cases, alternative models or estimation methods may be used. The quantile regression model, which considers the quantile of the dependent variable instead of its mean, is one of the alternative models. Moreover, while LR and robust alternatives focus on the mean of the response variable at each value of the predictors, the quantile regression provides more details about the probability distribution of the response variable and explains the effect of the predictors on quantiles of the response. In this study, we have considered the Turkey Household Budget Survey data. Data is taken from the TurkStat. We have observed that heteroscedasticity in residuals and moderately skewed distribution of residuals. Thus, we have considered quantile regression analysis for the Turkey Household Budget Survey data. The obtained results show that quantile regression estimates, which are calculated on the basis of quantiles and divided into 5 different expenditure groups, are quite different from OLS estimates. Rather than estimating the whole group with a single model based on the expected value, the use of quantile regression provides more precise and detailed results.

KEYWORDS: Turkey Household Budget Survey data, quantile regression, comparative study

O-20 Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation

Oral Abstract / Statistics & Probability

Luigi D'Ambra¹, Pietro Amenta², Antonello D'Ambra³,

¹University of Naples Federico II, ²University of Sannio, ³University of Campania ,

It is well known that the Pearson statistic χ^2 can perform poorly in studying the association between ordinal categorical variables. Taguchi's and Hirotsu's statistics have been introduced in the literature as simple alternatives to Pearson's chi-squared test for contingency tables with ordered categorical variables. The aim of this paper is to shed new light on these statistics, stressing their interpretations and characteristics, providing in this way new and different interpretations of these statistics. Moreover, a theoretical scheme is developed showing the links between the different proposals and classes of cumulative chi-squared statistical tests, starting from a unifying index of heterogeneity, unlikability and variability measures. Users of statistics may find it attractive to understand well the different proposals. Some decompositions of both statistics are also highlighted. This paper presents a case study of optimizing the polysilicon deposition process in a very large-scale integrated circuit, to identify the optimal combination of factor levels. It is obtained by means of the information coming from a correspondence analysis based on Taguchi's statistic and regression models for binary dependent variables. A new optimal combination of factor levels is obtained, different from many others proposed in the literature for this data.

KEYWORDS: Contingency table · Chi-squared statistic · Single and double cumulative chi-squared statistics · Likelihood ratio

O-21 A Panel Data Analysis: The Relationship Between Unemployment, Youth Unemployment and Economic Growth

Oral Abstract / Econometrics

Merve Altaylar¹, Hamdi Emeç²,

¹Social Sciences Institute / Dokuz Eylül University, ²Faculty of Economics and Administrative Sciences / Dokuz Eylül University,

A Panel Data Analysis: The Relationship Between Unemployment, Youth Unemployment and Economic Growth MERVE ALTAYLAR¹*, HAMDİ EMEÇ²* Social Sciences Institute, Dokuz Eylül University, Turkey mervealtaylar37@gmail.com Faculty of Economics and Administrative Sciences, Dokuz Eylül University, Turkey hamdi.emec@deu.edu.tr **ABSTRACT** This paper examines the relationship between unemployment, youth unemployment and economic growth in 20 OECD countries between 2000 and 2017. These variables are examined using static and dynamic panel time series techniques. He does not examine the relationship between these variables, as Okun suggested, and explores how unemployment and youth unemployment affect economic growth. Conventional panel unit root and panel break root and panel cointegration tests were used to investigate the differences of these variables on economic growth. As a result, youth unemployment is more susceptible to economic growth than unemployment, while developments in youth unemployment do not affect economic growth as well as unemployment. **Keywords:** Multiple Structural Break Panel Cointegration, Heterogeneous Panel, Cross-sectional Dependence, DOLS, PANKPSS

KEYWORDS: Multiple Structural Break Panel Cointegration, Heterogeneous Panel, Cross-sectional Dependence, DOLS, PANKPSS

O-22 Simpson’s Paradox: Literature Review and a Dataset for the Treatment of Acne Rosacea Patients in the Muğla Region

Oral Abstract / Statistics & Probability

Burcu Durmuş¹, Öznur İşçi Güneri¹, Aslı Akın Belli²,

¹Muğla Sitki Kocman University, ²Erdem Hospital,

When the data set is divided into groups and evaluated as a total in one study, dilemmas may be seen in the results. This situation is called as Simpson Paradox or Reversal Paradox. In this study, a literature search for Simpson Paradox and analysis for a new data set were performed. For this purpose, four studies in the literature were examined. Three of these studies are based on real-life data sets. The other data set is the “iris dataset” used in many analyzes in the literature. In addition to the literature, a data set from acne rosacea patients in the Muğla region was created. In this dataset, the location of the disease and the treatment methods applied were examined. As a result of the study, it was determined that acne rosacea disease showed Simpson paradox and the data set was included in the study. Besides the data sets, the mathematical definition of paradox is also mentioned in the study. At the end of the study, the situations in which the Simpson paradox is emerged and how to solve it were explained and its importance was emphasized.

KEYWORDS: Acne Rosacea, Different Treatment, Simpson’s Paradox

O-23 Multicomponent Stress-Strength Reliability Estimation based on the Standard Two-Sided Power Distribution

Çağatay ÇETİNKAYA^{1*}, Ali İhsan GENÇ²

Abstract – In this study, we consider the standard two-sided power (STSP) distribution with regard to stress-strength reliability for a multicomponent; s -out-of- k : G system. An s -out-of- k : G system is formed by independent and identically distributed X_1, X_2, \dots, X_n strength components and each component is exposed to a common random stress Y . Such a system survives only if at least s out of k ($1 < s < k$) strengths exceed the stress. Under the STSP distributed stress and strength variables, maximum likelihood and Bayesian estimation procedures are used to estimate the reliability of such a system. As approximate confidence intervals, parametric and nonparametric bootstrap methods for the maximum likelihood estimates and the highest posterior density (HPD) confidence interval by using the Markov Chain Monte Carlo (MCMC) technique for the Bayes estimates are obtained. Finally, performances of the estimators are illustrated with a real dataset.

KEYWORDS: Bayesian Estimation, Maximum Likelihood, Multicomponent Stress-Strength Reliability, Standard Two-Sided Power Distribution

O-24 A New Lifetime Distribution Based on the Triangular Kernel

Oral Abstract / Statistics & Probability

İsmet Birbiçer¹, Ali İhsan Genç¹,

¹Cukurova University, Faculty of Science and Letters, Department of Statistics,

In this study, we introduce a new lifetime model like Birnbaum-Saunders distribution. Several statistical properties of this distribution including hazard rate, moments and distribution of order statistics are studied. This new distribution has a scale parameter and a shape parameter that affects the shape of the distribution from reversed J-shaped to a unimodal one. We discuss four different methods of estimation of the parameters. These methods include classical moments estimation, percentile estimation, least squares estimation and the maximum likelihood estimation. Some of these methods require an iterative process and thus we give algorithms for the estimation. Since the distribution function of the new model is in an explicit form, the percentile estimators of the parameters can be obtained explicitly. A simulation study is performed to see the performances of the estimators. A real data application comparing our proposed model with the existing ones is also given.

KEYWORDS: Lifetime Distribution, Triangular Distribution, Reversed J-Shaped Distribution

O-25 A SIMULATION STUDY ON THE UNBALANCED DESIGN PROPERTIES OF THE GENERALIZED P-VALUE BASED TESTS

Oral Abstract / Statistics & Probability

Mustafa CAVUS¹, Berna YAZICI¹,

¹Eskisehir Technical University, Department of Statistics,

Several procedures are proposed for testing equality of group means under heteroscedasticity in the literature. Generalized p-value method is used to conduct powerful tests in the presence of nuisance parameters for unequal group variances in recent decades, such as Weerahandi Generalized F-test, Parametric Bootstrap test, Fiducial Approach test and Alvandi et al. Generalized F-test. The unbalanced design may also cause the heteroscedasticity besides the unequal group variances in testing the equality of group means. In this study, the unbalanced design properties of the generalized p-value based tests are investigated in terms of Type I error probability and power of the test. A Monte-Carlo simulation study is conducted under various unbalanced design scenarios. As a result of the study, the properties of the tests are discussed and some useful comments are given to the researchers.

KEYWORDS: unbalanced, penalized power, heteroscedasticity, ANOVA

O-26 Imputation of Missing Observations in Longitudinal Data via Neural Network

Oral Abstract / Biostatistics

Marwa BenGhoul¹, Berna Yazıcı¹,

¹Eskişehir Technical University,

Longitudinal data is a data which tracks the same information over subjects on different timepoints, it is mostly used in the pharmaceutical industry. Such type of data has common problem known as attrition or missing data. Despite several researches highlighted the cruciality of knowing the missingness mechanism before the imputation method, it still ignored. Recently, Neural Network (NN) has started to get more attention as a technic of data imputation for Missing Completely at Random data. Spite of many researches investigated this approach on classic clinical data, it still not highly applied specially for longitudinal data. Hence, this research consists on investigating the efficiency of NN, particularly multiplayer perceptron on handling the missing data. Based on recent researches results, the activation function in this paper will be a wavelet function. A generated longitudinal data based on a historical hypertension study published by National Institute Health is used. Four wavelet functions have been tested (Gaussian, Morlet, Meyer and Mexican Hat) as activation function. A comparison between the ad hoc imputation methods, Expectation Maximum (EM) and NN technic is provided. The SAS macros published by (Sarle, 1994) are updated and others are created to perform the NN for this research. NN approach, particularly with Morlet as an activation function, show interesting performance, better than the ad hoc imputation methods and with very slight difference from the EM.

KEYWORDS: longitudinal data, missing data, Neural Network, multilayer perceptron, wavelet function

O-27 An Application of New Stochastic Model Using Generalized Entropy Optimization Methods

Oral Abstract / Statistics & Probability

Aladdin Shamilov¹, Nihal İnce¹,

¹Eskisehir Technical University,

In this study, a new method to obtain approximate probability density function of random variable of solution of Stochastic Differential Equations (SDEs) by using Generalized Entropy Optimization Methods (GEOM) are developed. By starting given statistical data and Euler-Maruyama's (EM) method approximating SDE are constructed several trajectories of SDEs. The constructed trajectories allow to obtain random variable according to fixed time. The probability density function according to random variable which is solution of SDE at a fixed time is the solution of Fokker-Planck-Kolmogorov equation at this time. Probability density function of mentioned random variable is obtained by GEOM. An application of the developed method we have considered a biological data. Firstly, the formulas for estimating the parameters are given for SDE model. Then EM schemes are constructed via estimating the values of parameters for biological data. Biological data and its approximative EM values from model are illustrated in figures and tables. It is determined a lack of fit between biological data and approximative values of random variable using by χ^2 criteria, and also it is obtained R^2 , RMSE and entropy measure. Finally, approximative probability density functions of random variables of solutions of SDE model are constructed via pdfs of random variables in tables and figures by using GEOM. Moreover, obtained distributions using by GEOM can be used for assessment of the biological potential and the performance of biological systems. So, the present study may give different and useful insights to scientific areas.

KEYWORDS: Generalized Entropy Optimization Methods, Stochastic Differential Equations, Euler-Maruyama's method, Probability density function

O-28 A NEW METHOD BASED ON INTERQUARTILE RANGE TO FEATURE SELECTION FOR CLASSIFICATION IN BIG DATA

Oral Abstract / Statistics & Probability

Ahmet KOCATÜRK¹, Bülent ALTUNKAYNAK¹,

¹Gazi University,

In the analysis of large-scale gene expression data, it is important to identify a well-representing subset of the data. Specifying the subset can be done with the feature selection. In the feature selection, the models with the highest effect can be determined and the models with high accuracy can be obtained with few features. Feature selection is generally divided into two groups. First, wrapper methods are methods that seek combinations of features that maximize the accuracy of model. Genetic algorithms and simulated annealing methods are examples of wrapper methods. Second, filter methods examine the effect of each feature on the accuracy of model. Examples of filtering methods are Chi-square statistics, information gain and ReliefF. In this study, a new filtering method based on the interquartile range for feature selection (FSIQR) is proposed in gene expression data. The FSIQR method takes into account the overlapping areas of interquartile range from each class. The feature selection methods used by the FSIQR method are Chi-square, Information Gain, and ReliefF. Also, it is aimed to compare FSIQR method and other methods in terms of accuracy by using real data sets. Support vector machines and k-neighborhood method are used for accuracy rates and n-fold and leave-one-out methods are applied for cross validation. As a result, FSIQR method, which is a new filtering method based on interquartile range value for feature selection, has been found to reach a higher accuracy rate in many cases compared to other feature selection methods on real data.

KEYWORDS: Gene Expression Data, Feature Selection, Filter Method

O-29 RA-CUSUM chart based on LR-fuzzy data

Oral Abstract / Statistics & Probability

Afsaneh Rezaeifar¹, Bahram Sadeghpour Gildeh¹, G.R. Mohtashami Borzadaran¹,

¹Ferdowsi university of mashhad,

The cumulative sum chart is widely used in manufacturing processes, but in medical science and particularly for monitoring the performance of cardiac surgeon or a group of surgeons, the patients preoperative risk should be taken into account. So risk adjusted charting procedures gained attention. Risk-adjusted cumulative sum chart based on testing the odds of mortality was proposed in 2000, which takes the preoperative risk of patients into account. Since preoperative risk is vague and non-precise variable and the anesthesiologists after checking how many risk factors a patient has, determine the risk of mortality before the surgery as a linguistic term such as low, medium, high or others like that, it is better to be considered as a fuzzy number, which can be determined by using fuzzy logistic regression model. In this condition, we need a special chart to monitor the performance of surgeons based on these fuzzy data. In this paper we proposed RA-CUSUM chart based on LR-fuzzy data and then consider the conclusion on real data.

KEYWORDS: Preoperative risk, Monitoring surgical performance, Odds of mortality, RA-CUSUM control chart, Fuzzy logistic regression

O-30 SUPPORT VECTOR REGRESSION FOR WEATHER FORECASTING

Oral Abstract / Industrial Statistics & Engineering Applications

Neslihan Cevik¹, Ahmet Sermet Anagün¹,

¹Izmir University of Economics ,

Weather and precision of weather forecasts have a very important role in our daily lives especially in the field of transportation since it directly affects the quality and the safety of the service. In this study, the aim was to compare the forecast errors executed by different regression approaches. The data has been provided by Republic of Turkey Ministry of Agriculture and Forestry, General Directorate of Meteorology for Izmir Adnan Menderes Airport with eight independent variables and the average temperature as the dependent variable for the years 2015-2017. Some of the inputs are daily maximum and average wind speed and direction, daily maximum and average atmospheric pressure, daily average cloudiness, daily average relative humidity. In order to make the analysis reliable, some of the inputs were reorganized while some of them eliminated from the original data set. The data for 2015 and 2016 have been used for creating the regression models and 2017 has been used for testing purposes. All of the regression models were compared based on the RMSE values. Results show that Gaussian Regression with kernels; Matern 5/2, Rational Quadratic and Squared Exponential models have lower RMSE values compared with the SVR.

KEYWORDS: weather forecasting, support vector regression, multiple regression, nonlinear regression, data mining.

O-31 Robust Gene Co-Expression Network Analysis

Oral Abstract / Biostatistics

Aylin ALIN¹, Ayça ÖLMEZ², Gökhan KARAKÜLAH³,

¹Department of Statistics, Dokuz Eylül University, ²The Graduate School of Natural and Applied Science, Department of Statistics, Dokuz Eylül University, ³Izmir International Biomedicine and Genome Institute, Dokuz Eylül University,

Through the recent development of sequencing technologies, it has become easier to collect and reveal the information about genes. With these data, Gene co-expression networks (GCNs) can be constructed, and investigations that discover patterns and associations between the expression datasets of the high throughput genes and their products can be conducted. GCNs, which can be utilized for investigating causing diseases genes and revealing the role of genes in under different biological conditions, cannot be built by classical statistical methods (Pearson correlation coefficient, etc.) due to the complexity of the interested datasets. One of the proposed methods for the problems which occur as missing values, noise and multicollinearity problem is the Partial Least Squares estimator (PLS) of the regression. However, PLS is a very sensitive method for non - normally distributed error terms, leverage points and outliers. There is a robust alternative called Partial Robust M Regression (PRM). It provides robustness to not only non-normally distributed errors but also outliers and leverage points. PRM is a simple, realistic and informative method that can be used to create gene networks. Herein, robust Gene co expression networks based on PRM have been proposed and studied on a dataset including fore-, mid- and hindbrain of a developing mouse brain. The GCNs for different parts of developing brain were successfully created, and putative gene-gene interactions studied.

KEYWORDS: Gene co-expression network, partial robust m regression, robustness, outliers, leverage points.

O-32 Generalized Gamma Parameters Estimation with Cuckoo Search Algorithm

Oral Abstract / Statistics & Probability

Ali Mert¹, Rıdvan Temiz¹,

¹Ege University, Department of Statistics,

Generalized Gamma Distribution (GGD) is a very flexible and popular distribution because it transforms into many special statistical distributions such as Weibull, Exponential, Gamma, Pareto and Rayleigh. GGD, which is a widely studied topic in the literature due to its flexible structure, is also interesting due to its difficulty in estimating its parameters. The differentiation of the maximum likelihood function which is a widely employed approach is hard to perform because of the complexity of the function. Therefore, it is not possible to obtain formulas for GGD parameter estimation. Because of the fact, scientists prefer numerical solution instead of a closed formula. Meta – heuristic algorithm approach is one of the best in estimating GGD parameters. In this study, parameters estimation is performed by using the Cuckoo Search Algorithm (CSA). Since CSA is one of the most suitable methods for maximization problems, the maximum likelihood function of GGD is defined as the objective function and CSA is applied to the objective function. In order to measure the performance of the CSA, we conduct a simulation study. In the simulation phase, we generated random numbers from GGD with various parameters in different sample sizes (30, 50, 100 and 250). The performance of CSA is compared with the performance of the Genetic Algorithm. We run both algorithms 50 trials for each sample size and parameters combination. We utilized the best objective function value among trials and the standard deviation of objective function values as performance criteria. CSA is superior to Genetic Algorithm.

KEYWORDS: Generalized Gamma Distribution, Maximum Likelihood Estimation, Cuckoo Search Algorithm, Meta-Heuristic Algorithm, Optimization

O-33 A goal programming approach for the use of restricted data envelopment analysis as a tool in multi criteria decision analysis

Oral Abstract / Industrial Statistics & Engineering Applications

Esra Betül KINACI¹, Harun KINACI², Hasan BAL¹,

¹Gazi University, ²Erciyes University,

Multi-criteria decision making (MCDM) methods and Data Envelopment Analysis (DEA) allow for ranking between units using different methods on multiple feature units. The MCDM methods provide the solution of complex problems with conflicting characteristics by using the weight information related to the criteria. DEA performs this ranking through the definition of efficiency. The definition of the efficiency can basically be expressed as the ratio of the weighted outputs to the weighted inputs. In DEA, this ratio is converted to a linear model and resolved. At this stage, while optimizing the model, weighting is made according to input and output variables of DMUs. In the classical DEA has no restrictions on these weights and weights are free. However, in the MCDM processes the weight of the properties (criteria) of the units is important. A model that will be formed considering the relative importance levels of criterion weights will contribute to the use of DEA as a tool in the MCDM methods. Also, in classical DEAs, sometimes weights of variables can be assigned a value of zero. Multi-step linear model used in DEA can be very important. Multi-step linear model used in DEA can be solved by the goal programming approach. In this study, a DEA model considering the relationship between the variable weights obtained with the MCDM method was proposed and the model is solved as a goal programming problem. As a result, a problem in which the classical DEA assigns zero weight was solved by this method.

KEYWORDS: Data Envelopment Analysis, Multi Criteria Decision Making, Goal Programming

O-34 Comparison of Classification Algorithms on Different Data Sets

Oral Abstract / Statistics & Probability

Burcu Durmuş¹, Öznur İşçi Güneri¹, Nevin Güler Dincer¹,

¹Mugla Sitki Kocman University,

In this study, classification which is one of the most useful and popular data mining methods is discussed. The study investigates which algorithm has better performance over different data sets. For classification, Bayes, functions, decision tree, lazy techniques were examined. For this purpose, 50 different data sets from Machine Learning Repository were analyzed within the scope of classification. Weka Tools were used for analysis and the results of different classification algorithms were compared in terms of model performance criteria. As a result, LMT (Logistic Model Tree) and Random Forest performed best in decision tree algorithms. The performance criteria gave parallel results.

KEYWORDS: Classification Algorithms, Data Mining, Different Data Sets, Performance Evaluation, Weka

O-35 Classification of the Prices of Real Estate Using Machine Learning Methods

Oral Abstract / Statistics & Probability

Betül KAN KILINÇ¹, Yonca YAZIRLI²,

¹Eskisehir Technical University, Faculty of Science, Department of Statistics, ²Eskisehir Technical University, Institute of Graduate Programs, Department of Statistics,

As the information systems are growing day by day, it becomes easier to obtain bigger data and store it in systems. However, the data stored in the systems do not make sense of their own. Therefore, the analysis of the available data and the methods of predicting from this data play an important role for the decision makers. The process of obtaining useful information from huge amount of data can be done by data mining. One of the areas where data mining is used is the real estate platform. The aim of this study is to classify the housing unit prices of real estate for sale in Istanbul obtained from an online web source by using multinomial logit (MNL) model, support vector machines (SVM) and random forest (RF) methods. Classification algorithms have been to estimate the classifier performance. Data set is splitted as 70% for training and 30% for testing. For the model validation, 5-fold cross-validation technique is used. The accuracy and relevant performance metrics of the methods are compared. R Studio is used for all computations.

KEYWORDS: Data Mining, Classification, Cross Validation.

O-36 Improving Two Stage Two Parameter Ridge Estimator under Linear Restrictions

Oral Abstract / Statistics & Probability

Selma Toker¹, Nimet Özbay¹,

¹Çukurova University,

A two parameter estimator is more advantageous than a single parameter estimator because two parameters have two different benefits with regard to estimating structural coefficients. To take advantage of this idea, Toker (2018) has described a two stage two parameter ridge estimator in simultaneous equations model. This is such an estimator which mitigates the problem of multicollinearity with its first parameter and improves quality of fit with its second parameter. If some constraints are encountered in the model of simultaneous equations, restricted estimators can be more attractive than the classical ones. In this respect, we define restricted form of the two stage two parameter ridge estimator. While proposing our new restricted two stage two parameter ridge estimator, we have inspired by the notion in the paper of Üstündağ Şiray and Toker (2014). In addition, theoretical properties of this new estimator are investigated and an optimal biasing parameter is proposed. The best performed estimator is determined in terms of efficiency as a consequence of the empirical evaluations.

KEYWORDS: Linear restrictions, Mean square error, Multicollinearity, Simultaneous equations model, Two parameter estimator

O-37 Defining Some Adaptive Optimal Estimators for the Distributed Lag Model

Oral Abstract / Statistics & Probability

Nimet Özbay¹, Selma Toker¹,

¹Çukurova University,

Minimum mean square error estimators have a widespread usage to estimate the unknown coefficients of linear regression model. For practical purposes, adaptive forms of these estimators are preferable due to their attractive performances. Within this context, adaptive forms of the estimators that minimize mean square error can be evaluated in the distributed lag model, which is a dynamic model for time series data. In the estimation issue of the distributed lag model, Almon estimator is the foremost estimator depending on its unbiasedness and ease of application. However, in the existence of multicollinearity, the use of biased estimators, one of which is Almon ridge estimator, is inevitable. We take into consideration the Almon and Almon ridge estimators in this paper with the aim of defining two new adaptive optimal estimators. The performance of the proposed adaptive optimal Almon and adaptive optimal Almon ridge estimators are examined by means of a Monte Carlo simulation.

KEYWORDS: Adaptive optimal, Almon estimator, Almon ridge estimator, Distributed lag model, Mean square error

O-38 Determining the types of missing data under supervised statistical models

Oral Abstract / Statistics & Probability

Vladimir Vasić¹

¹University of Belgrade, Faculty of Economics,

In the modern business world, more and more data are analyzed. Especially the Customer Relationship Management department analyzes the behavior and preferences of customers. In order to have relevant information, companies often conduct surveys of their customers. In order for the company to get reliable information about their customers' preferences, it must allow clients to respond freely to the questions asked. Also, if the customer thinks that the question is intimate, he will not want to answer the question. A big mistake would be made in collecting data, if the customer is still forced to answer the given question. In this case, the customer gives an incorrect answer; and that's what we least want. For this reason, we must allow a customer, if he wants to, not to answer a question. In such real situations, we encounter the problem of missing data. How the missing data will be resolved depends on the mechanism of missing data. If the data are missing in a completely random way, then they can be solved by standard procedures. However, if the data are missing in a random manner then they must be addressed by modern methods of analyzing the missing data. If the data are missing in a non-random way then the problem of missing data cannot be resolved with quality; and such data should not be further analyzed statistically. For this reason it is very important to accurately determine the type of missing data; which will present the subject of research in this paper.

KEYWORDS: types of missing data, missing at random, missing completely at random, statistical tests, supervised statistical models

O-39 Ridge deviance residual charts for monitoring Poisson distributed data

Oral Abstract / Industrial Statistics & Engineering Applications

Ulduz Mammadova ¹, M. Revan Özkale¹,

¹Çukurova University,

Control charts based on regression models are appropriate for indicating the causes of a signal in a process. The residuals obtained from the regression models are usually for this propose. If a correlation exists among the predictors, then residuals considering the collinearity problem should be used. In this study, the control chart theory and the Poisson regression model are combined to yield an effective technique for controlling the process. Ridge deviance residuals for the Poisson regression model is defined in the case of multicollinearity. Then, Shewhart, CUSUM, and EWMA control charts are introduced based on the ridge deviance residuals for the Poisson regression model. The performance of the new approach is illustrated through a real data study in which plastic plywood process is examined.

KEYWORDS: Statistical Process Control, Poisson regression, Ridge estimator, Residual control charts

O-40 PERFORMANCE COMPARISON OF MACHINE LEARNING METHODS AND TRADITIONAL TIME SERIES METHODS FOR FORECASTING

Ozancan Özdemir^{1*}, Ceylan Yozgatlıgil²

¹Department of Statistics, Middle East Technical University, Turkey, ozancan@metu.edu.tr

²Department of Statistics, Middle East Technical University, Turkey, ceylan@metu.edu.tr

One of the main objectives of the time series analysis is forecasting. In this study, we use different approaches in time series modelling. In addition to traditional forecasting methods for time series data set which are namely ARIMA and exponential smoothing, the forecasts are also obtained by using eight different machine learning methods. These methods are Random Forest, Support Vector Regression, XGBoosting, Bayesian Neural Network, Light GBM, Long Short Term Memory Neural Network and Multilayer Perception. [1] It is also known that time series generally have both linear and nonlinear patterns. In order to deal with this data structure, a hybrid methodology which combines both linear and nonlinear components was proposed by Zhang. [2] In this methodology, predicted values of a time series can be obtained by summing both linear and nonlinear component. In this study, hybrid models are constructed by using both machine learning methods for nonlinear pattern and statistical methods for linear pattern. Therefore, the forecasts are also obtained using hybrid time series models. The data set used in this study is monthly time series used in M4 Competition. After observing the results of studies, the performance and impact of all methods are discussed.

KEYWORDS : Forecasting; Time series analysis; Hybrid method; Machine Learning; M4 Competition

References

[1] Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3): e0194889. <https://doi.org/10.1371/journal.pone.0194889>

[2] Zhang, G.P. (2003), Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing (50), 159-175.

O-41 Clustering with Genetic Algorithm: A Simulation Study

Oral Abstract / Statistics & Probability

Erkut TEKELİ¹, Özlem AKAY¹, Güzin YÜKSEL¹,

¹Cukurova University,

Clustering Analysis is an important unsupervised classification technique which is used to discover patterns and associations within data. Genetic algorithms are randomized search and optimization techniques guided by the principles of evolution and natural genetics, having a large amount of implicit parallelism. A good GA explores the search space properly as well as exploits the better solutions to find the globally optimal solution. Hence, genetic algorithm is suitable for clustering task. One of the major problems in cluster analysis is that different clustering methods can and do generate different solutions for the same data set. Therefore, in this study, it is aimed to ensure that the units are located in the correct clusters by using genetic algorithm. For this purpose, a new fitness function has been defined. The proposed algorithm was tested by simulation study. In the simulation study, R programming language was used and random samples were generated for different sample sizes, cluster numbers and different outliers. In order to compare the performance of the new clustering algorithm, also K-means method which is one of the most popular methods in clustering was used with it. According to the clustering results, classification ratios were calculated for the two methods. Results of the analysis showed that the proposed algorithm can generate better clustering results than k-means clustering algorithms. Hence, in this algorithm, the use of fitness function ensure for converge to the global optimum. This study is supported by the Scientific Research Projects Unit of Çukurova University. (Grant Number: FBA-2018-10438)

KEYWORDS: clustering, genetic algorithm, fitness function

O-42 Regression Analyses or Decision Trees?

Oral Abstract / Econometrics

Burcu Kocarık Gacar¹, Ipek Deveci Kocakoc²,

¹Manisa Celal Bayar University, Econometrics Dpt., ²Dokuz Eylül University, Econometrics Dpt.,

Decision trees, which are included in data mining techniques, are used as a method in estimating future data trends or forming important data classes by creating classification and regression models in the form of a tree according to the structure of the data sets. Similarly, logistic regression is another technique used for classification purposes. Logistic regression is an alternative to linear regression analysis when dependent variable is binary. In linear regression analysis, the value of the dependent variable is estimated and in logistic regression analysis, the probability of one of the dependent variable values is estimated. There are many statistical techniques used for classification and regression. This study means to give a lead about which technique should be used for which research question. Classification trees, logistic regression, linear regression and regression trees were applied on the same data set and suggestions were made on which technique would be more appropriate in which situations. After the models were verified, the performances of the models were evaluated and their predictive power was compared. This analysis aims to reveal similarities and differences between the models predicted by these four methods.

KEYWORDS: linear regression, logistic regression, regression trees, classification trees

O-43 Fitting One Life Expectancy at Birth Data to Stochastic Differential Equation

Oral Abstract / Statistics & Probability

Aladdin Shamilov¹, Sevda Ozdemir Çalikuşu², Fevzi Erdogan²,

¹Eskisehir Technical University, ²Van Yuzuncu Yil University,

In this study one life expectancy at birth data is investigated by Stochastic Differential Equation Modeling (SDEM). Firstly, parameters of SDE which occur in mentioned biological problem are estimated by using the maximum likelihood procedure. Then, we have obtained reasonable Stochastic Differential Equation (SDE) based on the given biological data. Moreover, by applying Euler-Maruyama Approximation Method trajectories of SDE are achieved. The performances of trajectories are established by Chi-Square criteria, Root Mean Square Error (RMSE) value. The results are acquired by using statistical software R-Studio. These results are also corroborated by graphical representation.

KEYWORDS: Ito stochastic differential equation, euler-maruyama approximation method, maximum likelihood

O-44 ECONOMETRIC ANALYSIS OF THE EFFECT OF OPEC OIL POLICIES ON ECONOMIC GROWTH

Oral Abstract / Econometrics

FATİH ÇEMREK¹, HÜSEYİN NACİ BAYRAÇ¹,

¹ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ,

Members of the Organization of Petroleum Exporting Countries (OPEC) have an important role in shaping the global oil market today. Since the industrial and export structure in these countries is largely dependent on oil, member economies are greatly affected by the fluctuations in oil prices. OPEC is an important cartel actor that determines global oil supply and thus oil prices. The main objective of OPEC is to establish a price system that will serve the interests of the supply and demanders in the oil market and guarantee the oil gains of the producer countries. OPEC countries have about 2/3 of the world's oil reserves and meet 1/3 of the daily oil production. OPEC's policies have a significant impact on spot and futures prices. The changes in oil prices are closely related to all countries and especially the countries with high level of oil importers and countries with significant reserves are significantly affected by these fluctuations in price. Although oil prices are determined according to supply and demand conditions in the market, different variables also affect the price. The conditions of the world economy, developments related to alternative energy sources and economic and social changes in the OPEC (Non-OPEC) and Non-OPEC countries constitute some of these variables. In this study, the relationship between annual average crude oil price (nominal value USD) and oil production (Gross Domestic Product (Million \$)) is investigated by panel data analysis.

KEYWORDS: OPEC, Petrol Price, OPEC Policies, Economic Growth, Panel Data Analysis

1 Doç. Dr. Fatih Çemrek, Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, fcemrek@ogu.edu.tr

Dr. Öğr. Üyesi Hüseyin Naci Bayraç, Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi İktisat Bölümü, nbayrac@ogu.edu.tr,

1 This study is supported by ESOGU Scientific Research Project Commission as (Project) 2018-2456

O-45 The Change of The Factors Determining Happiness in Turkey

Oral Abstract / Econometrics

Eda Yalçın Kavacan¹

¹Pamukkale University,

Happiness can simply be defined as the subjective well-being and elements provided us to enjoy life. Happiness and the determinants of happiness have been questioned since ancient times. The concept of happiness has become increasingly important in the science of economics rather than psychology and sociology. As a result, the concept of happiness economics has emerged and the studies in the related field have become increasingly important. There are different life satisfaction surveys aiming to measure the happiness levels of countries and enabling international comparisons. The Life Satisfaction Survey conducted by the Turkey Statistical Institute reveals the overall profile of what level of life satisfaction and happiness for Turkey. The main aim of the study was to examine the effects of determinants of happiness changes over the years in Turkey. For this purpose, the factors determining happiness in Turkey was examined by using the data of Life Satisfaction Survey for the period from 2008 to 2018. Considered as determinants of happiness socioeconomic variables such as household income, age, gender, education level, health status was used and the econometric analyses were performed by using generalized ordered logit model. With the obtained findings, the change of the factors determining happiness in Turkey over time has been evaluated in detail.

KEYWORDS: Happiness, Happiness Economics, Generalized Ordered Logit Model, Logit Model.

O-46 A new extended symmetry model for square contingency table

Oral Abstract / Statistics & Probability

Gökçen Altun¹,

¹Bartın University,

In this study, extended symmetry model is proposed to model the non-symmetric structures of square contingency tables. The proposed model tests the equality of local odds ratios between the one side of the main diagonal and corresponding other side. Two real data sets are analyzed to demonstrate the usefulness of new model. The proposed model is compared with twenty-five models introduced for analyzing the square contingency tables for both symmetric and non-symmetric structures. The model selection criteria are used to decide best fitted model among others. The results show that the proposed model provides the best fitting performance than other existing models for square contingency tables.

KEYWORDS: Square contingency tables, Ordinal category, Local odds, Non-symmetric structure.

O-47 Bayesian Parameter Estimation of Akash Distribution

Oral Abstract / Statistics & Probability

İlhan USTA¹, Merve AKDEDE²,

¹Eskisehir Technical University, ²Usak University,

Akash distribution is a new one-parameter lifetime distribution introduced by Shanker (2015) as an alternative to Lindley distribution. However, the Bayesian estimation of the parameter of Akash distribution has not yet been studied in the literature. Therefore, this study considers the Bayesian approach to estimate the parameter of Akash distribution. We derive the Bayes estimators under squared error loss function and general entropy loss function. Since the Bayes estimators cannot be obtained in close-form, Lindley's method is employed to attain the approximate Bayes estimates of the parameter. The performance of the proposed estimators is compared with the maximum likelihood estimator via a Monte Carlo simulation study. One real life data set has been analyzed for illustrative purposes.

KEYWORDS: Akash Distribution, Bayesian parameter estimation, Lindley approximation, Squared error loss, General entropy loss

O-48 E-Bayesian Estimation for Topp-Leone Distribution

Oral Abstract / Statistics & Probability

İlhan USTA¹, Merve AKDEDE²,

¹Eskisehir Technical University, ²Usak University,

In this study, we are concerned with the E-Bayesian (the expectation of Bayesian estimate) method, the maximum likelihood and the Bayesian estimation methods for estimating the shape parameter of Topp-Leone distribution. The Bayesian and E-Bayesian estimators under squared error (symmetric) and LINEX (asymmetric) loss functions are derived by applying Lindley approximation technique. The performances of the proposed E-Bayesian estimators are also compared with the corresponding maximum likelihood and Bayesian estimators in terms of bias and mean squared error through an extensive simulation study.

KEYWORDS: Topp-Leone Distribution, E-Bayesian estimation, Bayesian estimation, Squared error loss, LINEX loss

O-49 HEGY Seasonal Unit Root Test: An Application for Agricultural Products Producer Price Index

Oral Abstract / Econometrics

Okan KÜRES¹, Fatih ÇEMREK¹,

¹Eskişehir Osmangazi Üniversitesi,

In the study, the existence of seasonal unit roots were investigated by using quarter term data of Turkey's (2003:1-2018:4) agricultural products producer price. The study was performed with HEGY test used to determine seasonal unit root. The HEGY test proposed by Hylleberg, Engle, Granger and Yoo in 1990 is one of the most commonly used tests in the literature. According to the test results obtained; For the Agricultural Products Producer Price Index series, the presence of seasonal unit root at zero frequency was determined.

KEYWORDS: HEGY Test, Seasonal Unit Root Test, Agricultural Products Producer Price Index

O-50 Additive Gaussian Process Modeling and Novel Sparse Bayesian Regression with Applications in Business and Industry

Oral Abstract / Statistics & Probability

Juergen Pilz¹, Konstantin Posch¹, Maximilian Arbeiter²,

¹Alpen-Adria-University Klagenfurt, ²Alpen-Adria University Klagenfurt,

In this talk we first focus on the use of Gaussian Processes (GP's) for the approximation of computer models. Whereas GP's have proved to be useful for the analysis of spatially correlated data, these models cannot be simply transferred to analyze complex computer models. A very desirable attribute of such surrogate models is a high flexibility for making them applicable to a large class of engineering science problems while still obtaining interpretable results. To achieve this goal we use Gaussian Processes as basis functions of an additive model. Another desirable property of a surrogate model is numerical stability, which can be particularly challenging when it comes to estimating the correlation function parameters. To assure robust correlation matrices we use a Bayesian approach with a reference prior being assigned to each component of the additive model. The additive model structure also allows us to reduce the high-dimensional optimization problems to a few subroutines of lower dimension. We illustrate our findings by modeling the magnetic field of a magnetic linear position detection system. We then go on to deal with high-dimensional (linear) regression problems and introduce a novel Bayesian approach to the problem of variable selection in such models. In particular, we present a hierarchical setting which allows for direct specification of a-priori beliefs about the number of non-zero regression coefficients. To guarantee numerical stability, we adopt a g-prior with an additional ridge parameter for the unknown regression coefficients. In order to simulate from the joint posterior distribution an intelligent random walk Metropolis-Hastings algorithm, which is able to switch between different models, is proposed. Testing our algorithm on real and simulated data illustrates that it performs at least on par and often even better than other well established methods.

KEYWORDS: Sparse Regression, Variable Selction, Gaussian Processes, Surrogate Modelling, Posterior predictive uncertainty

O-51 Estimating Species Diversity Components of Various Forest Stand Types in The Lake Districts using a Draft Software for Biodiversity Estimation (BİÇEB)

Oral Abstract / Biostatistics

Ahmet MERT¹, Kürşad ÖZKAN¹, Ecir Uğur KÜÇÜKSİLLE², Halil SÜEL³, Serkan GÜLSOY¹, Murat BAŞAR⁴, Mehmet Güvenç NEGİZ³,

¹Faculty of Forestry, Isparta University of Applied Sciences, Isparta, Turkey, ²Computer Engineering Department, Faculty of Engineering, Süleyman Demirel University, Isparta, Turkey, ³Sütçüler Prof. Dr. Hasan Gürbüz Vocational School, Isparta University of Applied Sciences, 32900, Isparta, Turkey, ⁴Republic of Turkey General Directorate of Forestry, Ankara, Turkey,

The present study was carried out to estimate the species diversity components (i.e., alpha diversity (α), beta diversity (β) and gamma diversity (γ)) of plant communities taken from different forest stand types in the Lake District, Turkey. Whittaker, Simpson, Shannon and Legendre & De Cáceres equations were employed in defining the diversity components. All computations were done by a draft software called as Biyolojik Çeşitlilik Hesaplama Programı (BİÇEP). The results of diversity components of forest stand types were then compared to each other and evaluated from ecological point of view.

KEYWORDS: complexity, entropy, environmental factors, plant diversity, statistical methods

O-53 Modeling bivariate survival data in the presence of right censoring by Archimedean Copula approach

Oral Abstract / Statistics & Probability

ECE GORCEĞİZ¹, BURCU HUDAVERDI UCER¹,

¹Dokuz Eylül University,

Abstract - Modeling dependence structure of a bivariate survival data is one of the main issues in biomedical studies. Survival copula deals with such a lifetime data and is used for modeling and understanding the distributional structure. In this study, we consider modeling and analysing the bivariate survival data in the presence of right censoring using Archimedean copula functions. We use Emura et al.(2010) goodness-of-fit testing procedure for the model selection. First, we examine the heart transplant data and model the dependence structure between waiting time for transplant and post-transplant survival time to see the co-movements of these variables. Second, we examine the diabetic retinopathy data and model the dependence between the survival times of the two eyes of the same patient in case of laser photocoagulation treatment.

KEYWORDS: Copula, right censoring, bivariate survival data, Archimedean copula, survival copula

O-54 Theoretical and practical aspects in the conditional quantile estimation for dependent spatial functional data

Oral Abstract / Statistics & Probability

Fahimah Alawadhi¹, Ali Laksaci², Mustapha Rachdi³,

¹Kuwait University, ²King Khalid University, ³Grenoble Alps University,

Nowadays, there are many applied fields for which data are available from monitoring stations, and measurements are provided at relatively close times. This type of data could be considered as coming from a spatio-functional variable. In this contribution, we are interested in spatial prediction for this kind of data (spatial and functional) using the prediction model from conditional quantiles. Indeed, this model has many advantages over the classical model of regression. In particular, (i) it is more robust than the classical regression, (ii) it performs well in different forms (functional linear model, partial linear models or non-parametric models) and (iii) it provides a more informative pointwise predictive interval. Therefore, the main objective of this work is to build a new estimator for this spatio-functional model, using functional linear local estimation ideas and then compare it to existing competitive models in the literature. In particular, we establish some asymptotic convergence results, including the asymptotic normality of the proposed estimator, under some general mixing conditions. Finally, in order to validate/confirm the usefulness of this work and the progress that it realizes on the practical level, we experiment the obtained results first on simulated data, and then on real ecological spatial data.

KEYWORDS: Spatial function, estimation, conditional quantile.

**O-55 RESPONSE SURFACE APPROXIMATION TO STRESS-STRENGTH RELIABILITY
UNDER DEPENDENCY STRUCTURE**

Oral Abstract / Statistics & Probability

Gözde KUS¹, Sevcan DEMİR ATALAY¹,

¹Ege University Department of Statistics,

Stress-strength models are commonly used to specify the reliability of components or systems. One can express reliability of a component in stress-strength models as $P(Y>W)$, in which Y is the strength and W is the stress variables of the component. In general, analyses are performed on the assumption that the distributions of strength or stress variables of the components in the system are known. The present study aims to specify an approximation for the distribution of the strength variable of the components by response surface methodology using some factors that affect the strength variable. Second order polynomial model is used to fit a regression model for the strength variable in response surface methodology. Moreover, dependency structure between the stress and strength variables is examined by using Copula Theory. Hence, the reliability function is improved under the dependency structure. In this way, reliability optimization of the component and determining the design point which makes the reliability optimum may be possible.

KEYWORDS: Stress-Strength Reliability, Response Surface Methodology, Optimization, Copula Theory

O-56 Evaluation of The Reduction Algorithms Based on Rough Set Theory –An Application

Oral Abstract / Statistics & Probability

Yonca YAZIRLI¹, Betül KAN KILINÇ²,

¹Eskisehir Technical University, Institute of Graduate Programs, Department of Statistics, ²Eskisehir Technical University, Faculty of Science, Department of Statistics,

Attribute reduction is one of the essential problems in the field of data mining, machine learning and pattern recognition for decision makers. Different methods have been developed because of handling this problem. One of the most capable approaches used for this purpose is offered by the Rough Set Theory. In this paper, we evaluate reduction algorithms based on rough set theory for efficient classification with a minimum set of attributes for real estate in Istanbul. Johnson' s, Genetic Algorithm and Dynamic reducts were evaluated by using the voting method. The performances of reduction methods were compared to classify the unit housing price of real estate per ₺/m² in Istanbul. The results showed that genetic algorithm achieved a better performance by using voting classifier than others.

KEYWORDS: Attribute reduction, rough set theory, classification, real estate.

O-57 Parameter Estimation for the k-th Extreme Value Distribution

Oral Abstract / Statistics & Probability

Talha Arslan¹, Sukru Acitas², Birdal Senoglu³,

¹Van Yüzüncü Yıl University, ²Eskisehir Technical University, ³Ankara University,

The k-th Extreme Value (EV_k) distribution proposed by Gumbel (1935) is defined as asymptotical distribution of k-th extremes and used for modelling these extreme observations. The location and scale parameters of EV_k distribution are mostly estimated using the maximum likelihood (ML) method. Numerical methods should be performed to obtain ML estimates of parameters of EV_k distribution since likelihood equations include intractable terms. However, using numerical methods may be problematic and convergence is not guaranteed. In this study, we therefore utilize Tiku's (1967) modified maximum likelihood (MML) method and thus resulting MML estimators are explicitly formulated. See also Tiku and Akkaya (2004) and Aydin (2017) in the context of MML estimation when k=1. To investigate the effect of different values of k on the performances of the ML and MML estimators a Monte-Carlo simulation study is carried out. Results show that the MML estimators are as efficient as the ML estimators. Therefore, the MML estimators can be preferred to the ML estimators if focus is computational ease besides efficiency. References Aydin, D. (2017). Estimation of the lower and upper quantiles of Gumbel distribution: An application to wind speed data. *Applied Ecology and Environmental Research* 16, 1-15. Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré* 5(2), 115-158. Tiku, M. L. (1967). Estimating the mean and standard deviation from a censored normal sample. *Biometrika* 54, 155-165. Tiku, M.L. and Akkaya, A.D. (2004). *Robust Estimation and Hypothesis Testing*. New Age International Publishers (Wiley Eastern), New Delhi.

KEYWORDS: k-th Extreme Value distribution, Maximum Likelihood, Modified Maximum Likelihood, Efficiency

O-58 Investigation of Patients' Persistence Rate of Antibiotic Prescription by Sensitive Question Method; A Cross Sectional Study

Oral Abstract / Biostatistics

Robab Ahmadian ¹, Ilker Ercan², Yesim Uncu³, Ozlem Toluk¹, Fatma Ezgi Can⁴,

¹Bursa Uludag University, Institute of Health Sciences, Department of Biostatistics, ²Bursa Uludag University, Faculty of Medicine, Department of Biostatistics, ³Bursa Uludag University, Faculty of Medicine, Department of Family Medicine, ⁴Izmir Katip Celebi University, Faculty of Medicine, Department of Biostatistics,

Sensitive question methods are used when the subject under investigation is illegal or undesirable, and respondents are unwilling to give truthful answers to the questions. In our study intended to antibiotics prescription, we investigated insistence of patients and their try to force physician for prescription, by using sensitive question methods. There are many methods account for sensitive subjects, we used crosswise model which is one of the sensitive question methods. The crosswise model has developed for sensitive questions which participants are hesitant to answer directly. In this method, participants receive two questions which one of them is a sensitive question and another one is a non-sensitive. The probability of the non-sensitive question is known to the researcher. To answer the questions, participants are asked to select "A" if both items have the same answer or they are asked to select "B" if the two items do not have the same answer, in this way the sensitive question is not asked directly. For this study a cross-sectional survey using a validated self-administered questionnaire was held in Bursa city. For the sensitive question we asked respondents if they force the physician to prescribe antibiotics with complaining of unrealistic health symptoms. The non-sensitive item asked respondents about their birthday: "Is your birthday in January, February, or March?" In our study, a total of 1050 questionnaires was applied to 1050 patients by Uludag University. As a result of application, 30% of the patients were forced the physician to prescribe antibiotics.

KEYWORDS: Crosswise Method, Antibiotic Usage, Sensitive Question. Survey Method

O-59 Evaluation of User Experience Effects on Ergonomic Behavior with using Rough-SWARA Method

Oral Abstract / Industrial Statistics & Engineering Applications

Sercan Madanlar¹, Şebnem Demirkol Akyol²,

¹The Graduate School of Natural And Applied Sciences, ²Department of Industrial Engineering,

Product ergonomics became more essential factor for consumers day by day as a result of increased product range and competition between firms. In product selection decision, user experience about past product usage is an important factor because it recalls good and bad properties of the products. In this study, we develop a participant experiment to evaluate user experience effects on product selection and ergonomic behavior by using Rough-SWARA method. Since it is hard to determine the relative comparison of criteria as percent by participants, this method is suitable for this particular problem. The pruning shear is selected as an investigated product because it is easy to find a participant who has never used that product as a result of urbanization. The experiment is performed with non-expert thirty-two consumers who have evaluated the pruning shears with a total of nine criteria and eleven different types of pruning shears. The experiment consists of usage and evaluation of eleven different pruning shears by each participant. Participants have ranked the criteria twice which represents the following pair-wise matrices: ranking before experiment and ranking after experiment. Also, there is an option for consumers not to change the first ranking. In this case, the first ranking is transferred to the second matrix. Thus, these two matrices are compared according to criteria weight and major changes are observed between these criteria according to product usage experience. Also, this study shows the demonstration of how easy to apply Rough-SWARA method for participants and experts who are evaluating the criteria.

KEYWORDS: Rough-SWARA, User experience, Product ergonomics, Weight determination

O-60 Unit Lindley-Weibull Distribution and It's Some Properties

Oral Abstract / Statistics & Probability

Coşkun Kuş¹, Kadir Karakaya¹, Buğra Saraçoğlu¹, İsmail Kınacı¹,

¹Selçuk University,

In this talk, a new lifetime distribution is introduced. Motivation is provided to obtain the distribution. The closed form expressions of density function, cumulative distribution function, survival function and hazard rate are given. The raw moments are also presented. Several methods are used to estimate unknown parameters. A numerical example is provided to close the presentation.

KEYWORDS: Lindley Distribution, Weibull Distribution, Compounding, Monte Carlo Simulation, Estimation

O-61 Several Applications of New Generalized Entropy Optimization Methods in Survival Data Analysis

Oral Abstract / Statistics & Probability

Aladdin Shamilov¹, Nihal İnce¹, Sevda Özdemir Çalkıuşu²,

¹Eskisehir Technical University, ²Van Yüzüncü Yıl University,

In this paper, survival data analysis is realized by applying new Generalized Entropy Optimization Methods (GEOM) for solving Entropy Optimization Problems (EOP) consisting of optimizing a given entropy optimization measure subject to constraints generated by given moment vector functions. Mentioned problems in the form of GEOP2, GEOP3 based on GEOP1 have Generalized Entropy Optimization Distributions: GEOD2 in the form of $\text{Min}_{\top} \text{DMaxEnt}$, $\text{Max}_{\top} \text{DMaxEnt}$; GEOD3 in the form of $\text{Min}_{\top} \text{HMinxEnt}$, $\text{Max}_{\top} \text{HMinxEnt}$, where H is the Jaynes optimization measure, D is Kullback-Leibler optimization measure. It should be noted that formulation of GEOP1 uses only one optimization measure (H or D), however each of formulations of GEOP2, GEOP3 uses two measures H, D together. For this reason, survival data analysis by GEOD2 and GEOD3 acquires a new significance. In this research, given survival data is examined as application of developed new method. The performances of GEOD2 and GEOD3 are established by Chi-Square criteria, Root Mean Square Error (RMSE) criteria, H and D measures.

KEYWORDS: Generalized Entropy Optimization Methods, Jaynes optimization measure, Kullback-Leibler optimization measure, Survival data analysis

O-62 Combining Binary and Continuous Biomarkers

Oral Abstract / Biostatistics

Robab Ahmadian¹, Ilker Ercan¹, Deniz Sigirli¹, Abdulmecit Yildiz¹,

¹Uludag University,

The current biomarker combination methods have been developed for combining continuous markers. however, the combination of binary and continuous biomarkers will lead to a more precise conclusion in medical decisions. therefore, in this study, it was discussed the combination of binary and continuous biomarkers. The suggested combination for binary biomarkers was created by an approach using Youden's J statistic for combining binary biomarkers. The proposed approach will facilitate binary and continuous biomarker combinations. A real data of Contrast-induced nephropathy (CIN) was used to compare the performance of our proposed combination approach by parametric AUC calculation method. Also, a simulation study was conducted to compare the performance of our proposed combination approach according to different sample sizes. Both in the analysis of real data and the simulation studies for different samples, the proposed approach has been shown to yield favorable results and higher area under the curve.

KEYWORDS: Binary biomarker, biomarker combination, Roc curves, AUC

O-64 A Panel Data Analysis: The Relationship Between Unemployment, Youth Unemployment and Growth

Oral Abstract / Econometrics

Merve Altaylar¹, Hamdi Emeç ²,

¹Social Sciences Institute / Dokuz Eylül University, ²Faculty of Economics and Administrative Sciences,

This paper examines the relationship between unemployment, youth unemployment and economic growth in 20 OECD countries between 2000 and 2017. These variables are examined using static and dynamic panel time series techniques. He does not examine the relationship between these variables, as Okun suggested, and explores how unemployment and youth unemployment affect economic growth. Conventional panel unit root and panel break root and panel cointegration tests were used to investigate the differences of these variables on economic growth. As a result, youth unemployment is more susceptible to economic growth than unemployment, while developments in youth unemployment do not affect economic growth as well as unemployment.

KEYWORDS: Multiple Structural Break Panel Cointegration, Heterogeneous Panel, Cross-sectional Dependence, DOLS, PANKPSS

O-65 The Modelling of Earthquake Events Based on Bivariate Extreme Value Theory

Oral Abstract / Statistics & Probability

Gamze ÖZEL¹,

¹Hacettepe University,

The objective of this paper is to provide estimation for the probability of such extreme events where the mainshock and the largest aftershocks exceed certain thresholds. Two approaches are illustrated and compared – a parametric approach based on previously observed stochastic laws in earthquake data, and a non-parametric approach based on bivariate extreme value theory. We analyze the earthquake data from the North Anatolian Fault Zone (NAFZ) in Turkey during 1965–2018 and show that the two approaches provide unifying results.

KEYWORDS: extreme value, natural event, parametric approach

O-66 Healthcare Tourism Demand and an Empirical Analysis for Istanbul

Oral Abstract / Industrial Statistics & Engineering Applications

Erkan Işıklı¹, Bilgesu Bayır¹,

¹ISTANBUL TECHNICAL UNIVERSITY,

Healthcare and long waiting lists in developed countries. More people are traveling nowadays to developing countries to receive quality healthcare at lower costs. All the decision making and planning in any industry eventually rely on reliable forecasts. Thus, building a model that accurately reflects the nature and attributes of healthcare tourism demand has a significant impact on the host country's economy. This study attempts to incorporate various time series techniques to provide accurate forecasts for international demand for healthcare tourism to a particular destination, namely Istanbul. Since there is a paucity of research specifically focusing on the empirical investigation of healthcare tourism, studies on the modelling and forecasting of tourism demand since 2010, in general, are perused and four main methods (Holt-Winters, ETS, STL, and SARIMA) are fitted based on the results of the literature review. In addition, various combinations of these models are formed to improve forecast accuracy. SARIMA is found to perform usually better among the four main models and the combination approach that utilizes constrained linear regression is found to have the best overall accuracy in terms of Maximum Absolute Percentage Error and Root Mean Squared Error.

KEYWORDS: Healthcare tourism demand, service systems, time series forecasting, combination forecasts

O-67 Some tests and comparisons for homogeneity of variances

Oral Abstract / Statistics & Probability

Nilgün Nursu ÖZTÜRK¹, Hamza GAMGAM¹, Bülent ALTUNKAYNAK¹,

¹Gazi University Department of Statistics ,

The equality of variances is one of the basic assumptions of classical variance analysis. There are many test statistics known in the literature to determine whether this assumption is met or not. The aim of this study is to introduce the Bartlett (B), Hartley (H), Levene (L), Fligner-Killeen (FK) and Bootstrap (FRMD) tests for testing the homogeneity hypothesis of variances and compare these tests. Using Monte Carlo simulation method, these tests were compared in terms of type I error rate and power under various scenarios. With these comparisons, it was aimed to determine the test suitable for testing the homogeneity hypothesis of variances in both normal distribution and some nonnormal distributions.

KEYWORDS: Simulation, Analysis of variance, Power, Bootstrap

O-68 The Performance of the VSI Tukey-Exponentially Weighted Moving Average Control Chart

Oral Abstract / Statistics & Probability

Selcem ADSIZ¹, Burecu AYTAÇOĞLU¹,

¹Ege University,

Control chart is one of the most important tools in statistical process control. It is used for monitoring the process and determining whether the underlying process is in control or not. Traditional control charts depend on the normality assumption. However, in real life, the normality assumption may not be satisfied. One of the control charts which was proposed recently is the Tukey-Exponentially Weighted Moving Average (Tukey-EWMA) control chart. Tukey-EWMA control chart is designed by combining the feature of Tukey control chart with the classical EWMA control chart. Therefore, this chart uses the quartiles and inter-quartile range in the calculation of the control limits and it was shown to be quite efficient at detecting process shifts especially for skewed distributions. In this study, Variable Sampling Interval (VSI) Tukey-EWMA control chart is proposed and its statistical performance is investigated. In the design of VSI, warning limits are included in addition to the control limits of the existing chart. If the last control statistic is between the warning limits and control limits, next sample is drawn in a shorter time because of the possibility of an out of control situation. Since the sampling intervals are variable, average time to signal (ATS) is used as a performance measure. In order to compute ATS values for several situations, Markov chain approach is implemented by the help of the R software package.

KEYWORDS: VSI,ATS,Tukey-EWMA

O-69 A Metaheuristic Algorithm For In-Plant Milk-Run System

Oral Abstract / Industrial Statistics & Engineering Applications

Islam Altin¹, Aydin Sipahioglu¹,

¹Eskisehir Osmangazi University,

Milk-run, a cyclic material delivering system, aims to increase the efficiency of transportation and supply chain based on lean logistics perspective. There are two kinds of milk-run systems as supplier and in-plant milk-run system in the literature. In-plant milk-run system that has growing appeals with Industry 4.0 concept, is applied to manage process of delivering materials from warehouse to assembly stations in plants. This system is implemented using Automated Guided Vehicles (AGV), which provide automated materials handling in plant. However, a challenging problem arises in determining milk-run routes and periods allowed to be different for each AGV. Since this problem is quite difficult to handle with exact solution methods, Iterated Local Search algorithm with dynamic penalty function is developed in this study. Dynamic penalty function increases efficiency of the proposed algorithm in terms of avoiding local optima by keeping the properties of the last two solutions in memory. Moreover, it prevents exceeding vehicle capacities in obtained solution. In order to evaluate the performance of the proposed algorithm, test problems with different scales derived from the literature are used. The computational results show that the suggested algorithm is efficient to obtain both milk-run routes and periods for each AGV, in a reasonable computational time.

KEYWORDS: Logistics, In-Plant Milk-Run System, Automated Guided Vehicles, Iterated Local Search, Dynamic Penalty Function

O-70 A NEW RISK ASSESSMENT FOR THE RIGHT-SKEWED PROCESSES

Oral Abstract / Industrial Statistics & Engineering Applications

Melis Zeybek¹, Onur Köksoy¹,

¹Ege University,

Poorly operated production units and incomplete designed products cause major incidents involving monetary and social losses. Quality experts commonly utilise loss functions to develop new methodologies for quality improvement. The widespread use of loss functions in industrial applications has increased their popularity among statisticians and engineers due to their different loss-handling features. They mainly emphasize the importance of being on desired target with a small variation to reduce all types of manufacturing costs. This paper presents a new member of the inverted probability loss family. We propose a loss function by inverting the density first introduced by Zeybek and Köksoy (2018) for the responses under gamma noise effect. The proposed loss function has a right-skewed structure along with the range of $(-\infty, \infty)$. The important features of the proposed loss function are discussed under some process distributions of interest.

KEYWORDS: asymmetric quality loss, risk assessment, inverted probability loss family, right-skewed processes

O-71 Analysis of Internal Migration in Izmir with a Binary Logit Model

Oral Abstract / Econometrics

Tuba İlhan¹, Şenay Üçdoğruk Birecikli¹,

¹Dokuz Eylül Üniversitesi,

With the help of the Household Labor Force Survey data for 2014-2017, it has aimed to determine the internal migration to İzmir. In the Household Labor Force Survey, a data set was created based on the question “Which of these settlements you have previously resided? ”.A second model was established by adding İzmir's labor force participation rate and GDP per capita data to the data. The wage income of the individual has been added to the data set by deducting the price effect and taking the logarithm. Binary logit model has used for analysis. A second model was established by adding İzmir's labor force participation rate and GDP per capita rate to the data. According to the results of the binary logit model, which gives the possibility of migration to İzmir, migration decreased from 2014 to 2017. There has a linear relationship between education and age. Another finding is that married individuals are 16% more likely to migrate than unmarried individuals. According to the variables added to Model 2, the increase in İzmir's per capita gross domestic product increased the probability of migration to İzmir by 38%. Labor force participation rate leads to an increase of 1% on the probability of immigration to İzmir.

KEYWORDS: Migration, Internal Migration, Binary Logit Model

O-72 The Effect of Multicollinearity on the Estimators of the Regression Coefficients

Oral Abstract / Statistics & Probability

FİLİZ KARADAĞ¹, HAKAN SAVAŞ SAZAK¹,

¹Ege University, Department of Statistics,

Multicollinearity is the existence of linear relationships among two or more independent (explanatory) variables. The presence of multicollinearity poses a serious problem for regression analysis studies. This problem leads to unstable estimates of the regression coefficients and causes some serious problem in validation and interpretation of the regression model. In this study, the effect of multicollinearity on the regression coefficient estimators is examined with a simulation study. In the simulation study, in case of multicollinearity, the least squares estimation method and a robust method (using MATLAB Robustfit command which uses the Tukey M-estimators as default) are compared and the answer to the question of whether there is a difference in the variances of the regression coefficients between the estimations based on the least squares method and the estimations using robust methods is investigated. In other words, we are investigating that whether robust methods can be a remedy against multicollinearity. The simulation results show that the robust methods are also badly affected by multicollinearity because multicollinearity does not influence the residuals and the robust methods mostly produce solutions for the outliers in the residuals. At the end of the study, a real-life application is given. We also discuss the causes of multicollinearity and its detection methods and the possible remedies.

KEYWORDS: Regression analysis, Multicollinearity, Least squares method, Robust methods.

O-73 M- Estimation Use Pearson Type IV Distribution Weight Function in Robust Regression

Oral Abstract / Statistics & Probability

YASİN BÜYÜKKÖR¹, HATEM ÇOBAN², ALİ KEMAL ŞEHİRLİOĞLU²,

¹KARAMANOĞLU MEHMETBEY ÜNİVERSİTESİ, İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ, ²DOKUZ EYLÜL ÜNİVERSİTESİ, İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ, EKONOMETRİ BÖLÜMÜ,

In many regression applications, the distribution of errors is considered normal and the Least Squares (OLS) method is used to estimate parameters. However, in practice, even if the distribution of errors is assumed to be normal, residuals are not generally normally distributed. If the data contains outliers or there are observations which suspected to be outlier, the assumption of normality is violated and parameter estimates are biased. Many researchers use robust methods when such problems occur. One of these methods is M-estimators. Traditional M-estimators can easily be used when the data is symmetrical. However, traditional M-estimators can not achieve a good solution if the data has skewness and excess kurtosis. The differential equation of the Pearson Type IV distribution, which provides a better solution for both symmetric and asymmetric distributions, can be used as the Influence Function (IF). The commonly used M-estimators do not take into account the skewness and kurtosis measures of the data, while the Pearson Type IV distribution used in the study contains the skewness and kurtosis parameters. In this study, it is shown that Pearson differential equation can be used as Influence Function. By using the probability density function of Pearson Type IV distribution, the Objective Function, Influence Function and Weight functions are obtained. For parameter estimation, Iteratively Re-Weighted Least Squares Estimation (IRWLS) method is used and parameter estimations are made on many real data sets. In addition, simulation studies with different scenarios are performed. The method used is compared with the performance of other M-estimators.

KEYWORDS: M- Estimation, Robust Regression, Pearson Type IV Distribution, Iteratively Re-Weighted Least Squares

O-74 ARTIFICIAL NEURAL NETWORK APPROACH TO RESPONSE SURFACE MODEL FOR UPPER LIMB PERFORMANCE IN PATIENTS WITH CHRONIC NECK PAIN

Oral Abstract / Statistics & Probability

Leyla Bakacak Karabeni¹, Serpil Aktaş Altunay¹,

¹Hacettepe University,

Response surface model (RSM) is used to detect the variable values that make the response variable maximum or minimum. Besides, the effect of exploratory variables on the response variable is determined. Thus, this method can be referred as a combination of regression analysis and optimization. RSM is mostly used in many fields such as industry and chemistry. However, it has limited application in the field of health. The upper limb performance assessment is a two-stage assessment of upper limb contributions to task performance. In this study, the upper limb performance of chronic neck pain patients is examined on 63 patients. The score of Hand20 questionnaire identifying the performance of upper limb is assigned as response variable. Input variables are taken as the variables related the pain-rating scales of patients at rest or in activity. The central composite model is implemented to estimate the second-degree polynomial model. The artificial neural network approach is also applied to upper limb performance data. The root mean square error, correlation coefficients and standard error of prediction are obtained from evaluating the experimental and predicted values of both models. The comparative analysis for both models is made on the prediction accuracy. The optimum point of this region is estimated for the upper limb performance.

KEYWORDS: response surface model, optimization, artificial neural network, upper limb performance

O-75 DIAGNOSTIC META-ANALYSIS: AN APPLICATION IN DENTISTRY

Oral Abstract / Biostatistics

Merve PARMAKSIZ¹, Hayal BOYACIOĞLU¹, Pelin GÜNERİ²,

¹Ege University Department of Statistics, ²Ege University Department of Oral Diagnosis and Radiology,

In the study application of diagnostic meta-analysis in dentistry using various software programs were examined. Meta-analysis is a systematic review of a focused topic in the literature that provides a quantitative estimate for the effect of a treatment intervention or exposure. In this scope different estimation methods comprising DerSimonian and Laird (DL), Restricted Maksimum Likelihood (REML), Sidik and Jonkman (SJ), Hedges and Olkin (HO), Maksimum Likelihood (ML), Paule and Mandel (PM) were compared. In the implementation part, effectiveness of clinical oral examination (COE) in predicting the diagnosis of histological dysplasia or oral squamous cell carcinoma (OSCC) was studied.

KEYWORDS: Meta-Analysis, Diagnostic Test, Diagnostic Odds Ratio (DOR), DerSimonian and Laird (DL)

O-76 Quasi-Maximum Likelihood Estimator based on Moyal Distribution for Censored Data

Oral Abstract / Statistics & Probability

Ismail Yenilmez¹, Ilhan Usta¹, Yeliz Mert Kantar¹,

¹Eskisehir Technical University,

The censored data, in which the observed value of some variable is partially known, is related to data-gathering mechanism. In the context of regression analysis, while Ordinary Least Squares method (OLS) is used for full data, Tobit estimator is one of the well-accepted estimation methods for censored data. It is known that OLS gives biased and inconsistent results at different censorship levels and points. Tobit uses the maximum likelihood (ML) method under the assumption that the errors are based on the normal distribution. Tobit is known to be consistent and effective if the assumption of normality is maintained. However, many examples can be presented where the assumption of normality is not provided. At this stage, one of the various alternative methods is Quasi-Maximum Likelihood estimators (Q-ML). In this study, the Q-ML based on Moyal Distribution (MD) is introduced for censored data. The simulation results show that if the errors are not normal, Q-ML based on MD has a smaller bias and mean square error (MSE) value.

KEYWORDS: Censored data, Tobit, Quasi-Maximum Likelihood estimator, Moyal Distribution

O-77 A semi-parametric method to detect outbreaks in syndromic surveillance

Oral Abstract / Biostatistics

İmren Saygır Yılmaz¹, Eralp Doğu¹, Dursun Aydın¹,

¹Mugla Sitki Kocman University,

In public health practices, it is aimed to identify as soon as possible a symptom that can produce an outbreak. For the contagious disease, outbreak periods vary every year. This complicates the using of parametric regression models that require a number of assumptions. So, in this study, as an alternative to the parametric model called as Adaptive Regression Model with Sliding Baseline, a “Semi-parametric Regression Model with Sliding Baseline is proposed. A simulation study was performed by using a set of scenarios for performance comparison and One-Sided CUSUM is applied to the residues obtained from these two models. The performance evaluation criteria introduced within the scope of this study are compared by considering sensitivity, specificity and accuracy criteria. Finally, the advantages and disadvantages of the proposed method are presented at the end of the simulation study. Acknowledgement This study is supported by Mugla Sitki Kocman University, Scientific research project office (BAP) with project number 15/160.

KEYWORDS: Public Health, Syndromic Surveillance, Statistical Process Control, CUSUM

**O-78 DETERMINATION OF FACTORS AFFECTING LOCAL SELECTION RESULTS
THROUGH CLASSIFICATION TREES**

Oral Abstract / Statistics & Probability

İpek Deveci Kocakoç¹, İstem Köymen Keser¹,

¹Dokuz Eylül Üniversitesi,

The aim of this study is to estimate the relationship between the winning party in local elections and the social, economic, geographical and demographic characteristics of the provinces by using a Classification Tree (CART). In this study, economic, social, geographical and demographic variables were accepted as independent variables and the winning parties according to 2014 and 2019 local elections were accepted as dependent variables. The aim of the study is to find out which variables affect the voting decision the most, to try to identify a model that will lead the political campaigns and to reveal the significant differences between the two elections. Among the many classification algorithms, the R-coded C5.0 algorithm was used. Compared to more advanced and sophisticated machine learning models, decision trees under C5.0 generally work almost the same way, but the use of the C5.0 algorithm is preferable in the study because it is easier to understand and use and accommodates multi-class problems. The C5.0 algorithm determines the separation criterion with the largest information gain in each decision node and optimum separation takes place. In the study, it was observed that in 2014 local elections especially household size and age distribution were prominent, and in 2019 local elections, the weight of the effects on the distribution of votes in educational and economic variables increased. In addition, the provinces with high population density were also discussed in the study and the positions of the parties were determined in terms of the variables involved in multivariate classification techniques.

KEYWORDS: Classification Trees, C5.0 algorithm, Voting distribution

**O-79 FORECASTING THE NUMBER OF PATIENTS ARRIVING AT A HOSPITAL
EMERGENCY DEPARTMENT**

Oral Abstract / Statistics & Probability

ASLI KILIC¹, MURAT ERSEL¹,

¹EGE UNIVERSITY,

Emergency departments play an important role in providing timely treatment to patients and for this reason, they are considered among the basic and most important elements of health systems. Since the number of patients arriving at emergency departments fluctuates over time, it has a stochastic behavior and this situation complicates the estimation of the demand for emergency medical services. Inaccurate estimation of the demand for medical services leads to ineffective use of limited resources, thus prolonging waiting times for treatment and sometimes even loss of life. For this reason, effective management of emergency medical services with limited labor and material resources is a priority for hospital managers. In this study, based on the number of patients arriving at the emergency department of a training and research hospital, the number of patient visits is tried to be predicted using statistical forecasting techniques. The obtained forecasting model is intended to be used as a supportive tool for management decisions in terms of resource and business planning.

KEYWORDS: Emergency department patient arrivals, forecasting, modeling

O-80 Estimation of right-censored time-series with semi-parametric regression model

Oral Abstract / Statistics & Probability

Ersin Yılmaz¹, Dursun Aydın¹,

¹Mugla Sitki Kocman University,

This paper aims to estimate the right-censored time series with penalized spline method based on two different censorship solution techniques, k-nearest neighbours (kNN) imputation method (Batista and Monard, 2002) and Kaplan-Meier weights (KMW) (Kaplan and Meier, 1958; Miller, 1976; Stute, 1993). In this paper, the optimal modelling procedure is tried to conduct by using appropriate methods in terms of both estimations and solving the censorship. It is proved by Aydın and Yılmaz (2018) that penalized spline method is stronger than other smoothing methods on semi-parametric estimation of the right-censored data. In addition, kNN imputation and Kaplan-Meier weights have their own advantages and disadvantages. The purpose of this study is to find out which method gives better results when right-censored time-series are considered. A simulation study is carried out and results are presented to see the performance of methods. Accordingly, kNN imputation method is detected as a better method for its fully nonparametric nature. Also, KMW gives satisfying results for lower censoring levels.

KEYWORDS: Censored time-series, kNN imputation, Kaplan-Meier weights, penalized splines

O-81 Quantification of verbal assessments using hesitant fuzzy sets: Computing with words

Oral Abstract / Statistics & Probability

Murat Alper Basaran¹,

¹Alanya Alaaddin Keykubat Üniversitesi,

Quantification of verbal assessments using hesitant fuzzy sets: Computing with words Verbal assessments are widely used in researches ranging from social sciences to engineering fields. Generally, verbal assessments are needed for situations whose defining attributes are ill-defined or too complex to define in terms of quantitative manner. Hence, verbal statements or linguistic terms are instead are used to measure and to quantify those attributes. While fuzzy sets proposed by Zadeh was the first to be employed in those constructions, hesitant fuzzy sets based on fuzzy set theory are widely used tools and provide more comprehensive findings in various applications. The verbal statements or linguistic terms are transformed into membership functions to make computations, which are called computing with words, In this manuscript, a new method and its implementation will be provided using hesitant fuzzy sets.

KEYWORDS: Hesitant fuzzy sets, computing with words, verbal statements

O-82 An Open Source Decision Tree Interface for MATLAB

Oral Abstract / Statistics & Probability

Ipek Deveci Kocakoç¹, Metin Öner²,

¹Dokuz Eylül Üniversitesi, ²Celal Bayar Üniversitesi,

Decision trees help us to understand different combinations of data attributes that produce the desired result. Decision trees are used to enhance a record with additional data sources to optimize a process for a better result. The structure of the decision tree reflects the structure that can be hidden in the data. In these tree structures, leaves represent class labels, and branches represent a combination of properties (features) leading to these class labels. Decision trees are widely used as a predictive model and are matched with observations of an item to conclude the target value of the item. There are many algorithms for regression and classification type of decision trees. The purpose of this study is to code a Matlab interface that can easily be used for forming decision trees. Although Matlab has a Classification Toolbox, it is a commercial one and not free of charge. Algorithms such as Decision Tree (DT), Random Forest (RF), VIBES, and Gradient Boosting (GB) are combined in one Matlab GUI with their parameter options. GUI is designed to be used in both English and Turkish. Available results of algorithms such as confusion matrix, cross validation accuracy, ROC curves, feature importance scores and sample trees are given by the GUI.

KEYWORDS: Decision tree, classification tree, regression tree, Open source, MATLAB

O-83 Evaluations of the Mean Residual Lifetime Function of a Multi-state System

Oral Abstract / Statistics & Probability

Funda Iscioglu¹

¹Ege University, Department of Statistics,

Mean residual lifetime function is one of the important performance characteristics in reliability and survival analysis. This characteristic is well studied in binary system structures. However it has attracted the attention of many researchers studying multi-state system structures recently. In multi-state system modelling the system or the components can have more than just two states. Therefore different from binary case, the reliability measure for multi-state systems should consider all those states that are above some intermediate threshold state. Thus this makes the mean residual lifetime evaluation an important research problem. In this study we consider a one unit system having just three-states for simplicity, where “0,1 and 2” indicates failure, partial functioning and perfect functioning states, respectively. We deal with the evaluation problem of the mean residual lifetime function for this system in case of the lifetimes spent at state “1” and at state “2” are dependent. Different lifetime distributions are considered as well. We showed how the dependency parameter affects the mean residual lifetime functions considering different states. Besides we discussed the results when the lifetimes spend at each state are independent.

KEYWORDS: mean residual lifetime function, multi-state sytem, dependency, lifetime distribution

O-84 The Relationship between Health, Education Expenditures and Economic Growth: The Case of NUTS-1 Regions in Turkey

Oral Abstract / Econometrics

Aygül Anavatan¹, Zerife Yıldırım²,

¹Pamukkale University, ²Harran University,

Education and health expenditures are among the factors that increase human capital. The contributions to human capital are expected to increase economic growth, as well. It is examined as to whether the causality exists, or not, for the period covering the years 2004-2019 in Turkey according to the Nomenclature of Territorial Units for Statistics (NUTS) at level 1. The aim of the study is to reveal the relationship between health and education expenditures and economic growth. The three models were established related to investigate affect health expenditure, education expenditure and economic growth and these models have been investigated. In the first stage, the stationary analysis was performed for variables according to Levin Lin Chu, Hadri, Breitung, Im Pesaran Shin, Fisher ADF, and Fisher PP unit root tests. In the second stage, the existence of a long-run relationship between variables was investigated by Pedroni panel cointegration and Kao panel cointegration test. After that, the causality relationships between variables was investigated by the bootstrap panel causality test. Finally, in the line with an established model both for panel and regions long-run coefficient between variables were estimated by Fully Modified Ordinary Least Squares (FMOLS) and Dynamic Ordinary Least Squares (DOLS).

KEYWORDS: Economic Growth, Health Expenditure, Education Expenditure, Fully Modified Ordinary Least Squares (FMOLS), Dynamic Ordinary Least Squares (DOLS)

O-85 Reliability Analysis of Phased Mission System Using Markov Approach with Repairable Components

Oral Abstract / Statistics & Probability

Sibel Yılmaz¹, Özge Elmastaş Gültekin¹

¹Ege University, Department of Statistics,

Most reliability analysis techniques suppose that the systems operate in a single phase mission. However, many missions in nature have multiple phases. The mission of system usually consists of multiple phases in operation is called Phased Mission Systems. Phased mission systems have wide applications in engineering practices, nuclear power, chemistry, textile and many fields. In these systems, system structure, success criteria and failure rate of the components may vary from phase to phase. There are many methods to perform reliability analysis of these systems such as Fault Tree Analysis (FTA), Binary Decision Diagrams (BDD), Markov Methods, etc. The methods that have been developed to analyze the phased mission systems can be categorized for non-repairable or repairable systems. In this study, different system structures with repairable components in phased mission systems are discussed. New assumptions regarding repair and failure of the components are considered for these system structures and Markov approach is used for calculating the reliability of the systems.

KEYWORDS: Reliability analysis, phased mission systems, Markov approach, repairable components

O-86 A Golden Ratio Control Chart for Monitoring the Process Mean

Oral Abstract / Industrial Statistics & Engineering Applications

Elif KOZAN¹, Onur KÖKSOY¹,

¹Ege University,

The monitoring of the process mean is usually accomplished by the quality control charts. This work presents a new control chart based on the well-known “Golden ratio”. The Golden ratio (GR) control chart directly incorporates with all the information in the sequence of sample values by plotting the weighed values of the sample via the golden ratio number and median. When the small shifts are important, the Shewart control chart may not be a good option for monitoring the process mean; however, the exponentially weighted moving average (EWMA) control chart is known to be effective in the literature. This novel GR control chart, which is thought to be effective in small shifts, may be used alternatively to EWMA. In this paper, GR control chart is compared with the Shewart and EWMA control charts. Also, the proposed procedure and its merits are illustrated with an example.

KEYWORDS: Golden ratio number, Median, Shewart control chart, EWMA

O-87 On Using Structural Patterns in Data for Classification

Oral Abstract / Statistics & Probability

Guvenc Arslan¹, Bergen Karabulut¹, Halil Murat Ünver¹,

¹Kırıkkale University,

In recent years one may observe many new approaches in learning algorithms. For classification, there are now some interesting approaches such as semi-supervised algorithms, algorithms that learn distance functions, and various extensions and generalizations of support vector machines. In this study we propose a new clustering algorithm that uses similarities only and is used as an intermediate step for classification. The motivation for this combined approach is to obtain information from the data set that can be used for classification. After obtaining a clustering of the data set with the proposed clustering algorithm, we apply different strategies for classification. The results on some data sets show that this approach can have some advantages. For example, when using support vector machines the size of the training set is reduced while at the same time comparable performance results are obtained with a smaller number of support vectors.

KEYWORDS: structural pattern, clustering, classification, support vector machine

POSTER ABSTRACT

P-01 Time Series Analysis of Rice Prices using Box-Jenkins ARIMA Methodology in Hargeisa, Somalia

Poster Abstract / Statistics & Probability

Abdishakur Ismeal Adam¹, Vedide Rezan Uslu¹,

¹Ondokuz Mayıs University,

The international prices of agricultural commodities have been increasing considerably. This upward trend, which may cause a new food crisis, has attracted the attention of the world. Several explanations for these movements in prices have been provided by analysts, researchers, and development institutions. The main purpose of this study was to determine and get forecasts of rice prices in Hargeisa, Somalia by using Box-Jenkins ARIMA modeling. Rice prices in Hargeisa were examined in order to identify if it is stationary or not. In order to check if it is stationary we have used time series plot, correlograms and done Augmented Dickey-Fuller test. The results revealed that the data is non-stationary. We have used some approaches such as taking differences to make the data stationary. After getting it stationary we have determined some time series Box-Jenkins models as candidates. After that the determined models were compared with respect to the model accuracy criteria such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Then we have found the best model fitted well to the data set. After doing diagnostic checking we have calculated the forecasts. And all of the results of whole analysis were presented. The outcome of this study can aid both Somalia government and policy makers in making optimal production decisions and in managing overall price risks.

KEYWORDS: Agricultural Commodities, Trend, Box-Jenkins ARIMA modeling, Stationary, Somalia

P-02 Nonparametric Modelling Via Wavelet Smoothing

Poster Abstract / Statistics & Probability

BERNA YAZICI¹, MARWA BENGHOUL¹, MUSTAFA ÇAVUŞ¹,

¹ESKİŞEHİR TEKNİK ÜN. ,

Sometimes, the parametric features of the model can be violated or be too restrictive in some applications. If the parametric model is applied inappropriately, the conclusions from the analysis may be misled. Therefore, the nonparametric models are applied. For the nonparametric features, it is crucial to apply a smoothing approach, the most popular ones are smoothing splines, penalized splines, kernel approaches, and regression polynomial splines. Recently, wavelets decomposition has been known as a powerful mathematical tool applied in various domains such as image processing, audio processing, signal denoising and mentioned as a smoothing approach. Indeed, this paper has an objective to figure out the efficiency of wavelets analysis as a smoothing approach. A longitudinal data produced by the AIDS Clinical Trials Group, sponsored by NIAID/NIH will be used. Different smoothing methods have been used to be compared with the wavelet's decomposition. The results show that the wavelet decomposition demonstrate an impressive capacity as the classic known methods.

KEYWORDS: wavelets decomposition, smoothing, nonparametric model, longitudinal data

FULL TEXTS

O-04 Investigation by Simple Correspondence Analysis of Regional Distribution of Hospitals and Beds in Turkey

Ezgi Güler^{1*}, Gülşen Akman² and Zerrin Aladağ³

¹*Industrial Engineering/ Faculty Of Engineering, Bilecik Şeyh Edebali University, Turkey*

²*Industrial Engineering / Faculty Of Engineering, Kocaeli University, Turkey, akmang@kocaeli.edu.tr*

³*Industrial Engineering / Faculty Of Engineering, Kocaeli University, Turkey, zaladag@kocaeli.edu.tr*

Abstract

Correspondence Analysis is a technique that allows interpretation for categorical variables, facilitating the interpretation of the relationships or similarities and differences between row and column variables in cross tables, and illustrating changes in a graphical dimension. The main objective of the method, which has two types; simple and multiple correspondence analysis; to show the relationship between variables and categories of variables graphically and to reduce the size of the cross-tables with simple factors to obtain this representation. Basic investments in the health sector require a relational analysis of the current situation. In this study, the number of hospitals and beds of all regions in 2017 was obtained from the most recently published TURKSTAT (TÜİK) database and simple correspondence analysis was performed on the data. In the quantitative sense, which hospital types are compatible with which regions and the reasons for this situation were interpreted. As a sub-investigation, a separate simple application analysis was conducted in order to examine the distribution of patient beds according to regions divided into hospital categories. SPSS software was used in all analyzes. The findings of the analysis were interpreted comparatively.

Keywords: *Correspondence Analysis, Categorical Data Analysis, Health.*

1. Introduction

Most of the researches in health sciences contain qualitative variables. The chi-square analysis, which is frequently used in the evaluation of qualitative data, may be insufficient in analyzing large contingency tables. In such cases, correspondence analysis may be suggested as a better method. (Özgür et al., 2017). The analysis of the contingency table, is a very important component of multivariate statistics with many different types of analysis dedicated solely to this type of data set (Beh, E., 2007). Correspondence analysis is a descriptive multivariate statistical technique used in cases where the relationships between variables are examined with two or more dimensional crosstabs (contingency table). As a result of this analysis, the relationships between the categories of each variable are interpreted by analyzing graphically. The two main objectives of the correspondence analysis are; to show graphically the relationship between row and column categories in crosstabs, to develop simple factors that provide this representation and to reduce the size of crosstabs. In correspondence analysis, the data must be obtained categorically or analyzed by making the data categorically.

The purpose of health services; to meet the health needs of the society in high quality, at the desired time and at the lowest possible cost. Developing technology, increasing costs and patient complaints in the health sector necessitate the service to be performed in a more advanced and planned way (Zerenler and

Öğüt, 2007). Basic investments in the health sector require a relational analysis of the current situation. In this study, the number of hospital facilities and bed numbers of all provinces in 2017 were obtained from the most recently published TurkStat database. A simple correspondence analysis was performed on the data. In the quantitative sense, which hospital types are compatible with which regions and the reasons for this situation were interpreted. As a sub-examination, a simple simple correspondence analysis was conducted to examine the distribution of the bed numbers by region. “SPSS” software was used in all analyzes. The findings of the analysis were interpreted comparatively.

2. Materials and Methods

Correspondence analysis is a multivariate analysis technique that graphically shows the variability between the categories of categorical and continuous variables that can be tabulated in dimensions in a less dimensional space environment with the help of inertia calculated from chi-square distances or euclidean distances between the categories (Johnson and Wichern, 2007). Correspondence analysis is performed with the frequency type data belonging to the categories of the cross-tabulated X and Y variables of type $r \times c$ or the multidimensional X, Y and Z variables of type $r \times c \times m$.

The solution process of the correspondence analysis is divided into 2 stages and each stage consists of 3 stages. First, the profiles, weights and distances are determined during the analysis, taking into account the categories of one of the variations. Row and column profiles are calculated from the frequency or frequency table obtained according to the categories of variables and categories are given in the coordinate plane with the help of the obtained coordinates (Kılıç A.F., 2016).

In the graphs obtained as a result of the correspondence analysis, the categories of variables with high or significant relationships are close to each other and the categories of variables with low relationship are far from each other. If there is a similar relationship between variable categories, the distance between them is small and if there is an inverse relationship, the distance between them is greater (Özkoç, 2013). Basic concepts used in correspondence analysis; crosstab, profiles, weights, inertia, coordinates and eigenvalues, dimension. These concepts are briefly introduced below:

Crosstab: Rows and columns contain variable categories and frequencies observed in each cell.

Profiles: Correspondence analysis starts with the conversion of frequencies in the cross table into ratios. At this stage, the percentages obtained according to row totals are called row profiles and the percentages obtained according to column totals are called column profiles.

Weights: The concept of weight is calculated by the total sum of the rows or columns. The purpose of weighting is to ensure that each response contributes equally to each profile point.

Inertia: In the correspondence analysis, total inertia is used as a measure of the total variance in a table. This term is used for variance and refers to Chi-Square distances. Total inertia is a measure of the spacing of the distribution of profile points around the center of gravity with a mass of the Euclidean distance.

Coordinates and Eigenvalues: A point in the correspondence analysis is defined as the categories of variables in the analysis. The coordinates provide information about the position of the points in the dimensions and are interpreted as the relative position of the points on the dimensions. The eigenvalue is the total inertia divided by dimensions.

Dimensions: In correspondence analysis, the aim is to obtain the least possible graph to explain the relationship between row and column variables. In the correspondence analysis, the number of dimensions is determined as $\min[(\text{number of rows}, \text{number of columns}) - 1]$ (Kılıç A.F., 2016).

In this study, an analysis was made in terms of the distribution of the number of hospital facilities and the number of beds to the regions in order to supply the demands of the health sector and the patient demands. The relationship between regions and types of hospitals in terms of the number of hospital facilities and available bed capacities will be examined. For this purpose, simple correspondence analysis was performed. TurkStat's current database provides the total number of hospital facilities and beds for each province in 2017. All provinces were gathered into regions and categories were created for the "region" variable. The categories of the existing "hospital type" variable were identified as University Hospitals, Private Hospitals and Hospitals under the Ministry of Health. Table 1 shows the number of hospital facilities for each hospital category in 7 regions.

There are 2 variables in the research problem: “hospital type and region”. The total number of categories is 10 (Marmara Region, Black Sea Region, Aegean Region, Mediterranean Region, Eastern Anatolia Region, Southeast Anatolia Region, Central Anatolia Region/Hospitals under the Ministry of Health, University Hospitals, Private Hospitals).

In the first problem of the research, the distribution of hospital types in terms of the number of hospital facilities was examined. The correspondence table is located in Table 1. The variables "hospital_type and region" are assigned as rows and columns, and categories are defined.

Table 1. The Distribution Of Hospital Types In Terms Of The Number Of Hospital Facilities

region	hospital_type			
	Hospitals under the Ministry of Health	University Hospitals	Private Hospitals	Active Margin
Mediterranean	81	8	88	177
Southeastern Anatolia	78	3	47	128
Aegean	118	7	69	194
Eastern Anatolia	105	5	21	131
Black Sea	172	5	33	210
Central Anatolia	162	18	80	260
Marmara	163	22	233	418
Active Margin	879	68	571	1518

The summary table of SPSS analysis results is shown in Table 2.

Table 2. Summary Table

Summary								
Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,319	,102			,959	,959	,023	,002
2	,066	,004			,041	1,000	,027	
Total		,106	161,520	,000 ^a	1,000	1,000		

a. 12 degrees of freedom

The table shows how much of the variability is explained. These ratios are used when deciding on the number of dimensions. The eigenvalues of the dimensions are 31.9% and 6.6% respectively. The fact that inertia is different from 0 indicates that there is a relationship between rows and columns. Hospital types vary according to regions.

The graphical representation for the first problem is shown in Figure 1.

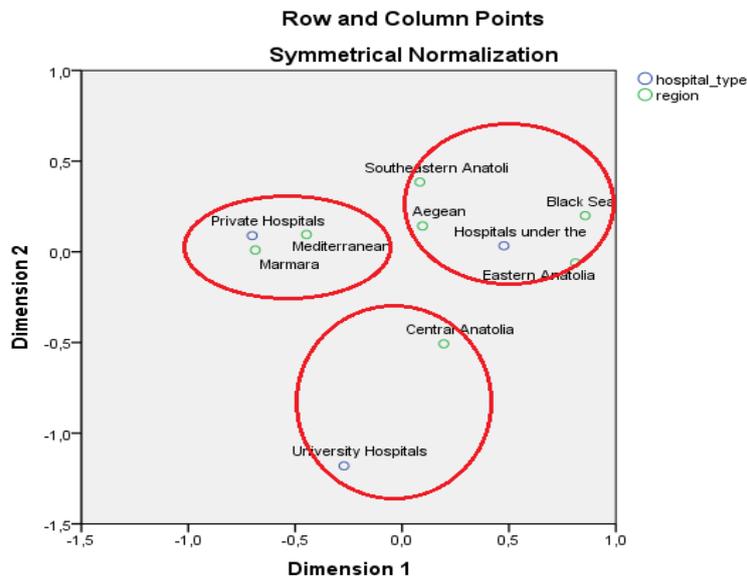


Figure 1. Graphical Representation

This situation may indicate that the structure in the Central Anatolia Region is more prone to academic competence in the field of health. It can be concluded that the facilities provided to private institutions and entrepreneurs providing health services are wider in the Mediterranean and Marmara regions and that these regions are more compatible with such investment decisions. Hospitals under the Ministry of Health, which is frequently preferred by patients and health personnel, can be said to have an intensive distribution in other regions.

As a second sub-research problem, after the simple correspondence analysis based on the number of hospital facilities criteria, correspondence analysis was performed for the number of beds of the same variable.

A comparison was made for 2 different quantitative health care criteria (number of hospital facilities, number of beds).

The distribution of hospital types in terms of the number of beds was examined. The correspondence table is located in Table 3.

Table 3. The Distribution Of Hospital Types In Terms Of The Number Of Beds

region	hospital_type			
	Hospitals under the Ministry of Health	University Hospitals	Private Hospitals	Active Margin
Mediterranean	16829	5186	7300	29315
Southeastern Anatolia	12369	2872	4947	20188
Aegean	18599	5628	5686	29913
Eastern Anatolia	13669	2465	1596	17730
Black Sea	18466	3255	3151	24872
Central Anatolia	25510	7057	8938	41505
Marmara	33897	12861	15582	62340
Active Margin	139339	39324	47200	225863

The summary table of SPSS analysis results is shown in Table 4.

Table 4. Summary Table

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	,154	,024			,927	,927	,002	-,022
2	,043	,002			,073	1,000	,002	
Total		,026	5784,938	,000 ^a	1,000	1,000		

a. 12 degrees of freedom

The table shows how much of the variability is explained. These ratios are used when deciding on the number of dimensions. The eigenvalues of the dimensions are 15.4% and 4.3% respectively. The fact that Inertia is different from 0 indicates that there is a relationship between rows and columns. Hospital types vary according to regions.

The graphical representation for the second problem is shown in Figure 2.

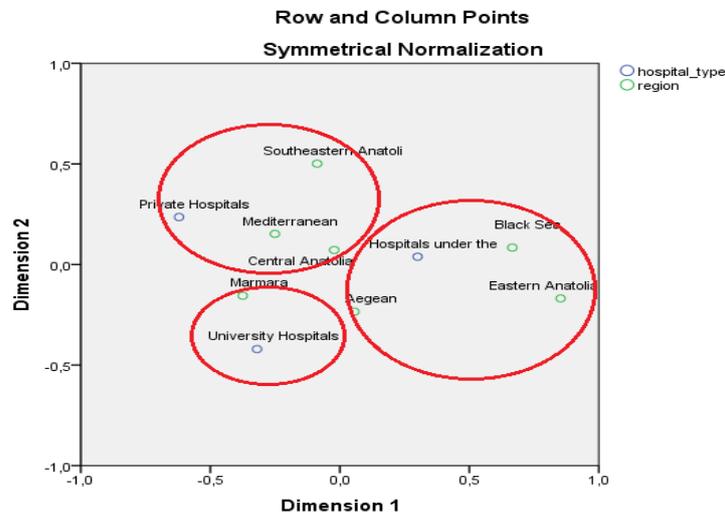


Figure 2. Graphical Representation

When the number of beds is examined, it is concluded that private hospitals are compatible with Southeastern Anatolia, Mediterranean and Central Anatolia Regions, university hospitals are compatible with Marmara Region and hospitals under the Ministry of Health are compatible with other regions.

3. Conclusion

According to the results of the analysis, the compatibility of hospital types with the regions varied for 2 different health care criteria. Private hospitals in the Mediterranean and Marmara Regions; university hospitals in the Central Anatolia Region; the distribution of hospitals under the Ministry of Health in Aegean, Eastern Anatolia, Southeastern Anatolia and Black Sea Regions is related and intensive.

In this context, it can be said that the facilities provided to private institutions and entrepreneurs providing health services in the Mediterranean and Marmara regions are wider and that these regions are more compatible for such investment decisions. The Central Anatolia Region " Hospitals under the Ministry of Health " or "Private hospitals" rather than "University Hospitals" is an area compatible with.

When the number of beds categorized according to hospital types in the regions is examined, the number of beds in Private Hospitals is concentrated in Central Anatolia, Mediterranean and Southeast Anatolia Regions. Unlike the first analysis, Central and Marmara Regions were replaced by Private Hospitals and University Hospitals. The density of the number of beds in Southeast Anatolia Region is higher in Private Universities.

In general, the number of hospital facilities for University Hospitals is compatible with the Central Anatolia Region. Marmara Region is more suitable for University Hospitals in terms of the number of beds.

The facilities of the Hospitals under the Ministry of Health were found to be compatible with the Southeastern Anatolia Region, while this region is compatible for private hospitals in terms of the number of beds. Bed capacities in the same hospital categories vary according to the number of facilities for 3 regions (Central Anatolia, Marmara and Southeastern Anatolia). This finding shows the change in the number of hospital facilities and number of beds according to regions.

Different variables (health facilities in the regions, type of budget allocated, health personnel, health equipment and device type, etc.) can be included in the study in order to determine the factors related to the distribution of the resources in the field of health and to present the current situation. A large number of simple correspondence analyzes and multiple correspondence analyzes can be performed.

References

Beh, E.J. (2004), “Simple Correspondence Analysis: A Bibliographic Review”, *International Statistical Review*, vol.72, no. 2, pp.257-284.

Clausen, S.E. (1998), *Applied correspondence analysis: An introduction*, 1st ed., Sage Publications, California,US.

Johnson, R. A., Wichern, D. W. (2007), “Applied Multivariate Statistical Analysis” 1st ed. Prentice Hall, New Jersey, US.

Kılıç, A.F. (2016), “Uyum Analizi (Correspondence Analysis)”, Adıyaman Üniversitesi YBS Ansiklopedi, 3rd ed., Adıyaman, Türkiye.

Özgür, E.G., Bekiroğlu, N., Baydemir, C. (2017), “Sağlık Bilimlerinde Çoklu Uyum Analizi Ve Uygulaması”, Kocaeli Üniversitesi Sağlık Bilimleri Dergisi, vol. 3, no. 2, pp.9-18.

Özkoç H. (2013), “Hastaların Sağlık Kurumu Tercihlerini Etkileyen Faktörlerin Belirlenmesi: Uygunluk Analizi Ve Nested Logit Model”, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, vol.15, no.2, pp.267-280.

Van De Geer, J. (1993), Multivariate Analysis Of Categorical Data: Applications, 1st ed., Sage Publications Inc, USA.

Zerenler, M., Ögüt, A. (2007), “Sağlık Sektöründe Algılanan Hizmet Kalitesi Ve Hastane Tercih Nedenleri Araştırması: Konya Örneği”, Selçuk Üniversitesi Sosyal Bilimler Dergisi, vol. 18, pp. 501-519.

O-05 A Regression Analysis for Predicting the Amount of Yearly Greenhouse Gas Emissions in Turkey

Merve AVCI¹ and Banu YETKİN EKREN²

¹Department of Industrial Engineering, Yasar University, Turkey, merve_7024@hotmail.com

²Department of Industrial Engineering, Yasar University, Turkey, banu.ekren@yasar.edu.tr

Abstract – In this paper, we study a stepwise multiple linear regression analysis in order to predict the amount of yearly greenhouse gas (GHG) emissions based on weather related data in Turkey. Greenhouse gas is a compound of gaseous capable of absorbing infrared radiation, resulting with trapping and holding heat in the atmosphere. By the increased heat in the atmosphere, greenhouse gases cause greenhouse effect, which ultimately leads to global warming. By the presented study, we both investigate the statistically significant factors affecting the yearly GHG emission amount and the existence of a proper regression function that is able to predict the yearly GHG emission amount accurately based on those significant factors. Thus, while the output is considered to be the yearly GHG emission amount, the input factors are considered to be: annual average weather temperature ($^{\circ}C$), relative humidity (%), hours of sunshine (h), average sea water temperature ($^{\circ}C$), and rainfall (mm/m^2). By those input factors, we also search whether or not there is relationship of yearly GHG emission amount on climate change. We use real data for the analysis which are obtained from the Turkish Statistical Institute’s web site as well as Turkish state Meteorological Service for the years of 1990 and 2017. The results show that there is significant effect of GHG emission on climate change and there exist a good fit regression function estimating the yearly GHG emission amount based on those climate-based input factors.

Keywords – Regression, greenhouse gas emission, stepwise regression, climate

1. Introduction

The climate change has become an important global issue that is largely caused by increased greenhouse gas (GHG) emissions resulting from human activities (Gething and Puckett, 2013). According to the "2018 Global Carbon Budget" report, the yearly global GHG emissions are estimated to be 37 billion tons in 2018 that is an increase of 2.7% from 2017 (Corinne et al., 2018). If no further actions are taken to reduce the GHG emissions, global warming is likely to exceed $2^{\circ}C$ above the pre-industrialized levels (Zheng et al., 2019). This issue would also have a significant impact on the world's landscape and sea levels also affecting the economic and social development of countries all over the world. Hence, it is important to identify efficient ways to manage GHG emission to decrease its negative effect. One main issues for this target may be predicting the GHG emissions level accurately to take possible precautions when there is significant increase.

The 2014 Intergovernmental Plan of Climate Change (IPCC) report declares that since these gases comprise 76%, 16%, 6%, and 2% of total global GHG emissions, the most important GHGs are: CO_2 , CH_4 , N_2O , and F -gases, respectively (Myhre et al., 2013). Hence, in this study as GHG emissions, we consider the total amount of those gases oscillated to the atmosphere. We study the data for the years from 1990 to 2017. Since it is estimated that there is significant relationship of GHG emissions on climate change, in this work we also seek and prove this relationship. We utilize a multiple linear regression

analysis, specifically stepwise regression, to predict the amount of yearly GHG emissions (in millions of tones) based on weather related data.

In the regression function while the dependent variable is considered as the amount of yearly GHG emissions (millions of tones), the independent variables are considered to be: annual average weather temperature ($^{\circ}\text{C}$), relative humidity (%), hours of sunshine (h), average sea water temperature ($^{\circ}\text{C}$), and rainfall (mm/m^2). Thus, we also search whether or not there is relationship of GHG emission amount on weather. The study is completed for Turkey case by using the real data from the Turkish Statistical Institute's web site as well as the Turkish state Meteorological Service for the years of 1990 and 2017.

An extensive literature review study is completed on greenhouse gas emission profiles, dynamics, and climate change mitigation efforts across the key climate change players by Zheng et al. (2019). In that literature, some studies focus on the analysis of possible realization of the emission mitigation targets. The largest part of literature focuses on China and its emission reduction targets for 2020 (Koo et al., 2014; Liu et al., 2014; Hao et al., 2015). Yi et al., (2016) combined a scenario-based analysis and a decomposition approach in order to present that carbon emission reduction target can be met by 2020. It is also shown that energy structure, population and economic output are the significant drivers in that change.

For forecasting the GHG emissions, a number of energy consumption-based greenhouse gas emission models have also been developed (Morita et al., 1994). Mikiko et al. (2000) developed an Asian-Pacific Integrated Model (AIM) to predict the GHG emission and evaluate policy measures to reduce it. Two socio-economic scenarios are assumed for the prediction. By the study, it is found that it is possible to mitigate carbon dioxide emissions without scaling back productive activities or standards of living.

Reijnders and Huijbregts, (2008) have studied the calculation of the emission of carbon-based greenhouse gases in kg CO_2 equivalent per kg palm oil, the additional input of energy for producing palm oil based biodiesel and specific fuel consumption for biodiesel.

Kavoosi et al. (2012) studied linear and non-linear equations based on the global oil, natural gas, coal, and primary energy consumption figures to forecast CO_2 emission by using Genetic Algorithm (GA).

Sun nad Liu (2016) presented a least squares support vector machine (LSSVM) to predict different types of CO_2 emissions in China. By the case studies, the proposed model is shown to have higher accuracy in predicting China's CO_2 emissions.

Abdel (2013) proposed an Artificial Neural Network model (ANN) to train and test the yearly CO_2 emission for the years of 1982-2000 and 2003-2010, respectively. They considered four input data including global oil, natural gas, coal, and primary energy consumption for time series forecasting.

Fang et al. (2018) used an improved Gaussian process regression method and particle swarm algorithm to predict carbon emissions.

Different from the existing studies, in this work we study a multiple linear regression for predicting the amount of yearly GHG emissions based on weather related data for Turkey. We utilize stepwise regression methodology also to test whether or not there is relationship of between GHG emission amount on weather related values: annual average weather temperature ($^{\circ}\text{C}$), relative humidity (%), hours of sunshine (h), average sea water temperature ($^{\circ}\text{C}$), and rainfall (mm/m^2).

2. Materials and Methods

Development of a regression function facilitates a better understanding of the true relationship between the input variables and the output. Regression analysis is a statistical tool for the investigation of relationships between the output and input variables. Usually, one seeks to ascertain the causal effect of one variable on another e.g., the efficiency of a system increases by the number of workers, or demand increases by the price, etc. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables on the variable that they influence.

The most suitable model in multiple linear regression analysis is determined by the classical and stepwise methods. To describe the dependent variable changes with the independent variable, also identified as response and explanatory variables, stepwise selection, reverse selection, and forward selection methods are employed in the multiple linear regression models. In some models of multiple linear regression analysis, some of the independent variables may have a limited contribution. To exclude these insignificance variables, in a process called "variable selection," the independent variable must be determined to explain the dependent variable properly. Many studies have been conducted by researchers such as Cox and Snell (1974), Myers (1990) and Thompson (1978) on the regression variable or most suitable model selection methods.

We employ the popular stepwise regression method to evaluate the relationship of input and output variables. *Stepwise regression* is the combination of *forward selection* and *backward elimination* methods as well as step-by-step selection (Montgomery, 1996). The purpose of the stepwise regression method is to find a meaningful subset of independent input variables which predict the dependent variable correctly. At each iteration, the terms that must be included or excluded in the model are reassessed using their partial F statistics. The term excluded in the model with the largest partial F statistic larger than F_{IN} is added to the model. The term included in the model with the smallest partial F statistic, smaller than F_{OUT} is removed from the model. Terms can enter the model and be removed from the model more than once.

2.1 Stepwise regression analysis

For the regression analysis, we obtained the data from the the Turkish Statistical Institute's web site as well as Turkish state Meteorological Service for the years of 1990 and 2017. The regarding data along with their units are summarized in Table 1.

Table 1. Utilized data for the stepwise regression analysis

Year	Dependent Variable (Output)	Independent Variables (Inputs)				
	GHG (millions of tones)	Annual Avg. Weather Temperature ($^{\circ}C$)	Relative Humidity (%)	Hours of Sunshine (h)	Avg. Sea Water Temperature ($^{\circ}C$)	Rainfall (mm/m ²)
1990	219.2	12.9	63.6	7.20	17.63	501.6
1991	226.6	12.7	66.7	6.40	17.60	646.5
1992	232.8	11.4	64.7	6.60	17.13	578.8
1993	240.1	12.3	64.5	6.80	17.08	545.2
1994	234.1	13.7	64.8	7.00	17.68	644.3
1995	247.6	13.1	65.8	7.00	17.20	635.7
1996	267.2	13.3	66.2	6.70	17.20	682.8
1997	278.6	12.5	65.6	6.80	17.10	684.5
1998	280.3	13.8	65.4	6.90	17.68	704.3
1999	277.8	14.1	63.4	7.10	17.95	551.4
2000	298.9	13.1	62.7	7.20	17.98	581.4
2001	280.4	14.2	63.0	7.00	18.18	694.2
2002	286.1	13.2	64.2	7.00	18.18	634.0
2003	305.6	13.2	63.7	6.80	17.48	664.4
2004	315.0	13.2	62.2	6.90	17.75	607.4
2005	337.2	13.3	63.2	6.70	17.80	637.2
2006	358.2	13.3	63.6	6.80	18.23	607.4
2007	391.4	13.8	61.3	6.80	18.23	596.7
2008	387.6	13.6	61.0	6.90	18.10	493.1
2009	395.5	13.7	63.8	6.50	18.13	793.8
2010	398.7	15.1	62.9	6.50	18.20	703.0
2011	427.6	12.8	63.1	6.70	17.98	642.2
2012	446.9	13.8	62.1	6.80	18.15	695.2
2013	439.0	13.8	59.6	6.80	18.28	547.0
2014	458.0	14.5	62.6	6.50	18.58	641.6
2015	472.2	13.8	62.4	6.50	18.65	637.8
2016	498.5	14.0	61.1	6.60	18.45	605.7
2017	526.3	13.7	61.5	6.50	18.53	536.4

First, we check the existence of correlation between yearly GHG emissions amount and each input variable. After observing that each input variable is at least moderately correlated with the GHG emissions, we included each input factor as well as their 2-way, 3-way, 4-way and 5-way interaction terms in the regression analysis. The stepwise regression result is summarized in Figure 1 that is a snapshot from the Minitab 17 stepwise regression result.

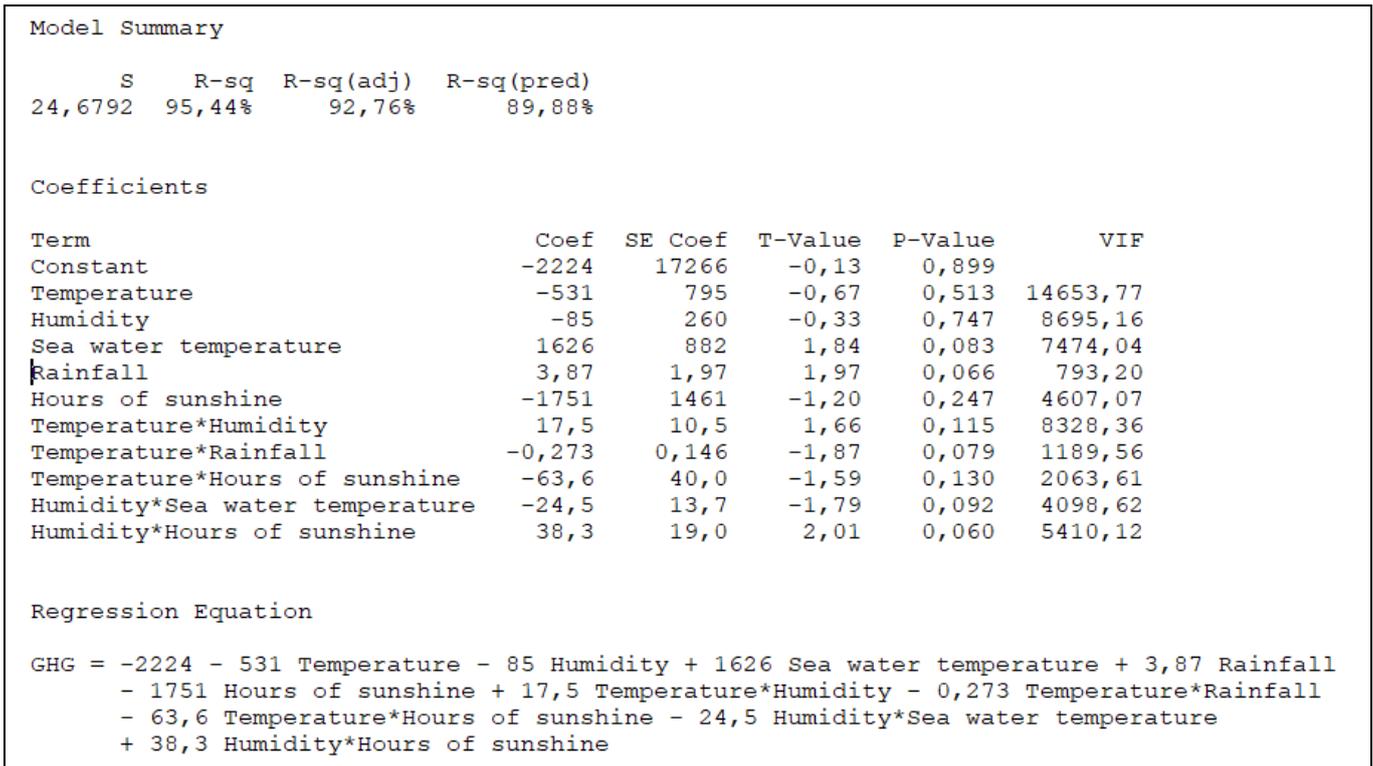


Figure 1. Stepwise regression result produced by Minitab 17

The Figure 1 result shows that a good fit regression is obtained at the R^2 value of 95.44%. In a regression, R^2 value provides a measure of how well outputs are likely to be predicted by the regression model. The bigger the value, the better fit the model. However, since the R^2 value tend to increases by the addition of a new input variable to the function, only considering the R^2 value would not be adequate to evaluate a regression function a good fit one. That’s why, it is also better to evaluate the R^2_{adj} value simultaneously. If it is significantly lower than R^2 , it normally means that one or more explanatory variables are missing. Thus, it is preferred that the R^2_{adj} value is big and close enough to the R^2 . From Figure 1, since these two R^2 and R^2_{adj} values are large and close enough to each other, it can be accepted to be a good fit function. The good function is summarized under the “Regression Equation” subtitle.

The regression model adequacy requires that residuals should be normally distributed, they should have a mean of zero and have a constant variance. Figure 2 shows the residual plots produced from the above regression function. According to that figure, we assume that the regression model adequacy is met.

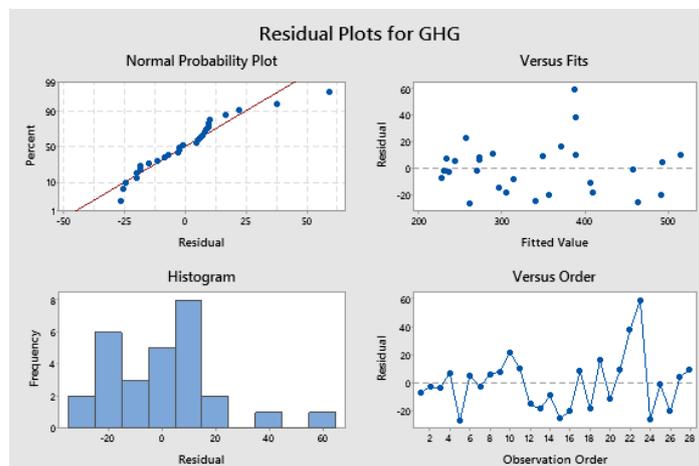


Figure 2. Residual plots for the regression result

3. Conclusion

In this work, we study a multiple linear regression by using stepwise regression method to predict the yearly greenhouse gas (GHG) emission amount based on weather related data in Turkey. In the regression analysis, the output is considered to be the yearly GHG emission amount (millions of tones), the input factors are considered to be: annual average weather temperature ($^{\circ}C$), relative humidity (%), hours of sunshine (h), average sea water temperature ($^{\circ}C$), and rainfall (mm/m^2). One can predict the yearly GHG emission amount by using those weather related input data. For the function fit, the real data which are obtained from the Turkish Statistical Institute’s web site as well as Turkish state Meteorological Service for the years of 1990 and 2017 are utilized. The results show that there is significant effect of GHG emission on climate change (e.g. weather related data) and there exist a good fit regression function that is able to predict the yearly GHG emission amount based on those climate-based input factors. The stepwise regression procedure’s proposed function has the R^2 and R_{adj}^2 values large and close enough to each other (e.g., 95.44% and 92.76%, respectively) as well as its residual plots meet the regression model adequacy. Hence, the function is accepted to be a good fit function that can be used to predict the future GHG emission amount based on the weather related input data.

References

- Baareh, A.K. (2013). “Solving the Carbon Dioxide emission estimation problem: an artificial neural network model”, *Journal of Software Engineering and Applications*, vol. 6, no. 7, pp.338-342.
- Corinne, L.Q., Andrew, R.M., Friedlingstein. P., (2018). “Global carbon Budget”, *Earth System Science Data*, vol. 10, no. 4, pp.2141-2194.
- Cox, D. R., Snell, E.J. (1974). “The choice of variables in observational studies”, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 23, no. 1, pp.51-59.
- Fang, D.B., Zhang, X.L., Yu, Q. (2018). “A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression”, *Journal of Cleaner Production*, vol.173, pp.143–150.
- Gething, B., Puckett, K. (2013). *Design for Climate Change*, RIBA Publishing. 2013.
- Hao, H., Geng, Y., Li, W., Guo, B.H. (2015). “Energy consumption and GHG emissions from China’s freight transport sector: Scenarios through 2050, 2015”, *Energy Policy*, vol. 85, pp.94-101.
- Kavoosi, H., Saidi, M.H., Kavoosi, M., Bohrng, M. (2012). “Forecast global Carbon Dioxide emission by use of genetic algorithm”, *International Journal of Computer Science*, vol. 9, no. 5, pp.418-427.
- Koo, C., Kim, H., Hong, T. (2014). “Framework for the analysis of the low-carbon scenario 2020 to achieve the national carbon emissions reduction target: focused on educational facilities”, *Energy Policy*, vol. 73, pp.356-367.
- Liu, L.W., Zong, H.J., Zhao, E.D., Chen, C.X., Wang J.Z. (2014). “Can China realize its carbon emission reduction goal in 2020: from the prespective of thermal poser development”, *Applied Energy*, vol. 124, pp.199-212.

- Mikiko, K., Yuzuru, M., Tsuneyuki, M. (2000). “The AIM/end-use model and its application to forecast Japanese carbon dioxide emissions”, *European Journal of Operational Research*, vol. 122, pp.416-425.
- Montgomery, D.C. (1996). *Design and Analysis of Experiments*, 4th ed., New York: John Wiley & Sons.
- Morita, T., Matsuoka, Y., Penna, I., Kainuma, M. (1994). “Global Carbon Dioxide Emission Scenarios and Their Basic Assumptions -1994 Survey”, CGER-I011-'94, Center for Global Environmental Research, Tsukuba, 77 pages.
- Myers, R.H. (1990). *Classical and Modern Regression with Applications*, 2nd ed. PWS-Kent Publishers, Boston.
- Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., Zhang, H. (2013). Anthropogenic and Natural Radiative Forcing. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Myhre, G., Shindell, D., Pongratz, J. (2014). Anthropogenic and Natural Radiative Forcing. In: Stocker. Thomas (ed.) : *Climate change 2013 : the physical science basis; Working Group I contribution to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press. pp. 659-740.
- Reijnders, L., Huijbregts, M.A.J. (2008). “Palm oil and the emission of carbon-based greenhouse gases”, *Journal of Cleaner Production*, vol. 16, no. 4, pp. 477-482.
- Sun, W., Liu, M. (2016). “Prediction and analysis of the three major industries and residential consumption CO₂ emissions based on least squares support vector machine in China”, *Journal of Cleaner Production*, vol. 122, pp.144-153.
- Thompson, M.L. (1978a). “Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review*, vol. 46, no. 1, pp. 1-19.
- Thompson, M.L. (1978b). “Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples”, *International Statistical Review*, vol. 46, no. 2, pp. 129-146.
- Yi, B.W., Xu, J.H., Fan, Y. (2016). “Determining factors and diverse scenarios of CO₂ emissions intensity reduction to achieve the 40-45% target by 2020 in China-A historical and prospective analysis for the period 2005-2020”, *Journal of Clean Production*, vol. 122, pp.87-101.
- Zheng, X., Streimikiene, D., Balezentis, T., Mardani, A., Cavallaro, F., Liao, H. (2019). “A review of greenhouse gas emission profiles, dynamics. and climatechange mitigation efforts across the key climate change players”, *Journal of Cleaner Production*, vol. 234, pp. 1113-1133.

O-07 The Impact of Primary Production of Renewable Energy on Labor Force: EU-28 Panel Data Analysis

SELENA KANTARMACI^{1*}, ŞENAY ÜÇDOĞRUK BİRECİKLİ²

¹ *Econometrics, Social Sciences Institute, Dokuz Eylül University, Turkey, selenakantarmaci@gmail.com*

² *Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylül University, Turkey s.ucdogruk@deu.edu.tr*

Abstract – The increase in the use of non-renewable energy sources in the world leads to an increase in prices, an increase in emission problems and depletion of resources. As a result, countries have turned to the use of renewable energy sources. The effect of "fuel of the future" on economic growth or labor force or another development criterion have been investigated by most studies, but primary production has not been considered. This study seeks to examine relationships between primary production of renewable energy and labor force. We employed the Fully Modified Ordinary Least Square (FMOLS) regression model to sample of EU-28 countries for the period 2006-2016 by using panel data analysis. Our main empirical findings reveal that primary production of renewable energy is associated with a positive and statistically significant impact on labor force in EU-28 for the period 2006–2016. These findings are important for the development of energy policies and employment issues. Also reveal contribution of primary production of renewable energy to employment in the future.

Keywords – *Renewable Energy, Primary Production, EU-28, Panel Data Analysis, Labor Force*

1. Introduction

As the world has become increasingly difficult to supply non-renewable energy sources, the emission problem and the increase in prices have led many countries to turn to renewable energy sources. More than one reason, renewable energy sources have become more important for countries, such as the fact that the source is bound to nature and is endless, the contribution to emissions is much less than non-renewable energy resources, the need for imports is not much and development of technologies for using resources. The aim of the study is to examine the contribution of primary production of renewable energy to the labor force by using panel data. The data used in the analysis were obtained from Eurostat and World Bank WDI database.

The word "energy" consists of the ancient Greek term $\epsilon\nu$ = active and $\epsilon\rho\gamma\omega\nu$ = work. Energy in physics is explained as the capacity to do business. Economically, energy includes the production and consumption of all energy resources. The international unit of energy measurement is Joule (Öztürk, 2013). Primary energy production is the production of ready-to-use products from energy sources. It is used when natural resources are used or in biofuel production. Converting energy from one form to another will not be primary energy production. For example, electricity or heat energy obtained in thermal power plants where primary energy sources are burned is not primary production. Renewable energy sources are biomass, wave, tidal, solar, hydraulic, wind, geothermal, natural energy sources which have a more positive impact on human health and environment than depleted energy sources. In a more general definition, renewable energy sources, which are free and natural in nature without the need for processing, are not based on fossils, produce minimum carbon dioxide emissions while supplying

electricity, constantly update themselves under natural conditions and are much healthier for the environment than non-renewable energy sources. According to the energy statistics of EU-28 countries with the data obtained by Eurostat and the calculated percentages, the amount of renewable energy in primary energy production has increased year by year as desired, with two exceptions. The first exception was from 101375.0 Ktoe to 99908.2 Ktoe when moving from 2001 to 2002, and the second exception was from 169160.9 Ktoe to 165795.3 Ktoe when moving from 2010 to 2011. Apart from these years, there is a continuous increase. Renewable energy has a 7% share in primary production with 71802.4 Ktoe in 1990, 10.4% with 98198.2 Ktoe in 2000, 20.1% with 169160.9 Ktoe in 2010 and 27.8% with 210708.0 Ktoe in 2016. Energy is seen as an important factor for the sustainable development of the economy. The need for energy resulting from increasing world population jeopardizes the level of economic welfare that countries want to achieve. Energy sources should be diversified and carbon emissions should be taken into consideration in order to provide the energy required by the society in a low-cost, clean and reliable way. There are multiple energy strategies and energy roadmaps for the purposes that European countries try to achieve according to the target years. Some of these are 2020, 2030 and 2050 energy strategies for the target year. The common goal is to reduce emissions, increase the consumption of renewable energy sources, focus on energy technologies and have a say in the energy market (Eurostat, 2018).

The empirical analysis is mostly based on the set of standard panel cointegration tools, such as Pedroni's cointegration tests and the Fully Modified Ordinary Least Squares (FMOLS) estimator. We follow the standard procedure of time series modelling, which consists of two methodological blocks: unit root testing as a prerequisite of the proper choice of modelling approach (VAR/VECM) and proper model specification and estimation within the selected modelling approach. However, instead of a single country, we use a panel dataset, whose analysis, in the context of time series modelling, differs to a univariate case in terms of unit root tests and estimation methods. The results of this paper suggest the existence of a robust and statistically significant long-run cointegrating relationship between primary production of renewable energy and LF.

2. Materials and Methods

This paper follows a panel of EU-28 for the period 2006-2016, acquired from the World Development Indicators (WDI) of the World Bank and the Eurostat. The multivariate framework includes real gross domestic product (Real GDP) as a proxy for economic growth, gross fixed capital formation (GFCF), total labor force (LF), primary production of renewable energy production (RNW). These variables are exhibited in Table 1 below.

The main aim examined in this paper is to determine whether primary production of renewable energy has an increasing effect on employment in the EU-28. In order to analyse the employment benefits of primary production of renewable energy, we have used real GDP, GFCF and total LF. These variables are relevant in view of the fact that in the traditional production function model. In the panel data analysis, the dimension for the 10-year period between 2006 and 2016 is shown as $T = 11$ and the cross-sectional dimension as $N = 28$ for 28 EU countries. We applied strongly balanced panel dataset. Variables were transformed to logarithmic terms before analysis. Firstly, IPS, Fisher ADF and Fisher PP unit root tests were applied for stationarity. Then, VAR model was established to determine long term effects for Johansen cointegration after determined the appropriate lag criterion between $I(1)$ series. In order to test cointegration relationships, test results that examine all assumptions were observed and appropriate

cointegration assumption was selected based on Akaike and Schwarz criteria. Then the error correction model VECM was established with appropriate lags for short term effects. For the accuracy and reliability of the obtained results, reverse roots, autocorrelation and heteroscedasticity test were performed. As a result of heteroscedasticity problem, Pedroni Cointegration test which resistant to this problem was applied. In order to estimate the long-term relationship coefficients according to these tests, the FMOLS model which is resistant to autocorrelation, endogeneity, heteroscedasticity was used. The determination of the stationarity of variables, co-integration and FMOLS regressions are all carried out using the econometrics software E-views.

Variab le	Period	Source	Description
Real GDP	2006– 2016	World Developm ent Indicators	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2010 U.S. dollars. Dollar figures for GDP are converted from domestic currencies using 2010 official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used. (constant 2010 US\$)
GFCF	2006– 2016	World Developm ent Indicators	Gross fixed capital formation (formerly gross domestic fixed investment) includes land improvements (fences, ditches, drains, and so on); plant, machinery, and equipment purchases; and the construction of roads, railways, and the like, including schools, offices, hospitals, private residential dwellings, and commercial and industrial buildings. According to the 1993 SNA, net acquisitions of valuables are also considered capital formation. Data are in constant 2010 U.S. dollars. (constant 2010 US\$)
Total LF	2006– 2016	World Developm ent Indicators	Labor force comprises people ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers. Not everyone who works is included, however. Unpaid workers, family workers, and students are often omitted, and some countries do not count members of the armed forces. Labor force size tends to vary during the year as seasonal workers enter and leave.
RNW	2006– 2016	Eurostat	Primary production of renewable energy by type. Primary production of biomass, hydropower, geothermal energy, wind and solar energy are included in renewable energies. (Ktoe)

Table 1. Description of the variables

2.1 Panel Unit Root Tests

In order to obtain reliable results in the analysis and to determine the degree of cointegration relationship, unit root test was applied to the series. Spurious regression occurs when working with series which has unit root. Bias in R squared and F statistical results is a negativity seen in models established with non-stationary series. Cross sections used in panel data analysis are generally heterogeneous. Rejecting this heterogeneity leads to inconsistent estimates of parameters. The most valid way of reflecting heterogeneity is to assume that the constant or slope parameters are heterogeneous and select estimation methods in this context (Tatoğlu 2016). Cross-sectional dependence tests are used to determine appropriate unit root tests. However, the cross-sectional dependence is not seen as a problem in micro panels containing several years (Hoeckl, 2007). According to Baltagi, cross-sectional dependence is considered a problem in time-sized macro panels exceeding 20 and 30, whereas no such problem is seen in micro panels containing several years (Baltagi, 1998).

Cross-sectional dependence, heterogeneity were examined in the study and Pesaran's CD test was used since the cross-sectional dimension was larger than the time dimension in accordance with the analysis. In the model, the cross-sectional dependence probability value was 0.9723. Since the probability values are greater than 0.05, H_0 hypothesis under the assumption that there is no horizontal cross-section dependence cannot be rejected. In this case, IPS, Fisher ADF, Fisher PP unit root tests which contains autocorrelation coefficient of each cross section unit in first generation panel root class was used. In Stata, F test was performed with unit-based results by using the "reshape" command. According to the result, the H_0 hypothesis was rejected by calculating value greater than the table value so the estimation methods proposed for heterogeneous panels will be used. As a result of panel unit root tests, all series are stationary in the first difference.

2.2 Johansen Cointegration

The cointegration examines the stability of the level combinations of the series, but also enables the search and interpretation of the long-term equilibrium when a positive result is obtained. If the series is cointegrated, each variable in the panel is not affected by any shock but by the stochastic trend (Tari and Yıldırım, 2009). Johansen Cointegration is based on VAR. The maximum eigenvalue and trace test statistics examine the existence of cointegration relations. The null hypothesis is constructed as the equality of rank to r or less than r . If the statistics are more than the critical value, the null hypothesis is not accepted and the existence of the cointegration relationship is confirmed. Akaike-Schwarz-Hannah Quinn criterion, final prediction error and LR test statistics were used in order to find the appropriate lags in VAR. The fact that it is based on annual data is the reason why the maximum lag length is set to four. The appropriate lag length for the model is chosen 4 based on the Akaike information criterion. After estimating VAR (4) model, Johansen cointegration phase was started. Null hypothesis “ H_0 : There is no cointegration relationship between series.” the alternative hypothesis “ H_1 : Series include cointegration relationship.” In order to select the appropriate assumption according to the Akaike information criterion, the summaries of the five assumptions were estimated by the Eviews 8 program and the assumption which includes intercept, ignored deterministic trend was selected. Result of Johansen Cointegration is exhibited in Table 2 below:

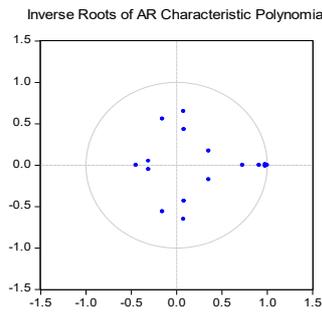
Table 2. Result of Johansen Cointegration

Hypothesized No.		Max-Eigen	0.05 Critical	Prob.
$H_0 : r = 0$	$H_1 : r \geq 1$	54.81509	28.58808	0.0000
$H_0 : r = 1$	$H_1 : r \geq 2$	22.59948	22.29962	0.0454
$H_0 : r = 2$	$H_1 : r \geq 3$	6.792214	15.89210	0.6948
$H_0 : r = 3$	$H_1 : r \geq 4$	0.250645	9.164546	0.9998
Max-eigenvalue test indicates 2 cointegrating eqn(s) at the 0.05 level				
Hypothesized No.		Trace Statistic	0.05 Critical	Prob.
$H_0 : r = 0$	$H_1 : r = 1$	84.45743	54.07904	0.0000
$H_0 : r \leq 1$	$H_1 : r = 2$	29.64234	35.19275	0.1754
$H_0 : r \leq 2$	$H_1 : r = 3$	7.042859	20.26184	0.8945
$H_0 : r \leq 3$	$H_1 : r = 4$	0.250645	9.164546	0.9998
Trace test indicates 1 cointegrating eqn(s) at the 0.05 level				

If the trace test statistic or eigenvalue test statistic is greater than the critical table value at 5% significance level, H_0 hypotheses are rejected. When the results are examined, both the trace and eigenvalue tests reject the hypothesis H_0 , which assumes no cointegration relationship. Maximum Eigenvalue test found two cointegration relationships between series, while trace test found one cointegration relationship. In such cases, it is correct to give importance to trace test statistics since trace test statistics are formed by considering even the smallest eigenvalues (Kasa, 1992; Serletis and King, 1997). Johansen and Juselius argued that trace test should be used when results of these two tests were in any conflict. According to the results of cointegration analysis, cointegration relationship was found between the series. However, although the series have long-term relationships, imbalances can be seen in the short term. The VECM working condition is that the coefficient for the first lag of the error term is between 0 and -1. The statistical significance of this coefficient is necessary to talk about the effect in the short term. As a result of VECM, independent variables have a short-term effect on labor force. 0.15% of the imbalance created by any shock caused by independent variables during the year will be corrected or eliminated until the end of the year.

There are model validation requirements for all these results. These are the characteristic inverse roots of the VAR model, LM autocorrelation test and White heteroscedasticity test. When the characteristic inverse roots of VAR (4) model were examined, no roots were found outside the unit circle. This shows that the test results are stationary and consistent. Another validation requirement is autocorrelation testing. LM test was used for this. The null hypothesis is established that there is no autocorrelation problem for the LM test, and H_0 cannot be rejected if the probability value is greater than critical significance levels. According to the results, since the probability value at the 4th lag is greater than 0.05, it does not include autocorrelation. The final validation requirement is a test of heteroscedasticity. White test was used for this purpose. The null hypothesis was established as homoscedasticity and the alternative hypothesis was as heteroscedasticity. If the probability value is greater than 0.05, H_0 cannot be rejected. According to the results, heteroscedasticity problem was found in VAR model. As a result of heteroscedasticity, VAR models can cause bias problems in standard errors and variances. As a result, Pedroni cointegration which can examine the long-term relationship separately or together for the sections that robust to heteroscedasticity will be used. The model validation conditions are given in Table 3 below:

Table 3. Model Validation Conditions



VAR Residual Serial Correlation		
Lags	LM-Stat	Prob
1	98.22489	0.0000
2	46.05639	0.0001
3	40.63873	0.0006
4	25.84826	0.0562

White Heteroskedasticity Tests			
	Chi-sq	df	Prob.
No Cross Terms	614.4220	320	0.0000
With Cross Terms	1575.380	1190	0.0000

2.3 Pedroni Cointegration

Since the heteroscedasticity problem was found in VAR models, Pedroni cointegration analysis which is resistant to this problem, was used. The null hypothesis in Pedroni cointegration is established as there is no cointegration relationship between the series, and the alternative hypothesis is established that there is a cointegration relationship between the series. Pedroni contains 4 residual-based test statistics within dimension and 3 residual-based test statistics between dimension. In Monte Carlo simulations, Pedroni found the Panel ADF or Group ADF statistics were more consistent than the others for small samples with 20 or fewer periods (Pedroni, 2004). If the probability values are less than the critical significance levels, H_0 hypothesis is rejected. According to Pedroni cointegration test results, four of the seven residual based test statistics reject the H_0 hypothesis at 5% significance level. The other three cannot reject the H_0 hypothesis according to 5% significance level. Group ADF and Panel ADF test statistic results reject the hypothesis H_0 which includes the assumption that there is no cointegration relationship. In this case, a long-term relationship exists between the series forming the model. Result of Pedroni Cointegration is exhibited in Table 4 below:

Table 4. Result of Pedroni Cointegration

	Statistic	Prob.	Weighted	Prob.
within-dimension				
Panel v-Statistic	0.361567	0.3588	0.022070	0.4912
Panel rho-Statistic	2.628763	0.9957	2.491411	0.9936
Panel PP-Statistic	-3.105778*	0.0009	-3.528986*	0.0002
Panel ADF-Statistic	-2.163613*	0.0152	-2.918780*	0.0018
between-dimension				
Group rho-Statistic	4.964967	1.0000		
Group PP-Statistic	-6.454225*	0.0000		
Group ADF-Statistic	-3.966974*	0.0000		

2.4 Fully Modified Ordinary Least Squares (FMOLS)

The next step is to estimate the unbiased and final coefficients of the relationship. The FMOLS recommended by Pedroni will be used. The most important advantage of the method is that it corrects and outputs the bias that may arise from the problems of endogeneity, heteroscedasticity and autocorrelation. As a result of Monte Carlo simulations, Pedroni observed that FMOLS was efficient in small samples. As a result, RNW was found to have a statistically significant and positive effect on labor force, represented by a coefficient of 0.075. GFCF was found to have a statistically significant and positive effect on labor force, represented by a coefficient of 0.136. As a result, GDP was found to have a statistically significant and negative effect on labor force, represented by a coefficient of -0.27. FMOLS output is exhibited in Table 5 below:

Table 5. FMOLS Output

Bartlett Kernel, Bandwith Method Newey-West Automatic Lag Specification FMOLS				
	Coefficient	Std. Error	t-Statistic	Prob.
GDP	-0.276912	0.011743	-23.58036	0.0000
RNW	0.075145	0.001479	50.79413	0.0000
GFCF	0.136680	0.004526	30.19770	0.0000
R-squared	0.999724			
Adjusted R-	0.999690			
Bartlett Kernel, Bandwith Method Newey-West Automatic Lag Specification Grouped FMOLS				
	Coefficient	Std. Error	t-Statistic	Prob.
GDP	0.467466	0.093544	4.997267	0.0000
RNW	0.300389	0.020861	14.39968	0.0000
GFCF	0.317698	0.035665	8.907923	0.0000

3. Conclusion

The aim of the study is to investigate the contribution of primary production of renewable energy to the labor force. Panel data analysis was performed for EU-28 which includes Poland, Portugal, Romania, Slovakia, Slovenia, Greece, Austria, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, Finland, France, Cyprus, Croatia, Netherlands, Ireland, Ireland, Spain, Sweden, Italy, Latvia, Lithuania, Luxembourg, Hungary, Malta. In the model, the 1% increase in the primary production of renewable energy increases the labor force by 0.07%, the 1% increase in the gross fixed capital formation increases the labor force by 0.13%, and the 1% increase in the real gross domestic product decreases the labor force by 0.27% in the long term. The negative effect can be attributed to the income effect and technological unemployment. The factors that determine labor supply in economic theory are the choice between leisure and work. Labor supply is an increasing function of wages, ie labor supply increases with the increase in real wages. But the increase in wages can also increase the choice of leisure time. The increase in wages increases the opportunity cost of leisure time spent. As a result, the individual reduces his demand for leisure time and increases the supply of labor, ie, hours of work, with the substitution effect. Another situation is that the increase in wages increases individual's income level and achieves the target income level. In this case, leisure time is seen as normal goods demanded by the increase in income. Once the individual is relatively enriched, he / she reduces the working hours, ie the supply of labor, with the income effect (Rıfkin, 1996; Ünsal, 2017). Since FMOLS resulted in high R square values, Variance Inflation Factor values were examined and the values were lower than five. This result shows that multicollinearity is not important. In the literature, Singh, Nyuur and Richmond's 2019 study in developed countries showed that 1% increase in real gross domestic product increases labor force by 0.21%, 1% increase in gross fixed capital formation increases labor force by 0.09%, 1% increase in renewable energy production increases labor force by 0.23% and 1% increase in fossil fuels electrical energy consumption increases labor force by 0.02%. In developing countries, 1% increase in gross domestic product increases labor force by 0.14%, 1% increase in gross fixed capital formation decreases labor force by 0.01%, 1% increase in renewable energy production decreases labor force by 0.02%, 1% increase in fossil fuel-induced electrical energy consumption increases labor force by 0.19%. The study can be improved by extending the time dimension, subdividing renewable energy sources into subclasses, comparing country groups, expanding models or comparing models. GMM and Causality Analysis are the methods for planned to improve the study in the future.

The effect of "fuel of the future" on economic growth or labor force or another development criterion have been investigated by most studies, but primary production has not been considered. These findings have contributed to the literature on the inter-linkages between primary production of renewable energy and labor force. In addition, there have been few studies on the impact of renewable energy production on labor force. This research provides new and rich insights into the channels through which primary production of renewable energy influences employment. Over the years, the development of renewable energy technologies, difficulties in obtaining non-renewable resources and price increases will improve the relationship between renewable energy and employment. In this context, countries' capital investments in renewable energy technologies should be increased, the use of fossil fuels should be reduced, and suitable areas for renewable energy resources facilities should be determined. Cooperation

between EU-28 countries about renewable energy, data transfers, support projects and common projects should be intensified.

References

- Baltagi, B. (1998). *Econometrics* . Berlin: Springer-Verlag Berlin Heidelberg
- Eurostat: Energy Statistics (2018) https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Energy_statistics__an_overview#Primary_energy_production
- Hoechle, D. (2007). ‘‘Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence’’, *The Stata Journal*. pp.1-31.
- Kasa, K. (1992). ‘‘Common stochastic trends in international stock markets’’, *Journal of Monetary Economics*. vol.29(1), pp.95-124.
- King, A. S. (1997). ‘‘Common Stochastic Trends and Convergence of European Union Stock Markets’’, *The Manchester School*. vol.65(1), pp.44-57.
- Öztürk, H. (2013). *Yenilenebilir Enerji Kaynakları*. İstanbul: Birsen Yayınevi.
- Pedroni, P. (2004). *Econometric Theory*, 20, 2004, 597–625+ Printed in the United States of America
- Rifkin, J. (1996). *La fin du travail*. Paris: Ed. La Decouverte.
- Singh, N., Nyuur, R. ve Richmond, B. (2019). Renewable Energy Development as a Driver of Economic Growth:Evidence from Multivariate Panel Data Analysis, *Sustainability*. 11(8):1-18.
- Tarı, R., Yıldırım, D. Ç. (2009). ‘‘Döviz Kuru Belirsizliğinin İhracata Etkisi: Türkiye İçin Bir Uygulama’’, *Celal Bayar Üniversitesi Yönetim ve Ekonomi Dergisi*. vol.11(2), pp.95-105.
- Tatoğlu, F. Y. (2016). *Panel Veri Ekonometrisi*. İstanbul: Beta Yayınları.
- The World Bank. (2019). WDI: <https://databank.worldbank.org/source/world-development-indicators#>, (24.04.2019)
- Ünsal, E. M. (2017). *Mikro İktisat* . Ankara: Murat Yayınları.

O-09 Inflation Targeting and Taylor Rule Model for Developed and Developing Countries: A Panel Data Analysis

HANDE ERK*, HAMDİ EMEÇ²

Econometrics, Social Sciences Institute, Dokuz Eylül University, Turkey, handeerk1@gmail.com

Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylül University, Turkey hamdi.emec@deu.edu.tr

Abstract – Since its introduction in 1990, the inflation targeting regime has become a monetary policy strategy adopted by many developed and developing countries. Aiming the price stability and targeting nominal interest rates in this direction, this strategy is expected to have stronger effects on the expectation of monetary policy when a rule based approach is followed. The purpose of this study is to test policy interest rates in the inflation targeting regime, based on the rule based approach with Taylor’s Rule and provide a comparison to Developed and Developing countries. For this purpose, data from 10 countries, 5 of which are developed and 5 of which are developing, have been used for the period 2008Q1 – 2018Q4. Developed countries are modelled with Original Taylor Rule and developing countries are modelled with Expanded Taylor Rule. Panel Data Analysis technique was used to determine the macroeconomic variables that are thought to have an impact on policy interest rates. According to the findings, while real interest rate, inflation rate and output gap variables are effective in determining policy in developed countries, real interest rate, output gap and inflation deficit variables are effective in determining policy in developed countries.

Keywords – *Taylor Rule, Hodrick Prescott Filter, Policy Interest Rate, Panel Data Analysis, Heterogeneity*

1. Introduction

Nowadays, the ultimate goal of many Central Banks is to achieve and maintain price stability. Sustainable price stability is expected to provide both economic stability and social welfare. To this end, central banks have adopted various monetary policy strategies over the years. Inflation targeting strategy, which directly targets inflation itself in order to achieve price stability, was put into practice by the Central Bank of New Zealand in 1990 for the first time and successful results were achieved.

Today, many central banks in developed and developing countries have adopted inflation targeting strategy.

Under the inflation targeting regime, the central bank shares a quantified inflation target with the public. This target can be in the form of a dot or in a band gap. The Central Bank undertakes that inflation will be kept at the target level or within the target range at the end of a certain target period. All monetary policy instruments are used to reach the inflation rate targeted by the Central Bank. Short-term interest rates are used as monetary policy instruments (TCMB, 2006).

There are certain prerequisites for economies to be implemented and successfully maintained by the inflation targeting regime. These prerequisites can be listed as follows. Central banks should have Vehicle

Independence in order for central banks to take necessary measures quickly in case of threats against price stability,

Since the functional implementation of monetary policy and changes in policy interest rate affect the general course of the economy, there should be a developed and stable financial system (Cuaresma ve Gnan, 2008), Financial Discipline should be ensured and the technical infrastructure needs to be improved.

Monetary policy rules based on inflation targets are widely used in the literature. The best known example is the Taylor rule. Taylor (1993) assumes that for the flexible exchange rate regime, each central bank sets its short-term interest rate target according to changes in price level and target output. The simple interest determination model created by Taylor for Fed's use is as follows (Taylor, 1993);

$$i_t = i_t^* + \pi_t + \alpha(\pi_t - \pi_t^*) + \beta(y_t - y_t^*), \alpha > 0 \text{ and } \beta > 0 \quad (1)$$

Where i_t denotes the real interest rate in the period t , i_t^* denotes the real interest rate, π_t and π_t^* respectively represent the actual and targeted inflation rate, y_t and y_t^* represent the actual and potential production level. α is the inflation deficit coefficient and β is the output gap coefficient. $y_t - y_t^*$ is the production deficit in period t and $\pi_t - \pi_t^*$ is the deviation of inflation in the period t from the target value.

Since the level of exchange rates is of particular importance for developing countries due to their commitment to international trade, the Taylor Rule's neglect of exchange rates has led to criticism (Roger ve Restrepo, 2009). Therefore, the Simple Taylor Rule has been enlarged considering the real exchange rate deficit for developing countries (Taylor, 2001):

$$i_t = i_t^* + \pi_t + \alpha(\pi_t - \pi_t^*) + \beta(y_t - y_t^*) + \delta(e_t - e_t^*), \alpha > 0, \beta > 0 \text{ and } \delta > 0 \quad (2)$$

Where δ is the exchange rate response coefficient.

2. Materials and Methods

In this study, for 2008Q1 and 2008Q4 periods, data from 10 countries were used, 5 of which are Developed (New Zealand, Israel, Sweden, England, Finland) and 5 of which are Developing Countries (Turkey, Mexico, Hungary, Brazil, Poland). It is aimed to see the impact of the 2008 economic crisis in determining the sample interval. Within the scope of Taylor Rule, nominal interest rate, real interest rate, realized inflation rate, output gap, deviation of inflation from the targeted value and exchange rate deficit were used. The abbreviations of the variables are presented in table 1.

Table 1. Abbreviations and Variables

Abbreviations	Variables
I_{it}	Nominal Interest Rate
r_{it}	Real Interest Rate
π_{it}	Actual Inflation

$y_{it} - y_{it}^*$	Output Gap
$\pi_{it} - \pi_{it}^*$	Deviation From Inflation Target
$e_{it} - e_{it}^*$	Exchange Rate Gap

Interbank Ratio data are used to represent the dependent variable nominal interest rate. Interbank rate data for Finland fred. com, and for other countries from the OECD database. The real interest rate is obtained by adjusting the nominal interest rate from inflation. The formula given by equation (3) is used in the calculation of real interest rates (Demir, 2019).

$$r = [(1 + I)/(1 + \pi)] - 1 \tag{3}$$

Where r is the real interest rate, I is the nominal interest rate and π is the inflation expectation.

CPI (2015 = 100) series were used to represent Actual inflation. CPI series data are obtained from OECD database. The inflation deficit variable is obtained by subtracting the 12-month inflation expectation from the annual percentage of inflation series. Inflation expectation data were obtained from the central bank database of each country. Monthly expectation series were converted to quarterly data by taking 3-month averages. The annual percentage of inflation series were obtained from the OECD database. Hodrick Prescott Filter (HP) was applied to the industrial production index (2015 = 100) series in order to obtain the output gap and the obtained potential production index series was derived from the Industrial production index series. Industrial production index series data are obtained from OECD database. In order to obtain the exchange rate gap variable, the HP filter is used as in the output gap variable. The HP Filter was applied to the real effective exchange rate based on the CPI and the potential exchange rate series was subtracted from the real effective exchange rate series. Real effective exchange rate series are obtained from Eurostat database.

In the analysis of the study, ‘Panel Data Analysis Technique’ which allows time dimension as well as unit size was used. After the correlation between the units with Breusch - Pagan LM and Pesaran - Ullah - Yamagata LM tests, homogeneity of the parameters was examined by F test. The unit root analysis of all variables to be used in the study was performed with Im, Pesaran – Shin (IPS) and Cross-Sectionally Augmented Im, Pesaran – Shin (CIPS) tests considering heterogeneity and cross-sectional dependence. In the modeling of developed countries sample, the Original Taylor Equation given by equation (1) is taken as reference, while the sample of developing countries is modeled with the Extended Taylor Equation which takes into account the exchange rates given by equation (2). Models for both country groups were estimated with the ‘Extended Average Group Estimator’, which is one of the heterogeneous and inter-unit correlated panel data estimation methods.

2.1 Test of Cross-Sectional Dependence

If the time dimension is greater than the number of observations in panel data models, Breusch and Pagan (1980) and Pesaran, Ulah, Yamagata (2008) tests are recommended. Otherwise, Friedman (1937), Frees (1995) and Pesaran (2004) horizontal section dependency tests may be used (De Hoyos ve Sarafidis, 2006). In the study, since T = 44 and N = 5, Breusch and Pagan (1980) and Pesaran, Ulah, Yamagata (2008) tests were preferred. For each test, H₀ hypothesis is “no cross-section dependency” and H₁ hypothesis “has cross-section dependency”. The results obtained are presented in Table 2.

Table 2. Cross-Sectional Dependence Test Results

Sampling	Test	Statistic	P-Probability
Developed Countries	Breusch - Pagan LM	42.38	0.0000
	Pesaran, Ullah - Yamagata LM	24.37	0.0000
Developing Countries	Breusch - Pagan LM	76.32	0.0000
	Pesaran, Ullah - Yamagata LM	49.3	0.0000

2.2 Homogeneity

According to the test results given in Table 2 above, the p - probability value of each test is less than 0.05 for the samples of developed and developing countries. Therefore the H_0 hypothesis, which states that there is no horizontal cross-section dependency, is rejected.

Yerdelen (2018) stated that while the coefficients are heterogeneous, estimation with the assumption of homogeneity will lead to deviant results. The basic method used to test the homogeneity of coefficients is the standard F Test (Peseran ve Yamagata, 2008). The null hypothesis states that the parameters are homogeneous and therefore the classical model will be valid, whereas the alternative hypothesis states that the parameters are heterogeneous, ie heterogeneous panel data models should be studied (Yerdelen, 2018). In this study, homogeneity of the parameters was tested by F test. F Test’s results are presented in Table 3.

Table 3. Result of F Test

Sampling	F - Statistic	F-Table
Developed Countries	31.42	1.56
Developing Countries	1.70	1.56

According to the F test results given in Table 3, the F statistical value calculated for both samples is greater than the F table value. The hypothesis H_0 , which expresses the validity of the classical model, is rejected.

2.3 Panel Unit Root Tests

In this study, heterogeneity and cross sectional dependence were determined according to the parameters for the samples of developed and developing countries. For this reason, IPS and CIPS were applied from 2nd Generation panel unit root tests which took heterogeneity and cross-sectional dependence into consideration. Table 4 presents the results of the IPS and CIPs panel unit root tests for the sample of developed countries and Table 5 presents the results of the IPS and CIPs panel unit root tests for the sample of developing countries.

According to IPS test results, each variable is stable at the level. According to the CIPS test results, variables other than the real interest rate are stable at the level. Although the real interest rate variable is not stable at the level according to the CIPS test, it is used at the level as it is stable according to the IPS test result.

Table 4. IPS and CIPS Test’s Results – Developed Countries

Im, Pesaran ve Shin (IPS)Test				
Variables	W Statistic - P-Probability			
	Fixed Trendless	P-Probability	Fixed Trend	P-Probability
\hat{i}_{it}	-2.004	0.022	-2.465	0.006
\hat{i}_{it}^*	-7.261	0.000	-6.945	0.000
π_{it}	-3.751	0.000	-2.132	0.016
$y_{it} - y_{it}^*$	-4.557	0.000	-3.093	0.001
$\pi_{it} - \pi_{it}^*$	-5.009	0.000	-4.422	0.000
Cross-Sectionally Augmented Im, Pesaran ve Shin (CIPS)Test				
Variables	CIPS-Statistic	%10	%5	%1
\hat{i}_{it}	-2.825	-2.21	-2.33	-2.55
\hat{i}_{it}^*	-1.654	-2.21	-2.33	-2.55
π_{it}	-2.967	-2.21	-2.33	-2.55
$y_{it} - y_{it}^*$	-4.055	-2.21	-2.33	-2.55
$\pi_{it} - \pi_{it}^*$	-3.516	-2.21	-2.33	-2.55

Table 5. IPS and CIPS Test’s Result – Developing Countries

Im, Pesaran ve Shin (IPS)Test				
Variables	W İstatistik - P Olasılık			
	Fixed Trendless	P-Probability	Fixed Trend	P-Probability
\hat{i}_{it}	-0.501	0.308	-0.633	0.263
$d(\hat{i}_{it})$	-6.449	0.000	-5.903	0.000
\hat{i}_{it}^*	-18.196	0.000	-14.431	0.000
π_{it}	-1.456	0.072	2.719	0.996
$y_{it} - y_{it}^*$	-3.145	0.000	-1.445	0.074
$\pi_{it} - \pi_{it}^*$	-3.079	0.001	-1.950	0.025
$e_{it} - e_{it}^*$	-6.211	0.000	-4.783	0.000
Cross-Sectionally Augmented Im, Pesaran ve Shin (CIPS)Test				
Variables	CIPS-Statistic	%10	%5	%1
\hat{i}_{it}	-0.337	-2.21	-2.33	-2.55
\hat{i}_{it}^*	-3.636	-2.21	-2.33	-2.55
π_{it}	-3.993	-2.21	-2.33	-2.55
$d(\pi_{it})$	-1.946	-2.21	-2.33	-2.55
$y_{it} - y_{it}^*$	-4.356	-2.21	-2.33	-2.55
$\pi_{it} - \pi_{it}^*$	-2.326	-2.21	-2.33	-2.55

$d(\Pi_{it})$: First difference of inflation deviation from the target

According to the results of the IPS test, the variables are generally stable at the level. While the Interbank ratio variable is not stationary in both fixed-trendless and fixed-trend models, the first difference is 1% of significance for both models. The inflation rate variable is not stationary in the fixed-trend model, but is stable at a 10% significance level in the fixed-trendless model and the first difference is stable according to both models.

According to CIPS Test results, Interbank Ratio variable and inflation rate variable are not stable at each level of 10%, 5% and 1% significance level but in the first difference they are stable at all three

levels of significance. Variables other than the interbank rate and inflation rate variables are stable at each of the 10%, 5% and 1% significance levels. Although the interbank rate and inflation rate variables are not stationary at the level according to IPS and CIPS test results, they are used in the first differences in the model since they are stationary in the first differences.

2.4 Extended Average Group Estimator (AMG)

Stability structures, parameter homogeneity and cross-sectional dependence of variables were examined for the sample of 10 countries, five developed and five developing countries. In this context, two different models are estimated in the determination of macroeconomic variables affecting policy interest rates within the framework of Taylor rule. The models were estimated with the Extended Average Group Estimator, which considers heterogeneity and cross-sectional dependence. Table 5 shows the results of the AMG of the developed countries and Table 6 shows the results of the AMG of the developing countries.

Table 5. AMG Results - Developed Countries

i_{it}	Coefficient	St.Dv.	Z - Statistic	P - Probability
i_{it}^*	1.681	0.566	2.97	0.003
π_{it}	0.175	0.077	2.28	0.023
$y_{it}-y_{it}^*$	0.045	0.010	4.32	0.000
$\pi_{it} - \pi_{it}^*$	0.020	0.035	0.58	0.562
Wald χ^2: 40.30		Prob: 0.0000		

According to the results of developed countries AMG estimation given in table 5; coefficients of real interest rate, inflation rate and output gap variables are consistent with the empirical findings of Taylor (1993) approach and the coefficients of real interest rate and output gap variables are 1% significance level and the inflation rate variable coefficient is statistically significant at 5% level. This result means that real interest rate, inflation rate and output gap variables are effective in determining policy rates in developed countries. Policy interest rate; One unit increase in real interest rate will increase by 1,681 units, 1 unit increase in inflation rate will increase by 0.175 units and one unit increase in output gap will increase by 0.045 units. The Wald test, which gives the significance of the model, is obtained as χ^2 value = 40.30 and Prob value = 0.000. Model is significant at 1% significance level.

Table 6. AMG Results – Developing Countries

$d(i_{it})$	Coefficient	St.Dv.	Z - Statistic	P - Probability
i_{it}	0.388	0.218	1.78	0.075
$d(\pi_{it})$	0.097	0.063	1.153	0.127
$y_{it}-y_{it}^*$	0.095	0.019	4.98	0.000
$\pi_{it} - \pi_{it}^*$	0.0539	0.027	1.98	0.048
$e_{it}-e_{it}^*$	-0.064	0.040	-1.58	0.114
Wald χ^2: 46.38		Prob: 0.0000		

According to the results of developing countries AMG estimation given in table 6; The coefficients of real interest rate, output gap and inflation deviation from the target are consistent with the empirical findings of Taylor (2001) approach and real interest rate variable coefficient is 10%, output gap variables coefficient is 1% and inflation deviation from target coefficient variable 5% is statistically significant. This result means that real interest rate, output gap and inflation deviation from the target variables are effective in determining policy rates in developed countries. Policy interest rate; One unit increase in real interest rate will increase by 0.388 units, one unit increase in output gap will increase by 0.095 units and one unit increase in inflation deviation from target will increase by 0.053 units. The Wald test, which gives the significance of the model, is obtained as χ^2 value = 46.38 and Prob value = 0.000. Model is significant at 1% significance level.

3. Conclusion

The aim of this study is to test the macroeconomic factors that are effective in determining the policy interest rates in the countries implementing the Inflation Targeting Regime under the Taylor Rule and to provide a comparison to Developed and Developing Countries. For 2008Q1 and 2008Q4 periods, data from 10 countries were used, 5 of which are Developed (New Zealand, Israel, Sweden, England, Finland) and 5 of which are Developing Countries (Turkey, Mexico, Hungary, Brazil, Poland). While the original Taylor Equation was taken as a reference in the modeling of developed countries, the sample of developing countries was modeled with the expanded Taylor Equation, which also takes into account the exchange rates. Macroeconomic factors which are effective in determining policy interest rates in developed countries were obtained as real interest rate, inflation rate and output gap. Developed countries' policies to increase the output level and keep inflation at the targeted level, especially with the 1929 crisis, coincide with the results obtained. In developing countries, real interest rate, inflation deficit and output gap coefficients were statistically significant and positive. It is possible for developing countries to be sensitive to the output gap due to their economic growth targets and to act responsive to the inflation deficit due to the inflation phenomenon that arises with the cost of production resulting from high exchange rates. Determining the policy interest rate by considering the economic growth in developing countries means keeping the real interest rates low. As the production capacity and income per capita have increased significantly in developed countries compared to developing countries, it is suggested that a policy with low interest rates should be pursued in order to preserve the current economic structure and maintain economic welfare.

References

- Cuaresma, J. C. ve Gnan, E. (2008). Four monetary policy strategies in comparison: how to deal with financial instability?. *Monetary Policy & the Economy*, (3), 65-102.
- De Hoyos, R. E., & Sarafidis, V. (2006). Testing for cross-sectional dependence in panel-data models. *The stata journal*, 6(4), 482-496.
- Demir. (2019). *Nominal ve Reel Faiz*
https://acikders.ankara.edu.tr/pluginfile.php/65963/mod_resource/content/0/1%20Nominal%20ve%20Reel%20Faiz.pdf

Roger S. , Restrepo J. (2009). The Role of the Exchange Rate in Inflation-Targeting Emerging Economies”, IMF Occasional Papers 267.

Tatođlu, F. Y. (2018). *İleri panel veri analizi: Stata uygulamalı*. Beta.

Taylor, J. B. (1993, December). Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy* (Vol. 39, pp. 195-214). North-Holland

Taylor, J. B. (2001). The role of the exchange rate in monetary-policy rules. *American Economic Review*, 91(2), 263-267.

TCMB.(2006). *Enflasyon Hedeflemesi Rejimi*.

[https://www.tcmb.gov.tr/wps/wcm/connect/07d5ced0-3f5c-4fa8-bd23-](https://www.tcmb.gov.tr/wps/wcm/connect/07d5ced0-3f5c-4fa8-bd23-619f6b3c1d6b/EnflasyonHedeflemesiRejimi.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-07d5ced0-3f5c-4fa8-bd23-619f6b3c1d6b-m5lkSAW)

[619f6b3c1d6b/EnflasyonHedeflemesiRejimi.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-07d5ced0-3f5c-4fa8-bd23-619f6b3c1d6b-m5lkSAW](https://www.tcmb.gov.tr/wps/wcm/connect/07d5ced0-3f5c-4fa8-bd23-619f6b3c1d6b/EnflasyonHedeflemesiRejimi.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-07d5ced0-3f5c-4fa8-bd23-619f6b3c1d6b-m5lkSAW), (30.07.2019).

O-12 How is the performance of the McNemar test to determine cut-off comparing to the Youden index and the minP methods for ordinal data?

A simulation study

Pervin Demir^{1*}, Afra Alkan¹ and Selcen Yüksel¹

¹*Biostatistic/ Ankara Yildirim Beyazit University, Ankara, Turkey*

pervin.demr@gmail.com

afra.alkan@gmail.com

selcenpehlivan@gmail.com

Abstract – It is important to determine the optimal cut-off point that differentiates the patients and healthy individuals for diagnostic tests with a continuous or an ordinal response. To get a cut-off point of a diagnostic test, all methodologies to address this issue uses the independent chi-square test results of 2x2 tables. Thus, the dependency of the results obtained from both the gold standard and the diagnostic test is ignored. In this study, we proposed the McNemar test, which considers the association between two dependent categorical variables to estimate the optimal cut-off point for the ordinal response test with five-point results. We evaluated the performance of this test statistic by a simulation design by considering the sample size and the balance of groups as simulation conditions and compared it to the Youden index method and the minimum P-value approach. The sample sizes were set 50, 100 and 200 per group in the balanced design and (50, 100), (50, 150) and (50, 200) for the diseased and non-diseased groups in the unbalanced design. For each scenario, 1000 MCMC repeats were generated. The median bias was 0 for the McNemar test and 1 for the other methods. The proportion of unbiased estimation was 46.0%, 40.0% and 73.9% for the Youden index method, the minimum P-value approach, and the McNemar test, respectively. The proportion of unbiased estimation in the McNemar test was higher in the unbalanced design compared to the balanced design.

Keywords – *ordinal data, optimal cut-off, Youden index, minimum P-value, McNemar test*

1. Introduction

Diagnostic tests, which play an important role in medical care, purpose to provide reliable information about the patient’s condition (Zhou et al., 2011). So, the interpretation of diagnostic tests is an important point to the health care providers. In assessing the performance of the diagnostic tests, it is desirable to know if the test results different for the two health states (the presence and absence of disease). The performance of a diagnostic test is examined by diagnostic accuracy studies.

While evaluating the test performance, it is necessary determining the optimal cut-off point which is used to classify individuals as positive or negative according to the test result. The receiver operating characteristic curve (ROC) is often used to determine the optimal cut-off point. It is a plot of the true-positive fraction (TPF, sensitivity) and the false-positive fraction (FPF, 1-specificity) for all possible cut-point values of the test (Pepe, 2003). The sensitivity and specificity are two basic measures of diagnostic accuracy, that is the probability of a diagnostic test to correctly identify those with the disease and that is the probability of the test to correctly identify those without the disease, respectively.

There are several methods to determine the optimal cut-off point of a diagnostic test (e.g. the Youden index, the point closest-to-(0, 1) corner in the ROC plane, the concordance probability, the minimum P-value approach). The choice of the appropriate methods to estimate the accuracy measures is a function of the type of data (Zhou et al., 2011). The type of the test results can be binary with only two results, ordinal taking on only a few ordered values and continuous taking on an unlimited number of values.

For all possible cut-off points of a continuous or ordinal diagnostic test, the 2x2 table are constructed where the column represents the true disease status and the row represents the dichotomized diagnostic test results. The Youden Index (J) method and the minimum P-value (minP) approach search the optimal cut-off point maximizing their objective functions by using these 2x2 tables. The minP approach is used under the assumption of the absence of association between the resulting dichotomous test and the binary outcome, thus it ignores the dependent structure of the data, whereas the McNemar test is applied to 2x2 contingency tables with matched pairs of subjects to determine whether the row and column marginal frequencies are equal (which is called marginal homogeneity). A non-significant p-value of the McNemar test implies that marginal homogeneity is supported.

In literature, there are some simulation studies that search the performance of these methods for the diagnostic test with scale results (Rota and Antolini, 2014; Hajian-Tilaki, 2017). For the ordinal response data, there is only one simulation study that compares the performance of the aforementioned four methods (Alkan et al, 2019). To our best knowledge, there is no study suggesting the McNemar test as a method to specify the optimal cut-off point and comparing its performance to other most performed methods in the literature. Thus, the aim of this study is to investigate the performance of the McNemar test approach to determine the optimal cut-off point comparing to the Youden index and the minP methods for the ordinal diagnostic test with five-point responses.

2. Materials and Methods

Let X denote an ordinal test with five possible results (e.g. the Breast Imaging Reporting and Database System Score - BIRAD score in which 1: normal, 2: benign, 3: probably benign, 4: suspicious and 5: malignant) which is assumed to be related to the true disease status (binary outcome), where D and \bar{D} present the presence and the absence of the disease, respectively. The true-positive fraction $TPF(c)$ and the false-positive fraction $FPF(c)$ are respectively defined, at any given possible cut-off point c of X , as

$$TPF(c) = P(X > c|D) = S_D(c) \quad (1)$$

$$FPF(c) = P(X > c|\bar{D}) = S_{\bar{D}}(c) \quad (2)$$

Youden Index method (J):

The Youden index (Youden, 1950) is the maximum achievable value of the Youden function $J(c)$, defined as the difference between the population quantities $TPF(c)$ and $FPF(c)$ which are given in Equation 1 and 2, respectively.

$$J(c) = S_D(c) - S_{\bar{D}}(c) \quad (3)$$

The optimal cut-off point \hat{c}_j is the c that achieves the maximum of the Youden Function $\hat{J}(c)$ over all possible cut-off values of X .

The minimum P-value approach (minP):

The minimum P-value approach (Miller and Siegmund, 1982) is based on a systematic search of the optimal cut-off point (\hat{c}_{minP}) that achieves the minimum of the P-value of the Chi-square test statistic on the absence of association between the dichotomized biomarker and the binary true status, or, in other words, the maximum of the associated Chi-square statistic over all possible cut-off point values c of X . The Chi-square objective function is given in Equation 4.

$$CHI_1^2(c) = \frac{(S_D(c) - S_{\bar{D}}(c))^2}{\left(\frac{n_D S_D(c) + n_{\bar{D}} S_{\bar{D}}(c)}{n_D + n_{\bar{D}}}\right) \left(1 - \frac{n_D S_D(c) + n_{\bar{D}} S_{\bar{D}}(c)}{n_D + n_{\bar{D}}}\right) \left(\frac{1}{n_D} + \frac{1}{n_{\bar{D}}}\right)} \quad (4)$$

where n_D is the number of diseased subjects and $n_{\bar{D}}$ is the number of non-diseased subjects (Rota and Antolini, 2014).

McNemar test approach (McN): a new criterion for cut-off point selection

The McNemar’s test is applied to 2x2 contingency tables with paired dichotomous data (McNemar, 1947). It considers the association between two dependent categorical variables. Thus, we propose using this method to estimate the optimal cut-off point for the ordinal response test considering each possible cut-off point. The point c of X minimizing the 1 degree of freedom McNemar Chi-square statistic computed from the classification table below can be considered as the optimal cut-off point. In other words, the point where the McNemar’s Chi-square test value is close to zero will be the cut-off point (\hat{c}_{McN}).

Test result	Disease status		Total
	Absent (\bar{D})	Present (D)	
Positive ($X > c$)	n_{11}	n_{12}	$n_{1.}$
Negative ($X \leq c$)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1} = n_{\bar{D}}$	$n_{.2} = n_D$	n

$$McCHI_1^2(c) = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \quad (5)$$

The $McCHI_1^2(c)$ can be expressed as a population quantity by writing the observed classification table in terms of classification probabilities (1) and (2) in the Appendix. The resulting population McNemar’s Chi-square objective function is

$$McCHI_1^2(c) = \frac{([1 - S_D(c)] n_D - [S_{\bar{D}}(c)] n_{\bar{D}})^2}{([1 - S_D(c)] n_D) + ([S_{\bar{D}}(c)] n_{\bar{D}})} \quad (6)$$

We considered that X has five possible ordinal outcomes as 0: definitely negative, 1: probably negative, 2: suspicious, 3: probably positive and 4: definitely positive for the presence of the disease. Item response

theory was used to generate data to ensure that higher responders were more likely to be diseased. Because responses were given on a 0-4-point rating scale, the Rating Scale Model (RSM) was applied in data generation.

Assume that θ is the underlying latent trait related to what test measures and β is the item difficulty. The RSM is

$$\pi_{nc} = \frac{\exp \sum_{j=0}^c (\theta_n - (\beta + \tau_j))}{\sum_{k=0}^4 \exp \sum_{j=0}^k (\theta_n - (\beta + \tau_j))} \quad j = 0,1,2,3 \quad (7)$$

$$\tau_j \equiv 0, \exp \sum_{j=0}^0 (\theta_n - (\beta + \tau_j)) = 1$$

where π_{nc} is the probability of resulting in a c score for individual n, β is the item difficulty and τ_j is the threshold of the j^{th} category (Andrich, 1978).

Let us assume that the latent trait is normally distributed for non-diseased and diseased populations, respectively as $\theta_{\bar{D}} \sim N(\mu_{\bar{D}} = 0, \sigma_{\bar{D}}^2 = 1)$ and $\theta_D \sim N(\mu_D, \sigma_D^2 = 1)$ (Figure 1). These two distributions intersect at $\mu_D/2$ resulting in the optimal cut-off point for the latent trait to discriminate the diseased subjects from those without the disease (Rota and Antolini, 2014).

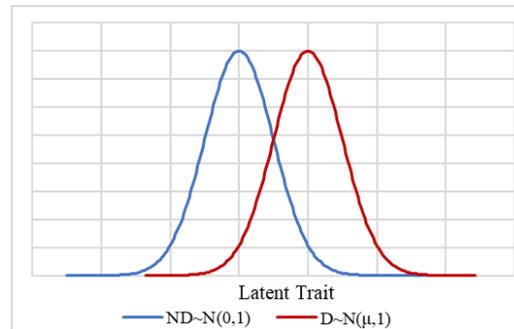


Figure 1. The distribution of the latent trait in disease and non-diseases population (ND: Non-diseased population, D: Diseased population)

The test result of a person with the latent trait of $\mu_D/2$ is estimated by (7) with specified item parameters, which gives the optimal cut-off point for the ordinal response test.

Simulation design

The simulation was performed via R language (ver. 3.4.4) and RStudio (Version 1.1.463 – © 2009-2018 RStudio, Inc) (R Core Team, 2018). We considered balanced and unbalanced designs. The sample sizes of the diseased and non-diseased samples were set 50, 100 and 200 in the balanced design, and (50, 100), (50, 150) and (50, 200) in the unbalanced design. Each scenario was performed with 1000 MCMC repeats.

First, the latent trait of non-diseased samples was randomly generated from $N(0,1)$. μ_D is set to equal $\{0.51, 1.05, 1.68, 2.56\}$ for the diseased samples, resulting the optimum cut-off points of latent trait as $\{0.255, 0.575, 0.84, 1.28\}$. Then, the item difficulty of the ordinal test was set to 0.25 and the category

thresholds as $\{-2.25, -0.75, 0.75, 2.25\}$. The response probabilities of a person with the latent trait of $\mu_D/2$ were estimated by substituting the difficulty and category thresholds in (7). The test score with the highest response probability was considered as the true cut-off points of the test, which were $\{2, 2, 2, 3\}$.

The test scores of the samples from RSM were obtained via the *genPattern()* function of the *catR* package (Magis and Raiche, 2012) by using the latent trait and the item parameters.

In the next step, the \hat{c}_j and \hat{c}_{minP} were estimated via *optimal.cutpoints()* function of *OptimalCutpoints* package (Lopez-Raton et al., 2014). The \hat{c}_{McN} was obtained by using a new function that was constructed from the *mcnemar.test ()* function of *stats* package (R Core Team, 2018). In the case of multiple values satisfying the corresponding conditions, the minimum of these values was selected as the estimated cut-off point. The difference between the estimated cut-off points and the corresponding true cut-off point was defined as bias. Each step was performed in the same manner for the unbalanced design. The design of the simulation is summarized in the following flow chart (Figure 2).

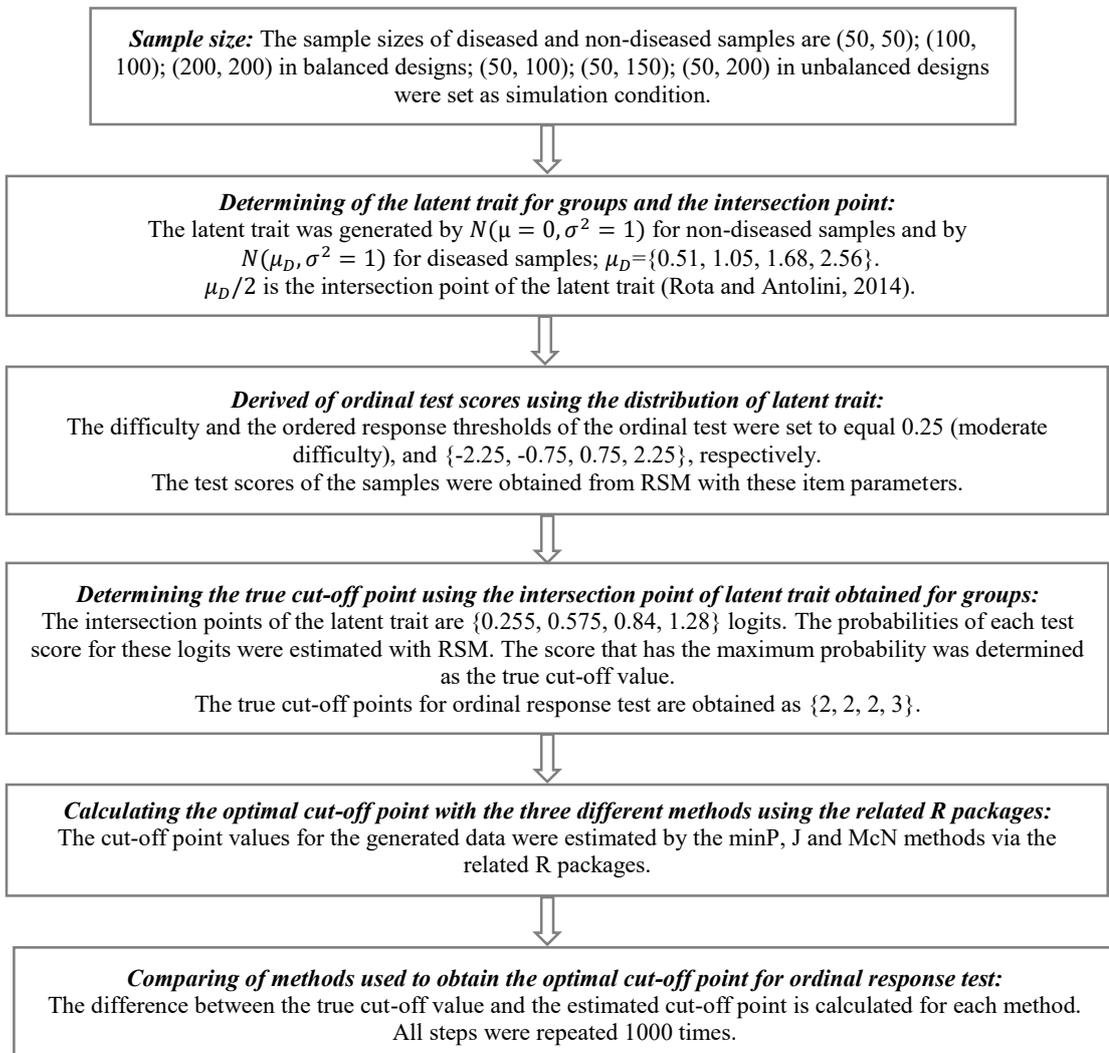


Figure 2. The steps of the simulation study

3. Results

The median bias was 0 for the McN and 1 for the other methods (Figure 3). The ranges of bias in the J and minP methods were wider than the range of bias in the McN in both unbalanced and balanced design (Figure 4). The range of bias in the McN became narrower and the median bias of the minP method decreased to 0 in the balanced design.

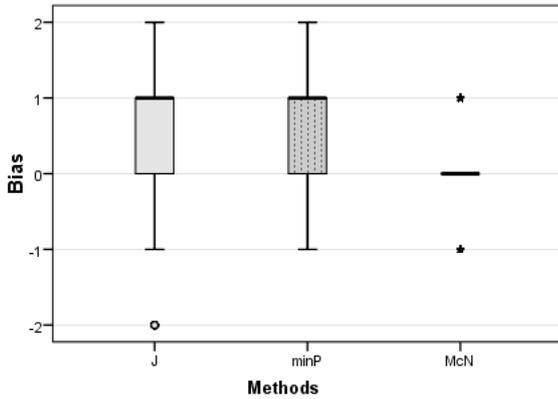


Figure 3. The bias distribution of each method

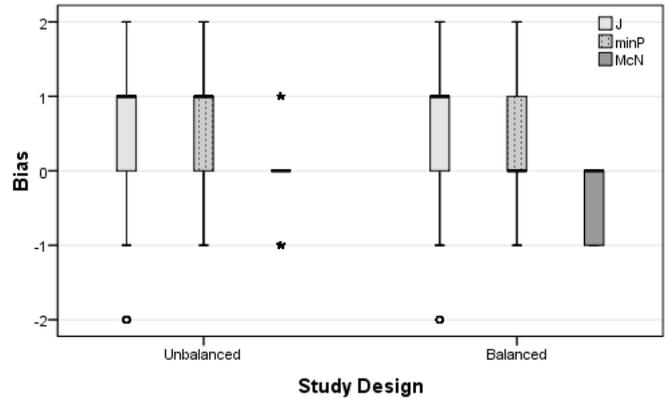


Figure 4. The bias distribution of each method based on the study design

In the unbalanced design, the bias distribution in all three methods did not change when the sample size of the non-diseased group was increased (Figure 5). The ranges of bias in the McN and minP became narrower than the J for all sample sizes in unbalanced design and for $n=50$ in the balanced design. The sample size didn't change the bias distribution and the median of bias for the minP and McN in the balanced design; however, the range of bias in J method became narrower. The median bias of the minP and McN in all sample sizes was 0 but the median bias of the J increased to 1 from 0 when the sample size was 100 and 200.

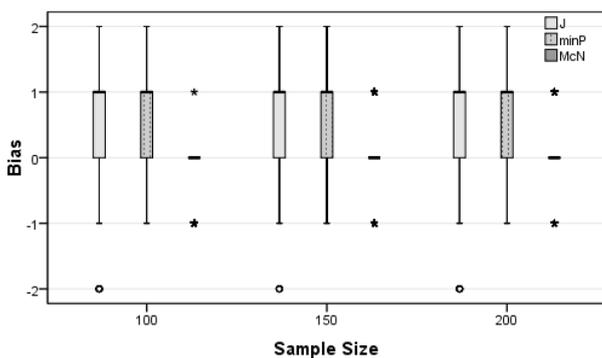


Figure 5. The bias distribution of each method based on the sample size in the **unbalanced** design

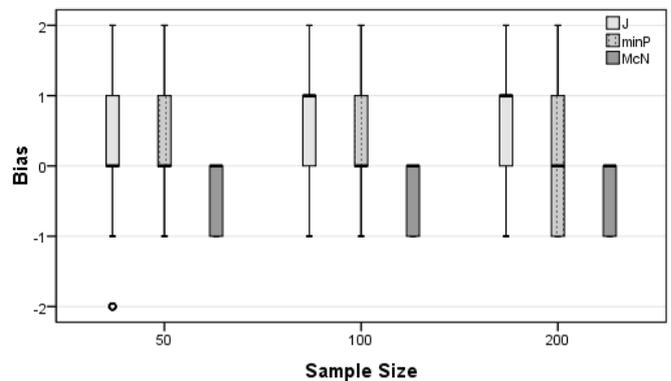


Figure 6. The bias distribution of each method based on the sample size in the **balanced** design

The proportion of underestimation (UE), unbiased (UB) and overestimation (OE) in each method is given Table 1. The proportion of UB estimation was 46.0%, 40.0% and 73.9% for the J, the minP and the McN, respectively. The proportion of UB estimation in the McN was higher in the unbalanced design compared

to the balanced design in general and considering all sample sizes. The UE proportion of McN was higher than the other two methods in balanced design; on the other hand, the OE was not dictated for McN in the balanced design.

Table 1. The proportion (%) of underestimation, unbiased and overestimation in each method

	Underestimation			Unbiased			Overestimation		
	J	minP	McN	J	minP	McN	J	minP	McN
All	1.3	8.4	23.0	46.0	40.0	73.9	52.7	51.6	3.1
Unbalanced	1.5	3.2	8.2	45.9	38.4	85.5	52.6	58.4	6.3
(50, 100)	1.7	4.0	19.5	45.9	39.3	80.4	52.4	56.7	0.1
(50, 150)	1.4	3.0	4.8	46.2	37.6	92.9	52.4	59.4	2.3
(50, 200)	1.3	2.4	0.5	45.7	38.4	83.1	53.0	59.2	16.4
Balanced	1.2	13.6	37.8	46.0	41.5	62.2	52.8	44.9	0.0
(50, 50)	2.6	5.3	39.0	48.1	46.3	61.0	49.3	48.4	0.0
(100, 100)	0.9	8.4	37.6	46.1	42.6	62.4	53.0	49.0	0.0
(200, 200)	0.2	27.2	36.8	43.8	35.8	63.2	56.0	37.0	0.0

4. Discussion

In this study, we compared the performance of the McNemar test (McN) to the Youden Index (J) and the minP methods to determine the optimal cut-off point for the ordinal diagnostic test with five-point responses. In each scenario of the simulation study except the balanced design with 50 samples, the median bias of the Youden Index method was 1. The range of bias was narrower in the balanced design with 100 and 200 samples for the Youden Index method compared to the other conditions. The proportion of underestimation was lower in the Youden Index method than those of the other methods, but the proportion of overestimation was higher in the Youden Index method than the proportion of overestimation obtained for the McNemar approach. The proportion of unbiased was lower than 50% for the Youden Index.

The simulation conditions didn't affect the range of bias in the minP approach, while the median bias in the balanced design decreased to zero. The proportion of underestimation was lower and the proportion of the overestimation was higher than the McNemar. The proportion of unbiased was lower than 50% and the other methods. The median bias of the McNemar approach was lower than the median bias for the Youden Index and the minP approach in unbalanced design. The range of bias in the McNemar was narrower than the range of bias in the other two methods in all conditions. The proportion of underestimation was higher and the overestimation was lower than others. The proportion of unbiased was higher than 50% and the proportion obtained from the other two methods. The simulation design for this study was conducted under similar conditions to Rota and Antolini (2014). According to their study results, they did not recommend the use of the minP approach for cut point finding for a continuous diagnostic test. In our study, the proportion of unbiased for the minP approach was lower than the other two methods examined in the study.

Since the studies examining the performance of these three methods for ordinal responses are lacking in the literature, especially there are no studies using the McNemar test. So, we can only compare the results with our previous simulation study which investigates the performance of the optimal cut-off methods,

which are generally used for diagnostic tests with a continuous response, for tests with an ordinal response (Alkan et al., 2019). When the range of bias in both balanced and unbalanced scenarios are examined to evaluate the performance of four methods, the range of bias for the J and minP approach was wider than finding for the other two methods (the point closest to (0,1) corner in the ROC plane and concordance probability methods). When the median values are examined, the minP value method is better than others in the balanced scenario in their study. Similarly, the median bias for the minP approach in this study is lower than the median bias of the J method in the balanced design.

5. Conclusion

The McNemar has less bias according to the other two methods. In the unbalanced design, the performance of McNemar is better than the Youden Index and minP methods. The median bias is similar in the McNemar and minP approach in balanced design with each sample size, while the range of bias in the McNemar was narrower than the range of bias in the minP approach. The proportion of unbiased estimation in the McNemar was higher in the unbalanced design compared to the balanced design. The proportion of unbiased estimation in McNemar was higher than those of both the Youden Index and minP approach. As a result, we have shown that the McNemar approach performs well and outperforms both the Youden Index and the minP approach.

In literature, the McNemar test application in obtaining the optimal cut-off for a diagnostic test with ordinal response is lacking. Thus, the performance of the McNemar approach to determine the optimal cut-off point for an ordinal response test can be examined by comparing with the other methods used determining the cut-off point under different simulation conditions (e.g. different response categories).

Acknowledgment

The authors declared no potential conflicts of interest with respect to the authorship and publication of this article. The authors received no financial support for the research or publication of this article. This work accepted to be supported by TUBITAK's program to support participation in domestic scientific activities (2224-B).

References

- Alkan, A., Yüksel, S., Demir, P. (2019). "Which of the favorite optimal cut-off determination methods is preferable for the ordinal response data? A simulation study", Proceedings of the 13th International Conference on Applied Statistics, Bucharest, Romania. pp.6-12.
- Andrich, D. (1978). "A rating formulation for ordered response categories.", *Psychometrika*, vol. 43, no.4, pp. 561-573.
- Habibzadeh, F., Habibzadeh, P., Yadollahie, M. (2016). "On determining the most appropriate test cut-off value: the case of tests with continuous results.", *Biochemia Medica*, vol. 26, no.3, pp.297–307, doi: 10.11613/BM.2016.034.

Hajian-Tilaki, K. (2017). “The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation.”, *Statistical Methods in Medical Research*, vol. 27, no. 8, pp. 2374-2383, doi: 10.1177/0962280216680383.

Liu, X. (2012). “Classification accuracy and cut point selection.”, *Statist. Med.*, vol.31, pp.2676–2686, doi: 10.1002/sim.4509.

Lopez-Raton, M., Rodriguez-Alvarez, M.X, Cadarso-Suarez, C., Gude-Sampedro, F. (2014). “OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests.” *Journal of Statistical Software* vol:61, no:8, pp.1-36.

Magis, D., Raiche, G. (2012). “Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR.” *Journal of Statistical Software*, vol.48, no.8, pp.1-31. doi:10.18637/jss.v048.i08.

McNemar, Q. (1947). “Note on the sampling error of the difference between correlated proportions or percentages.” *Psychometrika*, Vol.12, no. 2, pp. 153-157.

Miller, R., Siegmund, D. (1982). “Maximally selected Chi square statistics.” *Biometrics*, vol. 38, no.4, pp.1011–1016, doi: 10.2307/2529881.

Pepe, MS. (2003). “The Statistical Evaluation of Medical Tests for Classification and Prediction.” Oxford University Press: New York.

Perkins, N.J., Schisterman, E.F. (2006). “The inconsistency of optimal cutpoints obtained using two criteria based on the receiver operating characteristic curve.” *American Journal of Epidemiology*, vol. 163, no. 7, pp.670–675.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Rota, M., Antolini, L. (2014). “Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers.”, *Computational Statistics and Data Analysis*, vol.69, pp.1–14 ,doi: 10.1016/j.csda.2013.07.015.

Youden, WJ. (1950). “Index for rating diagnostic tests.” *Cancer*, vol.3, no. 1, pp.32–35.

Zhou, X.H., Obuchowski, N.A., McClish, D.K. (2011). “Statistical Methods in Diagnostic Medicine.” Wiley: New York.

Appendix: Theoretical expression of the McNemar test approach

Let X denote an ordinal test with five possible results which is assumed to be related to the true disease status (binary outcome), where D and \bar{D} present the presence and the absence of the disease, respectively.

In diagnostic accuracy studies, the diagnostic test result obtained from all possible cut-off point (c) and the disease status can be represented by a 2x2 table as shown below. Each cell indicates the number of observed outcomes in each category.

Test result	Disease status		Total
	Absent (\bar{D})	Present (D)	
Positive ($X > c$)	n_{11}	n_{12}	$n_{1.}$
Negative ($X \leq c$)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1} = n_{\bar{D}}$	$n_{.2} = n_D$	n

McNemar test is used to test the marginal homogeneity for a correlated binary outcome. The 1 degree of freedom McNemar Chi-square statistic obtained from this classification table is,

$$McCHI_1^2(c) = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$$

The sensitivity (Sen), the specificity (Sp), the false-negative fraction (FNF) and the false-positive fraction (FPF) which are the statistical measures of the performance of a classification test are respectively defined, at any given possible cut-off c of X, as

$$Sen(c) = \frac{n_{22}}{n_D} = S_D(c) \quad , \quad Sp(c) = \frac{n_{11}}{n_{\bar{D}}} = 1 - S_{\bar{D}}(c)$$

$$FNF(c) = \frac{n_{12}}{n_D} = 1 - S_D(c) \quad , \quad FPF(c) = \frac{n_{21}}{n_{\bar{D}}} = S_{\bar{D}}(c)$$

The $McCHI_1^2(c)$ can be expressed as a population quantity by writing the observed classification table in terms of this classification probabilities. Thus, the McNemar's Chi-square objective function in population was obtained as

$$McCHI_1^2(c) = \frac{([1 - S_D(c)] n_D - [S_{\bar{D}}(c)] n_{\bar{D}})^2}{([1 - S_D(c)] n_D + [S_{\bar{D}}(c)] n_{\bar{D}})} .$$

O-13 Evaluation of Social Progress Performance of European Union Countries and Turkey by Data Envelopment Analysis

¹Esra Betül KINACI*, ²Hasan BAL and ³Ihsan ALP

¹Statistics, Gazi University, Turkey, esrakinaci@gazi.edu.tr

²Statistics, Gazi University, Turkey, hasanbal@gazi.edu.tr

³Statistics, Gazi University, Turkey, ihsanalp@gazi.edu.tr

Abstract – States and societies always tend to increase the level of development of their country. In general, the growth in the economies of the countries constitutes the perception that there is an increase in the level of development of the society. The economy have an absolute impact on the level of development of society. However, it would not be right to make such a judgment considering only the economy so many other factors have to be considered. These factors include human development, the ability to meet the basic human needs of citizens, and the opportunities that enable citizens to improve and maintain their quality of life by creating space for themselves. One of the studies to examine this complex structure is the Social Progress Index (SPI). This index is created taking into account measurable factors in many areas, including housing, nutrition, rights and education. Thus the index assess social progress performance levels of countries. The aim of this study is to evaluate Performance of Social Development the European Union countries and Turkey in a different method, taking into account variables used in SPI. Data envelopment analysis (DEA), which is a linear programming based method, was used in the study. The results obtained from the analysis were compared with SPI and the similarities were evaluated with statistical methods.

Keywords – Data Envelopment Analysis, Social Progress Index, Efficiency

1. Introduction

Economic, political and social development of societies has led to the concept of social development. In the simplest terms, social development means providing an environment in which the people who make up the society will live well and be happy. In daily life, it is not possible to differentiate between economic and social development levels of countries. The increase in income is an important factor for development but it is insufficient to measure the level of social development alone. Because social development refers to the quality of life of people, the opportunities they have access to, as well as their rights, individual freedoms and opportunities for self-development. Amartya Sen, a Nobel Prize-winning economist, argued in his book *Development as Freedom* that the concept of development is not only about economic indicators but also related to the increase in the opportunities of countries such as education and health and the adoption of democratic institutions. Under the leadership of economists who are aware of this deficiency, the United Nations Human Development Index (HDI) has been developed. This index is composed of variables such as literacy rate, infant mortality rate, average life expectancy adult education period, schooling rate and per capita income. HDI is calculated for 189 countries and examines the basic dimensions of human development, including health education and income. HDI is frequently used in academic studies and international research, but it has emerged over time in new indices that examine social development from different perspectives. The Social Progress Index (SPI) is one of these alternatives. Unlike the others, the Social Index includes factors such as political freedoms, security and transparent governance. While calculating SPI, the average of 3 sub-indices, Basic Human

Needs, Fundamentals of Welfare and Opportunities are taken. Basic human needs measure the fulfilment of issues such as nutrition, shelter, access to clean water, and security needed for individuals to survive. Basics of Welfare measures basic education, access to information and communication channels, and environmental sustainability. Opportunities sub-index assesses individual and social rights, anti-discrimination and access to higher education

DEA can be defined as a linear programming-based method used to evaluate the relative effectiveness of decision points, responsible for producing outputs or outputs using inputs with different measurement units. Decision-making units subject to analysis should be of similar type and have similar functions for the same objective. With this analysis, a Decision Making Unit (DMU) can be compared with others. As with the least squares method, the method focuses on effective boundaries rather than central tendencies and forms a discrete plane covering effective observations. Therefore, DEA evaluates each observation separately and compares the current observation to only the nearest efficiency units.

This study aims to measure the social development levels of the countries subject to the analysis with DEA using SPI factors. DEA calculates the relative effectiveness of DMUs and provides information to the user about the improvements necessary for inefficiency units to be efficiency. It also allows to rank DMU's efficiency scores. In this way, it helps the countries in the rankings to produce policies in order to improve their social development levels

2. Materials and Methods

Data used in the application taken from 2018 Social Progress Imperative (<https://www.socialprogress.org>). Basic Human Needs, Foundation of Wellbeing and Opportunity are the three main objectives of the SPI. These three objectives are divided into 10 sub-categories covering high priority social policies such as nutrition and basic human needs, housing, water and sanitation, personal safety and access to education. 33 indicators, which are calculated from country-level data and fall under these 10 subject categories, form the basis of the subject categories. Each main title is divided into sub-groups. The main objectives, categories and indicators are given in Table 1.

The data used in the calculation of SPI scores in the SPI studies conducted each year change in each calculation year. The data in Table 1 is for 2018. The SPI scores obtained from the data in Table 1 were compared with the efficiency scores using DEA

In SPI, some variables were obtained by subjective methods. For this reason, these variables are not included in the application. 33 indicators belonging to 10 categories given in Table 1 are taken as output. The reason for using these indicators as output variables is that each country has obtained the values related to these indicators with certain inputs. Since these input values are not known, the model's input is a virtual input and is set to 1 for all DMUs. With this assumption, the output-oriented CCR model is used. The reason for this is to indicate how much the outputs should be increased while the inputs are fixed. DEA was performed for each category. The averages of the categories were calculated and the scores of the main topics were formed. Finally, the social development scores of the countries were calculated by taking the average on the basis of basic headings. The model used for Basic Human Needs is presented in Table 2. The other is modelled similarly.

When the defined model is applied to all decision points, efficiency scores are obtained for each decision point. In the classical DEA method, there is no possibility of ranking among the efficiency DMUs. However, in the literature referred to as super-efficiency method Andersen and Petersen (1993) developed by the method of effective decision-making units can be listed in itself. Therefore, the output-oriented CCR model was accompanied by a super-efficiency model.

Table 1. Headings for SPI Calculation Category and Indicators

Social Progress Index		
1.Basic Human Needs	2FOUNDATIONS OF WELLBEING	3.OPPORTUNITY
1.1.Nutrition and Basic Medical Care	2.1.ACCESS TO BASIC KNOWLEDGE	3.1.PERSONAL RIGHTS
-Undernourishment (% of pop.)	-Adult literacy rate (% of pop. aged 15+)	-Political rights (0=no rights; 40=full rights)
-Maternal mortality rate (deaths/100,000 live births)	-Primary school enrolment (% of children)	-Freedom of expression (0=no freedom; 1=full freedom)
-Child mortality rate (deaths/1,000 live births)	-Secondary school enrolment (% of children)	-Freedom of religion (0=no freedom; 4=full freedom)
-Child stunting (% of children)	-Gender parity in secondary enrolment (girls/boys)	-Access to justice (0=non-existent; 1=observed)
-Deaths from infectious diseases (deaths/100,000)	-Access to quality education (0=low; 4=high)	-Property rights for women (0=no right; 5=full rights)
1.2.Water and Sanitation	2.2.HEALTH AND WELLNESS	3.2.PERSONAL FREEDOM AND CHOICE
-Access to at least basic drinking water (% of pop.)	-Life expectancy at 60 (years)	-Vulnerable employment (% of employees)
-Access to piped water (% of pop.)	-Premature deaths from non-communicable diseases (deaths/100,000)	-Early marriage (% of women)
-Access to at least basic sanitation facilities (% of pop.)	-Access to essential health services (0=none; 100=full coverage)	-Satisfied demand for contraception (% of women)
-Rural open defecation (% of pop.)	-Access to quality healthcare (0=unequal; 4=equal)	-Corruption (0=high; 100=low)
1.3.Shelter	2.3.Environmental Quality	3.3.Inclusiveness
-Access to electricity (% of pop.)	-Outdoor air pollution attributable deaths (deaths/100,000)	-Acceptance of gays and lesbians (0=low; 100=high)
-Quality of electricity supply (1=low; 7=high)	-Wastewater treatment (% of wastewater)	-Discrimination and violence against minorities (0=low; 10=high)
-Household air pollution attributable deaths (deaths/100,000)	-Greenhouse gas emissions (CO2 equivalents per GDP)	-Equality of political power by gender (0=unequal power; 4=equal power)
1.4.Personal Safety	-Biome protection	-Equality of political power by socioeconomic position (0=unequal power; 4=equal power)
-Homicide rate (deaths/100,000)		-Equality of political power by social group (0=unequal power; 4=equal power)
-Perceived criminality (1=low; 5=high)		3.4.ACCESS TO ADVANCED EDUCATION
-Political killings and torture (0=low freedom; 1=high freedom)		-Years of tertiary schooling
Traffic deaths (deaths/100,000)		-Women's average years in school
		-Globally ranked universities (points)
		-Percent of tertiary students enrolled in globally ranked universities

Table 2. DEA model for Basic Human Needs

Categories	Model	Variables
Nutrition and Basic Medical Care	Output	Undernourishment
		Maternal mortality rate
		Child mortality rate
		Child stunting
		Deaths from infectious diseases
	Input	Taken to 1 for all Decision Making Units
Water and Sanitation	Output	Access to at least basic drinking water
		Access to piped water
		Access to at least basic sanitation facilities
		Rural open defecation
	Input	Taken to 1 for all Decision Making Units
	Shelter	Output
Household air pollution attributable deaths		
Input		Taken to 1 for all Decision Making Units
Personal Safety	Output	Homicide rate
		Traffic deaths
	Input	Taken to 1 for all Decision Making Units

In the study, since there is no control over the inputs and maximizing the output variables is important for the effectiveness of DMU's ,the output-oriented CCR model is used. The model proposed by Charnes et. al (1978) for the decision point is as follows.

$$\begin{aligned}
 \text{Min } h_0 &= \sum_{r=1}^n v_r x_r \\
 \sum_{i=1}^m u_r y_r &= 1 \\
 -\sum_{r=1}^n u_r y_r + \sum_{i=1}^m v_i x_i &\geq 0 \\
 v_i \geq 0; \quad u_r \geq 0; \quad i &= 1, \dots, m \quad r = 1, \dots, n
 \end{aligned}
 \tag{1}$$

3. Application

The data were analyzed in the EMS package program. The objective of the output-oriented CCR model is to maximize the output with fixed input. However, some of the output variables used are negative and such variables are undesirable outputs. For this reason, undesirable output variables are added to the model by dividing them into one. Thus, the model minimizes such variables. The countries of the European Union member countries and Turkey in the data set, and each one has been selected as DMU. Table 3 shows the efficiency scores of countries based on sub-indices.

DEA scores of effective decision making units appear as 1 and below 1 in the output-oriented super efficiency model. As a result of the analysis, Austria, Netherlands, Sweden and United Kingdom are efficient units in Basic Human Needs category. In the Wellbeing category, Austria, Ireland and Belgium are efficient units. Finally, in the Opportunity category, Ireland, Luxemburg, Sweden and United Kingdom are efficient units. Efficiency scores were obtained for the social development levels of the country by taking the average of DEA scores in the categories.

Table 4 shows the efficiency scores and rankings of the countries. According to the results, United Kingdom, Ireland and Luxemburg share the first three places. Austria, Finland and the Netherlands follow them. Lithuania, Bulgaria and Turkey are the last three countries. EMS scores and rankings of efficiency and scores and rankings of SPI were compared in SPSS Program. Spearman Correlation was used to compare ranking, Pearson Correlation was used to compare scores.

Table 3. Efficiency scores for the categories

Countries	Basic Human Needs		Foundations of Wellbeing		Opportunity	
	DEA Score	Rank	DEA Score	Rank	DEA Score	Rank
Austria	89,14%	1	98,85%	1	112,53%	23
Belgium	126,65%	21	99,85%	3	100,81%	6
Bulgaria	139,70%	24	116,33%	26	128,60%	26
Cyprus	137,87%	22	103,99%	12	106,70%	19
Czech Republic	111,74%	12	103,94%	11	105,70%	18
Denmark	103,17%	8	100,60%	5	105,50%	17
Estonia	120,20%	17	104,98%	16	110,71%	22
Finland	102,10%	6	100,85%	6	100,23%	5
France	115,23%	14	104,78%	15	102,91%	11
Germany	103,15%	7	101,51%	8	108,63%	20
Greece	116,67%	16	110,12%	23	102,90%	10
Hungary	121,77%	19	110,07%	22	104,25%	15
Ireland	102,07%	5	99,37%	2	94,77%	3
Italy	112,95%	13	102,56%	10	104,06%	13
Latvia	144,89%	25	109,07%	21	102,39%	9
Lithuania	149,64%	26	105,66%	17	116,80%	24
Luxembourg	105,40%	10	100,01%	4	93,07%	2
Netherlands	97,87%	2	101,21%	7	104,64%	16
Poland	110,83%	11	105,94%	18	109,58%	21
Portugal	120,55%	18	106,36%	20	120,46%	25
Romania	139,40%	23	110,81%	24	101,44%	7
Slovakia	123,15%	20	106,16%	19	103,81%	12
Slovenia	116,47%	15	104,12%	13	102,21%	8
Spain	103,33%	9	104,41%	14	104,14%	14
Sweden	99,77%	3	116,88%	27	99,98%	4
Turkey	161,53%	27	115,77%	25	131,63%	27
United Kingdom	99,86%	4	102,36%	9	79,86%	1

Table 4. scores and ranking of efficiency scores

Countries	Efficiency Score	Rank	Countries	Efficiency Score	Rank
United Kingdom	94,03%	1	Poland	108,78%	15
Ireland	98,74%	2	Belgium	109,10%	16
Luxembourg	99,49%	3	Greece	109,90%	17
Austria	100,17%	4	Slovakia	111,04%	18
Finland	101,06%	5	Estonia	111,96%	19
Netherlands	101,24%	6	Hungary	112,03%	20
Denmark	103,09%	7	Portugal	115,79%	21
Spain	103,96%	8	Cyprus	116,19%	22
Germany	104,43%	9	Romania	117,21%	23
Sweden	105,54%	10	Latvia	118,78%	24
Italy	106,52%	11	Lithuania	124,03%	25
Czech Republic	107,13%	12	Bulgaria	128,21%	26
Slovenia	107,60%	13	Turkey	136,31%	27
France	107,64%	14			

Table 5. Compression of efficiency ranking of DEA and ranking of SPI.

DMU	Super Efficiency Scores	Ranking of DEA	Scores of SPI	Ranking of SPI
Austria	100,17%	4	86,76	12
Belgium	109,10%	16	87,39	10
Bulgaria	128,21%	26	76,27	25
Cyprus	116,19%	22	82,85	18
Czech Republic	107,13%	12	84,66	16
Denmark	103,09%	7	89,96	1
Estonia	111,96%	19	83,49	17
Finland	101,06%	5	89,77	2
France	107,64%	14	87,88	9
Germany	104,43%	9	89,21	5
Greece	109,90%	17	82,59	19
Hungary	112,03%	20	80,11	23
Ireland	98,74%	2	88,82	7
Italy	106,52%	11	86,04	13
Latvia	118,78%	24	79,25	24
Lithuania	124,03%	25	81,86	20
Luxembourg	99,49%	3	89,27	4
Netherlands	101,24%	6	89,34	3
Poland	108,78%	15	81,21	21
Portugal	115,79%	21	85,36	15
Romania	117,21%	23	74,51	26
Slovakia	111,04%	18	80,34	22
Slovenia	107,60%	13	85,50	14
Spain	103,96%	8	87,11	11
Sweden	105,54%	10	88,99	6
Turkey	136,31%	27	66,81	27
United Kingdom	94,03%	1	88,74	8

The following hypotheses have been established to compare rankings

H₀: There is no significant relationship between efficiency rankings and SPI rankings.

H₁: There is a significant relationship between the efficiency rankings and SPI rankings.

Spearman Correlation was used to find out the degree of the relationship between two related values. Since the significance value is less than 0.01, H₀ hypothesis is rejected at this level and the correlation coefficient between the two variables is significant. In addition, the degree of relationship between the two rankings is $r = 0.856$. Therefore, it can be said that there is a strong positive relationship between them.

Then, the presence and degree of correlation between efficiency scores and SPI scores were measured. The following hypotheses were established to establish the relationship between the scores and Pearson Correlation was used for hypothesis testing.

H₀: There was no significant relationship between efficiency scores and SPI scores of the countries.

H₁: There is a significant relationship between efficiency scores and SPI scores of countries.

H₀ hypothesis was rejected because the level of significance was less than 0.01. In other words, there is a significant relationship between DEA scores and SPI scores. In addition, the degree of the level of the relationship $r = -0.874$ shows a high level of relationship. In the analysis, the number r indicates the negative direction. This is due to the fact that the country with the lowest super-efficiency score is the most effective, whereas the SPI is the opposite.

4. Conclusion

The data set contains three categories that make up the Social Progress Index. In these categories, there are again sub-categories within themselves. Within the scope of the study, the performances of the countries on social development were analyzed in the titles of Nutrition and Basic Medical Care, Water and Sanitation, Shelter, Personal Safety, Access to Basic Knowledge, Access to Information and Communications, Health and Wellness, Environmental Quality, Personal Freedom and Access to Advanced Education. The efficiency scores of the categories were obtained by taking the average of the efficiency scores in the sub-categories. Finally, efficiency scores were established by means of the average efficiency scores of the categories. In this way, social development performance of 27 European Union countries in which Turkey has been assessed under the DEA.

According to the results of the analysis, United Kingdom ranks first in the DEA ranking, followed by Ireland and Luxemburg. These countries are also efficient countries. Austria has the highest score among inactive countries. According to DEA in Turkey are the most recent results. Turkey is followed by Bulgaria and Lithonia. Turkey Nutrition and Basic Medical Care, Shelter, is efficient on Environmental Quality and Personal Freedom category. It was observed that there was a high correlation between efficiency scores and rankings and SPI scores and rankings.

Weight assignments are made indicating the importance of the outputs in the results found with DEA. In addition, DEA gives information on which output should change in order to be similar to efficient DMUs. In this way, the study is guiding for the policies that can be formed. In addition, it may be recommended to apply to other country groups as a future study and to evaluate efficacy with different DEA methods in the literature.

References

- Andersen, P. and Petersen, N. C. (1993) “A procedure for ranking efficient units in data envelopment analysis” *Management Science*, vol. 39, no. 10, pp.1261-1264.
- Charnes, A., Cooper, W.W. ve Rhodes, E. (1978). ”Measuring the efficiency on decision making units”. *European Journal of Operational Research*, vol. 2, pp.429–499.
- Sen A., (1999). “Development as freedom”, Oxford University Press, ISBN-13:978-0-19-289330-7
- Sökmen, A. (2014). “Sosyal gelişme endeksi Türkiye için ne ifade ediyor?”, *Türkiye Ekonomi Politikaları Araştırma Vakfı, Değerlendirme Notu*, pp.1-11.

O-17 Predicting the Price of Real Estate Using Decision Tree Approach

Simay Mirgen¹, Betül Kan-Kılınc^{2*}

¹ Institute of Graduate Programs, Eskisehir Technical University, Turkey, E-mail: simaymirgen@hotmail.com

²Department of Statistics, Eskisehir Technical University, Turkey, E-mail: bkan@eskisehir.edu.tr

Abstract – In this study, the relationship between the real estate and its properties are investigated by using a decision tree that is built from the training set including 70% of dataset. Using a public housing platform as a case study, the real estate for sale in Eskisehir have been collected. The existence of some real estate characteristics is recorded as 1 and 0 elsewhere. The dependent variable is the unit price of real estate for 280 observations that is classified in three categories, such as cheap, moderate and expensive. For model validation 5-fold cross validation is used and evaluation metrics are summarized for both train and test data. Addition, the built tree model identified the most significant characteristics of real estate in determining the unit price.

Keywords – *splitting; real estate, tree algorithm*

1. Introduction

Decision tree models can be used for visual analysis of the statistical data, to determine the relationship between variables and to make predictions from the data (Timofeev, 2004, Zeiles et al. 2005). Morgan and Sonquist (1963) developed a recursive partitioning strategy (AID - Automatic Interaction Detection) for a continuous response. Recursive partitioning (RP) models were introduced in the book “Classification and Regression Trees” by Breiman et al. (1984). RP models have also been developed in the machine learning basis, by Quinlan (1986) and C4.5 (1993) algorithms being among the most widely recognized.

RF methods have become popular and often used tools for non-parametric regression and classification in many scientific fields. As CART does not assume any underlying relationship between the dependent variable and the predictors, therefore the model can be easily analysed and interpreted. In tree models, the paths that arise with the answers from the questions of the independent variable form the tree branches visually and show the independent variables or variables that affect the dependent variable. The answers to the questions asked to the independent variable are seen as tree branches so that even if the data set is very complex, the variables affecting the dependent variable and the importance of these variables in the model can be examined visually.

There is a great various of studies concerning of CART applications. For instance, a regression tree is employed to check if property price can be modeled by its attributes, spatial boundary or by spatial coordinates with the same level of precision by Meland et al. (2016). To accomplish their aim, the estimation of real estate property values impacted by the physical characteristics, local amenities and politic-economic factors are examined. The determinants of house prices examined by Fan et al. (2006).

Using the real estate market in Singapore, the paper shows homebuyers are more concerned about basic housing characteristics of two-and three room flats.

In this study, the classification tree is applied to a real data set, including a categorical response and covariates. Our aim is to determine the significant predictors on housing unit price and obtain the performance success of classification tree by using 5-fold cross validation. In this paper, empirical results on the determinants of real estate value in Eskişehir area are used to predict whether a house unit price is expensive, less expensive or cheap given the information like age, amenities, number of rooms, number of bathrooms..., etc.

This paper is organized as follows. Recursive partitioning is described first. Next section introduces the data collected from a website. Then, classification tree algorithm is performed to classify a three classes response. Last section concludes.

2. Materials and Methods

The classification and regression tree (CART) is a non-parametric statistical method used to analyze and estimate the values of both categorical and continuous dependent variables, since it does not require any assumptions for the data set.

The response variable y depends on a vector of p predictors $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and it is modeled with Eq. (1). Assume that there are N samples of \mathbf{y} and \mathbf{x} , $\{y_i, x_i\}_{i=1}^N$. Let $\{R_j\}_{j=1}^S$ be a set of S disjoint subregions of $D \subset \mathbb{R}^p$ such that $D = \cup_{j=1}^S R_j$. The recursive partitioning estimates of the unknown function $f(\mathbf{x})$ at \mathbf{x} with

$$\hat{f}(\mathbf{x}) = \hat{f}_j(\mathbf{x}) \text{ for } \mathbf{x} \in R_j \quad (1)$$

where the function $\hat{f}_j(\mathbf{x})$ estimates the unknown function over the R_j^{th} subregion of D . In recursive partitioning, $\hat{f}_j(\mathbf{x})$ is frequently taken to be the constant function (Morgan, 1963; Breiman, 1984; Quinlan, 1986). Hence, $\hat{f}_j(\mathbf{x})$ is taken as given in Eq. (2)

$$\hat{f}_j(\mathbf{x}) = c_j \quad \forall \mathbf{x} \in R_j \quad (2)$$

where c_j is chosen to minimize the i^{th} component of the residual squared error,

$$rse[\hat{f}_j(\mathbf{x})] = \min_{c_j} \sum_{\mathbf{x} \in R_j} (y_i - c_j)^2 \quad (3)$$

Since the subregions of the domain D are disjoint, each c_j will be the sample mean of the y_i 's whose $\{x_i\}_{i=1}^N \in R_j$. It is noted that the number of the partition does not change during the partition but the number of disjoint subregions that partition D (Stevens, 1991).

The distribution of the tree shows the partition structure. The root shows the total number of observations that results the largest information gain. Subsequent splits refer to the same statistics for the associated data subsets. The terminal nodes represent the remaining observations and the prediction for the outcome (represents the class label).

The core algorithm for building decision trees called ID3 is developed by Quinlan. It employs a top-down, greedy search through the space of possible branches with no backtracking. This algorithm uses Entropy and Information Gain to construct a decision tree.

a) Entropy for one attribute:

$$E(S) = \sum_{i=1}^c -p_i \log_2(p_i) \quad (4)$$

b) Entropy for two attributes:

$$E(T, X) = \sum_{c \in X} P(c)E(c) \quad (5)$$

The information gain is computed as given Eq(6) that based on the decrease in entropy after the dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (6)$$

3. Application

The data published from October 2018 to May 2019 on a widely website was used in the analysis. The data includes the characteristics of 280 houses for sale in Eskişehir. These features are price of houses, land, age, type of room, bedroom, number of bathrooms, garage, social environments around (hospital, school, shopping center, etc.), balcony status, the floor of the apartment, the elevator status (elevator) and the district where the apartment is located.

The ratio (house unit price, TL/m²) calculated by dividing house prices by square meters is used as the dependent variable. Dependent variable housing unit price (TL/m²) is classified into three categories in order structure. To determine the category borders, Central Bank' s published the housing unit prices in Turkey (TL/m²) for 2018 on March 11, 2019 was used (EVDS). The minimum price 2.118,52 (TL/m²) and the maximum 2.314,83 (TL/m²) were used to create the ordered categories such as cheap, expensive, and moderate elsewhere. The variables room, land, bedroom and bathroom are continuous ones whereas the garage, amenities and district are the categorical predictors. The district is classified in two classes: Odunpazari and Tepebaşı. The apartments with garage or amenities are recorded as 1 and 0 elsewhere.

In order to measure the classification success of the models obtained by both methods, the data set was divided into three sets such as training (50%), validation (25%) and test (25%). Training and validation data sets were used for modeling whereas 5-fold cross validation was used to avoid overfitting. The results obtained from the model were 73.5%, 72.9% and 74.1% of accuracy respectively.

The tree that represents a series of splits from top of the tree is given in Figure 1. Each split is based on classifying the predictor variable that is submitted by a series of tests to determine the class label.

Decision nodes are the questions like “What’s the age of the flat? “How large is the flat?”, “Does it have a garage?” Algorithm runs this using a hierarchical structure.

In Figure 1, the classification tree for housing unit price is presented. The first node level splits into two parts, one with variable elevator with and the other without elevator. The second node level splits, similarly, each part into two parts, with variable district consisting of Odunpazarı and the other one with Tepebaşı. The terminal nodes are node 2, node 3 and node 4. The percentage of the expensive apartments with elevator in Odunpazarı is obtained as 41.2% whereas the percentage of the expensive apartments in Tepebaşı is obtained as 25.9%. The percentage of the cheap apartments with elevator in Tepebaşı is 48.2% whereas the percentage of the cheap apartments in Odunpazarı is 29.4%. Also, the percentage of the cheap apartments without elevator is almost 60%.

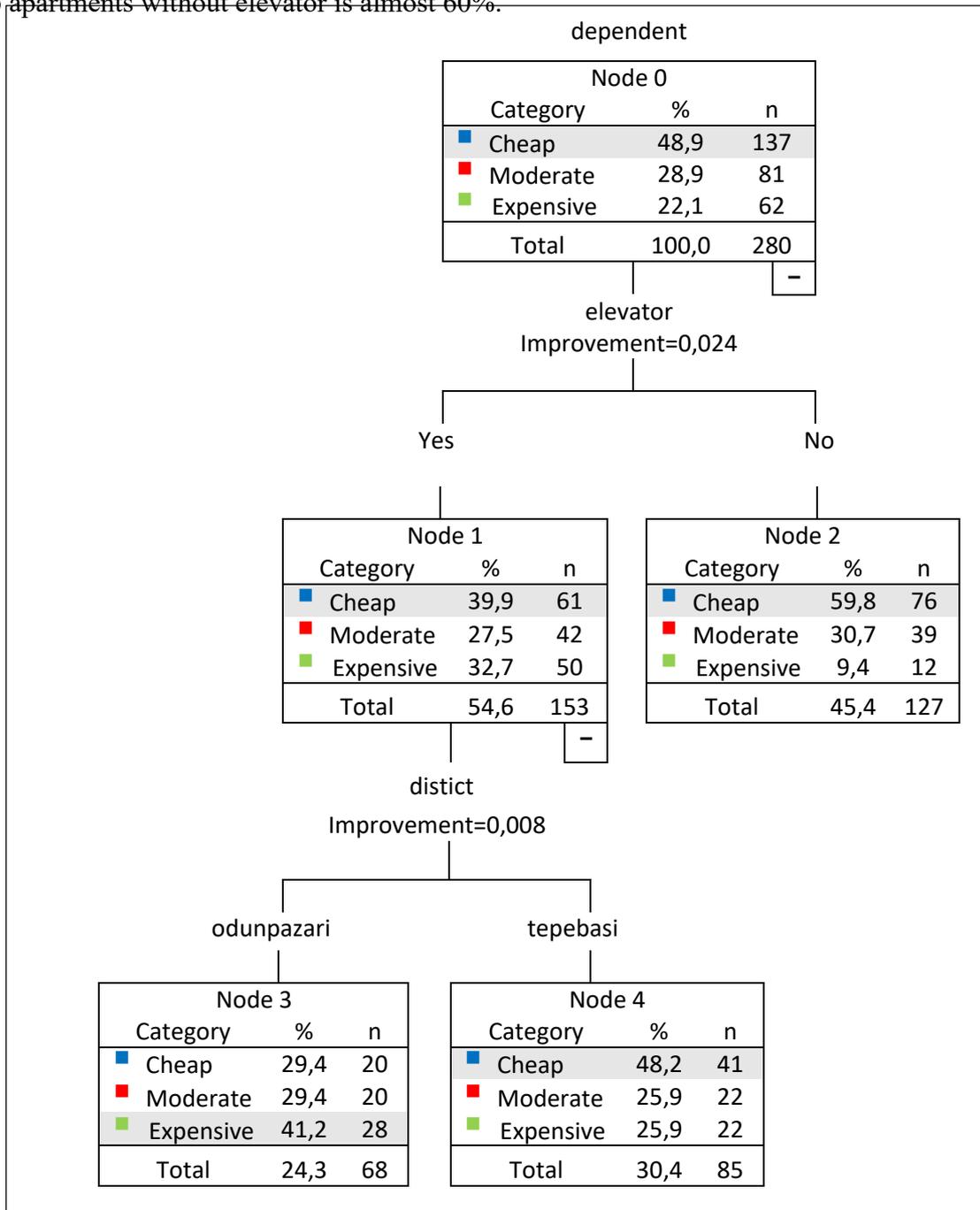


Figure 1: Classification tree for housing unit prices

Conclusion

In this study, the potential variables that may affect housing unit prices in Eskişehir province are considered for comparisons. The most significant variables affecting the housing unit price are observed as elevator and district. The 5-fold cross validation is used to avoid overfitting and the results from the test data set showed that, classification tree is sufficiently applied to the housing price data set. To extend this study, researchers may use the alternative methods to performance comparison or study a larger size of records.

References

- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth International Group, Belmont CA.
- Fan, G.Z., Ong, S.E., Koh, H.C. (2006). “Determinants of House Price”, *Urban Studies*, 43(12). 2301-2315.
- Meland, E., Hunter A., Barry, M. (2016). “Identification of locational influence on real property values using data mining methods”, *European Journal of Geography*,
- Morgan, J.N. and Sonquist, J.A. (1963). “Problems in the Analysis of Survey Data, and a Proposal”, *Journal of the American Statistical Association*, 58, 415-434.
- Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, 1, 81-106.
- Quinlan, J. R. (1993) *C4.5: Tools for Machine Learning*, Morgan Kauffman, San Mateo, CA.
- Stevens, J.G. (1991). *An Investigation of Multivariate Adaptive Regression Splines for Modelling and Analysis of Univariate and Semi-Multivariate Time Series Systems*, PhD Thesis.
- Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*. A Master Thesis Presented. Berlin: Humbolt University, CASE.
- Zeileis, A., Hothorn, T., Hornik, K. (2005). “Model-based Recursive Partitioning”. Research Report Series 19, Institut für Statistik und Mathematik, WU Vienna University of Economics and Business, Vienna.

O-18 Anxiety and Attitudes towards Biostatistics and Scientific Research Methods Courses of Students in a Dental School

Adnan Karaibrahimoğlu^{1*}, Karaoğlu, Nazan² and Karabekiroğlu, Said³

¹*Biostatistics & Medical Informatics Dept/Faculty of Medicine, Süleyman Demirel University, Turkey,*

adnankaraibrahim@gmail.com

²*Family Practice Dept/ Faculty of Medicine, Necmettin Erbakan University, Turkey,*

drnkaraoglu@gmail.com

³*Restorative Dentistry and Treatment Dept/Faculty of Dentistry, Necmettin Erbakan University, Turkey,*

dentisaid@hotmail.com

Abstract – Attitude and anxiety are two important terms related to psychology. However, there are often used in education. College students sometimes feel anxious and develop an attitude for some courses because of many reasons. The aim of this study is to determine the anxiety and attitude level of students in a dental school during three semesters consecutively. After approval of the ethics committee, two separate scales which are Statistics Attitude Scale (SAS) and Scientific Research Methods Attitude Scale (SRMAS) with shown the reliability and validity were applied to the volunteer 152 first-year students. The data collection process took three years to compare the different terms since the students take the biostatistics course only in the first year of their academic education. There was a significant difference in the anxiety level of both statistics and scientific research methods between the terms. The age, gender, high school type, economic level, and the residence location were not affecting factors of the attitude and anxiety. The Biostatistics and scientific research methods courses should be given to the health sciences students decreasing the level of anxiety by more clinical examples, applications and various statistical softwares.

Keywords – *Anxiety, attitude, statistics, scientific research methods, dental school*

1. Introduction

The behavior that individuals develop in the case of any phenomenon is called attitude. It is unique to individuals and ensures that the tendency developed against ideas or institutions continues consistently (Zanakis & Valenzi, 1997). Anxiety manifests itself as indeterminate fear and distress in the individual due to the stimuli present. Unlike the concepts of fear or phobia, anxiety is used to describe situations that are disliked and whose physiological effects are not evident (Ergüven, Işık, & Kılınc, 2013). At the same time, anxiety is the observable emotional reactions that cause an individual to feel tense in the face of threat (Doğan & Çoban, 2009). There are two types of state and trait anxiety. If the state of tension in an individual's subjective experiences persists for a certain period of time, the state is called anxiety (Faber & Drexler, 2019). However, as a result of the intensification of state anxiety, it becomes a form of constant anxiety and becomes increasingly dangerous for the individual (Büyüköztürk, 1997). Statistics is one of the courses that almost every student has to face during the university years. As it is a common language of science, almost all departments offer courses related to statistics or statistics only (Mwebesa, Novembrieta, & Musinguzi, 2018). They are confronted in the process of master or doctorate education since the basis of academic research is depended on the research methods. In addition, statistical

information is often needed during the preparation of the thesis or final paper. However, it is reported that students have anxiety in statistical education which is so important for scientific development and academic achievement. Statistical anxiety is defined as the anxiety, negative attitude, pressure, mental obsession or reservation felt for the statistical course taken at any level or in any department (Paechter, Macher, Martskvishvili, & Wimmer, 2017). Attitudes towards statistical or scientific research courses are expected to be positive. It is stated that the optimal level of anxiety increases the success, but the increase in the level results in mental distress and physiological effects in the individual and results in failure (Townsend, Moore, Tuck, & Wilton, 1998). Although the awareness about statistical anxiety started in the 1980s, it is observed that the publications in this field increased in the 2000s. Studies related to the anxiety and attitudes of students who took statistics courses in different departments such as social sciences, natural sciences, engineering, and medicine were conducted (Onwuegbuzie, 2000). Although a few studies have been conducted for medical students and trainees in the field of health, no statistical anxiety study has been found for the students in the dental school.

The aim of this study is to measure the attitudes and anxiety levels of students of a faculty of dentistry towards biostatistics and scientific research methods courses and determine the effects of various demographic characteristics on anxiety.

2. Materials and Methods

Describe in detail the materials and methods used when conducting the study. The citations you make from different sources must be given and referenced in references.

2.1 Population of the Study

The population of the study consisted of the first year students of a dental school in a university. The school accepts approximately 70 students each year. In order to make comparisons according to different years, information was obtained from freshman students for three years.

2.2 Methodology of the Study

This is a cross - sectional survey application. A questionnaire was prepared to measure the attitudes and concerns of biostatistics and scientific research methods courses of dentistry faculty students. The questionnaire consists of three parts: demographic characteristics, statistical anxiety scale (İYKÖ) and attitude towards scientific research methods course (BAYD-TÖ). The İYKÖ scale consists of 33 items and was evaluated as "1- disagree" and "5- totally agree" at the 5-point Likert level. A total of 18 items were reversed because they contained negative expressions. Validity and reliability studies were performed for the original scale and Cronbach's alpha = 0.927. As a result of exploratory factor analysis (EFA), the scale consisted of 5 sub-dimensions and the sub-dimensions were named as the relationship of statistics with professional life, statistical anxiety-fear, enjoyment of statistics, the importance of statistics and perceived statistical difficulty (Yaşar, 2014b). The BAYD-T scale consisted of 20 items and 5 items were reverse coded because they contained negative expressions. Scale application was evaluated at 5-point Likert level. As a result of the validity and reliability analyzes, the internal consistency coefficient of the scale was found to be 0.917 and four factors were obtained from the EFA. These dimensions were named as the importance of scientific research, cognitive self-confidence, interest-positive attitude and

the relationship between daily life and profession. As a result of Confirmatory Factor Analysis (CFA), the correlation values between the factors were found to be high and significant (Yaşar, 2014a).

2.3 Data Collection

After the ethical approval of the non-drug research ethics committee of the faculty (Date: 10.12.2014 and No: 2014/003), a survey was started for the students. In the first year, a total of 52 students were surveyed on a voluntary basis. In the first year, six weeks had passed since the Biostatistics course started. In the first week of the second year Biostatistics course, a total of 75 students were surveyed. The third year was applied to 25 students in the last week of the semester. It was not possible to reach a sufficient number of students as it was before the end of the term and final exams.

2.4 Statistical Analyses

Statistical analysis of the study was performed with SPSS 20.0 (IBM Inc., Chicago, IL, USA) program. Descriptive measures were presented as frequency (percentage) and mean \pm SD. The mean scores of the general and sub-dimensions of the scales were calculated and compared according to demographic characteristics. As the scale scores showed normal distribution, Student's t-test was used for comparison of two independent groups and one-way analysis of variance was used for multiple groups. Cronbach's alpha values were calculated by reliability analysis. Exploratory factor analysis was performed using the Varimax rotation method. Spearman's Rho correlation analysis was used to determine the relationships between the factors. In the whole study, type-I error value was taken as 5% and $p < 0.05$ value was accepted as statistically significant.

3. Results

A total of 152 students participated in the study. Of these, 34.2% were first year students, 49.3% were second year students and the rest were third year students. More than half of the participants were female (57.9%; $n = 88$) and 42.1% were male. The majority of the participants came from outside the province of Konya (69.5%). Most of the students were Anatolian/Science high school graduates (88.7%) and the rest was graduated from the private high schools. The majority of the students had moderate economic income (72.4%) and the rest (25.7%) were good. Only two participants declared a poor economic level. Most of them stated that they prefer to study in the dentistry department of their own initiatives (86.6%). However, almost half of the participants (48.7%) stated that they would choose dentistry again if they had the chance to choose again. The age range of the participating students was determined as 18-25 years. The median values of the university exam scores to achieve the dental school were 445, 479 and 448, respectively.

As a result of the reliability analyzes, Cronbach's alpha value was found to be 0.907 and scientific research methods attitude scale reliability value was found to be 0.899. Extraction (h^2) values were all over 0.50. The principal components analysis using the Varimax rotation method yielded 6 factors for İYKÖ. However, since the variance value explained by the last of the factors was 3%, it was seen that it could be collected in 5 sub-dimensions and a result close to the original scale was obtained (KMO = 0.864). KMO value obtained for BAYD-TÖ was 0.873, and the number of factors obtained was found to

be four. The mean value of the İYKÖ total scale score was 2.77 ± 0.57 (1.12-4.24) and the mean of BAYD-TÖ total scale score was 2.59 ± 0.67 (1.10-4.40). When the questionnaire scores of the subscales were examined, it was seen that the relationship between statistics and professional life had the highest score (3.08 ± 0.76). Then, respectively, the fear of statistics (2.90 ± 0.59), the importance of statistics (2.89 ± 0.77), perceived statistical difficulty (2.84 ± 0.71) and statistical anxiety-fear (2.16 ± 0.81) dimensions were calculated. Total scale scores were compared between years. The value for the second year was significantly lower than the other years for both İYKÖ ($p=0.011$) and BAYD-TÖ ($p = 0.020$) (Figure 1 and Figure 2).

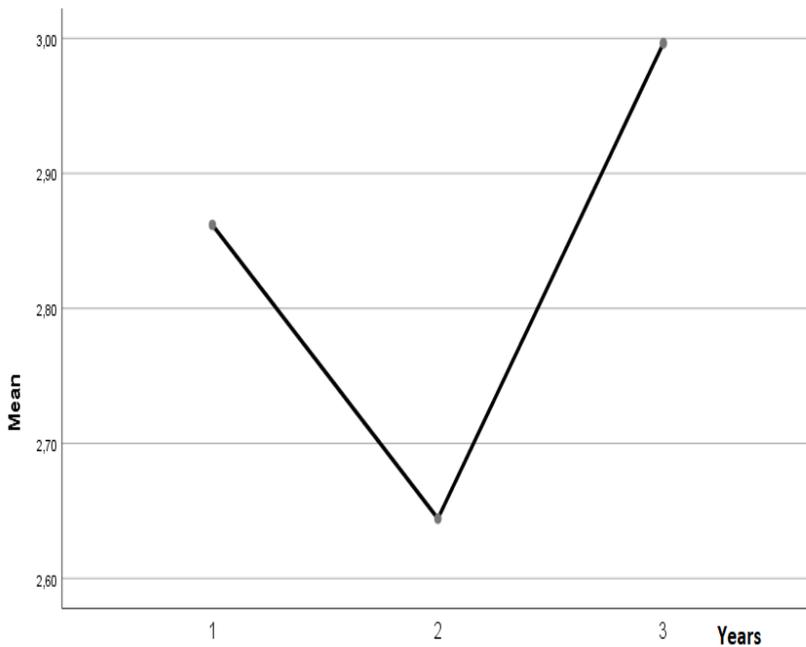


Figure.1 Anxiety level towards statistics with respect to three years (out of 5.00 points)

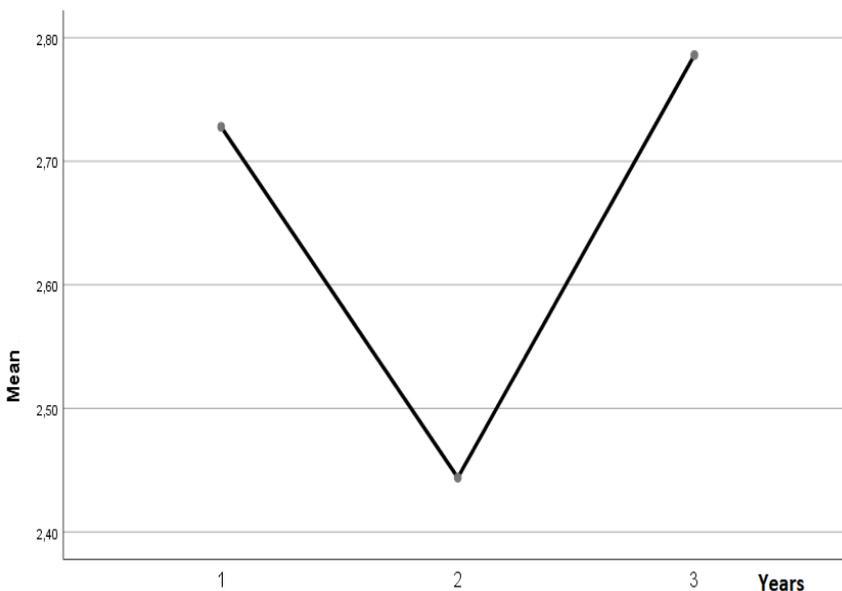


Figure.2 Anxiety level towards scientific research methods with respect to three years (out of 5.00 points)

In the comparisons according to demographic characteristics, it was observed that anxiety or attitude scores did not differ in general. In the first year surveys, both the anxiety about statistics and the attitude scores towards the scientific research methods course were not different between the genders. In addition, the fact that the student comes from outside the province, the type of high school he or she graduated from, the economic level and the idea of choosing dental school again did not affect the anxiety and attitude. However, the level of statistical anxiety was significantly lower in students who did not choose the dental school on their own initiatives ($p = 0.025$). The level of attitudes towards scientific research was also low ($p = 0.097$). In the second year, negative attitudes towards scientific research methods were significantly higher in male students ($p = 0.045$). It was observed that the family away from the province, the type of high school graduated, the economic level, the willingness to choose dentistry and the desire to choose dentistry again did not affect the statistical anxiety and attitude towards scientific research. In the third year, there was no significant difference between the students' demographic characteristics in terms of statistical anxiety and attitude towards scientific research. There was a low and negative correlation between university entrance exam score and statistical anxiety ($r = -0.25$) and attitude towards scientific research ($r = -0.24$).

When the subscales of the scales were compared by years, the relationship between statistics and occupational life and the importance of statistics were not significantly different, while the fear of statistics, enjoyment of statistics and perceived statistical difficulty were found to be significantly lower in the second year participants. In addition, the importance of scientific research, cognitive self-confidence, attitudes and the relationship between scientific research and profession were found to be significantly lower in the second year participants. No significant relationship was found between sub-dimensions and gender or economic level. It was understood that the demographic characteristics such as being from outside the province, the reason for choosing a dental school and the idea of choosing dentistry again did not affect the scores of the sub-dimensions. Only those who do not prefer to choose the dentistry again on their own initiatives will have a significantly lower perception of the relationship between statistics and professional life ($p = 0.037$). A significant and high positive correlation was calculated between the two scales ($r = 0.73$; $p < 0.001$). Therefore, there was a negative and low correlation between all sub-dimensions and university exam scores for dentistry.

4. Discussion and Conclusion

Numerical courses, especially mathematics and statistics, create undoubtedly a concern and fear for students (Baloğlu, 2003; Sloopmaeckers, Kerremans, & Adriaensen, 2012; Onwuegbuzie & Wilson, 2003). The students' attitude towards the statistical course and anxiety towards the course affect the success negatively after a certain level (Onwuegbuzie, 2004). Statistical anxiety has been tried to be measured at very different educational levels and in different faculties (Faber, Drexler, Stappert, & Eichhorn, 2018). Not only the attitudes towards anxiety and its sub-dimensions were measured, but also the relationships between many factors such as anxiety and achievement, age, gender, perception, social and cultural environment were examined (Birenbaum & Eylath, 2011; Shahram, 2011). On a study conducted with only 151 female students in educational sciences, it was reported that statistical anxiety was related to numerical ability (Birenbaum & Eylath, 2011). In another study, 26 students randomly

selected from different non-mathematics-based departments showed that both the less anxious and the more anxious students were more successful at the end of the final exam with unlimited time (Onwuegbuzie & Seaman, 1995). 394 students from the department of business administration participated in the study in which the barriers were determined for success in statistics. It was stated that content difficulty and the importance given to statistics emerged as the most important barrier and cognitive capacity could be solved by frequent repetition of statistical modules (Chifurira, Mudhombo, & Chikobvu, 2014). In a study conducted with 435 students who took a statistics course in the first year of education in a university, it was found that low statistical perception caused negative attitudes, and positively correlated with an increase in competence, desire, effort and success (Ncube & Moroke, 2015). In a study conducted with 431 students studying in the behavioral sciences, it was seen that statistics and mathematics anxiety were parallel, and both exam anxiety and content constraint factors came to the fore for both concepts (Zeidner, 1991). In a study designed with 76 students from different departments as pre-test and post-test, the STARS scale was used and it was found that the effect of the trainer's interest and immediacy on statistical anxiety was measured (Williams, 2010). In the study conducted with 472 participants who took a statistics course in health sciences departments, the relationship between statistical anxiety and exam anxiety, statistical course performance, attitude towards statistics and trait anxiety was examined. Numerical skills were found to reduce statistical anxiety and increase test success (Sesé, Jiménez, Montaña, & Palmer, 2015).

Scale development studies related to statistical anxiety are sufficiently high. However, some of them have been used more frequently than others. In a study designed by Earp as a Ph.D. thesis, SAM (Statistics Anxiety Measure) scale was created with the participation of 131 undergraduate students and 215 graduate students, and five sub-dimensions were obtained (Earp, 2007). Psychometric evaluation of the STARS scale was conducted with 400 freshman and sophomore students, and attitude towards statistics was found to be negative (Papousek et al., 2012). In a scale development study conducted in Germany with 113 graduate students, anxiety, overcoming and emotional state dimensions related to statistical anxiety were studied (Faber et al., 2018). In a study conducted in the UK, the adaptation of STARS (Statistics Anxiety Rating Scale) was conducted with 650 psychology students and six sub-dimensions of the scale were found to be successful in measuring statistical anxiety (Hanna, Shevlin, & Dempster, 2019). In the study conducted with 83 sophomore students in the Sociology Department, SAT-34 scale was adapted in Russian. Six dimensions were obtained under the headings of statistics in business life, statistics in daily life, expectations, interests, efforts, and difficulties. It was observed that students generally developed negative attitudes toward statistics (Khavenson & Orel, 2014). Although the number of studies on anxiety and attitude towards statistics is quite high, there are few studies on anxiety and attitude towards scientific research methods. In a study conducted with 113 undergraduate and 93 graduate students studying in the department of educational sciences, 12-item one-dimensional Research Anxiety Scale (AYKÖ) was developed. It was found that the anxiety level of undergraduate students was slightly higher than that of graduate students (Büyükoztürk, 1997). In a study conducted with the participation of 142 junior students studying in the field of educational sciences, it was found that research experience and success in the course affected the anxiety about the research, and gender was not effective (Büyükoztürk, 1999). In a study conducted with social sciences students in the following years, a four-dimensional scale was obtained by conducting the Attitude Scale Towards Scientific Research Methods and tried to eliminate the deficiency in this area (Yaşar, 2014a).

In this study, anxiety and attitude study were conducted by using the freshman dentistry students together with the İYKÖ and BAYD-TÖ scales together. Similar to previous studies, it was found that factors such as gender, coming from different regions, age and economic status did not affect the anxiety level in general. The only effective factor was the voluntary selection of dental school. It was seen that the students who chose the dental school for other reasons had higher anxiety and negative attitude. The most important factor related to the difference between anxiety levels was the different periods. While the anxiety level was lower and more positive attitude score was obtained for the students in the periods in which a more interactive educational environment was established in the course, the students of the second period in the non-interactive and non-immediatist educational environment had more anxiety points. In addition, the biostatistics curriculum was redesigned and the need for numerical ability and mathematics skills was minimized. This situation caused the level of anxiety to be lower than the previous studies, and the attitude towards statistics was more positive. Anxiety level was lower in students with higher university entrance scores. In addition, a significantly high correlation was found between the anxiety and attitude towards statistics and the attitude towards scientific research methods.

Although it is reported that there is a high level of anxiety in students studying in social sciences or health sciences, it is observed that the level of anxiety is low in this study conducted with dental students. It can be concluded that the level of anxiety about statistics may be reduced with applications such as being more comprehensible and simpler the curriculum, devoting time to computer applications in biostatistics course, allowing more time to solve exercises and establishing a reliable communication with the students. Therefore, positive developments in the attitude towards both statistical and scientific research methods can be achieved.

References

- Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences*, 34: 855–865.
- Birenbaum, M., & Eylath, S. (2011). Who is afraid of statistics? Correlates of statistics anxiety among students of educational sciences. *Educational Research*, 36(1): 93–98. <https://doi.org/10.1080/0013188940360110>
- Büyüköztürk, Ş. (1997). Araştırmaya Yönelik Kaygı Ölçeğinin Geliştirilmesi. *Eğitim Yönetimi*, 3(4): 453–464.
- Büyüköztürk, Ş. (1999). Araştırmaya Yönelik Kaygı ile Cinsiyet, Araştırma Deneyimi ve Araştırma Başarısı Arasındaki İlişki. *Eğitim ve Bilim*, 23(112): 29–34.
- Chifurira, R., Mudhombo, I., & Chikobvu, M. (2014). Surrendering with Statistics ! Are University Students Stooping in with a Deficit ? A Preliminary Analysis of the Constraints at a University in a Developing Country. *Mediterranean Journal of Social Sciences*. 5(3):234-242. <https://doi.org/10.5901/mjss.2014.v5n3p234>
- Doğan, T., & Çoban, A. E. (2009). EĞİTİM FAKÜLTESİ ÖĞRENCİLERİNİN ÖĞRETMENLİK MESLEĞİNE YÖNELİK TUTUMLARI İLE KAYGI DÜZEYLERİ ARASINDAKİ İLİŞKİNİN İNCELENMESİ. *Eğitim ve Bilim*, 34(153): 157–168.

- Earp, M. (2007). Development and Validation of the Statistics Anxiety Measure. *Doctoral Dissertation*-Unpublished, College of Education, University of Denver.
- Ergüven, S. S., Işık, B., & Kılınç, Y. (2013). Diş hekimliği fakültesi birinci sınıf öğrencileri ile son sınıf öğrencilerinin dental kaygı-korku düzeylerinin karşılaştırmalı olarak değerlendirilmesi. *Acta Odontologica Tursica*, 30(2):70-76
- Faber, G., & Drexler, H. (2019). Predicting Education Science Students ' Statistics Anxiety : The Role of Prior Experiences Within a Framework of Domain-Specific Motivation Constructs. *High Learn. Res. Commun.*, 9(1): 10–27. <https://doi.org/10.18870/hlrc.v9i1.435>
- Faber, G., Drexler, H., Stappert, M. A. A., & Eichhorn, B. A. J. (2018). Measuring Education Science Students ' Statistics Anxiety Conceptual Framework , Methodological Considerations, and Empirical Analyses, *Research report*. <https://doi.org/10.13140/RG.2.2.18109.31201>
- Hanna, D., Shevlin, M., & Dempster, M. (2019). The structure of the Statistics Anxiety Rating Scale : A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences*, 45(2008): 68–74.
- Khavenson, T. E., & Orel, E. A. (2014). Dispositional Factors of Attitudes Towards Statistics in Social Science Students : Perseverance and Academic Motivation. *Journal of the Higher School of Economics: Psychology*, 11(3): 37–54.
- Mwebesa, E., Novembrieta, S., & Musinguzi, D. (2018). Antecedents of statistics anxiety in a higher education system. *Preprints* (August): 1–12. <https://doi.org/10.20944/preprints201808.0101.v1>
- Ncube, B., & Moroke, N. D. (2015). STUDENTS ' PERCEPTIONS AND ATTITUDES TOWARDS STATISTICS IN SOUTH AFRICAN UNIVERSITY : AN EXPLORATORY FACTOR ANALYSIS APPROACH. *Journal of Governance and Regulation*, 4(3): 231–240.
- Onwuegbuzie, A. J. (2000). Statistics anxiety and the role of self-perceptions. *Journal of Educational Research*, 93(5): 323–330. <https://doi.org/10.1080/00220670009598724>
- Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assessment and Evaluation in Higher Education*, 29(1): 3–19. <https://doi.org/10.1080/0260293042000160384>
- Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *Journal of Experimental Education*, 63(2): 115–124. <https://doi.org/10.1080/00220973.1995.9943816>
- Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education*, 8(2): 195–209. <https://doi.org/10.1080/1356251032000052447>
- Paechter, M., Macher, D., Martskvishvili, K., & Wimmer, S. (2017). Mathematics Anxiety and Statistics Anxiety . Shared but Also Unshared Components and Antagonistic Contributions to Performance in Statistics. *Front Psychol.* 8(1196): 1–13. <https://doi.org/10.3389/fpsyg.2017.01196>
- Papousek, I., Ruggeri, K., MacHer, D., Paechter, M., Heene, M., Weiss, E. M., ... Freudenthaler, H. H. (2012). Psychometric evaluation and experimental validation of the statistics anxiety rating scale. *Journal of Personality Assessment*, 94(1): 82–91. <https://doi.org/10.1080/00223891.2011.627959>

- Sesé, A., Jiménez, R., Montaña, J., & Palmer, A. (2015). Can Attitude toward Statistics and Statistics Anxiety Explain Students ' Performance ? Can Attitude Toward Statistics and Statistics Anxiety Explain Students ' Performance ?. *Revista de Psicodidactica*, 20(2):285-304. <https://doi.org/10.1387/RevPsicodidact.13080>
- Shahram, V. (2011). Canonical correlation analysis of procrastination, learning strategies and statistics anxiety among Iranian female college students. *Procedia - Social and Behavioral Sciences*, 30:1620–1624. <https://doi.org/10.1016/j.sbspro.2011.10.314>
- Slotmaeckers, K., Kerremans, B., & Adriaensen, J. (2012). TOO AFRAID TO LEARN?! ATTITUDES TOWARDS STATISTICS AS A BARRIER TO LEARNING STATISTICS AND TO ACQUIRING QUANTITATIVE SKILLS. *Politics*, 34(2), 1–12.
- Townsend, M. A. R., Moore, D. W., Tuck, B. F., & Wilton, K. M. (1998). Self-concept and anxiety in university students studying social science statistics within a co-operative learning structure. *Educational Psychology*, 18(1): 41–54. <https://doi.org/10.1080/0144341980180103>
- Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education*, 18(2): 1–18. <https://doi.org/10.1080/10691898.2010.11889495>
- Yaşar, M. (2014a). Bilimsel Araştırma Yöntemleri Dersine Yönelik Tutum Ölçeği Geliştirme Çalışması: Geçerlik ve Güvenirlik. *Eğitim Bilimleri ve Araştırmaları Dergisi*, 4(2): 109–129.
- Yaşar, M. (2014b). İstatistiğe Yönelik Tutum Ölçeği : Geçerlilik ve Güvenirlik Çalışması Attitudes Toward Statistics Scale : Validity and Reliability Study. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 36: 59–75.
- Zanakis, S. H., & Valenzi, E. R. (1997). Student Anxiety and Attitudes in Business Statistics. *Journal of Education for Business*, 73(1): 10–16. <https://doi.org/10.1080/08832329709601608>
- Zeidner, M. (1991). Statistics and Mathematics Anxiety in Social Science Students: Some Interesting Parallels. *British Journal of Educational Psychology*, 61(3): 319–328. <https://doi.org/10.1111/j.2044-8279.1991.tb00989.x>

O-19 Analysis of Turkey Household Budget Survey Data with Quantile Regression

Ismail YENILMEZ¹, Yeliz MERT KANTAR^{2*}

¹Department of Statistics, Eskisehir Technical University, Turkey, ismailyenilmez@eskisehir.edu.tr

²Department of Statistics, Eskisehir Technical University, Turkey, ymert@eskisehir.edu.tr

Abstract – The linear regression (LR) models the relationship between independent variable(s) and the conditional mean of a dependent variable. Ordinary least squares (OLS) estimation is the best estimation method for regression model under the certain assumptions such as homoscedasticity and normality. However, if these assumptions are not satisfied for LR and/or outliers are detected in data, LR is not suitable to model data. Such cases, alternative models or estimation methods may be used. The quantile regression model, which considers the quantile of the dependent variable instead of its mean, is one of the alternative models. Moreover, while LR and robust alternatives focus on the mean of the response variable at each value of the predictors, the quantile regression provides more details about the probability distribution of the response variable and explains the effect of the predictors on quantiles of the response. In this study, we have considered the Turkey Household Budget Survey data. Data is taken from the TurkStat. We have observed that heteroscedasticity in residuals and moderately skewed distribution of residuals. Thus, we have considered quantile regression analysis for the Turkey Household Budget Survey data. The obtained results show that quantile regression estimates, which are calculated on the basis of quantiles and divided into 5 different expenditure groups, are quite different from OLS estimates. Rather than estimating the whole group with a single model based on the expected value, the use of quantile regression provides more precise and detailed results.

Keywords – Turkey Household Budget Survey data, quantile regression, comparative study

1. Introduction

Nowadays, data stacks are increasing exponentially. As data size and complexity increase, it becomes more difficult to estimate and explain to statistical model based on such data. When existing methods sometimes fail to estimate and explain model due to certain reasons, new methods or new models are needed. An example of new models is the quantile regression (QR). Ordinary least squares (OLS) regression, called as classical linear regression (LR), is well-known and widely-used statistical technique. However, it is possible to talk about its superiority under certain assumptions. Recently, the QR is considered as an alternative method for cases where the assumptions of OLS regression (or LR) are not met.

QR is based on conditional quantile functions. This statistical method is introduced by Koenker and Bassett (1978). In pioneering study of Koenker and Bassett, a minimization problem yielding the ordinary sample quantiles is generalized to the linear model generating a new class of statistics called as *regression quantiles*. The QR regression estimate models for the conditional median function and other conditional quantile functions. Koenker and Hallock (2001) stated that Boscovich laid the foundations of idea of QR in the 18th century. Laplace and Edgeworth were among those who produced ideas for QR. It can be said

that the literature is developed under the hegemony of LR for a long time. Since the mid-nineteenth century, important theoretical findings on QR have been provided. From the perspective of the QR, quantiles, ranks and optimization terms are discussed by Koenker (2005). Likelihood based inference for QR using the asymmetric Laplace distribution is introduced by Sanchez et al. (2013). In another similar study, likelihood process using skewed distributions is proposed for QR by Galarza et al. (2017). A framework is introduced for QR in binary longitudinal data by Rahman and Vossmeier (2019). For QR, the maximum likelihood estimators (MLE) is studied under the unimodal or bimodal distribution for QR by Gomez et al. (2019). Goodness-of-fit procedures for QR is introduced by Koenker and Machado (1999). QR have been used for different applications. Concepts such as wage trend, earning function, sample selection, birth outcomes, etc. have been examined by QR. These researches have been compiled and presented by Fitzenberger et al. (2002). In addition, in the majority of the theoretical studies mentioned above, applications for real life data are found besides simulation.

Studies in which expenditure is modeled by QR are also available in the literature. A simple but basic example is presented by Koenker and Hallock (2001) for Engel's classic empirical application data in the economics. The main determinant of tourist expenditure is investigated by Marrocu et al. (2015) using LR and QR models. The motivation of this study is the modeling of household expenditures from different aspects. For this purpose, the household budget survey (HBS) data has been used. Considering that this modeling differs according to the groups, it is useful to consider the expenditures from different perspectives. QR is a good procedure for this purpose.

The paper is organized as follows: Section 2 briefly reviews QR and also LR and QR have been introduced from the theoretical perspective. Section 3 presents the results of analysis of Turkey HBS data with QR. Finally, the obtained results are presented and discussed in conclusion section.

2. Quantile Regression

θ -Quantile (q_θ) of a continuous random variable Y can be written as follows:

$$F_Y(q_\theta) = \Pr(Y \leq q_\theta) = \theta \quad (1)$$

where $\theta \in (0,1)$ and $F_Y(.)$ is cumulative distribution function (CDF). Quantile can be defined as probability that random variable Y is smaller than or equal to q_θ is equal to θ . $F_Y(.)$ is invertible (continuity of $F_Y(.)$ is assumed).

$$F_Y^{-1}(\theta) = q_\theta \quad (2)$$

In OLS procedure for classical LR, estimates of the parameters are obtained by minimizing the least sum of squares function. In ML procedure, parameters estimates are obtained by maximizing the likelihood function. In the QR model, *absolute error function* is minimized instead of minimizing *least squares error function*, thus, it is more robust than the LR model.

The LR model can be written as

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, p \quad (3)$$

β_j can be estimated by minimizing least squares optimization problem as follows

$$\widehat{\beta}_j = \underset{\beta_j}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \quad (4)$$

where $\widehat{\beta}_j$ is an estimate of the parameter of the LR model.

For the quantile level θ of the response, the regression model can be written as

$$Q_\theta(y_i) = \beta_0(\theta) + \beta_1(\theta)x_{i1} + \dots + \beta_p(\theta)x_{ip}, \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, p \quad (5)$$

β_j can be estimated by QR procedure as follows

$$\widehat{\beta}_j(\theta) = \underset{\beta_j(\theta)}{\operatorname{argmin}} \sum_{i=1}^n \rho_\theta(y_i - \beta_0(\theta) - \sum_{j=1}^p \beta_j(\theta)x_{ij}) \quad (6)$$

where $\rho_\theta(u) = \theta \max(u, 0) + (1 - \theta)\max(-u, 0)$ is called as check loss/loss function/check function in different studies. Regression coefficients are often obtained for the certain quantile levels (median; $\theta = 0.5$, lower and upper quartiles; $\theta = 0.25$ and $\theta = 0.75$ respectively and deciles; $\theta = 0.1, 0.2, \dots, 0.9$), even if the values of the quantile level are between 0 and 1 ($\theta \in (0,1)$). In particular, for $\theta = 0.5$, the well-known the *median regression* model is obtained.

2.1 Differences Between Linear and Quantile Regression

An excellent comparison table is presented by Rodriguez and Yao (2017). Conditional mean $E[Y|X]$ is estimated in LR whereas conditional quantiles ($Q_\theta[Y|X]$) are estimated in QR. LR often requires distribution assumption whereas QR does not require any distribution assumption. In fact, the assumption of distribution within the LR is optional. However, the distribution assumption may be required for most of the parametric estimators in order to have important properties such as minimum variance and unbiased. The problem of heteroscedastic variance can be solved with the help of the monotone transformation ($h(\cdot)$) in QR. However, $E[Y|X]$ under transformation does not preserved. QR is a robust method whereas LR is sensitive to outliers.

3. Analysis of HBS Data with Quantile Regression

OLS regression is designed to estimate conditional mean models whereas QR is formed to estimate conditional quantile models. As a result, it has been stated by Fitzenberger et al. (2002) that QR offer a more comprehensive view. It is logical to analyze the data by sub-groups. In particular, modeling and interpreting of variables such as expenditures may require studies from different quantiles. Thus, different expenditure groups are taken into consideration. In this study, the average monthly expenditures of persons have been examined by using QR.

3.1 Data

Data have been taken from Turkey Statistical Institute (TURKSTAT). The data set consists of 12166 observations. Dependent variable is the average monthly expenditures (EXPEN), independent variables are size of house m^2 (SHOUS), household size (HHSIZ), average monthly disposable income (DINC), number of refrigerator (NREF), number of TV (NTV), number of computer (NCOM), number of mobile phone (NMP) and number of car (NCAR). The correlation matrix is presented in Table 1. When the

correlation coefficients are examined, it is seen that there is no high correlations between the dependent variable and the independent variables, except between HHSIZ and NMP. Estimates based on OLS and QR have been obtained for this data belonging to 2017.

Table 1. Correlation matrix for variables

	EXPEN	SHOUS	HHSIZ	DINC	NREF	NTV	NCOM	NMP	NCAR
EXPEN	1								
SHOUS	0.2904	1							
HHSIZ	0.1277	0.166	1						
DINC	0.4606	0.1936	-0.3233	1					
NREF	0.1307	0.1759	0.0408	0.1013	1				
NTV	0.3381	0.2466	0.0693	0.2343	0.1367	1			
NCOM	0.3814	0.1934	0.0696	0.2566	0.1055	0.322	1		
NMP	0.2906	0.2137	0.6008	-0.0956	0.1081	0.1979	0.286	1	
NCAR	0.4217	0.2347	0.1104	0.2201	0.1172	0.2348	0.2627	0.2511	1

3.2 Empirical Results

It can be observed that expenditures differ in certain quantiles. The structure of the data is examined with the fraction of the data presented in Figure 1. STATA and RStudio are used in the analyzes. Table 2 presents the results of LR and QR. Five different quantile levels ($\theta = 0.05, \theta = 0.25, \theta = 0.5, \theta = 0.75, \theta = 0.95$) have been analyzed to represent very low, low, moderate, high and very high expenditures, respectively. All explanatories, except the NREF, are significant in all models. Thus, NREF may be excluded from the model, the model is estimated without NREF. There is no change in the significance of the coefficients and minor changes between 1% and 1‰ were observed in the coefficients. These results are available upon request of the authors. Median regression ($\theta = 0.50$) analysis is presented in Table 3. Authors do not report other QR analysis tables here for the sake of brevity.

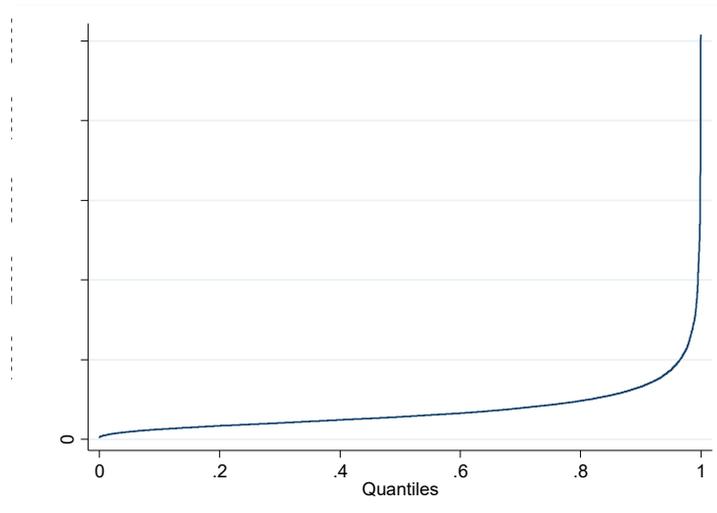


Figure 1. Fraction of the expenditure

Table 2. Results of LR and QR for different quantile levels

	OLS	QR_0.05	QR_0.25	QR_0.50	QR_0.75	QR_0.95
SHOUS	5.443*** (8.45)	1.877*** (4.18)	2.441*** (-6.56)	3.222*** (-7.2)	3.489*** (-4.21)	6.292** (-3.09)
HHSIZ	220.1*** (14.25)	68.56*** (6.36)	140.9*** (-15.77)	232.5*** (-21.65)	348.2*** (-17.5)	529.9*** (-10.84)
DINC	0.858*** (49.66)	0.288*** (23.95)	0.568*** (-56.85)	0.853*** (-71.03)	1.361*** (-61.22)	2.245*** (-41.11)
NREF	157.7 (1.09)	21.30 (0.21)	101 (-1.2)	41.42 (-0.41)	38.22 (-0.2)	-73.32 (-0.16)
NTV	477.9*** (13.49)	176.3*** (7.14)	230.7*** (-11.27)	253.7*** (-10.3)	278.2*** (-6.1)	603.0*** (-5.38)
NCOM	588.3*** (17.19)	345.2*** (14.47)	421.7*** (-21.32)	489.7*** (-20.58)	591.9*** (-13.43)	534.6*** (-4.94)
NMP	329.4*** (13.47)	192.7*** (11.30)	256.7*** (-18.16)	271.2*** (-15.94)	295.9*** (-9.39)	349.9*** (-4.52)
NCAR	1166.5*** (28.60)	293.4*** (10.32)	470.4*** (-19.96)	695.0*** (-24.51)	1207.9*** (-23)	2744.9*** (-21.28)
(Intercept)	-1108.2*** (-7.25)	-213.1* (-2.00)	-483.1*** (-5.46)	-643.6*** (-6.05)	-974.1*** (-4.95)	-1401.7** (-2.90)
N	12166	12166	12166	12166	12166	12166

t statistics in parentheses; * p<0.05, ** p<0.01, *** p<0.001

Coefficients for the median regression using STATA in Table 2 coincide with the median regression results using R in Table 3. The results presented in Table 2 and the comparisons based on this table are satisfactory. However, all tables can be provided upon request of the authors. Moreover, LR estimate is indicated by the red solid line; estimates of QR for different quantile levels are plotted as the darker dashed line in Figure 2. Graphs presented in Figure 2 provide a great summary for comparing LR estimates with the QR estimates obtained for different quantile levels.

Table 3. Results of QR for ($\theta = 0.5$)

	Coeff.	Std. Err.	t value	Pr(> t)
(Intercept)	-643.64487	34.84675	-18.47073	0.00000
SHOUS	3.22215	0.38389	8.39345	0.00000
HHSIZ	232.53972	9.47469	24.54327	0.00000
DINC	0.85288	0.02641	32.29953	0.00000
NREF	41.41722	34.11055	1.21421	0.22469
NTV	253.72132	21.75720	11.66149	0.00000
NCOM	489.73031	28.19202	17.37124	0.00000
NMP	271.15980	13.50700	20.07549	0.00000
NCAR	694.96860	31.33834	22.17630	0.00000

There is only one set of coefficients for OLS. For instance, for each additional SHOUS brings 5.44 more Turkish Lira (₺) in expenditures. 1.87 is SHOUS coefficient for QR at the 0.05 quantile level. This is interpreted as at the 5 quantile level for the expenditure, each additional number SHOUS brings ₺1.87 more in expenditures. QR estimates is much lower value than OLS estimates. It can be said that a lower value would be for the low quantiles. In contrast, for QR at the 0.95 quantile level, SHOUS coefficient is 6.29. At the 0.95 quantile level for the expenditure, each additional number SHOUS brings ₺6.29 more in

expenditures. Other explanatory can be interpreted in a similar manner. Instead of long interpretations, Figure 2 can be used for simple and understandable interpretations for the sake of brevity.

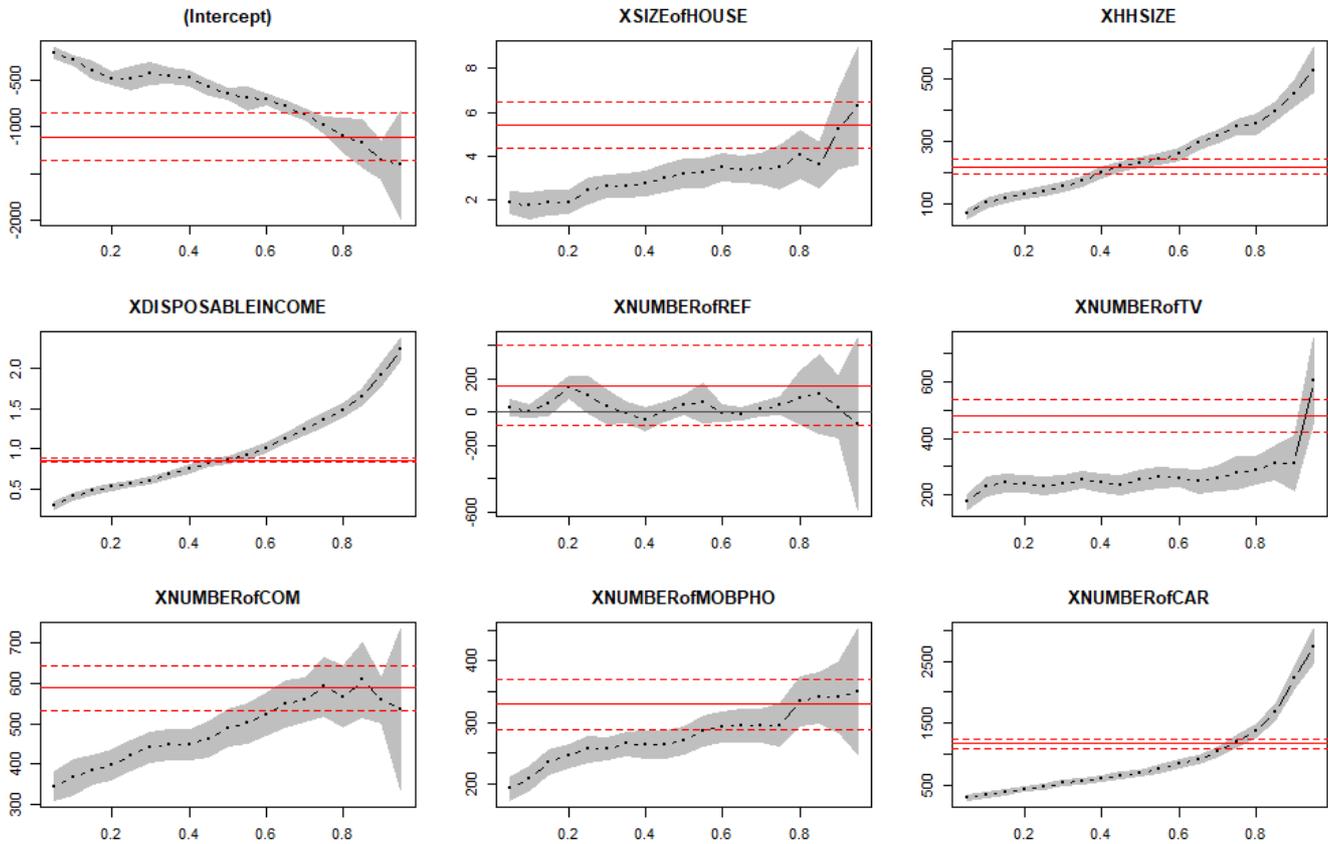


Figure 2. Estimates and confidence intervals

Comparison can be made using the confidence intervals (CIs) presented for the 5% level of significance. For instance, CIs of OLS and QR do not intersect until approximately 0.75 quantile level for SHOUS. So SHOUS coefficient is also significantly different than OLS until approximately 0.75 quantile level. For the HHSIZ independent variable, CIs of OLS and QR intersect between approximately 0.40 and 0.60 quantile levels. Therefore, in other cases, it can be said that the QR estimates differ from the OLS estimates. For DINC, the situation is similar to that of HHSIZ. For the NTV independent variable, there is an intersection with OLS estimates after the 0.90 quantile level. In this context, the importance of QR estimates is seen more clearly. However, when all of the different levels of QR coefficients and CIs are covered by OLS, the QR procedure, which requires more intensive processing than OLS, may be omitted. A comment is made similar to the comments made for variable HHSIZ and DINC within the NCAR variable. The only difference for NCAR is the level of quantile that intersects with CIs of OLS.

In addition to the difference between the coefficient estimates obtained from any quantile levels and the coefficient estimates obtained from the OLS procedure, estimates of QR can also be compared among themselves. Joint test of equality of slopes is conducted. Additionally, significant results are found for 5 models $F_{0.001;1,32} = 13.11 < 138.87$, pairwise binary comparisons are made. Other results are available upon request of the authors. QR tends to move out of OLS's CIs for heteroscedasticity, non-normal, skewed, and multimodal distribution situations. The findings obtained up to this stage is supported by Breusch-Pagan/Cook-Weisberg test results obtained as $\chi^2_{(0.001;8)} = 37.69 < 943.62$ for heteroscedasticity.

4. Conclusion

As a matter of fact, LR focuses on the mean. The inadequacy of estimations and interpretations based on the conditional *mean* of the response is a serious constraint for real life data. A process based on quantiles provides significant advantages for modeling real-life data. QR does not require assumptions for the distribution of residues. It also presents *different aspects* of the relationship between dependent and independent variables. Mosteller and Tukey (1977) are stated that regression based on mean presents incomplete picture. In addition, several different regression curves corresponding to the various percentage points of the distributions gives a more complete picture. Koenker, who has important researches in the literature on QR, states that QR is a comprehensive way to complete the regression picture. In this study, Turkey's HBS data is used and modeled with LR and QR. The results support the hypothesis that different coefficients can be obtained for different expenditure groups. For a comprehensive view, QR can be seen as an alternative. In the further studies, it is aimed to study the QR in the case of censored data.

Acknowledgment

This study was supported by Eskisehir Technical University Scientific Research Projects Commission under the grant no: 19ADP093. We would like to thank to TURKSTAT for providing the Household Budget Survey (HBS) data.

References

- Koenker, R., Bassett, G. (1978). “Regression Quantiles”, *Econometrica*, vol. 46, no. 1, pp.33-50.
- Koenker, R., Hallock, K. F. (2001). “Quantile Regression”, *Journal of Economic Perspectives*, Vol. 15, Number 4, pp. 143-156.
- Koenker, R., and Machado, J. (1999). “Goodness of fit and related inference processes for quantile regression”, *Journal of the American Statistical Association*, 94, 1296-1310.
- Sánchez, B.L., Lachos, H.V., Labra, V.F. (2013). “Likelihood Based Inference for Quantile Regression Using the Asymmetric Laplace Distribution”, *Journal of Statistical Computation and Simulation* 81, 1565–1578.
- Fitzenberger, B., Koenker, R. and Machado, J. (2002). “Economic Applications of Quantile Regression”, New York, NY: Physica-Verlag Heidelberg.
- Koenker, R. (2005). “Quantile Regression”, New York, NY: Cambridge University Press.

Galarza, C.E.; Lachos, V.H.; Barbosa, C.; Castro, L.M. (2017). “Robust quantile regression using a generalized class of skewed distributions”. *Stat* 2017, 6, 113–130.

Rahman, M. A., Vossmeier, A. (2019). “Estimation and applications of quantile regression for binary longitudinal data”, *Advanced in Econometrics*, Vol 40B.

Gómez, Y.M.; Gómez-Déniz, E.; Venegas, O.; Gallardo, D.I.; Gómez, H.W. (2019). An Asymmetric Bimodal Distribution with Application to Quantile Regression. *Symmetry*, 11, 899.

Mosteller, F., Tukey, J.. (1977). “*Data Analysis and Regression: A Second Course in Statistics*”. Reading, Mass.: Addison-Wesley.

Marrocu, E., Paci, R., Zara, A., (2015), Micro-economic determinants of tourist expenditure: A quantile regression approach, *Tourism and Management* Vol. 50, pp. 13-30.

Rodriguez, Robert N., and Yonggang Yao. (2017). *Five Things You Should Know about Quantile Regression*. Paper SAS525–2017. Cary: SAS Institute Inc

O-20 Statistical vs. Metaheuristic Techniques in Parametric Optimisation of Industrial Processes

Tatjana Sibalija^{1*}

¹*Faculty of Information Technology; Faculty of Management; Belgrade Metropolitan University; Serbia,*

tsibalija@gmail.com

Abstract – The parametric process optimisation aims to find a setting of process control factors that meets requirements for the response mean value and minimise variation simultaneously. The problem becomes more complex when a process is characterised by multiple responses, which are typically correlated. Therefore, the parametric optimisation methods are of the utmost importance for improving quality of industrial processes. The most frequently used statistical methods (Taguchi method and its modifications; response surface methodology), and non-conventional methods based on metaheuristic search algorithms (genetic algorithm; simulated annealing; particle swarm optimisation) are discussed in this paper. The implementation of both types of methods was critically appraised in terms of their peculiarities, benefits, shortcomings and applicability for certain type of problems.

Keywords – *process parameter optimisation; Taguchi method; response surface methodology (RSM); genetic algorithm (GA); simulated annealing (SA); particle swarm optimisation (PSO)*

1. Introduction

In industrial engineering, the process parameters design is one of the major tasks that highly affect the performance indicators, such as quality, cost and time to market. Taguchi identified three major types of factors that affect any process (Taguchi 1986): (i) Control factors (\mathbf{x}) are used to control the process. Each control factor can take multiple values, called levels. (ii) Signal factors (\mathbf{M}) have a very high and direct effect on the response. However, since signal factors cannot be easily identified for a vast majority of processes, they will not be considered in the further analysis. (iii) Noise factors (\mathbf{N}) negatively affect the process performance; they cannot be controlled or their control would be too costly or impractical. They cause the response y deviation from the target value that leads to a quality loss.

The robust parameter design aims to determine the optimal values of process control parameters in order to: (i) achieve the desired values of the responses, and (ii) decrease the noise factor effects that cause the process variability. The problem also includes constraints, typically the process parameters limits and other process condition constraints.

Due to an increased dynamicity at the global market, modern processes have become very complex, involving a large number of control factors and multiple outputs, highly non-linear and mainly unknown interdependencies.

2. Techniques for Parametric Process Optimisation

Various methods have been employed for the parametric process optimisation: conventional ones (based on the mathematical or statistical procedures), and non-conventional based on the artificial intelligence techniques. The methods based on the experimental design and on the metaheuristic algorithms have been

proven as the most effective and most applicable for a variety of processes, so they will be analysed further (Sibalija and Majstorovic, 2016).

2.1 Statistical Techniques for Parametric Process Optimisation

RSM includes the designed experimentation, typically using a full factorial. Based on the experimental data, the response surfaces are developed to map the process parameters vs. responses. This presents an input for optimisation, performed by hill climbing or descending tools. The original RSM addresses only the response mean, so the dual-response design is proposed where the response variation is added as a separate response (Myers and Montgomery, 2002). Despite a wide applicability, the following shortcomings were noticed: it gives indefinite saddle function in a quadratic model for processes with more than 3 responses; it could be easily trapped in a local optima for highly non-linear processes and processes with a large number of process parameters; simultaneous optimisation of both mean and variance would double the number of responses, which is inconvenient for a practical application.

The Taguchi’s robust parameter design relies on the designed experimentation based on the orthogonal arrays. The experimental analysis is based on the signal to noise ratio (SNR), presenting a ratio between an average response generated by control factors and variability caused by noise factors. The quality loss (QL) is formulated as a loss encountered by the user if the product response deviates from the desired value (Taguchi, 1986):

$$QL = K \cdot MSD = K \cdot \left\{ \begin{array}{ll} \frac{1}{n} \sum_{i=1}^n y_i^2 & \dots \text{ for } STB \\ \frac{1}{n} \sum_{i=1}^n (y_i - t)^2 = \frac{n-1}{n} s^2 + (\bar{y} - t)^2 & \dots \text{ for } NTB \\ \frac{1}{n} \sum_{i=1}^n \frac{1}{y_i^2} & \dots \text{ for } LTB \end{array} \right\} \quad (1)$$

where y is response; n is the sample size; \bar{y} is response mean in a sample; s^2 is sample variance; three type of responses are: nominal-the-best (NTB), larger-the-better (LTB), smaller-the-better (STB); QL is average loss per unit; MSD is mean square deviation in a sample; K is coefficient; t is target value for NTB response. Since the traditional method does not simultaneously address multiple responses, many approaches have been proposed in the last two decades:

- a. Weights are assigned to individual SNR or QL values to establish a single performance measure, but the weights assignment is performed subjectively. Besides, correlations among responses are not considered.
- b. Principal component analysis (PCA) is applied to transform SNRs or QLs into set of uncorrelated principal components (PCs), where only components with eigenvalue higher than 1 are considered (Su and Tong, 1997; Fung and Kang, 2005). Some approaches involved all PCs, but performed PCA on a raw response data (Liao 2006), which is misleading since the response specifications in the SNR analysis are not taken into account.
- c. Grey relational analysis (GRA) is applied over SNRs or QLs to establish a single performance measure, i.e. a grey relational grade, but assuming equal weights for all responses (Lung et al., 2007). To overcome this problem, Dubey & Yadava (2008) performed PCA on SNRs, and then applied GRA on PCs including all PCs.
- d. Desirability function approach (DFA) is applied on SNRs or QLs to develop a single performance measure. However, the correlations among responses are not considered. Moreover, the DFA-based

methods have been frequently criticised since they do not always provide an optimal solution and require a known process model.

Besides, the general limitation of the Taguchi-based, as well as DFA- and GRA- based approaches is that they consider only the discrete values of process parameters (control factors) used in the experimental trials..

2.2 Metaheuristic Techniques for Parametric Process Optimisation

Metaheuristics are inspired by the natural phenomena of survival of the fittest individuals. They are divided into two major groups: (i) point-to-point search (e.g. SA); (ii) population-based search (evolutionary algorithms – EAs, such as GA and PSO). GA, SA and PSO are the most frequently used metaheuristics in parametric process design.

GA starts with an initial population with n individuals (chromosomes). The objective, i.e. fitness function $f(x)$ of each chromosome is evaluated, to select the ones suitable for the next generation. The new population is formed using the following: scaling, selection of the parent chromosomes according to their scaled fitness values, crossover (combines two parents to form a child), and mutation (small random modifications of genes). Then, a migration is performed; the worst individuals are replaced with the best individuals from another subpopulation. Finally, the fitness of chromosomes in the new population is evaluated. The procedure continues until the termination condition is met (e.g. predefined number of iterations and/or no significant improvement in the objective function).

SA starts with an initial point at an initial temperature. New points are randomly produced in a proximity of the old one. Evaluation of the fitness function $f(x)$ is performed, and new points are accepted based on the defined probability of acceptance function: worse point could be accepted to decrease $f(x)$, but also better point could be accepted to extend a search. The initial temperature is a significant for the probability of acceptance; it has to be high enough to provide movement to any neighbourhood. An annealing schedule (temperature function) controls the temperature decrease to avoid local optimum in the beginning of search. Reannealing is performed after a certain number of accepted points to raise the temperature; search is resumed until the termination criterion is met (similar to GA).

PSO is a major algorithm of the swarm intelligence subgroup of EAs. An initial swarm is generated with positions x_i and velocities v_i . The objective function is evaluated to find the particle best position ($pbest$), the swarm best position ($gbest$) and the corresponding $f(pbest)$ and $f(gbest)$. The particle velocity is updated based on the previous velocity and the particle cognition and social components, which are controlled by the inertia weight and the self-adjustment and social adjustment learning factors (c_1 and c_2). The particle position is updated based on the previous one and the updated velocity. The objective function is evaluated: If $f(x_i) < f(pbest)$, then $pbest = x_i$; If $f(x_i) < f(gbest)$, then $f(gbest) = f(x_i)$ and $gbest = x_i$. The swarm is updated and procedure continues until the stopping condition is met.

3. Intelligent System for Multiresponse Robust Process Design (IS-MR-RPD)

This section proposes an advanced method for parametric process optimisation, based on the statistical and metaheuristic techniques. First, the statistical approach (the factor effects) is applied: responses are expressed using QL values and normalised to obtain NQLs; PCA is applied over NQLs to obtain set of uncorrelated PCs from the correlated NQLs; then GRA is applied on PCs to integrate them into a single process performance measure (grade relational grade γ), based on their weights ω from PCA. Therefore, a

process performance measure γ is obtained in a fully objective manner, taking into account all PCs to enclose the total variability. The effects of the control factors on γ are calculated, and the factor levels with the highest impact are selected as optimal (Sibaliija, 2020).

Artificial neural networks (ANNs) are used to map the process measure γ vs. process control factors, presenting an objective function for the metaheuristic algorithm that finds the optimal values of process control factors to maximise γ value. Two major algorithms are benchmarked: GA and SA (Sibaliija and Majstorovic, 2016).

The IS-MR-RPD application is depicted using the study performed to optimise the parameters of Nd:YAG laser drilling in processing Nimonic 263 sheets. Two process parameters, i.e. control factors (pulse frequency – f and duration - t) were varied on 5 levels; seven responses were observed (4 responses of LTB type; 2 responses of STB type; one NTB response). After performing experiment (Table 1), QL values for each response were calculated by formula (1), for the sample of 3 units. Application of PCA on NQLs resulted with the following independent PCs:

$$\begin{aligned}
 Y_1(k) &= 0.37 \cdot NQL_{Den_k} + 0.50 \cdot NQL_{Dex_k} - 0.28 \cdot NQL_{Cen_k} - 0.49 \cdot NQL_{Cex_k} - 0.47 \cdot NQL_{AR_k} + \\
 &0.26 \cdot NQL_{\theta_k} + 0.06 \cdot NQL_{Sa_k} \\
 &\dots \\
 Y_7(k) &= -0.07 \cdot NQL_{Den_k} + 0.75 \cdot NQL_{Dex_k} + 0.03 \cdot NQL_{Cen_k} + 0.01 \cdot NQL_{Cex_k} + 0.64 \cdot NQL_{AR_k} - \\
 &0.13 \cdot NQL_{\theta_k} + 0.03 \cdot NQL_{Sa_k}
 \end{aligned} \tag{2}$$

The common approach would involve only 91.2% of the total variance (first two PCs; eigenvalues >1). GRA was applied on the seven PCs using their weights from PCA ([0.554; 0.358; 0.065; 0.014; 0.006; 0.003; 0.001]), enclosing all PCs to compute the grey relational coefficient ξ and grey relation grade γ_k (Table 1):

$$\begin{aligned}
 \xi_i(k) &= \xi(z_0(k) \ z_i(k)) = \frac{\min_i \min_k |Z_i(k) - Z_0(k)| + \zeta \max_i \max_k |Z_i(k) - Z_0(k)|}{|Z_i(k) - Z_0(k)| + \zeta \max_i \max_k |Z_i(k) - Z_0(k)|} \\
 \gamma_k &= \gamma(z_0(k) \ z_i(k)) = \sum_{i=1}^p \omega_i \cdot \xi_i(k) = \sum_{i=1}^p \omega_i \cdot \xi(z_0(k) \ z_i(k)) \ ; \quad \sum_{i=1}^p \omega_i = 1
 \end{aligned} \tag{3}$$

where z_0 is ideal sequence, z_i is comparative sequence; ω_i is weight of the i th index.

Several ANNs with different number of neurones in the hidden layer were train to develop the process model. The ANN 2-15-1 resulted with the smallest mean square error ($MSE=1.28 \cdot 10^{-6}$) and the highest correlation between original data and the network output ($R=0.98$), so it was selected as the process model (Figure 1). Nine GAs and 36 SAs were run with different algorithm-specific parameters settings. As seen from Table 2, the best SA and the best GA found the same solution in almost the same number of iterations. Results of SAs are less disperse than of GAs, showing favourable robustness of SA. SA overperformed GA in this study, since robustness is an essential issue.

Table 1. Laser drilling optimisation: experimental plan, responses, NQLs, PCs and process performance measure

No	Factor levels	Response values	Normalised quality losses (NQLs)	Principal components $Y_j(k)$ $j=1, \dots, 7; k=1, \dots, 15$	γ_k $k=1, \dots, 15$
----	---------------	-----------------	----------------------------------	--	--------------------------------

f	t	Dent	Dex	Cen	Cex	AR	θ	Sa	NQL ₁	Y ₂	Y ₃	Y ₄	Y ₅	Y ₆	Y ₇									
		Dent	Dex	Cen	Cex	AR	θ	Sa	Dent	Dex	Cen	Cex	AR	θ	Sa	(k)	(k)	(k)	(k)	(k)	(k)	(k)	(k)	
1	1	5	433	255	0.94	0.88	2.77	7.2	0.02	0.28	0.71	0.98	1.00	1.00	0.92	0.11	-0.53	0.03	-0.14	0.66	-0.51	1.27	1.08	0.561
2	1	4	428	249	0.94	0.87	2.80	7.2	0.02	0.12	0.75	0.98	1.00	0.98	0.94	0.08	-0.56	0.10	-0.02	0.75	-0.47	1.23	1.11	0.514
...
1	3	1	394	216	0.94	0.95	3.05	7.2	0.06	1.00	1.00	0.97	0.85	0.83	0.93	0.63	0.06	-0.52	-0.32	0.43	-0.82	1.54	1.15	0.612
5																								

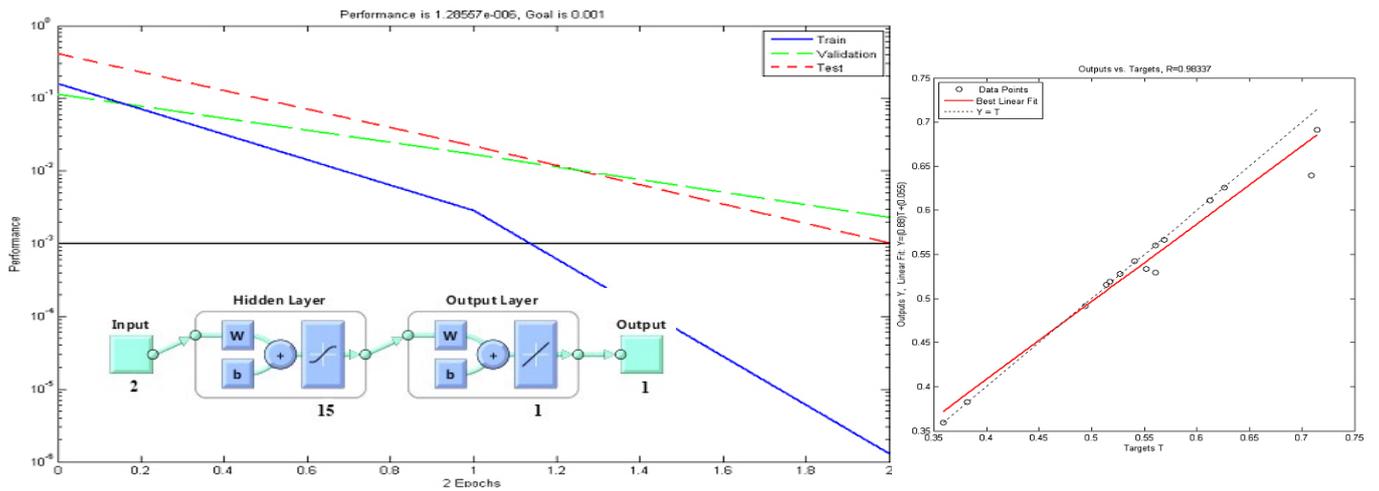


Figure 1. Laser drilling optimisation: characteristics, training and regression plot of the selected neural process model

Table 2. Laser drilling optimisation: summary of GA and SA results

Optimisation algorithm	GA	SA
Range of the objective function (process performance measure γ)	0.7133÷0.75230	0.7508÷0.75230
Range of the optimal process parameters (control factors)	[7.00÷7.50; 0.5]	[7.44÷7.57; 0.5÷0.502]
Maximal objective (process performance measure γ)	0.75230	0.75230
The optimal process parameters (control factors)	[7.5; 0.5]	[7.5; 0.5]
The iteration number to reach the maximal objective (convergence speed)	4	5

4. Statistical vs. Metaheuristic Techniques in Process Parametric Optimization

4.1 Comparison Based on the IS-MR-RPD Implementation

The above procedure was applied on 7 studies, and benchmarked with the most common approaches from the literature. From Table 3, the following conclusions could be drawn regarding the results of the statistical approaches:

- The proposed factor effects approach performed significantly better than the other tested approaches.
- RSM showed the worst results or no result in studies with a large number of responses.
- The Sung & Tong’s (1997) and Fung & Kang’s (2005) methods performed with varying results. They consider only PCs with eigenvalue > 1, thus including only a part of variability (e.g. study 4: 51.6% of the total variability).

- Liao’s approach (2006) enclosed the total variability, but the response requirements (e.g., NTB, STL, LTB) were not addressed at all. In overall, it performed worse than the above two approaches.
- The approach proposed by Dubey & Yadava (2008) enclosed all PCs but it was based on SNR values. It showed varying results, which demonstrates the importance of expressing responses via QLs rather than SNRs.
- Lung’s method (2007) did not show good results due to two issues: (i) assumption that all responses are of equal importance; (ii) response correlations are not addressed (GRA was applied on the response data directly).

Effectiveness of the proposed factor effects approach is well demonstrated in all studies, in dealing with a large number of responses and complex interrelations, and in dealing with row data from control charts (study 4).

From these six studies, the recommendations for the GA and SA settings are drawn (Sibalija and Majstorovic, 2016). SA overperformed GA in respect to the algorithm robustness and quality of the obtained solution; the convergence speed, in average, was equal. Since SA was better than GA, in the last study (study 7), PSO was benchmarked with SA, showing significantly better results than SA for all three criteria. PSO application on other six studies is ongoing.

4.2 Comparison Based on the Comprehensive Literature Review

A thorough literature review has been performed that included a few hundred of papers published in highly ranked scientific journals from the soft computing field and from the industrial and manufacturing field (Sibalija, 2018; Sibalija, 2019). It was focused on the comparison among the most frequent metaheuristic algorithms, also addressing their comparison with the statistical techniques. Metaheuristics were compared in terms of the quality of the obtained solution and the convergence speed. The analysis of the algorithm robustness was not reported in the cited papers.

The following conclusions were drawn regarding comparison between metaheuristics and statistical methods:

- Metaheuristics performed better than RSM in terms of the solution accuracy/quality (over 10 studies reported this comparison). SA was the most frequently benchmarked with RSM, and in all studies it showed better results.
- Metaheuristics performed better than the approaches based on the Taguchi method. GA and PSO were the most frequently benchmarked with the Taguchi method. However, this is not a “fair comparison” since the Taguchi method cannot search over the continual space of solutions, in contrast to metaheuristics.
- Metaheuristics demonstrated better results than the approached based on DFA and GRA.

Comparison among metaheuristic algorithms showed the following:

- On a larger sample (over 60 studies), it has been demonstrated that SA and PSO performed better than GA. PSO showed better convergence speed than GA in all studies. SA, in average, converged slightly faster than GA. PSO outperformed SA for both criteria. Therefore, PSO clearly outperformed SA and GA for both criteria.
- On a smaller sample (3 to 10 studies), it has been concluded that SA showed approximately equal performance as Hoopoe heuristic (HH) and harmony search (HS), and performance inferior to the other algorithms. PSO outperformed all metaheuristics (ant colony optimisation - ACO, artificial bee colony - ABC, HH, HS, scatter search - SS), except teaching-learning based algorithm (TLBO, that

found better solution and showed equal convergence speed as PSO), and cuckoo optimisation algorithm (COA, that was better than PSO for both criteria).

These results should be taken conservatively since the sample for comparison was small and none of the cited studies considered robustness of the algorithm (sensitivity to its own parameters tuning). The algorithm robustness is an essential issue, especially since the repeatability of the results of metaheuristic algorithms cannot be guaranteed.

Table 3. Summary of the comparative analysis of results obtained by different methods in seven case studies

Case study	Method	RSM	Su & Tong (1997)	Fung & Kang (2005)	Liao (2006)	Dubey &Yadava (2008)	Lung et al. (2007)	The factor effects approach	IS-MR-RPD with GA	IS-MR-RPD with SA
1: copper (50 μm) wire bonding–weld side (Sibalija & Majstorovic, 2009)	Process measure γ	0.7610	0.7371	0.6043	0.6043	/	/	0.7655	0.7673	0.7673
2: copper (50 μm) wire bonding – bond side (Sibalija et al., 2011a)	Process measure γ	0.8514	0.9286	0.9613	0.9286	/	/	0.9613	0.96870	0.97054
3: gold (70 μm) wire bonding (Sibalija & Majstorovic, 2010)	M1: Process measure γ	0.6303	0.6303	0.5637	0.5768	/	/	0.6395	0.88120	0.88120
	M2: Process measure γ	0.6379	0.5736	0.6379	0.5876	/	/	0.6463	0.71280	0.75801
4: enamelling process, using historical data (Sibalija et al., 2011b)	Process measure γ	no result	0.4368	0.7456	0.4598	/	/	0.7647	0.82114	0.87285
5: laser drilling of Nimonic 263 sheets (Sibalija et al., 2011c)	Process measure γ	0.5215	0.5514	0.5514	0.6125	0.514	0.514	0.7133	0.7523	0.7523
6: laser shock peening of Nimonic 263 sheets (Sibalija et al., 2014)	Process measure γ	no result	0.6034	0.6779	0.3766	0.5736	0.6779	0.8153	0.93052	0.93052
	Method	RSM	Su & Tong (1997)	Fung & Kang (2005)	Liao (2006)	Dubey &Yadava (2008)	Lung et al. (2007)	The factor effects approach	IS-MR-RPD with SA	IS-MR-RPD with PSO
7: laser cutting of Nimonic 263 sheets (Sibalija et al., 2019)	Process measure γ	0.8413	0.6861	0.6861	0.6856	0.8178	0.8178	0.8754	0.900762	0.900825

5. Conclusion

This paper presented a comprehensive comparison of the most frequently statistical and metaheuristic techniques is a parametric optimisation of industrial processes. For single response problems, both techniques have been widely and successfully applied. The advantage of RSM in comparison to Taguchi method is that it searches through a continuous space of solution. The recommendation for future RSM application is to always include process variability, by applying a dual response surface. From the other side, the advantage of Taguchi method is that it encloses both the process location and dispersion. Metaheuristic algorithms were successful in resolving single response problems, largely due to a search through a continuous space of solution. The recommendation for metaheuristics application refers to a proper algorithm-specific parameters tuning.

For multiresponse problems, the optimisation procedure and interpretation are more complex. RSM can successfully solve problems with 2 or 3 responses, but it showed inferior results for problems with a large number of responses. The traditional Taguchi method is unable to deal with multiple responses, so various approaches has been developed to address this issue. Their shortcoming were analysed above, and

the factor effects approach is proposed to overcome their deficiencies. The recommendations are: address the correlations among responses; include the total variability of original responses; express responses via QL rather than SNR, to adequately address the user preferences.

Regarding metaheuristics, the literature review revealed that the best results were achieved by COA and TLBO, followed by PSO that was significantly better than GA and SA (based on a large sample of studies), and also overperformed ACO, HS, SS, HH. Metaheuristics were very effective in resolving multiresponse problems. However, there are issues to be addressed (Sibaliija, 2019): (i) the algorithm robustness; (ii) allocation of weights for individual responses, if a single objective function is formed; (iii) trade-offs among multiple Pareto fronts, if the Pareto front method is used to deal with more than 3 responses; (iv) addressing both the process mean and variability.

References

- Dubey, A.K., Yadava, V. (2008). “Multi-objective optimization of Nd:YAG laser cutting of nickelbased superalloy sheet using orthogonal array with principal component analysis”, *Opt Lasers Eng*, vo.46, pp.124-132.
- Fung, C.P., Kang, P.C. (2005). “Multi-response optimization in friction properties of PBT composites using Taguchi method and principle component analysis”, *J Mater Process Technol*, vo.17, pp.602–610.
- Liao, H.C. (2006). “Multi-response optimization using weighted principal component”, *Int J Adv Manuf Technol*, vol.27, pp. 720–725.
- Lung, K.P., Che CW, Shien LW, Hai FS (2007). “Optimizing multiple quality characteristics via Taguchi method-based grey analysis”, *J Mater Process Technol*, vol.182, pp.107-116.
- Myers, R.H., Montgomery D.C. (2002). *Response surface methodology: process and product optimization using designed experiments*. Wiley, New York, US.
- Sibaliija, T., Majstorovic, V., (2009). “Multi-response Optimisation of Thermosonic Copper Wire Bonding Process with correlated responses”, *Int J Adv Manuf Technol*, vol. 42, pp. 363-371, doi: 10.1007/s00170-008-1595-1.
- Sibaliija, T., Majstorovic, V. (2010). “Novel Approach to Multi-Response Optimisation for Correlated Responses”, *FME Transactions*, vol. 38, pp. 39-48.
- Sibaliija, T., Majstorovic, V., Miljkovic, Z. (2011a). “An intelligent approach to robust multi-response process design”, *Int J Prod Res*, vol. 49, pp. 5079-5097, doi:10.1080/00207543.2010.511476.
- Sibaliija, T., Majstorovic, V., Sokovic, M. (2011b). “Taguchi-Based and Intelligent Optimisation of a MultiResponse Process Using Historical Data”, *Strojniški vestnik - Journal of Mechanical Engineering*, vol. 57, pp. 357-365, doi:10.5545/sv-jme.2010.061.
- Sibaliija, T., Petronic, S., Majstorovic, V., Prokic-Cvetkovic, R., Milosavljevic, A. (2011c). “Multi-response design of Nd:YAG laser drilling of Ni-based superalloy sheets using Taguchi’s quality loss function, multivariate statistical methods and artificial intelligence”, *Int J Adv Manuf Technol*, vol. 54, pp. 537–552, doi:10.1007/s00170-010-2945-3.
- Sibaliija, T., Petronic, S., Majstorovic, V., Milosavljevic, A. (2014). “Modelling and optimisation of laser shock peening using an integrated simulated annealing-based method”, *Int J Adv Manuf Technol*, vol. 73, pp.1141-1158.
- Sibaliija T., Majstorovic V. (2016). *The advanced multiresponse process optimisation. An intelligent and integrated approach*, doi:10.1007/978-3-319-19255-0, Springer, Switzerland.

- Sibaliija, T., (2018). “Application of Simulated Annealing in Process Optimization: A Review”, in Alex Scollen and Thomas Hargraves (Eds.): *Simulated Annealing: Introduction, Applications and Theory*, pp. 1-48, Nova Science Publishers, US.
- Sibaliija, T., Petronic, S., Milovanovic, D. (2019). “Experimental Optimization of Nimonic 263 Laser Cutting Using a Particle Swarm Approach”, *Metals*, vol. 9, doi: 10.3390/met9111147.
- Sibaliija, T., (2019). “Particle Swarm Optimisation in Designing Parameters of Manufacturing Processes: a Review (2008-2018)”, *Applied Soft Computing*, vol. 84, pp.105743, doi:10.1016/j.asoc.2019.105743.
- Sibaliija, T., (2020). “The Quality Loss Function-Based Approach for Discrete Multiresponse Process Parameters Optimization”, in Tatjana Sibaliija (Ed.): *A Closer Look at Loss Function*, Nova Science Publishers, US – *expected on 02/2020*.
- Su, C.T., Tong, L.I. (1997). “Multi-response robust design by principal component analysis”, *Total Qual Manag*, vol.8, pp. 409–416.
- Taguchi, G., (1986). *Introduction to quality engineering*. Asian Productivity Organization, UNIPUB, New York, US.

O-21 A Panel Data Analysis: The Relationship Between Unemployment, Youth Unemployment and Economic Growth

Merve ALTAYLAR^{1*} and Hamdi EMEÇ²

¹*Econometrics, Social Sciences Institute, Dokuz Eylül University, Turkey, mervealtaylar37@gmail.com*

² *Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylül University, Turkey, hamdi.emec@deu.edu.tr*

Abstract *This paper examines the relationship between unemployment, youth unemployment and economic growth in 20 OECD countries between 2000 and 2017. These variables are examined using static and dynamic panel time series techniques. He does not examine the relationship between these variables, as Okun suggested, and explores how unemployment and youth unemployment affect economic growth. Conventional panel unit root and panel break root and panel cointegration tests were used to investigate the differences of these variables on economic growth. As a result, youth unemployment is more susceptible to economic growth than unemployment, while developments in youth unemployment do not affect economic growth as well as unemployment.*

Keywords – *Multiple Structural Break Panel Cointegration, Heterogeneous Panel, Cross-sectional Dependence, DOLS, PANKPSS*

1. Introduction

Unemployment, which has become a global problem, has far more serious effects, especially for some groups in terms of causes and consequences. Young people are the leading groups. Thus, the phenomenon of youth unemployment is now becoming accepted among the global problems in the economic literature. When the unemployment indicators of developed or developing countries are analyzed, it is seen that youth unemployment rates are at a much higher level compared to the total unemployment rates. If moderate economic growth is achieved in order to reduce unemployment, the overall level of unemployment will decrease in relation to the increase in economic activity level, thus providing a constructive impact on youth unemployment. Okun's Law (1962), which is accepted as one of the basic laws in economics, analyzes the relationship between unemployment and economic growth. Okun's Law (1962) suggests that there is an inverse relationship between macroeconomic aggregates based on the model.

The aim of this study is to examine the relationship between unemployment and youth unemployment and economic growth.

For this purpose, the macroeconomic relationship between unemployment, youth unemployment and GDP variables was examined by panel data analysis methods with the data of 24 OECD member countries between 2000-2017.

2. Materials and Methods

In this study, three macroeconomic indicators such as GDP, unemployment rate and youth unemployment rate were analyzed and the scope of the application was decided in 24 OECD¹ countries. Thus, a panel data set with cross-section and time dimension was formed. The relevant variables were included in the analysis between 2000 and 2017 and quarterly and the data were collected from the OECD database.

Detailed information on macroeconomic variables is shown in Table 1 below.

Table 1. Variables

Variables	Series	Data Type	Period
Unemployment Rate	Seasonally Adjusted	Level Rate	2000 Q1-2017 Q4
Youth Unemployment Rate	Seasonally Adjusted	Level Rate	2000 Q1-2017 Q4
GDP	Seasonally Adjusted	Currency \$ Dollar	2000 Q1-2017 Q4

Table 2 shows the descriptive statistics of the three variables subject to analysis (Gross Domestic Product, unemployment rate and youth unemployment rate). According to these indicators, the panel data set consists of 24 sections and 72 time dimensions for all three macroeconomic variables, showing a balanced and long panel ($T > N$).

Table 2. Descriptive Statistics of Macroeconomic Variables

Variables	Obs. Per Group (N)	Time (T)	Observation	Mean
GDP	24	72	1724	1213162
Unemployment Rate	24	72	1724	8.01
Youth Unemployment Rate	24	72	1724	18.07

In this study, whether the macroeconomic variables are stationary or not is analyzed by methods appropriate to the structure of the panel time series in order to avoid spurious regression problems. However, since the analysis was carried out with panel time series, the concept of cross-sectional dependence specific to this structure gained priority. In this way, cross-sectional dependence was tested with Pesaran CD test for three macroeconomic variables and the results showed that all three variables had cross-sectional dependency problems. Thus, macroeconomic variables were subjected to the Pesaran CIPS and PANKPSS panel stationary tests, which are the second generation panel unit root tests which have the assumption of cross-sectional dependence. According to the common results of panel unit root and stationarity analysis tests performed for the whole panel, it was found that all three variables were not stationary at the level. When the first differences of these variables were taken and the same process was repeated, the variables became stationary. After the panel unit root and stationary analysis, it was examined whether the Unemployment Rate-GDP model is cointegrated. In the continuation of the analysis, panel cointegration analyzes were performed for the second model, the GDP-Unemployment Rate model, the third model for the Youth Unemployment Rate-GDP model, and finally for the GDP-Youth Unemployment Rate model. However, in this study, because of the analysis of panel time series,

¹ There are 36 member countries of the OECD (OECD, www.oecd.org, 09.06.2019). Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, Greece, Hungary, Ireland, Italy, South Korea, Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Spain, England, USA, Estonia, Israel and Slovenia. Due to limitations in the data sets Chile, France, Germany, Iceland, Japan, Latvia, Lithuania, Luxembourg, Mexico, Sweden, Switzerland and Turkey were excluded from the study. Among these countries, Hungary and Poland have the status of developing countries while other countries have the status of developed countries.

the concept of parameter heterogeneity has gained importance in addition to the cross-sectional dependence specific to this structure. For the four models examined, cross-sectional dependence was determined by Breusch Pagan LM test; parameter homogeneity was tested with Swamy S test. The results revealed cross-sectional dependence and parameter heterogeneity in all four models. The cointegration relations in these models were investigated by using Westerlund (2009) panel cointegration test with Multiple Structural Breaks, which has the assumption of cross-sectional dependence, parameter heterogeneity and structural break. Thus, in order to estimate long-term relationships, the panel cointegration model estimator DOLS, which is the assumption of cross-sectional dependence and parameter heterogeneity, was preferred, and the model without cointegration relation was estimated with the AMG estimator. Stata 14 and Gauss 10 software were used during the analyzes.

2.1 Testing the Cross-sectional Dependency

The cross-sectional dependence, defined as the interaction between the units that make up the panel (eg households, firms, countries, etc.), can be regarded as the equivalent of the series correlation in time series. This can occur in behavioral interactions between individuals, consumers in a community, or firms working in the same sector. It can also be caused by unobservable common factors or common shocks that are very common in macroeconomics. As with the correlation problem observed in time series, cross-sectional dependence leads to loss of productivity in the OLS estimator and may result in the invalidation of traditional t tests and F tests using standard variance covariance estimators, and in some cases even inaccuracy. For this reason, it is emphasized that it is wise to test cross-sectional dependence before starting the analysis (Baltagi and Kao, 2012: 137). In this section, Pesaran CD test is introduced to investigate cross-sectional dependence.

Table 3. Pesaran CD Test Results

Variables	CD-Test Stat.	Correlation Coefficient	Prob.
Unemployment Rate	32.65	0.232	0.000*
Youth Unemployment Rate	34.76	0.247	0.000*
GDP	103.62	0.735	0.000*

Table 3 shows the results of the Pesaran CD Test for measuring cross-sectional dependence for all variables examined.

According to these results, the null hypothesis, which states that there is no dependency between units' errors in all three variables at 5% significance level, was rejected and it was found that there was a cross-sectional dependency problem in the panel. Cross-sectional dependence is usually; country, region, city etc. This problem is expected to be encountered when working with units. The mean correlation coefficients of the variables were 23% for the unemployment series, 24% for the youth unemployment series and 73% for the GDP series. Therefore, first generation panel unit root tests are inadequate in analyzing these variables since they do not consider the cross-sectional dependence problem. Therefore, in this study, second generation panel unit root tests which take the cross-sectional dependence problem into consideration were preferred.

2.2 Panel Unit Root and Stationary Tests

Unit root presence testing has become a common practice among researchers in time series analysis. However, it is stated that unit root research in panel datasets is more recent in the literature than time series. In the literature, panel unit root tests are divided into two groups according to cross-sectional dependence. The first generation panel unit root tests work with the assumption that there is no cross-sectional dependence, while the second generation panel unit root tests work with the presence of cross-sectional dependence (Chen and Lu, 2003:343).

The assumption that there is no cross-sectional dependence when analyzing with panel data is seen as a very strict restriction in applied research. For this reason, second generation panel unit root tests were developed considering cross-sectional dependence (Levin, Lin and Chu, 2002: 14).

CIPS panel unit root test developed by Pesaran (2007) is based on the logic of modeling cross-sectional dependence through factors. In his study, which uses the horizontal cross-sectional mean of the individual series forming the sections as a tool variable for the factors not observed in the model, he suggested that this approach eliminates the cross-sectional dependence. It is suggested that ADF regression is extended with the horizontal cross-sectional mean and delayed values of the series and when this first regression difference is taken, it eliminates the cross-sectional dependence. When traditional unit root tests are used, it can be concluded that the series is not stationary if there is a structural break in the series. For this case, Lee and Strazicich (2003) found that the element that disrupts the stationary of the time series is in fact caused by structural breaks, thus concluding that the time series is actually stationary. To solve this problem, Carrion-i Silvestre et al. (2005) developed the PANKPSS test. This test has the null hypothesis that the panel is stationary and has a test statistic that allows for the presence of multiple structural breaks, taking into account the problem of cross-sectional dependence. Two different characteristics are considered depending on the fixed term and / or structural breaks that occur in the trend. The statistical model is flexible enough to allow the number of fractures and times to differ between sections (Carrion-i-Silvestre vd., 2005: 162). Table 4 shows the results of panel unit root and stationarity analysis for all three macroeconomic variables.

Table 4. Panel Unit Root and Stationary Tests

Unemployment Rate					GDP					Youth Unemployment Rate				
Equation	Constant		Constant and Trend		Equation	Constant		Constant and Trend		Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.	Unit Root Tests	Stat.	Prob.	Stat.	Prob.	Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	4.295	1.0000	5.385	1.0000	CIPS	2.500	0.994	2.723	0.9970	CIPS	2.790	0.997	2.685	0.9960
PANKPSS	-	-	4.251	0.000	PANKPSS	-	-	14.243	0.000	PANKPSS	-	-	11.244	0.000
Δ Unemployment Rate					Δ GDP					Δ Youth Unemployment Rate				
Equation	Constant		Constant and Trend		Equation	Constant		Constant and Trend		Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.	Unit Root Tests	Stat.	Prob.	Stat.	Prob.	Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	-13.697	0.0000	-13.697	0.0000	CIPS	-18.089	0.000	-17.167	0.0000	CIPS	-19.165	0.0000	-18.802	0.0000
PANKPSS	-	-	-0.512	0.697	PANKPSS	-	-	-0.503	0.692	PANKPSS	-	-	-0.212	0.584

Table 4 shows the CIPS and PANKPSS test results of the unemployment variable. According to these results, it was found that the unemployment variable was not stationary at the level. However, when the difference of unemployment variable from the first order is taken, it has been found that this variable becomes stationary. Table 4 (part two) shows the CIPS and PANKPSS test results of the GDP. According to these results, it was found that the GDP variable was not stationary at the level. However, when the difference of GDP from the first order is taken, it has been found that this variable becomes stationary. Table 4(part three) shows the CIPS and PANKPSS test results of the youth unemployment variable. According to these results, it was found that the unemployment variable was not stationary at the level. However, when the difference of youth unemployment variable from the first order is taken, it has been

found that this variable becomes stationary. In this case, it is seen that all three variables are not stationary at the level, but the first difference of these variables is stationary.

2.3 Testing the Parameter Heterogeneity

Another important concept in panel data is parameter heterogeneity. When working with panel data, first of all, it must be tested whether the regression parameters of the sections forming the panel are specific to the relevant section. If there is heterogeneity in the regression parameters, tests and / or estimation methods that take this heterogeneity into consideration should be preferred. Otherwise, choosing prediction methods that ignore such heterogeneity may lead to inconsistent or meaningless estimation of parameters (Tatoğlu, 2018: 51).

In this study, Swamy S test was used to test parameter heterogeneity. The first studies for testing homogeneity in panel data analysis studies were made by Swamy (1970). In Swamy's S test, the null hypothesis states that the parameters of the sample units examined are homogeneous (Tatoğlu, 2017: 247).

Table 5 shows the results of the parameter heterogeneity test.

Table 5. Swamy S Test for Parameter Heterogeneity

Models	Parameter Heterogeneity Test
Model 1 (Unemployment Rate-GDP)	Chi-Square Stat. 28578.56
	Chi-Square Prob. 0.0000*
Model 2 (GDP-Unemployment Rate)	Chi-Square Stat. 9.300
	Chi-Square Prob. 0.0000*
Model 3 (Youth Unemployment Rate-GDP)	Chi-Square Stat. 25197.41
	Chi-Square Prob. 0.0000*
Model 4 (GDP-Youth Unemployment Rate)	Chi-Square Stat. 7.100
	Chi-Square Prob. 0.0000*

According to the test results, the null hypothesis that the parameters of 5% significance level was homogeneous was rejected. Parameter heterogeneity is observed in these models. Therefore, second generation panel cointegration tests that take into account parameter heterogeneity should be selected.

2.4 Panel Cointegration Test With Sutstructural Breaks

Panel cointegration test with structural break developed by Westerlund (2009) considers cross-sectional dependence, parameter heterogeneity and multiple structural breaks. This test allows up to three structural breaks in the model. The test also includes a constant or a constant and a break in trend. Table 6 shows the results of the Westerlund Panel Cointegration Test with Multiple Structural Breaks for the four models examined.

Table 6. Westerlund (2009) Panel Cointegration Test with Multiple Structural Breaks

Models	Coef.	Bootstrap Prob.
Model 1 (Unemployment Rate-GDP)	1.412	0.004*
Model 2 (GDP-Unemployment Rate)	10.553	0.180
Model 3 (Youth Unemployment Rate-GDP)	6.501	0.840
Model 4 (GDP-Youth Unemployment Rate)	7.405	0.880

According to the results shown in Table 6, the null hypothesis stating that there is a cointegration relationship at 5% significance level is rejected. Thus, considering the breaks, there is no evidence of the cointegration relationship in the Unemployment Rate-GDP model. The null hypothesis, which states that there is a cointegration relationship at 5% significance level according to the calculated probability for the second model (GDP-Unemployment Rate), could not be rejected. In this case, there is a cointegration relationship between the two variables and it is necessary to estimate the cointegration model in order not to lose this long-term synchronous relationship structure. According to the results of cointegration analysis for the third and fourth models, the null hypothesis stating that there is a cointegration relationship at 5% significance level could not be rejected. Thus, a long-term relationship was found between the variables of the third and fourth models. Therefore, three of these models (2,3,4) should be estimated within the framework of panel cointegration model.

2.5 Dynamic Ordinary Least Squares (DOLS) Estimator

In the literature, there are many estimation methods used for estimating the panel cointegration model for the variables that have been determined to have cointegration relations between them. Dynamic Least Squares (DOLS) estimator is among the estimators that can be preferred in the presence of parameter heterogeneity and cross-sectional dependence in estimating the long-term relationship between the cointegrating variables. In this estimator, firstly, the cointegration model is estimated for each section and in the next step these estimation results are combined for the whole panel with the Pesaran and Smith Mean Group (MG) approach.

Table 7. Long Term Parameters (DOLS Estimator)

Models	Coef.	t Stat.
Model 2 (GDP-Unemployment Rate)	-0.168	-31.34*
Model 3 (Youth Unemployment Rate-GDP)	-2.281	-29.32*
Model 4 (GDP-Youth Unemployment Rate)	-0.165	-28.38*

For the GDP-Unemployment Rate model (2), the t-statistic calculated at 5% significance level is greater than the critical value, so that the long-term parameter is statistically significant. According to this result, when unemployment increased by 1%, GDP displayed an average decrease of approximately 0.17% in the whole panel. Therefore, it can be said that the increase in unemployment causes a decrease in GDP for the examined countries.

According to the results of the second model, long-term coefficient was found to be statistically significant and there was an inverse relationship between youth unemployment and GDP. This situation is in harmony with the economic literature. 1% increase in GDP variable reduces youth unemployment by 2.28% on average.

According to the results of the panel-based cointegration model, a statistically significant negative correlation was found between GDP and Youth Unemployment Rate variables. The 1% increase in youth unemployment has a 0.16% reduction in GDP.

2.6 Augmented Mean Group (AMG) Estimator

There are many estimators in the literature in order to estimate panel data models. This estimator, developed by Eberhardt and Teal (2010) and Bond and Eberhardt (2009) in this section, is used especially for estimating panel time series. It is resistant to cross-sectional dependence, parameter homogeneity and non-stationary variables. At this stage, the first model, which was not found to be cointegration regression, was estimated with the AMG estimator. Table 8 shows the panel-based regression results.

Table 8. Regression Coefficients (AMG Estimator)

Model 1 (Unemployment Rate-GDP)		
Variables	Coef.	Prob.
Slope	-1.062	0.001**
Common Factor	1.069	0.000*
Chi-Square	-	0.0009*

According to the results shown in Table 8, the model (chi-square probability) is a statistically significant and reliable regression. According to the estimation results, both variables had a statistically significant effect on the dependent variable. The 1% increase in GDP has a 1.06% reduction in unemployment rate.

3. Conclusion

The four models examined in this study were tested with the Westerlund (2009) panel cointegration test with Multiple Structural Breaks, which investigated the cointegration relationship under the assumption of cross-sectional dependence, parameter heterogeneity and structural breaks. With this test, the cointegration relationship could not be reached only in the Unemployment Rate-GDP model of the four models examined, and the cointegration relationship was found in the other three models. As a result of the analyzes, 1% percentage point increase in GDP decreased the unemployment rate by 1.06%, while %1 percentage point increase in GDP reduced the youth unemployment rate by 2.29%. Thus, it was found that youth unemployment is almost two and a half times more sensitive to economic growth than unemployment. At the same time, 1% increase in the unemployment rate reduces GDP by 0.17%, while 1% increase in youth unemployment reduces GDP by 0.16%. The conclusion from these rates is that youth unemployment is a very fragile structure and if it is reduced by countries with the right policies, it will affect economic growth as much as unemployment.

References

Baltagi, B. H. , Feng, Q. and Kao, C. (2012). “A Lagrange Multiplier Test for Cross-Sectional Dependence in a Fixed Effects Panel Data Model”, *Journal of Econometrics*, vol. 1, pp. 164-177.

Carrion-i-Silvestre, J. , Castro, T. and López-Bazo, E. (2005). “Breaking The Panels: An Application to the GDP Per Capita”, *The Econometrics Journal*, vol.2, pp.159-175.

Chen, J. ve Lu, W. (2003). “Panel Unit Root Tests of Firm Size and Its Growth.*Applied Economics Letters*”, vol.10, pp.343-345.

Levin, A. , C.F. Lin and C-S.J. Chu (2002). “Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties”, *Journal of Econometrics*, vol.108 pp.1–24.

Tatođlu, F. Y. (2017). “Panel Zaman Serileri Analizi: Stata Uygulamalı”, İstanbul: Beta

Tatođlu, F.Y. (2018). “Avrupa Ülkelerinde Okun Yasasının Çok Boyutlu Panel Veri Modelleri İle Analizi”, *Yönetim ve Çalışma Dergisi*, vol.1, pp. 43-56.

Westerlund, J. and Basher, S. (2009). “Panel Cointegration and The Monetary Exchange Rate Model. *Economic Modelling*”, vol.26, pp.506-513.

O-22 Simpson’s Paradox: Literature Review and a Dataset for the Treatment of Acne Rosacea Patients in the Muğla Region

Burcu DURMUŞ^{1*}, Öznur İŞÇİ GÜNERİ² and Aslı AKIN BELLİ³

¹Department of Statistics, Muğla Sıtkı Koçman University, Turkey, burcudurmus@mu.edu.tr

²Department of Statistics, Muğla Sıtkı Koçman University, Turkey, oznur.isci@mu.edu.tr

³Department of Dermatology, Erdem Hospital, İstanbul, dr_asliakin@hotmail.com

Abstract – When the data set is divided into groups and evaluated as a total in one study, dilemmas may be seen in the results. This situation is called as Simpson Paradox or Reversal Paradox. In this study, a literature search for Simpson Paradox and analysis for a new data set were performed. For this purpose, four studies in the literature were examined. Three of these studies are based on real-life data sets. The other data set is the “iris dataset” used in many analyzes in the literature. In addition to the literature, a data set from acne rosacea patients in the Muğla region was created. In this dataset, the location of the disease and the treatment methods applied were examined. As a result of the study, it was determined that acne rosacea disease showed Simpson paradox and the data set was included in the study. Besides the data sets, the mathematical definition of paradox is also mentioned in the study. At the end of the study, the situations in which the Simpson paradox is emerged and how to solve it were explained and its importance was emphasized.

Keywords – Acne Rosacea, Different Treatment, Yule-Simpson Effect, Simpson’s Paradox.

1. Introduction

There are the situations that a question has no definite answer or solution. These questions may seem as contradicting each other, be meaningless, or in the vicious cycle. The paradox emerges at these points. Mathematically, many paradoxes are available. Simpson's Paradox or Reversal Paradox, one of the paradoxes, is the situation that conflicting interpretations arise when the data set is divided into different groups. The Paradox was first introduced by Yule in 1903. It was written for the first time in 1951 by Simpson. Therefore, it is also known as the Yule-Simpson Effect (Kock, 2015; Chambaz et al., 2017). The effects of the paradox are largely observed in multivariate statistical analyzes. Mathematical Definition of Simpson’s Paradox;

Let's handle 3 events as A, B, and C randomly. Consider A', B' and C' as complementary events. When events B and C are dependent,

$$\begin{aligned} P(A|B\&C) &> P(A|B'\&C) \\ P(A|B\&C') &> P(A|B'\&C') \end{aligned} \quad (1)$$

as shaped. But if

$$P(A|B) \leq P(A|B') \quad (2)$$

it clearly demonstrates how equation 1 and equation 2 paradoxes emerged in random events (Fitelson, 2017; Maa, 2015). An arithmetical way to represent the paradox is as in equation 3:

$$\frac{a}{b} < \frac{A}{B}, \quad \frac{c}{d} < \frac{C}{D} \quad \text{and} \quad (3)$$

$$\frac{a+c}{b+d} < \frac{A+C}{B+D}$$

An arithmetic example is given by equation 4 (Zalta et. Al., 2016);

$$\frac{1}{5} < \frac{2}{8}, \quad \frac{6}{8} < \frac{4}{5} \quad \text{and} \quad (4)$$

$$\frac{7}{13} > \frac{6}{13}$$

as obtained. Table 1 provides a sample table for two groups.

Table 1: 2x2 Simpson’s Paradox

	Group 1	Group 2
Central	A	B
All	C	D
Both	A+C	B+D

The Simpson Paradox is often encountered in real life. In this study, how Simpson Paradox is seen in regional disease treatments is discussed. The study reminds us that the data available is not always complete and it emphasizes of better analysis of data acquisition processes.

2. Materials and Methods

The study focused on the areas where the Simpson Paradox is seen, its mathematical meaning, the examples in the literature, and how it is observed for a real data set (Acne Rosacea Patients).

2.1 Simpson's Paradox in Literature

Wagner (1982), in his study of the Simpson Paradox, discussed two examples in which paradox emerged. In the sample of “Income Tax Rates”, different income groups were examined for 1974 and 1978. Information about the study is given in Table 2. When the table is examined, it can be said that the change of tax rates in income groups by years creates a paradox.

Taylor and Mickel (2014) considered that there is a discrimination in the state of California. Therefore, they examined the sources of funding in the region. In the study, they performed one and two variable analyzes. In their first analysis, it was revealed that there was a discrimination. However, the second detailed examination revealed that there was no discrimination and the results included only Simpson's paradox. Table 3 presents the results of the paradox. The study shows great importance in eliminating discrimination turmoils in societies.

Table 2. Results of Income Tax Rates

Adjusted Gross Income	1974			1978		
	Income	Tax	Tax Rate	Income	Tax	Tax Rate
Under \$ 5.000	41.651.643	2.244.467	0.054	19.879.622	689.318	0.035
5,000 to \$ 9.999	146.400.740	13.646.348	0.093	122.853.315	8.819.461	0.072
10,000 to \$ 14.999	192.688.922	21.449.597	0.111	171.858.024	17.155.758	0.100
15,000 to \$ 99.999	470.010.790	75.038.230	0.160	865.037.814	137.860.951	0.159
100,000 or more	29.427.152	11.311.672	0.384	62.806.159	24.051.698	0.383
Total	880.179.247	123.690.314		1.242.434.934	188.577.186	
Overall Tax Rate			0.141			0.152

Table 3. Average Expenditures for California

Age	Hispanic	White non-Hispanic
0-5	\$ 1.393	\$ 1.367
6-12	\$ 2.312	\$ 2.052
13-17	\$ 3.955	\$ 3.904
18-21	\$ 9.960	\$ 10.133
22-50	\$ 40.924	\$ 40.188
51+	\$ 55.585	\$ 52.670
Consumers	\$ 11.066	\$ 24.698

Maa (2015) also presented a Simpson's paradox between GDP (Gross Domestic Product) per capita and overall GDP in a study.

In another study, Xu et. al. (2018) tried to understand Simpson's Paradox on 3 different data sets. One of the data sets discussed is the Iris data set, which is very common in the literature. The results for the iris data set are given in Figure 1. It is seen that there is Simpson paradox between species and general trends.

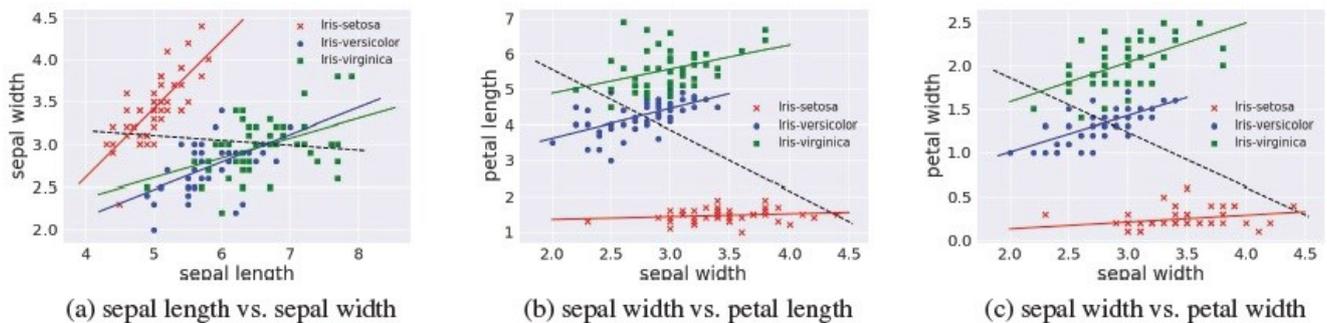


Figure 1. Results of Iris

Cankar (2018) carried out examinations in two classes (class A and class B) and two categories (girls and boys) to determine the mathematical ability of the students who took the Pisa exam. It was seen that boys in class A were more successful than girls and boys in class B were more successful than girls. However,

when the total value is examined, it was concluded that girls were more successful than boys by 2 points. The results of the hypothesis are given in Table 4.

Table 4. Results of Pisa

	Boys	Girls	Difference
Class A	23.6	20.3	-3.3
Class B	13.7	10.4	-3.3
Total	16.0	18.0	+2.0

As can be seen from Table 4, the Simpson’s Paradox appears when the data set is divided into different groups. In order to answer the questions of the problem, firstly, how the data will be handled must be determined. Separately or total? Combining data can cause other problems. Therefore, the process that constitutes the causal model of the data should be well known. This and similar studies reveal the importance of the Simpson’s paradox and data acquisition processes.

3. Application

Imagine that a disease has been observed in the particular area of your skin and different treatments have been applied for the disease. The results that appeared low at the end of the treatment gave higher results in total. Can this situation really happen? Let's analyze the situation with Simpson's paradox.

The data set consisted of the patients who applied to Muğla Sıtkı Koçman University Research Hospital between January 2015 and June 2015. Thirty-seven patients diagnosed with acne rosacea after clinical examination and the controls were recorded (Belli et al., 2016). Table 5 shows the paradox encountered.

Table 5. Detection of Simpson’s Paradox

	Topical Metronidazole	Oral Tetracycline
Central Face	0.75	0.80
All the Face	0.20	0.33
Both	0.53	0.36

The 37 patients included in the data set were divided into two groups according to the location of the disease. Topical metronidazole or oral tetracycline treatments were applied to the patients. Table 4 presents the rates of patients with different treatment methods depending on the location of the disease. The ratio of patients that topical metronidazole applied and with central involvement was 0.75; The rate of the disease which is on all face is 0.20. Overall, the rate of patients receiving this treatment was 0.53. When it was examined for oral tetracycline treatment, the rate of the patients with central face involvement was 0.80; the rate of the disease which is on all face is 0.33. In total, this ratio is 0.36. When the patient rates are analyzed according to the location of the disease, the rates of patients treated with topical metronidazole are lower than those of oral tetracycline treated with.

However, the results are reversed when only the treatment rates are examined without any location selection. When reversed, the rate of patient treated with oral tetracycline appears to be lower. Figure 2 was created to examine the variation of the proportions given in Table 5.

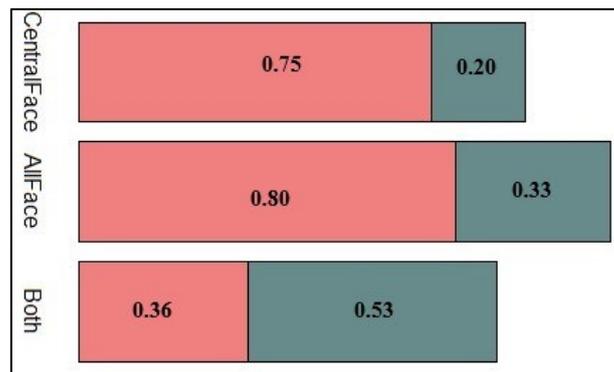


Figure 2. Bar Plot of Treatment Methods according to Disease Location

4. Conclusion

To avoid dilemma because of Simpson paradox, the paradox must be solved. For this purpose, the questions should be answered that such as how the data were produced, which factors have affected the data, whether the data should be grouped, which analyzes are needed, and what the future studies are aiming. In the study of Xu et al.(2018), we can intuitively see that sepal length values affect sepal width values. However, when the “iris species” are included in the study, the proportions of increases or decreases create confusion. In other words, when we look at the total data, the sepal width parameter is ignored. In the case of “Acne Rosacea”, this means that the region where the disease is seen is ignored.

The Simpson paradox may not always be easy to understand. In cases that require grouping (age ranges, educational level, etc.), attention should be paid to the Simpson paradox. Because there may be contradictions between groups and total numbers. Although the paradox cannot always be prevented, it can be controlled. As a first step, it should not be forgotten that the data studied are not total data. The second stage is based on the data collection process. The process of data collection or production should be well planned. If the causal model of the data is interpreted, other factors affecting the outcome can also be found.

In conclusion, this study is statistically significant for several reasons.

- It encourages the development of skills such as critical thinking and analysis.
- It emphasizes that rational thinking and questioning ability is the best way to influence.
- It emphasizes the importance of analyzing any situation.
- The mathematical definition of the paradox gives a different perspective to the concept of probability.
- It changes readers' perspectives against allegations of discrimination in our society.
- It forces the reader to search for different meanings behind the display.

References

Kock, N. (2015). “How Likely is Simpson’s Paradox in Path Models?”, International Journal of e-Collaboration, vol.11, no.1, pp.1-7.

Chambaz, A., Drouet I., Memetea, S. (2017). “Simpson’s Paradox, A Tale of Causality”, <https://hal.archives-ouvertes.fr/hal-01664904>.

Cankar, G. (2018). “Demonstration of Simpson’s Paradox in PISA 2015 Data: Confusing Differences between Boys and Girls”, *Orbis Scholae*, vol.12, no.2, pp.125-140.

Wagner, C.H. (1982). “Simpson’s Paradox in Real Life”, *The American Statistician*, vol.36, no.1, pp.46-48.

Xu, C., Brown, S.M., Grant, C. (2018). “Detecting Simpson’s Paradox”, *The Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31)*, Florida, America.

Taylor, S.A., Mickel, A.E. (2014). “Simpson's Paradox: A Data Set and Discrimination Case Study Exercise”, *Journal of Statistics Education*, vol.22, no.1, pp.1-18.

Maa, Y.Z. (2015). “Simpson’s Paradox in GDP and per capita GDP Growths”, *Empir Econ*, vol.49, pp.1301-1315.

Fitelson, B. (2017). “Confirmation, Causation, and Simpson's Paradox”, *2016 Episteme Conference*, vol.14, no.3, pp.297-309, United Kingdom.

Zalta, E.N., Nodelman, U., Alben, G., Anderson, R.L. (2016). “Simpson's Paradox”, *Metaphysics Research Lab, CSLI, Stanford University, California, USA*.

Belli, A.A., Gök, S.Ö., Akbaba, G., Etgu, F., Doğan, G. (2016). “The Relationship between Rosacea and Insulin Resistance and Metabolic Syndrome”, *Eur J Dermatol*, vol.36, no.3, pp.260-264.

O-25 A Simulation Study on the Unbalanced Design Properties of the Generalized p-value Based Tests

Mustafa CAVUS^{1*} and Berna YAZICI²

¹Department of Statistics, Eskisehir Technical University, Turkey, mustafacavus@eskisehir.edu.tr

²Department of Statistics, Eskisehir Technical University, Turkey, bbaloglu@eskisehir.edu.tr

Abstract – Several procedures are proposed for testing equality of group means under heteroscedasticity in the literature. Generalized p-value method is used to conduct powerful tests in the presence of nuisance parameters for unequal group variances in recent decades, such as Weerahandi Generalized F-test, Parametric Bootstrap test, Fiducial Approach test and the Alvandi et al. Generalized F-test. The unbalanced design may also cause the heteroscedasticity besides the unequal group variances in testing the equality of group means. In this study, the unbalanced design properties of the generalized p-value based tests are investigated in terms of Type I error probability and penalized power of the test. A Monte-Carlo simulation study is conducted under various unbalanced design scenarios for a different number of groups. As a result of the study, the properties of the tests are discussed and some useful comments are given to the researchers.

Keywords – unbalanced, penalized power, heteroscedasticity, ANOVA

1. Introduction

There are many procedures are improved to solve the generalized Behrens-Fisher problem. Weerahandi (1995) proposed the Generalized F-test based on the generalized p-value method. Krishnamoorthy et al. (2007) developed the Parametric Bootstrap test for testing equality of group means. The Fiducial Approach test is given by Li et al. (2011). Also, the new generalized p-value test is proposed by Alvandi et al. (2012). In many articles, their performances are investigated for several heteroscedasticity level, sample size and number of groups. The effect of the unbalanced design is also considered. However, they are really limited on the diversity of scenarios in simulations. Gamage and Weerahandi (1998) conducted a limited simulation study of the unbalanced case. They indicated that the GF test does not control the Type I error probability and requiring further research in this area. Krishnamoorthy et al. (2007) investigated the performance of the GF and PB test in unbalanced design, but it is not a systematic simulation study provides the useful information about the unbalancedness properties of the tests. Gokpinar and Gokpinar (2012) investigated the performance of the GF and PB test for the unbalanced small and moderate samples. However, they use the only one scenario and it is not enough to talk about the effect of unbalancedness.

In this study, an extensive simulation study is conducted to investigate the performance of the generalized p-value based tests under several unbalanced cases in terms of penalized power and Type I error probability. The paper is organized as follows. The generalized p-value method is introduced in Section 2 and the methods based on the generalized p-value methods are detailed. The penalized power is used instead of power of the test is introduced in Section 3. The results of the Monte-Carlo simulations are presented in Section 4. In the last section, the performance of the tests is discussed in details.

2. Generalized p-value

Tsui and Weerahandi (1989) introduced generalized inference method and generalized p-value is used to derive test statistics in the presence of nuisance parameter ξ . Let x denote the observed value of X and a function of $(X: x, \delta, \xi)$ be $T = T(X: x, \delta, \xi)$. It is called generalized pivotal quantity if it has the following two conditions:

- The distribution of $T(X: x, \delta, \xi)$ is free unknown parameters.
- The observed value of $T(X: x, \delta, \xi)$ is free of the nuisance parameter ξ .

For further details on the concepts of generalized p-values, see Tsuia and Weerahandi (1989).

2.1 Generalized F-Test

Weerahandi (1995) proposed the test statistic in (2) using the generalized p-value approach. Consider the standardized sum of squares between groups,

$$T_G(S_1^2, S_2^2, \dots, S_k^2) = \sum_{i=1}^k \frac{n_i}{S_i^2} \bar{X}_i - \frac{[\sum_{i=1}^k n_i \bar{X}_i / S_i^2]^2}{\sum_{i=1}^k n_i / S_i^2} \quad (1)$$

Assume $U_i \sim \chi_{ni-1}^2$ random sample,

$$T_{GF} = \sum_{i=1}^k (n_i U_i / v_i^2) \bar{x}_i^2 - \frac{[\sum_{i=1}^k (n_i U_i / v_i^2) \bar{x}_i]^2}{\sum_{i=1}^k n_i U_i / v_i^2} \quad (2)$$

where $v_i^2 = (n_i - 1)S_i^2$. The H_0 is rejected when $T_{GF} > T_G$.

2.2 Parametric Bootstrap Test

Krishnamoorthy et al. (2007) proposed a procedure to test the equality of group means under heteroscedasticity. Assume $Z_i \sim N(0,1)$ and $U_i \sim \chi_{ni-1}^2$ random samples, the test statistic of the PB test is computed as in (3).

$$T_{PB} = \sum_{i=1}^k \frac{Z_i^2 (n_i - 1)}{U_i} - \frac{[\sum_{i=1}^k \sqrt{n_i} Z_i (n_i - 1) / S_i U_i]^2}{\sum_{i=1}^k n_i (n_i - 1) / S_i^2 U_i} \quad (3)$$

The H_0 is rejected when $T_{PB} > T_G$.

2.3 Fiducial Approach Test

Li et al. (2011) proposed the test statistic in (4). Assume $t_i \sim t_{ni-1}$ random sample,

$$T_{FA} = \sum_{i=1}^k t_i^2 - \frac{\left[\sum_{i=1}^k \frac{\sqrt{n_i}}{S_i} t_i \right]^2}{\sum_{i=1}^k \frac{n_i}{S_i^2}} \quad (4)$$

The H_0 is rejected when $T_{FA} > T_G$.

2.4 Alvandi et al. Generalized Test

Alvandi et al. (2012) proposed the test statistic in (5) as an alternative of the GF test. Assume $U_i \sim \chi_{ni-1}^2$ random sample,

$$T_{AGF} = \sum_{i=1}^k \frac{n_i - 1}{U_i} (\bar{X}_i - q_i \tilde{X})^2 \quad (5)$$

where $q_i = \sqrt{\frac{n_i/s_i^2}{\sum_{i=1}^k n_i/s_i^2}}$ and $\tilde{X} = \sum_{i=1}^k q_i \bar{X}_i$. The H_0 is rejected when $T_{AGF} > T_G$.

3. Penalized Power

Monte-Carlo simulation studies are used to compare the performance of the tests in terms of power and Type I error probability. However, any comparison of the powers is invalid when Type I error probabilities are different. Cavus et al. (2019) proposed the penalized power approach in (6) to compare the power of the tests when Type I error probabilities are different.

$$\gamma = \frac{1 - \beta}{\sqrt{1 + \left| 1 - \frac{\alpha_i}{\alpha_0} \right|}} \quad (6)$$

where β is Type II error rate, α_i is Type I error of the test and α_0 is the nominal level. Penalized power adjusts the power function with the square root of the percentile deviation between Type I error probability and the nominal level. Thus, penalized power is used to compare the power of the tests in the simulation studies.

4. Monte-Carlo Simulation Study

A simulation study is conducted to investigate the effect of unbalanced design on the tests in terms of penalized power and Type I error probability. To measure the effect of unbalanced design, the sample sizes are considered balanced to unbalanced in the scenarios. The number of groups is fixed as $k = 3$. The results of the Type I error probabilities of the tests for small, moderate and large samples are tabulated in Table 1.

Table 1. Type I error probabilities of the tests

n_i	GF	PB	FA	AGF
10,10,10	0.0528	0.0540	0.0470	0.0916
9,10,11	0.0456	0.0478	0.0398	0.0792
7,10,13	0.0428	0.0480	0.0354	0.0704
5,10,15	0.0412	0.0492	0.0328	0.0580
30,30,30	0.0490	0.0488	0.0466	0.1058
27,30,33	0.0486	0.0502	0.0468	0.1004
21,30,39	0.0500	0.0524	0.0488	0.0928
15,30,45	0.0474	0.0496	0.0450	0.0734
50,50,50	0.0456	0.0458	0.0448	0.1076
45,50,55	0.0518	0.0512	0.0508	0.1078
35,50,65	0.0532	0.0542	0.0516	0.0986
25,50,75	0.0470	0.0486	0.0446	0.0782

According to the Table 1, the PB test is the best to control the Type I error probability for the nominal level $\alpha = 0.05$. While the GF is better than the PB test in balanced designs, the PB is superior in unbalancedness increases. The FA test is the most negatively affected test from the unbalancedness. It does not control the Type I error probability for small samples. Also, the Type I error probability of the AGF test is affected positively when the unbalancedness increases.

The penalized power results are adjusted the power with respect to the deviation of Type I error probability from the nominal level given in Tables 2, 3 and 4.

Table 2. Penalized power of the tests for small samples

n_i	Δ	GF	PB	FA	AGF
10,10,10	0.2	0.3536	0.3576	0.3339	0.3843
	0.5	0.9539	0.9461	0.9511	0.7366
	0.8	0.9731	0.9623	0.9713	0.7388
9,10,11	0.2	0.3536	0.3670	0.3122	0.4179
	0.5	0.9505	0.9711	0.9022	0.7938
	0.8	0.9587	0.9787	0.9114	0.7946
7,10,13	0.2	0.4088	0.4291	0.3408	0.4883
	0.5	0.9325	0.9784	0.8766	0.8427
	0.8	0.9349	0.9806	0.8798	0.8427
5,10,15	0.2	0.4166	0.4514	0.3312	0.5140
	0.5	0.9212	0.9911	0.8603	0.9285
	0.8	0.9221	0.9921	0.8626	0.9285

Table 3. Penalized power of the tests for moderate samples

n_i	Δ	GF	PB	FA	AGF
30,30,30	0.2	0.8499	0.8516	0.8277	0.6553
	0.5	0.9901	0.9882	0.9676	0.6875
	0.8	0.9901	0.9882	0.9676	0.6875
27,30,33	0.2	0.8808	0.8932	0.8638	0.6796
	0.5	0.9863	0.9980	0.9695	0.7057
	0.8	0.9863	0.9980	0.9695	0.7057
21,30,39	0.2	0.9318	0.9112	0.9192	0.7177
	0.5	1	0.9768	0.9882	0.7340
	0.8	1	0.9768	0.9882	0.7340
15,30,45	0.2	0.9243	0.9442	0.8984	0.8161
	0.5	0.9750	0.9960	0.9535	0.8253
	0.8	0.9750	0.9960	0.9535	0.8253

Table 4. Penalized power of the tests for large samples

n_i	Δ	GF	PB	FA	AGF
50,50,50	0.2	0.9416	0.9434	0.9344	0.6790
	0.5	0.9587	0.9605	0.9517	0.6817
	0.8	0.9587	0.9605	0.9517	0.6817
45,50,55	0.2	0.9721	0.9781	0.9820	0.6804
	0.5	0.9825	0.9882	0.9921	0.6810
	0.8	0.9825	0.9882	0.9921	0.6810
35,50,65	0.2	0.9666	0.9576	0.9810	0.7118
	0.5	0.9695	0.9605	0.9844	0.7121
	0.8	0.9695	0.9605	0.9844	0.7121
25,50,75	0.2	0.9682	0.9831	0.9470	0.7991
	0.5	0.9713	0.9863	0.9500	0.7996
	0.8	0.9713	0.9862	0.9500	0.7996

According to the results in Table 2, the PB test is superior than others. Also, the GF test is better than the FA test in terms of penalized power. In Tables 3 and 4, it is clearly seen that there is no difference between the performance of the test in higher effect sizes. The results of the AGF tests in Tables 3 and 4 are not considered because it can't control the Type I error probability.

Acknowledgment

This study is supported in part by the Scientific Research Projects commission of Eskisehir Technical University under the research project grant no.19ADP049.

5. Conclusions

The performances of the tests are investigated in a simulation study under several unbalanced cases in terms of penalized power and Type I error probability. The unbalanced scenarios are fixed to balanced to higher level of unbalanced cases. According to the results in Table 1, the PB test is the best to control the Type I error probability, also the GF and FA test is fairly well performed for moderate and large samples.

The penalized power properties of the tests are about similar for moderate and large samples except the AGF test. It should not be considered in penalized power comparisons because of the poor performance on the controlling Type I error probability. Finally, it is concluded that the PB test is must be chosen for testing equality of several group means under unequal variances in case of higher level of unbalancedness. For moderate and large samples, the AGF test is should not be used.

References

- Cavus, M., Yazıcı, B., Sezer, A. (2017) “Modified tests for comparison of group means under heteroskedasticity and non-normality caused by outlier(s)”, *Hacettepe Journal of Mathematics and Statistics*, 46 (3): 493-510.
- Cavus, M., and B. Yazici. (2019). “doex: The One-Way Heteroscedastic Anova Tests.” <https://CRAN.R-project.org/package=doex>.
- Cavus, M., Yazıcı, B., Sezer, A. (2019) “Penalized power approach to compare the power of the tests when Type I error probabilities are different”, *Communications in Statistics-Computation and Simulation*, <https://doi.org/10.1080/03610918.2019.1588310>.
- Gamage, J., and Weerahand, S. (1998). “Size Performance of Some Tests in One-Way Anova.” *Communications in Statistics - Simulation and Computation* 27 (3): 625–40.
- Gokpinar, E. Y., and Gokpinar, F. (2012). “A test based on the computational approach for equality of means under the unequal variance assumption” *Hacettepe Journal of Mathematics and Statistics*, 41 (4): 605–613.
- Krishnamoorthy, K., Lu, F., Mathew, T. (2007). “A parametric bootstrap approach for anova with unequal variances: Fixed and random models”, *Computational Statistics and Data Analysis*, vol. 51, pp.5731-5742.
- Li, X., J. Wang, and Liang, H. (2011). “Comparison of Several Means: A Fiducial Based Approach.” *Computational Statistics and Data Analysis* 55 (5): 1993–2002. <https://www.sciencedirect.com/science/article/pii/S0167947310004779>.
- Sadooghi-Alvandi, S. M., Jafari, A. A., Mardani-Fard, H. A. (2012). “One-way ANOVA with unequal variances”, *Communications in Statistics: Theory and Methods*, vol. 41, pp.4200-4221.
- Tsui, K., and Weerahandi, S. (1989). “Generalized P-Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters.” *Journal of the American Statistical Association* 84: 602–7. <https://www.jstor.org/stable/2289949>.
- Weerahandi, S. (1995). “ANOVA under unequal error variances”, *Biometrics*, vol. 51, pp.589-599.

O-26 Imputation of Missing Observations in Longitudinal Data via Neural Network

Marwa BenGhoul^{1*}, Berna Yazıcı²

¹ Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, Turkey, benghoulmarwa@gmail.com

² Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, Turkey, bbaloglu@eskisehir.edu.tr

Abstract – Longitudinal data is a data which tracks the same information over subjects on different timepoints, it is mostly used in the pharmaceutical industry. Such type of data has common problem known as attrition or missing data. Despite several research studies highlighted the cruciality of knowing the missingness mechanism before the imputation method, it still ignored. Recently, Neural Network (NN) has started to get more attention as a technic of data imputation for Missing Completely at Random data (MCAR). Spite of many studies investigated this approach on classic clinical data, it still not highly applied specially for longitudinal data. Hence, this research consists in investigating the efficiency of NN, particularly Multilayer Perceptron (MLP) on handling the missing data. Based on recent researches results, the activation function in this paper will be a wavelet function. Four wavelet functions have been tested (Gaussian, Morlet, Meyer and Mexican Hat) as activation function. A comparison between the ad hoc imputation methods, Expectation Maximum (EM) and the proposed approach is provided. NN approach with wavelet functions as activation functions, particularly Morlet function, show interesting performance, better than the ad hoc imputation methods and with very slight difference from the EM.

Keywords – *longitudinal data, missing data, neural network, multilayer perceptron wavelet function, imputation*

1. Introduction

Longitudinal data is a data which tracks the same information over subjects on different timepoints, it is mostly used in the pharmaceutical industry to follow the dynamic development of the subject. At least two repeated observations should be recorded to get a longitudinal data (Fitzmaurice et al. ,2014; Hedeker and Gibbon, 2006). Such type of data has common problem known as attrition or missing data due to treatment discontinuation, loss of follow-up, death etc.

Despite several research studies highlighted the cruciality of knowing the type of mechanism before the imputation method, it still ignored. Hedeker and Gibbon (2006) defined the missing data mechanism as what characterizes the reasons for the missing data. Wood et al. (2004) have considered the missingness mechanism highly required to be known before handling missing data, as the performance of longitudinal data analysis models can depend critically on it. These mechanisms as explained by Hedeker and Gibbons (2006) answered the fundamental question of why data are missing. They have not been considered before Rubin (1976) who introduced the corresponding notation which still commonly used. Recently, Neural Network (NN) has started to get more attention as a technic of data imputation for Missing Completely at Random (MCAR) data. Spite of many researchers investigated this approach on classic clinical data, it still not vastly applied specially for longitudinal data.

During the last two decades, a mathematical tool in signal decomposition, known as wavelets, has started to be applied in different fields such as signal denoising, smoothing approaches, outliers handling etc. and as an activation function for the NN. Hence, based on recent researches results, this research consists in investigating the efficiency of NN, particularly Multilayer Perceptron (MLP) on handling the missing data by applying the wavelet function as activation function.

A generated longitudinal data based on a historical hypertension study published by National Institute Health will be used. Four wavelet functions will be tested (Gaussian, Morlet, Meyer and Mexican Hat) as activation functions. A comparison between the ad hoc imputation methods, Expectation Maximum (EM) and NN approach will be provided. SAS macros published by Sarle (1994) are updated and others are created to perform the NN for this research. NN approach, particularly with Morlet as an activation function, show interesting performance, better than the ad hoc imputation methods and with very slight difference from the EM.

The remainder of the paper will be organized in 4 sections: Second section is devoted to the theoretical framework examining the wavelets functions and the algorithm of addressing missing data via neural network. Third section defines the generated dataset. Section four summaries the findings and discuss the results. The final section sums up this research.

2. Materials and Methods

2.1 Wavelets analysis

Wavelets analysis has been settled as an extension of Fourier concept allowing simultaneously the analysis of the frequential domain and the temporal one without losing any aspect of the information. (Meyer, 1990; Daubechies, 1992; Graps,1995)

As definition, wavelet is a small wave which is located in time and frequency. Indeed, it is a mathematic function allowing the decomposition of a signal or a variable into different components via time at a specific frequency. Also, wavelet is stated as a *mathematical microscope* (Arneodo et al.,1995; Muzy et al., 1994) due to its ability in showing the weak transients and peculiarities in the time series, it uses the optics of the microscope, its magnification varies with the scale factor.

The following admissibility requirement is the only specific condition for the wavelet:

$$C_{\Psi} = \int_0^{\infty} \frac{|\Psi(f)|}{f} df < \infty, \text{ defining } \Psi(f) = \int_{-\infty}^{+\infty} \psi(t)e^{-2\pi ift} dt \text{ the Fourier transform.}$$

This condition facilitates the reconstruction of the signal after the decompositon (Percival & Walden, 2000). f is a frequency function for $\Psi(f)$, ψ is known as the mother wavelet. To ensure that $C_{\Psi} < \infty$, the following conditions, related to the mother wavelet, are required:

$$\begin{aligned} \psi(0) &= 0, \text{ or } \int_{-\infty}^{\infty} \psi(t)dt = 0 \\ \int_{-\infty}^{\infty} |\psi(t)|^2 dt &= 1, \text{ representing the energy unit.} \end{aligned}$$

The following wavelets functions will be used as activation functions.

Gaussian wavelet function $\Psi(t) = \frac{t}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$, Morlet wavelet function $\Psi(t) = \cos(1.75t)(-\frac{t^2}{2})$, Meyer wavelet function (approximate formula) $\Psi(t) = 35t^4 - 84t^5 + 70t^6 - 20t^7$, Mexican Hat (Mexihat) wavelet function $\Psi(t) = c(1 - t^2)\exp(-\frac{t^2}{2})$, $c = \frac{2}{\sqrt{3}}\pi^{-1/4}$.

2.2 Addressing Missing data via Neural Network

Handling missing methods data can be classified into four main categories: complete cases, interpolation, imputation and modelling.

Several researches have studied addressing the missing data via neural network such as (Yoon & Lee, 1999; Sonnberger & Maine, 2000). As noted by Silva-Ramírez et al. (2011), for most of the studies, training the NN is applied only on the not missing and for each variable separately. Indeed, the target value is observed, so the generated NN is proceeded for imputing the missing observations (Koikkalainen, 2002; Eurodit, 2005). Silva-Ramírez et al. (2011) used the Multilayer perceptron to impute data on MCAR, they proceed the training subset in a way that it includes observed and missing data. Smieja et al. (2018) addressed the missing data by NN, their proposed approach consists in replace typical neuron's response in the first hidden layer by its expected value.

Lipton et al. (2016) and Choi et al. (2015) used Recurrent Neural Networks (RNNs) by combining missing entries with the input or by applying simple imputations. In the same context, Che et al. (2016) developed a novel model called Gated Recurrent Unit (GRU) to better predict the missing patterns via RNN.

As wavelet analysis has gained a lot of attention during the last two decades, it has been applied also for imputation missing data. Mondal and Percival (2010) applied the wavelet variance analysis (defined as a scale-based decomposition of the process variance) for gappy time series. Heaton and Silverman (2008) proposed a new approach of imputation by utilizing the expected sparse representation of a surface in a wavelet or lifting scheme basis. Rocha et al. (2010) utilized a wavelet scheme to reconstruct the missing sections in time series signals based on the estimation of the available sections. Altan & Ustundag (2012) reconstructed the missing observations of a meteorological data via the wavelet transform. Lilly (2017) developed an interesting platform to analyze and visualize time-localized events in noisy time series using wavelets analysis.

It is indispensable to highlight that there are many studies that applied the Wavelet Neural Network (WNN) to impute missing data. WNN concept was invented by Zhang and Benveniste (1992), it is based on the wavelet transform and it is considered as an alternative to feedforward NN for approximating random nonlinear functions. Then, several studies have utilized WNN for different applications such as Yu & Chen (2007) who they used the WNN to differentiate six different beat types in ECG signal and Dong et al. (2008) who applied fuzzy WNN and irregular sets to forecast fault diagnosis accuracy of power transformers. Panigrahi et al. (2012) investigated how to remove and interpolate the missing data by WNN. Wang et al. (2013) explored the WNN by applying multiple wavelet Functions in target threat assessment.

Based on the works of Silva-Ramírez et al. (2011), Eurodit (2005), Panigrahi et al. (2012) and Wang et al. (2013), this research consists in imputing the missing data for MCAR in longitudinal data by using the

predictions, the novelty here is that different wavelet mother functions will be tested as activation functions for the neural network and not as a neuron. The algorithm to impute the missing data via the predictions of NN has mainly two steps: the training of the wavelet network then the imputation step. The training dataset includes complete observations (no missing data), the wavelet functions present the activation function. The missing observation will be imputed with the corresponding predicted value. Here, filling the gaps with the predictions is not a single imputation as the predictions are not equal. The proposed algorithm is addressed as Multilayer Perceptron for Imputation of Missing Data (MLPIMID) in BenGhoul and Yazıcı (2019). For this paper, within the MLPIMID algorithm, the activation functions will be wavelet functions not sigmoid.

Figure 1 summarizes the MLPIMID algorithm proposed in BenGhoul and Yazıcı (2019).

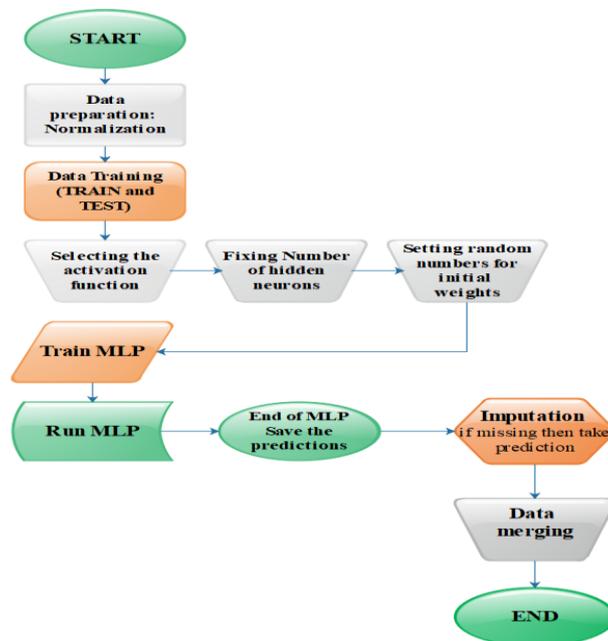


Figure 1. Flowchart of MLPIMID algorithm

2.3 Dataset

A generated dataset, based on a previous clinical study, will be utilized in this paper. The concept of generating data by referring previous clinical studies, was adopted from previous research papers (Chen et al., 2008 and Psioda, 2012). Framingham Heart Study provided by National Institutes of Health (NIH) is considered as the previous study to be used for data generating. The step after data generation consists in randomly imposing different percentages of missingness (5%, 10%, 15%, 20%, 25% and 30%) within the output variable of incident hypertension i.e. the first variable Y1 will have 5% of missing observations, the second variable Y2 will include 10% of missing observations etc. The method of filling the gaps with the NN predictions will be tested for each missingness level.

3. Results and discussion

This section aims to provide the results of the proposed algorithm. Six missingness levels have been imposed for the variable Incident Hypertension, four mother wavelet functions will be applied separately. Root Mean Squared Error (RMSE) is an estimate of the prediction error using the unknown optimal weights in the population. Root Final Prediction Error (RFPE) is an estimate of the prediction error using the weights estimated from the sample to make predictions for the population. Table 1 resumes RMSE and RFPE for each network fitted to the hypertension dataset for the different missingness levels and by activation function.

Table 1. RMSE and RFPE of the network by activation function

	Activation function							
	Gaussian		Meyer		Mexican hat		Morlet	
	RFPE	RMSE	RFPE	RMSE	RFPE	RMSE	RFPE	RMSE
5% missing observations	0,4548	0,4205	0,7241	0,6696	0,3874	0,3582	0,2234	0,2066
10% missing observations	0,4613	0,4229	0,7365	0,6751	0,3975	0,3644	0,2258	0,2070
15% missing observations	0,4833	0,4388	0,7660	0,6954	0,3970	0,3604	0,2367	0,2149
20% missing observations	0,5066	0,4506	0,8044	0,7155	0,4231	0,3763	0,2487	0,2212
25% missing observations	0,5859	0,4944	0,9337	0,7879	0,5057	0,4268	0,2894	0,2442
30% missing observations	0,9328	0,7147	1,4783	1,1326	0,7679	0,5883	0,4577	0,3507

Referring to RFPE and RMSE, the best model is the MLP model that includes Morlet as activation function then MLP model with Mexican Hat as activation function.

3. Conclusion

Plenty approaches have addressed handling missing data via the multiple imputation, EM and more recently neural networks and wavelets analysis. Hence, this research has as primary objective combining the NN and wavelets analysis to handle missing data for Missing Completely at Random longitudinal data.

The proposed method has two stages: the first one concerns training a NN based on a non-missing data and use the predictions results to fill the gaps, then merge this data with the original one to get the imputed variable. It is not an interpolation by the mean as proposed in several studies that applied NN. Furthermore, the activation functions were used as the wavelets mother functions (Morlet, Meyer, Gaussian and Mexican Hat). A longitudinal generated data with different missingness levels, was utilized to test the performance of this method.

The results of the proposed method show interesting performance which is better than the ad hoc imputation methods and very slightly different from the EM notably when the Morlet wavelet function was used as activation function. Moreover, with increasing the percentage of missing data, the algorithm remained powerful.

Based on the presented results, further research studies are highly recommended to deep investigate the WNN, the NN with wavelets activation functions to deal with missingness in longitudinal data.

Acknowledgment

Authors are thankful to Özer Özdemir, Prof. at Eskişehir Technical University for providing valuable support related to the Development of the Neural Network Algorithm.

References

- Altan, N.T & Ustundag, B.B. (2012). Reconstruction of Missing Meteorological Data Using Wavelet Transform. IEEE, 2012 First International Conference on Agro- Geoinformatics (Agro-Geoinformatics). <https://doi.org/10.1109/Agro-Geoinformatics.2012.6311644>.
- Arneodo, A., Bacry, E., & Muzy, J.F. (1995). The thermodynamics of fractals revisited with wavelets, *Physica A* 213-232. [https://doi.org/10.1016/0378-4371\(94\)00163-N](https://doi.org/10.1016/0378-4371(94)00163-N).
- BenGhoul, M. & Yazıcı, B. (2019). Multilayer Perceptron for Imputation of Missing Data (MLPIMID) for Missing Completely at Random data (MCAR). Manuscript submitted for publication.
- Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. (2016). Recurrent Neural Network for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1), 6085. <https://doi.org/http://dx.doi.org/10.1038/s41598-018-24271-9>.
- Chen, J., Stock, S., & Deng, C. (2008). Sample Size Estimation Through Simulation of a Random Coefficient Model by Using SAS Abstract. *PharmaSUG*, 3.
- Choi, E. Bahadori, M.T, & Sun, J. (2015). Doctor AI: Predicting clinical events via recurrent neural networks. Cornell University. arXiv preprint arXiv:1511.05942.
- Daubechies, I (1992). Ten lectures on wavelets. First edition: Society for Industrial and Applied Mathematics.
- Dong, L., Xiao, D., Liang, Y., & Liu, Y. (2008). Rough set and fuzzy wavelet neural network integrated with least square weighted fusion algorithm-based fault diagnosis research for power transformers. *Electric Power Systems Research*, 78,129–136. <https://doi.org/10.1016/j.epsr.2006.12.013>.
- Eureedit, (2005). Interim report on evaluation criteria for statistical editing and imputation. <http://www.cs.york.ac.uk/eureedit>.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2014). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics, xxv, 701.
- Graps, A. (1995). An Introduction to Wavelets. *IEEE Computational Science and Engineering*, 2(2), 50-61. <https://doi.org/10.2307/2153134>.
- Heaton, T.J. & Silverman, B.W. (2008). A wavelet- or lifting-scheme-based imputation method. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 70(3), 567-587. <https://doi.org/10.1111/j.1467-9868.2007.00649.x>.

- Hedeker, D., & Gibbons, R. (2006). *Longitudinal Data Analysis*. 1st edition, New Jersey, Wiley-Blackwell.
- Howell, D. C. (2008). The Treatment of Missing Data. *The SAGE Handbook of Social Science Methodology*, 212–226. <https://doi.org/10.4135/9781848607958.n11>.
- Koikkalainen, P. (2002). Neural networks for editing and imputation. In *Data clean 2002 conference*. Retrieved from: <http://erin.it.jyu.fi/dataclean/abstracts/node30.html>.
- Kreindler, D.M & Lumsden, C.J. (2012). The effects of the irregular sample and missing data in time series analysis. *Nonlinear Dynamics Psychology and Life Sciences* 10(2), 187-214.
- Lilly, JM. (2017). *Element analysis: a wavelet-based method for analyzing time-localized events in noisy time series*. Royal Society Publishing, 473: 20160776. <http://dx.doi.org/10.1098/rspa.2016.0776>.
- Lipton, Z.C, Kale, C.K & Wetzel, R. (2016). Directly modeling missing data in sequences with RNN's: Improved classification of clinical time series. *Proceedings of the 1st Machine Learning for Healthcare Conference*, PMLR 56:253-270.
- Meyer, Y. (1990). *Ondelettes et Opérateurs*, vol. I–III. Paris: Hermann 1990.
- Mondal, D., & Percival, D.B. (2010). Wavelet variance analysis for gappy time series. *Ann Inst Stat Math*, 62:943–966. <https://doi.org/10.1007/s10463-008-0195-z>.
- Muzy, J.F., Bacry, E., & Arneodo, A., The multifractal formalism revisited with wavelets, *Int. J. Bifurcation Chaos*, 4(2), 245, 1994. <https://doi.org/10.1142/S0218127494000204>.
- Panigrahi, L., Das, K, Mishra, D. & Ranjan, R. (2012). Removal and Interpolation of Missing Values using Wavelet Neural Network for Heterogeneous Data Sets. *ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 1004-1009. ISBN: 978-1-4503-1196-0. <https://doi.org/10.1145/2345396.2345558>.
- Percival, D, Walden, A. (2000). *Wavelet Methods for Time Series Analysis*, Cambridge University Press. <https://doi.org/10.1017/CBO9780511841040>.
- Pickles, A. (2005). *Missing Data, Problems and Solutions*. *Encyclopedia of Social Measurement*.
- Psioda, M. (2012). *Random Effects Simulation for Sample Size Calculations Using SAS*. 1st edition, Carolina, University of North Carolina, Chapel Hill NC, p. 1–11.
- Quarta, L., *Une introduction (élémentaire) à la théorie des ondelettes*, 2nd edition, Belgium, Mons-Hainaut University, p. 4-18, 2001.
- Rocha, TR., Paredes, SP. & Henrique, JH (2010). A Wavelet Scheme for Reconstruction of Missing Sections in Time Series Signals. *Computing in Cardiology 2010*; 37:461–464. ISSN 0276–6574.
- Rubin, D. B. (1976). *Inference and Missing Data*. *Biometrika*. <https://doi.org/10.2307/2335739>.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, John. <https://doi.org/10.1002/9780470316696>.
- Sarle, S.W. (1994). *Neural Network Implementation in SAS Software*. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. SAS Institute Inc., Cary, NC, USA.

- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037//1082-989X.7.2.147>.
- Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M., Cubiles-de-la-Vega, M.D. (2011). Missing value imputation on missing completely at random data using multilayer perceptron. *ScienceDirect, Neural Networks* 24, 121–129.
- Smieja, M., Struski, L., Tabor, J., Zielinski, B. & Spurek, P. (2018). Processing of missing data by neural networks. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
- Sonnberger, H., & Maine, N. (2000). Editing and imputation in Eurostat. In Working paper N21, UN/ECE work session on statistical data editing. Conference of European statisticians.
- Wang, G., Guo, L. & Duan, H. (2013). Wavelet Neural Network Using Multiple Wavelet Functions in Target Threat Assessment. *The Scientific World Journal*, 7. <http://dx.doi.org/10.1155/2013/632437>.
- White, I.R, Royston, P. & Wood, A.M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4), 377–399. <https://doi.org/10.1002/sim.4067>.
- Wood, A. m., White, I. r., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4), 368–376.
- Yoon, S.Y & Lee, S.Y. (1999). Training Algorithm with Incomplete Data for Feed-Forward Neural Networks. *Neural Processing Letters* 10, 171–179. <https://doi.org/10.1023/A:1018772122605>.
- Yu, S.N., & Chen, Y.-H. (2007). Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recognition Letters*, 28, 1142–1150. <https://doi.org/10.1016/j.patrec.2007.01.017>.
- Zhang, Q. & Benveniste, A. (1992). Wavelet Networks. *IEEE Transactions on Neural Networks*, 3(6), 889-898.

O-28 A New Method Based on Interquartile Range to Feature Selection for Classification in Big Data

Ahmet Kocatürk^{1*}, Bülent Altunkaynak²

¹Department of Statistics, Gazi University, Ankara, ahmetkocaturk@gazi.edu.tr

²Department of Statistics, Gazi University, Ankara, bulenta@gazi.edu.tr

Abstract – In the analysis of large-scale gene expression data, it is important to identify a well-representing subset of the data. Specifying the subset can be done with the feature selection. Feature selection is generally divided into two groups. First, wrapper methods are methods that seek combinations of features that maximize the accuracy of model. Second, filter methods examine the effect of each feature on the accuracy of model. In this study, a new filtering method based on the interquartile range for feature selection (FSIQR) is proposed in gene expression data. The FSIQR method takes into account the overlapping areas of interquartile range from each class. The feature selection methods used by the FSIQR method are Chi-square, Information Gain, and ReliefF. Also, it is aimed to compare FSIQR method and other methods in terms of accuracy by using real data sets. Support vector machines and k-neighborhood method are used for accuracy rates and n-fold and leave-one-out methods are applied for cross validation. As a result, FSIQR method, which is a new filtering method based on interquartile range value for feature selection, has been found to reach a higher accuracy rate in many cases compared to other feature selection methods on real data.

Keywords – Gene Expression Data, Feature Selection, Filter Method

1. Introduction

The feature selection can be defined as specifying a subset that best represents the data to reduce the size of the data set. The aim is to select the features that have the most effect on the method to be applied. A small number of particularly high accuracy models can be achieved with reasonable computer time (Altunkaynak, 2019). Feature selection methods can be examined under two headings: wrapper and filter methods (John et al., 1994).

Wrapper methods are methods that generally seek combinations of features that will maximize the explanation / verification power of the model. Examples of wrapper methods are recursive feature elimination, genetic algorithms, and simulated annealing.

Filter methods are based on examining the effect of features on the model's explanation / verification power one by one. In a classification problem, the examination of the relationship between the class and each property in sequence and selecting the features with the highest relationship can be given as an example of filter methods. Examples of filter methods are Chi-square Statistics, Information Gain and ReliefF.

Gene expression data obtained with the use of microarray technology are very large number of feature (variable) data. Gene data provide important information about cancers, tumors and genetic diseases

(Giordani, 2009 ; Kasim et. al., 2016; Vicente, 2009). Therefore, the selection of important features is important for the analyzes to be conducted on these data.

In this study, a new filtering method (Feature Selection based on Interquartile Range, FSIQR) based on interquartile range value is proposed for feature selection in gene expression data. The proposed method takes into account the overlap areas of the interquartile range obtained from each class. The overlap areas of the properties that are important for the classification are expected to be small.

In study, FSIQR method, Chi-square, Information Gain, and ReliefF feature selection methods were compared with three real gene data. 10, 20, 40, 60, 80 and 100 genes were selected for each method and correct classification possibilities were obtained. Support vector machines and k-neighborhood method were used in the classification, while n-fold and leave-one-out methods were applied for cross validation. For the Chi-square, Information Gain, and ReliefF feature selection methods, the R-project v3.5.2 uses a *biocomb* package and a *datamicroarray* package for data sets. FSIQR algorithm is coded by the authors in R-project v3.5.2.

2. Feature Selection Methods

2.1 Chi-square

The Chi-square statistics used to investigate the relationship between categorical variables can also be used for feature selection. For this, Chi-square values between class and properties are calculated. The feature with the largest Chi-square value is the most distinctive on the class variable. The Chi-square value indicates the significance of the feature. The *Cramer V* correlation coefficient is used to find the normalized weight coefficients in the 0-1 range of each feature. That is, the *Cramer V* correlation coefficient in this feature selection is the weight coefficient indicating the significance of each feature.

In order to calculate Chi-square statistics, a cross table (two-dimensional frequency table) is created between the class and the feature. A cross table is created for the l and m feature, with k being the number of levels for the class as follows.

Table 1. Example of a cross table

Class	Feature of m				Total
	1	2	...	k	
1	n_{11}	n_{12}		n_{1k}	$n_{1.}$
2	n_{21}	n_{22}		n_{2k}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
l	n_{l1}	n_{l2}		n_{lk}	$n_{l.}$
Total	$n_{.1}$	$n_{.2}$...	$n_{.k}$	n

where n_{lk} values are called as observed frequencies. Expected frequencies are e_{lk} .

$$e_{lk} = n \frac{n_{.k}}{n} \frac{n_{l.}}{n} = \frac{n_{.k}n_{l.}}{n} \quad (1)$$

The chi-square statistic is defined as follows, while the observed frequencies and expected frequencies are present for each of the $l \times k$ cells.

$$\sum_{j=1}^l \sum_{i=1}^k \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \quad (2)$$

The *Cramer V* coefficient to be used for feature selection, with the chi-square statistic χ_h^2 calculated by this formula, is calculated by the following formula.

$$V = \sqrt{\frac{\chi_h^2}{n \times \min(l-1, k-1)}} \quad (3)$$

2.2 Information Gain

Quinlan (1986) suggested that the *Information Gain (IG)* value takes into account the entropy value among features.

$$IG(X) = H(Y) - H(Y/X) \quad (4)$$

where X and Y are features. The entropy values $H(Y)$ and $H(Y/X)$ are calculated using the following formulas.

$$H(Y) = - \sum_{y \in Y} p(y) \log_2(p(y)) \text{ and } H(Y/X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \log_2(p(y/x)) \quad (5)$$

2.3 ReliefF

The Relief algorithm developed by Kira and Rendell (1992) assigns a weight value between -1 (bad) and 1 (good) to each feature. In calculating the weight value, the distance to the nearest unit (H) in the class where each unit is located and to the nearest unit (M) in the other class is taken into consideration.

For a data with d features, set the features to $X = \{X_1, X_2, \dots, X_d\}$. Let the number of units at level $j = 1, 2, \dots, l$ including the number of level l belonging to the class j . In this case, the steps of the Relief algorithm for the X_i feature, with $n = n_1 + n_2 + \dots + n_l$ can be given as follows.

1. $k = 1$ and $w_i = 0$
2. Calculate weight coefficient $w_i = w_i + \frac{diff(R_k, H) + diff(R_k, M)}{n}$, where $diff(A, B) = \begin{cases} 0 & \text{if } A = B \\ 1 & \text{otherwise} \end{cases}$
3. $k = k + 1$, if $k > n$, go to step 6, otherwise go to step 2.
4. If $w_i \geq \theta$, select feature.

3. Proposed Method

Number of classes l and $j = 1, 2, \dots, l$. Let the first and third quarter values for class j of the property i be represented by Q_{ij}^1 and Q_{ij}^3 , respectively. Limits of interquartile range for i feature are $IQR_{ij} = [Q_{ij}^1, Q_{ij}^3]$. The algorithm calculates the overlap values of the *IQR* ranges of the classes for each feature. There should be little overlap between classes in high discriminatory features. The steps of the proposed *FSIQR* algorithm are:

For each features;

1. Calculate IQR_{ij} intervals for each class.

2. Sort ranges from small to large by lower boundaries (Q_{ij}^1).
3. Calculate overlap areas $OA_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^l \varphi_i(j, k)$, where $\varphi_i(j, k) = \begin{cases} Q_{ij}^3 - Q_{ik}^1 & \text{if } Q_{ik}^1 < Q_{ij}^3 \\ 0 & \text{otherwise} \end{cases}$
4. Calculate area coefficient $AC_i = \frac{OA_i}{\max(Q_{i1}^3, Q_{i2}^3, \dots, Q_{il}^3) - Q_{i1}^1}$
5. Calculate the normalized area coefficient $NAC_i = \frac{AC_i}{\max(AC_s)}$, $s = 1, 2, \dots, d$.
6. Calculate weight coefficient $w_i = 1 - NAC_i$, if $w_i \geq \theta$, select feature.

4. Application

4.1 Data Sets

Breast cancer: West et al. (2001) used by breast cancer data, 49 samples and 7129 genes is a data set. The data set has a structure of two classes depending on whether the estrogen receptor (estrogen receptor, ER) is (+) or (-).

NCI-60 cancer cells: Subramanian et al. (2005) used by the NCI-60 cancer cells data is a set of two classes consisting of 50 samples and 10100 genes. The class with 17 samples consisted of normal patients and the class with 33 samples consisted of patients carrying mutations in the gene.

CNS tumor: Pomeroy et al. (2002) used by CNS (Central Nervous System) tumor data consisting of 60 observations and 7128 gene is a set of two class data.

4.2 Classification algorithms

Support Vector Machine (SVM)

SVM classifies by creating optimal hyperplanes in the vector space of features to maximize the space between objects of different classes (Cortes and Vapnik, 1995). In this method, an iterative training algorithm is used to minimize the error function defined by $\Lambda(w)$ to construct an optimal hyperplane (Chandra and Gupta, 2011). For $\xi_i \geq 0$, $i = 1, 2, \dots, n$; $y_i[w'K(x_i) + b] \geq 1 - \xi_i$ under the constraint, the error function is defined as follows.

$$\Lambda(w) = \frac{1}{2}w'w + C \sum_{i=1}^n \xi_i$$

where, w coefficients vector, b constant and ξ_i noisy data are parameters that allow incorrect classification. For each instance i , the independent variables represented by class labels y_i are x_i s. K , is the kernel function that converts input data into a higher dimensional feature space and is used to create a nonlinear decision boundary. In this study, linear kernel function is used. Parameter C is used to check overfitting. The larger the C value, the more penalized the error. A standard SVM is used in two classes of data. Multi-class SVM methods have been developed for data with more than two classes (Chandra and Gupta, 2011; Chih-Wei and Chih-Jen, 2002).

k-Neighborhood Method

This algorithm, used in cases where the independent variables are quantitative, performs classification based on the distances between observations. Steps can be given as follows.

- Determination of k value (number of neighbors)

- Calculation of distances between observations
- Sorting of observations from small to large by distances
- Assigning the most repetitive class in k observations with the smallest distance

4.3 Cross Validation

n-Fold

In this method, data is randomly divided into n parts. Starting from the first part of the data, each piece is used as test data up to the n th part, respectively. For example, if $n = 5$, the data is divided into 5 parts and each part is used once as test data. The overall accuracy is the average of the accuracy obtained for each case.

Leave-One-Out

This method is a special case of the n -fold method. In the n -fold method, when the number n is equal to the number of observations, the method becomes leave-one-out method. In this case, each observation is reserved for the test, while the remaining $n - 1$ observation algorithm is used for training purposes. This method is also called LOOCV (Leave-one-out cross validation).

Table 2. Accurate classification rates for selected properties from 10 to 100 with respect to leave-one-out cross validation

Data sets	Classification	Method	Number of features selected					
			10	20	40	60	80	100
Breast Cancer 49×7129	SVM	Chi-square	81.63	81.63	81.63	91.84	91.84	89.80
		IG	85.71	83.67	87.76	89.80	89.80	93.88
		ReliefF	75.51	73.47	73.47	71.43	73.47	79.59
		FSIQR	83.67	85.71	81.63	91.84	87.75	93.88
	kNN	Chi-square	91.84	91.84	93.88	85.71	83.67	85.71
		IG	89.80	93.88	89.80	89.80	91.84	87.76
		ReliefF	77.55	73.47	71.43	75.51	75.51	71.43
		FSIQR	87.76	85.71	81.63	89.80	79.60	89.80
NCI-60 Cancer Cells 50×10100	DVM	Chi-square	84.00	88.00	86.00	88.00	90.00	90.00
		IG	84.00	80.00	84.00	84.00	88.00	94.00
		ReliefF	84.00	82.00	78.00	84.00	74.00	80.00
		FSIQR	86.00	86.00	84.00	86.00	92.00	88.00
	kNN	Chi-square	86.00	86.00	86.00	84.00	84.00	86.00
		IG	86.00	86.00	90.00	88.00	90.00	88.00
		ReliefF	78.00	74.00	78.00	80.00	78.00	74.00
		FSIQR	82.00	92.00	88.00	86.00	92.00	92.00
CNS Tumor 60×7128	DVM	Chi-square	80.00	68.33	66.67	68.33	73.33	86.67
		IG	70.00	80.00	71.67	76.67	81.67	88.33
		ReliefF	73.33	73.33	76.67	68.33	73.33	61.67
		FSIQR	81.67	80.00	86.67	83.33	83.33	88.33
	kNN	Chi-square	75.00	65.00	73.33	75.00	71.67	73.33
		IG	75.00	76.67	75.00	75.00	73.33	83.33
		ReliefF	76.67	71.67	68.33	71.67	71.67	68.33
		FSIQR	81.67	78.33	78.33	80.00	81.67	81.67

According to Table 2;

- In the Breast Cancer data set, the highest accuracy rate for SVM is 93.88% and this accuracy is achieved when 100 features of IG and FSIQR methods are selected. In addition, the FSIQR method achieves the highest accuracy for 20 and 60 features selected.

- In the NCI-6 Cancer Cells data set, the highest accuracy rate for kNN is 92%, and this accuracy is achieved only when the FSIQR method has selected 20, 80 and 100 features, respectively. In cases where 10, 40 and 60 features are selected, the IG method reached the highest accuracy rate.
- In the CNS Tumor dataset, it is seen that FSIQR method gives the highest accuracy rates for both SVM and kNN (except when 100 features are selected).

Table 3. Accurate classification rates for selected features from 10 to 100 based on 5-fold cross validity

Data sets	Classification	Method	Number of features selected					
			10	20	40	60	80	100
Breast Cancer 49×7129	SVM	Chi-square	83.56	89.78	83.78	92.00	91.78	88.00
		IG	85.56	88.00	85.78	89.78	89.78	91.78
		ReliefF	83.78	79.56	69.56	83.78	91.78	83.33
		FSIQR	81.78	83.78	81.78	92.00	92.00	94.00
	kNN	Chi-square	87.78	89.78	93.78	82.00	81.33	83.77
		IG	86.00	85.11	92.00	89.78	85.78	87.56
		ReliefF	73.56	74.89	67.56	75.33	75.78	78.00
		FSIQR	87.78	83.33	81.33	85.78	85.78	90.00
NCI-60 Cancer Cells 50×10100	SVM	Chi-square	87.88	85.88	83.66	87.70	85.52	89.96
		IG	87.52	85.88	84.28	88.36	89.92	96.08
		ReliefF	66.24	75.88	78.36	79.88	88.36	81.88
		FSIQR	96.18	87.29	84.46	83.96	91.74	87.70
	kNN	Chi-square	84.06	78.10	88.50	86.36	84.73	84.32
		IG	86.14	82.65	88.51	88.00	89.70	88.10
		ReliefF	76.24	75.92	77.62	72.10	80.65	82.14
		FSIQR	84.73	87.96	87.96	88.28	91.96	91.96
CNS Tumor 60×7128	SVM	Chi-square	78.33	68.33	68.65	66.67	75.00	81.64
		IG	71.47	79.98	75.19	74.76	81.67	83.13
		ReliefF	71.42	53.33	58.22	56.67	69.65	66.72
		FSIQR	83.69	74.78	85.00	81.16	86.67	86.47
	kNN	Chi-square	73.44	68.52	68.24	73.33	74.93	73.33
		IG	73.33	76.19	78.33	78.95	78.31	81.32
		ReliefF	66.45	67.98	65.00	68.24	66.67	71.88
		FSIQR	80.00	75.36	78.44	77.15	80.08	81.77

According to Table 3;

- In the Breast Cancer data set, the highest accuracy rate for SVM is 94% and this accuracy rate is reached when FSIQR method is selected 100 features. In addition, the proposed method is superior to other methods when selecting features 60 and 80 and achieves the second highest accuracy rate.
- In the NCI-6 Cancer Cells data set, the highest accuracy rate for kNN is 91.96% and this accuracy rate is only achieved when FSIQR method 80 and 100 features is selected. Furthermore, in cases where 20 and 60 features are selected, the proposed method has a higher accuracy rate than others.
- In the CNS Tumor dataset, the FSIQR method has a higher accuracy rate than 10, 40, 60, 80 and 100 for SVM, and 10, 40, 80 and 100 for kNN compared to other methods.

5. Conclusion

In this study, a new filtering method based on interquartile range value (FSIQR) was proposed for feature selection in gene expression data. In study, FSIQR method, Chi-square, Information Gain, and ReliefF feature selection methods were compared with three real gene data. According to the results, FSIQR was able to select the features that reached higher accuracy in many cases compared to other methods. According to these results, we recommend using FSIQR algorithm as a filtering method in feature selection.

References

- Altunkaynak, B. (2019). Veri Madenciliği Yöntemleri ve R Uygulamaları (2 ed.): Seçkin Yayıncılık.
- Chandra, B., & Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, 44(4), 529-535.
- Chih-Wei, H., & Chih-Jen, L. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415-425. doi:10.1109/72.991427
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297. doi:10.1023/a:1022627411411
- Giordani, I. (2009). Relational Clustering for Knowledge Discovery in Life Sciences. (PhD), University of Milano-Bicocca, Milan, Italy.
- John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In W. W. Cohen & H. Hirsh (Eds.), *Machine Learning Proceedings 1994* (pp. 121-129). San Francisco (CA): Morgan Kaufmann.
- Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., & Talloen, W. (2016). *Applied Biclustering Methods for Big and High-Dimensional Data Using R*: Chapman & Hall/CRC.
- Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In D. Sleeman & P. Edwards (Eds.), *Machine Learning Proceedings 1992* (pp. 249-256). San Francisco (CA): Morgan Kaufmann.
- Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., . . . Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415, 436. doi:10.1038/415436a
- Quinlan, J. R. (1986). Induction of decision trees. 1(1), 81-106. doi:10.1007/bf00116251
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545-15550. doi:10.1073/pnas.0506580102
- Vicente, R. S. (2009). Visual Analysis of Gene Expression Data by Means of Biclustering. (PhD), University of Salamanca, Salamanca.
- West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., . . . Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. 98(20), 11462-11467. doi:10.1073/pnas.201162998 %J *Proceedings of the National Academy of Sciences*.

O-29 RA-CUSUM chart based on LR-fuzzy data

Afsaneh Rezaeifar¹, Bahram Sadeghpour Gildeh^{2*} and G.R. Mohtashami Borzadaran³

¹Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran, afsaneh.Rezaeifar@mail.um.ac.ir

²Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran, sadeghpour@um.ac.ir

³Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran, grmohtashami@um.ac.ir

(Corresponding author should have an asterisk sign (*) if possible.)

Abstract – The cumulative sum chart is widely used in manufacturing processes, but in medical science and particularly for monitoring the performance of a cardiac surgeon or a group of surgeons, the patient’s preoperative risk should be taken into account. So risk-adjusted charting procedures have gained attention. A risk-adjusted cumulative sum (RA-CUSUM) chart based on testing the odds of mortality was proposed in 2000, which takes the preoperative risk of patients into account. Since the preoperative risk is a vague and non-precise variable and the anesthesiologists after checking how many risk factors a patient has, determine the risk of mortality before the surgery as a linguistic term such as low, medium, high or others like that, it is better to be considered as a fuzzy number, which can be determined by using the fuzzy logistic regression model. In this condition, we need a special chart to monitor the performance of surgeons based on these fuzzy data. In this paper we proposed the RA-CUSUM chart based on LR-fuzzy data and then consider the conclusion on real data.

Keywords – Preoperative risk; Surgical outcomes; Monitoring performance; RA-CUSUM; Fuzzy logistic regression.

1. Introduction

The cumulative sum (CUSUM) chart methodology was firstly developed by Page (1954) for monitoring manufacturing processes. Since this chart is sensitive to small changes (Montgomery (2007)), it is interested in researches. Application of the CUSUM for monitoring surgical performance was first proposed by Williams et al. (1992) but in that paper the covariate information such as patient characteristics, surgical team and others did not take into account.

In this situation all patients were assumed to have the same surgical risk. However, in the field of medical science, patients have different clinical presentations and prior risk. So the risk-adjusted charts for monitoring the performance of surgeons should be used.

Lovegrove et al. (1997, 1999) and Poloniecki et al. (1998) suggested the first charting procedure which was based on the difference between the preoperative risks and the surgical outcomes. This chart is named variable life-adjusted display (VLAD) that takes the risk of a patient into account but the problem of that is about lacking proper signaling rule. Steiner et al. (2000) proposed a risk-adjusted cumulative sum (RA-CUSUM) chart based on testing the odds of mortality that a patient dies. They showed that trainee surgeons were performing better than experienced surgeons on the standard CUSUM (without taking the risk of a patient into account) while the RA-CUSUM chart showed the opposite. They used

data set from the UK center for cardiac surgery and used Parsonnet score (Parsonnet et al. (1989)) to estimate the preoperative risk of patients by the logistic regression model.

Sherlaw-Johnson (2005) mapped the control limits of the RA-CUSUM chart onto the VLAD but the resulting signaling rule is complicated because the control limits change with the inclusion of data from every new surgical operation.

Gan et al. (2012) proposed a general RA-CUSUM chart in which the RA-CUSUM (proposed by Steiner in 2000) was a special case of that. They also assessed the sensitivities of the RA-CUSUM chart with respect to changes in the underlying risk distribution. The distribution was supposed to be beta families, but in real data the risk can have other distributions. The performance of the RA-CUSUM chart is explored in lots of articles such as Axelrod et al. (2009), Morton et al. (2008) and Chen et al. (2011). Grigg and Farewell (2004) and Woodall et al. (2015) provided an overview of risk-adjusted monitoring that includes the RA-CUSUM and some other methods and concluded that in most circumstances the RA-CUSUM is preferred.

In the Steiner model (2000) the risk was modeled by the logistic regression for which the response variable is that patient survival (alive/death). However in reality, we need to evaluate the risk based on the anesthesiologist decision before the surgery and just by knowing the risk factors for each patient. Actually, anesthesiologists after checking how many risk factors a patient has, determine the risk of mortality before the surgery as a linguistic term such as low, medium, high or others like that (Miller et al. (2014)). So, the preoperative risk is a vague and non-precise variable and can be considered as a fuzzy number. In this situation we need a fuzzy model to estimate the fuzzy risk. The best description of these kinds of observations is that they are fuzzy outputs.

There are several fuzzy regression models (see Buckley et al. (1999)). Here we have a nonlinear and fuzzy categorical output and crisp explanatory variables, so we use the fuzzy logistic regression based on the least squares approach proposed by Pourahmad et al. (2011). This model is appropriate to estimate fuzzy risk. There are useful properties for LR-fuzzy numbers, which make arithmetics more comfortable, besides that there are some methods for generating LR-fuzzy numbers for characteristics. Hence we assume that the risk of mortality is an LR-fuzzy number.

After risk determination, we need a special chart to monitor the performance of surgeons based on these LR-fuzzy data. Erginel and Şentürk (2016) proposed the fuzzy EWMA and fuzzy CUSUM control chart. They used the fuzzy median (midrange) transformation technique, which is integrated into the α -cut fuzzy median (midrange) but the chart is plotted with representative values of the fuzzy data. These representative values result in losing important information included in the original data. The CUSUM control chart for LR-fuzzy data was proposed by Wang and Hryniewicz (2013). Their model was derived from the standard CUSUM, which is designed based on a normal distribution. However it is not easy to determine the distribution of a fuzzy variable and it can be no normal. Inspiration by that, we propose the RA-CUSUM chart based on LR-fuzzy data in this paper.

In Section 2, we review some fuzzy definitions. In Section 3, the RA-CUSUM chart based on LR-fuzzy data is considered and in Section 4, we present the use of the RA-CUSUM chart based on LR-fuzzy data with a real data example.

2. Fuzzy logistic regression

In this section, we give a brief review on some of the fuzzy set theory definitions (given in Zimmermann (1991)) and fuzzy logistic regression.

Definition 2.1 Let E denotes a function space, such that $u \in E$ if and only if $u: \mathbb{R} \rightarrow [0,1]$ is a function which satisfies the following requirements:

- (i) normality: $u(x_0) = 1$ for some $x_0 \in \mathbb{R}$;
- (ii) u is a convex, that is,

$$u(\lambda x + (1 - \lambda)y) \geq \min\{u(x); u(y)\}, x, y \in \mathbb{R}, \lambda \in [0,1],$$

(iii) u is upper semicontinuons, that is,

$$\limsup_{x \rightarrow t} f(x) = f(t), \quad t \in \mathbb{R}.$$

(iv) $(u)_0 = \text{closure}\{t | t \in \mathbb{R}, u(t) > 0\}$ is compact.

The space E is called a fuzzy number space, each $u \in E$ is called a fuzzy number and x_0 is the core of u .

Definition 2.2 Let $u \in E$ and let $\alpha \in [0,1]$. Then

$$(u)_\alpha = \{t | u(t) \geq \alpha\}.$$

is called an α -cut or α -level sets of u .

Note: An especial case of fuzzy numbers, called an LR-fuzzy number is represented by

$$u(t) = \begin{cases} L\left(\frac{m-t}{a}\right), & t \leq m, \quad a > 0 \\ R\left(\frac{t-m}{b}\right), & t > m, \quad b > 0 \end{cases} \quad (1)$$

where L and R are non-increasing nonnegative real-valued functions defined over $[0,+\infty)$ which satisfy $L(0) = R(0) = 1$.

If $L(t) = R(t) = T(t)$, where $T(t) = \begin{cases} 1-t, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$

then $u(t)$ in (1) is a triangular fuzzy number. It is symbolically denoted by $(m, a, b)_T$. Also m is the mean value of u and a, b ($a > 0, b > 0$) are called the left and right spreads, respectively.

If $a = b$, then $u(t)$ is called a symmetric triangular fuzzy number and is denoted by $(m, s)_T$ symbolically.

Proposition 2.3 (LR-fuzzy number arithmetic) Let $M = (m, a, b)_{LR}$ and $N = (n, c, d)_{LR}$ be two fuzzy numbers and let $\lambda \in \mathbb{R}$. Then

- (i) $\lambda \otimes (m, a, b)_{LR} = \begin{cases} (\lambda m, \lambda a, \lambda b)_{LR}, & \lambda > 0, \\ (\lambda m, -\lambda b, -\lambda a)_{LR}, & \lambda < 0. \end{cases}$
- (ii) $M \oplus N = (m + n, a + c, b + d)_{LR}$
- (iii) $M \ominus N = (m - n, a + d, b + c)_{LR}$
- (iv) Let M, N be two positive LR-fuzzy numbers (defined on \mathbb{R}^+). Then
- (v)

$$\frac{M}{N} \simeq \left(\frac{m}{n}, \frac{dm + an}{n^2}, \frac{cm + bn}{n^2} \right)_{LR}.$$

The proposed division can be an LR-fuzzy number or not.

As mentioned in section 1, the risk of mortality is a linguistic term such as low, medium, high or others like that which are assigned by experts and is better to be considered as a fuzzy output. As a result we need a suitable model to estimate the risk when the output variable is fuzzy. Because the explanatory variables are crisp, we use fuzzy logistic regression based on the least squares method (Williams et al. 1992).

Due to the vague status of cases relative to the response categories, the probability of mortality cannot be calculated and modeled exactly based on explanatory variables, and so the odds of mortality are meaningless. In this situation, we should consider the possibility of mortality instead of probability. The possibilistic odds is defined as follows.

Definition 2.4 (Possibilistic Odds) Let \tilde{p}_n , $n = 1, 2, 3, \dots$, be the possibility of success for the n th patient, $\tilde{p}_n = \text{Poss}(Y_n \approx 1)$. It can be defined in two manners:

- (i) A real crisp value, $\tilde{p}_n \in \mathbb{R}$, $0 \leq \tilde{p}_n \leq 1$ or
- (ii) A linguistic term $\tilde{p}_n \in \{\dots, \text{low}, \text{medium}, \text{high}, \dots\}$. Then terms should be defined in such a way

that the union of their supports cover the whole range of $(0,1)$. Thus the ratio $\frac{\tilde{p}_n}{1-\tilde{p}_n}$ is considered as possibilistic odds of the n th case which detects the possibility of success relative to the possibility of non success.

(iii) Let the expert assign the linguistic term as $\tilde{p}_n = \{\dots, \text{low}, \text{medium}, \text{high}, \dots\}$, for $n = 1, 2, 3, \dots$ where \tilde{p}_n is the possibility of mortality for the n th patient and the observed outputs. Defining a suitable membership function for that is very important. After consulting with some experts and based on Miller’s Anesthesia reference book (Miller et al. (2014)) in the preoperative evaluation section, it was described as a triangular fuzzy number as follows:

$$\text{Very low}(x) = \begin{cases} 1 - \frac{0,004 - x}{0,004}, & 0 \leq x \leq 0,004, \\ 1 - \frac{x - 0,004}{0,005}, & 0,004 < x \leq 0,009, \end{cases}$$

$$\text{low}(x) = \begin{cases} 1 - \frac{0,009 - x}{0,005}, & 0,004 \leq x \leq 0,009, \\ 1 - \frac{x - 0,009}{0,061}, & 0,009 < x \leq 0,07, \end{cases}$$

$$\text{Medium}(x) = \begin{cases} 1 - \frac{0,07 - x}{0,061}, & 0,009 \leq x \leq 0,07, \\ 1 - \frac{x - 0,07}{0,04}, & 0,07 < x \leq 0,11, \end{cases}$$

$$\text{High}(x) = \begin{cases} 1 - \frac{0,11 - x}{0,04}, & 0,07 \leq x \leq 0,11, \\ 1 - \frac{x - 0,11}{0,889}, & 0,11 < x \leq 0,999, \end{cases}$$

$$\text{Very High}(x) = \begin{cases} 1 - \frac{0,999 - x}{0,889}, & 0,11 \leq x \leq 0,999, \\ 1 - \frac{x - 0,999}{0,001}, & 0,999 < x \leq 1. \end{cases}$$

Then by using the fuzzy logistic regression when the response variable (\tilde{p}_n) is fuzzy, the risk can be estimated by the following model:

$$\text{Logit}(\tilde{P}_n) = \tilde{\beta}_0 + \sum_{i=1}^m \tilde{\beta}_i u_{ni} \quad i = 1, 2, \dots, m, \quad (2)$$

where u_{ni} is the i th explanatory variable value for the n th patient. Thus $\tilde{\beta}_0 = (a_0, s_0)_T$ is intercept and $\tilde{\beta}_i = (a_i, s_i)_T$ is the i th coefficient of the model, where $i = 1, 2, \dots, m$. Also, $\tilde{\beta}_i$ s are treated as a fuzzy number, and $\text{Logit}(\tilde{P}_n)$ is the estimator of the logarithmic transformation of possibilistic odds and is equal to

$$\text{Logit}(\tilde{P}_n) = (f_n(a), f_n(s))_T, \quad (3)$$

where

$$f_n(a) = a_0 + \sum_{i=1}^m a_i u_{ni}, \quad (4)$$

$$f_n(s) = s_0 + \sum_{i=1}^m s_i u_{ni}, \quad (5)$$

Using the least squares method, the variables a_0, a_1, \dots, a_m and s_0, s_1, \dots, s_m are estimated, and then using the extension principle, $\tilde{P}_n = (m_n, l_n, r_n)_T$ in (2) as an LR-fuzzy number with l_n left width and r_n right width is obtained by:

$$m_n = \frac{\exp(f_n(a))}{1 + \exp(f_n(a))},$$

$$l_n = \frac{\exp(f_n(a))}{1 + \exp(f_n(a))} - \frac{\exp(f_n(a) - f_n(s))}{1 + \exp(f_n(a) - f_n(s))},$$

$$r_n = \frac{\exp(f_n(a) + f_n(s))}{1 + \exp(f_n(a) + f_n(s))} - \frac{\exp(f_n(a))}{1 + \exp(f_n(a))}.$$

To obtain the goodness-of-fit between observed values and the estimated values, the level of closeness for them should be determined.

Definition 2.5 Suppose $u, v \in E$. Then, the level of closeness for u and v is defined by (Gildeh and Gien (2002))

$$I_{UI} = \min \left(\frac{\text{Card}(u \cap v)}{\text{Card}(u)}, \frac{\text{Card}(u \cap v)}{\text{Card}(v)} \right),$$

where

$$\text{Card}(u) = \begin{cases} \int u(t)dt, & \text{continuous case,} \\ \sum u(t), & \text{discrete case.} \end{cases}$$

Moreover I_{UI} is observed that $0 \leq I_{UI} \leq 1$.

DEFINITION 2.6. For fuzzy logistic regression, the mean of I_{UI} is used as a measure of evaluating model's goodness-of-fit

$$\text{MCI} = \frac{1}{m} \sum_{i=1}^m I_{UI}(\tilde{p}_n, \tilde{P}_n).$$

In which $0 \leq \text{MCI} \leq 1$. So that the larger MCI corresponds with better goodness-of-fit.

In the following section, the mentioned control charts based on LR-fuzzy data are considered.

3. RA-CUSUM chart based on LR-fuzzy data

For the RA-CUSUM chart based on odds of mortality developed by Steiner et al. (2000), the hypothesis tests are considered as

$$\begin{cases} H_0: \text{odds of mortality for } n\text{th patient} = Q_0 \frac{p_n}{1-p_n} \\ H_A: \text{odds of mortality for } n\text{th patient} = Q_A \frac{p_n}{1-p_n} \end{cases} \quad (6)$$

where p_n is the risk of mortality for n patient. The ratio Q_0 is usually supposed to be one. It means under H_0 the odds of mortality equals what is expected by the risk model, while H_A corresponds to a worsening or improvement of performance. If we are searching for the deterioration (improvement) of the surgeon's performance, then Q_A should be more than one (less than one).

The RA-CUSUM chart for testing these hypotheses is obtained by plotting

$$S_n = \max\{0, S_{n-1} + W_n\} \quad (7)$$

against n , where

$$W_n = \begin{cases} \log \frac{(1 - p_n + Q_0 p_n) Q_A}{(1 - p_n + Q_A p_n) Q_0}, & \text{if the } n\text{th patient dies,} \\ \log \frac{1 - p_n + Q_0 p_n}{1 - p_n + Q_A p_n}, & \text{if the } n\text{th patient survives.} \end{cases}$$

The patient's score W_n is the logarithm of the likelihood ratio of densities. It can be viewed as a penalty-reward score given to the surgeon depending on the patient's preoperative risk and the outcome of the operation. Such that, if a patient dies, the penalty will be heavy if the patient's preoperative risk is low, and the penalty will be small if the patient's preoperative risk is high. Similarly, if a patient survives, the reward will be big if the patient's preoperative risk is high and the reward will be small if the patient's preoperative risk is low.

A signal is issued when $S_n > h$, where h is the upper control limit.

Considering the risk as a linguistic term, then the hypothesis (6) can be rewritten as

$$\begin{cases} H_0: \text{possibilistic odds of mortality for } n\text{th patient} = Q_0 \frac{\tilde{p}_n}{1 - \tilde{p}_n} \\ H_A: \text{possibilistic odds of mortality for } n\text{th patient} = Q_A \frac{\tilde{p}_n}{1 - \tilde{p}_n} \end{cases}$$

By using the fuzzy logistic regression, the fuzzy risk $\tilde{P}_n = (m_n, l_n, r_n)_T$ can be estimated. Then under H_0 , the possibilistic odds is an LR-fuzzy number as

$$\text{possibilistic odds} = \left(\frac{Q_0 m_n}{1 - m_n}, \frac{Q_0 l_n}{(1 - m_n)^2}, \frac{Q_0 r_n}{(1 - m_n)^2} \right)_T.$$

Therefore

$$\tilde{S}_n(s) = \max_{s=\max\{a,y\}} \{0(a), (\tilde{S}_{n-1} \oplus \tilde{W}_n)(y)\}, \quad s \in \mathbb{R}, \quad (8)$$

where 0 denotes an LR-fuzzy number with 0 width and

$$\tilde{W}_n = \begin{cases} \log \frac{(1 \ominus \tilde{p}_n \oplus Q_0 \odot \tilde{p}_n) \odot Q_A}{(1 \ominus p_n \oplus Q_A \odot \tilde{p}_n) \odot Q_0}, & \text{if the } n\text{th patient dies,} \\ \log \frac{1 \ominus \tilde{p}_n \oplus Q_0 \odot \tilde{p}_n}{1 \ominus \tilde{p}_n \oplus Q_A \odot \tilde{p}_n}, & \text{if the } n\text{th patient survives,} \end{cases} \quad (9)$$

Then, \tilde{W}_n is a fuzzy number and the α -cut of that is defined as,

$$W_n^-(\alpha) = \begin{cases} \log \left(\frac{Q_A}{(1 - m_n + Q_A m_n)} (C_1(\alpha - 1) + 1) \right), & \text{if the } n\text{th patient dies,} \\ \log \left(\frac{1}{(1 - m_n + Q_A m_n)} (C_1(\alpha - 1) + 1) \right), & \text{if the } n\text{th patient survives,} \end{cases}$$

$$W_n^+(\alpha) = \begin{cases} \log \left(\frac{Q_A}{(1 - m_n + Q_A m_n)} (C_2(1 - \alpha) + 1) \right), & \text{if the } n\text{th patient dies,} \\ \log \left(\frac{1}{(1 - m_n + Q_A m_n)} (C_2(1 - \alpha) + 1) \right), & \text{if the } n\text{th patient survives.} \end{cases}$$

where $\alpha \in [0,1]$ and

$$C_1 = \frac{(l_n + Q_A r_n) + (l_n + r_n)(1 - m_n + Q_A m_n)}{(1 - m_n + Q_A m_n)},$$

$$C_2 = \frac{(r_n + Q_A l_n) + (l_n + r_n)(1 - m_n + Q_A m_n)}{(1 - m_n + Q_A m_n)}.$$

Therefore,

$$S_n^-(\alpha) = \max\{0, S_{n-1}^-(\alpha) + W_n^-(\alpha)\},$$

$$S_n^+(\alpha) = \max\{0, S_{n-1}^+(\alpha) + W_n^+(\alpha)\}.$$

To detect that the process is in control or not, the α -cut set of the control limit needs to be specified. To find out that, firstly the distribution of risk (m_n), left width (l_n), and right width (r_n) are needed and then by knowing the first type error and using the simulation method, it is determined as $[h^-(\alpha), h^+(\alpha)]$.

As a result, we have two intervals $[S_n^-(\alpha), S_n^+(\alpha)]$ and $[h^-(\alpha), h^+(\alpha)]$ which are named S_n and H , respectively. If the interval S_n is smaller (larger) than the interval H and they do not have any intersection, then the process is in (out of) control. Otherwise, having an intersection makes the decision complicated. Thereupon, we need a measure that is suitable for all of the conditions which can happen.

We use the $D_{p,q}$ -distance with some differences (Gildeh and Gien(2002)). The distance measures how far two fuzzy numbers are from each other. The analytical properties of $D_{p,q}$ depend on the parameter p , while the parameter q characterizes the subjective weight attributed to the sides of the fuzzy numbers. In the definition of $D_{p,q}$ -distance, we need to integrate all the values of $\alpha \in [0,1]$, but since the chart is plotted for a specific value of α , we calculate the distance based on that. Therefore to compare S_n and H , we compute the distance between S_n and C and the distance between H and C , where C is a constant interval, i.e. $C = (C^-, C^+)$.

$$d_{S_n} = D_{p,q}(S_n, C) = \left((1-q) |S_n^-(\alpha) - C^-|^p + q |S_n^+(\alpha) - C^+|^p \right)^{1/p}, \quad (10)$$

$$d_H = D_{p,q}(H, C) = \left((1-q) |h^-(\alpha) - C^-|^p + q |h^+(\alpha) - C^+|^p \right)^{1/p}, \quad (11)$$

where $p \in [1, +\infty)$ and $q \in [0, 1]$. Thus, if $d_{S_n} > d_H$, then the process is out of control.

To evaluate a control chart, the average run length (ARL) is used (Montgomery (2007)). The ARL is the expected number of points that are observed until a signal takes place. The ARL that results when the process remains in control, denoted by ARL_0 , is used to determine the control limits. The ARL that results when the process is out of control, denoted by ARL_1 , is a performance metric that can be used to compare one control chart with another. There are different ways such as the simulation or Markov chains to calculate or approximate the ARL. Both provide approximations of the ARL. Simulation may be necessary if the calculation of transition probabilities between the Markov chains states is sufficiently complicated or intractable.

4. Application to real data

The data set contains outcomes of a patient with heart surgery and is based on 6994 operations, from a single surgical center over the seven-year period, 1992-1998, which were used by Steiner et al. (2000). The data consists of some information on each patient like a surgeon, type of procedure and the pre-operative variables, which comprise the Parsonnet score (The score is based on a combination of many other information like age, gender, hypertension and others).

In the data, 461 deaths occurred within 30 days of surgery, giving an overall mortality rate of 6.6%.

To identify the risk factors in phase I, we used the first two years of data (1992-1993) and began the monitoring from 1994, in phase II.

In the first two years, a total of 2218 surgeries were performed and 143 deaths were observed (mortality rate of 6.5%). The fuzzy logistic regression is used to estimate risk factors. As mentioned in section 2, $\text{Logit}(\tilde{P}_n) = (f_n(a), f_n(s))_T$ is the estimated output in this model. To reach that for the data, because of

not accessing to the result of experts decision as the possibility of mortality, we use the Buckley approach (2006) based on the fuzzy probabilities from a confidence interval. Then, by using the least squares method, $f_n(a)$ and $f_n(s)$ which are defined in (4) and (5), are obtained as follows:

$$f_n(a) = -3.528 + 0.0554 u_n,$$

$$f_n(s) = 0.1834 + 0.000014 u_n.$$

where u_n denotes the Parsonnet score for patient n , $n= 1, 2, \dots, 2218$. To evaluate the model, we used MCI, which is defined in definition (2.6).

$$MCI = \frac{1}{2218} \sum_{i=1}^{2218} I_{UI}(\check{P}_n, \tilde{P}_n) = 0.72.$$

Assume that we are interested in designing an RA-CUSUM chart which is optimal in detecting a deterioration in the performance with the possibilistic odds of mortality risk increased to two times the in-control possibilistic odds, that is $Q_A = 2$.

Therefore \tilde{W}_n in (8) and its α -cut for intended α can be calculated to reach d_{S_n} in (10). Since there is no reason to distinguish any side of fuzzy numbers, we use $D_{p,q}$ -distance when $p = 1$ and $q = 0.5$. Assume that that $ARL_0 = 10000$. the first and the end points of the interval H (the α -cut control limit) and the ARL_1 for $\alpha \in \{0.85, 1\}$ can be seen in Table1. To determine the control limits, as mentioned in the previous section, we used the distribution of core (m_n) and width (l_n and r_n) of fuzzy risks and the simulation method. Decreasing α increases the ambiguity, so we decided not to lower it than 0.85 and compare it when $\alpha=1$ (When W_n is crisp).

Table 1: The end points of α -cut control limits and ARL_1 (α) for the RA-CUSUM chart based on LR-fuzzy data

α	$h^-(\alpha)$	$h^+(\alpha)$	$ARL_1(\alpha)$
0.85	3.8413	7.0494	256
1	4.8085	4.8085	212

Figure1 shows the RA-CUSUM control chart based on LR-fuzzy data. When $\alpha = 1$, there are some out of control points in the chart. In general, the signal was not sustained and the process returned to its in-control level. Decreasing α has different interpretations. The $\alpha = 0.85$ in Figure1, means that we considered the α -cut set of \tilde{W}_n scores with this intended α instead of considering the scores as the crisp numbers (while the risks are fuzzy). In this case, the process becomes in control. Actually, there are some out of control points in crisp control chart, but by considering $\alpha = 0.85$, based on the position of side of fuzzy numbers and the value of $D_{p,q}$, they can become in control. So, decreasing α can make the chart more in control.

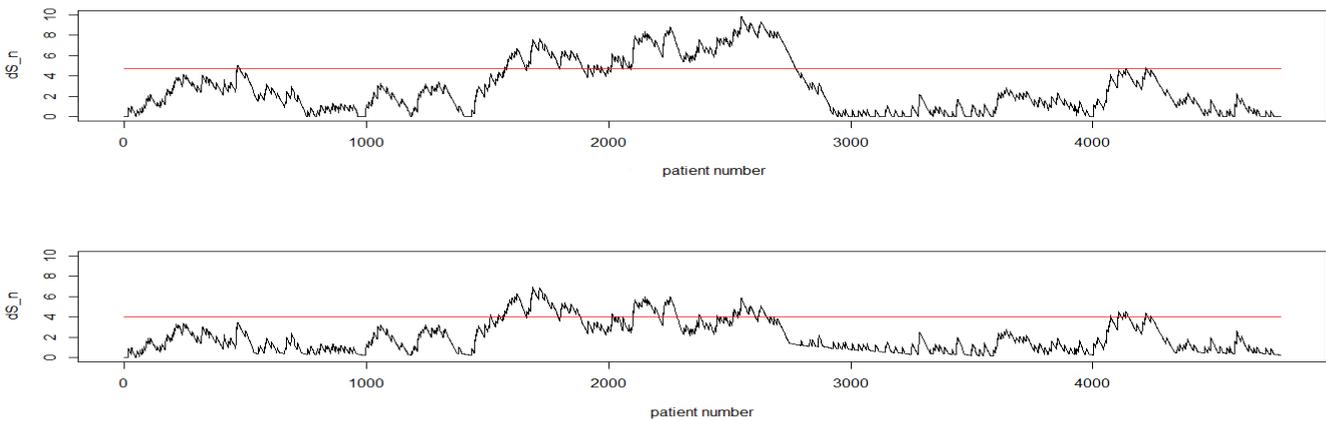


Figure 1: The RA-CUSUM control chart based on LR-fuzzy data when $\alpha=1$ (The above chart) and $\alpha=0.85$ (The latter chart)

5. Conclusions

Since the preoperative risk of surgery is a nonprecise and vague variable and a linguistic term such as low, medium, high or others like that, which are assigned by experts, it is better to be considered as a fuzzy number and determined by an appropriate membership function. As for fuzzy concepts, instead of a crisp value, we can consider numbers around it with less membership grade. From the perspective of α -cut, it can be considered as a short interval for each α , ($\alpha \in [0,1]$). The fuzzy preoperative risk can be determined by using a fuzzy model such as the fuzzy logistic regression model. In this case, we need a special chart to monitor the performance of the surgeons which its statistics and control limits are fuzzy numbers. For this purpose, we proposed the RA-CUSUM control chart based on LR-fuzzy data. The score used to build the chart statistics (\tilde{W}_n) is a fuzzy number and by considering its α -cut, we can have a short interval for each defined α (i.e. $[W_n^-(\alpha), W_n^+(\alpha)]$). Then by calculating the control statistics interval and their corresponded control limits interval and using the $D_{p,q}$ -distance, the chart based on LR-fuzzy data can be plotted. If $\alpha = 1$, then the behavior of the chart is the same as its corresponded crisp chart, but decreasing α or considering the \tilde{W}_n score in a more fuzzy environment, the result of the chart changes. Because of the flexibility implied in a fuzzy control process, the $ARL(\alpha)$ is almostly a decreasing function of α . So, It is better not taking α low.

References

- Axelrod, D.A., Kalbfleisch, J. D., Sun, R. J., Guidinger, M. K., Biswas, P., Levine, G. N., Arrington, C. J., Merion, R. M. (2009). “Innovations in the Assessment of Transplant Center Performance: Implications for Quality Improvement”, American Journal of Transplantation, vol. 9, no. (4Pt 2), pp. 959–969.
- Buckley, J. J., Feuring, T., Hayashi, Y. (1999). “Multivariate non-linear fuzzy regression: an evolutionary algorithm approach”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 7, no. 2, pp. 83–98.
- Buckley, J. J. (2006). “Fuzzy Probability and Statistics”, Springer.
- Chen, T. T., Chung, K. P., Hu, F. C., Fan, C. M., Yang, M. C. (2011). “The use of Statistical Process Control (Risk-Adjusted CUSUM, Risk-Adjusted RSPRT and CRAM with prediction Limits) for

- monitoring the outcomes of out-of-hospital cardiac arrest patient rescued by EMS system”. *Journal of Evaluation in Clinical Practice*, vol. 17, no. 1, pp.71–77.
- Erginel, N., Şentürk, S. (2016). “Fuzzy EWMA and Fuzzy CUSUM Control Charts”, *Fuzzy Statistical Decision-Making*. Springer, Cham, pp. 281-295.
- Gan, F. F., Lin, L., Loke, C. K. (2012). “Risk-adjusted cumulative sum charting procedures”, *Frontiers in Statistical Quality Control*, vol. 10, pp. 207–225.
- Gildeh, B. S., Gien, D. (2002). “A goodness of fit index to reliability analysis in fuzzy model”, 3rd WSEAS international conference on fuzzy sets and fuzzy systems, Interlaken, Switzerland, pp. 11–14.
- Grigg, O., Farewell, V. (2004). “An overview of risk-adjusted charts”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 167, no. 3, pp. 523–539.
- Lovegrove, J., Sherlaw-Johnson, C., Valencia, O., Treasure, T., Gallivan, S. (1999). “Monitoring the performance of cardiac surgeons”, *Journal of the Operational Research Society*, vol. 50, no. 7, pp. 684–689.
- Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C., Gallivan, S. (1997). “Monitoring the results of cardiac surgery by variable life-adjusted display”, *The Lancet*, vol. 350, no. 9085, pp. 1128–1130.
- Miller, R. D., Eriksson, I., Fleisher, L.A., et al. (2014). “Miller’s Anesthesia E-Book”, Elsevier Health Sciences.
- Montgomery, D. C. (2007). “Introduction to statistical quality control”, John Wiley and Sons.
- Morton, A. P., Clements, A. C., Doidge, S. R., Stackelroth, J., Curtis, M., Whitby, M. (2008). “Surveillance of healthcare acquired infections in Queensland, Australia: data and lessons learned in the first 5 years”, *Infection Control and Hospital Epidemiology*, vol. 29, no. 8, pp. 695–701.
- Page, E. S. (1954). “Continuous inspection schemes”, *Biometrika*, vol. 41, no. 1.2, pp. 100–115.
- Parsonnet, V., Dean, D., Bernstein, A. D. (1989). “A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease”, *Circulation*, vol. 79, no. (6 Pt 2), I3–12.
- Poloniecki, J., Valencia, O., Littlejohns, P. (1998). “Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery”, *Bmj*, vol. 316, no. 7146, pp. 1697–1700.
- Pourahmad, S., Ayatollahi, S. M. T., Taheri, S. M., Agahi, Z. H. (2011). “Fuzzy logistic regression based on the least squares approach with application in clinical studies”, *Computers and Mathematics with Applications*, vol. 62, no. 9, pp. 3353–3365.
- Sherlaw-Johnson, C. (2005). “A method for detecting runs of good and bad clinical outcomes on variable life-adjusted display (vrad) charts”, *Health care management science*, vol. 8, no. 1, pp. 61–65.
- Steiner, S. H., Cook, R. J., Farewell, V. T., Treasure, T. (2000). “Monitoring surgical performance using risk-adjusted cumulative sum charts”, *Biostatistics*, vol. 1, no. 4, pp. 441–452.
- urgen Zimmermann, H. J. (1991), “Fuzzy set theory| and its applications”.
- Wang, D., Hryniewicz, O. (2013). “The design of a cusum control chart for lr-fuzzy data”, *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, IEEE, pp. 175–180.
- Williams, S. M., Parry, B. R., Schlup, M. (1992). “Quality control: an application of the cusum”, *BMJ: British medical journal*, vol. 304, no. 6838, pp. 1359–1361.
- Woodall, W. H., Fogel, S. L., Steiner, S. H. (2015). “The monitoring and improvement of surgical-outcome quality”, *Journal of quality technology*, vol. 47, no. 4, pp. 383–399.

O-33 A goal programming approach for the use of restricted data envelopment analysis as a tool in multi criteria decision analysis

¹Esra Betül KINACI*, ²Harun KINACI and ³Hasan BAL

¹*Statistics, Gazi University, Turkey, esrakinaci@gazi.edu.tr*

²*Business, Erciyes University, Turkey, hkinaci@erciyes.edu.tr*

³*Statistics, Gazi University, Turkey, hasanbal@gazi.edu.tr*

Abstract – Multi-criteria decision making (MCDM) methods and Data Envelopment Analysis (DEA) allow for ranking between units using different methods on multiple feature units. The MCDM methods provide the solution of complex problems with conflicting characteristics by using the weight information related to the criteria. DEA performs this ranking through the definition of efficiency. In the classical DEA has no restrictions on these weights and weights are free. However, in the MCDM processes the weight of the properties (criteria) of the units is important. A model that will be formed considering the relative importance levels of criterion weights will contribute to the use of DEA as a tool in the MCDM methods. Also, in classical DEAs, sometimes weights of variables can be assigned a value of zero. Multi-step linear model used in DEA can be very important. Multi-step linear model used in DEA can be solved by the goal programming approach. In this study, a DEA model considering the relationship between the variable weights obtained with the MCDM method was proposed and the model is solved as a goal programming problem. As a result, a problem in which the classical DEA assigns zero weight was solved by this method.

Keywords – *Data Envelopment Analysis, Multi Criteria Decision Making, Goal Programming,*

1. Introduction

Today, with the developing technology, robot and robotic systems are used in many fields. Robots are assigned a number of important or dangerous tasks in these areas. The robots replace materials, components, tools and other special devices through control programs to complete the job. The selection of robots is a very important function for the enterprises, because the wrong selection of robots can adversely affect the profitability of the operation and its safety. There are several factors (criteria) for effectively selecting a robot. These factors can be objective or subjective. It is seen that there are many conflicting criteria which affect the decision of robot selection, such as repeatability, load capacity, speed, accuracy, utilization coefficient, program flexibility, memory capacity, supplier quality of service. These criteria can be divided into two as benefit and cost criteria. Mondal and Chakraborty (2013). Unsuitable robot selection will have negative effects on the operation. This will not only affect the productivity of the products, but also the quality and market reputation of the manufacturer. In addition, the implementation of a robot is capital intensive. For this reason, it is necessary to evaluate the effect of various selection criteria and to examine their applicability and performance carefully before production. For the measure of robot performance are used multi criteria decision making (MCDM) methods and optimization techniques in literature. Mondal and Chakraborty (2013) approach this problem with Data Envelopment Analysis (DEA). Goh et al. used the same problem MCDM techniques. Decision Analysis deals with situations where the decision maker must choose the best one among the alternatives, taking into account conflicting criteria at the same time.

The purpose of the MCDM approaches is to rank alternatives on the basis of specific criteria. Similarly, DEA performs a sorting between units using relative efficiency between units. In their study, Yılmaz and Yurdusev (2011) proposed a model that would meet the needs of the decision-maker and adapted these weights to the Data Envelopment model, taking into account the criterion weights. Bal and Örkücü (2007) analyzed DEA model with goal programming technique in order to increase the discrimination power in DEA. In this study, DEA was considered as Goal Programming problem considering the criteria weights and activity scores were calculated with the help of the obtained weights. The same problem was also solved with CODAS, one of the MCDM techniques, and the results were compared with each other.

2. Materials and Methods

2.1 DEA and DEA with Goal Programming Approach

DEA is a linear programming based method used to evaluate the relative effectiveness of decision points, responsible for producing outputs or outputs using inputs with different measurement units. The CCR model developed by Charnes, Cooper and Rhodes and the model that Li and Reeves improve to coincide with this model are presented below, respectively.

$$\begin{aligned}
 \text{Max } h_0 &= \sum_{r=1}^s u_r y_r \\
 \sum_{i=1}^m v_i x_i &= 1 \\
 \sum_{r=1}^s u_r y_r - \sum_{i=1}^m v_i x_i &\geq 0 \\
 u_r, v_i &\geq 0
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 \min d_o &\left(\text{or } \max \sum_{r=1}^s u_r y_{ro} \right) \\
 \min M \\
 \min \sum_{j=1}^n d_j \\
 \sum_{i=1}^m v_i x_{io} &= 1 \\
 \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} + d_j &= 0 \quad j = 1, 2, \dots, n \\
 M - d_j \\
 u_r &\geq 0, \quad r = 1, 2, \dots, s \\
 v_i &\geq 0, \quad i = 1, 2, \dots, m \\
 d_j &\geq 0, \quad j = 1, 2, \dots, n
 \end{aligned} \tag{2}$$

Where, u_r , weight given by the DMU to the r th output, v_i , weight given by the DMU to the r th input, y_{rj} , r th output for DMU_j, x_{ij} , i th input for DMU_j. d_j is a deviation variable for DMU_j, M is a maximum deviation variable. The larger the deviation variable, which is also bounded by interval $[0,1]$, the lesser efficient is DMU_o. The efficiency score is equal to $1-d_o$. The first objective function, $\min d_o$, is the classical DEA objective. Under the objective $\min d_o$, a DMU is efficient if and only if the value of d_o is zero. The second objective function, $\min M$, is a min max function minimizing the maximum deviation variable. The third objective function, $\min d_j$, is a min sum function minimizing the sum of the deviation variables. The constraints $M-d_j, j=1,2,\dots,n$ that define the maximum deviation M do not change the feasible region of decision variables as discussed in Li and Reeves

Bal and Örkcu Model have adapted the multi-criteria DEA model to the goal programming model as follows

$$\begin{aligned}
 \min \alpha &= \left(n_1 + p_1 + p_2, \sum_j n_{3j}, \sum_j d_j \right) \\
 \sum_{i=1}^m v_i x_{io} + n_1 - p_1 &= 1 \\
 \sum_{r=1}^s u_r y_{ro} + n_2 - p_2 &= 1 \\
 \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} + d_j &= 0 \quad j = 1, 2, \dots, n \\
 M - d_j + n_{3j} - p_{3j} &= 0, \quad j = 1, 2, \dots, n \\
 u_r &\geq 0, \quad r = 1, 2, \dots, s \\
 v_i &\geq 0, \quad i = 1, 2, \dots, m \\
 d_j &\geq 0, \quad j = 1, 2, \dots, n \\
 n_1, p_1, n_2, p_2 &\geq 0 \\
 n_{3j}, p_{3j} &\geq 0, \quad j = 1, 2, \dots, n
 \end{aligned} \tag{3}$$

Where for the DMU under evaluation, n_1 and p_1 are the unwanted deviation variables for the goal which constraints the weighted sum of inputs to unity, n_2 is the wanted deviation variable for the goal which makes the weighted sum of outputs less than or equal to unity, p_2 is the unwanted deviation variable for the goal which makes the weighted sum of out puts less than or equal to unity.

2.2 CODAS Method

The steps of the codas method are given below.

Step 1: construct the decision matrix $X = [x_{ij}]_{n \times m} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$, where $x_{ij} \geq 0$ refer to the

account value of i th alternative on j th criterion ($i \in \{1, 2, \dots, n\}$) and ($j \in \{1, 2, \dots, m\}$).

Step 2 normalized matrix is calculated. Linear normalization of performance values is used as follow,

$$n_{ij} = \begin{cases} \frac{x_{ij}}{\max_i x_{ij}} & ,if j \in N_b \\ \frac{\min_i x_{ij}}{x_{ij}} & ,if j \in N_c \end{cases} \quad (4)$$

where N_b and N_c donates the sets of benefit and cost criteria. respectively.

Step 3. the weighted normalized decision matrix is calculated as follow.

$$r_{ij} = w_j n_{ij} \quad (5)$$

Where $w_j (0 < w_j < 1)$ represent the weight of j th criterion. and $\sum_{j=1}^m w_j = 1$

Step 4. negative ideal solutions are obtained with the following equation

$$ns = [ns_j]_{1 \times m} \quad (6)$$

$$ns_j = \min_i r_{ij} \quad (7)$$

Step 5 Euclidean and Taxicab distances are calculated by the equation below.

$$E_i = \sqrt{\sum_{j=1}^m (r_{ij} - ns_j)^2} \quad (8)$$

$$T_i = \sum_{j=1}^m |r_{ij} - ns_j| \quad (9)$$

Step 6. The relative assignment matrix associated with the following equation is calculated.

$$Ra = [h_{ik}]_{n \times n} \quad (10)$$

$$h_{ik} = (E_i - E_k) + (\psi(E_i - E_k))x(T_i - T_k) \quad (11)$$

where $k \in \{1, 2, \dots, n\}$ and ψ is called a threshold function. This function is related to the Euclidean distance between the two alternatives and is defined as follows:

$$\psi = \begin{cases} 1 & if |x| \geq \tau \\ 0 & if |x| < \tau \end{cases} \quad (12)$$

where, τ is called the threshold parameter. It can be set decision-maker. This parameter takes values between 0.01 and 0.05. In this study. In our study, this parameter was used as 0.02.

Step 7. the assessment score is calculated for each alternative. shown as follows:

$$H_i = \sum_{k=1}^n h_{ik} \tag{13}$$

Step 8. The rating score of the alternatives is sorted by decreasing values H_i is the best choice among the alternatives.

3. Application

In this study, industrial robot selection problem of Chakraborty and Zavadskas (2014) was used. 7 alternative robots with 5 criteria were evaluated. The criteria were determined as repeatability (RE), load capacity (LC), maximum speed (MTS), memory capacity (MC), manipulator access (MR).in this here .LC, MTS, MC, MR are benefit variables, and RE is cost variable.

Table 1.. Rank of scores for models

DMU/ALTERNATIVES	DEA-CCR	GPDEA -CCR	GPDEA-CCR/w	CODAS
A1	1	1	3	3
A2	5	7	1	1
A3	2	4	2	2
A4	4	5	4	5
A5	6	2	7	7
A6	3	3	6	6
A7	7	6	5	4

Spearman Rho (r) correlation test was used to understand the relationship between the rankings obtained by the applied methods and results are given in table 2.

Table 1. Correlations of models

	GPDEA/w	CODAS	DEA	GPDEA
GPDEA/w	1	.964	.393	-.429
CODAS		1	.535	.294
DEA			1	.215
GPDEA				1

As can be seen, GPDEA model, which takes criterion weights into consideration, is highly correlated with MCDM technique. It can be said that the proposed GPDEA/w model can be used as an alternative tool in multi-criteria decision problems and can play an impressive role in the decision process.

4. Conclusion

In this study, the utility of DEA in multi-criteria decision making problems was investigated. In Data Envelopment Analyses, criteria weights do not contain any restrictions. However, these weights should be taken into consideration in order to use DEA in MCDM problems. A DEA problem can be solved by using goal programming method to improve the discrimination power of the model. With this approach DEA can be used as a supportive factor in the decision process in Multi Criteria Decision. We can also determine efficiency levels according to the relative weight of the criteria. As a different study, the contribution of DEA methods to the MCDM process can be increased by using other MCDM techniques and improving the constraints on the weight of the proposed HPDEA/w model

References

- Bal, H., Örkçü, H. H. (2007). “A goal programming approach to weight dispersion problem in data envelopment analysis”, *G.U. Journal of Science*, vol. 20, no.4, pp.117-125.
- Chakraborty, S., Zavadskas, E.K. (2014), “Applications of WASPAS method in manufacturing decision making”, *Informatica*, vol. 25, no. 1, pp.1-20.
- Charnes, A., Cooper, W. W., Rhodes, E. (1978). “Measuring the efficiency of decision making units”, *European Journal of Operation Research*, vol. 2, no. 6, pp.429-444.
- Li, X. B., Reeves, G. R., (1999). “A multi criteria approach to data envelopment analysis”, *European Journal of Operational Research*, vol. 115, no. 3, pp.507-517.
- Mondal, S., Chakraborty, S. (2013). “A solution to robot selection problems using data envelopment analysis”, *International Journal of Industrial Engineering Computations*, vol. 4, pp.355-372.
- Yılmaz, B., Yurdusev, M. A. (2011). “Use of data envelopment analysis as a multi criteria decision tool-a case of irrigation management”, *Mathematical and Computational Applications*, vol. 16, no. 3, pp.669-679.

O-34 Comparison of Classification Algorithms on Different Data Sets

Öznur İŞÇİ GÜNERİ¹, Burcu DURMUŞ^{2*} and Nevin GÜLER DİNCER³

¹ Department of Statistics, Muğla Sıtkı Koçman University, Turkey, oznur.isci@mu.edu.tr

² Department of Statistics, Muğla Sıtkı Koçman University, Turkey, burcudurmus@mu.edu.tr

³ Department of Statistics, Muğla Sıtkı Koçman University, Turkey, nguler@mu.edu.tr

Abstract – In this study, classification which is one of the most useful and popular data mining methods is discussed. The study investigates which algorithm has better performance over different data sets. For classification, Bayes, functions, decision tree, lazy techniques were examined. For this purpose, 50 different data sets from Machine Learning Repository were analyzed within the scope of classification. Weka Tools were used for analysis and the results of different classification algorithms were compared in terms of model performance criteria. As a result, LMT (Logistic Model Tree) and Random Forest performed best in decision tree algorithms. The performance criteria gave parallel results.

Keywords – Classification Algorithms, Data Mining, Different Data Sets, Performance Evaluation, Weka.

1. Introduction

Conceptually, data is an unprocessed form of any information recorded. Data mining can be defined as discovering and obtaining useful, previously unknown information from a population data. In other words, data mining is the process of discovering information from a large database. Data mining includes data analysis techniques such as statistics, artificial intelligence, database management systems, machine learning.

Data mining methods are divided into two as forecasting and descriptive methods. The most commonly used methods of estimation are classification and regression, and clustering and association analysis for descriptive methods. Classification is a method of assigning new information to a predefined class. The important point here is that the classes are known in advance. The clustering is to group the data in the database in a meaningful and connected manner. The association analysis examines the relationships between the data in the database and investigates which events can take place simultaneously. The association analysis is also known as ‘Basket Analysis’. Regression is the prediction of the behavior of an event with a model.

The need to analyze large data in line with the needs of today is increasing day by day. Data mining, which is used in many areas from health to education, from banking to economy, has become much more practical and meaningful with the development of computer technologies. In this respect, many software programs have been developed for data mining. WEKA, R, Knime, SPSS, MATLAB are just a few of these programs (Zupan and Demsar, 2008).

In the literature, there are many studies to compare data mining algorithms (Sewaiwar and Verma, 2015). WEKA program was preferred in the majority of these studies. In this study, more than one data set (50 different data sets) were taken into consideration and algorithm results were compared.

2. Materials and Methods

For a set of data whose classes are specific, the process of deciding which class the new member belongs to is called classification. Classification consists of two stages: training and testing. In the training stage, classification model is created by using learning data with educational data. In the test phase, the test data is applied to the model and the classes of the test data are estimated.

Data mining methods can be used in almost any field. Amin et al. (2013) conducted a study with MATLAB for predicting heart disease using significant risk factors. In the study using neural networks and genetic algorithms, they predicted the risk of heart disease with 89% accuracy. Zarate and Lewis (2016) aimed to determine the best algorithm in their study to estimate the duration of anesthesia of the southern elephants. The prediction of the duration is to ensure that researchers who study the species avoid risky situations. As a result, Random Tree was found to be the best classification algorithm with 98.79% accuracy.

2.1 Model Performance Evaluation

The most important misconception when evaluating the classification studies is to take into account the correct classification rate. However, there are other criteria to be looked at (Sokolova and Lapalme, 2009). Model performance criteria can be examined with the confusion matrix given in Table 1.

Table 1. Confusion Matrix

	Predicted Class		
	Class=1	Class=2	
Actual Class	Class=1	TP	FN
	Class=2	FP	TN

TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative

2.1.1 Accuracy

Accuracy is a measurement obtained by proportioning the number of accurately classified observations to the total number of samples. This value is calculated as in equation (1).

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

2.1.2 Error Rate

It is found by the ratio of the number of incorrectly classified observations to the total number of observations. The equation required for the calculation is given by equation (2).

$$ER = \frac{FP+FN}{TP+TN+FP+FN} \quad (2)$$

2.1.3 Precision

The concept of precision can be defined as the ratio of the number of positive samples found as class 1 to the number of samples with class 1. Equation is calculated as in (3).

$$PRE = \frac{TP}{TP+FP} \quad (3)$$

2.1.4 Recall

It is the ratio of the number of correctly classified positive observations to the total number of positive observations. The given value sensitivity ratio is calculated by equation (4).

$$REC = \frac{TP}{TP+FN} \quad (4)$$

2.2 Data Training and Testing Process

2.2.1 Training Data Set

It is a set of data used to estimate the classification model with the help of learning algorithms in the process of classification.

2.2.2 Test Data Set

The test data set is used to estimate the classes of observations through the classification model estimated in the training process.

2.2.3 Cross-Validation

Cross validation is a technique used in model selection. In this technique, the dataset k is subdivided into subgroups. A group is used as the test set, the others as training sets. This calculation repeats k times. Studies generally use 10-fold cross validation (Temel et. Al., 2012).

2.2.4 Percentage Split

In this method, a certain percentage of data (eg 66%) is used for education. The remaining data are estimated using classes as test data.

2.3 Classification with WEKA Tools

In the study, the Weka program developed with the Java language by Waikato University was used (<https://www.cs.waikato.ac.nz/ml/weka/>). Weka is an open source, free and user friendly program. It includes many data mining methods such as classification, association analysis, regression, artificial neural network algorithms. Data can be entered to Weka with formats such as Arff, csv, Xarff.

2.4 Application

50 different data sets were taken from the UC Irvine Machine Learning Store (<https://archive.ics.uci.edu/ml/datasets.php>). The data is arranged in the arff file format and transferred to the Weka program. 10-fold cross validation method and algorithm selection was performed in the classification window. Analyzes were made through 10 different algorithms.

Table 2. Used Data Sets

No	Data Set	Instances	No	Data Set	Instances	No	Data Set	Instances
1	Abalone	4177	18	Glass	214	35	Skin	245057
2	Adult	30162	19	Haberman	306	36	Spambase	4601
3	Balance	625	20	Hayes-roth	132	37	Statlog(ger)	1000
4	Ballons	16	21	Image	210	38	Statlog(land)	2000
5	Blood	748	22	Indian liver	583	39	Statlog(shuttle)	14500
6	Breast-cancer	286	23	Ionosphere	351	40	Teach	151
7	Car	1728	24	Iris	150	41	Tic-tac-toe	958
8	Chess	3196	25	Lens	24	42	User	258
9	Connect	67557	26	Letter	20000	43	Vertebral2	310
10	Contraceptive	1473	27	Liver	345	44	Vertebral3	310
11	Credit	653	28	Lymphography	148	45	Vowel	990
12	Dermatology	366	29	Magic	19020	46	Wifi	2000
13	Diagnosis	120	30	Meta	264	47	Wine	178
14	Diyabet	768	31	Mushroom	5644	48	Wine-white	4898
15	Ecoli	336	32	Nursery	12960	49	Yeast	1484
16	Fertility	100	33	Poker	25010	50	Zoo	101
17	Flag	194	34	Post-operative	87			

When the data sets given in Table 2 were analyzed through classification algorithms, Correctly Classify Instance Rate (CCIR) results in Figure 1 were reached. When the graph is examined, it is seen that the CCIR values of the LMT and Random Tree algorithms are higher, in other words, they are close to 100. Decision Stump algorithm does not give very good results in general.

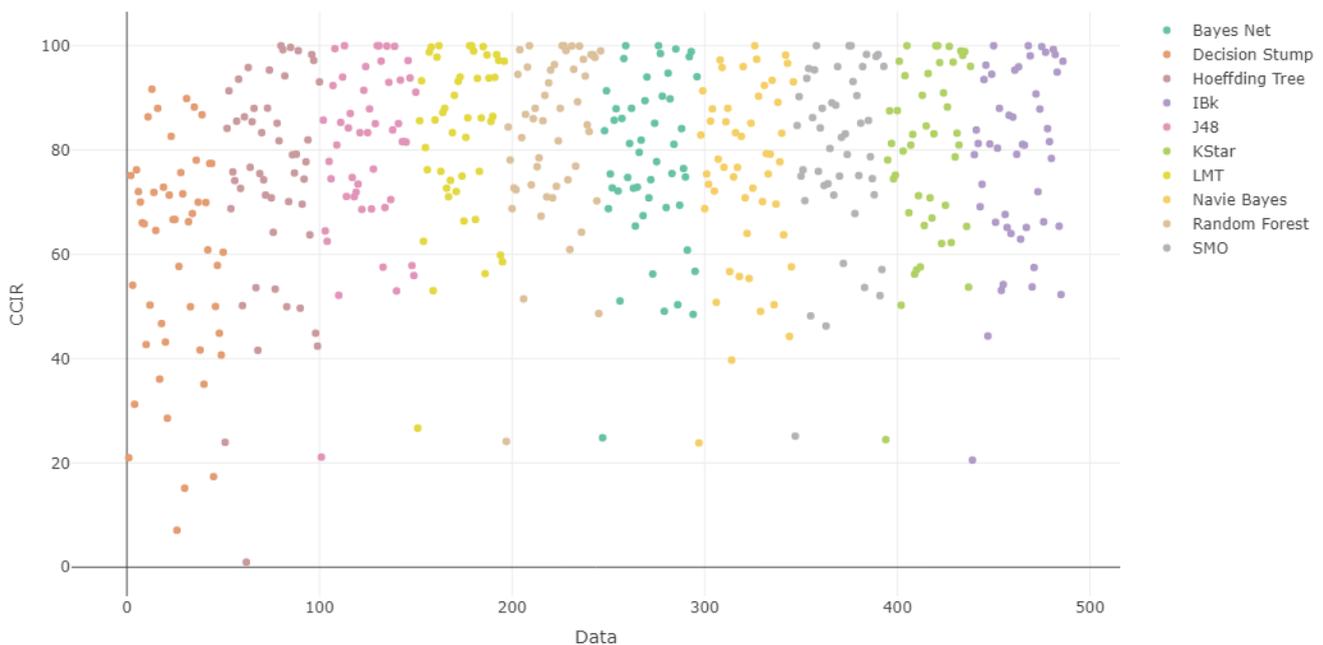


Figure 1. CCIR Values of Data Sets

Specifically, if you want to calculate the average CCIR values for each algorithm, the results given in Figure 2 are reached. These results support the results shown in Figure 1.

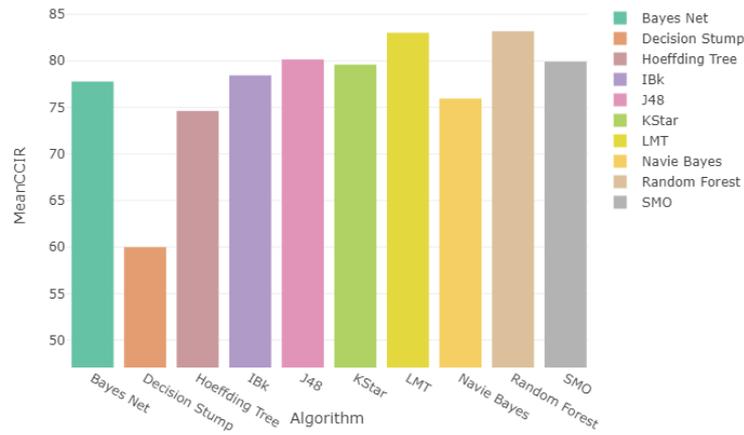


Figure 2. Mean CCIR

The biggest misconception in the classification studies is that the model success is only based on accuracy and other criteria are ignored. However, evaluation of other criteria is very important in model performance. For this reason, *Root Mean Square Error* (RMSE) values and Kappa coefficients frequently used in determining the distance between the estimated values and the real values were studied. Figure 3 shows the RMSE averages of algorithms and Kappa Statistics values in Figure 4.

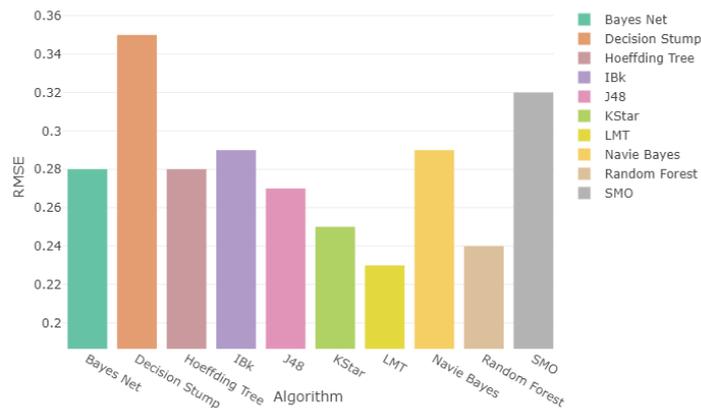


Figure 3. Mean RMSE

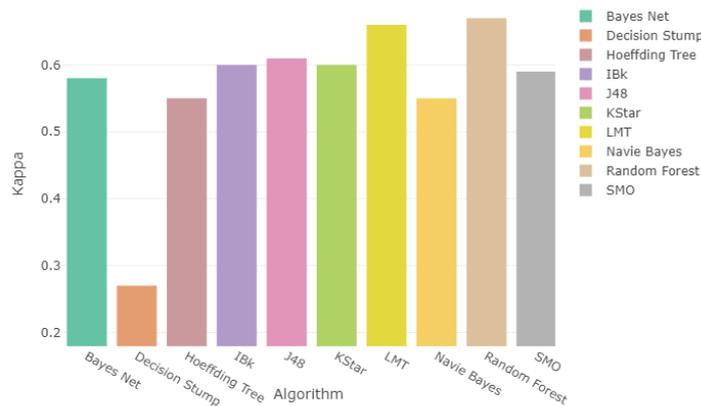


Figure 4. Mean Kappa Coefficients

RMSE is the measure of the distance between the estimated value and the actual value (Chai and Draxler, 2014). Therefore, when these values are small, it means that the model performance is high. When the Mean RMSE graph is analyzed, parallel to the previous results, while Decision Stump algorithm shows low performance LMT and Random Forest algorithms are much better.

The Kappa coefficient is used to indicate the harmony between the estimated value and the actual value (McHugh, 2012). As the Kappa value approaches 1, the adjustment value increases. Mean Kappa Coefficients graph shows the best fit values in LMT and Random Forest algorithms.

3. Conclusion

When the results of the study are examined, it is seen that LMT and Random Forest algorithms produced the best results with 83.01% and 83.17% accuracy, respectively. These algorithms are followed by the J48 algorithm with 80.14%. The Kappa coefficients obtained for the data showed the highest compatibility with Random Forest, LMT and J48 with 0.67, 0.66, 0.61 ratios, respectively. When RMSE values were considered, LMT and Random Forest algorithms showed better performance with 0.23 and 0.24, respectively. As a result, the results were found to be in the same direction for all three criteria.

There are many methods and algorithms used to access meaningful information in data mining. Many studies on the performance of these algorithms are available in the literature. However, these studies were carried out on only a few data sets. In this study, Weka tool was used to create models and to evaluate successes. With respect to the algorithm performance in 50 different data sets, general investigations have been made both in detail with Figure 1 and in terms of mean values. The results were parallel to each other.

In addition to the analyzes, different studies can be done to improve the model performance of the algorithms with low performance, to examine the speed of the algorithms in time, or to model performance of algorithms with different methods (such as Percentage Split).

References

Amin, S.U., Agarwal, K., Beg, R. (2013). “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies, Thuckalay, Tamil Nadu, India.

Chai, T., Draxler, R.R. (2014). “Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?”, Geoscientific Model Development Discussions, vol.7, pp.1247-1250.

<https://archive.ics.uci.edu/ml/datasets.php>

<https://www.cs.waikato.ac.nz/ml/weka/>

McHugh, M.L. (2012). “Interrater Reliability: The Kappa Statistic”, Biochemia Medica, vol.22, pp.276-282.

Sewaivar, P., Verma, K.K. (2015). “Comparative Study of Various Decision Tree Classification Algorithm Using WEKA”, International Journal of Emerging Research in Management & Technology, vol.4, pp.87-91.

Sokolova, M., Lapalme, G. (2009). “A Systematic Analysis of Performance Measures for Classification Tasks”, Information Processing and Management, vol.45, pp.427-437.

Temel, G.O., Erdoğan, S., Ankaralı, H. (2012). “Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı”, *Bilişim Teknolojileri Dergisi*, vol.5, pp.1-7.

Zarate, M.D., Lewis, M.N. (2016). “Estimate of the Anesthesia Stage in Southern Elephant Seals Using WEKA Data Mining Tool”, *International Journal of Applied Information Systems*, vol.11, pp.48-52.

Zupan, B., Demsar, J. (2008). “Open-Source Tools for Data Mining”, *Clinics in Laboratory Medicine*, vol.28, pp.37-54.

O-35 Classification of the Prices of Real Estate Using Machine Learning Methods

Betül KAN KILINÇ^{1*} and Yonca YAZIRLI²

¹*Department of Statistics, Eskisehir Technical University, Turkey, bkan@eskisehir.edu.tr*

²*Institute of Graduate Program, Statistics Program, Eskisehir Technical University, Turkey, yoncayazirli@gmail.com*

Abstract – As the information systems are growing day by day, it becomes easier to obtain bigger data and store it in systems. However, the data stored in the systems do not make sense of their own. Therefore, the analysis of the available data and the methods of predicting from this data play an important role for the decision makers. The process of obtaining useful information from huge amount of data can be done by data mining. One of the areas where data mining is used is the real estate platform. The aim of this study is to classify the housing unit prices of real estate for sale in Istanbul obtained from an online web source by using multinomial logit (MNL) model, support vector machines (SVM) and random forest (RF) methods. Classification algorithms have been to estimate the classifier performance. Data set is splitted as 70% for training and 30% for testing. For the model validation, 5-fold cross-validation technique is used. The accuracy and relevant performance metrics of the methods are compared. R Studio is used for all computations.

Keywords – *Data Mining, Classification, Cross Validation*

1. Introduction

Data mining is the process to discover interesting knowledge from large amounts of data (Han and Kamber, 2006). It is a combination of statistics, machine learning and computing. There are various types of studies in which machine learning algorithms and statistical techniques are used together. Kusiak (2006) introduced data mining in two categories. The first includes neural network and regression analysis and

Tomiazzi et al. (2019) investigated the increasing use of pesticides in agriculture and leading to a public health problem. They evaluate the possible ototoxic effects of exposure to pesticides and/or cigarettes in Brazilian farmers, through basic audiological evaluation and high-frequency auditory thresholds. Artificial neural network (ANN), k-NN and SVM classify the exposure groups according to audible frequency range.

Rajathi et al. (2019) proposed using an ensemble of classifiers for evaluating chronic liver disease with computed tomography images. The hybrid whale optimization algorithm with simulated annealing (WOA-SA) is used in selecting an optimal set of features to accurately classify. The proposed method compares SVM, k-NN and RF methods for the success of classification performance. The result of the proposed method was the error rate, which was 1.90% and it is the most successful algorithms.

Qi et al. (2018) developed an additional data fusion strategy based on low-level data fusion for two portions (cap and stipe) and mid-level data fusion for two spectroscopic techniques (UV and FTIR) to distinguish porcini mushrooms from different species and origins. In addition, they compared four mathematical algorithms of partial least squares discriminant analysis (PLS-DA), k-NN, genetic algorithm and SVM (GA-SVM) and RF for the discrimination to propose the best one. GA-SVM has best discrimination for porcini discrimination because this method has best performance accuracy.

Albayati and Altamimi (2019) proposed a smart system (FBChecker) that enables users to check if any Facebook profile is fake. FBChecker utilizes the data mining approach to analyze and classify a set of behavioral and informational attributes provided in the personal profiles. They investigate these attributes using four supervised data mining algorithms (k-NN, decision tree, SVM and Naïve Bayes) to determine how successfully they can detect the fake profiles. Results showed that SVM outperforms other classifiers with an accuracy rate of 98.0%.

Aggarwal (2019) had proposed an approach to estimate the decision makers choice behavior through preference learning, pointing to the complementarity of MNL model and preference learning and advocating their cross-fertilization.

Chen et al. (2018) compared decision tree, RF and Naïve bayes classification algorithms for landslide susceptibility assessment in the Longhai area of China. The three models were compared by using the area under the receiver operating characteristic (AUROC) curve, standard error, 95% confidence interval, accuracy, precision, recall, and F measure. According to these performance evaluation criteria RF is best classification algorithm for landslide susceptibility mapping.

One of the areas where data mining is used is the real estate sector. Hromada (2015) represents a software that is used for real estate evaluation and mapping and analyzing of real estate advertisements published on the internet in the Czech Republic from year 2007 until today. The software gathers price offers concerning sale or rental of apartments, houses, business properties and building lots for each half year. The author evaluates results as a steady long-term decrease of real estate market prices since the second quarter of 2008.

Chogle et al. (2017) had aimed to develop a real estate web application using Microsoft ASP .NET and SQL 2008. They used Naïve bayes classification algorithms for the price prediction. It helped to fulfill customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate.

Asilkan et al. (2012) examined the applicability of Hedonic Regression (HR) and ANN models in housing market in Tirana. Prediction of the implicit prices of the housing attributes determined the price of the complete house according to HR model. The house which contains the desired attributes has greater price than the others. The house price estimated by using ANN model based on a particular input set. ANN model obtained more successful result than HR model.

Liu and Zong (2017) proposed Twin Support Vector Regression based on data mining and large data for second-hand real estate price forecasting. It compared with traditional support vector model, the results of experiment showed that the proposed method achieves higher predictive performance.

The paper is structured as follows. In Section 2, the related works on the data mining solutions are reviewed. Section 3 classification methods which are multinomial logit (MNL) model, support vector machine (SVM) and random forest (RF) are explained. The motivation and the application are given in Section 4. The last section concludes the paper.

2. Classification Methods

2.1 Multinomial Logit Model (MNL)

Logistic regression is used to model binary response variables. Unlike a binary logit model, in which a dependent variable has only a binary choice (e.g., presence/ absence of a characteristic), the dependent variable in MNL model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category. In MNL model, the dependent variable is distributed as multinomial distribution.

The basic model formula pairs each category with a reference category. Software usually sets the last category (c) as the reference. For $c = 3$, for instance, the model uses $\log\left(\frac{\pi_1}{\pi_2}\right)$ and $\log\left(\frac{\pi_2}{\pi_3}\right)$. Conditional on the response falling in category j or in category c , $\log\left(\frac{\pi_j}{\pi_c}\right)$ is the log odds that the response is j . The reference category logit model with an explanatory variable x is

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x, \quad j = 1, 2, \dots, c - 1. \quad (1)$$

The model has $c - 1$ equations, with separate parameters for each. The effects vary according to the category paired with the reference. For p explanatory variables, this model extends to

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \dots + \beta_{jp}x_p, \quad j = 1, 2, \dots, c - 1. \quad (2)$$

Different logits have different effects in model for each explanatory variable. Software for multicategory logit models fits all $(c - 1)$ equations (2) simultaneously, using the Fisher scoring iterative algorithm. The reference category is arbitrary and the same maximum likelihood parameter estimates occur for a pair of categories no matter which reference category you use.

For the reference category logit model (2), the response probabilities relate to the model parameters by

$$\pi_j = \frac{\exp(\alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \dots + \beta_{jp}x_p)}{\sum_{h=1}^c \exp(\alpha_h + \beta_{h1}x_1 + \beta_{h2}x_2 + \dots + \beta_{hp}x_p)}, \quad j = 1, 2, \dots, c. \quad (3)$$

The denominator is the same for each π_j , and the numerators for various j sum to the denominator, hence $\sum_j \pi_j = 1$. The parameters equal zero for whichever category j is the reference in the logit expressions. (Agresti, 2019) (Wang, 2005).

2.2 Support Vector Machines (SVM)

SVM originally proposed by (Cortes and Vapnik,1995) and it is widely used for classification problems. The theoretical basis of SVM generates from statistical learning theory. It aims to find an optimal hyperplane that separates maximum margin between classes (Burges,1998). Margin is defined as the perpendicular distance between the decision boundary and the closest of data points. The support vector are the data points which are closest to the hyperplane. It is possible to distinguish linearly separable data, but it is not easy to separate the nonlinearly separable data. Therefore, SVM uses the kernel function to transform the input data into a higher dimensional space that is larger than its size. The optimal hyperplane is constructed with respect to maximum margin. Hence, the separation of classes is more easily obtained.

2.3 Random Forest (RF)

RF is one widely used method in data mining. It is similar to classification trees called Classification and Regression Trees (CART). RF uses an ensemble (i.e., forest) of decision tree predictors such that each tree have no relationship with each other and with the same distribution for all trees in the forest (Breiman et al., 1998). The trees are grown to maximum size (e.g. 2000 trees). Then it combines the predictions from all trees. Also, it is an effective method for estimating missing data.

3. Cross-Validation and Evaluation Criteria

The k-fold cross-validation method is proposed to improve the generalization ability of the model. Firstly, data are randomly partitioned into k mutually exclusive subsets or “folds” D_1, D_2, \dots, D_k each of approximately equal size. The k^{th} group of data is selected as a test sample and the remaining $k-1$ groups of data are used to train (Han and Kamber, 2006). Training and testing are performed k times, thus accuracy calculated for each fold. Finally, the global accuracy is obtained by taking the average of the accuracy of each fold. The global accuracy is the success of correctly predicted values by the model and it is calculated by confusion matrix with respect to three classes given *Table 1*.

Table 1: Confusion Matrix with Three Classes

		ACTUAL CLASS			
		A	B	C	Total
PREDICTE D CLASS	A	TP _A	E _{BA}	E _{CA}	D= TP _A + E _{BA} + E _{CA}
	B	E _{AB}	TP _B	E _{CB}	F= E _{AB} +TP _B + + E _{CB}
	C	E _{AC}	E _{BC}	TP _C	G= E _{AC} + E _{BC} + TP _C
	Total	H= TP _A + E _{AB} + E _{AC}	I= E _{BA} + TP _B + E _{BC}	J= E _{CA} + E _{CB} + TP _C	T=D+F+G+H+I+J

The false negative in the A class (FN_A) is the sum of E_{AB} and E_{AC} (FN_A= E_{AB} + E_{AC}) which implies the sum of all class A samples that were incorrectly classified as class B or C. Shortly, FN of any class which is located in a column can be calculated by adding the errors in that class/column whereas the false

positive for any predicted class which is located in a row represents the sum of all errors in that row. For instance, the false positive in class A (FP_A) is calculated as follows, $FP_A = E_{BA} + E_{CA}$. In addition, the true negative in the A class (TN_A) is the sum of TP_B and TP_C , $TN_A = TP_B + TP_C$.

$$Global\ Accuracy = \frac{Number\ of\ values\ correctly\ classified}{Number\ of\ total\ predictions} \times 100\% \quad (6)$$

In application, 5-fold cross validation is used for measuring the performance of classification algorithms.

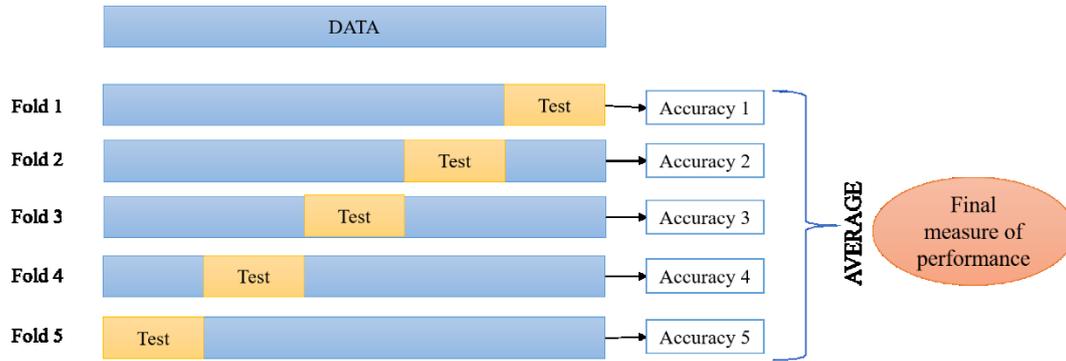


Figure 1 5-fold cross validation

In addition, kappa, sensitivity, specificity, positive prediction value and negative prediction value were used to compare the performance of the models.

For the classification model, accuracy rate shows how often the classifier is correct and is given in Eq. (6) (Hastie et al.,2008).

Cohen’ s Kappa coefficient (k) is frequently used to measure inter- rater reliability for categorical items and is given by

$$k = \frac{P_0 - P_e}{1 - P_e} = 1 - \frac{1 - P_0}{1 - P_e} \quad (7)$$

where $P_0 = \frac{TP_A + TP_B + TP_C}{T}$ is the ratio between predicted and actual and, $P_e = \left(\frac{H}{T} \times \frac{D}{T}\right) + \left(\frac{I}{T} \times \frac{F}{T}\right) + \left(\frac{J}{T} \times \frac{G}{T}\right)$ is the hypothetical probability of chance agreement (Hoehler,2000) (Tang et al,2015).

Sensitivity is the proportion of true positives that are correctly identified by the test for each class. For example, sensitivity for class A is calculated by $\frac{TP_A}{H}$. Specificity is defined as proportion of true negatives that are correctly predicted by the test for each class and for specificity for class A is obtained by $\frac{TP_B + TP_C}{I + J}$ (Altman and Bland, 1994).

Positive predictive value (PPV) is given by

$$PPV = \frac{Sensitivity \times Prevalence}{(Sensitivity \times Prevalence) + (1 - Specificity) \times (1 - Prevalence)} \quad (8)$$

and negative predictive value (NPV) is calculated by

$$NPV = \frac{Sensitivity \times (1 - Prevalence)}{(Sensitivity \times Prevalence) + (1 - Specificity) \times (1 - Prevalence)} \quad (9)$$

where *Prevalence* for class A is calculated by $\frac{H}{T}$.

4. Application

We collected the advertisements of real estate for Istanbul between 9 October- 13 December 2018 from an online popular shopping platform in which people can sell and buy real estates, car, variety of goods and services. The data contains the sale prices of 250 real estate for residential purposes. Addition to that, the number of rooms, age of building, number of floor, elevator, and bathroom for each real estate are recorded as well. The property of garage and balcony is categorized as 1 and 0 elsewhere. In addition, the convenience point for a real estate is called as amenities which is coded as 1 for yes, 0 elsewhere. One of the explanatory variable is the district of the real estate. The data is collected from 5 different districts of Istanbul. Hence the the variable district is classified in five classes. The dependent variable is the price of real estate that is converted to a categorical variable. The determine the class intervals, housing unit prices for Turkey (TL/m²) in 2018 are used (EVDS, Data Center, 14 February, 2019). According to that, if the price is larger than 2315,17 then it is classified as 2, if smaller than 2315,17 and larger than 2118,52 then it is classified as 1, and 0 elsewhere.

Table 2: Accuracy and Kappa rates for MN, SVM and RF

	Accuracy	Kappa
MNL	0.757	0.190
SVM	0.797	0.366
RF	0.821	0.423

From Table 2, it is evident that accuracy of RF is larger as compared to MNL, SVM and RF. The measure of accuracy being the frequency of correct classifier is obtained by RF as 82.1 %. Similar results hold for Kappa statistic. Accuracy is always the easiest metric to use for comparison however it does not show type of errors the classifier does. The results from the confusion matrix that is another metric used to measure the performance of a classification algorithm are given in Table 3.

Table 3: Predictive performance (%) indicators for the MNL, SVM and RF

	prices	Sensitivity	Specificity	PPV	NPV	Prevalence
MNL	Expensive	0.912	0.236	0.800	0.876	0.770
	Modarate	0.000	0.986	0.000	0.982	0.013
	Cheap	0.251	0.931	0.499	0.818	0.216
SVM	Expensive	0.912	0.412	0.838	0.588	0.770
	Modarate	0.000	1.000	Na	0.983	0.013
	Cheap	0.437	0.913	0.588	0.854	0.216
RF	Expensive	0.938	0.427	0.846	0.676	0.770
	Modarate	0.000	1.000	Na	0.983	0.013
	Cheap	0.453	0.939	0.676	0.861	0.216

To detect how sensitive (recall) is the classifier in positive instances or how often the predictions are correct when the actual value is positive, sensitivity metric of higher percent for expensive class is desired which leads us to the RF model (93.8%). This means that 93.8% of the actual observations in the data set of 250 real estates are correctly predicted for price group, only 6.2% of the real estates are incorrectly predicted to be other classes by RF model. On the other hand, all classification models perform

inadequate scores for detection of moderate class whereas specificity score's of these models perform better for detection of true negatives. Generally, RF outperforms others according to sensitivity and specificity for each class. MNL model poorly detect the instances of moderate class that were correctly classified for moderate class (0.00), however it performs better for NPV score (98.2%). RF model predicts sample of expensive class as an expensive class more accurate than SVM and MNL models. In addition, prevalence scores are equal for all models.

5. Conclusion

In this study, the housing unit prices of real estate for sale in different districts of Istanbul obtained from an online web source is classified by using multinomial logit (MNL) model, support vector machines (SVM) and random forest (RF) methods. For this purpose, two hundred and fifty real estate and its properties are investigated. The performance of all methods are compared with respect to accuracy, Kappa value, and relevant evaluation metrics. Results indicate that the overall accuracy produced by RF yields a better performance than others.

References

- Aggarwal, M. (2019). “Preferences-based learning of multinomial logit model”, *Knowledge and Information Systems*, vol.59, pp. 523–538.
- Albayati, M. B., Altamimi, A. M. (2019). “An empirical study for detecting fake facebook profiles using supervised mining techniques”, *Informatica*, vol43, pp. 77–86.
- Altman, D. G., Bland, J.M. (1994). “Statistics Notes: Diagnostic tests 1: sensitivity and specificity”, *British Medical Journal*. vol. 308, pp. 1552.
- Asilkan, O., Faqolli, A., Gerdecı, A., Cico, B. (2012). “Estimating the market values of houses in Tirana using data mining”, *AWER Procedia Information Technology & Computer Science*, vol. 1, pp. 1224-1234.
- Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. (1998). *Classification and Regression Trees*, Chapman and Hall/CRC Press, Florida.
- Burges, C.J.C (1998). “A tutorial on support vector machines for pattern recognition”, *Data Mining and Knowledge Discovery*, vol.2, no. 2, pp. 121–167.
- Chan, Y.H. (2005). “Biostatistics 305. Multinomial logistic regression”, *Basic Statistics For Doctors*, *Singapore Med J*. Vol.46, no.6, pp.259-269.
- Chen, W., Zhang, S., Li, R., Shahabi, H. (2018). “Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling”, *Science of the Total Environment*, vol. 644, pp. 1006–1018.
- Chogle, A., Khaire, P., Gaud, A., Jain, J. (2017). “House price forecasting using data mining techniques”, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, no. 12, pp. 81-90.

- Cortes, C., Vapnik, V. (1995). “Support-vector networks”, *Machine Learning*, vol.20, no.3, pp. 273-297.
- Fawcett, Tom (2006). "An Introduction to ROC Analysis", *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874.
- Han, J. and Kamber, M. (2006). “Data Mining: Concepts and Techniques”, 2nd ed., Elsevier, San Francisco, California.
- Hastie T, Tibshirani R, Friedman J. (2008). “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, 2nd ed., Springer, Standford, California.
- Hoehler, F.K. (2000). “Bias and prevalence effects in kappa viewed in terms of sensitivity and specificity”. *Journal of Clinical Epidemiology*, vol.53, no. 5, pp.499-503.
- Hromoda, E. (2015). “Mapping of real estate prices using data mining techniques”, *Procedia Engineering* vol. 123, pp. 233 – 240.
- Liu, G., Zong, X. (2017). “Research of Second-hand Real Estate Price Forecasting Based on Data Mining”, *IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference*, Chengdu, China.
- Tang, W., Hu, J., Zhang, H., Wu, P., He, H. (2015). “Kappa coefficient: a popular measure of rater agreement”. *Shanghai Arch. Psychiatry*, vol. 27, no. 1, pp.62–67.
- Tomiazzi, J. S., Pereira, D. R., Judai, M. A., Antunes, P. A., Favareto, A. P. A. (2019). “Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke”, *Environmental Science and Pollution Research*, vol. 26, pp. 6481–6491.
- Rajathi, G. I., Jiji, G. W. (2019). “Chronic Liver Disease Classification Using Hybrid Whale Optimization with Simulated Annealing and Ensemble Classifier”, *Symmetry*, vol. 11, no.1, pp. 33-54.
- Qi, L. M., Li, J. Q, Liu, H. G., Li, T., Wang, Y. Z. (2018). “An additional data fusion strategy for the discrimination of porcini mushrooms from different species and origins in combination with four mathematical algorithms”, *Food & Function*, vol.9, pp. 5903–5911.
- EVDS, Data Center <https://evds2.tcmb.gov.tr/index.php?evds/serieMarket>, Accessed on:14.02.2019.
- Zeng, A., Pan, D., Zheng, Q. L., Peng, H. (2006). “Knowledge acquisition based on rough set theory and principal component analysis”, *IEEE Intelligent Systems*, vol. 21, issue 2, pp. 78-85.

O-36 Improving Two Stage Two Parameter Ridge Estimator under Linear Restrictions

Selma Toker^{1*} and Nimet Özbay²

¹Department of Statistics, Çukurova University, Turkey, stoker@cu.edu.tr

²Department of Statistics, Çukurova University, Turkey, nturker@cu.edu.tr

Abstract – A two parameter estimator is more advantageous than a single parameter estimator because two parameters have two different benefits with regard to estimating structural coefficients. To take advantage of this idea, Toker (2018) has described a two stage two parameter ridge estimator in simultaneous equations model. This is such an estimator which mitigates the problem of multicollinearity with its first parameter and improves quality of fit with its second parameter. If some constraints are encountered in the model of simultaneous equations, restricted estimators can be more attractive than the classical ones. In this respect, we define restricted form of the two stage two parameter ridge estimator. While proposing our new restricted two stage two parameter ridge estimator, we have inspired by the notion in the paper of Üstündağ Şiray and Toker (2014). In addition, theoretical properties of this new estimator are investigated and an optimal biasing parameter is proposed. The best performed estimator is determined in terms of efficiency as a consequence of the empirical evaluations.

Keywords –Linear restrictions, Mean square error, Multicollinearity, Simultaneous equations model, Two parameter estimator

1. Introduction

The basis of the class of simultaneous equations model emerges in a data generation process which is related to more than one equation interacting together to generate the observed data. On the contrary of a single equation model where a dependent variable is a function of independent variables, in each simultaneous equation, other dependent variables exist among the independent variables. These variables are named endogenous and predetermined variables. Endogenous variables are random variables and have interactions in the model. Predetermined variables consist of exogenous, lagged exogenous and lagged endogenous variables. Being non-random variables, exogenous variables are determined outside the system of simultaneous equations. Lagged endogenous variables are the lagged values of the endogenous variables of some past periods. Also, lagged exogenous variables seem as lagged values of the exogenous variables in the system.

Let us introduce the matrix form of the simultaneous equations model as follows:

$$Y\Gamma + XB = U. \quad (1)$$

In model (1), Y and X are matrices of observations with the sizes of $T \times M$ and $T \times K$, respectively. Γ and B are the matrices of structural coefficients corresponding to the sizes of $M \times M$ and $K \times M$. In addition, U is the matrix of structural disturbances having the dimension of $T \times M$.

By postmultiplying both sides of matrix notation (1) by a nonsingular matrix, Γ^{-1} , the reduced form equation is derived as follows:

$$Y = X\Pi + V, \tag{2}$$

where

$$\Pi = -B\Gamma^{-1} \tag{3}$$

and

$$V = U\Gamma^{-1} \tag{4}$$

are the reduced form coefficients.

The structural coefficients describe the direct effects of variables on one another. Before these coefficients can be meaningfully estimated, identification problem, which examines the relationship between the structural equations and the reduced form of these equations, should be solved. By using different restriction techniques, one of which is the zero restrictions criterion, the identifiability status can be determined. While employing this criterion, let us consider any equation of the system (say first equation) as follows:

$$y_1 = Y_1\gamma_1 + X_1\beta_1 + u_1. \tag{5}$$

In this first equation, there exist $m_1 + 1$ included and $m_1^* = M - m_1 - 1$ excluded jointly dependent variables and K_1 included and $K_1^* = K - K_1$ excluded predetermined variables. $Y = [y_1 \ Y_1 \ Y_1^*]$ and $X = [X_1 \ X_1^*]$ are variables in which Y_1, Y_1^*, X_1 and X_1^* have the dimensions of $T \times m_1, T \times m_1^*, T \times K_1$ and $T \times K_1^*$, respectively. $\gamma_{.1} = [1 \ -\gamma_1 \ 0]'$ and $\beta_{.1} = [-\beta_1 \ 0]'$ are variables which includes γ_1 and β_1 having the dimensions of $m_1 \times 1$ and $K_1 \times 1$, respectively. The last term u_1 is the first column of U .

The partition of the reduced form equation is written as

$$\begin{aligned} [y_1 \ Y_1 \ Y_1^*] &= [X_1 \ X_1^*] \begin{bmatrix} \pi_{11} & \Pi_{11} & \Pi_{11}^* \\ \pi_{21} & \Pi_{21} & \Pi_{21}^* \end{bmatrix} \\ &+ [v_1 \ V_1 \ V_1^*] \end{aligned} \tag{6}$$

and the following equations are derived

$$y_1 = X\pi_1 + v_1 \tag{7}$$

and

$$Y_1 = X\Pi_1 + V_1, \tag{8}$$

where $\pi_1 = [\pi_{11} \ \pi_{21}]'$ and $\Pi_1 = [\Pi_{11} \ \Pi_{21}]'$ and $\pi_{11}, \pi_{21}, \Pi_{11}, \Pi_{21}, v_1$ and V_1 are the variables with the dimensions of $K_1 \times 1, K_1^* \times 1, K_1 \times m_1, K_1^* \times m_1, T \times 1$ and $T \times m_1$, respectively.

To express the identifiability relationship between the structural parameters and the reduced form parameters for the first equation, we utilize the reduced form coefficients (3) and (4). After turning these equations to $\Pi\Gamma + B = 0$ and $V\Gamma = U$, we get only the first column of Γ, B and U as follows:

$$\Pi\gamma_{.1} + \beta_{.1} = \begin{bmatrix} \pi_{11} & \Pi_{11} & \Pi_{11}^* \\ \pi_{21} & \Pi_{21} & \Pi_{21}^* \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_1 \\ 0 \end{bmatrix} + \begin{bmatrix} -\beta_1 \\ 0 \end{bmatrix} = 0', \tag{9}$$

and

$$V\gamma_{.1} = [v_1 \ V_1 \ V_1^*] \begin{bmatrix} 1 \\ -\gamma_1 \\ 0 \end{bmatrix} = u_1. \tag{10}$$

In the next step, the relations below are derived from the equations (9) and (10)

$$\pi_{11} = \Pi_{11}\gamma_1 + \beta_1, \tag{11}$$

$$\pi_{21} = \Pi_{21}\gamma_1, \tag{12}$$

and

$$v_1 = V_1\gamma_1 + u_1. \tag{13}$$

We rewrite the first equation of the system (5) as follows:

$$\begin{aligned} y_1 &= Y_1\gamma_1 + X_1\beta_1 + u_1 \\ &= Z_1\delta_1 + u_1, \end{aligned} \tag{14}$$

where $Z_1 = [Y_1 \ X_1]_{T \times p_1}$, $\delta_1 = [\gamma_1 \ \beta_1]_{p_1 \times 1}'$ and $p_1 = m_1 + K_1$.

The next step constitutes the form below by replacing the equations (13) and (8) one by one in the structural equation (14):

$$\begin{aligned}
 y_1 &= Z_1 \delta_1 + u_1 \\
 &= [Y_1 \quad X_1] \begin{bmatrix} \gamma_1 \\ \beta_1 \end{bmatrix} + u_1 \\
 &= Y_1 \gamma_1 + X_1 \beta_1 + v_1 - V_1 \gamma_1 \\
 &= [Y_1 - V_1 \quad X_1] \begin{bmatrix} \gamma_1 \\ \beta_1 \end{bmatrix} + v_1 \\
 &= [X\Pi_1 + V_1 - V_1 \quad X_1] \begin{bmatrix} \gamma_1 \\ \beta_1 \end{bmatrix} + v_1 \\
 &= [X\Pi_1 \quad X_1] \begin{bmatrix} \gamma_1 \\ \beta_1 \end{bmatrix} + v_1.
 \end{aligned} \tag{15}$$

Let us simplify the equation (15) to its final form below

$$y_1 = \bar{Z}_1 \delta_1 + v_1, \tag{16}$$

where $\bar{Z}_1 = E(Z_1) = [X\Pi_1 \quad X_1]$, $\delta_1 = [\gamma_1 \quad \beta_1]_{p_1 \times 1}$, $E(v_1) = 0$ and $E(v_1 v_1') = \sigma^2 I$.

By the way of standardization of variables in the model (16), $\bar{Z}_1' \bar{Z}_1$ and $\bar{Z}_1' y_1$ turn to correlation matrix of the regressors and the vector of correlations of the dependent variable, respectively.

According to the identification status of the model (just or over identified), we determine a proper estimation method. When we come across just or over identified equations, two stage least squares (TSLS) estimator is used. This estimator is an alternative to ordinary least squares (OLS) estimator since the OLS estimates of structural coefficients are biased and inconsistent.

We can explain the process of the TSLS method as follows. In a first stage regression, instrumental variables are produced in which each endogenous variable is regressed against the entire set of exogenous variables. Next, the second stage regression is performed for the structural equations by using the OLS estimates of the first stage regression as instruments for the observed values of the endogenous variables.

Let us consider the least squares objective function below for implementation of the TSLS method with the standardized variables:

$$\begin{aligned}
 S^2 &= \|v_1\|^2 \\
 &= (y_1 - \bar{Z}_1 \delta_1)' (y_1 - \bar{Z}_1 \delta_1) \\
 &= 1 - 2\delta_1' \bar{Z}_1' y_1 + \delta_1' \bar{Z}_1' \bar{Z}_1 \delta_1.
 \end{aligned} \tag{17}$$

The objective function above is used for the minimization of the sum of squared deviations. The following normal system of the equations is obtained after minimizing the objective function (17) for the vector δ_1 :

$$\bar{Z}_1' \bar{Z}_1 \delta_1 = \bar{Z}_1' y_1. \tag{18}$$

The solution regarding this normal system becomes the TSLS estimator and here is the definition:

$$\delta_1^{LS} = (\bar{Z}'_1 \bar{Z}_1)^{-1} \bar{Z}'_1 y_1. \quad (19)$$

In this form of estimator, \bar{Z}_1 is not known, hence, we use the estimator of Π_1 below

$$\hat{\Pi}_1 = (X'X)^{-1} X'Y_1 \quad (20)$$

at the first stage and then \hat{Z}_1 becomes

$$\hat{Z}_1 = [X\hat{\Pi}_1 \quad X_1]. \quad (21)$$

Finally, we reach the TSLS estimator in practice as follows:

$$\hat{\delta}_1^{LS} = (\hat{Z}'_1 \hat{Z}_1)^{-1} \hat{Z}'_1 y_1. \quad (22)$$

The quality of the model can be determined with the coefficient of multiple determination. With this respect, coefficient of multiple determination will be defined by utilizing the equations (17) and (18) as follows:

$$\begin{aligned} R^2 &= 1 - S^2 \\ &= 2\delta'_1 \bar{Z}'_1 y_1 - \delta'_1 \bar{Z}'_1 \bar{Z}_1 \delta_1 \\ &= \delta'_1 \bar{Z}'_1 y_1 \\ &= \delta'_1 \bar{Z}'_1 \bar{Z}_1 \delta_1. \end{aligned} \quad (23)$$

As for the orthogonality between each regressor and the error vector, the equation (18) can be rearranged as

$$\bar{Z}'_1 v_1 = 0. \quad (24)$$

To test the model quality by the F statistics, a beneficial decomposition of R^2 and S^2 should be found. By means of the orthogonality property in the equation (24), definition of R^2 in the equation (23) and the definition of S^2 in the equation (17), the following decomposition is derived as follows:

$$1 = R^2 + S^2. \quad (25)$$

When multicollinearity exists, the matrix $\bar{Z}'_1 \bar{Z}_1$ is ill conditioned, so the variances of the estimated TSLS coefficients of structural parameters inflate, the estimated TSLS coefficients have wrong signs and hence these estimates become statistically insignificant. Because of the hardship we run into while estimating the coefficients of structural parameters with TSLS method in the existence of multicollinearity, we should try to find alternative methods. To this end, an alternative estimator, ridge estimator (RE) of Hoerl and Kennard (1970), was recommended by Vinod and Ullah (1981) for estimating parameters in simultaneous equations model.

To apply two stage ridge regression, we can consider the following objective function:

$$\begin{aligned} S^2 &= \|v_1\|^2 + k\|\delta_1\|^2 \\ &= 1 - 2\delta_1'\bar{Z}'_1y_1 + \delta_1'\bar{Z}'_1\bar{Z}_1\delta_1 + k\delta_1'\delta_1, \end{aligned} \quad (26)$$

where $k > 0$ is biasing parameter. By the way of minimization of the equation (26) according to vector δ_1 , a normal system of equations

$$(\bar{Z}'_1\bar{Z}_1 + kI)\delta_1 = \bar{Z}'_1y_1 \quad (27)$$

is obtained. The solution concerning the equation (27) resulted in two stage RE

$$\delta_1^{RE} = (\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1. \quad (28)$$

For the practical use of the two stage RE, the following form is used:

$$\hat{\delta}_1^{RE} = \left(\hat{Z}'_1\hat{Z}_1 + kI\right)^{-1}\hat{Z}'_1y_1. \quad (29)$$

Even though the two stage RE is beneficial for recovering multicollinearity, the quality of fit of a regular regression is deficient for the result of the two stage RE. The relations (23)-(25) provided for the TSLS estimator are not provided for the two stage RE. These negative drawbacks lead us to deal with some other estimators for simultaneous equations model. Some various estimators in the literature are as follows: Two stage Liu estimator of Toker et al. (2018), two stage two parameter estimator of Özbay and Toker (2018), two stage two parameter ridge estimator (TPRE) of Toker (2018) and two stage modified RE of Toker and Özbay (2018a). Since a two parameter estimator is more advantageous than a single parameter estimator, we prefer to investigate a two parameter estimator in simultaneous equations model. Within this context, we will deal with the two stage TPRE of Toker (2018). Toker (2018) have described this estimator by following the way of defining the TPRE of Lipovetsky and Conklin (2005) and Lipovetsky (2006) in linear regression model. The objective function of the two stage RE (26) is generalized and then the objective function of the two stage TPRE is constituted as

$$S^2 = \|v_1\|^2 + k_1\|\delta_1\|^2 + k_2\|\bar{Z}'_1y_1 - \delta_1\|^2 + k_3\|y'_1v_1\|^2. \quad (30)$$

While generalizing the objective function, two additional terms (third and the fourth terms in eq. 30) are used. By means of the third term, the estimate of the coefficient vector becomes closer to the paired correlations (\bar{Z}'_1y_1). The fourth term is for maximizing the coefficient of multiple determination and it is achieved by minimizing the residuals due to $y'_1v_1 = 1 - R^2$.

After minimizing the objective function in the equation (30) for δ_1 , the following simple form of the normal system of equations can be written :

$$(\bar{Z}'_1\bar{Z}_1 + kI)\delta_1 = q\bar{Z}'_1y_1, \quad (31)$$

where $k = k_1 + k_2 + k_3\bar{Z}'_1y_1y'_1\bar{Z}_1$ and $q = 1 + k_2 + k_3$. The solution related to this system becomes the two stage TPRES

$$\delta_1^{TPRES} = q(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1, \quad (32)$$

where k and q are two new constant parameters.

The operational form of the two stage TPRES is found with

$$\hat{\delta}_1^{TPRES} = q\left(\hat{Z}'_1\hat{Z}_1 + kI\right)^{-1}\hat{Z}'_1y_1. \quad (33)$$

The unknown parameter q is found by Toker (2018) via maximizing the coefficient of multiple determination and it is formulated as follows

$$q = \frac{y'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1+kI)^{-1}\bar{Z}'_1y_1}{y'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1+kI)^{-1}\bar{Z}'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1+kI)^{-1}\bar{Z}'_1y_1}. \quad (34)$$

The second parameter q improves quality of fit and at the same time, the first parameter k is for recovering the multicollinearity. In addition to these advantages of two biasing parameters, the two stage TPRES satisfies the properties in the equations (23)-(25) as in the TSLS estimator.

When we encounter some restrictions in a simultaneous equations model, the restricted estimators can be more appealing than unrestricted ones in the simultaneous equations model. The restricted estimators of Zellner et al. (1988), Toker and Kaçiranlar (2017) and Toker and Özbay (2018b) become a solution to the multicollinearity in the simultaneous equations model. Specially, Toker and Özbay (2018b) introduced restricted TSLS estimator and restricted two stage RE by inspiring from the paper of Groß (2003) in linear regression model. By this context, the following exact linear restrictions on the coefficients are taken into consideration:

$$r_1 = R_1\delta_1, \quad (35)$$

where r_1 is a $G \times 1$ vector and R_1 is a $G \times p_1$ matrix of rank $G < p_1$. Under these restrictions, the restricted TSLS estimator is given by Toker and Özbay (2018) as follows:

$$\delta_1^{RLS} = \delta_1^{LS} - (\bar{Z}'_1\bar{Z}_1)^{-1}R'_1[R_1(\bar{Z}'_1\bar{Z}_1)^{-1}R'_1]^{-1}(R_1\delta_1^{LS} - r_1). \quad (36)$$

To employ the restricted TSLS estimator in practice, the form of the estimator below is used:

$$\hat{\delta}_1^{RLS} = \hat{\delta}_1^{LS} - \left(\hat{Z}'_1\hat{Z}_1\right)^{-1}R'_1\left[R_1\left(\hat{Z}'_1\hat{Z}_1\right)^{-1}R'_1\right]^{-1}(R_1\hat{\delta}_1^{LS} - r_1). \quad (37)$$

It is evident that the restricted TSLS estimator depends on the TSLS estimator. Thus, when the simultaneous equations model is exposed to multicollinearity, the restricted TSLS estimator displays a

poor estimation performance because of this dependence. Toker and Özbay (2018b) offered the restricted two stage RE as an alternative to the restricted TLSLS estimator as follows:

$$\delta_1^{RRE} = \delta_1(k, \delta_1^0) - (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1 [R_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1]^{-1} (R_1 \delta_1(k, \delta_1^0) - r_1), \quad (38)$$

where

$$\delta_1(k, \delta_1^0) = (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} (\bar{Z}'_1 y_1 + k\delta_1^0) \quad (39)$$

is the two stage modified RE of Toker and Özbay (2018a) with the prior information vector $\delta_1^0 = R'_1 (R'_1 R_1)^{-1} r_1$ which is the shortest vector satisfying the restrictions.

The operational form of this estimator is found as

$$\hat{\delta}_1^{RRE} = \hat{\delta}_1(k, \delta_1^0) - (\hat{Z}'_1 \hat{Z}_1 + kI)^{-1} R'_1 \left[R_1 (\hat{Z}'_1 \hat{Z}_1 + kI)^{-1} R'_1 \right]^{-1} (R_1 \hat{\delta}_1(k, \delta_1^0) - r_1), \quad (40)$$

where

$$\hat{\delta}_1(k, \delta_1^0) = (\hat{Z}'_1 \hat{Z}_1 + kI)^{-1} (\hat{Z}'_1 y_1 + k\delta_1^0). \quad (41)$$

Based on the recent papers in literature related to the developments about restricted versions of the estimators in the simultaneous equations model, we feel that the restricted versions of the two parameter estimators are worth to examine. We will fulfill this idea in the next section to propose restricted two stage TPRES.

2. Defining the New Restricted Estimator

In this section, we will carry out the linear restrictions issue to the two stage TPRES for defining restricted two stage TPRES. The similar consideration, namely implementing the restrictions to the TPRES, was performed by Üstündağ Şiray and Toker (2014) for linear regression model. In defining our new estimator in the simultaneous equations model, we act by utilizing this idea of Üstündağ Şiray and Toker (2014).

In the existence of prior information, we will first propose a two stage modified two parameter ridge estimator (MTPRES) with respect to this information. To this end, let us constitute the objective function below:

$$S^2 = \|v_1\|^2 + k_1(\|\delta_1 - \delta_1^0\|^2 - a) + k_2(\|\bar{Z}'_1 y_1 - (\delta_1 - \delta_1^0)\|^2 - a) + k_3 \left(\|y'_1 (y_1 - \bar{Z}_1 (\delta_1 - \delta_1^0))\|^2 - a \right) \quad (42)$$

where k_1, k_2, k_3 and a are the constants and $\delta_1^0 = R_1'(R_1'R_1)^{-1}r_1$ satisfies the linear restrictions. Let us get the derivative of this objective function with respect to δ_1 as follows:

$$\begin{aligned} \frac{\partial S^2}{\partial \delta_1} = & -2y_1'\bar{Z}_1 + 2\delta_1'\bar{Z}_1'\bar{Z}_1 + 2k_1\delta_1 - 2k_1\delta_1^0 - 2k_2y_1'\bar{Z}_1 + 2k_2\delta_1' \\ & - 2k_2\delta_1^0 - 2k_3y_1'\bar{Z}_1 + 2k_3\delta_1'\bar{Z}_1'y_1y_1'\bar{Z}_1 - 2k_3\delta_1^0\bar{Z}_1'y_1y_1'\bar{Z}_1. \end{aligned} \quad (43)$$

Then we equate the equation (43) to zero and obtain the corresponding solution as

$$\delta_1^{MTPRE} = q(\bar{Z}_1'\bar{Z}_1 + kI)^{-1}(\bar{Z}_1'y_1 + (k/q)\delta_1^0), \quad (44)$$

where $k = k_1 + k_2 + k_3\bar{Z}_1'y_1y_1'\bar{Z}_1$ and $q = 1 + k_2 + k_3$. We call this new estimator two stage MTPRE. The two stage MTPRE contains the two stage modified RE (39) in the case of $q = 1$ and the two stage TPRES (32) when $\delta_1^0 = 0$.

To use this estimator in application, we give the following form:

$$\hat{\delta}_1^{MTPRE} = q(\hat{Z}_1'\hat{Z}_1 + kI)^{-1}(\hat{Z}_1'y_1 + (k/q)\delta_1^0). \quad (45)$$

As for the definition of the restricted two stage TPRES, we will impose the linear restrictions on the parameter space and obtain the objective function below:

$$\begin{aligned} S^2 = & \|v_1\|^2 + k_1(\|\delta_1 - \delta_1^0\|^2 - a) + k_2(\|\bar{Z}_1'y_1 - (\delta_1 - \delta_1^0)\|^2 - a) \\ & + k_3(\|y_1'(y_1 - \bar{Z}_1(\delta_1 - \delta_1^0))\|^2 - a) + \lambda(R_1\delta_1 - r_1), \end{aligned} \quad (46)$$

where k_1, k_2, k_3 and a are constants and $\delta_1^0 = R_1'(R_1'R_1)^{-1}r_1$ is the prior information which satisfies the linear restrictions. Let us differentiate the equation (46) with respect to δ_1 and λ and then these derivatives result in

$$\begin{aligned} \frac{\partial S^2}{\partial \delta_1} = & -2y_1'\bar{Z}_1 + 2\delta_1'\bar{Z}_1'\bar{Z}_1 + 2k_1\delta_1 - 2k_1\delta_1^0 - 2k_2y_1'\bar{Z}_1 + 2k_2\delta_1' \\ & - 2k_2\delta_1^0 - 2k_3y_1'\bar{Z}_1 + 2k_3\delta_1'\bar{Z}_1'y_1y_1'\bar{Z}_1 - 2k_3\delta_1^0\bar{Z}_1'y_1y_1'\bar{Z}_1 + R_1'\lambda, \end{aligned} \quad (47)$$

$$\frac{\partial S^2}{\partial \lambda} = R_1\delta_1 - r_1. \quad (48)$$

The normal system of equations will be reached after equating the derivatives (47) and (48) to zero. After some algebraic operations, this normal system of equations are simplified as follows:

$$(\bar{Z}'_1\bar{Z}_1 + kI)\delta_1 = q\bar{Z}'_1y_1 + k\delta_1^0 - R'_1\lambda, \quad (49)$$

where $k = k_1 + k_2 + k_3\bar{Z}'_1y_1y'_1\bar{Z}_1$ and $q = 1 + k_2 + k_3$.

As a solution of equation (49) we obtain

$$\tilde{\delta}_1 = \delta_1^{MTPRE} - (\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1\lambda. \quad (50)$$

By premultiplying the equation (50) by R_1 , we obtain

$$R_1\tilde{\delta}_1 = R_1\delta_1^{MTPRE} - R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1\lambda. \quad (51)$$

Finally the simple form of the equation (51) is attained as

$$R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1\lambda = R_1\delta_1^{MTPRE} - r_1. \quad (52)$$

By solving the equation (52) for λ , the estimator of λ will be derived as

$$\hat{\lambda} = (R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1)^{-1}(R_1\delta_1^{MTPRE} - r_1). \quad (53)$$

$\hat{\lambda}$ will be replaced in the equation (50) then the final form of the restricted two stage TPRES is derived as follows:

$$\delta_1^{RTPRES} = \delta_1^{MTPRES} - (\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1[R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1]^{-1}(R_1\delta_1^{MTPRES} - r_1). \quad (54)$$

The restricted two stage TPRES is a general estimator which includes restricted two stage RE (38) when $q = 1$ and the restricted TSLS estimator (36) when $k = 0$ and $q = 1$ as special cases.

The operational form of the new estimator will be defined as

$$\hat{\delta}_1^{RTPRES} = \hat{\delta}_1^{MTPRES} - \left(\hat{Z}'_1\hat{Z}_1 + kI\right)^{-1}R'_1\left[R_1\left(\hat{Z}'_1\hat{Z}_1 + kI\right)^{-1}R'_1\right]^{-1}(R_1\hat{\delta}_1^{MTPRES} - r_1). \quad (55)$$

3. Determination of the biasing parameter

In our newly defined estimators, δ_1^{MTPRES} and δ_1^{RTPRES} , the main mission of the second parameter q is improving the quality of fit of the model. Therefore, the selection issue of this parameter is so significant. Taking account of q selection method of Üstündağ Şiray and Toker (2014), we introduce the following procedure to select q . We will try to find an optimal q which maximizes the general expression of the coefficient of multiple determination in the equation (23) for the restricted two stage TPRES.

By putting δ_1^{MTPRE} in the equation (54), an alternative form of the restricted two stage TPRES is written as

$$\delta_1^{RTPRE} = q(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1 - q(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\theta(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1 + Ar_1, \quad (56)$$

where $\theta = R'_1[R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1]^{-1}R_1$ and $A = (\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1[R_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}R'_1]^{-1}$. Then by letting $M_k = (\bar{Z}'_1\bar{Z}_1 + kI)^{-1} - (\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\theta(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}$, the restricted two stage TPRES is rewritten in the simpler form below:

$$\delta_1^{RTPRE} = qM_k\bar{Z}'_1y_1 + Ar_1. \quad (57)$$

For the purpose of obtaining an optimal q , we firstly employ the general formula in the equation (23) to propose the coefficient of multiple determination for the restricted two stage TPRES as follows:

$$R^2 = 2[qy'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1} - qy'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\theta(\bar{Z}'_1\bar{Z}_1 + kI)^{-1} + r'_1A']\bar{Z}'_1y_1 - [qy'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1} - qy'_1\bar{Z}_1(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\theta(\bar{Z}'_1\bar{Z}_1 + kI)^{-1} + r'_1A']\bar{Z}'_1\bar{Z}_1 \times [q(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1 - q(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\theta(\bar{Z}'_1\bar{Z}_1 + kI)^{-1}\bar{Z}'_1y_1 + Ar_1]. \quad (58)$$

R^2 formula in the equation (58) can be simplified as

$$R^2 = 2[qy'_1\bar{Z}_1M_k + r'_1A']\bar{Z}'_1y_1 - [qy'_1\bar{Z}_1M_k + r'_1A']\bar{Z}'_1\bar{Z}_1[qM_k\bar{Z}'_1y_1 + Ar_1] = 2qy'_1\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r'_1A'\bar{Z}'_1y_1 - q^2y'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - qy'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 - qr'_1A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - r'_1A'\bar{Z}'_1\bar{Z}_1Ar_1. \quad (59)$$

By differentiating the equation (59) with respect to q

$$\frac{\partial R^2}{\partial q} = 2y'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - 2qy'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - y'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 - r'_1A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 \quad (60)$$

and then solving it for q , we obtain the optimal q value which maximizes the R^2 of the restricted two stage TPRES as follows:

$$q = \frac{y'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - r'_1A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}. \quad (61)$$

After the determination of the optimal q parameter, we proceed with demonstration of some useful properties of our new restricted two stage TPRES. Let us substitute q in the equation (61) in the equation (57) and hence the equation (57) becomes

$$\delta_1^{RTPRE} = \frac{y'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - r'_1A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} M_k\bar{Z}'_1y_1 + Ar_1. \quad (62)$$

From the equation (61), we get the following relation

$$qy'_1\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 = y'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - r'_1A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 \quad (63)$$

and using this relation the R^2 formula in the equation (59) turns to

$$\begin{aligned}
 R^2 &= 2qy_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r_1'A'\bar{Z}'_1y_1 - q(y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1) \\
 &\quad - qy_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 - qr_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= qy_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r_1'A'\bar{Z}'_1y_1 - qy_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 - r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1.
 \end{aligned} \tag{64}$$

If we put the optimal q value (61) into the equation (64), we obtain

$$\begin{aligned}
 R^2 &= \frac{y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r_1'A'\bar{Z}'_1y_1 \\
 &\quad - \frac{y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 - r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1. \\
 &= \frac{y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r_1'A'\bar{Z}'_1y_1 \\
 &\quad - \frac{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1y_1'\bar{Z}_1Ar_1 - r_1'A'\bar{Z}'_1y_1y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1y_1'\bar{Z}_1r_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} - r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= \frac{y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + 2r_1'A'\bar{Z}'_1y_1 - y_1'\bar{Z}_1Ar_1 \\
 &\quad + r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 - r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= y_1'\bar{Z}_1\delta_1^{RTPRE}.
 \end{aligned} \tag{65}$$

In addition to this, we observe

$$\begin{aligned}
 (\delta_1^{RTPRE})'\bar{Z}_1\bar{Z}_1(\delta_1^{RTPRE}) &= (qM_k\bar{Z}'_1y_1 + Ar_1)'\bar{Z}_1\bar{Z}_1(qM_k\bar{Z}'_1y_1 + Ar_1) \\
 &= q^2y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 + qy_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 + qr_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1 + \\
 &\quad r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= qy_1'\bar{Z}_1M_k\bar{Z}'_1y_1 + qy_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 + r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= R^2 - 2r_1'A'\bar{Z}'_1y_1 + 2qy_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 + 2r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= R^2 - 2r_1'A'\bar{Z}'_1y_1 + 2\frac{y_1'\bar{Z}_1M_k\bar{Z}'_1y_1 - r_1'A'\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1Ar_1 + \\
 &\quad 2r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= R^2 - 2r_1'A'\bar{Z}'_1y_1 + 2\frac{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1y_1'\bar{Z}_1Ar_1 - r_1'A'\bar{Z}'_1y_1y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1y_1'\bar{Z}_1Ar_1}{y_1'\bar{Z}_1M_k\bar{Z}'_1\bar{Z}_1M_k\bar{Z}'_1y_1} \\
 &\quad + 2r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= R^2 - 2r_1'A'\bar{Z}'_1y_1 + 2y_1'\bar{Z}_1Ar_1 - 2r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 + 2r_1'A'\bar{Z}'_1\bar{Z}_1Ar_1 \\
 &= R^2.
 \end{aligned} \tag{66}$$

From the equations (65) and (66) we can easily see that

$$y_1'\bar{Z}_1\delta_1^{RTPRE} = (\delta_1^{RTPRE})'\bar{Z}_1\bar{Z}_1(\delta_1^{RTPRE}) \tag{67}$$

which is similar to the relation (23). Furthermore, this equality yields the orthogonality property in the equation (24). Hence, the relations (23)-(25) will be satisfied for restricted two stage TPRES as are satisfied for the TSLS estimator and the two stage TPRES.

4. Mean Square Error of the New Estimators

Mean square error criterion, which indicates the performance of an estimator, is frequently preferable for comparison of the estimators. By the way we will also use this criterion to determine the best performed estimator.

The matrix mean square error (MSE) of $\bar{\delta}_1$, which is an estimator of δ_1 , is defined to be

$$MSE(\bar{\delta}_1) = V(\bar{\delta}_1) + B(\bar{\delta}_1)B(\bar{\delta}_1)', \quad (68)$$

where the variance-covariance matrix and the bias are

$$Var(\bar{\delta}_1) = E \left[(\bar{\delta}_1 - E(\bar{\delta}_1)) (\bar{\delta}_1 - E(\bar{\delta}_1))' \right] \quad (69)$$

and

$$Bias(\bar{\delta}_1) = E(\bar{\delta}_1) - \delta_1. \quad (70)$$

The scalar mean square error (mse) of $\bar{\delta}_1$ are as follows:

$$mse(\bar{\delta}_1) = tr \left(MSE(\bar{\delta}_1) \right). \quad (71)$$

Let us give the MSEs of the above mentioned estimators. The first one to give is the MSE of two stage MTPRE. The variance, bias and finally the MSE of this estimator are defined as follows:

$$Var(\delta_1^{MTPRE}) = q^2 \sigma^2 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \bar{Z}'_1 \bar{Z}_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1}, \quad (72)$$

$$Bias(\delta_1^{MTPRE}) = (q(\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} k \delta_1^0, \quad (73)$$

$$\begin{aligned} MSE(\delta_1^{MTPRE}) &= q^2 \sigma^2 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \bar{Z}'_1 \bar{Z}_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \\ &\quad + [(q(\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} k \delta_1^0] \\ &\quad \times [(q(\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} k \delta_1^0]'. \end{aligned} \quad (74)$$

Before writing the MSE of the restricted two stage TPRES, let us express it in a simple form:

$$\delta_1^{RTPRE} = qM_k \bar{Z}'_1 y_1 + kM_k \delta_1^0 + (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1 [R_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1]^{-1} r_1, \quad (75)$$

where $M_k = (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} - (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1 [R_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1} R'_1]^{-1} R_1 (\bar{Z}'_1 \bar{Z}_1 + kI)^{-1}$. By replacing r_1 with $R_1 \delta_1$, δ_1^{RTPRE} can be rewritten as

$$\begin{aligned} \delta_1^{RTPRE} &= qM_k \bar{Z}'_1 y_1 + kM_k \delta_1^0 + \delta_1^0 - M_k (\bar{Z}'_1 \bar{Z}_1 + kI) \delta_1^0 \\ &= qM_k \bar{Z}'_1 y_1 + \delta_1^0 - M_k \bar{Z}'_1 \bar{Z}_1 \delta_1^0. \end{aligned} \quad (76)$$

To obtain the variance, bias and finally the MSE for the restricted two stage TPRES, we use the form of the estimator in the equation (76) and these are given in the subsequent formulas:

$$Var(\delta_1^{RTPRE}) = q^2 \sigma^2 M_k \bar{Z}'_1 \bar{Z}_1 M_k, \quad (77)$$

$$Bias(\delta_1^{RTPRE}) = (qM_k \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (I - M_k \bar{Z}'_1 \bar{Z}_1) \delta_1^0, \quad (78)$$

$$\begin{aligned} MSE(\delta_1^{RTPRE}) &= q^2 \sigma^2 M_k \bar{Z}'_1 \bar{Z}_1 M_k \\ &\quad + [(qM_k \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (I - M_k \bar{Z}'_1 \bar{Z}_1) \delta_1^0] \\ &\quad \times [(qM_k \bar{Z}'_1 \bar{Z}_1 - I) \delta_1 + (I - M_k \bar{Z}'_1 \bar{Z}_1) \delta_1^0]'. \end{aligned} \quad (79)$$

When $q = 1$, the MSE of the restricted two stage TPRES turns to the MSE of the restricted two stage RE. Since the MSE formulas of the foregoing estimators are so complicated to compare, we prefer to perform the comparison in a numerical way with regard to the mse. In the next section, we will do an application for examining the performance of the two stage MTPRES and the restricted two stage TPRES as against the existing estimators.

5. Data Analysis

We consider the model which is one of the three Keynesian macroeconomic models of the U.S. economy declared in the book of Economic Fluctuations published in 1950 by Lawrence Klein (Klein, 1950). We take account of Model I based on the data in the period between 1921 and 1941. There are three identity equations called national accounts, profit and loss account and change in capital stock in addition to three behavioral equations named consumption, investment and demand for labor equations. The equations of Klein's Model I are set out below:

$$(80) \quad \text{Consumption} \quad : C_t = \delta_0 + \delta_1 P_t + \delta_2 P_{t-1} + \delta_3 (W_t + W'_t) + u_t,$$

$$(81) \quad \text{Investment} \quad : I_t = \delta'_0 + \delta'_1 P_t + \delta'_2 P_{t-1} + \delta'_3 K_{t-1} + u'_t,$$

$$(82) \quad \text{Demand for labor} \quad : W_t = \delta''_0 + \delta''_1 X_t + \delta''_2 X_{t-1} + \delta''_3 (t - 1931) + u''_t$$

$$(83) \quad \text{National Accounts} \quad : X_t = C_t + I_t + G_t,$$

$$(84) \quad \text{Profit and Loss Account} : P_t = X_t - W_t - T_t,$$

$$(85) \quad \text{Change in Capital stock} : K_t = K_{t-1} + I_t.$$

Herein, C is aggregate consumption, P is total profits, W is total wages paid by private industry, W' is the government wage bill, I is net investment, K is the stock of capital goods, X is the total production of private industry, G is government nonwage expenditure and T is business taxes in a year. The six endogenous variables are C, P, W, I, K and X . The eight predetermined variables are $G, T, t, W', K_{-1}, P_{-1}, X_{-1}$ and intercept.

The consumption, investment and demand for labor equations can be written in the form of the equation (14) as follows:

$$\begin{bmatrix} C_1 \\ \vdots \\ C_{21} \end{bmatrix} = \begin{bmatrix} P_1 & W_1 + W'_1 & \vdots & 1 & P_0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{21} & W_{21} + W'_{21} & \vdots & 1 & P_{20} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_3 \\ \dots \\ \delta_0 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_{21} \end{bmatrix}, \quad (86)$$

$$\begin{bmatrix} I_1 \\ \vdots \\ I_{21} \end{bmatrix} = \begin{bmatrix} P_1 & \vdots & 1 & P_0 & K_0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{21} & \vdots & 1 & P_{20} & K_{20} \end{bmatrix} \begin{bmatrix} \delta'_1 \\ \dots \\ \delta'_0 \\ \delta'_2 \\ \delta'_3 \end{bmatrix} + \begin{bmatrix} u'_1 \\ \vdots \\ u'_{21} \end{bmatrix}$$

(87)

and

$$\begin{bmatrix} W_1 \\ \vdots \\ W_{21} \end{bmatrix} = \begin{bmatrix} X_1 & \vdots & 1 & X_0 & -10 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{21} & \vdots & 1 & X_{20} & 10 \end{bmatrix} \begin{bmatrix} \delta''_1 \\ \dots \\ \delta''_0 \\ \delta''_2 \\ \delta''_3 \end{bmatrix} + \begin{bmatrix} u''_1 \\ \vdots \\ u''_{21} \end{bmatrix}, \quad (88)$$

where the subscript 21 refers to the annual observations of the period 1921 to 1941. Z_1 consists of two submatrices one of which is Y_1 that contains the values of the explanatory jointly dependent variables and the other one is X_1 that contains the values of the explanatory predetermined variables. For numerical evaluation, we arbitrarily choose $R_1 = \begin{bmatrix} 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}$. To satisfy the exact linear restrictions given by the equation (35), r_1 is computed as $[0.1482 \quad -0.7547]'$, $[-0.0836 \quad -0.9606]'$ and $[1.6065 \quad 0.6599]'$ corresponding to the consumption, investment and demand for labor equations. The prior vectors related to the equations (80)-(82) are respectively calculated as $\delta_1^0 = [0.1999 \quad 0.7030 \quad -0.2516]'$, $\delta_1^0 = [0.1183 \quad 0.7587 \quad -0.3202]'$ and $\delta_1^0 = [0.6933 \quad 0.2533 \quad 0.2200]'$.

We illustrate the estimated mse values of the restricted estimators in Figures 1-3 concerned to the equations of consumption, investment and demand for labor, respectively. These figures are drawn for the values of k in the interval $[0, 3]$ with the increment of 0.01. As is seen in Figure 1, the restricted two stage TPRE outperforms the restricted two stage RE and the restricted TSLS estimator for the given interval of k , under the given restrictions and prior information. Based on the Figure 2, it is observed that the restricted two stage TPRE is always superior to the restricted TSLS estimator. In the meantime, the line for our new estimator goes below the line for the restricted two stage RE for most cases. Namely, when k is approximately greater than 0.5, the restricted two stage TPRE is better than the restricted two stage RE with regard to the mse under the given restrictions and the prior information. Taking account of Figure 3, the restricted two stage TPRE always gives overwhelming results in the sense of mse in comparison to the results of the existing restricted estimators under the given restrictions and the prior information.

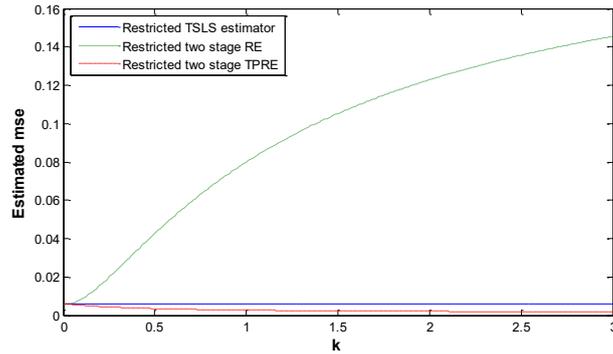


Figure 1. Estimated mse values of the restricted TSLS estimator, restricted two stage RE and restricted two stage TPRE for consumption equation

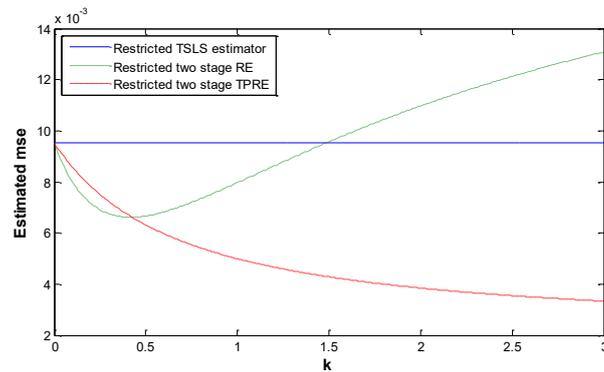


Figure 2. Estimated mse values of the restricted TSLS estimator, restricted two stage RE and restricted two stage TPRE for investment equation

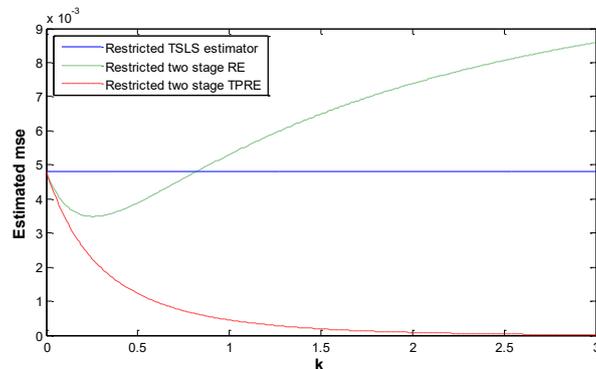


Figure 3. Estimated mse values of the restricted TSLS estimator, restricted two stage RE and restricted two stage TPRE for demand for labor equation

After drawing three figures above to support the theoretical results with graphical representations, we can look into the comparisons of the estimators by means of the the tables. Tables 1-3 demonstrates the estimated mse results of the TSLS estimator, restricted TSLS estimator, two stage RE, restricted two stage RE, two stage TPRE, two stage MTPRE and restricted two stage TPRE and the optimal q values for different values of k according to the given restrictions and the prior information. Table 1 is for the consumption equation, Table 2 is constituted for the investment equation and at last the Table 3 gives the results for the demand for labor equation. Since recovering multicollinearity becomes easier with the help of the linear restrictions, superiority of the restricted two stage TPRE to its counterparts is expected. As an evidence to this expectation, the restricted TSLS estimator is superior to the TSLS estimator, the restricted two stage RE is superior to the two stage RE and also, the restricted two stage TPRE is superior

to the two stage TPRE for both three equations according to the results obtained from the Tables 1-3. Additionally, as a general inference, it is concluded that the restricted two stage TPRE is the best comparing against its competitors both for the equations of consumption and demand for labor. Specially, for the equation of investment, the estimated mse values of the restricted two stage TPRE are smallest when k is approximately greater than 0.5. In fact, the outcomes are attained for certain restrictions and prior information. Within this context, the outcomes could change for different restrictions and prior information.

Table 1. Estimated values of q and mse for different values of k by consumption equation

	$k = 0.05$	$k = 0.1$	$k = 0.2$	$k = 0.3$	$k = 0.4$	$k = 0.5$	$k = 0.7$	$k = 0.9$	$k = 1$	$k = 3$
\hat{q}	1.03482	1.06964	1.13928	1.20891	1.27855	1.34819	1.48747	1.62674	1.69638	3.08914
TOLS estimator	0.04010	0.04010	0.04010	0.04010	0.04010	0.04010	0.04010	0.04010	0.04010	0.04010
Restricted TOLS estimator	0.00612	0.00612	0.00612	0.00612	0.00612	0.00612	0.00612	0.00612	0.00612	0.00612
Two stage RE	0.03130	0.03804	0.06629	0.09918	0.13120	0.16090	0.21277	0.25596	0.27491	0.48188
Restricted two stage RE	0.00637	0.00842	0.01550	0.02434	0.03359	0.04265	0.05928	0.07362	0.07998	0.14545
Two stage TPRE	0.03012	0.03111	0.04292	0.05787	0.07258	0.08620	0.10975	0.12903	0.13739	0.22330
Two stage MTPRE	0.06056	0.10632	0.19102	0.25724	0.31054	0.35553	0.43048	0.49350	0.52208	0.90677
Restricted two stage TPRE	0.00565	0.00525	0.00464	0.00418	0.00383	0.00355	0.00313	0.00284	0.00273	0.00185

Table 2. Estimated values of q and mse for different values of k by investment equation

	$k = 0.05$	$k = 0.1$	$k = 0.2$	$k = 0.3$	$k = 0.4$	$k = 0.5$	$k = 0.7$	$k = 0.9$	$k = 1$	$k = 3$
\hat{q}	1.02429	1.04858	1.09716	1.14573	1.19431	1.24289	1.34004	1.43720	1.48578	2.45733
TOLS estimator	0.14312	0.14312	0.14312	0.14312	0.14312	0.14312	0.14312	0.14312	0.14312	0.14312
Restricted TOLS estimator	0.00951	0.00951	0.00951	0.00951	0.00951	0.00951	0.00951	0.00951	0.00951	0.00951
Two stage RE	0.09009	0.10111	0.12974	0.15217	0.17055	0.18665	0.21504	0.24030	0.25206	0.41021
Restricted two stage RE	0.00864	0.00799	0.00715	0.00675	0.00662	0.00667	0.00706	0.00765	0.00797	0.01308
Two stage TPRE	0.09020	0.09826	0.12015	0.13506	0.14520	0.15254	0.16260	0.16934	0.17200	0.19460
Two stage MTPRE	0.06254	0.04428	0.03954	0.04594	0.05635	0.06882	0.09683	0.12638	0.14118	0.37257
Restricted two stage TPRE	0.00901	0.00857	0.00782	0.00722	0.00673	0.00632	0.00567	0.00519	0.00500	0.00333

Table 3. Estimated values of q and mse for different values of k by demand for labor equation

	$k = 0.05$	$k = 0.1$	$k = 0.2$	$k = 0.3$	$k = 0.4$	$k = 0.5$	$k = 0.7$	$k = 0.9$	$k = 1$	$k = 3$
\hat{q}	0.98812	0.97623	0.95247	0.92870	0.90493	0.88117	0.83364	0.78610	0.76234	0.28701
TLS estimator	0.03988	0.03988	0.03988	0.03988	0.03988	0.03988	0.03988	0.03988	0.03988	0.03988
Restricted TLS estimator	0.00481	0.00481	0.00481	0.00481	0.00481	0.00481	0.00481	0.00481	0.00481	0.00481
Two stage RE	0.03931	0.05613	0.08545	0.10660	0.12299	0.13663	0.15940	0.17878	0.18766	0.31077
Restricted two stage RE	0.00423	0.00386	0.00352	0.00350	0.00365	0.00388	0.00445	0.00503	0.00530	0.00859
Two stage TPRES	0.04002	0.05866	0.09251	0.11905	0.14153	0.16183	0.19904	0.23368	0.25029	0.50569
Two stage MTPRES	0.02233	0.01869	0.01943	0.02272	0.02673	0.03104	0.03997	0.04893	0.05335	0.12303
Restricted two stage TPRES	0.00407	0.00347	0.00259	0.00198	0.00154	0.00122	0.00080	0.00054	0.00045	0.00001

6. Concluding Remarks

In this paper, our fundamental aim is to take the advantage of applying the restrictions to a two parameter estimator in a simultaneous equations model. Two stage TPRES of Toker (2018) is beneficial in eliminating the multicollinearity with its first parameter and improving quality of fit with its second parameter. As a contribution to the literature, we propose the modified and restricted forms of the two stage TPRES by incorporating restrictions to the model of simultaneous equations. Also, an optimal value of the second parameter q is determined by maximizing the coefficient of multiple determination. Thanks to this optimal value, we succeed in developing the performance of the restricted two stage TPRES. Moreover, theoretical properties of the newly defined estimators are investigated. Our numerical findings from the data analysis demonstrate that the outperformance of the restricted two stage TPRES is noteworthy in dealing with multicollinearity. As future directions, we can offer proposing the restricted versions of the different two parameter estimators when someone come across some constraints in the simultaneous equations model.

Acknowledgment

This paper is supported by Çukurova University Scientific Research Projects Unit with Project Number: FBA-2019-11884.

References

- Groß, J. (2003), “Restricted Ridge Estimation”, *Statistics&Probability Letters*, vol. 65, pp.57–64.
- Hoerl A.E., Kennard R.W. (1970). “Ridge regression: Biased estimation for nonorthogonal problems”, *Technometrics*, vol. 12, no. 1, pp.55-67.

- Klein L.R. (1950). "Economic Fluctuation in the United States, 1921-1941", John Wiley and Sons, Inc., New York.
- Lipovetsky, S., Conklin, W.M. (2005). "Ridge regression in two parameter solution", *Applied Stochastic Models in Business and Industry*, vol. 21, pp.525-540.
- Lipovetsky, S. (2006). "Two parameter ridge regression and its convergence to the eventual pairwise model", *Mathematical and Computer Modelling*, vol. 44, pp.304-318.
- Özbay, N., Toker, S. (2018). "Multicollinearity in simultaneous equations system: Evaluation of estimation performance of two-parameter estimator", *Computational and Applied Mathematics*, vol. 37, no. 4, pp.5334-5357.
- Toker, S., Kaçiranlar, S. (2017). "Ridge estimator with correlated errors and two stage ridge estimator under inequality restrictions", *Communications in Statistics-Theory and Methods*, vol. 46, no. 3, pp.1407-1421.
- Toker, S. (2018). "Investigating the two parameter analysis of Lipovetsky for simultaneous systems", *Statistical Papers*, DOI: 10.1007/s00362-018-1021-1.
- Toker, S., Özbay, N. (2018a). "Evaluation of two stage modified ridge estimator and its performance", *Sakarya Journal of Science*, vol. 22, no. 6, pp.1631-1637.
- Toker, S., Özbay, N. (2018b). "Investigation of two stage ridge estimator under linear constraints", *International Conference on Multidisciplinary Sciences (ICOMUS 2018)*, 15-16 December, İstanbul, Turkey.
- Toker, S., Kaçiranlar, S. and Güler, H. (2018). "Two-stage Liu estimator in a simultaneous equations model", *Journal of Statistical Computation and Simulation*, vol. 88, no. 11, pp.2066-2088.
- Üstündağ Şiray, G., Toker, S. (2014). "Restricted two parameter ridge estimator", *Australian&New Zealand Journal of Statistics*, vol. 55, no. 4, pp. 455-469.
- Vinod, H.D., Ullah, A. (1981). "Recent Advances in Regression Methods", Marcel Dekker, New York, Inc.
- Zellner, A., Bauwens, L. and Van Dijk, H.K. (1988). "Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods", *Journal of Econometrics*, vol. 38, pp.39-72.

O-37 Defining Some Adaptive Optimal Estimators for the Distributed Lag Model

Nimet Özbay^{1*} and Selma Toker²

¹Department of Statistics, Çukurova University, Turkey, nturker@cu.edu.tr

²Department of Statistics, Çukurova University, Turkey, stoker@cu.edu.tr

Abstract – Minimum mean square error estimators have a widespread usage to estimate the unknown coefficients of linear regression model. For practical purposes, adaptive forms of these estimators are preferable due to their attractive performances. Within this context, adaptive forms of the estimators that minimize mean square error can be evaluated in the distributed lag model, which is a dynamic model for time series data. In the estimation issue of the distributed lag model, Almon estimator is the foremost estimator depending on its unbiasedness and ease of application. However, in the existence of multicollinearity, the use of biased estimators, one of which is Almon ridge estimator, is inevitable. We take into consideration the Almon and Almon ridge estimators in this paper with the aim of defining two new adaptive optimal estimators. The performance of the proposed adaptive optimal Almon and adaptive optimal Almon ridge estimators are examined by means of a Monte Carlo simulation.

Keywords – Adaptive optimal, Almon estimator, Almon ridge estimator, Distributed lag model, Mean square error

1. Introduction

In economic sense, a variation in the magnitude of an explanatory variable can often affect the response in some future time periods because of the dynamic structure of the economic data as well. The reason for observing such an effect is that any variation in economic variables can take a long time. This type cause and effect relationship reveals in distributed lag models. The model in question is classified as finite and infinite distributed lag models related to their number of lag. We express the finite distributed lag model as follows:

$$y_t = \sum_{i=0}^p \beta_i x_{t-i} + u_t, \quad t = p + 1, \dots, T, \quad (1)$$

where x_{t-i} shows t -th observation of the i -th period lag value of the explanatory variable, y_t is the t -th observation on the dependent variable depending on both x_t and some past values of x_t , the unknown distributed lag coefficient is shown with β_i and the disturbance term for the t -th observation is u_t which is distributed as $IN(0, \sigma^2)$. Herein, p represents the lag length.

With the aim of presenting the matrix notation of this model, we use the expression below:

$$y = X\beta + u, \tag{2}$$

where

$$X = \begin{bmatrix} x_{p+1} & x_p & \dots & x_1 \\ x_{p+2} & x_{p+1} & \dots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_T & x_{T-1} & \dots & x_{T-p} \end{bmatrix}, \tag{3}$$

$$y' = (y_{p+1}, y_{p+2}, \dots, y_T), \tag{4}$$

$$\beta' = (\beta_0, \beta_1, \dots, \beta_p) \tag{5}$$

and

$$u' = (u_{p+1}, u_{p+2}, \dots, u_T). \tag{6}$$

In the case that all of the model assumptions are satisfied, ordinary least squares (OLS) estimator $\hat{\beta} = (X'X)^{-1}X'y$ is a convenient estimator. On the other hand, this estimator is the reason for high variances when multicollinearity is faced. The problem of multicollinearity often occurs in the distributed lag model in consequence of high collinearity among the invariably lagged values of the same explanatory variable. In the existence of this problem, some methods, one of which is the incorporation of extraneous information by specifying a lag distribution, are inevitably recommended (see also Fisher, 1937; Gujarati, 1999; Kennedy, 2003; Vinod and Ullah, 1981).

Almon polynomial lag distribution (Almon, 1965) is a popular tool to struggle against the multicollinearity. The procedure of Almon includes placing some extra information into the model. That is, the prior information of the lag weights is specified as a form of a polynomial of degree r as follows:

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \dots + \alpha_r i^r, \quad p \geq r \geq 0. \tag{7}$$

For the purpose of simplification, the equation (7) is expressed as

$$\beta = A\alpha, \tag{8}$$

where

$$\alpha' = (\alpha_0, \alpha_1, \dots, \alpha_r) \tag{9}$$

and

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & p & p^2 & \dots & p^r \end{bmatrix} \tag{10}$$

are $(r + 1) \times 1$ vector and $(p + 1) \times (r + 1)$ matrix, respectively. It is supposed that the ranks of matrices X and A are $(p + 1) < (T - p)$ and $(r + 1) < (p + 1)$, respectively. To apply the Almon

technique, the form of lag weights in the equation (8) is substituted in the model (2) to obtain the Almon model below:

$$y = Z\alpha + u, \quad (11)$$

where $Z = XA$. By employing the OLS method to the model (11), the associated estimator, namely the Almon estimator (AE), is defined to be

$$\hat{\alpha}_{AE} = (Z'Z)^{-1}Z'y. \quad (12)$$

The AE of β in the model (2) is derived as follows:

$$\hat{\beta}_{AE} = A\hat{\alpha}_{AE}. \quad (13)$$

When we encounter the multicollinearity in the distributed lag model, the performance of the AE becomes poor. The large values in $(Z'Z)^{-1}$ yield large variances and inflated confidence intervals of the coefficients estimated by the AE. There have been many biased estimation methods to recover the problem of multicollinearity (see Vinod and Ullah, 1981; Chanda and Maddala, 1984; Gültay and Kaçiranlar, 2015; Güler et al., 2017; Özbay and Kaçiranlar, 2017; Özbay and Toker, 2018a; Özbay and Toker, 2018b; Özbay, 2019; Özbay and Toker, 2019a, Özbay and Toker, 2019b). One of the biased estimation methods that was initially proposed for linear regression model is ridge regression of Hoerl and Kennard (1970). The ridge estimator can be evaluated as an alternative interpretation of the polynomial given in the equation (8) (see Yeo and Trivedi, 1989). Therefore, the ridge estimator for the Almon model (11) is called Almon ridge estimator (ARE) which is defined to be

$$\hat{\alpha}_{ARE} = (Z'Z + kI)^{-1}Z'y, \quad k \geq 0. \quad (14)$$

The ARE for β in the model (2) is obtained as

$$\hat{\beta}_{ARE} = A\hat{\alpha}_{ARE}. \quad (15)$$

Despite its usefulness in mitigating the multicollinearity, some troubles concerned about the selection of the biasing parameter k emerge in the usage of the ARE.

2. Defining Adaptive Optimal Almon Estimators

In the literature, the researchers have been going on to search and develop biased estimation methods that will improve the mean square error criterion in the linear regression model. Estimators that minimize the mean square error were defined by Theil (1958, 1971), Toutenburg (1968), Rao (1971, 1973, 1999) and Schaffrin (1985, 1986, 1987). The adaptive forms of these estimators were investigated by the authors Farebrother (1975), Dwivedi and Srivastava (1978), Vinod (1976, 1980), Stahlecker and Trenkler (1985), Tracy and Srivastava (1994), Wan and Chaturvedi (2000) and Özbay and Kaçiranlar (2017). At this point, as an effective solution to the estimation problem, estimators which make the mean square error minimum and their adaptive forms can be introduced for the distributed lag model when unbiasedness as well as linearity is neglected.

Let C be an arbitrary matrix with $(r + 1) \times (T - p)$ dimension. Then consider the following class of linear estimators for the model (11)

$$\hat{\alpha}_{CLE} = Cy. \tag{16}$$

The matrix mean square error (MSE) of $\hat{\alpha}_{CLE}$ is found by

$$\begin{aligned} MSE(\hat{\alpha}_{CLE}) &= E(\hat{\alpha}_{CLE} - \alpha)(\hat{\alpha}_{CLE} - \alpha)' \\ &= E(CZ\alpha + Cu - \alpha)(CZ\alpha + Cu - \alpha)' \\ &= (CZ - I)\alpha\alpha'(CZ - I)' + \sigma^2 CC'. \end{aligned} \tag{17}$$

The matrix C for which (17) becomes minimum is derived by solving

$$\frac{\partial MSE(\hat{\alpha}_{CLE})}{\partial C} = 2C(Z\alpha\alpha'Z' + \sigma^2 I) - 2\alpha\alpha'Z' = 0 \tag{18}$$

for C . This is resulted in $C = \alpha\alpha'Z'(Z\alpha\alpha'Z' + \sigma^2 I)^{-1}$. By substituting C in the equation (16), we define the optimal AE as follows:

$$\hat{\alpha}_{OAE} = \alpha\alpha'Z'(Z\alpha\alpha'Z' + \sigma^2 I)^{-1}y. \tag{19}$$

An alternative form of the equation (19) can be written as

$$\begin{aligned} \hat{\alpha}_{OAE} &= \frac{\alpha\alpha'Z'}{\sigma^2} [I - Z\alpha(\sigma^2 + \alpha'Z'Z\alpha)^{-1}\alpha'Z']y \\ &= \frac{1}{\sigma^2} \left(\alpha'Z'y - \frac{\alpha'Z'Z\alpha\alpha'Z'y}{\sigma^2 + \alpha'Z'Z\alpha} \right) \alpha \\ &= \left(\frac{\alpha'Z'y}{\sigma^2 + \alpha'Z'Z\alpha} \right) \alpha. \end{aligned} \tag{20}$$

Note that since $\hat{\alpha}_{OAE}$ depends upon the unknown α and σ^2 , it is practically useless. We can replace α and σ^2 with the unbiased estimators $\hat{\alpha}_{AE}$ and $\hat{\sigma}^2$, respectively. As a consequence, we propose adaptive optimal Almon estimator (AOAE) as follows:

$$\hat{\alpha}_{AOAE} = \left(\frac{\hat{\alpha}'_{AE}Z'y}{\hat{\sigma}^2 + \hat{\alpha}'_{AE}Z'Z\hat{\alpha}_{AE}} \right) \hat{\alpha}_{AE}. \tag{21}$$

The AOAE for β in the model (2) is written as

$$\hat{\beta}_{AOAE} = A\hat{\alpha}_{AOAE}. \tag{22}$$

As is seen, the AOAE represented by the equation (21) shrinks the AE shown by the equation (12). Based on the general advantages of shrinkage estimators, we can say that estimators which have smaller variance and bias are obtained. Therefore, the AOAE will be more effective than the AE in the solution of the estimation problem for the model of distributed lag. However, since the AOAE contains estimates for unknown parameters, the ability to make the MSE minimum may vary. For this reason, the performance analysis will be carried out with the simulation study in the next section.

Another issue arises from the requirement to investigate the behavior of the AOAE in the presence of multicollinearity. The form of the AOAE contains the AE and hence the lack of resilience of the AE to

multicollinearity also adversely affects the AOAE. Therefore, the ARE given by the equation (14) can be utilized to reduce the sensitivity of the estimator given by the equation (21) to the multicollinearity. Because, while defining the ARE, it is aimed to overcome the problem of multicollinearity by adding a constant to the diagonal elements of the matrix $Z'Z$. The adaptive optimal Almon ridge estimator (AOARE) is thus defined as follows:

$$\hat{\alpha}_{AOARE} = \left(\frac{\hat{\alpha}'_{ARE} Z' y}{\hat{\sigma}^2 + \hat{\alpha}'_{ARE} Z' Z \hat{\alpha}_{ARE}} \right) \hat{\alpha}_{ARE}. \quad (23)$$

Herein, to estimate β in the model (2), the form of the AOARE is given below:

$$\hat{\beta}_{AOARE} = A \hat{\alpha}_{AOARE}. \quad (24)$$

After these definitions, a simulation study will be performed in the next section to examine the performance of two new estimators in the model of distributed lag.

3. Monte Carlo Experiment

This experiment aims to measure relative efficiencies of the mentioned estimators with regard to scalar mean square error (mse) criterion. Frost (1975), Güler et. al. (2017) and Özbay and Kaçiranlar (2017) built a typical form for the distributed lag model for the purpose of conducting a Monte Carlo simulation study. By following these authors, we can write the model below to generate observations:

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \dots + \beta_p x_{t-p} + u_t, \quad (25)$$

$$x_1 = v_1, \quad (26a)$$

$$x_t = w x_{t-1} + (1 - w^2)^{1/2} v_t \text{ for } t \geq 2, \quad (26b)$$

where w is the correlation between the explanatory variables. The error terms are independent pseudo-normal deviates so that $u_t \sim N(0, \sigma^2)$ and $v_t \sim N(0,1)$. Three different levels of σ are considered as 1, 10 and 100. For each replication, the data can be generated by holding values of u_t for $t = 12, \dots, 71$ and v_t for $t = 1, \dots, 71$.

After generating the values of x_t with the help of the equations (26a) and (26b), the values of y_t will be obtained from the equation (25) taking into account $T = 60$ observations with lags of 11 periods for each trial. Four different values of w corresponding to 0.75, 0.85, 0.95 and 0.99 are used. At first, we choose the parameter values of β according to $\beta' \beta = 1$ and the results for this experiment are given in Table 1. Next, we use three different experiments named corresponding to Experiment I, Experiment II and Experiment III. Chanda and Maddala (1984) suggested these three different parameter vectors which constitute the experiments that address different shapes of lag distribution for real life application. The shape of the Experiment I represents damped oscillation, the second experiment shows a declining line

shape and the last experiment forms a humped shape. The specific coefficients for these experiments are summarized in Table 2 and the outcomes for each experiment are represented in Tables 3-5, respectively. The initial generation of the variables is made from the model (2). For the transition to the Almon model in the equation (11), we constitute the matrix $Z = XA$ where the polynomial degree for A is chosen as $r = 2$. To select the biasing parameter of the AOARE, we suggest the estimator $\hat{k} = \frac{\hat{\sigma}^2}{[\max(\hat{\alpha}_{AE})]^2}$ by following the method of Hoerl and Kennard (1970). These findings are utilized for computing the estimated coefficients of the estimators for the model (11). Then, we use the relation given by the equation (8) to compute estimated coefficients of the model (2).

The performance of the estimators are compared according to the mse criterion defined as

$$mse(\tilde{\beta}) = \frac{1}{MCN} \sum_{mci=1}^{MCN} (\tilde{\beta} - \beta)' (\tilde{\beta} - \beta), \tag{27}$$

where $\tilde{\beta}$ is an estimator of the parameter β and MCN represents the number of Monte Carlo replications which is taken as 10000.

The results for the relative performances of the estimated values of the mse are given in the Table 1 when we choose the parameter values of β according to $\beta'\beta = 1$. Besides, the results related to the relative performances are indicated in Tables 3-5 depending on the three experiments.

The formulas and notations which are used for the relative performances in the tables are given as follows:

$$R_1 = \frac{mse(\hat{\beta}_{AOAE})}{mse(\hat{\beta}_{AE})}, R_2 = \frac{mse(\hat{\beta}_{AOAE})}{mse(\hat{\beta}_{ARE})}, R_3 = \frac{mse(\hat{\beta}_{AOARE})}{mse(\hat{\beta}_{AE})}, R_4 = \frac{mse(\hat{\beta}_{AOARE})}{mse(\hat{\beta}_{ARE})}, R_5 = \frac{mse(\hat{\beta}_{AOARE})}{mse(\hat{\beta}_{AOAE})}. \tag{28}$$

To determine that the newly defined estimators perform better, the values of R_1 to R_5 are expected to be smaller than one.

Table 1. Relative mse values of the estimators

ρ	σ	R_1	R_2	R_3	R_4	R_5
0.75	1	0.99566	1.09273	0.91127	1.00011	0.91524
	10	0.77434	1.45037	0.47781	0.89496	0.61705
	100	0.62667	1.31889	0.35760	0.75262	0.57065
0.85	1	0.99654	1.07069	0.93076	1.00002	0.93399
	10	0.79662	1.60601	0.45033	0.90789	0.56530
	100	0.62740	1.42666	0.33818	0.76900	0.53902
0.95	1	0.99715	1.12259	0.88799	0.99971	0.89053
	10	0.81463	1.82712	0.40773	0.91450	0.50051
	100	0.62802	1.55444	0.31921	0.79010	0.50828
0.99	1	0.99675	1.48009	0.67278	0.99901	0.67497
	10	0.80044	1.92757	0.37596	0.90537	0.46970
	100	0.62790	1.58674	0.31547	0.79722	0.50243

Table 2. True regression coefficients used in the Monte Carlo experiments (Chanda and Maddala, 1984)

Coefficients	Experiment I	Experiment II	Experiment III
β_0	11.60	12.23	1.095
β_1	-11.0	10.98	2.620
β_2	9.66	10.21	2.263
β_3	-9.29	8.65	3.173
β_4	8.38	7.61	5.400
β_5	-6.0	6.45	6.165
β_6	7.11	6.44	4.495
β_7	-4.43	4.69	3.839
β_8	5.2	3.99	3.199
β_9	-3.09	2.66	2.751
β_{10}	2.31	2.00	0.973
β_{11}	-1.16	2.07	-0.131

Table 3. Relative mse values of the estimators for Experiment I

ρ	σ	R_1	R_2	R_3	R_4	R_5
0.75	1	1.00042	0.99634	1.00423	1.00013	1.00380
	10	1.00027	0.99511	1.00525	1.00006	1.00498
	100	0.88304	1.05184	0.80157	0.95480	0.90774
0.85	1	1.00017	0.99733	1.00288	1.00004	1.00272
	10	1.00000	0.99536	1.00465	0.99999	1.00465
	100	0.86642	1.08161	0.76341	0.95301	0.88111
0.95	1	1.00003	0.99816	1.00188	1.00000	1.00185
	10	0.99959	0.99683	1.00266	0.99989	1.00308
	100	0.79449	1.19711	0.61520	0.92696	0.77433
0.99	1	1.00001	0.99814	1.00187	1.00000	1.00186
	10	0.99761	1.01790	0.97952	0.99944	0.98186
	100	0.69349	1.43583	0.41598	0.86127	0.59984

Table 4. Relative mse values of the estimators for Experiment II

ρ	σ	R_1	R_2	R_3	R_4	R_5
0.75	1	1.00000	0.99973	1.00027	1.00000	1.00027
	10	0.99947	0.98671	1.01295	1.00001	1.01348
	100	0.92947	1.06375	0.86366	0.98843	0.92919
0.85	1	1.00000	0.99968	1.00032	1.00000	1.00032
	10	0.99951	0.98279	1.01702	1.00001	1.01752
	100	0.93982	1.17249	0.79116	0.98703	0.84183
0.95	1	1.00000	0.99939	1.00061	1.00000	1.00061
	10	0.99951	0.96502	1.03575	1.00000	1.03625
	100	0.94882	1.48921	0.62669	0.98361	0.66049
0.99	1	1.00000	0.99767	1.00233	1.00000	1.00234
	10	0.99940	0.90936	1.09897	0.99997	1.09963
	100	0.94312	1.90703	0.48308	0.97681	0.51222

Table 5. Relative mse values of the estimators for Experiment III

ρ	σ	R_1	R_2	R_3	R_4	R_5
0.75	1	0.99997	1.00027	0.99970	1.00001	0.99973
	10	0.99780	1.07574	0.92759	1.00005	0.92964
	100	0.80091	1.41469	0.52337	0.92445	0.65347
0.85	1	0.99998	1.00030	0.99968	1.00000	0.99970
	10	0.99820	1.11314	0.89674	0.99999	0.89835
	100	0.81912	1.58717	0.48097	0.93196	0.58719
0.95	1	0.99999	1.00069	0.99931	1.00000	0.99932
	10	0.99815	1.27578	0.78235	0.99996	0.78381
	100	0.82949	1.83990	0.41889	0.92914	0.50499
0.99	1	0.99999	1.00619	0.99384	1.00000	0.99385
	10	0.99728	1.64016	0.60783	0.99966	0.60949
	100	0.81051	1.95169	0.37935	0.91347	0.46804

Due to the results in Table 1, R_1 , R_3 , R_4 and R_5 are generally smaller than one. So, we can state that the AOAE and the AOARE are better than the AE while the AOARE is superior to the ARE and the AOAE. That is, one of our newly proposed estimator, the AOARE, is the best one when we take account of the relative efficiency.

The findings of the Tables 3-4 show that the experiments indicating a damped oscillation and a declining line shape may not be preferable to examine the relative performances of the mentioned estimators. Even though there exist some cases in which the AOAE and the AOARE are preferable, a general outperformance cannot be inferred. On the other hand, the results obtained from the Table 5 in which the experiment forms a humped shape satisfy the expectations. Namely, since R_1 , R_3 and R_5 are always smaller than one, the AOAE outperforms the AE and the performance of the AOARE is better than the AE and the AOAE.

By the increase of σ , the outcomes for R_1 , R_3 , R_4 and R_5 decrease in the Table 1 and Table 4, generally. These results show us that even if σ increases, newly defined estimators continue to be effective over the others in the sense of the mse.

4. Conclusion

In the distributed lag model, biased estimation methods have been frequently preferred to mitigate the problem of multicollinearity. However, the estimation methods that minimize the mean square error criterion for the linear regression model have not been applied to the distributed lag model. This study will be a pioneer in the application of the method of minimizing the mean square error criterion in the Almon model. As a result, two new efficient estimators are defined by means of two approaches. The product of the first approach, the AOAE, is adopted from the paper of Farebrother (1975). Afterwards, the AOARE, which we expect to be more successful in comparison to the AOAE in dealing with the problem of multicollinearity, is defined. Indeed, empirical results show that the AOAE and AOARE generally outperform the AE while the AOARE often performs better than the ARE and the AOAE. Because of these satisfactory results, newly defined estimators are recommended to use in the model of distributed lag. From this point of view, different new variations of estimators that minimize the mean square error can be tested by the researchers for this model. In addition, sample properties of these estimators can be examined.

Acknowledgment

This paper is supported by Çukurova University Scientific Research Projects Unit with Project Number: FBA-2018-10169.

References

- Almon, S. (1965). “The distributed lag between capital appropriations and expenditures”, *Econometrica*, vol. 33, no. 1, pp.178-196.
- Chanda, A.K., Maddala, G.S. (1984). “Ridge estimators for distributed lag models”, *Communications in Statistics-Theory and Methods*, vol. 13, no. 2, pp.217-225.
- Dwivedi, T.D., Srivastava, V.K. (1978). “On the Minimum Mean Squared Error Estimators in a Regression Model”, *Communications in Statistics - Theory and Methods*, vol. 7, pp.487-494.
- Farebrother, R.W. (1975). “The minimum mean square error linear estimator and ridge regression”, *Technometrics*, vol. 17, pp.127-128.
- Fisher, I. (1937). “Income in Theory and Income Taxation Practice”, *Econometrica*, vol. 5, no. 1, pp.1-55.
- Frost, P.A. (1975). “Some Properties of the Almon Lag Technique When One Searches for Degree of

- Polynomial and Lag”, *Journal of the American Statistical Association*, vol. 70, no. 351, pp.606–612.
- Gujarati, D.N. (1999). *Basic Econometrics*, New York, McGraw-Hill.
- Güler, H., Gültay, B., Kaçiranlar, S. (2017). “Comparisons of the alternative biased estimators for the distributed lag models”, *Communications in Statistics-Simulation and Computation*, vol. 46, no. 4, pp.3306-3318.
- Gültay, B., Kaçiranlar, S. (2015). “Mean square error comparisons of the alternative estimators for the distributed lag models”, *Hacettepe Journal of Mathematics and Statistics*, vol. 44, no. 5, pp.1215-1233.
- Hoerl, A.E., Kennard, R.W. (1970). “Ridge regression: Biased estimation for nonorthogonal problems”, *Technometrics*, vol. 12, no. 1, pp.55-67.
- Kennedy, P.A. (2003). *Guide to Econometrics*, Cambridge, MIT Press.
- Özbay, N., Kaçiranlar, S. (2017). “The Almon two parameter estimator for the distributed lag models”, *Journal of Statistical Computation and Simulation*, vol. 87, no. 4, pp.834–843.
- Özbay, N., Toker, S. (2018a). “Almon modelinde çapraz geçerlilik kriterinin kullanımı ile ridge ve Liu tahmin edicilerin ön tahmin performansının geliştirilmesi”, 3rd International Mediterranean Science and Engineering Congress (IMSEC 2018), 24-26 October, ADANA.
- Özbay, N., Toker, S. (2018b). “Implementation of linear constraints in distributed lag model”, *International Conference on Multidisciplinary Sciences (ICOMUS 2018)*, 15-16 December, İstanbul.
- Özbay, N. (2019). “Two-parameter ridge estimation for the coefficients of Almon distributed lag model”, *Iranian Journal of Science and Technology, Transactions A: Science*, vol. 43, no. A4, pp. 1819-1828.
- Özbay, N., Toker, S. (2019a). “Efficiency of Mansson’s method: Some numerical findings about the role of biasing parameter in the estimation of distributed lag model”, *Communications in Statistics - Simulation and Computation*, <https://doi.org/10.1080/03610918.2018.1517215>.
- Özbay, N., Toker, S. (2019b). “Determination of Biasing Parameters for Almon Liu Type Estimator via a Mathematical Programming Approach”, *6th IFS and Contemporary Mathematics Conference (IFSCOM2019)*, 7-10 June, Mersin.
- Rao, C.R. (1971). “Unified Theory of Linear Estimation”, *Sankhya A*, vol. 33, pp.371–394.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, NewYork.
- Rao, C.R., Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*, Springer-Verlag, Berlin.

Schaffrin, B. (1985). “A Note on Linear Prediction within a Gauss–Markov Model Linearized with respect to a Random Approximation”, In Proceedings of the First International Tampere Seminar on Linear Statistical Models and Their Applications, ed. by Pukkila T. and Puntanen S. Tampere, Finland, Department of Mathematical Sciences/Statistics, A(138), pp.285–300.

Schaffrin, B. (1986). “New Estimation/Prediction Techniques for the Determination of Crustal Deformations in the Presence of Geophysical Prior Information”, *Technometrics*, vol. 130, pp.361–367.

Schaffrin, B. (1987). “Less Sensitive Tests by Introducing Stochastic Linear Hypotheses”, In Proceedings of the Second International Tampere Conference in Statistics, eds. by Pukkila, T. and Puntanen S. Tampere, Finland, Department of Mathematical Sciences/Statistics, Report A(184), pp.647–664.

Stahlecker, P., Trenkler, G. (1985). “On Heterogeneous Versions of the Best Linear and the Ridge Estimator”, In Proceedings of the First International Tampere Seminar on Linear Statistical Models and Their Applications, ed. by Pukkila, T. and Puntanen S. Tampere, Finland, Department of Mathematical Sciences/Statistics, Report A (138), pp.301–322.

Theil, H. (1958). *Economic Forecasts and Policy*, North-Holland, Amsterdam.

Theil, H. (1971). *Principles of Econometrics*, Wiley, New York.

Toutenburg, H. (1968). “Vorhersage im Allgemeinen Linearen Regressions Modell mit Zusatz Information über die Koeffizienten”, *Operationsforschung Mathematische Statistik*, vol. 1, pp.107–120.

Tracy, D.S., Srivastava, A.K. (1994), “Comparison of operational variants of best homogeneous and heterogeneous estimators in linear regression”, *Communications in Statistics - Theory and Methods*, vol. 23, pp.2313–2322.

Vinod, H.D. (1976). “Simulation and Extension of a Minimum Mean Estimator in Comparison with Stein’s”, *Technometrics*, vol. 18, pp.491–496.

Vinod, H.D. (1980). “Improved Stein-rule Estimator for Regression Problems”, *Journal of Econometrics*, vol. 12, pp.143–150.

Vinod, H.D., Ullah, A. (1981). *Recent Advances in Regression Methods*, Marcel Dekker, New York, Inc.

Wan, A.T.K., Chaturvedi, A. (2000). “Operational Variants of the Minimum Mean Squared Error Estimator in Linear Regression Models with Non-Spherical Disturbances”, *The Annals of Mathematical Statistics*, vol. 52, pp.332–342.

Yeo, S.J., Trivedi, P.K. (1989). “On Using Ridge Type Estimators for a Distributed Lag Model”, *Oxford Bulletin of Economics and Statistics*, vol. 51, no. 1, pp.85-90.

O-38 Determining the types of missing data under supervised statistical models

Vladimir Vasić

Faculty of Economics, University of Belgrade, Serbia, E-mail: vladimir@ekof.bg.ac.rs

Abstract – In the modern business world, more and more data is being analysed. Especially the customer relationship management department analyses customer behaviour and preferences. In order to have relevant information, companies often conduct surveys of their customers. In order for a company to obtain reliable information about its customers' preferences, it must enable its customers to answer the questions asked freely. Also, if customers think the question is intimate, they will not want to answer the question. It would be a big mistake in collection the data if the customer is forced to answer the given question. In this case, customer give an incorrect answer; and that's we least want. For this reason, we must allow the customers not to answer the question if they do not want to. In such situations, the problem of missing data arises. How the missing data will be resolved depends on the nature of missing data. If the data are missing in a completely random way, then they can be solved by standard procedures. However, if the data are missing in a random manner, then they must be addressed by modern methods of analysing the missing data. If the data are missing in a non-random way then the problem of missing data cannot be resolved with quality; and such data should not be further analysed statistically. For this reason it is very important to accurately determine the type of missing data; which will present the subject of research in this paper.

Keywords – *types of missing data, missing at random, missing completely at random, statistical tests, supervised statistical models*

1. Introduction

The presence of missing data, when collecting information from customers is quite common at conducting various surveys. Companies that want to improve their business must consider the affinities and preferences of customers. The best way to get information about customers' affinities and preferences is to conduct a survey. In surveys, customers generally answer all the questions, but there are also questions that some customers do not answer. On that way, we come to the problem of "missing data".

It is very important that the problem of missing data is addressed properly. If we do not pay enough attention to this issue, it is likely that we will get biased results when analysing the data. And that's what we least want. For this reason, it is necessary to solve the problem of missing data in a correct way.

The first step in solving the missing data problem, is to determine the nature of missing data. With regard to the nature of missing data (Fitzmaurice et al. (2015)), it has long been established that there are three types: missing data completely at random, missing data at random, and non-missing data. Only, if data is missing completely in a random way; solving missing data can be handled by traditional methods. Traditional methods include deleting an observation that has missing data, or inserting an arithmetic mean in place of missing data.

However, if the data are not missing in a completely random way, then using traditional methods to deal with missing data, is wrong. If the data are missing in a non-random manner, then any further statistical analysis of the given data set is incorrect. Suppose that if data are missing in a random way, then using modern methods of analysis of missing data, it is possible to perform statistical analysis in the correct way.

From all that has been said so far, it is important that (if the data are missing) the nature of missing data is determined accurately. This is precisely the theme of this paper, to determine exactly the nature of missing data, of data set, we want to model with supervised statistical models. In general, supervised statistical models have a target variable in addition to the predictor variables.

2. Literate Review

Regarding the supervised statistical models, the nature of missing data was determined as follows (Tabachnick and Fidell (2014)): it is first tested whether missing data are missing in a completely random way (Little and Rubin (2002)). If this hypothesis is rejected, further testing is undertaken, which tests, whether missing data are missing in a random manner. Depending of rejection or non-rejection a given hypothesis, it is concluded that the data are missing in a non-random manner or that the data are missing in a random manner.

It is already common in the literature, to test whether missing data are missing in a completely random way, with Little's *MCAR* (*Missing Completely At Random*) test (IBM (2017a)). If tested null hypothesis was rejected, it goes to testing, whether the missing data are missing in a random manner. The essence is that the available data is a simple random sample of the "complete" data set. If this is fulfilled, then the missing data has the same distribution as the available data. (Enders, C. (2010)).

The equality of distributions implies that the group of available data and the group of missing data have the same arithmetic mean as the variance. Kim and Bentler (2002) called this property the homogeneity of means and covariances. Some authors check the homogeneity of both, means and covariances (Thommes and Enders (2007)), while other authors found it difficult to assume that the data are homogeneous in arithmetic means and heterogeneous in covariances. For this reason, they examine the homogeneity of arithmetic means only (IBM (2017a)). Homogeneity of arithmetic means, is tested by univariate *t-tests*.

3. Research Methodology

So, univariate *t-test* for testing the equality of arithmetic means of two independent samples, is used to perform the *MAR* (*Missing at Random*) test. Two independent samples for each variable are formed based on unavailability of data of another variable in the analysis. If variable (noted as $X_i, i = 1, \dots, n$) has missing data, then unavailability indicator is formed (noted as M_i) that takes a value of 1 if the data is unavailable, and a value of 0 if the data is available. The indicated unavailability indicator may be presented in the form

$$M_i = \begin{cases} 0, & \text{if } X_i \text{ is response} \\ 1, & \text{if } X_i \text{ is missing} \end{cases}, i = 1, \dots, n \quad (1)$$

So, to check if a variable (X_i) has missing data that missing in a random way; we first design an unavailability indicator (M_i) for a given variable. Then we use this indicator at every another variable, for splitting into two independent samples; and then we use the univariate t -test, for testing the homogeneity of arithmetic means.

A given t -test has certain disadvantages, for example, created independent samples can be very unbalanced (Enders, C. (2010)) (the sample representing the variable availability is much larger than the sample representing the variable unavailability). Then the independent sample, which representing the variable unavailability, can have only a few elements. Also, the limitation of testing arithmetic means of two independent samples with t -test, is the fact that, even the data are missing an *MNAR* (*Missing Not At Random*) manner, the two independent samples, can have the same arithmetic means (Enders, C. (2010)).

For this reason, it is much better to suggest another way of testing the *MAR*-type of missing data. This is exactly the content of this paper, to suggest another way of testing. Another way of testing, whether the nature of missing data is *MAR*-type, could be a non-parametric test of two independent samples Mann-Whitney U test. The great advantage of this non-parametric test, over the univariate t -test is that, the Mann-Whitney test does not test the equality of arithmetic means; it tests just what is needed; that is, are the two distributions equal. So there is no need to go the roundabout, to suggest that if the two distributions are equal, then the arithmetic means and the variances are the same; we can provide directly by nonparametric two independent samples Mann-Whitney test, to tests the equality of the two distributions.

An additional important suggestion would be to use the exact method instead of the generally accepted asymptotic method to calculate the p -value at the non-parametric Mann-Whitney test. The advantage of using the exact method for calculating p -values is that (Mehta and Patel (2017)) it calculates reliable values, even in situations where the conditions for applying the t -test are not met, as the imbalance of samples. So, by using the exact method of calculating the p -value of the non-parametric Mann-Whitney test, we would overcome all the shortcomings and limitations of the t -test, which is used in practice.

Of course, in addition to the obvious advantages of the proposed process, there are some difficulties in implementing the proposed procedure. Namely, the application of the exact method of calculating the p -value is very demanding, that is, computationally intensive, based on all possible permutations of the realized set of two independent samples, which we need to test.

3.1 The Exact Method of Calculating p -Value at Nonparametric Test

For the nonparametric two independent samples Mann-Whitney test, suppose the two independent samples are sizes n_1 and n_2 . We can present the null hypothesis, which we are testing, as

$$H_0: F_1(x) = F_2(x) \quad (2)$$

which meaning that these two independent samples have the same distributions.

In the Mann-Whitney test, the original realized value is replaced with ranked values, i.e. the value x_i is replaced with $rank(x_i)$. Further, to calculate the exact p -value, we need to know the exact probability of the realized two independent samples with ranked values (Mehta and Patel (2017)). The given probability is given by the expression

$$h(w) = \frac{\prod_{i=1}^2 n_i!}{(n_1+n_2)!} \quad (3)$$

where n_1 and n_2 represent the sizes of the first and second independent sample, respectively; and design w represents the realization of given two independent samples with ranked values instead of original values; or any permutation of the same. Then the probability distribution of the Mann-Whitney statistics U to take the concrete value u , is

$$Pr(U = u) = \sum_{U=u} h(w) \quad (4)$$

where summation is performed with respect to all possible permutations of the design w . The Mann-Whitney statistic U is defined by formula (6). Similarly, the exact p -value (for the right-tail of distribution for U statistics) is

$$p_1 = Pr(U \geq u) = \sum_{U \geq u} h(w). \quad (5)$$

The previously defined formulas represent the general formulas for obtaining the exact p -value for two independent samples for non-parametric tests. Of course, each particular two independent samples non-parametric test, has a special numerical algorithm, for the exact p -value calculation. This is the content of the next sub-chapter.

3.2 Missing at Random Test for Supervised Statistical models

In this section, we will present a methodology for the exact calculation of p -values in two independent samples non-parametric Mann-Whitney test. Using the given calculated p -value, we will be able to test the null hypothesis that missing data are missing in a random manner. At the beginning of the calculation, we first need to calculate the sums of the ranks. We do this, by combining both samples into one sample, which elements are sorted by size. Then we rank the elements in a unified sample. If there the same elements, it take the average rank for them. After that (IBM (2017a)), we calculate the sum of ranks, for each independent samples (group 1 (or G_1) and group 2 (or G_2))

$$S_1 = \sum_{i \in G_1 \geq u} rank(x_i; D_1 \cup D_2) \cdot I(x_i \in D_1) \quad \text{and} \quad S_2 = \sum_{i \in G_2 \geq u} rank(x_i; D_1 \cup D_2) \cdot I(x_i \in D_2) \quad (5)$$

where D_i represents observations from the i -th independent sample ($i = 1, 2$), while $I(\cdot)$ is an indicator that takes values 0 if the expression in parentheses is incorrect, and value 1 if the expression in parentheses is correct.

The next step is to calculate the test statistic as well as the corresponding p -value. The Mann-Whitney U test statistic for the first independent sample (or for group 1) is given by the expression (IBM (2017a))

$$\begin{aligned}
 U &= \sum_{i \in G_1} \sum_{j \in G_2} I(x_i < x_j) + \frac{1}{2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} I(x_i = x_j) \\
 &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - S_1 = S_2 - \frac{n_2 \cdot (n_2 + 1)}{2}
 \end{aligned}
 \tag{6}$$

where n_1 and n_2 represent the sizes of the first and second independent samples, respectively.

After defining the test statistics U , the next step is to calculate the p -value. To do that, it will need the sampling frequency of test statistics U for the value u , and sampling size i and j of independent samples (in notation $f_{i,j}(u)$). It will be use the following property, by which the frequency distribution of test statistics U can be expressed over the sum of the lower order frequency distributions, as

$$f_{n_1, n_2}(u) = f_{n_1-1, n_2}(u - n_2) + f_{n_1, n_2-1}(u). \tag{7}$$

Each lower order distribution is symmetric with respect to the various values of the U test statistic. Also the sum of the lower order distributions gives a result that is also symmetric.

It starting with calculation the frequency distribution of the test statistics U (in the notation $f_{n_1, n_2}(u)$) with the known distribution of the test statistics U for $i = 1$. (The result will be the same, if it starts with the distribution of the test statistic U for $j = 1$). Next, it will be use the property given by expression (7), as well as the symmetry property, to derive the frequency distribution of the test statistics U , for $i = 2$. Then, this procedure will repeat until deriving the frequency distribution for $i = n_1$.

After calculating the frequency distribution for test statistics U (IBM (2017a)), we are able to calculate (two sided) exact p -values using the following expressions

$$p_1 = \begin{cases} \sum_{u=0}^{trunc(U)} f_{n_1, n_2}(u), & \text{if } U \leq \frac{n_1 \cdot n_2 + 1}{2} \\ 1 - \sum_{u=0}^{trunc(U)} f_{n_1, n_2}(u), & \text{if } U > \frac{n_1 \cdot n_2 + 1}{2} \end{cases}$$

$$p - value = 2 \cdot p_1$$

where $trunc(U)$ represents the integer value of test statistic U .

4. Data and Empirical Implementation

The data used in this paper represent the responses of potential clients in one of the leading commercial foreign banks in the Republic of Serbia. The survey was conducted in May 2019 at the main branch of a given foreign commercial bank for the city of Novi Sad. An anonymous, very short, cash loan product

survey was offered to all potential clients of the bank, who asked bank's employees about the terms and conditions of receiving a cash loan product. In particular, CRM (Customer Relation Management) analysts intended to examine the cash repayment period that potential bank customers are most interested in. They were also interested in whether the number of months of repayment of a cash loan could be affected by a client's basic socio-economic characteristics. For this reason, the survey itself had 6 simple questions, which are: "Appropriate repayment period for cash loans (in months)" (noted as Y, this is target variable); then "Age (in years)" (noted as X1, this is predictor variable), then "Years at current address" (noted as X2, this is also predictor variable), then "Household income (in thousands)" (noted as X3, this is also predictor variable (logarithm transform)), then "Years with current employer" (noted as X4, this is also predictor variable), and the last question "Number of people in household (noted as X5, this is also predictor variable).

For this research, it was planned that the sample size would be 200; which were collected. However, not all respondents answered on all the survey questions, so the problem of missing data existing. Table 1 shows the number and percentage of missing data, where variables are sorted in ascending order, relative to the percentage of missing data.

Table 1. Variable Summary

		X3	X2	X4	X1	Y	X5
Missing	N	37	29	22	7	7	4
	Percent	18.5%	14.5%	11.0%	3.5%	3.5%	2.0%
Valid N		163	171	178	193	193	196

It can be noted that the respondents did not answer the question X3 the most, with 18.5%; while the question with the least missing answers is X5 with only 2%.

The first step in statistical analysis would be to examine the nature of missing data. In this step, the first activity would be, to test the null hypothesis, which states that, the nature of missing data is *MCAR*-type. To perform this test, as well as all other statistical calculations, command syntax of the IBM SPSS Statistics software (IBM (2017b)) were used. The test results are: *Chi-Square* = 95.465, *DF* = 72, *Sig.* = 0.034, so we reject the null hypothesis, that the nature of missing data is *MCAR*-type. Based on this, it can be conclude that the classical way of solving the problem of missing data; such as deleting incomplete cases, or inserting an arithmetic mean (instead of missing value), would be – wrong.

After rejected hypothesis, it can be start with testing next null hypothesis, that the nature of missing data is *MAR*-type. As it described, it will be use a non-parametric Mann-Whitney test with exact *p-value* calculation, instead of *t-test*. The results of testing are shown in Table 2. For each variable (with missing data), an indicator for missing data is created, labelled M_X#. Also for each variable with missing data, testing was applied against the remaining variables (from supervised statistical model). Now, the significance level will be divided by the number of testing (the Bonferroni correction). So, significance level instead of classic 0.05 would be $0.05 / 5 = 0.01$.

Analysing the exact *p-value* with respect to the Bonferroni correction significance level, it can be conclude that the nature of the missing data is not *MNAR*-type, because the incompleteness of predictor variables are not related to the target variable. Also, it can be seen why the nature of data unavailability is

not *MCAR*-type, which is the same as in the previous *Little's MCAR test*. The reason is a causality of incompleteness X3 (which is a question about income) with X1 (which is question "Age"). Causality is logical, because the people who earning more (they usually are seniors, i.e. older) they more likely escape question about earnings.

Table 2. Test Statistics and exact p-values

	Y	X1	X2	X3	X4	X5
Grouping Variable: M_X1						
Mann-Whitney U	549.500	-	281.000	454.000	507.000	661.500
Exact Sig. (2-tailed)	0.935	-	0.227	0.886	0.945	1.000
Grouping Variable: M_X2						
Mann-Whitney U	2074.500	1861.000	-	1384.000	1959.000	2244.500
Exact Sig. (2-tailed)	0.391	0.159	-	0.185	0.749	0.690
Grouping Variable: M_X3						
Mann-Whitney U	2435.000	1860.000	2142.000	-	2169.000	2672.000
Exact Sig. (2-tailed)	0.197	0.001	0.747	-	0.223	0.371
Grouping Variable: M_X4						
Mann-Whitney U	1858.000	1635.000	1366.000	1390.000	-	1344.500
Exact Sig. (2-tailed)	0.927	0.484	0.492	0.842	-	0.073
Grouping Variable: M_X5						
Mann-Whitney U	193.500	296.500	124.000	199.500	99.000	-
Exact Sig. (2-tailed)	0.097	0.476	0.139	0.213	0.319	-

5. Conclusion

The process of solving missing data (which is usually the first stages of any statistical research), it is necessary to determine exactly the nature of data unavailability. If this done correctly, unavailability can be properly addressed, and the upcoming statistical findings (results) will be correct. So far, testing the *MAR*-type of data unavailability, has been conducted with *t-test*, which has limitations (as shown by numerous researchers). All these limitations, can be solved with non-parametric Mann-Whitney test, with *p-value* exactly calculated.

Predictors in supervised statistical models were numerical, so including categorical predictors in model, will be direction of further research.

Acknowledgment

The work was created as a result of research within the project of the Ministry of Education, Science and Technological Development of the Republic of Serbia No. 179005.

References

- Enders, C. (2010). *Applied Missing Data Analysis*. The Guilford Press, New York.
- Fitzmaurice, G. at al. (2015). "Missing Data: Introduction and Statistical Preliminaries", in Molenbergs, G. et al. (ed.) (2015). *Handbook of Missing Data Methodology*. CRC Press, Boca Raton, pp. 3-22

IBM (2017a). *IBM SPSS Statistics Algorithms*. IBM Corporation, Armonk.

IBM (2017b). *IBM SPSS Statistics 25 Command Syntax Reference*. IBM Corporation, Armonk.

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67, 609–624.

Little, R. & Rubin, D. (2002). *Statistical Analysis with Missing Data* (2 ed.). Wiley, Chichester.

Mehta and Patel (2017). *IBM SPSS Exact Tests 25*. IBM Corporation, Armonk.

Tabachnick, B. & Fidell, L. (2014). *Using Multivariate Statistics* (6 ed.). Pearson, Harlow.

Thoemmes, F., & Enders, C. K. (2007). *A structural equation model for testing whether data are missing completely at random*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Fitting One Life Expectancy at Birth Data Ito Stochastic Differential Equation

Aladdin Shamilov¹, Sevda Ozdemir Calikusu^{2*} and Fevzi Erdogan³

¹*Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey,
asamilov@eskisehir.edu.tr*

²*Accountancy and Tax Department, Ozalp Vocational School, Van Yuzuncu Yil University, Turkey,
sevdaoazdemir@yyu.edu.tr*

³*Department of Econometrics, Faculty of Economics and Administrative Sciences, Van Yuzuncu Yil
University, Turkey, ferdogan@yyu.edu.tr*

Abstract –In this study one life expectancy at birth data is investigated by Stochastic Differential Equation Modeling (SDEM). Firstly, parameters of SDE which occur in mentioned biological problem are estimated by using the maximum likelihood procedure. Then, we have obtained reasonable Stochastic Differential Equation (SDE) based on the given biological data. Moreover, by applying Euler-Maruyama Approximation Method trajectories of SDE are achieved. The performances of trajectories are established by Chi-Square criteria, Root Mean Square Error (RMSE) value. The results are acquired by using statistical software R-Studio. These results are also corroborated by graphical representation.

Keywords –Itô stochastic differential equation, Euler-Maruyama approximation method, Maximum likelihood

1. Introduction

It is known that SDEMs are more realistic mathematical model than normal differential equation models of the situation, Qksendal (2000). For this reason, Stochastic Differential Equation (SDE) models play a prominent role in a range of application areas, including biology, Allen (2003), chemistry, Gardiner (1985), economics and finance, Mikosch (1998), epidemiology, mechanics, microelectronics, Desmond and Higham (2001).

In this research, we have examined by SDEM the data of the average number of years a newborn child would live if current mortality patterns were to stay to same in Turkey between 1919 and 2018.

2. Material and Methods

2.1 Euler-Maruyama Method

An Ito stochastic differential equation on the interval $[0, T]$ has the form

$$dX(t) = f(t, X(t); \theta)dt + g(t, X(t); \theta)dW(t) \quad (1)$$

is considered where $\theta \in \mathbb{R}^m$ is a vector of parameters that are unknown for $0 \leq t \leq T$ where $X(0) = X_0$, $X_0 \in H_{RV}$, H_{RV} is the Hilbert space of random variables and $W(t)$ is the Wiener process.

As the exact solution to a stochastic differential equation is generally difficult to obtain, it is useful to be able to approximate the solution. Euler-Maruyama method is a simple numerical method. When applied to (1), Euler’s method has the form

$$X_{j+1,i}^{(m)} \left(t_{i-1} + (j + 1) \frac{\Delta t}{K} \right) = X_{j,i}^{(m)} \left(t_{i-1} + j \frac{\Delta t}{K} \right) + f \left(t_{i-1} + j \frac{\Delta t}{K}, X_{j,i}^{(m)} \left(t_{i-1} + j \frac{\Delta t}{K} \right) \right) \frac{\Delta t}{K} + g \left(t_{i-1} + j \frac{\Delta t}{K}, X_{j,i}^{(m)} \left(t_{i-1} + j \frac{\Delta t}{K} \right) \right) \sqrt{\frac{\Delta t}{K}} \eta_{j,i}^{(m)} \quad (2)$$

for $i = 0, 1, 2, \dots, N - 1$ and $j = 0, 1, 2, \dots, K - 1$ where $\Delta t = \frac{T}{N}$, $t_i = i\Delta t$, and $\Delta t_i = t_{i+1} - t_i = \frac{\Delta T}{K}$, $\eta_{j,i}^{(m)} \sim N(0, \Delta t)$, $\Delta W(t, w) = W(t_{i+1}, w) - W(t_i, w)$, $\Delta W(t, w) \sim N(0, \Delta t_i)$, and m indicates a simulation number. This means that by changing number K m times then m simulations realized, Shamilov (2012).

In this research, the extreme values of approximation trajectories $X_{j+1,i}^{(m)}$ by applying Euler-Maruyama Approximation Method are achieved.

2.2 Parameter Estimation for Stochastic Differential Equations

The problem is to find an estimate of the vector θ given these $N + 1$ data points. Two estimation methods are a maximum likelihood estimation method and anonparametric estimation method, Allen (2007).

2.3 General model of population biology

At population biology the SDE has the following form

$$dX = \theta_1 X(t) dt + \sqrt{\theta_2 X(t)} dW(t) \quad (3)$$

(3) equation is commonly seen in mathematical models of population dynamics.

In our investigation, using the maximum likelihood estimation method for equation (1), $\hat{\theta}_1 = 0.0119$ and $\hat{\theta}_2 = 0.048$ are obtained.

The acceptancy of equation (3) with parameters θ_1 and θ_2 to consider in our study biological data is provided with Chi-square value equal to 80.22382 of Goodness of fit test.

3. An application

One of the important application areas of stochastic differential equations is population biology. In this study, by assuming that remains the same death rates between 1919 and 2018, the average number of years a newborn child would live in Turkey (life expectancy) if current mortality patterns were to stay the

same are examined. This data set is accessed with the help of open access <https://www.gapminder.org/data/> and the data set is shown in Table 1.

Table1. The average number of years a newborn child would live in Turkey (life expectancy)

yea r	tim e	yea r	tim e	yea r	Tim e	yea r	Tim e	yea r	tim e	yea r	tim e	yea r	tim e	yea r	Tim e
191	31.	193	37.	194	37.8	195	47.5	197	57.	198	67.	199	71.	201	78.2
9	3	2	1	5	39.3	8	48.3	1	8	4	5	7	9	0	78.2
192	30.	193	37.	194	40.8	195	49.1	197	58.	198	68	199	72.	201	78.6
0	3	3	2	6	41.5	9	49.9	2	5	5	68.	8	6	1	78.4
192	30.	193	37.	194	42.3	196	50.8	197	59	198	6	199	72	201	78.8
1	4	4	3	7	43.1	0	51.6	3	59.	6	69.	9	74.	2	79
192	34.	193	37.	194	43.3	196	52.4	197	6	198	9	200	1	201	79.1
2	6	5	4	8	43.8	1	53.3	4	60.	7	70.	0	74.	3	79.3
192	36.	193	37.	194	44.3	196	54.1	197	2	198	4	200	8	201	79.6
3	7	6	5	9	44.9	2	54.9	5	60.	8	70.	1	75.	4	
192	36.	193	37.	195	45.5	196	55.6	197	9	198	1	200	4	201	
4	7	7	6	0	46.1	3	56.3	6	62	9	70.	2	75.	5	
192	36.	193	36.	195	46.8	196	57	197	62.	199	2	200	5	201	
5	8	8	8	1		4		7	9	0	70.	3	76.	6	
192	36.	193	35.	195		196		197	63.	199	2	200	3	201	
6	8	9	9	2		5		8	3	1	70.	4	76.	7	
192	36.	194	35.	195		196		197	64.	199	5	200	7	201	
7	8	0	1	3		6		9	2	2	70.	5	77.	8	
192	36.	194	34.	195		196		198	65	199	6	200	2		
8	9	1	2	4		7		0	65.	3	70.	6	78.		
192	36.	194	33.	195		196		198	8	199	7	200	1		
9	9	2	4	5		8		1	66.	4	70.	7	78.		
193	36.	194	34.	195		196		198	6	199	9	200	8		
0	9	3	9	6		9		2		5	71.	8	78.		
193	37	194	36.	195		197		198		199	4	200	2		
1		4	3	7		0		3		6		9			

In solving our problem we shall use following stages of investigation. Firstly, parameters of SDE which occur in mentioned biological problem are estimated by using the maximum likelihood procedure. Then, we have obtained reasonable Stochastic Differential Equation (SDE) based on the given biological data. Moreover, by applying Euler-Maruyama Approximation Method trajectories of SDE are achieved and they are demonstrated in Figure 1.

Using the maximum likelihood estimation method, $\hat{\theta}_1 = 0.0119$ and $\hat{\theta}_2 = 0.048$ are obtained. If these estimated parameters are considered in the SDEM (3), $dX(t)$ has the following form

$$dX(t) = 0.0119 X(t)dt + \sqrt{0.943X(t)} dW_1(t), \quad X_0 = 31.3.$$

Furthermore, the average number of years a newborn child would live in Turkey (life expectancy) and the approximate EM values of the data set $\hat{X}(t_i)$, $i = 1, 2, \dots, 100$. Approximate trajectories taken randomly according to $\hat{X}(t_i)$ for $i = 30$, $i = 100$ are given in Table 2 and Table 3, respectively.

Table 2. The average number of years a newborn child would live in Turkey (life expectancy) and the approximate $\hat{X}(t_{30})$ EM values of the data set

Tim e	EM	Tim e	EM	tim e	EM	tim e	EM	tim e	EM	tim e	EM
31.3	31.3	37.5	37.4319	44.3	43.8181	57	56.5038	69.9	68.5597	76.3	75.4733
30.3	31.2773	37.6	9	44.9	6	57.8	3	70.4	1	76.7	9
30.4	0	36.8	37.4936	45.5	44.2900	58.5	56.8638	70.1	69.9700	77.2	76.2486
34.6	30.2054	35.9	2	46.1	2	59	3	70.2	5	78.1	9
36.7	2	35.1	37.6593	46.8	44.9881	59.6	57.7488	70.2	70.4150	78.8	76.6492
36.7	30.4395	34.2	5	47.5	3	60.2	4	70.5	5	78.2	3
36.8	2	33.4	36.6516	48.3	45.4426	60.9	58.6746	70.6	70.1293	78.2	77.2685
36.8	34.7566	34.9	4	49.1	6	62	0	70.7	6	78.2	0
36.8	7	36.3	35.8753	49.9	46.1837	62.9	59.0122	70.9	70.2725	78.6	78.2204
36.9	36.6909	37.8	2	50.8	4	63.3	6	71.4	3	78.4	7
36.9	9	39.3	35.2504	51.6	46.7707	64.2	59.3485	71.9	70.1875	78.8	78.8010
36.9	36.7360	40.8	9	52.4	6	65	2	72.6	5	79	0
37	4	41.5	34.2594	53.3	47.5901	65.8	60.2167	72	70.5379	79.1	78.1479
37.1	36.6690	42.3	5	54.1	1	66.6	5	74.1	2	79.3	1
37.2	9	43.1	33.3535	54.9	48.2949	67.5	60.9191	74.8	70.6221	79.6	78.0995
37.3	36.8766	43.3	4	55.6	6	68	7	75.4	1	7	7
37.4	8	43.8	34.8620	56.3	49.1716	68.6	61.9523	75.5	70.6313	78.2547	
	36.8855		5		2		2		7	3	
	9		36.2868		50.0286		62.9291		70.9979	78.5052	
	36.8962		7		9		8		3	9	
	3		37.8198		50.9688		63.2269		71.2915	78.4060	
	36.9584		0		8		8		2	5	
	2		39.2947		51.6018		64.2154		71.7443	79.0136	
	36.8963		2		5		5		5	9	
	8		40.7532		52.3132		64.9246		72.6159	79.0733	
	36.9494		0		9		7		4	4	
	5		41.7335		53.3034		65.6691		71.9345	79.0794	
	37.1470		7		6		8		4	9	
	9		42.2330		54.0823		66.5476		74.1337	79.2648	
	37.2673		1		7		6		8	8	
	7		43.1094		54.8817		67.2708		74.8996		
	37.2525		7		9		7		7		
	5		43.2549		55.6632		68.1054		75.1895		
			5		1		9		3		

Table 3. The average number of years a newborn child would live in Turkey (life expectancy) and the approximate $\hat{X}(t_{100})$ EM values of the data set

Tim e	EM	Tim e	EM	tim e	EM	tim e	EM	tim e	EM	tim e	EM
31.3	31.3000	37.5	37.2238	44.3	43.6894	57	56.1467	69.9	68.8042	76.3	75.6302
30.3	0	37.6	0	44.9	0	57.8	4	70.4	4	76.7	1
30.4	30.3000	36.8	37.4567	45.5	44.2223	58.5	57.0903	70.1	69.9478	77.2	76.2834
34.6	0	35.9	1	46.1	6	59	3	70.2	9	78.1	7
36.7	30.3057	35.1	37.5349	46.8	44.9189	59.6	57.5612	70.2	70.8944	78.8	76.9450
36.7	1	34.2	0	47.5	6	60.2	5	70.5	8	78.2	5
36.8	30.4041	33.4	36.8629	48.3	45.5182	60.9	58.5227	70.6	70.1058	78.2	77.1924
36.8	7	34.9	4	49.1	7	62	8	70.7	0	78.2	6
36.8	34.5081	36.3	35.9198	49.9	46.0255	62.9	58.8776	70.9	70.3124	78.6	78.1204
36.9	4	37.8	7	50.8	8	63.3	8	71.4	8	78.4	6
36.9	36.7524	39.3	35.0673	51.6	46.6994	64.2	59.7730	71.9	70.0208	78.8	78.7521
36.9	4	40.8	7	52.4	6	65	1	72.6	0	79	5
37	36.9901	41.5	34.1280	53.3	47.4819	65.8	60.1833	72	70.5734	79.1	77.9246
37.1	7	42.3	9	54.1	2	66.6	3	74.1	2	79.3	5
37.2	36.8033	43.1	33.3762	54.9	48.3358	67.5	60.3629	74.8	70.1363	79.6	78.1291
37.3	5	43.3	5	55.6	7	68	6	75.4	3		5
37.4	36.6556	43.8	34.9055	56.3	49.2006	68.6	62.1757	75.5	70.6952		78.1030
	9		8		2		9		2		9
	36.9320		36.3890		49.7619		63.0441		71.0416		78.7250
	6		5		1		8		1		4
	36.7105		38.2185		50.6916		63.1877		71.2256		78.4265
	2		1		9		4		6		6
	36.8681		39.2277		51.6191		64.2581		72.0392		78.7565
	4		0		2		6		0		8
	36.6850		41.0145		52.4842		64.9031		72.8080		78.9885
	4		3		1		7		9		8
	36.9237		41.3329		53.0706		65.6661		72.1620		78.9200
	0		2		1		8		9		6
	36.9619		42.5434		54.2630		66.5069		73.9591		79.4473
	5		6		4		7		1		1
	37.0951		43.0889		54.9880		67.2981		74.7831		
	8		1		4		8		3		
	37.2988		43.3687		55.7568		67.9688		75.5420		
	8		7		9		8		7		

The performances of SDE are established by Chi-Square criteria, Root Mean Square Error (RMSE) criteria. The results are acquired by using statistical software R-Studio. These results are also corroborated by graphical representation in Figure 1.

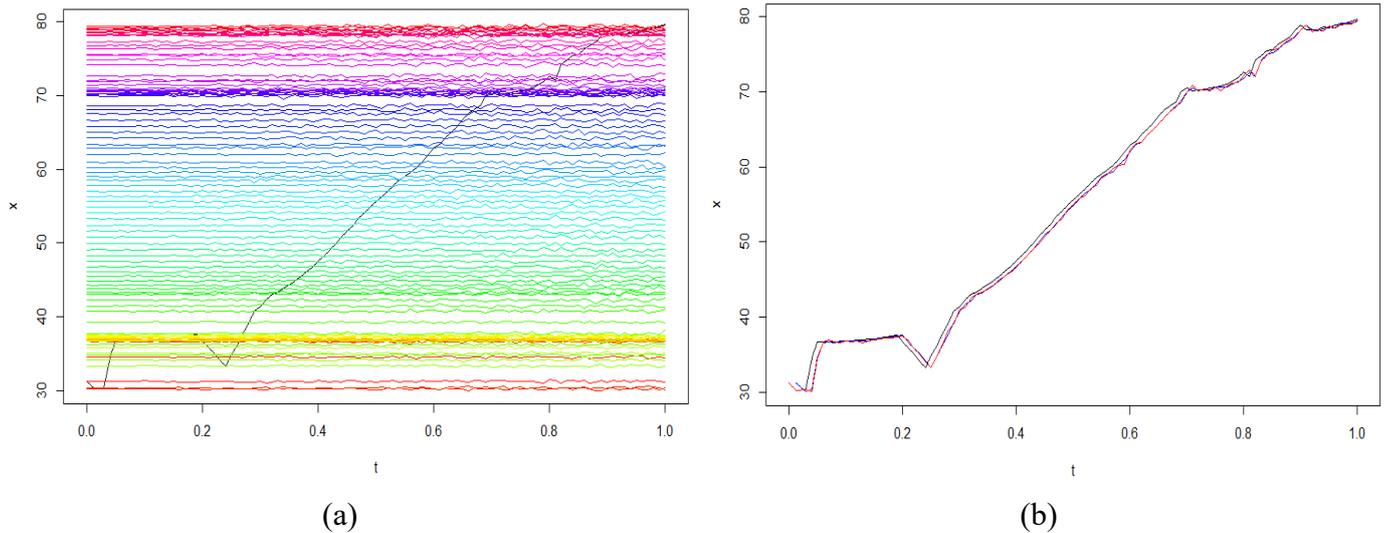


Figure 1. (a) The average number of years a newborn child would live in Turkey (black line) and forward and backward approximate EM trajectories (colored lines) of the SDD

(b) The average number of years a newborn child would live in Turkey (black line) and its EM trajectory starting from $\hat{X}(t_{30})$ (blue line) and $\hat{X}(t_{100})$ (red line)

3. Conclusion

In this study the average number of years a newborn child would live in Turkey (life expectancy) at birth data is investigated by Stochastic Differential Equation Modeling (SDEM). Firstly, parameters of SDE which occur in mentioned biological problem are estimated $\hat{\theta}_1 = 0.0119$ and $\hat{\theta}_2 = 0.048$ by using the maximum likelihood procedure. Then, we have obtained reasonable Stochastic Differential Equation (SDE) based on the given biological data. The acceptancy of equation (3) with parameters θ_1 and θ_2 to consider in our study biological data is provided with Chi-square value equal to 80.22382 of Goodness of fit test. Moreover, by applying Euler-Maruyama Approximation Method trajectories of SDE are achieved. The performances of trajectories according to $\hat{X}(t_{30})$ and $\hat{X}(t_{100})$ are established by Chi-Square criteria. The acceptancy of Chi-Square criteria for $\hat{X}(t_{30})$ and $\hat{X}(t_{100})$ is realized by values of $\chi^2 = 0.0145$; 0.0431, respectively. According to $\hat{X}(t_{30})$ and $\hat{X}(t_{100})$, Root Mean Square Error (RMSE) values are examined 0.0014 and 0.0024, respectively.

References

- Allen, E. J. 2007. Modeling with Itô Stochastic Differential Equations. *Springer*, Dordrecht.
- Mikosch, T. (1998). Elementary Stochastic Calculus, with Finance in View, World Scientific Publishing Co. Pte. Ltd., 224 pp,
- Allen, L.J.S. (2003). An Introduction to Stochastic Processes with Applications to Biology. Pearson Education Inc., Upper Saddle River, New Jersey.

Bak, J., Nielsen A. and Madsen H. (1999). “Goodness of fit of stochastic differential equations”, P. Linde, A. Holm (Eds.), Symposium iAnvendtStatistik, Copenhagen Business School, Copenhagen, Denmark.

Gardiner C. W. (1985). Hand book of Stochastic Methods for Physics, Chemistry and the Natural Sciences, Springer, Third Edition.

Higham, D.J. (2001). “An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations”, SIAM Review, 43, 525-546.

Kloeden P.E. and Platen E. (1995). Numerical Solution of Stochastic Differential Equations, Springer.

Qsendal B (2000). Stochastic Differential Equations : An Introduction with Applications, Fifth Edition, Corrected Printing, Springer-Verlag Heidelberg New York.

Shamilov A. (2014). Probability Theory with Conceptional Interpretations and Applications. Nobel Publishing House, Turkey, p.439.

Shamilov A. (2012). Differential Equations with Theory and Solved Problems, Nobel Publishing House, Turkey, p.310.

Shamilov A., Bozdog B. (2011). “Hisse Senetlerinin Fiyatlandırılması için Yeni Bir Stokastik Model Önerisi”, Journal of Statistics Research, Turkey, 8/2, 21-26.

Shamilov A. (2007). Measurement Theory, Probability and Lebesgue Integral, Anadolu University Publications, Eskisehir.

O-49 Hegy Seasonal Unit Root Test: an Application for Agricultural Products-Producer Price Index

Okan KÜREŞ^{1*}, Fatih ÇEMREK²

¹Statistic, Eskişehir Osmangazi University, Turkey, okankures@gmail.com

²Statistic, Eskişehir Osmangazi University, Turkey, fcemrek@gmail.com

Abstract

In the study, the existence of seasonal unit roots were investigated by using quarter term data of Turkey's (2003:1-2018:4) agricultural products producer price. The study was performed with HEGY test used to determine seasonal unit root. The HEGY test proposed by Hylleberg, Engle, Granger and Yoo in 1990 is one of the most commonly used tests in the literature. According to the test results obtained; For the Agricultural-PPI series, the presence of seasonal unit root at zero frequency was determined.

Key Words: HEGY Test, Seasonal Unit Root Test, Agricultural Products Producer Price Index

1. Introduction

Time series may show some changes by being affected by seasonal factors in certain periods. These changes that occur at regular intervals are called as seasonal changes or seasonality. Seasonal effects generally occur in the analysis of monthly and quarterly data. Some factors such as calendar events, holidays, weather condition, salary periods; lead to changes in series in these periods (Sevüktekin and Çınar, 2017; Bozkurt, 2007).

During the separation of the series; seasonal component also emerge together with trend, cyclical and irregular components. While there is a correlation between periods in successive monthly or quarterly time series, it is also possible to have a relationship between successive monthly and quarterly series. In such cases, seasonal variations need to be analyzed in order to model the series (Akgül, 2003).

2. Seasonal Time Series

The priority in investigating seasonal effects in time series is to determine whether the series is stochastic or deterministic. A time series can have either one deterministic or stochastic structure, or both components. On the other hand, the seasonality structure of some time series may not be expressed by both components (Saraçoğlu, 1997; Nazem, 1988).

Deterministic seasonality is a component that is valid in the long term, but whose effect disappears after a period of time, can affect the time series negatively or positively and the intensity measurement can be made exactly. This intensity is an effect that can shows an increase or decrease effect over the years. The seasonality of the series which has deterministic component can be eliminated by taking the difference of the range with the appropriate distance (Nazem, 1988).

On the other hand, stochastic seasonality has permanent shocks unlike deterministic seasonality. In a series with stochastic seasonality, the shock in a certain t-period affects not only the value of the series in that period, but also the subsequent period values. Modeling a series with stochastic seasonal effects is more complex than modeling deterministic seasonality because it requires a number of statistical parameters for analysis (Saraçoğlu, 1997; Nazem, 1988).

In the application phase, in addition to series having a regular and strong seasonality structure, series whose seasonality varies from year to year are also encountered. On the other hand, it is possible that these series which is varying from year to year, may also vary in periods when they receive maximum and minimum values. In such cases, there are two different views regarding the seasonality of the series.

Firstly, there is the view that accepts the seasonal effect as the "parasite that pollutes the economic data". According to this view, the parasite does not need to be explained and therefore it is more appropriate to use "seasonally corrected data" in the analysis phase. Such data are converted into seasonally adjusted series and published as "seasonally adjusted data".

The second view is "the way that seasonal change time series basically follows", and in the case of such series, seasonal variation in the model does not need to be explained. In other words, instead of any adjustment, it is considered appropriate to explain the relationships within the model.

In case of trend and/or seasonality effect in the series, it is possible to reach the "trend-adjusted" and "seasonally adjusted" data by following the first opinion and to make studies with the models using these data. The fluctuations encountered in these series can be annihilated by "seasonal adjustments". In this context, if trend and seasonality in the series show deterministic characteristics, it is recommended to use trend(time) and dummy variables respectively in the models. On the other hand, in the case of presence of stochastic seasonality in the series, the dummy variable and trend approach is not appropriate. It is stated that the difference receiving process is suitable for such series (Akgül, 2003).

3. Seasonal Unit Root Test

In order to obtain statistical results, the series must comply with some assumptions. The most important of these assumptions is that the series is stationary. If the series is not stationary, a stationarizing process needs to be made. Seasonality must always be taken into account in the difference-taking process, which is one of the most commonly used methods during the stagnation of a series which has seasonality (Göktaş, 2007).

The unit root in the seasonal frequencies of a time series is called seasonal unit root. In seasonal unit roots, a seasonal filter method, which is expressed by $(1 - B^4)$ for quarterly series and $(1 - B^{12})$ for monthly series, is used. B in this formula refers to the delay operator. A unit root at zero frequency has a lasting effect on the level value of the series. At the seasonal unit root, a shock applied to the series has a lasting effect in the seasonal course of the series (Leong, 1997).

There are several tests and methods used to control the seasonal unit root in time series. The most commonly used test is the Hylleberg-Engle-Granger-Yoo (HEGY) test developed for quarterly data, and the HEGY seasonal unit root test was examined in the next section of the study.

3.1. Hylleberg-Engle-Granger-Yoo (HEGY) Seasonal Unit Root Test

The Hylleberg-Engle-Granger-Yoo (HEGY) Test (1990) is a method that explains whether any series contains a unit root or not and, if the unit contains root, which type of seasonal effect leads to this unit root.

Seasonal time series are expressed in three different ways. The first is the process in which seasonality is accepted deterministic. At this stage, the interested variable is estimated by using dummy variables representing the seasonality.

$$Y_t = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 \quad (1)$$

Equation (1) states that a time series consisting of quarterly expressions is estimated with the help of three artificial variables.

The second process is the stationary seasonal process.

$$\vartheta(B) = x_t = e_t \quad (2)$$

The series is said to be a stationary series, since all of the unit roots that make $\vartheta(B) = 0$ are located outside the circle.

In the third and final process, the seasonal unit root is expressed through an autoregressive process. In this process, the following equations are reached for the quarterly series;

$$\begin{aligned} (1 - B^4)x_t = e_t &= (1 - B)(1 + B + B^2 + B^3)x_t \\ &= (1 - B)(1 + B)(1 + B^2)x_t \end{aligned} \quad (3)$$

Here, when the expression $(1 - B^4)$ is factored, four roots occur. $(1 - B)$; represents zero frequency in the long run, $(1 + B)$; represents $(1/2)$ integrated component $(1/2)$ at a frequency of six months; $(1 + B^2)$; represents an integrated component at frequencies $1/4$ and $3/4$. In the light of these findings; Four roots which are (1) , (-1) , $(-i)$ and $(+i)$ are obtained. $(-i)$ and $(+i)$ are considered as annual cycle since the unit roots cannot be separated.

Because the equation used in HEGY test is;

$$(1 - B^4)Y_t = \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \varepsilon_t \quad (4)$$

in the equation(4),

$$\begin{aligned} Y_{1,t} &= (1 + B + B^2 + B^3)x_t \\ Y_{2,t} &= (1 - B + B^2 - B^3)x_t \\ Y_{3,t} &= -(1 - B^2)x_t \\ Y_{4,t} &= (1 - B^4)x_t \end{aligned} \quad (5)$$

the equation has above values (Hylleberg vd., 1990).

In the generated model, π_1, π_2, π_3 and π_4 express the frequencies 0, 1/2, 1/4 and 3/4, respectively (Franses and Segers, 2010).

Variable of $Y_{1,t}$ is removed from unit root at 1/4, 1/2 and 3/4 frequencies. Five models can be constructed as alternative to the regression equation in Equation (4) by adding trend, constant term and seasonal dummy variable.

Model 1: There is no deterministic variable in the equation.

$$Y_{4,t} = \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \sum_{i=1}^k \beta_i Y_{4,t-i} + \varepsilon_t \quad (6)$$

Model 2: Only the constant term is added to the equation.

$$Y_{4,t} = \alpha_0 + \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \sum_{i=1}^k \beta_i Y_{4,t-i} + \varepsilon_t \quad (7)$$

Model 3: Seasonal dummy variables are added to the equation with constant term.

$$Y_{4,t} = \alpha_0 + \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \sum_{i=1}^k \beta_i Y_{4,t-i} + \varepsilon_t \quad (8)$$

Model 4: The equation has a constant term and a trend.

$$Y_{4,t} = \alpha_0 + \delta t + \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \sum_{i=1}^k \beta_i Y_{4,t-i} + \varepsilon_t \quad (9)$$

Model 5: The constant term, trend and seasonal dummy variable are included in the equation.

$$Y_{4,t} = \alpha_0 + \delta t + \pi_1 Y_{1,t-1} + \pi_2 Y_{2,t-1} + \pi_3 Y_{3,t-2} + \pi_4 Y_{3,t-1} + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \sum_{i=1}^k \beta_i Y_{4,t-i} + \varepsilon_t \quad (10)$$

The three hypotheses in the HEGY test can be expressed as follows;

$$1-) H_0: \pi_1 = 0$$

$$H_a: \pi_1 < 0$$

The t-test statistic is used to test the first hypothesis and if the hypothesis H_0 cannot be rejected, It can be deduced that "0 frequency has a unit root in the series.". In this case $(1 - B)$ operator is used to stabilize the series.

$$2-) H_0: \pi_2 = 0$$

$$H_a: \pi_2 < 0$$

T-test statistic is used to test the second hypothesis just like the first hypothesis. In the case of irrefutability of null hypothesis testing the equivalence of π_2 of $Y_{2,t}$ variable to 0, It can be deduced that there is a unit root in the $\frac{1}{2}$ frequency of the series and to stabilize the series $(1 + B)$ operator is used.

$$3-) H_0: \pi_3 = \pi_4 = 0$$

$$H_a: \pi_3 \neq 0 \text{ ve/veya } H_a: \pi_4 \neq 0$$

F test statistic is used for the third hypothesis. If the null hypothesis that tests the whether π_3 and π_4 coefficients of the lagged values of variable $Y_{3,t}$ together equal to zero cannot be rejected, then "unit root exists at frequencies $1/4$ and $3/4$." the result is reached.

If one of π_3 and / or π_4 in the series with π_2 is different from zero, there is no seasonal unit root in the series. This requires rejection of the second and third hypotheses. The fact that all π coefficients are non-zero means that there is no unit root in the series, i.e the series is stationary.

4. Data Set and Method

The data used in the study are producer price index of agricultural products (2010=100). Data covering the period 2003:1-2018:4 were obtained by conversion of data that are published monthly until the period 2017:4 on the official site of Turkey Statistical Institute (TurkStat) to quarterly data. The data for 2018 were also obtained by collecting and compiling the bulletins published as monthly by TurkStat. Eviews 10 program was used in the analysis of this study.

Before the determination of the seasonal frequencies, the logarithm of the Agriculture-PPI variable was taken in order to see the change more clearly. For the detection of frequencies; equation (10), which includes seasonal dummy variable, constant term and trend, is used as base. In terms of using quarterly data, the maximum number of lag is determined as 4 by the Schwarz Information Criteria.

In the next section; $\pi_1, \pi_2, \pi_3, \pi_4$ and (π_3, π_4) values will be compared with 0.10 and 0.05 error margins and the presence of seasonal unit root will be investigated.

5. Findings

Natural logarithm of agricultural product producer price index data was taken and then by taking the first differences of them seasonal effects and trend were observed.

Figure 1. Agricultural-PPI value chart



Figure(1) shows the quarterly data of agricultural product producer price index starting from 2003: 1 period to 2018: 4 period. When the Figure (1) is examined, it is seen that the Agricultural-PPI values tend to increase and do not show a stable process.

Figure 2. The graph of the first differences of Agricultural-PPI data

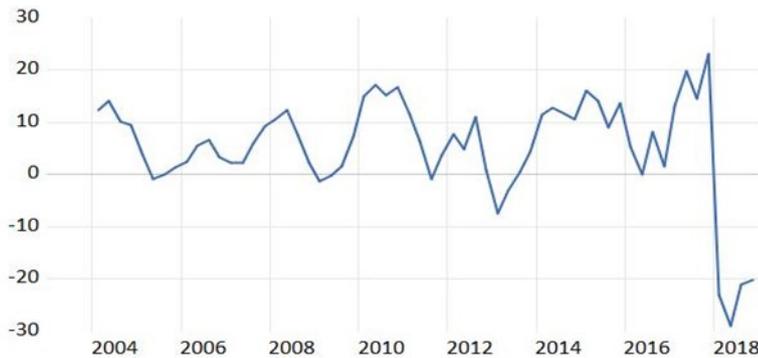


Figure (2) shows the graph of the first differences of the logarithm of Agricultural-PPI data. When this chart is examined; there is a fluctuating change with seasonal variables. After taking the first differences, it can be said that the graph has a relatively static identity.

In this series, it can be said that the seasonality process generally shows similarities. Therefore, constant terms, trends and seasonal dummy variables are included in the model and we can test whether the Agricultural-PPI data includes seasonal unit root. In order to obtain a definite result from this similarity, HEGY test was used in the next section.

Table 1. Seasonal unit root test results with HEGY test

HEGY seasonal unit root test for quarterly Agricultural-PPI values			
Number of observations: 60			
Deterministic variables: Seasonal dummy variable + constant term + trend			
Lags: 1 2 3 4			
	Test statistics	critical value (%5)	critical value (%10)
t[π_1]	-3.308	-3.668	-3.335
t[π_2]	-5.162	-3.048	-2.707
t[π_3]	-4.468	-3.618	-3.248
t[π_4]	-3.590	-1.917	-1.487
F[3-4]	22.838	6.562	5.405
F[2-4]	51.478	6.067	5.130
F[1-4]	38.609	6.516	5.703

The actual presence of seasonal unit roots in the agricultural PPI series sheds more light on possible dramatic effect of a non-optimal econometric technique may have on an economic theory test. When HEGY unit root test results were analyzed in detail, the following results were obtained.

The alternative hypothesis $H_a: \pi_1 < 0$ were tested versus the null hypothesis $H_0: \pi_1 = 0$. Since $t_h = -3.308 > t_t = -3.668$ and the null hypothesis of the coefficient π_1 cannot be rejected, it can be concluded that "there is a unit root at 0 frequency".

The alternative hypothesis $H_a: \pi_2 < 0$ was tested against the null hypothesis $H_0: \pi_2 = 0$ and it was found that $t_h = -5.162 < t_t = -3.048$. The null hypothesis of the coefficient π_1 is rejected and the result is that "there is no seasonal unit root at 1/2 frequency".

The alternative hypothesis $H_a: \pi_3 \neq 0$ and/or $H_a: \pi_4 \neq 0$ was tested against the null hypothesis $H_0: \pi_3 = \pi_4 = 0$. Since $F_h = 22.838 > F_t = 6.562$, the null hypothesis of the coefficient " π_3 and π_4 " is rejected. According to this result, it is said that "there is no seasonal unit root at 1/4 and 3/4 frequencies".

6. Conclusion

In this study, it is examined whether the agricultural producer price index series, which includes quarterly data for 2003: 1-2018: 4 period, has seasonal unit root or not. HEGY test developed by Hylleberg, Engle, Granger and Yoo (1990) was used to determine seasonal unit roots and frequency periods.

HEGY test is a method which is based on auxiliary regression, allows to determine whether there is a unit root in the series at certain frequencies and is frequently used in the examination of seasonal unit root.

As a result of the analyzes performed with the data obtained from TurkStat website; According to the test results for the Agricultural-PPI series, seasonal unit root was determined at zero frequency. It is seen that in Semi-annual frequency and 1/4 and 3/4 frequency the presence of seasonal unit root is not in question.

References

- Akgül, I., 2003, Zaman Serilerinin Analizi ve Arıma Modelleri, Der Yayınları, İstanbul, ss. 177-186.
- Bozkurt, H., 2007, Zaman Serileri Analizi, Ekin Kitabevi, Bursa.
- Franses, P. H., Segers, R., 2010, Seasonality in Revisions of Macroeconomic Data, Journal of Official Statistics, 26(2), ss. 361–369.
- Göktaş, P., 2007, Mevsimsel Zaman Serilerinde Birim Kök Testlerinin Karşılaştırmasına İlişkin Bir Simülasyon Çalışması, Yüksek Lisans Tezi, Ankara Üniversitesi.
- Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, B.S., 1990, Seasonal Integration and Cointegration, Journal of Econometrics, 44, 215-238.
- Leong, K., 1997, Seasonal Integration in Economic Time Series, Mathematics and Computers in Simulations, 18:3, ss. 413-419.

- Nazem, S.M. 1988. Applied Time Series Analysis for Business and Economic Forecasting, Marcel Dekker, Inc., Newyok and Basel.
- Saaçođlu, B., 1997, Türkiye'nin Milli Geliri ve Zaman Serisi Modelleri Yardımıyla Daimi Gelirinin Tahmin Edilmesi, Hazine Müsteşarlığı Araştırma-İnceleme Dizisi, 12, Ankara.
- Sevüktekin, M., Çınar, M., 2017, Ekonometrik Zaman Serileri Analizi, Dora Yayıncılık, Ankara.

O-51 Estimating Species Diversity Components of Various Forest Stand Types in The Lake Districts using a Draft Software for Biodiversity Estimation (BİÇEB)

Ahmet MERT^{1*}, Kürşad ÖZKAN¹, Ecir Uğur KÜÇÜKSİLLE², Halil SÜEL³, Serkan GÜLSOY¹, Murat BAŞAR⁴ and Mehmet Güvenç NEGİZ³

¹*Faculty of Forestry, Isparta University of Applied Sciences, 32260 Isparta, Turkey*

²*Computer Engineering Department, Faculty of Engineering, Süleyman Demirel University, Isparta, Turkey*

³*Sütçüler Prof. Dr. Hasan Gürbüz Vocational School, Isparta University of Applied Sciences, 32900, Isparta, Turkey*

⁴*Republic of Turkey General Directorate of Forestry, Ankara, Turkey*

* *ahmetmert@isparta.edu.tr*

Abstract – The present study was carried out to estimate the species diversity components (i.e., alpha diversity (α), beta diversity (β) and gamma diversity (γ)) of plant communities taken from different forest stand types in the Lake District, Turkey. Whittaker, Simpson, Shannon and Legendre & De Cáceres equations were employed in defining the diversity components. All computations were done by a draft software called as Biyolojik Çeşitlilik Hesaplama Programı (BİÇEB). The results of diversity components of forest stand types were then compared to each other and evaluated from ecological point of view.

Keywords – *complexity, entropy, environmental factors, plant diversity, statistical methods*

1. Introduction

There are many functions that ecosystems offer for living organisms. Biological diversity is one of the most important ecosystem functions (Mace et al., 2012). Although biodiversity does not show an equal distribution on the earth, it is reported that the highest species richness is in tropical ecosystems (especially in humid tropical forests) (Dirzo and Raven, 2003). In recent years, as a result of human activities such as fire and wood production, natural ecosystems in the tropics, especially forests, have been damaged or destroyed (Cochrane et al., 1999). In Turkey as well, forest lands have been damaged for long due to human-driven factors such as deforestation, the need for firewood and timber, grazing and fire, on the one hand, and some natural geological events such as landslides, erosion, earthquakes, etc., on the other hand (Uslu, 1973). Therefore, just like in the case of tropical forests, biodiversity is in decline in Turkey in recent years.

The importance attached to the decrease in biodiversity as a result of ongoing forest destruction has increased in the last a few years. In this way, consensus has been reached to take biodiversity protection measures in forest areas. As a matter of fact, after the United Nations Conference on Environment and Development held in Rio de Janeiro in 1992, it was emphasized that biodiversity is a key factor in forest management and policy (Summit, E., 1992). Immediately after this date, many countries began to prepare action plans for the conservation and sustainability of biodiversity. Furthermore, the duties and responsibilities regarding biodiversity have been better understood in recent years by those who prepare and implement forest management plans, and there has been a significant increase in scientific studies (Ferris and Humphrey, 1999). In other words, studies conducted in forests, which are leading among the

richest terrestrial ecosystems in terms of biodiversity, have become a critical task for countries at a local, national or global scale (Pimm and Raven, 2000; Gao et al., 2014).

In biodiversity studies conducted in forest ecosystems, plant species diversity can be said to be the most common study model representing the biodiversity (Sauberer et al., 2004; Bräuniger et al., 2010) because it is possible to indirectly represent taxonomic and genetic diversity in the field through a study on plant species diversity. For this reason, biodiversity studies in forest areas were mostly carried out with plant species diversity calculations at alpha (α), beta (β) and gamma (γ) levels. In this context, the number of different species in a region or an area indicates species richness, which is the simplest level of alpha diversity. Alpha diversity can be calculated more accurately with alternative indices. Beta diversity, on the other hand, is a measure of the variation in species composition between areas. As a product of alpha diversity, gamma diversity is the sum of species richness in a region (Dirzo and Raven, 2003). All these species diversity calculations can be performed over a large number of alternative indices (Özkan, 2016).

Studies are needed to be conducted in the forest lands of Turkey using biological diversity indices. It will thusly be possible to obtain important information about biological diversity, which has become a critical task in recent years. Biodiversity studies should be given priority especially in forest areas where endemic, rare or endangered species are concentrated. Therefore, this study was carried out in the Yukarıgökdere district, which has an important place among Mediterranean forests especially in terms of endemic species (Gundogdu, 2010; Özkan and Negiz, 2011; Özkan and Berger, 2014). In the study, plant species diversity calculations for forest stand types dominated by four different tree species were carried out through a software called “Biyolojik Çeşitlilik Hesaplama Programı (BİÇEP)”. Thus, with this study, both forest stand types with potentially high biodiversity in Yukarıgökdere district were identified and also the BİÇEP software was tested in diversity calculations. As a result, it was aimed to obtain information on forest stand types for planning and organization related to management of diversity in the district.

2. Materials and Methods

2.1 Material

The study area is Yukarıgökdere district in Isparta, which is situated between the northern latitudes of 37° 35' – 37°-50' and the eastern longitudes of 30° 50' -30° 25' within the M25 b4-c1 map sections (Figure 1).

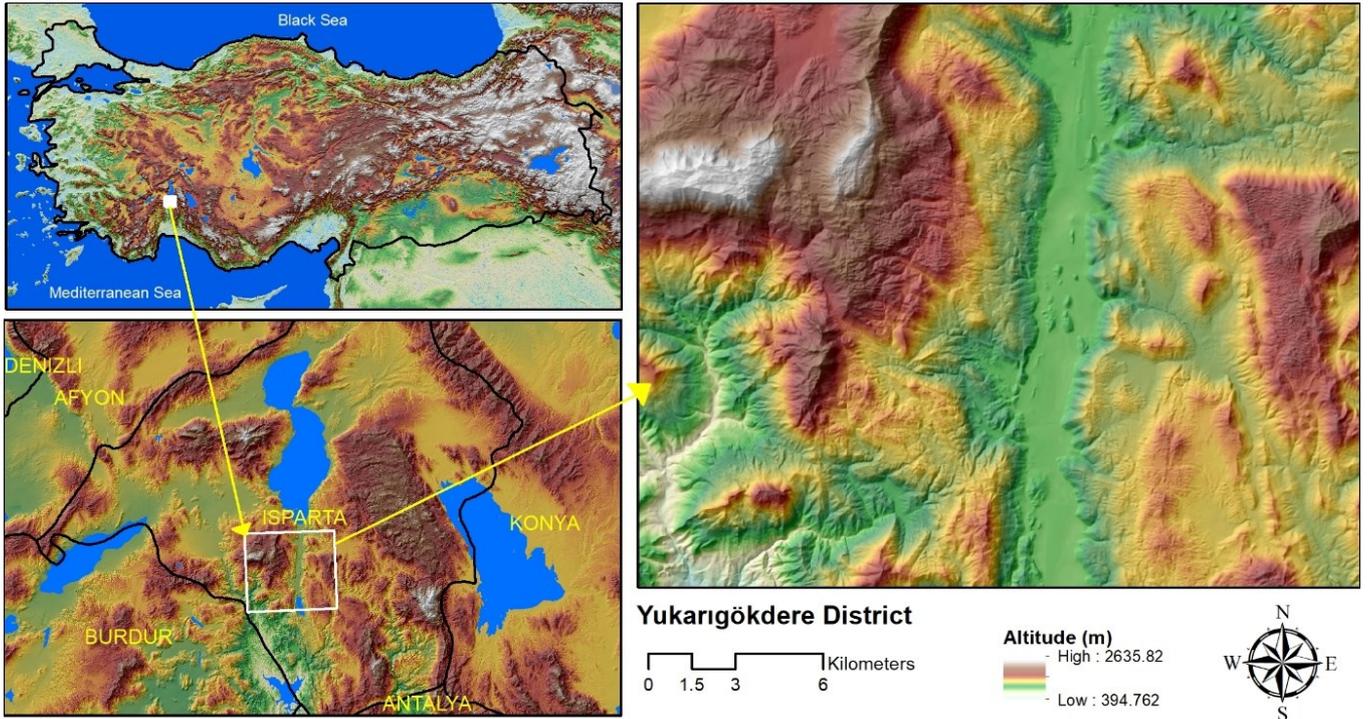


Figure 1. The study area: Yukarıgökdere district

The study area consists of a total of 14.667 ha including 10.899 ha of forest and 3.768 ha of open area. It is surrounded by Eyüpler Village and Karatepe Hill (1679 m) at its north, Çukurköy Village at its south, Mount Asacak (1720 m) at its west and the Eğirdir-Sütçüler highway at its east. Kasnak Oak Forest Nature Reserve, which covers an area of 1300 ha, is also located in this district that is situated between the heights of 900 m and 1900 m approximately.

According to the 1993-1915 data from Eğirdir (Isparta) Meteorological Station, which is the closest station to the district, the average annual temperature is 13.1°C, the hottest months are July and August with 23.8°C, the coldest month is January with 3.3°C and the average annual rainfall is 744.4 mm (DMI, 2015). Although the Mediterranean climate is dominant in the district in terms of climatic characteristics, a transitional climate towards continental climate is also observed in part (Negiz, 2009).

Geological formations of the study area are generally allochthonous units. These formations are mostly represented by ophiolitic mélangé on Mesozoic basement, and thick conglomerate and limestone layers upwards, while polygenic limestone and collapsed limestone are also rarely observed (Atayeter, 2005). In terms of structure, the soil of Yukarıgökdere district is clay loam, salt-free, slightly alkaline, chalky, phosphorus-poor, potassium-rich and poor in organic matter. Thus, it is seen that most of the study area consists of red-brown Mediterranean soils followed by limeless brown forest soil and brown forest soil, respectively (KHGM, 1994; KHIİM, 1997).

Phytogeographically, the study area is located in the Mediterranean flora region. However, Iran-Turan flora elements are also frequently encountered in the district due to the fact that the flora regions with continental character can easily interpenetrate and that there are important floristic relationships between the Mediterranean and Iran-Turan regions (Yaltırık and Efe, 1989). In a study conducted by Özen and Fakir (2015) on the flora of Yukarıgökdere district, 78 families, 253 genera and 442 taxa were identified

in the district. Of these taxa, 69 were reported to be endemic. According to another study carried out, the most densely distributed species in the district were Crimean juniper (*Juniperus excelsa*), prickly juniper (*Juniperus oxycedrus*), Turkish oak (*Quercus cerris*) and storax (*Styrax officinalis*) (Özkan and Negiz; 2011).

The study was carried out with a total of 44 sample plot data and these data were readily supplied from the sample plots studied by Negiz (2009). These data include plant species, their cover-abundance values determined according to the Braun Blanquet scale, and topographic (altitude, slope, surface stoniness) and soil depth parameters. Of the selected sample plots, 13 belong to Taurus cedar, 13 to Anatolian black pine, 11 to Kasnak oak and 7 to Red pine stands. Plant coverage inventory data of the plant species were then translated into numerical values as suggested by Westhoff and Maarel (1973). Thus, the data matrix was made available for diversity calculations.

2.2 Species Diversity Calculations

The sample plots-species data table was transferred to the BİÇEP software and species diversity calculations were made at alpha, beta and gamma levels using the indices given below.

Species richness (S) was calculated first in the alpha diversity calculation (Peet, 1974).

$$S = \sum_{i=1}^S S_i \quad 1$$

Shannon-Wiener entropy (H') (Shannon, 1948) and Simpson index (λ) (Simpson, 1949) were utilized to calculate alpha species diversity using abundance data.

$$H' = -\sum p_i \ln p_i \quad 2$$

$$\lambda = \sum_{i=1}^s p_i^2 \quad 3$$

In equation 2 and equation 3, p_i represents the proportional values of the species.

Beta diversity can be determined using both binary (presence-absence) and countable data. Beta diversity based on presence-absence data was determined by Whittaker's (1960) (β_w) equation.

$$\beta_w = \frac{\gamma}{\bar{\alpha}} - 1 \quad 4$$

where γ represents the gamma diversity of the upper community and $\bar{\alpha}$ represents the alpha diversity average of the lower communities that make up the upper community.

Beta diversity (β_{total}) (Legendre and De Cáceres, 2013) as the variance of community data, universal beta diversity with Simpson index (D_T) (Lande, 1996) and universal beta diversity with Shannon index (H_{within}) (Lande, 1996) were employed respectively for the calculation of universal beta diversity with countable data. These formulas are listed below.

$$\beta_{total} = SS_{total}/(n-1) \quad 5$$

As shown in the formula above, total beta diversity (β_{total}) is obtained by dividing the total distribution (SS_{total}) value by the degree of freedom ($n-1$). The n in the formula represents the number of sample plots in the data matrix.

$$D_T = D_{within} + \bar{D}_{with} \quad 6$$

D_T represents the total species diversity of the upper community, while D_{within} represents Simpson beta diversity and \bar{D}_{with} represents the average of the alpha diversity values of the lower communities.

$$H_{within} = - \sum_i \bar{p}_i \ln \bar{p}_i - \sum_j q_j H_j \quad 7$$

where \bar{p}_i is the average of the proportional values of the species. H_j represents the Shannon index values of each lower community in the upper community, q_j is the proportional weight and its value is equal to 1/lower community.

The gamma diversity calculation was calculated over the total frequency of the species determined in each forest stand type.

3. Results

As a result of the land inventory carried out in the study, 71 different plant species in sample plots in which Kasnak oak dominated, 67 in Taurus cedar, 51 in Anatolian black pine, and 35 in red pine sample plots, respectively, were determined. These values were gamma values of species richness in the stand where each different species was dominant. As a result of the alpha diversity measurements, the findings related to the alpha diversity calculations determined in the sample plots studied in each forest stand type are given in Table 1.

In beta diversity calculations, Whittaker beta diversity (β_w), that was applied for binary (presence-absence) data type, beta diversity as a variance of community data (β_{total}), universal beta diversity by Simpson index (D_T) and universal beta diversity by Shannon index (H_{within}) were calculated, respectively (Table 2).

Table 1. Alpha diversity values of the sample plots within the forest stand types

	Sample Plot	H'	λ	S		Sample Plot	H'	λ	S
Taurus cedar	cd1	2,507	0,904	15	Anatolian black pine	ck1	2,116	0,858	10
	cd2	2,946	0,940	22		ck2	2,272	0,885	11
	cd3	1,926	0,832	8		ck3	1,657	0,781	6
	cd4	2,722	0,927	17		ck4	2,660	0,917	17
	cd5	2,393	0,891	13		ck5	2,222	0,881	10
	cd6	2,717	0,924	17		ck6	3,009	0,942	23
	cd7	2,693	0,920	17		ck7	2,503	0,905	14
	cd8	2,354	0,890	12		ck8	2,269	0,883	11
	cd9	2,983	0,945	21		ck9	2,538	0,911	14
	cd10	2,572	0,920	14		ck10	2,264	0,878	11
	cd11	2,630	0,921	15		ck11	2,661	0,920	16
	cd12	2,620	0,919	15		ck12	1,973	0,845	8
	cd13	2,205	0,878	10		ck13	2,321	0,894	11
	Average	2,559	0,908	15,1		Average	2,343	0,885	12,5
	Sample Plot	H'	λ	S		Sample Plot	H'	λ	S
Kasnak oak	qv1	3,336	0,960	31	Red pine	cz1	2,272	0,885	11
	qv2	2,430	0,882	15		cz2	2,409	0,896	13
	qv3	2,204	0,875	10		cz3	2,635	0,921	16
	qv4	2,717	0,924	17		cz4	2,507	0,904	15
	qv5	2,693	0,920	17		cz5	2,296	0,889	11
	qv6	2,860	0,937	19		cz6	1,657	0,781	6
	qv7	2,538	0,913	14		cz7	2,222	0,881	10
	qv8	2,572	0,920	14		Average	2,286	0,880	11,7
	qv9	2,472	0,906	13					
	qv10	2,630	0,921	15					
	qv11	2,350	0,885	12					
	Average	2,618	0,913	16,1					

Table 2. Beta diversity values calculated for different forest stand types

	β_w	β_{total}	D_T	H_{within}
Taurus cedar	3,444	0,499	0,951	1,007
Anatolian black pine	3,093	0,473	0,937	0,865
Kasnak oak	3,412	0,523	0,956	2,549
Red pine	1,988	0,382	0,919	0,668

3. Conclusion and Discussion

Determining the general plant species diversity in forests based on stand types is important for obtaining indicators of biological diversity (Chirici et al., 2012; Gao et al., 2014). Thus, the stand types in forests (species), structural parameters (forest canopy cover, old tree and deadwood ratio etc.) and functional parameters (productivity, food chain, forest destruction etc.) make it possible to obtain information about

biological diversity (Noss, 1990; Ferris and Humphrey, 1999; Larsson, 2001). In fact, it was argued that certain species or combinations of species especially at a stand level can be a good indicator of biological diversity (Noss, 1990). It is possible to come across some important studies in the literature. In an exemplary study conducted on the subject, species composition and plant species diversity in *Fagus sylvatica* natural forests and *Picea abies* forestation areas were compared and although no significant difference was found in species diversity, it was determined that there were significant differences in plant species composition (Máliš et al., 2010). In another study on the subject, poplar stands were determined to have a higher vascular plant species richness compared to pine and spruce stands (Reich et al., 2001).

In all of the studies on stand type-species diversity relationships, it is stated that plant species diversity can be a good indicator especially representing general biological diversity (Sauberer et al., 2004; Bräuniger et al., 2010). It is also believed that the results obtained from such studies can provide low cost, practical and ecologically meaningful information for those who prepare and implement forest management plans (Ferris and Humphrey, 1999). It is important to conduct these studies especially in areas with high potential for endemic and rare species. In the present study, which was carried out in Yukarıgökdere district where the Kasnak oak (*Q. vulcanica*) as one of the most important endemic tree species of Turkey has formed a stand, plant species diversity was determined for alpha (α), beta (β) and gamma (γ) levels for four different stand types. The results showed that plant species richness in Kasnak oak stands was higher than Anatolian black pine, Taurus cedar and red pine stands in all three plant species diversity levels.

The stand types with the lowest diversity in the area are located at lower altitudes dominated by red pine species. Although the red pine is the most widespread tree species in the Mediterranean region, difficult site conditions such as high temperature and shallow soil are present in many areas where this species is distributed in the district (Çelik and Özkan, 2015). It is also observed that the red pine stands that are near settlements as a result of being at lower elevations are under intense human pressure and this situation has affected the diversity. Additionally, the study found the diversity of Anatolian black pine and Taurus cedar stands to be very similar to each other, and this situation is caused by the similarity of the ecological conditions of these two species in the district.

Consequently, plant species diversity calculations were made for four different stand types in the present study, and significant differences were found based on the results obtained. Therefore, the relationships between the stand types involved in forest management plans and diversity can be made more functional by increasing such studies. Moreover, a number of other structural parameters such as tree diameter and length in stands, vertical canopy closure in intermediate and lower layers, degree of crown cover and dead wood density can be included in such studies to obtain information about important indicators of biodiversity.

Acknowledgment

In this study, Draft Software for Biodiversity Estimation (BİÇEP) was used. We would like to thank TÜBİTAK for financial support to the software with the project numbered TÜBİTAK-117O983.

References

- Asfaw, A. G. (2018). “Woody species composition, diversity and vegetation structure of dry afro-montane forest, Ethiopia”, *Journal of Agriculture and Ecology Research International*, pp. 1-20.
- Atayeter, Y. (2005). Aksu Çayı Havzası'nın Jeomorfolojisi, Fakülte Kitabevi Yayınları, No:55, Coğrafya Dizisi :1, Isparta.
- Bräuniger, C., Knapp, S., Kühn, I., Klotz, S. (2010). “Testing taxonomic and landscape surrogates for biodiversity in an urban setting”, *Landscape and Urban Planning*, vol. 97, no. 4, pp.283-295.
- Chirici, G., McRoberts, R.E., Winter, S., Bertini, R., Brändli, U.B., Asensio, I.A., Bastrup-Birk, A., Rondeux, J., Barsoum, N., Marchetti, M. (2012). “National forest inventory contributions to forest biodiversity monitoring”, *For. Sci.*, vol. 58, pp. 257–268.
- Cochrane, M. A., Alencar, A., Schulze, M. D., Souza, C. M., Nepstad, D. C., Lefebvre, P., Davidson, E. A. (1999). “Positive feedbacks in the fire dynamic of closed canopy tropical forests”, *Science*, vol. 284, no. 5421, pp. 1832-1835.
- Çelik, H., Özkan, K. (2015). “Antalya Ovacık Dağı Yöresi'nde kızılçam (*Pinus brutia* Ten.)'ın gelişimi ile yetişme ortamı özellikleri arasındaki ilişkiler”, *Journal of Natural & Applied Sciences*, vol. 19, no. 2, pp. 190-197.
- Dirzo, R., Raven, P. H. (2003). “Global state of biodiversity and loss. Annual review of Environment and Resources”, vol. 28, pp. 137–167.
- DMİ, (2016). Devlet Meteoroloji İşleri Genel Müdürlüğü, Türkiye Meteorolojik Veri Arşiv Sistemi (TMVAS). 1993-2015 Yılları arası Sinoptik Klima ve Otomatik istasyon verilerini değerlendirme raporu (Sayısal veri), Disket I. Ankara.
- Gao, T., Hedblom, M., Emilsson, T., Nielsen, A. B. (2014). “The role of forest stand structure as biodiversity indicator”, *Forest Ecology and Management*, vol. 330, pp. 82-93.
- Gundogdu, E. (2010). “Conservation strategies on bird and mammal species in Yukarigokdere-Isparta, Turkey”, *Journal of Animal and Veterinary Advances*, vol. 9, no. 9, pp. 1338-1344.
- KHGM, (1994). Isparta İli Arazi Varlığı. T.C. Başbakanlık Köy Hizmetleri Genel Müdürlüğü Yayınları, 97, Ankara.
- KHIİM, (1997). Toprak ve Su Tahlil Laboratuvarı Verileri. Köy Hizmetleri Isparta İl Müdürlüğü, Isparta.
- Larsson, T.B. (2001). “Biodiversity evaluation tools for European forests”. *Criteria and Indicators for Sustainable Forest Management at the Forest Management Unit Level*, pp. 75-81.

Lande R. (1996), “Statistics and partitioning of species diversity and similarity among multiple communities”, *Oikos*, vol. 76, no. 1, pp. 5-13.

Legendre P., De Cáceres M. (2013). “Beta diversity as the variance of community data: dissimilarity coefficients and partitioning”, *Ecology Letters*, vol. 16, no.8, pp. 951-963.

Mace, G.M., Norris, K., Fitter, A.H. (2012). “Biodiversity and ecosystem services: a multilayered relationship”, *Trends Ecol. Evol.*, vol. 27, pp. 19–26.

Máliš, F., Vladovič, J., Čaboun, V., & Vodálová, A. (2010). “The influence of *Picea abies* on herb vegetation in forest plant communities of the Veporské vrchy Mts.”, *Journal of Forest Science*, vol. 56 no. 2, pp. 58-67.

Negiz, M.G., (2009). Isparta-Yukarıgökdere (Eğirdir) Yöresi'ndeki Odunsu Vejetasyonun Sınıflandırılması ve Haritalanması, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Orman Mühendisliği Anabilim Dalı Yüksek Lisans Tezi, 122 p. Isparta.

Nguyen, T. V., Mitlohner, R., Bich, N. V., Do, T.V. (2015). “Environmental factors affecting the abundance and presence of tree species in a tropical lowland limestone and non-limestone forest in Ben En National Park, Vietnam”, *Journal of Forest and Environmental Science*, vol. 31, no. 3, pp. 177-191.

Noss, R.F. (1990). “Indicators for monitoring biodiversity: a hierarchical approach”, *Conserv. Biol.*, vol. 4, pp. 355–364.

Özdamar, K. (1999). Paket programlar ile istatistiksel veri analizi-1: SPSS-MINITAB. Kaan Kitabevi, Eskişehir.

Özen, M., Fakir, H. (2015). “Isparta kasnak meşesi tabiatı koruma alanı ve çevresinin florası”, *Journal of Natural & Applied Sciences*, vol. 19, no. 3, pp. 48-65.

Ferris, R., Humphrey, J.W. (1999). “A review of potential biodiversity indicators for application in British forests”, *Forestry* vol. 72, pp. 313–328.

Özkan, K. 2003. Beyşehir Gölü Havzası'nın Yetiştirme Ortamı Özellikleri ve Sınıflandırılması, İstanbul Üniversitesi Fen Bilimleri Enstitüsü Doktora Tezi, 189p. İstanbul.

Özkan, K. (2016). Biyolojik Çeşitlilik Bileşenleri (α , β ve γ) Nasıl Ölçülür. Süleyman Demirel Üniversitesi, Orman Fakültesi Yayın No: 98, 142 p., Isparta.

Özkan, K., Berger, U. (2014). “Predicting the potential distribution of plant diversity in the Yukarıgökdere forest district of the Mediterranean region”, *Polish Journal of Ecology*, vol. 62, pp. 441–454.

Özkan, K., Negiz, M.G. (2011). “Isparta Yukarıgökdere Yöresi'ndeki odunsu vejetasyonun hiyerarşik yöntemlerle sınıflandırılması ve haritalanması”, *SDU Orman Fakültesi Dergisi*, vol. 12, pp. 27-33.

- Peet, R.K. (1974), “The measurement of species diversity”, *Ann. Rev. Ecol. System.*, vol. 5, pp. 285-307.
- Pimm, S. L., Raven, P. (2000). “Biodiversity: extinction by numbers”, *Nature*, vol. 403, no. 6772, pp. 843-845.
- Reich, P. B., Bakken, P., Carlson, D., Frelich, L. E., Friedman, S. K., Grigal, D. F. (2001). “Influence of logging, fire, and forest type on biodiversity and productivity in southern boreal forests”, *Ecology*, vol. 82, no. 10, pp. 2731-2748.
- Sauberer, N., Zulka, K. P., Abensperg-Traun, M., Berg, H. M., Bieringer, G., Milasowszky, N., Moser, D., Plutzer, C., Pollheimer, M., Storch, C., Tröstl, R., Zechmeister, H., Grabherr, G. (2004). “Surrogate taxa for biodiversity in agricultural landscapes of eastern Austria”, *Biological Conservation*, vol. 117, no. 2, pp. 181-190.
- Shannon C. E. (1948), “A mathematical theory of communication”, *The Bell System Technical Journal*, vol. 27, pp. 379-423.
- Simpson, E. H., (1949), “Measurement of diversity”, *Nature*, pp. 163: 688.
- Summit, E. (1992). *The United Nations Conference on Environment and Development*. Rio de Janeiro, pp. 3-14.
- Uslu, S. (1973). “Türkiye’de orman tahribatı ve doğurduğu problemler”, *İstanbul Üniversitesi Orman Fakültesi Dergisi*, pp. 40-47.
- Westhoff, V. van der Maarel, E. (1973) *The Braun-Blanquet Approach*. In: Whittaker, R.H., Ed., *Ordination and Classification of Communities*, Dr. W. Junk, Dordrecht, pp. 617-626.
- Whittaker R., H. (1960), “Vegetation of the Siskiyou Mountains, Oregon and California”, *Ecological Monographs*, vol. 30, no.3, pp. 279-338.

O-52 Estimating Community Rarity by creating the Locked Matrix and using the Tsallis Entropy

Kürşad Özkan^{1*}

¹ Faculty of Forestry, Isparta University of Applied Sciences, 32260 Isparta, Turkey

* kursadozkan@isparta.edu.tr

Abstract – Tsallis entropy (qS) is the generalized form of Shannon entropy. q is the Tsallis index. By using various values of the Tsallis index (q), a diversity profile of a community can be created. q^* represents the value corresponding to the point in the profile where relative differences among qS and ${}^qS^{max}$ values are maximized. Evenness and entropy corresponding to q^* value are symbolized as ${}^{q^*}E$ ve ${}^{q^*}S$ respectively.

The present study was carried out to indicate how to calculate rarity at the community level from the locked matrix (Matrix H) derived from inventory data (Matrix A). Tsallis entropy and its' related equations were applied for the calculations. ${}^{q^*}E$ and ${}^{q^*}S$ were symbolized as ${}_{H}^{q^*}E$ and ${}_{H}^{q^*}S$ since the computations were based on Matrix H.

A hypothetical ecological community data including 5 communities and 15 species was used to explain how to create a locked matrix. q^* , ${}_{H}^{q^*}E$ and ${}_{H}^{q^*}S$ values were defined for each of the hypothetical communities. Findings indicated that q^* and ${}_{H}^{q^*}E$ are favorable parameters to describe the community rarity.

Keywords – biodiversity, ecology, ecosystem components, entropy, generalization

1. Introduction

An enormous number of indices have been proposed to estimate species diversity. The most popular basic diversity indices are species richness (the number of species), Simpson index and Shannon entropy (Özkan, 2016a). Shannon entropy was first introduced by Claude Shannon (Shannon, 1948). Afterwards, its' parametric generalizations were improved by many others such as A. Rényi (Rényi, 1961), C. Tsallis (Tsallis, 1988) and G. Kaniadakis (Kaniadakis, 2001).

Tsallis entropy contains S ($q = 0$), H ($q \rightarrow 1$) and $1 - D$ ($q = 2$). It has also a bias-corrected form (Macron et al. 2014). It is therefore an available equation to estimate species diversity. Besides Mendes et al. (2008) explored that q^* or Eq^* obtained from Tsallis entropy may represent community rarity. In their approach, calculations of q^* and Eq^* of a community are based solely on the species data taken from that community. However, as pointed out by Özkan (2016b), Fattorini (2008), Leroy et al.(2013), Leroy et al.(2015), Hussain et al. (2008), Palmer et al. (2002), Dennis et al. (2000), and Borges et al., 2000, estimating species rarity of a community is related not only to species data of that community but also to those of the other communities. It means that the rarity value of a community is not independent from species data of the other communities.

By considering all ecological communities and using the proposed index of Mendes et al. (2008), an accurate and reasonable approach might be offered in estimation of the rarity of a target ecological community.

The present paper addresses how to integrate the unified index of Mendes et al. (2008) and species data of the other communities by offering the equations and applying them to a hypothetical data.

2. Material and Method

2.1 Method

The core equation employed in the present study is Tsallis entropy (qS) which is given by (Tsallis, 1988; Macron et al., 2014):

$${}^qS = \frac{1 - \sum_{i=1}^W p_i^q}{q-1} = - \sum_{i=1}^S p_i^q \ln_q p_i \quad (1)$$

Where W is the number of states. p_i is the probability of the state i and, q the Tsallis index.

The parameters regarding community rarity are estimated across an inventory data matrix and a locked matrix.

Figure 1 explains how to create a locked matrix. First of all, an inventory data matrix (Matrix A) (Figure 1A) is converted to presence absence species data matrix (Matrix B) (Figure 1B). From that matrix, the frequency values of each species are defined (the column nearby Matrix B). Later, Species frequency values are loaded to the relevant cells. Thus Matrix C is created (Figure 1C). Matrix C is the first half of the main matrix. Additionally two matrices (Matrix D and Matrix E) are prepared (Figure 1D and Figure 1E). All the cells of Matrix D are composed of maximum f_{i+} value whereas negative value matrix (Matrix E) is derived from matrix B. Both of the matrices (D and E matrices) are used to create the other half of the main matrix (Matrix F) (Figure 1F). Lastly matrix C and matrix F are combined into one matrix (Matrix G) (Figure 1G). Matrix H is the locked matrix including the relative values derived from matrix G where the total relative values of each community is equal to 1 (Figure 1H).

The probability values are computed for each community (C_j) from the locked matrix as follows:

$$\dot{p}_i = \frac{x_i^0 f_{i+}}{\sum x_i^0 f_{i+} + \sum (\max f_{i+} + 1 - x_i^0)} \quad (2)$$

$$\dot{p}'_i = \frac{\max f_{i+} + 1 - x_i^0}{\sum x_i^0 f_{i+} + \sum (\max f_{i+} + 1 - x_i^0)} \quad (3)$$

$$\sum \dot{p}_i + \sum \dot{p}'_i = \sum \hat{p}_i = 1 \quad (4)$$

Computations of ${}^q_H S$, ${}^q_H E$ and q^* are based on the idea of Mendes at all. (2008). According to this idea, first of all ${}^q_H S$ and ${}^q_H S^{max}$ profiles should be formed. The family of ${}^q_H E$ profile is derived from both of them. Note that the symbol H found in the entropic equations represents “matrix H”.

The following equations are used to create the profile curves of ${}^q_H S^{max}$, ${}^q_H S$ and ${}^q_H E$ for each community (C_j).

$${}^q_H S = \frac{1 - \sum_{i=1}^W \hat{p}_i^q}{q-1} \quad (5)$$

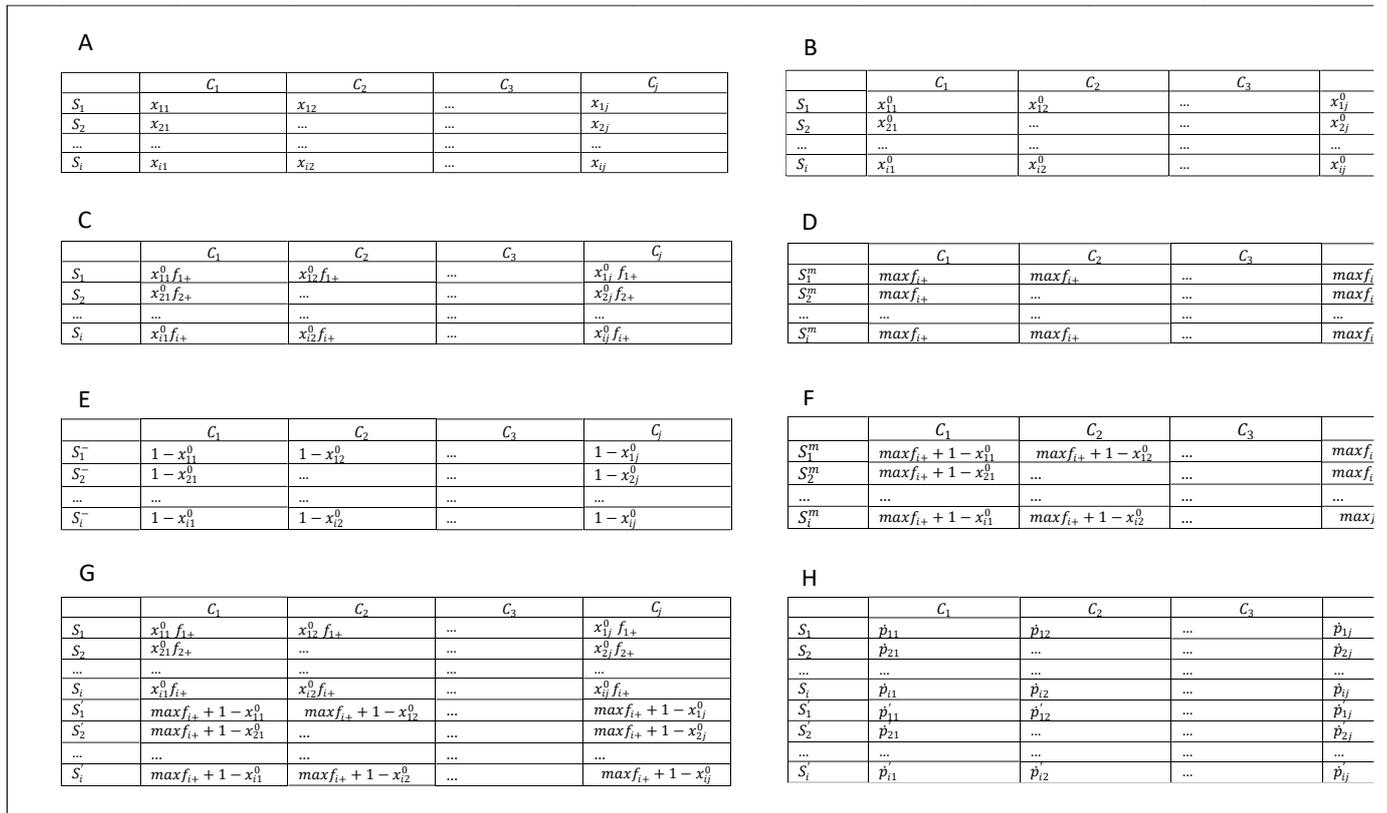
$${}^q_H S^{max} = \frac{1 - W^{1-q}}{q-1} \quad (6)$$

$${}^q_H E = \frac{{}^q_H S}{{}^q_H S^{max}} \quad (7)$$

Where W is the number of states of a given community j (the number of species of C_j + the total number of the species including all the plots (S'_i)). \hat{p}_i^q is the probability of the state i . q denotes entropic index. ${}^q_H S$, ${}^q_H S^{max}$ and ${}^q_H E$ are the Tsallis entropy, the maximum value of ${}^q_H S$ and evenness index labelled by q respectively.

As explained by Mendes et al. (2008), each family ${}^q_H E$ contains a minimum value for each C_j . A maximum contrast between ${}^q_H S^{max}$ and ${}^q_H S$ occurs at the q value corresponding to the minimum value of ${}^q_H E$ (q^*). In other words, q^* represents the position in the curve where relative differences among ${}^q_H S$ and ${}^q_H S^{max}$ values are maximized. ${}^q_H E$ and ${}^q_H S$ corresponding to q^* are ${}^{q^*}_H E$ and ${}^{q^*}_H S$ respectively. Since computations are based on Matrix H , those can be symbolized as ${}^{q^*}_H E$ and ${}^{q^*}_H S$.

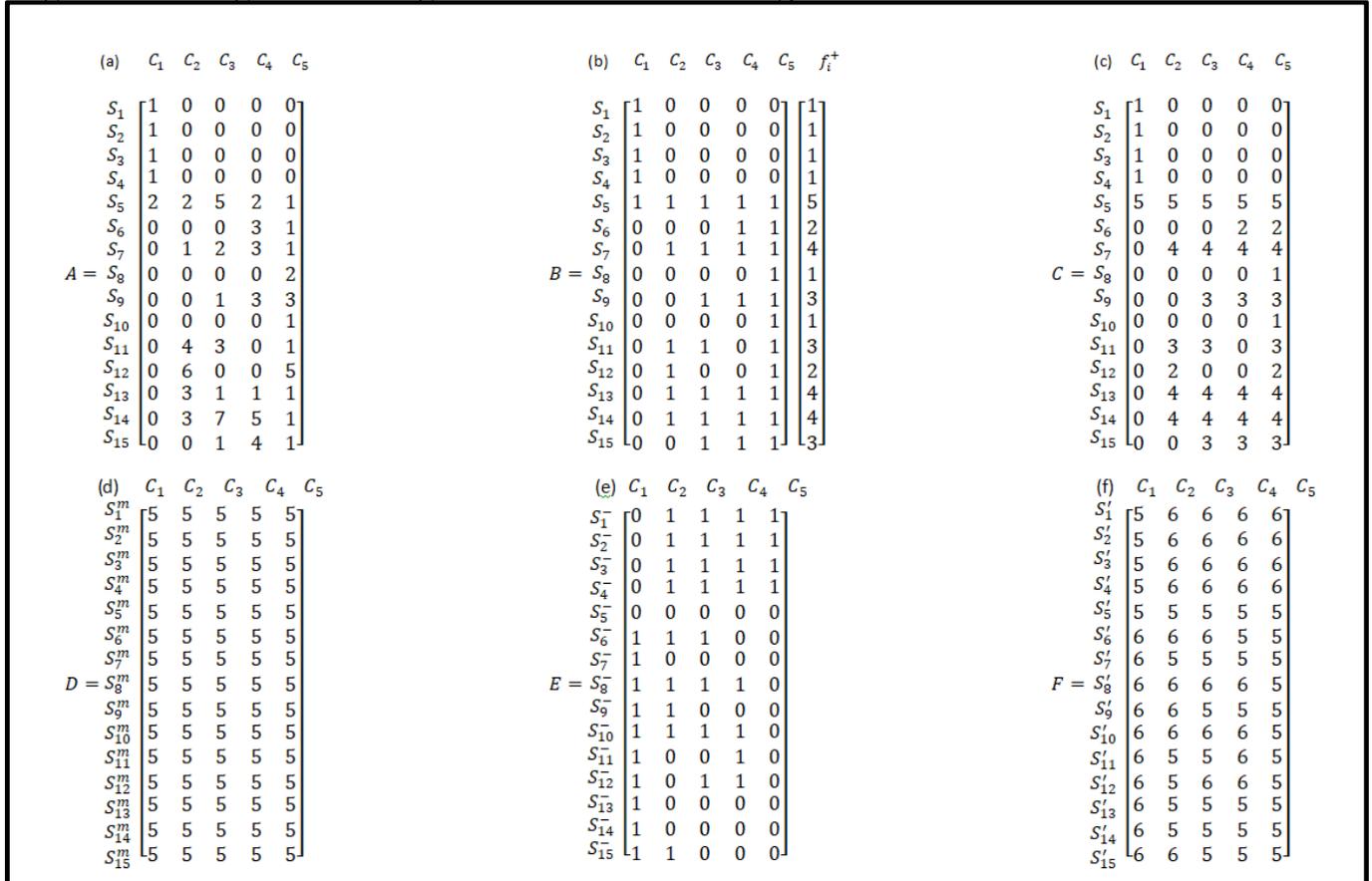
Figure 1. The process of creating a locked matrix



2.2 Material

A hypothetical data example is given in Figure 2. This table indicates 5 communities and 15 species with their abundance values. There are 5 species in community 1 (C_1). Four of them (S_1, S_2, S_3 and S_4) are singletons and specific to C_1 . C_2 has 6 species with only one singleton. C_3 has 3 singletons whereas C_4 has only one singleton. C_3 and C_4 include 7 species. C_5 has the most number of species. Among them, eight species are singletons. Besides S_8 and S_{10} are only found in C_5 .

Figure 2. The stages for creating the locked matrix of the hypothetical data



3. Results and Discussion

Matrix B, prepared from Matrix A (Figure 2a) includes presence absence data of the species (x_{ij}^0) (Figure 2b). The cells of Matrix C include the values of $x_{ij}^0 f_{i+}$ (Figure 2c). Maximum f_{i+} obtained from Matrix B is equal to 5. Hence the value of all cells found in Matrix D is equal to 5 (Figure 2d). Matrix E is inverse of Matrix B ($1 - x_{ij}^0$) (Figure 2e). The cell values of Matrix F are the products of $\max f_{i+} - x_{ij}^0$ (Figure 2f). Matrix G composed of Matrix C and Matrix F is the locked matrix (Figure 2c and Figure 2f). The last matrix including probability values of the locked matrix (Matrix G) that corresponds to Matrix H (Table 1). The size of Matrix H is 5X30 and, it is essential for all computations of hypothetical data.

Table 1. The proportional values of Matrix G (Matrix H)

	C_1	C_2	C_3	C_4	C_5
\dot{p}_1	0,0106	0	0	0	0
\dot{p}_2	0,0106	0	0	0	0
\dot{p}_3	0,0106	0	0	0	0
\dot{p}_4	0,0106	0	0	0	0
\dot{p}_5	0,0532	0,0472	0,0459	0,0463	0,045
\dot{p}_6	0	0	0	0,0185	0,018
\dot{p}_7	0	0,0377	0,0367	0,037	0,036
\dot{p}_8	0	0	0	0	0,009
\dot{p}_9	0	0	0,0275	0,0278	0,027
\dot{p}_{10}	0	0	0	0	0,009
\dot{p}_{11}	0	0,0283	0,0275	0	0,027
\dot{p}_{12}	0	0,0189	0	0	0,018
\dot{p}_{13}	0	0,0377	0,0367	0,037	0,036
\dot{p}_{14}	0	0,0377	0,0367	0,037	0,036
\dot{p}_{15}	0	0	0,0275	0,0278	0,027
\dot{p}'_1	0,0532	0,0566	0,055	0,0556	0,0541
\dot{p}'_2	0,0532	0,0566	0,055	0,0556	0,0541
\dot{p}'_3	0,0532	0,0566	0,055	0,0556	0,0541
\dot{p}'_4	0,0532	0,0566	0,055	0,0556	0,0541
\dot{p}'_5	0,0532	0,0472	0,0459	0,0463	0,045
\dot{p}'_6	0,0638	0,0566	0,055	0,0463	0,045
\dot{p}'_7	0,0638	0,0472	0,0459	0,0463	0,045
\dot{p}'_8	0,0638	0,0566	0,055	0,0556	0,045
\dot{p}'_9	0,0638	0,0566	0,0459	0,0463	0,045
\dot{p}'_{10}	0,0638	0,0566	0,055	0,0556	0,045
\dot{p}'_{11}	0,0638	0,0472	0,0459	0,0556	0,045
\dot{p}'_{12}	0,0638	0,0472	0,055	0,0556	0,045
\dot{p}'_{13}	0,0638	0,0472	0,0459	0,0463	0,045
\dot{p}'_{14}	0,0638	0,0472	0,0459	0,0463	0,045
\dot{p}'_{15}	0,0638	0,0566	0,0459	0,0463	0,045

It is well known that qS increases with decreasing q value for a given community when using Tsallis entropy. As can be seen in Table 2 and Figure 3c, C_3 has the maximum q^* value corresponding to the minimum ${}^q_H S$ value. That is expected result for C_3 since it does not contain a rare species, in particular, a unique species. Even though C_1 contains the least number of species, It harbors the most number of unique species because four of them are only found in C_1 . C_1 is therefore the most specific plot as well. As a result of this, its' q^* value corresponds to 0.61. The q^* value of C_5 is 0.63. That community also deserves such a q^* value because S_8 and S_{10} are unique to C_5 . C_5 has a greater ${}^q_H S$ value than it of C_1 . It is not unexpected result because the number of the species also plays important role over the value of qS and, C_5 is the most richness community with 11 species whereas C_1 contains only 5 species (Table 2, Figure 2a).

Table 2. The results of hypothetical community data

Plots	q^*	q_{HjE}^*	q_{HjS}^*	S
C_1	0.61	0.95598076	5.43348672	5
C_2	0.69	0.99031649	5.01462050	6
C_3	0.71	0.99193508	4.96230959	7
C_4	0.69	0.98942845	5.12926103	7
C_5	0.63	0.97451370	6.15896239	11

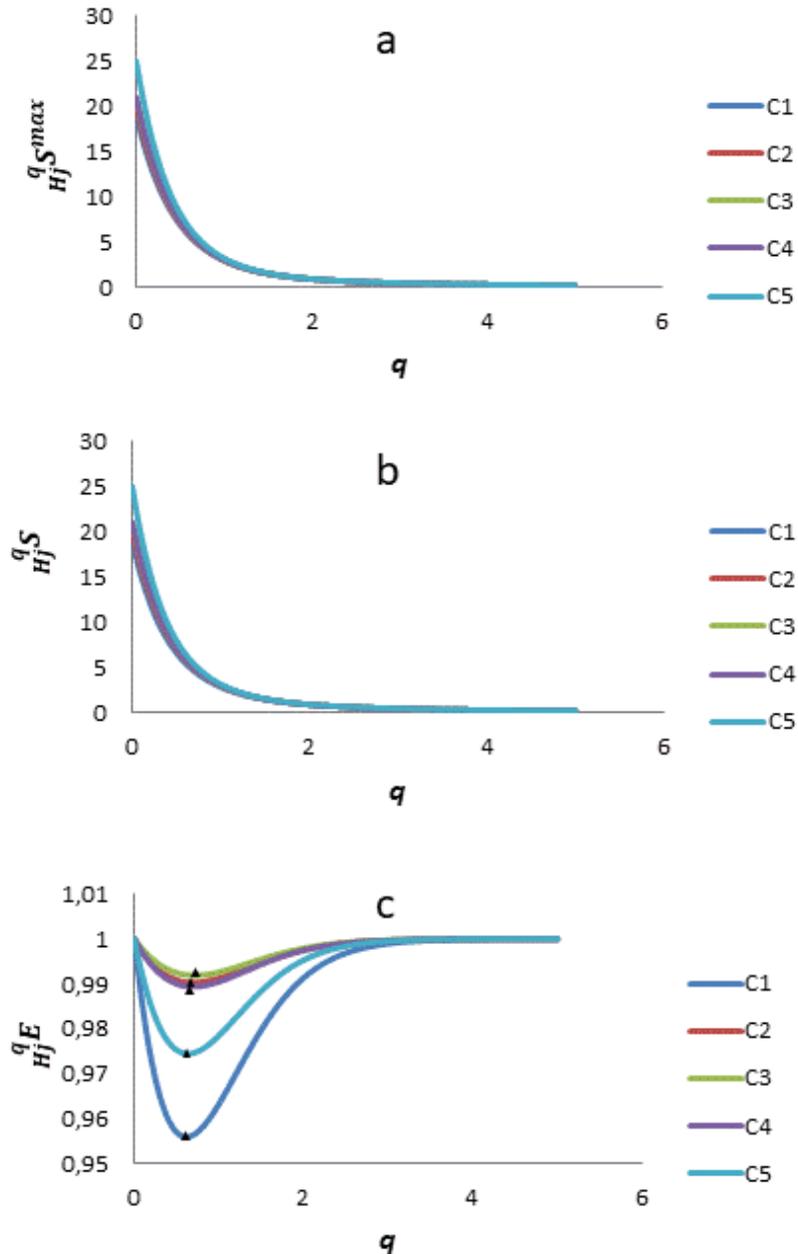


Figure 3. The profile curves of q_{HjS}^{max} , q_{HjS} and q_{HjE} . q_{HjE} exhibits a minimum for $q=q^*$ (triangles)

4. Conclusions

Knowing to diversity and rarity of the ecological communities are vital for preparing accurate and reliable ecosystem based management plans (Özkan, 2016b). With regard to biodiversity, the majority of the estimation issues have been solved owing to the bias-corrected equations (Chao and Shen, 2003), the generalized entropic forms (Macron et al. 2014) and the notion of sample coverage (Chao and Jost, 2012). It is however not possible to tell such a thing in defining community rarity since the metrics employed for measuring rarity ignores undetected species, generalization and standardization. This gap can be solely filled by improving accurate and robust approaches or indices.

The present study does not focus on bias reduction or the notion of sample coverage for measuring community rarity but, suggests a locked matrix approach in order to activate Tsallis entropy. Tsallis entropy has a bias corrected form (Macron et al. 2014). For measuring rarity, this approach might therefore give us an opportunity to improve more robust metrics considering bias reduction at the community level in the future.

To test the proposed approach, a hypothetical community data was used. According to q^* values, the ordering of rarity values is $C_1 > C_5 > C_2 = C_4 > C_3$ from the rarest (the most specific) community to the most common (the most ordinary) one. This ordering is sensible because four species found in C_1 are both singletons and specific to C_1 . C_5 is the second rarest community with two specific species and 7 singletons. Besides the ordering of the ${}^q_H E$ values shows parallelism with the ordering of q^* values. It means ${}^q_H E$ may represent community rarity as well.

As can be seen in Table 2, there is not a significant agreement between community rarity (q^*) and community diversity(S). However, that is not unexpected result since rarity is a different notion from diversity.

As a conclusion, both q^* and ${}^q_H E$ seems to be promise to evaluate community rarity in ecology. However this is a new proposed approach. Further studies should therefore be done to confirm this approach.

References

- Borges, P.A.V., Serrano, A.R., Quartau, J.A. (2000). “Ranking the Azorean Natural Forest Reserves for conservation using their endemic arthropods”, *Journal of Insect Conservation*, vol. 4, pp.129-147.
- Chao, A, Shen, T.J. (2003). “Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample”, *Environmental and Ecological Statistics*, vol. 10, no. 4, pp.429-443.

- Chao, A., Jost, L. (2012). “Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size”, *Ecology*, vol. 93, no.2, pp.2533-2547.
- Dennis, R.L.H., Shreeve, T.G., Olivier, A., Coutsis, G.J. (2000). “Contemporary geography dominates butterfly diversity gradients within the Aegean archipelago (Lepidoptera: Papilionoidea, Hesperioidea)”, *Journal of Biogeography*, vol. 27, pp.1365-1383.
- Fattorini, S. (2008). “How island geography and shape may influence species rarity and biodiversity loss in a relict fauna: a case study of Mediterranean Beetles”, *The Open Conservation Biology Journal*, vol. 2, pp. 11-20.
- Hussain, M. S., Sultana, A., Khan, J.A., Khan, A. (2008). “Species composition and community structure stands in Kumaon Himalaya, Uttarakhand, India”, *Tropical Ecology*, vol. 49, no. 2, pp. 167-181.
- Kaniadakis, G. (2001). “Non-linear kinetics underlying generalized statistics”, *Physica A* vol. 296 no.3-4, pp.405-425.
- Leroy, B., Canard, A., Ysnel, F. (2013). “Integrating multiple scales in rarity assessments of invertebrate taxa”, *Diversity and Distributions*, vol. 19, pp. 794-803.
- Leroy, B., Petillon, J., Gallon, R., Canard, A., Ysnel, F. (2015). “Improving occurrence-based rarity metrics in conservation studies by including multiple rarity cut-off points”, *Insect Conservation and Diversity*, vol. 5, pp. 159-168.
- Marcon, E, Scotti, I, Hérault, B, Rossi, V, Lang, G. (2014). “Generalization of the partitioning of shannon diversity. PLoS ONE, vol. 9, no. 3, pp. 1-8 (e90289).
- Mendes, R.S., Evangelista, L.R., Thomas, S. M., Agostinho, A.A., Gomes, L.C. (2008). “A unified index to measure ecological diversity and species rarity”, *Ecography*, vol. 31, no: 4, pp. 450-456.
- Özkan, K. (2016a). How to measure biodiversity components (α , β , γ) ?, Suleyman Demirel University, Faculty of Forestry, 98, 142 p., Isparta.
- Özkan, K. (2016b). “On the way of only one fundamental information layer for everything within new paradigm sense: ecosystem qualification mapping”, *Journal of the Faculty of Forestry, İstanbul University*, vol. 66, no. 2, pp. 410-444.
- Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S., Wohlgemuth, T. (2002). “Quantitative tools for perfecting species lists”, *Environmetrics*, vol. 13, pp. 121-137.
- Rényi, A., (1961). “On measures of entropy and information”. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, The Regents of the University of California.
- Shannon, C.E., (1948). “A mathematical theory of communication”, *Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423.
- Tsallis, C. (1988). “Possible generalization of Boltzmann-Gibbs statistics”, *Journal of Statistical Physics*, vol. 52, no. 1, pp. 479-487.

O-53 Modeling bivariate survival data in the presence of right censoring by Archimedean Copula approach

Author: Ece Gorcegiz¹, Burcu Hudaverdi Ucer¹

Dokuz Eylul University, Department of Statistics, İzmir¹

Abstract - Modeling dependence structure of a bivariate survival data is one of the main issues in biomedical studies. Survival copula deals with such a lifetime data and is used for modeling and understanding the distributional structure. In this study, we consider modeling and analysing the bivariate survival data in the presence of right censoring using Archimedean copula functions. We use Emura et al.(2010) goodness-of-fit testing procedure for the model selection. First, we examine the heart transplant data and model the dependence structure between waiting time for transplant and post-transplant survival time to see the co-movements of these variables. Second, we examine the diabetic retinopathy data and model the dependence between the survival times of the two eyes of the same patient in case of laser photocoagulation treatment.

Key Words: *Copula, right censoring, bivariate survival data, Archimedean copula, survival copula*

1 Introduction

There has been growing interest in modeling bivariate and multivariate survival data. In the biomedical area, the dependence structure of the bivariate survival data has been studied by many researchers. Copulas are key tools to analyse the dependence structure. A bivariate survival function can be expressed as the composition of marginal survival functions and a bivariate copula. Since a copula is a great deal of flexibility in modeling bivariate survival data, it provides an effective approach for understanding and modeling the dependent random variables and so the dependence structure.

Copula models examine the joint distribution in two ways: the dependence between the variables and the marginal distributions of the individual variables. Hence the dependence structure can be tackled separately from the marginal distributions and the copulas achieve pure dependence structure in random variables. One-parameter family of copulas is preferred for modeling dependence. The Archimedean family of copulas is an important family of copulas. The Archimedean copulas can be reduced to a single univariate distribution function called as Kendall distribution function, $K(\cdot)$. Wang and Wells (2000) propose model selection procedures for bivariate survival models for censored data and a goodness-of-fit-based model selection methodology in study. In this study, we investigate the goodness-of-fit testing procedure of Archimedean class for bivariate survival data.

Let (X, Y) be a random pair of bivariate survival time with bivariate survival function $S(X, Y)$, so the copula can be written as

$$S(X, Y) = C(S(X), S(Y)) \quad (1)$$

where $S(x)$ and $S(y)$ are marginal survival functions and $C(u, v): [0, 1]^2 \rightarrow [0, 1]$ is the copula function. For Archimedean copulas, φ is a generator function which is continuous and strictly decreasing $\varphi_\alpha(\cdot): [0, 1] \rightarrow [0, \infty]$ with a dependence parameter α . Then, the copula $C(u, v) = \varphi_\alpha^{-1}(\varphi_\alpha(u) + \varphi_\alpha(v))$ is called Archimedean copula (Nelsen (2006)). In this study, we deal with three members (Gumbel, Frank and Clayton) of the Archimedean copula family which have various dependence characteristics.

Kendall's tau (τ) is mainly used in copulas to measure the dependence level of the bivariate random variables, and the dependence parameters of copulas can be obtained in terms of Kendall's tau measure.

It can also be expressed as

$$\rho_\tau = 4 \int_0^1 \int_0^1 C_\theta(u, v) c_\theta(u, v) du dv - 1, \quad (2)$$

where $c_\theta(u, v) = \frac{\partial^2 C_\theta(u, v)}{\partial u \partial v}$. The local odds ratio function is proposed by Oakes (1989) and its definition;

$$\theta^*(x, y) = \frac{\partial^2 P(X > x, Y > y) / \partial x \partial y}{\partial P(X > x, Y > y) / \partial x \partial P(X > x, Y > y) / \partial y} \quad (3)$$

$$= \frac{P(\Delta_{ij} = 1 | \tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)}{P(\Delta_{ij} = 0 | \tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)} \quad (4)$$

where $\tilde{X}_{ij} = \dots, X_j$ and $\tilde{Y}_{ij} = \dots, Y_j$, and $a \wedge b = \min(a, b)$. In Eq. (3), Emura and Wang, Hung (2011) indicated that a cross-ratio function can be introduced the characterization results and inferential procedures for censored data and uncensored data by real data. But in Oakes Emura and Wang, Hung (2011) study, the procedures developed work well with survival censored data. If the pair (X, Y) is independent at (x, y) then $\theta^*(x, y)$ is a bivariate function measuring local dependence and equal 1. And also it is indicated that $\theta^*(x, y) = \theta\{F(x, y)\}$ for the Archimedean copula class. $\theta\{F(x, y)\}$ is univariate function and it can be obtained as

$$\theta_\alpha(v) = -v \left\{ \frac{[\phi'_\alpha(v)]}{[\phi'_\alpha(v)]} \right\}, \quad (4)$$

Genest and Rivest (1993) indicated that $K(v)$ is related to $\phi_\alpha(v)$ through the differential equation

$$\lambda(v) = v - K(v) = \frac{\phi_\alpha(v)}{\phi'_\alpha(v)} \quad (5)$$

where $\phi'_\alpha(v) = \partial\phi_\alpha(v)/\partial v$ and $v = S(X, Y)$ (Wang and Wels (2000)). And also it is showed that the function $\phi_\alpha(v)$ can be obtained by the univariate function $K(v) \equiv P\{F(X, Y) \leq v\}$ where $K(v)$ and $\lambda(v)$ are related with the generator function $\phi_\alpha(v)$ and determine the dependence structure of the Archimedean copula class.

After the parameter estimation process of some Archimedean copulas, the goodness-of-fit testing procedure is applied to check which copula fits best to the given data. The goodness-of-fit test provides simple graphical tools and numerical techniques for selecting an appropriate model, estimating its parameters, and checking its goodness-of-fit. The goodness-of-fit tests implement that the unknown copula (C) , belongs to a parametric family, (C_0) of copulas, that is, $H_0 = C \in C_0$ versus $H_1 = C \notin C_0$. Graphical methods, error statistics, and formal goodness of fit statistics are used to measure the adequacy of this hypothesis. In this study, we follow the goodness-of-fit procedure for Archimedean copula proposed by Emura et al. (2010). In Section 2, the testing procedure is given for both the censored and the uncensored data. In Section 3, the procedure is applied to the heart transplant data and diabetic data sets. Section 4 is devoted to the conclusion.

2 The Goodness-of-Fit Test Procedure

In this section, we deal with uncensored data firstly, then we modify the model for the right censored data. The main hypothesis is given by

$$H_0 : C(u, v) = \phi_\alpha^{-1} [\phi_\alpha(u) + \phi_\alpha(v)] \text{ for some } \alpha \in \mathfrak{R},$$

where the alternative hypothesis is any other copula.

2.1. The GOF procedure for the uncensored data

Let $\{(X_i, Y_i); (i=1, \dots, n)\}$ be the uncensored data, and Δ_{ij} be the concordance indicator, then

$$U_k(\alpha) = \sum_{i < j} W_k(\tilde{X}_{ij}, \tilde{Y}_{ij}) \left[E_{[-ij]} \tilde{X}_{ij} \quad x, \tilde{Y}_{ij} \quad y \right] \quad (6)$$

$$U_k(\alpha) = \sum_{i < j} W_k(\tilde{X}_{ij}, \tilde{Y}_{ij}) \left[\frac{\theta_\alpha \{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}}{\theta_\alpha \{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}} \right] \quad (7)$$

where $W_k(\tilde{x}, \tilde{y})$ is weight function and $\hat{S}(x, y)$ is an estimator of $S(x, y)$. $\hat{S}(x, y)$ is called empirical survival function and it can be expressed as $\hat{S}(x, y) = \frac{1}{n} \sum_i I(X_i \geq x, Y_i \geq y)$. In this GoF

procedure, two weight functions are used to obtain $\hat{\alpha}_k$ solve $U_k(\alpha) = 0$ for $k=1,2$. When H_0 is true, $n^{1/2}\{\hat{\alpha}_1 - \hat{\alpha}_2\}$ converges to a mean zero normal distribution. The power of the test depends to choose two weight functions, see for more details Emura et al. (2010).

2.2. The weight functions

Shih (1998) compared the unweighted and weighted concordance estimators of the association parameter and indicated that if proposed model is fit, the difference of these two estimates converges to zero in his study. Using the idea of the likelihood approach of Clayton (1978), the estimating function can be written in terms of $U_k(\alpha)$. Define the set of grid points as follow,

$$\psi = \left\{ (x, y) \left| \sum_{i=1}^n I(X_i = x, Y_i \geq y) = 1, \sum_{i=1}^n I(X_i \geq x, Y_i = y) = 1 \right. \right\} \quad (8)$$

Also let $D(x, y) = \sum_i I(X_i = x, Y_i = y)$ be the number of observed failures at (x, y) , and $R(x, y) = r = \sum_{i=1}^n I(X_i \geq x, Y_i \geq y)$ measures the number of risk at $(x, y) \in \psi$. Then, $D(x, y)$ is distributed Bernoulli with the success probability

$$P\{D(x, y) = 1 | R(x, y) = r, (x, y) \in \psi\} = \frac{\theta_\alpha \{S(x, y)\}}{r - 1 + \theta_\alpha \{S(x, y)\}} \quad (9)$$

By using conditional probability, the likelihood function can be obtained as

$$L(\alpha) = \prod_{(x,y) \in \psi} \left[\frac{\theta_\alpha \{S(x, y)\}}{R(x, y) - 1 + \theta_\alpha \{S(x, y)\}} \right]^{D(x,y)} \times \left[\frac{R(x, y) - 1}{R(x, y) - 1 + \theta_\alpha \{S(x, y)\}} \right]^{1 - D(x,y)} \quad (10)$$

with all points in ψ under the independence assumption among the grids.

Then, the estimating function is

$$U_1(\alpha) = \sum_{i < j} \frac{\dot{\ell}_{\alpha, i} S(\tilde{X}_{ij}, \tilde{Y}_{ij}) \tilde{Y}_{ij} - \alpha \dot{\ell}_{\alpha, i} S(\tilde{X}_{ij}, \tilde{Y}_{ij}) \tilde{X}_{ij}}{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \} R_{ij}} - \sum_{ij} \frac{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}}{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \} R_{ij}} \quad (11)$$

$$= \sum_{i < j} W_1(\tilde{X}_{ij}, \tilde{Y}_{ij}) \left[-ij \frac{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}}{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \} R_{ij}} \right] \quad (12)$$

where $R_{ij} = R(\tilde{x}_{-ij}, \tilde{y}_{-ij}, \dots, \tilde{x}_{-ij}, \tilde{y}_{-ij})$ and $\dot{\ell}_{\alpha}(\tilde{x}_{-ij}, \tilde{y}_{-ij}) = \frac{\partial}{\partial \alpha} \ell_{\alpha}(\tilde{x}_{-ij}, \tilde{y}_{-ij})$. We show the second estimating function

$$U_2(\alpha) = \sum_{i < j} \left[\Delta_{ij} - \frac{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}}{\theta_\alpha \{ \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}} \right] \quad (13)$$

By solving $U_1(\alpha) = 0$ and $U_2(\alpha) = 0$, we find $\hat{\alpha}_k$ for $k = 1, 2$ Frank copula is defined as $\phi_\alpha(v) = \log\{(1 - \alpha^{-1}) / (1 - \alpha^{-v})\}$ and with $\theta_\alpha \{S(x, y)\} = S(x, y) \log(\alpha) / (1 - e^{-S(x, y) \log(\alpha)})$

$$W_1(\tilde{X}_{ij}, \tilde{Y}_{ij}) = \frac{(\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}} + 1)}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}})) + 1} \quad (14)$$

Therefore,

$$U_1(\alpha) = \sum_{i < j} \frac{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}} + 1)}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}})) + 1} \times \left[\Delta_{ij} - \frac{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))})}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}) + 1} \right] \quad (15)$$

$$U_2(\alpha) = \sum_{i < j} \left[\Delta_{ij} - \frac{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))})}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}) + 1} \right] \quad (16)$$

2.3. U-statistics of Archimedean copulas

In this part, we obtained the U-statistics of some Archimedean copulas. The weighted and unweighted function approach to the true parameter value of Frank copula is

$$U_1(\alpha) = \sum_{i < j} \frac{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}} + 1}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{S_{X_{ij} Y_{ij}}} S_{X_{ij} Y_{ij}} \dots S_{X_{ij} Y_{ij}} e^{S_{X_{ij} Y_{ij}}})) + 1} \times \left[\Delta_{ij} - \frac{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))})}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}) + 1} \right] \quad (17)$$

$$U_2(\alpha) = \sum_{i < j} \left[\Delta_{ij} - \frac{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))})}{((\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij}) e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}) + 1} \right] \quad (18)$$

The weighted and unweighted function approach to the true parameter value of Gumbel copula is

$$U_1(\alpha) = \sum_{i < j} \frac{2 \log \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})}{\log \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})} \Delta_{ij} - \frac{e^{-S(\tilde{X}_{ij}, \tilde{Y}_{ij})}}{g \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})} \quad (19)$$

$$U_2(\alpha) = \sum_{i < j} \left[\Delta_{ij} - \frac{\log \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})}{2 \log \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})} \right] \quad (20)$$

The weighted and unweighted function approach to the true parameter value of Clayton copula is

$$U_1(\alpha) = \sum_{i < j} \frac{[(\alpha+1)+1]}{(\alpha+1) \hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})} \left[\Delta_{ij} - \frac{(\alpha+1)}{[(\alpha+1)+1]} \right] \quad (21)$$

$$U_2(\alpha) = \sum_{i < j} \left[\Delta_{ij} - \frac{(\alpha+1)}{[(\alpha+1)+1]} \right] \quad (22)$$

For, Frank copula, the asymptotic distribution of $n^{1/2}(\hat{\alpha}_1 - \alpha_2)$ and $n^{1/2}(\log \hat{\alpha}_1 - \log \hat{\alpha}_2)$ approximates to a normal distribution with mean zero and variance $\sigma^2 = 4E[h\{(X_1, Y_1), (X_2, Y_2)\}h\{(X_1, Y_1), (X_3, Y_3)\}]$. And also $(\log \hat{\alpha}_1 - \log \hat{\alpha}_2)$ converges zero in probability, so that $U_1(\alpha)$ and $U_2(\alpha)$ involve the estimator. However, this causes difficulty in technical. When we hold U -statistic;

$$\tilde{U}_{1, \alpha} = \sum_{i < j} \frac{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha} S(\tilde{X}_{ij}, \tilde{Y}_{ij})}{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha} S(\tilde{X}_{ij}, \tilde{Y}_{ij})} \Delta_{ij} - \frac{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha}}{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha}} \quad (23)$$

and

$$\tilde{U}_{2, \alpha} = \sum_{i < j} \left[\Delta_{ij} - \frac{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha}}{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^{-\alpha}} \right] \quad (24)$$

Lemma: Under the correct model and regularity conditions

$$n \binom{n}{2}^{-1} U_1(\alpha) = \binom{n}{2}^{-1} \tilde{U}_{1, \alpha} + o_p(1), \quad \binom{n}{2}^{-1} U_2(\alpha) = \binom{n}{2}^{-1} \tilde{U}_{2, \alpha} + o_p(1),$$

where $O_p(1)$ is uniform in α .

If the parameter α is a positive value, natural logarithm transformation can recuperate to normal approximation.

Lemma: Under the correct model and regularity conditions. When $\gamma = \log \alpha$ and $\hat{\gamma}_k = \log \hat{\alpha}_k$

$$n^{1/2}(\hat{\gamma}_1 - \hat{\gamma}_2) - n \cdot \binom{n}{2}^{-1} \sum_{i < j} h\{(X_i, Y_i), (X_j, Y_j)\} + O_p(1), \quad (25)$$

where the function h is symmetric and

$$h\{(X_i, Y_i), (X_j, Y_j)\} \equiv \frac{1}{\alpha} \left(\frac{\dot{\ell}_\alpha(S, \tilde{X}_{ij}, \tilde{Y}_{ij})}{A_L \theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij}), S, \tilde{X}_{ij}, \tilde{Y}_{ij}\}} - \frac{1}{A} \right) \left[\frac{\theta_\alpha \{S(\tilde{X}_{ij}, \tilde{Y}_{ij})\}}{\theta_\alpha \{S, \tilde{X}_{ij}, \tilde{Y}_{ij}\}} \right], \quad (26)$$

where $A \equiv E \left(\frac{\dot{\ell}_\alpha(S, \tilde{X}_{12}, \tilde{Y}_{12})}{\theta_\alpha \{S(\tilde{X}_{12}, \tilde{Y}_{12})\}} \right)$ and $A_L \equiv E \left(\frac{\dot{\ell}_\alpha(S, \tilde{X}_{12}, \tilde{Y}_{12})}{\theta_\alpha \{S(\tilde{X}_{12}, \tilde{Y}_{12}), S, \tilde{X}_{12}, \tilde{Y}_{12}\}} \right)$.

2.4. The testing procedure

When $|\hat{\gamma}_1 - \hat{\gamma}_2|/\hat{\sigma}$ is greater than $Z_{1-\alpha/2}$ for α - level test and $Z_{1-\alpha/2}$ is the p th percentile of the standard normal distribution. In Archimedean copula models, derivation of the variance estimator is difficult and also this formula becomes complex for right-censored data. Therefore, Emura et al. (2010) proposed the jackknife method for the variance estimation. The jackknife method is defined as

$$\sigma_{jack}^2 = \frac{n-1}{n} \sum_{i=1}^n \{ \hat{\gamma}_1^{(i)} - \hat{\gamma}_2^{(i)} - (\hat{\gamma}_1 - \hat{\gamma}_2) \}^2, \quad (27)$$

where $\hat{\gamma}_k^{(i)}$ is the estimator after deleting the i th observation and $\hat{\gamma}_1 - \hat{\gamma}_2 = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_1^{(i)} - \hat{\gamma}_2^{(i)}$.

2.5. In the case of right-censoring

Let (A_i, B_i) be the independently identically distributed bivariate censoring variables. (A_i, B_i) and (X_i, Y_i) are independent of each other. Through the right censoring process, it is applied that, $\tilde{Y}_i = \min(Y_i, B_i)$, $\delta_i^x = I(X_i \leq A_i)$ and $\delta_i^y = I(Y_i \leq B_i)$. The order of X_i and X_j is known if and only if $\tilde{x}_{-y} < \tilde{x}_{-y}$, where $\tilde{x}_{-y} = \tilde{x}_{-i} - \tilde{x}_{-j}$ and $\tilde{x}_{-y} = \tilde{x}_{-j} - \tilde{x}_{-i}$. Similarly, the order of Y_i

and Y_j is known if and only if $\tilde{x}_y = \tilde{y}$. Let Z_{ij} indicates whether the ordering relationship is certain or not. Then, then U- statistic is

$$U_k(\alpha) = \sum_{i < j} Z_{ij} W_k(\tilde{x}_{-y}, \tilde{y}, \dots) \left[\Delta_{ij} - \frac{\theta_\alpha \{ \hat{S}(\tilde{x}_{-y}, \tilde{y}) \}}{\theta_\alpha \{ \hat{S}(\tilde{x}_{-y}, \tilde{y}) \}} \right] \quad k=(1,2), \quad (28)$$

where $\hat{S}(x, y)$ is an estimator of $S(x, y)$ and $\Delta_{ij} = I\{(\tilde{X}_i - \tilde{X}_j)(\tilde{Y}_i - \tilde{Y}_j) > 0\}$. The weight function formula is

$$W_1(\tilde{x}_{-y}, \tilde{y}, \dots) = \frac{\theta_\alpha \{ \hat{S}(\tilde{x}_{-y}, \tilde{y}) \}}{\theta_\alpha \{ \hat{S}(\tilde{x}_{-y}, \tilde{y}) \}} \quad (29)$$

where $W_2(\tilde{x}_{-y}, \tilde{y}, \dots) = 1$ and $R_{ij} = \sum_{l=1}^n I(\tilde{X}_l \geq \tilde{x}_{-y} - \tilde{y})$.

To solve \hat{u}_k , the following equation is used;

$$\sum_{i < j} Z_{ij} W_k(\tilde{x}_{-y}, \tilde{y}, \dots) \left[\Delta_{ij} - \frac{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}}{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}} \right] = 0 \quad l=1,2,\dots \quad (30)$$

where \tilde{S} show the estimated value in the l-step and also defined as $\tilde{S}_{\alpha, \tilde{x}_{-y}, \tilde{y}}$.

As in the censored case, the results of asymptotic normality are valid and

$$\tilde{U}_{\alpha, \tilde{x}_{-y}, \tilde{y}} = \sum_{i < j} Z_{ij} \left[\Delta_{ij} - \frac{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}}{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}} \right], \quad (31)$$

U -statistic approximate to $U_k(\alpha)$ as in the above $U_2(\alpha)$ function.

$$\tilde{U}_{\alpha, \tilde{x}_{-y}, \tilde{y}} = \sum_{i < j} Z_{ij} \left[\Delta_{ij} - \frac{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}}{\theta_\alpha \{ \tilde{S}(\tilde{x}_{-y}, \tilde{y}) \}} \right] \quad (32)$$

3 Real Data Applications

Firstly, our data is obtained from heart transplant recipients from the Stanford heart transplant program (Crowley and Hu (1977)). We tackled 69 transplant data from 103 transplants and also our data is used for the uncensored case. We model the dependence structure between waiting time for transplant and post-transplant survival time to see the co-movements of these variables and how they affect each other. Table 2A shows the goodness-of-fit results of the copula models. For the Frank copula model, we obtain $\hat{\alpha}_1 = 2.1930472$ and $\hat{\alpha}_2 = 2.134277$. Also, we calculate τ -values based on $\hat{\alpha}$ values as

$\hat{\tau}_1 = 0.2328296$ and $\hat{\tau}_2 = 0.2271088$. As seen in Table 2A, the Frank copula has the best fit (p-value=0.3594) for heart transplant data. The Gumbel copula model cannot be rejected at the 5% significance level also, but its GoF statistic is higher than the Frank copula model. The model has symmetric dependence structure and transplant and post-transplant survival times do not have co-movements at the tails. Figures 1A-1C are constructed using λ -functions of Clayton, Gumbel and Frank copula model, respectively. In Figures, the left graph is for the empirical λ -function of the data, the middle one is for the theoretical λ -function and the right one is for both empirical and theoretical λ -functions. When Figures 1A-1C are examined, it can be concluded that the Frank copula model has again best fit to the data.

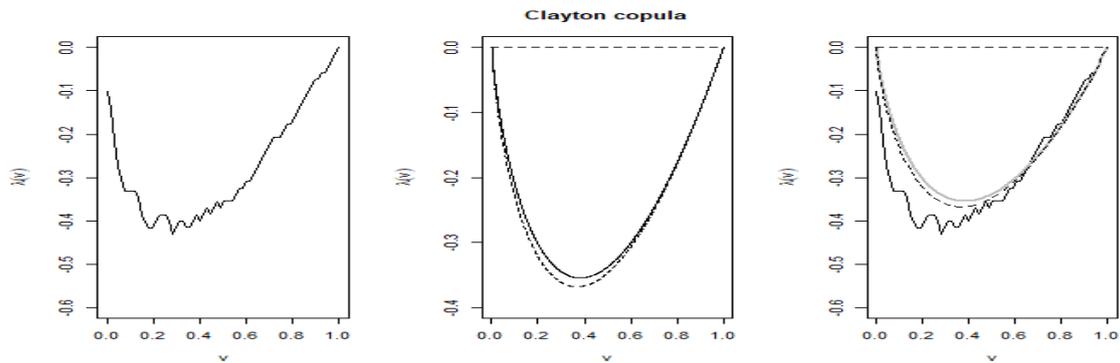


Fig. 1A: λ -functions based on the Stanford heart transplant program for Clayton copula model

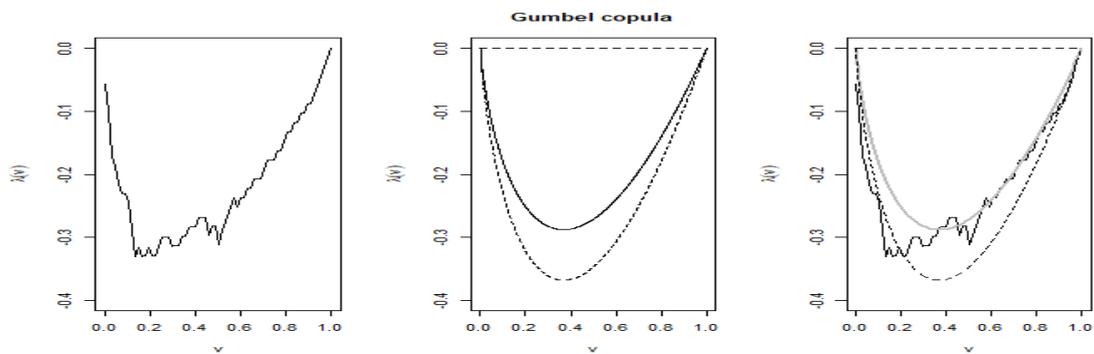


Fig. 1B: λ -functions based on the Stanford heart transplant program for Gumbel copula model

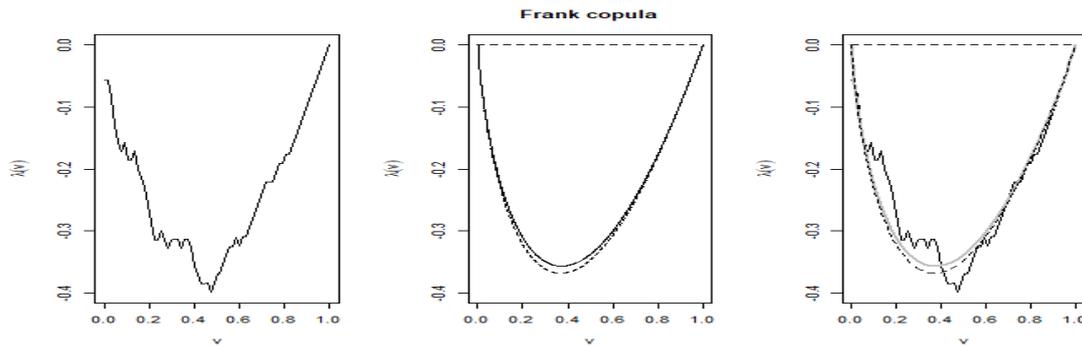


Fig. 1C: λ -functions based on the Stanford heart transplant program for Frank copula model

Secondly, our other data is obtained from diabetic retinopathy study (Manatunga and Oakes (1999)). We tackled 197 patients for our censored study. One eye of patients was randomly selected for photocoagulation treatment and an experimental treatment was applied to one eye, a standard treatment to the other eye, of each patient enrolled. Manatunga and Oakes (1999) concentrated the effect of laser- photocoagulation therapy, and whether this effect differs for adult onset and juvenile onset patients and association between the survival times of the two eyes of the same patient for their analyses in their study. We examine that of the 197 patients, 38 experienced failure in both eyes, 79 experienced failure in one eye and 80 experienced no failure. Considering censored status, we use the variables as treated and untreated eyes and also we deal with time to loss of vision or last follow-up of the patients as survival time.

In this data, we model the dependence structure between the survival times of the two eyes of the same patient after laser- photocoagulation therapy to see how they affect each other. In these two studies, we deal with Frank copula, Gumbel copula and Clayton copula models. The GoF results are given in Table 2B. When we examine Table 2B, the Frank copula model has best fit to our data (p-value=0.435 (Frank)). Also, for the Frank copula model, we obtain $\hat{\alpha}_1 = 1.5647446$ and $\hat{\alpha}_2 = 1.6743183$. The visual comparisons are also Figures 2A-2C via λ -functions. In Figures, the left graph is for the empirical λ -function of the data, the middle one is for the theoretical λ -function and the right one is for both empirical and theoretical λ -functions. When Figures 1A-1C are examined, it can be concluded that the Frank copula model has again best fit to the data.

Frank copula is important because it allows negative dependence between the marginals and also the dependence is symmetric in tails. Frank copula is “comprehensive” in the sense that Fréchet lower bound and Fréchet upper bound are included in the range of dependence. In theory, Frank copula can be applied to the model outcomes with strong positive or negative dependence. (Trivedi and Zimmer (2005))

Table 1:
Examples of Families of Archimedean Copulas

	ϕ	λ	τ
Clayton	$(S(\tilde{X}_{ij}, \tilde{Y}_{ij}))^\alpha$	$-S(\tilde{X}_{ij}, \tilde{Y}_{ij}) - S(\tilde{X}_{ij}, \tilde{Y}_{ij})^\alpha$	$\alpha/(\alpha+2)$
Frank	$\log\left(\frac{1 - \exp(-\alpha)}{1 - \exp(-\alpha S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}\right)$	$-\frac{1 - \exp(-\alpha S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}{\alpha \exp(-\alpha S(\tilde{X}_{ij}, \tilde{Y}_{ij}))} - \frac{1 - \exp(-\alpha)}{\alpha \exp(-\alpha S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}$	$1 + 4\{D_1(\alpha) - 1\}/\alpha$
Gumbel	$\{-\log(S(\tilde{X}_{ij}, \tilde{Y}_{ij}))\}^\alpha$	$S(\tilde{X}_{ij}, \tilde{Y}_{ij}) - \alpha S(\tilde{X}_{ij}, \tilde{Y}_{ij})^\alpha$	$\alpha/(\alpha+1)$

NOTE: The three families are indexed by a single real parameter α

* D_1 is the Debye function of order 1, $D_1(\alpha) = \int_0^\alpha \{t/\alpha(e^t - 1)\} dt$ (Genest and Rivest (1993)).

Table 2A:
The goodness-of-fit test results for three AC models based on the Stanford heart transplant program (Crowley and Hu (1977))

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$(\hat{\gamma}_1 - \hat{\gamma}_2)/\hat{\sigma}_{jack}$	<i>p value</i>
Clayton	0.1692321	0.6449004	-6.437196	0
Frank	2.1930472	2.134277	0.3676287	0.3594
Gumbel	1.2789026	1.2598375	0.9582767	0.1711

Table 2B:
The goodness-of-fit test results for three AC models based on the diabetic retinopathy study (Manatunga, A. K. and Oakes, D. (1999)).

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$(\hat{\gamma}_1 - \hat{\gamma}_2)/\hat{\sigma}_{jack}$	<i>p value</i>
Clayton	1.164139	1.246918	-1.00857	0.1587
Frank	3.897294	3.944172	-0.1755508	0.435
Gumbel	1.5647446	1.6743183	-2.603582	0.0047

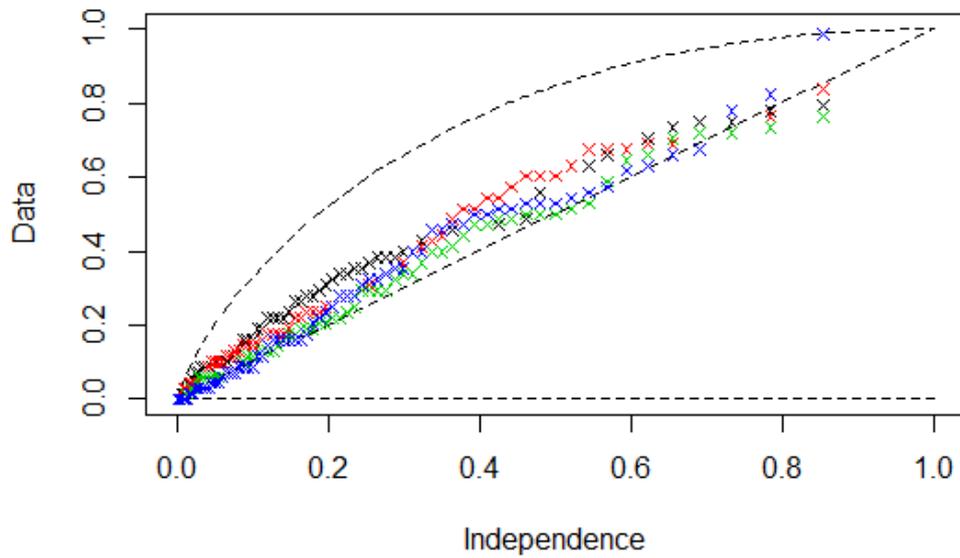


Fig. 2A: Kendall plot based on the Stanford heart transplant program
NOTE: Green (Frank copula), red (Clayton copula), blue (Gumbel copula)

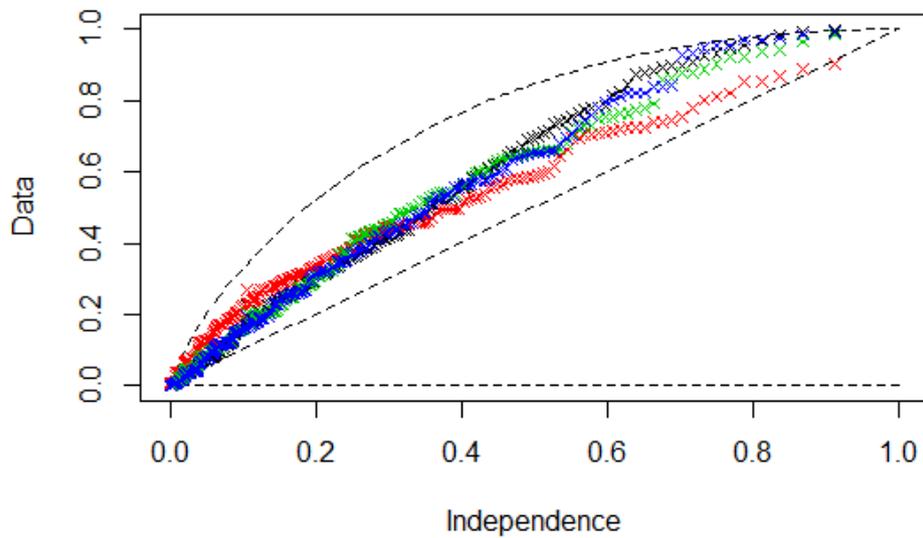


Fig. 2B: Kendall plot based on the diabetic retinopathy study
NOTE: Green (Frank copula), red (Clayton copula), blue (Gumbel copula)

Kendall distribution of the selected copula model for both of the data set is visualized in Figures 2A- 2B. It can also be concluded that Frank copula model has best fit to both Stanford heart transplant data and diabetic retinopathy data.

4 Conclusion

In this study, we deal with modeling and analyzing bivariate survival uncensored and right-censored data for three Archimedean copula models. We apply goodness-of-fit method proposed by Emura et al. (2010) to select the best appropriate Archimedean copula model to our data. Different from the conventional methods in survival analysis, this method yields a formal goodness-of-fit test using weight function based on conditional likelihood. In the presence of right-censoring for the data, deleting non-orderable pairs from the equation is used. This strategy was also used by Oakes (1982). We use Kendall distribution function and λ -functions to choose the best fit visually. We adapt to right-censored data to the U-statistics. Since the variance estimator is complicated for the right-censored data, the jack-knife estimator is obtained. For both of the heart transplant data and diabetic retinopathy data, Frank copula model is selected. The data sets have symmetric dependence structure.

References

1. Clayton, D.G., 1978. A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65, 141-151.
2. Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Am. Statist. Assoc.* 72, 27-36.
3. Emura, T., Lin, C.-W., and Wang, W. (2010). A goodness-of-fit test for Archimedean copula models in the presence of right censoring. *Computational Statistics & Data Analysis* 54(12):3033–3043.
4. Emura T, Wang W, Hung HN (2011) Semi-parametric inference for copula models for truncated data. *Stat Sin* 21:349–367
5. Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* 88, 1034-1043.

6. Manatunga, A.K., Oakes, D. (1999) Parametric Analysis of Matched Pair Survival Data. *Lifetime Data Analysis* 5, 371 – 387.
7. Nelsen RB. (2006) *An introduction to copulas*. Springer, New York.
8. Oakes, D. (1982), “A Model for Association in Bivariate Survival Data,” *Journal of the Royal Statistical Society. Ser. B*, 44, 414-422.
9. Shih, J.H., 1998. A goodness-of-fit test for association in a bivariate survival model. *Biometrika* 85, 189 200.
10. Wang, W., Wells, M., 2000. Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* 95, 6272.
11. J Trivedi PK, Zimmer DM. Copula modeling: an introduction for practitioners. *Found Trends Economet* 2005; 1(1):1–111.

O-57 Parameter Estimation for the k –th Extreme Value Distribution

Talha Arslan¹, Şükrü Acıtaş^{2*} and Birdal Şenoğlu³

¹*Department of Econometrics, Van Yüzüncü Yıl University, Turkey, mstalhaarlan@yyu.edu.tr*

²*Department of Statistics, Eskişehir Technical University, Turkey, sacitas@eskisehir.edu.tr*

³*Department of Statistics, Ankara University, Turkey, senoglu@science.ankara.edu.tr*

Abstract – The k –th Extreme Value (EV k) distribution is defined as asymptotical distribution of k –th extremes (Gumbel, 1935) and used for modelling these extreme observations. The location and scale parameters of EV k distribution are mostly estimated using the maximum likelihood (ML) method. Numerical methods should be performed to obtain ML estimates of parameters of EV k distribution since likelihood equations include intractable terms. However, using numerical methods may be problematic and convergence is not guaranteed. In this study, we therefore utilize Tiku’s (1967,1968) modified maximum likelihood (MML) method and thus resulting MML estimators are explicitly formulated. See also Tiku and Akkaya (2004) and Aydin (2017) in the context of MML estimation when $k = 1$. To investigate the effect of different values of k on the performances of the ML and MML

estimators, a Monte-Carlo simulation study is carried out. Results show that the MML estimators are as efficient as ML estimators. Therefore, the MML estimators can also be preferred to the ML estimators if our focus is computational ease besides efficiency.

Keywords – *k –th Extreme Value distribution, Maximum Likelihood, Modified Maximum Likelihood, Efficiency*

1. Introduction

The k –th Extreme Value (EV k) distribution has the following probability distribution function (pdf) and cumulative distribution function (cdf):

$$f(y; \mu, \sigma, k) = \frac{k^k}{\sigma \Gamma(k)} \exp \left[-k \left(\frac{y - \mu}{\sigma} \right) - k \exp \left(- \left(\frac{y - \mu}{\sigma} \right) \right) \right]; \quad y \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma, k > 0 \quad (1)$$

and

$$F(y; \mu, \sigma, k) = \Gamma \left[k \exp \left(- \left(\frac{y - \mu}{\sigma} \right) \right), k \right], \quad (2)$$

respectively. Here, μ is the location parameter, σ is the scale parameter, k is the shape parameter and $\Gamma(\cdot, \cdot)$ denotes the upper incomplete gamma function. The EV k distribution is obtained as asymptotic distribution of the k –th extremes (Gumbel, 1935). It should be noted that the EV k distribution can be considered as the special case of the generalized Gumbel distribution proposed by Demirhan (2018); see also LeDuc and Stevens (1977), Adeyemi and Ojo (2003). In Table 1, skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) values of the EV k distribution are tabulated for better understanding on the shape of the distribution for certain values of the shape parameter k . See also Figure 1 in which the pdf plots of the EV k distribution are illustrated.

Table 1. The skewness and kurtosis values of EV k distribution for certain values of k ($\mu = 0, \sigma = 1$).

k	0.5	3	6	9
$\sqrt{\beta_1}$	1.5351	0.6209	0.4247	0.3424
β_2	7.0000	3.7626	3.3597	3.2342

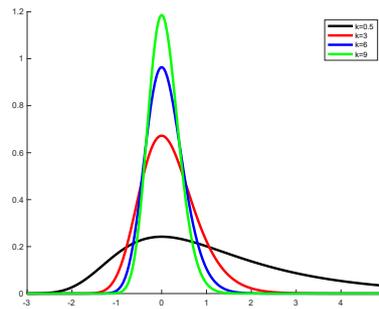


Figure 1. The pdf plots of EVk distribution for certain values of k ($\mu = 0, \sigma = 1$).

The maximum likelihood (ML) estimates of the parameters μ and σ cannot be obtained explicitly since likelihood equations of the corresponding parameters involve nonlinear functions of them. Therefore, iterative techniques, such as Newton Raphson, should be utilized to solve the likelihood equations simultaneously. However, it is known that using numerical methods may have the following problems: (i) non-convergence of iterations, (ii) convergence to multiple roots and (iii) convergence to wrong root, see e.g. Barnett (1966), Puthenpura and Sinha (1986) and Vaughan (1992).

The aim of this study is to obtain closed form estimators for both the location parameter μ and scale parameter σ by using the modified ML (MML) methodology originated by Tiku (1967,1968). It should be noted that Tiku and Akkaya (2004) and Aydin (2017) consider the MML estimation when $k = 1$. We here investigate the effect of the various values of k on the efficiencies of the estimators.

The rest of the paper is organized as follows. A brief descriptions of the ML and MML methodologies are given in Section 2. Section 3 is reserved to a Monte Carlo (MC) simulation study to show the performance of the proposed estimation methodology. The paper is ended with some concluding remarks.

2. Parameter Estimation

In this section, the ML and MML estimation for the location and scale parameters of the EVk distribution are provided, respectively. It should be noted that the shape parameter k is assumed to be known throughout the study.

Let Y_1, Y_2, \dots, Y_n be a random sample from the EVk distribution, then $\ln L$ function can be written as follows:

$$\ln L = n [k \ln k - \ln \sigma - \ln \Gamma(k)] - k \sum_{i=1}^n \left(\frac{y_i - \mu}{\sigma} \right) - k \sum_{i=1}^n \exp \left[- \left(\frac{y_i - \mu}{\sigma} \right) \right]. \quad (3)$$

The ML and MML estimators of the location and scale parameters of the EV k distribution are obtained in following subsections, respectively.

2.1 ML estimation

After taking derivatives of the $\ln L$ with respect to the parameters μ and σ and setting them equal to 0, we obtain likelihood equations. In other words, ML estimates of the parameters μ and σ are obtained by solving the following likelihood equations:

$$\frac{\partial \ln L}{\partial \mu} = \frac{k}{\sigma} \sum_{i=1}^n g(z_i) = 0 \quad (4)$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{k}{\sigma} \sum_{i=1}^n z_i g(z_i) = 0 \quad (5)$$

where $g(z_i) = 1 - \exp(-z_i)$ and $z_i = (y_i - \mu)/\sigma$, $i = 1, 2, \dots, n$. It should be noticed that the ML estimators of the location parameter μ and scale parameter σ cannot be obtained explicitly. Therefore, numerical methods such as Newton-Raphson should be performed.

2.1 MML estimation

In this study, the MML methodology originated by Tiku (1967,1968) is used to obtain the estimators of the location parameter μ and scale parameter σ , explicitly. There are two steps to obtain the MML estimators of μ and σ . They are given step by step as follows:

Step 1. Standardized observations are ordered in ascending form, i.e. $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$.

Step 2. $g(z_i) = 1 - \exp(-z_i)$ function is linearized around the expected values of the standardized ordered observations, i.e. $t_{(i)} = E(z_{(i)})$, by using the first two terms of Taylor series expansion:

$$g(z_{(i)}) \cong \alpha_i + \beta_i z_{(i)} \quad (i = 1, \dots, n)$$

where $\beta_i = \exp(-t_{(i)})$ and $\alpha_i = g(t_{(i)}) - \beta_i t_{(i)}$.

After replacing $g(\cdot)$ function with its linearized form in the equations (4) and (5), following modified likelihood equations are obtained:

$$\frac{\partial \ln L^*}{\partial \mu} = \frac{k}{\sigma} \sum_{i=1}^n [\alpha_i + \beta_i z_{(i)}] = 0 \quad (6)$$

and

$$\frac{\partial \ln L^*}{\partial \sigma} = -\frac{n}{\sigma} + \frac{k}{\sigma} \sum_{i=1}^n z_{(i)} [\alpha_i + \beta_i z_{(i)}] = 0. \quad (7)$$

Solutions of the equations (6) and (7) give the following MML estimators:

$$\hat{\mu}_{MML} = \bar{y}_w + \frac{\Delta}{m} \hat{\sigma}_{MML} \quad \text{and} \quad \hat{\sigma}_{MML} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}}. \quad (8)$$

Here,

$$\bar{y}_w = \frac{\sum_{i=1}^n \beta_i y_{(i)}}{\sum_{i=1}^n \beta_i}, \quad \Delta = \sum_{i=1}^n \alpha_i, \quad m = \sum_{i=1}^n \beta_i,$$

$$B = k \sum_{i=1}^n \alpha_i (y_{(i)} - \bar{y}_w), \quad C = k \sum_{i=1}^n \beta_i (y_{(i)} - \bar{y}_w)^2.$$

Remark 1. The denominator of $\hat{\sigma}_{MML}$ is replaced by $2\sqrt{n(n-1)}$ for bias correction.

Remark 2. Here, $t_{(i)} = E(z_{(i)})$ values cannot be obtained exactly. We therefore use the approximate values given by $t_{(i)} = Q(i/(n+1))$ where $Q(\cdot)$ is the quantile function of the EVk distribution given as follows:

$$Q(u) = \mu + \sigma \left\{ -\log \left(k^{-1} \Gamma^{-}(u, k) \right) \right\}, \quad 0 < u < 1.$$

Here, $\Gamma^{-}(\cdot, \cdot)$ denotes the inverse of the upper incomplete gamma function. The use of these approximate values does not affect the efficiency of the MML estimators adversely.

3. Simulation Study

In this section, we provide results of the MC simulation study to show performance of the proposed methodology. In our simulation setup, the k is taken to be 0.5, 3, 6 and 9. The data is generated from the EVk distribution based on three different sample sizes: $n = 30, 50$ and 100 . Without loss of generality μ and σ are taken to be 0 and 1, respectively. All the simulations are carried out for $\llbracket 100,000/n \rrbracket$ Monte-Carlo runs where $\llbracket \cdot \rrbracket$ denotes the integer value function. We use MATLAB2017a software for all computations. The performances of the estimators are evaluated using mean, variance and mean square error (MSE) criteria. We also use deficiency (DEF) criterion which is defined as the joint efficiencies of the estimators, i.e. $DEF = MSE(\hat{\mu}) + MSE(\hat{\sigma})$, see e.g. Kantar and Senoglu (2008). The results for parameter estimations are tabulated in Table 2.

It is clear from Table 2 that the biases of the ML estimator of μ are smaller than those of the corresponding MML estimator for all values of k . However, the MML estimator of σ is more preferable than the corresponding ML estimator in terms of their biases. The variances of the

ML and the MML estimators of μ are more or less the same and they are close to each other for larger values of the sample size. The variance of the ML estimator of σ is less than that of the MML estimator for all cases. According to the MSE criterion, the ML estimators of μ and σ are more desirable. However, the MML estimators have also promising results. Indeed, the MSEs of the MML estimators are almost the same as those of the ML estimators. This conclusion is also supported by the DEF values. In other words, DEF values of the ML and the MML estimators are close to each other but the ML is more preferable. It should also be noticed that the performance of the ML is better than that of MML estimator when $k = 0.5$. However, the performances of the ML and MML estimators are nearly the same for larger values of k according to the DEF criterion.

Table 2. Simulated mean, variance and MSE values for the estimators of μ and σ for certain values of k .

		$\hat{\mu}$			$\hat{\sigma}$			
n		Mean	Variance	MSE	Mean	Variance	MSE	DEF
$k = 0.5$								
30	ML	0.0258	0.0832	0.0839	0.9729	0.0224	0.0231	0.1070
	MML	0.0669	0.0847	0.0892	0.9895	0.0233	0.0234	0.1125
50	ML	0.0227	0.0634	0.0639	0.9823	0.0172	0.0175	0.0814
	MML	0.0543	0.0642	0.0672	0.9946	0.0177	0.0177	0.0849
100	ML	0.0201	0.0562	0.0566	0.9832	0.0149	0.0152	0.0718
	MML	0.0481	0.0571	0.0594	0.9939	0.0153	0.0153	0.0748
$k = 3$								
30	ML	0.0064	0.0117	0.0118	0.9747	0.0174	0.0180	0.0298
	MML	0.0151	0.0118	0.0120	0.9932	0.0182	0.0182	0.0302
50	ML	0.0014	0.0090	0.0090	0.9793	0.0131	0.0135	0.0225
	MML	0.0081	0.0090	0.0091	0.9934	0.0136	0.0136	0.0227
100	ML	0.0003	0.0078	0.0078	0.9819	0.0116	0.0119	0.0197
	MML	0.0063	0.0078	0.0079	0.9942	0.0119	0.0120	0.0198
$k = 6$								
30	ML	0.0041	0.0059	0.0059	0.9766	0.0177	0.0182	0.0241
	MML	0.0085	0.0059	0.0060	0.9945	0.0184	0.0184	0.0244
50	ML	0.0030	0.0068	0.0068	0.9802	0.0133	0.0136	0.0204
	MML	0.0081	0.0068	0.0069	0.9941	0.0137	0.0137	0.0206
100	ML	- 0.0001	0.0073	0.0073	0.9823	0.0115	0.0118	0.0191
	MML	0.0054	0.0073	0.0074	0.9947	0.0118	0.0118	0.0192
$k = 9$								
30	ML	0.0007	0.0038	0.0038	0.9746	0.0165	0.0171	0.0209
	MML	0.0036	0.0038	0.0038	0.9921	0.0171	0.0172	0.0210
50	ML	0.0021	0.0043	0.0043	0.9824	0.0128	0.0131	0.0174
	MML	0.0055	0.0043	0.0043	0.9959	0.0132	0.0132	0.0176
100	ML	0.0004	0.0047	0.0047	0.9838	0.0118	0.0121	0.0168
	MML	0.0041	0.0047	0.0047	0.9959	0.0122	0.0122	0.0169

4. Conclusion

In this study, the ML and MML estimations of the location and scale parameters of the EV k distribution are considered. Since the likelihood equations contain intractable terms, the ML estimators cannot be obtained explicitly. We therefore apply the MML methodology to obtain explicit estimators of the location and scale parameters. The case in which $k = 1$ is previously studied in the context of MML estimation, see e.g. Tiku and Akkaya (2004) and Aydin

(2017). Therefore, we investigate the effect of different values of k on the performances of the ML and MML estimators. For this purpose, the MC study is carried out. Results show that the ML estimators preferable to the MML estimators in terms of MSE and DEF criteria. Furthermore, the MML estimators are as efficient as ML estimators. The computation of ML estimators may be problematic in some cases, thus the MML estimators can also be preferred if the focus is computational ease as well as efficiency.

References

- Adeyemi, S. and Ojo, M.O (2003). “A generalization of the Gumbel distribution”. *Kragujevac J. Math.*, vol.25, pp. 19–29.
- Aydin, D. (2017). “Estimation of the lower and upper quantiles of Gumbel distribution: An application to wind speed data”, *Applied Ecology and Environmental Research*, vol. 16, pp. 1-15.
- Barnett, V.D. (1966). “Evaluation of the maximum likelihood estimator when the likelihood equation has multiple roots”. *Biometrika*, vol. 53, pp. 151-165.
- Demirhan, H. (2018). “A generalized Gumbel distribution and its parameter estimation”. *Communications in Statistics - Simulation and Computation*, vol. 47, no.10, pp. 2829-2848.
- Gumbel, E. J. (1935). “Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré*”, vol. 5, no. 2, pp. 115-158.
- Kantar, Y.M., Senoglu, B. (2008). “A comparative study for the location and scale parameters of the Weibull distribution with given shape parameter”, *Computers and Geosciences*, vol. 34, no. 12, pp. 1900–1909.
- LeDuc, S.K., Stevens, D.G. (1977). “Maximum Likelihood Estimates for Parameters of the m th Extreme Value Distribution”, *Journal of Applied Meteorology*, vol. 16 no. 3, pp. 251-254.
- Puthenpura, S., Sinha, N.K. (1986). “Modified maximum likelihood method for the robust estimation of system parameters from very noise data”. *Automatica*, vol. 22, pp. 231-235.
- Tiku, M. L. (1967). “Estimating the mean and standard deviation from a censored normal sample”, *Biometrika*, vol. 54, pp. 155-165.
- Tiku, M. L. (1968). “Estimating the parameters of Normal and Logistic distributions from censored samples”. *Aust. J. Stat.*, vol. 10, pp. 64-74.
- Tiku, M.L., Akkaya, A.D. 2004. “Robust Estimation and Hypothesis Testing”, *New Age International Publishers (Wiley Eastern)*, New Delhi, India.
- Vaughan, D.C. (1992). “On the Tiku-Suresh method of estimation”. *Commun. Stat.-Theory Meth.*, vol. 21, no. 2, pp. 451-469

O-61 Several Applications of New Generalized Entropy Optimization Methods in Survival Data Analysis

Aladdin Shamilov¹, Nihal İnce^{2*} and Sevda Ozdemir Calikusu³

¹Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey, asamilov@eskisehir.edu.tr

²Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey, nihalyilmaz@eskisehir.edu.tr

³Accountancy and Tax Department, Ozalp Vocational School, Van Yuzuncu Yil University, Turkey, sevdaozdemir@yyu.edu.tr

Abstract – In this paper, survival data analysis is realized by applying new Generalized Entropy Optimization Methods (GEOM) for solving Entropy Optimization Problems (EOP) consisting of optimizing a given entropy optimization measure subject to constraints generated by given moment vector functions. Mentioned problems in the form of GEOP2, GEOP3 based on GEOP1 have Generalized Entropy Optimization Distributions: GEOD2 in the form of $Min_{D} MaxEnt$, $Max_{D} MaxEnt$; GEOD3 in the form of $Min_{H} MinxEnt$, $Max_{H} MinxEnt$, where H is the Jaynes optimization measure, D is Kullback-Leibler optimization measure. It should be noted that formulation of GEOP1 uses only one optimization measure (H or D), however each of formulations of GEOP2, GEOP3 uses two measures H, D together. For this reason, survival data analysis by GEOD2 and GEOD3 acquires a new significance. In this research, given survival data is examined as application of developed new method. The performances of GEOD2 and GEOD3 are established by Chi-Square criteria, Root Mean Square Error (RMSE) criteria, H and D measures.

Keywords – Generalized entropy optimization methods, Jaynes optimization measure, Kullback-Leibler optimization measure, Survival data analysis

1. Introduction

Entropy Optimization Methods (EOM) are formulated and methods realizing these principles are suggested by Kapur and Kesavan (1992). EOM can be applied to different statistical problems especially in energy field, economy, survival data analysis by Shamilov et al. (2007, 2008, 2013 and 2017), Zhoi et. al. (2013). In this study, a generalization of entropy optimization problems is formulated as Generalized Entropy Optimization Problem (GEOP). The method of solving GEOP we call as Generalized Entropy Optimization Method (GEOM) and the solution of GEOP as Generalized Entropy Optimization Distribution (GEOD).

GEOM have proposed distributions in the form of the MinMaxEnt, the MaxMaxEnt, the MinMinxEnt and the MaxMinxEnt how close to or far from statistical data (or distribution) in

the sense of H or D measures by Shamilov (2006, 2009 and 2010). Therefore, GEOM can be more successfully applied in survival data analysis. Different aspects and methods of investigations of survival data analysis are considered. In particular, it is investigated several problems of hazard rate function estimation based on the maximum entropy principle. The potential applications include developing several classes of the maximum entropy distributions which can be used to model different data-generating distributions that satisfy certain information constraints on the hazard rate function by Ebrahimi (2000) and Joly et. al. (2012).

Our study consists of following sections. In Section 2, formulations of new Generalized Entropy Optimization Problems (GEOP2,3) represented by Shamilov and Ince (2016) are described. In Section 3, survival data analysis is fulfilled by applying GEOM. Finally, the main results obtained from the study are summarized and problems for future applications are expressed.

2. Materials and Methods

Entropy Optimization Problem (EOP) and Generalized Entropy Optimization Problem (GEOP1) are given in detail

by Shamilov (2009). Note that formulations of GEOP2,3 and GEOD2,3 are introduced by Shamilov and Ince (2016) are given in the following form.

GEOP2: Let $f^{(0)}(x)$ be given probability distribution of random variable X . H be Jaynes entropy measure, D be Kullback-Leibler measure and K be a set of given moment vector functions $g(x)$ generating moment vector conditions. It is required to choose moment vector functions $g^{(1)}, g^{(2)} \in K$ such that $g^{(1)}(x)$ generates MaxEnt distribution $f^{(1)}(x)$ Kullback-Leibler measure D of which has minimum value, $g^{(2)}(x)$ generates MaxEnt distribution $f^{(2)}(x)$ Kullback-Leibler measure D of which has maximum value on K . MaxEnt distribution generated by $g^{(1)} \in K$ we call as $Min_D MaxEnt$ distribution, MaxEnt distribution generated by $g^{(2)} \in K$ we call $Max_D MaxEnt$ distribution.

The $Min_D MaxEnt$ distribution represents the MaxEnt distribution $f^{(1)}(x)$ which closest to $f^{(0)}(x)$, $Max_D MaxEnt$ distribution represents the MaxEnt distribution $f^{(2)}(x)$ which furthest from $f^{(0)}(x)$ in the sense of D measure.

GEOP3: Let $f^{(0)}(x)$ be given probability distribution of random variable X . H be Jaynes entropy measure, D be Kullback-Leibler measure and K be a set of given moment vector functions $g(x)$ generating moment vector conditions. It is required to choose moment vector functions $g^{(1)}, g^{(2)} \in K$ such that $g^{(1)}(x)$ generates MinxEnt distribution $f^{(1)}(x)$ Jaynes measure H of which has minimum value, $g^{(2)}(x)$ generates MinxEnt distribution $f^{(2)}(x)$

Jaynes measure H of which has maximum value on K . MinxEnt distribution generated by $g^{(1)}(x)$ we call $Min_{H} MinxEnt$ distribution, MinxEnt distribution generated by $g^{(2)}(x)$ we call $Max_{H} MinxEnt$ distribution.

The $Min_{H} MinxEnt$ distribution represents the MinxEnt distribution $f^{(1)}(x)$ which closest to $f^{(0)}(x)$, $Max_{H} MinxEnt$ represents MinxEnt distribution $f^{(2)}(x)$ which furthest from $f^{(0)}(x)$ in the sense of H measure.

3. Application

3.1 Analysis of Results of GEOD2 and GEOD3

In this application, GEOM2 and GEOM3 new methods developed by Shamilov and Ince (2016) are applied in solving proper problems in survival data analysis. In the present research, the data of the life table for engine failure data (1980) is investigated by Deshpande and Purohit (2005).

For mentioned data, the experiment is planned for 200 numbers of patients surviving at beginning of interval but the presence of censoring from the planning patients 97 individuals stay out the experiment.

Table 1. The data of the life table for engine failure data (1980) and Observed and Corrected probabilities

Survival Time (year) t	Working at the beginning of interval n_i	Failed during the interval d_i	Censored during the interval c_i	Observed probabilities p_i	Corrected probabilities p_i^*
0-1	200	5	0	0.0485	0.0485
1-2	195	10	1	0.0971	0.1068
2-3	184	12	5	0.1165	0.1650
3-4	167	8	2	0.0777	0.0971
4-5	157	10	0	0.0971	0.0971
5-6	147	15	6	0.1456	0.2039
6-7	126	9	3	0.0874	0.1165
7-8	114	8	1	0.0777	0.0874
8-9	105	4	0	0.0388	0.0388
9-10	101	3	1	0.0291	0.0388

It should be noted that, the presence of censoring in the survival times leads to a situation where the sum of observation probabilities stands less than 1 for the survival data. For this reason, in solving many problems, it is required to supplement the sum of observation probabilities up to 1. Since the sum of observed probabilities p_i in Table 1 is 0.8155,

according to the number of censoring, supplementary probability $1-0.8155 = 0.1845$ is uniformly distributed to each censoring data and corrected probabilities p_i^* are obtained.

In our investigation as components of K_0 characterizing moment vector functions

$$g_1(x) = x, g_2(x) = x^2, g_3(x) = \ln x, g_4(x) = (\ln x)^2, g_5(x) = \ln(1 + x^2)$$

are chosen which are mostly used in statistics. Consequently, $K_0 = \{g_1, \dots, g_5\}$. For example, if $m = 2$ then $(g_0, g^{(1)}) = (1, x, \ln x)$, $g^{(1)} \in K_{0,2}$ gives the least value to $U(g)$ and $(g_0, g^{(2)}) = (1, x, (\ln x)^2)$, $g^{(2)} \in K_{0,2}$ gives the greatest value to $U(g)$.

3.1.1 The Performance of GEOD2

In order to obtain the performance of GEOD2, we use various criteria as Root Mean Square Error (RMSE), Chi-Square(χ^2), entropy values of distributions. The best distribution function can be determined according to the lowest values of RMSE, χ^2 , D measure. It is obtained by comparison of the $Min MaxEnt$ and $Max MaxEnt$ distributions by using the above given criteria.

Table 2. The obtained results for $(Min MaxEnt)_m^D$, $m = 1,2,3,4$

Distribution of $Min MaxEnt_D$	$D(p : q)$	Calculated value of Chi – Square	Table value of Chi – Square	RMSE
$(Min MaxEnt)_1^D$	0.3938	4.4310	$\chi_{8,\alpha}^2 = 15.51$	0.3158
$(Min MaxEnt)_2^D$	0.3310	1.8896	$\chi_{7,\alpha}^2 = 14.07$	0.1909
$(Min MaxEnt)_3^D$	0.3300	1.7787	$\chi_{6,\alpha}^2 = 12.59$	0.1799
$(Min MaxEnt)_4^D$	0.3193	1.6161	$\chi_{5,\alpha}^2 = 11.07$	0.1830

Table 3. The obtained results for $(Max MaxEnt)_m^D$, $m = 1,2,3,4$

Distribution of $Max MaxEnt_D$	$D(p : q)$	Calculated value of Chi – Square	Table value of Chi – Square	RMSE
$(Max MaxEnt)_1^D$	0.4438	5.3820	$\chi_{8,\alpha}^2 = 15.51$	0.3492
$(Max MaxEnt)_2^D$	0.4007	4.9233	$\chi_{7,\alpha}^2 = 14.07$	0.3492
$(Max MaxEnt)_3^D$	0.3458	2.2804	$\chi_{6,\alpha}^2 = 12.59$	0.2104
$(Max MaxEnt)_4^D$	0.3199	1.6383	$\chi_{5,\alpha}^2 = 11.07$	0.1888

From Tables 2,3, it is shown that all distributions $\left(\text{Min MaxEnt}\right)_m^D$ and $\left(\text{Max MaxEnt}\right)_m^D$, $m = 1,2,3,4$ are acceptable to survival data in the sense of χ^2 criteria. Also, in the sense of RMSE criteria each $\left(\text{Min MaxEnt}\right)_m^D$ distribution is better than corresponding $\left(\text{Max MaxEnt}\right)_m^D$ distribution. Therefore, $\left(\text{Min MaxEnt}\right)_1^D$ is nearer to survival data than $\left(\text{Max MaxEnt}\right)_1^D$ and $\left(\text{Max MaxEnt}\right)_2^D$ distributions; each $\left(\text{Min MaxEnt}\right)_m^D$ ($m = 2,3,4$) is better than all of $\left(\text{Max MaxEnt}\right)_m^D$ ($m = 1,2,3,4$) distributions. From these results follows that among of distributions $\left(\text{Min MaxEnt}\right)_m^D$, ($m = 1,2,3,4$) the distribution $\left(\text{Min MaxEnt}\right)_3^D$ is more suitable; among of distributions $\left(\text{Max MaxEnt}\right)_m^D$, ($m = 1,2,3,4$) the distribution $\left(\text{Max MaxEnt}\right)_4^D$ is more convenient for survival data in the sense of RMSE criteria.

It is indicated that the Min MaxEnt distribution is more suitable than the Max MaxEnt distribution in modeling survival data and the Min MaxEnt distributions are getting better while the number of moment constraints is added.

3.1.2 The Performance of GEOD3

In order to obtain the performance of the mentioned distributions. The parameters for the statistical analysis: RMSE and χ^2 and H are given in Tables 4,5 for all Min MinxEnt and Max MinxEnt distributions based on survival data.

Table 4. The obtained results for $\left(\text{Min MinxEnt}\right)_m^H$, $m = 1,2,3,4$

Distribution of Min MinxEnt_H	H	Calculated value of Chi – Square	Table value of Chi – Square	RMSE
$\left(\text{Min MinxEnt}\right)_1^H$	3.1912	0.3571	$\chi_{8,\alpha}^2 = 15.51$	0.0896
$\left(\text{Min MinxEnt}\right)_2^H$	3.1665	0.2996	$\chi_{7,\alpha}^2 = 14.07$	0.0839
$\left(\text{Min MinxEnt}\right)_3^H$	3.1666	0.3109	$\chi_{6,\alpha}^2 = 12.59$	0.0881
$\left(\text{Min MinxEnt}\right)_4^H$	3.1692	0.2924	$\chi_{5,\alpha}^2 = 11.07$	0.0797

Table 5. The obtained results for $(Max MinxEnt)_m^H$, $m = 1,2,3,4$

Distribution of $Max MinxEnt_H$	H	Calculated value of Chi – Square	Table value of Chi – Square	RMSE
$(Max MinxEnt)_1^H$	3.1926	0.3690	$\chi_{8,\alpha}^2 = 15.51$	0.0916
$(Max MinxEnt)_2^H$	3.1727	0.3282	$\chi_{7,\alpha}^2 = 14.07$	0.0892
$(Max MinxEnt)_3^H$	3.1696	0.2915	$\chi_{6,\alpha}^2 = 12.59$	0.0792
$(Max MinxEnt)_4^H$	3.1701	0.2909	$\chi_{5,\alpha}^2 = 11.07$	0.0791

From Tables 4,5, it is shown that all $(Min MinxEnt)_m^H$, $(Max MinxEnt)_m^H$, $m = 1,2,3,4$ distributions are acceptable to survival data in the sense of χ^2 criteria. Also, in the sense of RMSE criteria each $(Min MinxEnt)_m^H$ ($m = 1,2$) distribution is better than corresponding $(Max MinxEnt)_m^H$ ($m = 1,2$) distribution. But, the $Max MinxEnt_H$ distributions are getting better while the number of moment constraints is added. Therefore, $(Max MinxEnt)_m^H$ ($m = 3,4$) is nearer to survival data than $(Min MinxEnt)_m^H$ ($m = 3,4$) distributions. From these results follows that $(Max MinxEnt)_4^H$ distribution is more suitable for survival data than other distributions.

Finally, it can also be concluded that the GEOD3 fit better to different types of data than the GEOD2. These results are also supported by the illustrations in Figure 1 for each data. It can be observed in the all figures that the GEOD3 closely match the measured data much with better than the GEOD2, particularly for the survival data in Figure 1.

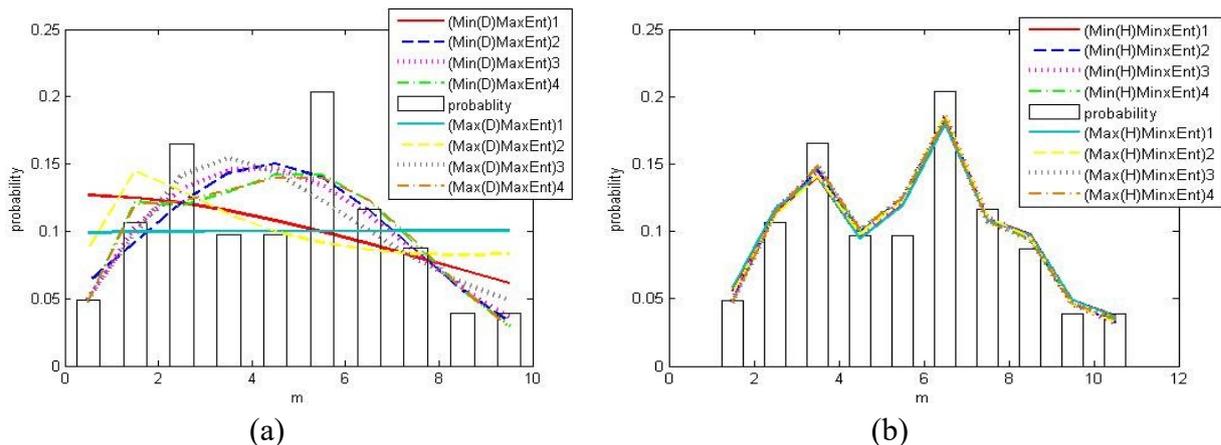


Figure 1. (a) Graphic of the comparison of $(Min(D)MaxEnt)_m$ and $(Max(D)MaxEnt)_m$, $m = 1,2,3,4$ distributions for survival data and histogram
 (b) Graphic of the comparison of $(Min(H)MinxEnt)_m$ and $(Max(H)MinxEnt)_m$, $m = 1,2,3,4$ distributions for survival data and histogram

4. Survival Expressions of $\left(\text{Max MinxEnt}\right)_H^4$ Distribution

In this section, $\left(\text{Max MinxEnt}\right)_H^4$ distribution is more presentable for the data of the life table for engine failure data (1980) given in Table 1 among GEOD2 and GEOD3 in the sense of RMSE criteria. $\left(\text{Max MinxEnt}\right)_H^4$ estimation of Probability Density Function $\hat{f}(t)$, Cumulative Distribution Function $\hat{F}(t)$, Survival Function $\hat{S}(t)$ and Hazard Rate $\hat{h}(t)$ is calculated using by formulations in Shamilov et. al. (2013) and obtained results are shown by graphical representation in Figure 2.

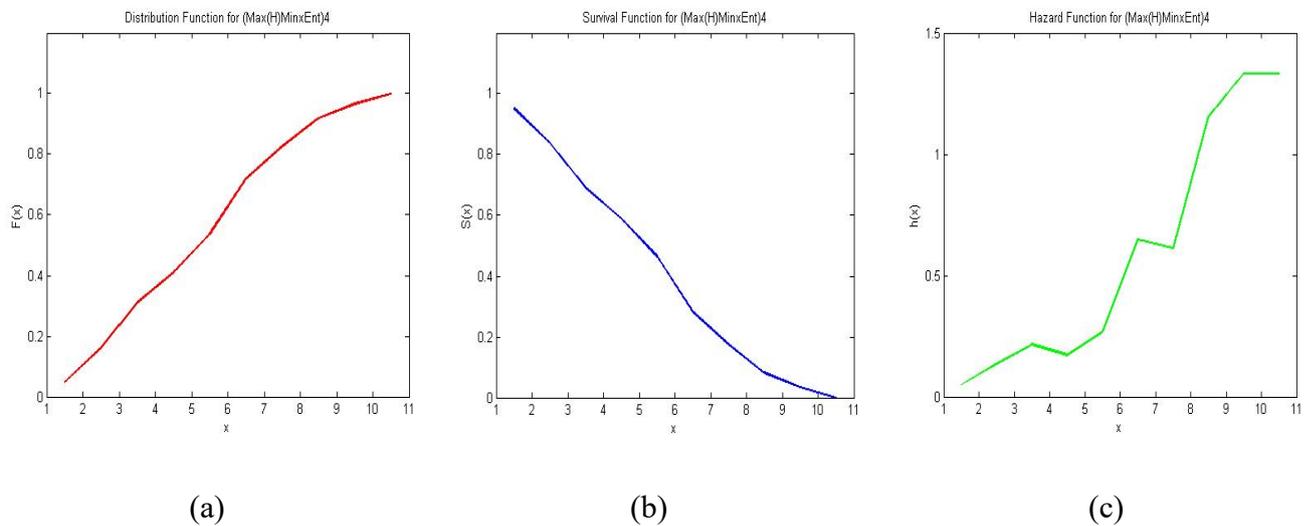


Figure 2. Survival expressions of distribution $\left(\text{Max MinxEnt}\right)_H^4$

5. Conclusion

In this study it is established that survival data analysis is realized by applying new Generalized Entropy Optimization Methods (GEOM2,3). Generalized Entropy Optimization Distributions (GEOD2,3) which are obtained on basis of Jaynes optimization measure (H) and Kullback-Leibler optimization measure (D) and supplementary optimization with respect to characterizing moment functions, more exactly represents the given statistical data. For this reason, survival data analysis by GEOD2,3 acquire a new significance. The performances of GEOD2,3 are established by Chi-Square criteria, RMSE criteria, H and D measure. The comparison of GEOD2 in the difference senses shows that along of GEOD2

distribution $\left(\text{Min MaxEnt}\right)_D^4$ is better than distribution $\left(\text{Max MaxEnt}\right)_D^4$. Also, the comparison of GEOD3 in the difference senses shows that along of GEOD3, distribution $\left(\text{Min MinxEnt}\right)_H^4$ is better than distribution $\left(\text{Max MinxEnt}\right)_H^4$. According to obtained distribution $\left(\text{Max MinxEnt}\right)_H^4$ estimator of Probability Density Function $\hat{f}(t)$, Cumulative Distribution Function $\hat{F}(t)$, Survival Function $\hat{S}(t)$ and Hazard Rate $\hat{h}(t)$ are evaluated and graphically illustrated. Our investigation indicates that GEOM2,3 in survival data analysis yields reasonable results.

References

- Deshpande J. V., Purohit S.G. (2005). Life Time Data: Statistical Models and Methods, Series on Quality, Reliability and Engineering Statistics, India.
- Kapur J.N., Kesavan H.K. (1992). Entropy Optimization Principles with Applications. Academic Press, New York, pp.408.
- Lee E. T., Wang J.W. (2003). Statistical Methods for Survival Data Analysis, Wiley-Interscience, Oklahoma.
- Nader Ebrahimi (2000). “The Maximum Entropy Method for Lifetime Distributions”, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), vol.62, no.2, pp. 236-243.
- Shamilov A. (2006). “A Development of Entropy Optimization Methods, Wseas Transactions on Mathematics”, vol.5, no.5, pp.568-575.
- Shamilov, A. (2007). “Generalized Entropy Optimization Problems and the Existence of Their Solutions”. Physica A: Statistical Mechanics and its Applications. vol.382, no.2, pp. 465-472.
- Shamilov A., Mert Kantar Y., Usta I. (2008). "Use of MinMaxEnt distributions defined on basis of MaxEnt method in windpowerstudy", Energy Conversion & Management, vol.49, pp.660-677.
- Shamilov A. (2009). Entropy, Information and Entropy Optimization. Turkey.
- Shamilov A. (2010). “Generalized entropy optimization problems with finite moment function sets”, Journal of Statistics and Management Systems, vol.13, no.3, pp. 595-603.

Shamilov A., Ince N. (2016). “On Several New Generalized Entropy Optimization Methods”, Turkey Clinics J Biostat, vol.8, no.2, pp. 110-115.

Shamilov A., Kalathilparmbil C., Ozdemir S. (2017). “An Application of Generalized Entropy Optimization Methods in Survival Data Analysis”, Journal of Modern Physics, vol.8, no.3, pp. 349-364.

O-64 A Panel Data Analysis: The Relationship Between Unemployment, Youth Unemployment and Economic Growth

Merve ALTAYLAR^{1*} and Hamdi EMEÇ²

¹*Econometrics, Social Sciences Institute, Dokuz Eylül University, Turkey,
mervealtaylar37@gmail.com*

²*Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylül University,
Turkey, hamdi.emec@deu.edu.tr*

Abstract *This paper examines the relationship between unemployment, youth unemployment and economic growth in 20 OECD countries between 2000 and 2017. These variables are examined using static and dynamic panel time series techniques. He does not examine the relationship between these variables, as Okun suggested, and explores how unemployment and youth unemployment affect economic growth. Conventional panel unit root and panel break root and panel cointegration tests were used to investigate the differences of these variables on economic growth. As a result, youth unemployment is more susceptible to economic growth than unemployment, while developments in youth unemployment do not affect economic growth as well as unemployment.*

Keywords – *Multiple Structural Break Panel Cointegration, Heterogeneous Panel, Cross-sectional Dependence, DOLS, PANKPSS*

1. Introduction

Unemployment, which has become a global problem, has far more serious effects, especially for some groups in terms of causes and consequences. Young people are the leading groups. Thus, the phenomenon of youth unemployment is now becoming accepted among the global problems in the economic literature. When the unemployment indicators of developed or developing countries are analyzed, it is seen that youth unemployment rates are at a much higher level compared to the total unemployment rates. If moderate economic growth is achieved in order to reduce unemployment, the overall level of unemployment will decrease in relation to the increase in economic activity level, thus providing a constructive impact on youth unemployment. Okun's Law (1962), which is accepted as one of the basic laws in economics, analyzes the relationship between unemployment and economic growth. Okun's Law (1962) suggests that there is an inverse relationship between macroeconomic aggregates based on the model.

The aim of this study is to examine the relationship between unemployment and youth unemployment and economic growth.

For this purpose, the macroeconomic relationship between unemployment, youth unemployment and GDP variables was examined by panel data analysis methods with the data of 24 OECD member countries between 2000-2017.

2. Materials and Methods

In this study, three macroeconomic indicators such as GDP, unemployment rate and youth unemployment rate were analyzed and the scope of the application was decided in 24 OECD² countries. Thus, a panel data set with cross-section and time dimension was formed. The relevant variables were included in the analysis between 2000 and 2017 and quarterly and the data were collected from the OECD database. Detailed information on macroeconomic variables is shown in Table 1 below.

Table 1. Variables

Variables	Series	Data Type	Period
Unemployment Rate	Seasonally Adjusted	Level Rate	2000 Q1-2017 Q4
Youth Unemployment Rate	Seasonally Adjusted	Level Rate	2000 Q1-2017 Q4
GDP	Seasonally Adjusted	Currency \$ Dollar	2000 Q1-2017 Q4

Table 2 shows the descriptive statistics of the three variables subject to analysis (Gross Domestic Product, unemployment rate and youth unemployment rate). According to these indicators, the panel data set consists of 24 sections and 72 time dimensions for all three macroeconomic variables, showing a balanced and long panel ($T > N$).

Table 2. Descriptive Statistics of Macroeconomic Variables

Variables	Obs. Per Group (N)	Time (T)	Observation	Mean
GDP	24	72	1724	1213162
Unemployment Rate	24	72	1724	8.01
Youth Unemployment Rate	24	72	1724	18.07

In this study, whether the macroeconomic variables are stationary or not is analyzed by methods appropriate to the structure of the panel time series in order to avoid spurious regression problems. However, since the analysis was carried out with panel time series, the concept of cross-sectional dependence specific to this structure gained priority. In this way, cross-sectional dependence was tested with Pesaran CD test for three macroeconomic variables and the results showed that all three variables had cross-sectional dependency problems. Thus, macroeconomic variables were subjected to the Pesaran CIPS and PANKPSS panel stationary tests, which are the second generation panel unit root tests which have the

² There are 36 member countries of the OECD (OECD, www.oecd.org, 09.06.2019). Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, Greece, Hungary, Ireland, Italy, South Korea, Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Spain, England, USA, Estonia, Israel and Slovenia. Due to limitations in the data sets Chile, France, Germany, Iceland, Japan, Latvia, Lithuania, Luxembourg, Mexico, Sweden, Switzerland and Turkey were excluded from the study. Among these countries, Hungary and Poland have the status of developing countries while other countries have the status of developed countries.

assumption of cross-sectional dependence. According to the common results of panel unit root and stationarity analysis tests performed for the whole panel, it was found that all three variables were not stationary at the level. When the first differences of these variables were taken and the same process was repeated, the variables became stationary. After the panel unit root and stationary analysis, it was examined whether the Unemployment Rate-GDP model is cointegrated. In the continuation of the analysis, panel cointegration analyzes were performed for the second model, the GDP-Unemployment Rate model, the third model for the Youth Unemployment Rate-GDP model, and finally for the GDP-Youth Unemployment Rate model. However, in this study, because of the analysis of panel time series, the concept of parameter heterogeneity has gained importance in addition to the cross-sectional dependence specific to this structure. For the four models examined, cross-sectional dependence was determined by Breusch Pagan LM test; parameter homogeneity was tested with Swamy S test. The results revealed cross-sectional dependence and parameter heterogeneity in all four models. The cointegration relations in these models were investigated by using Westerlund (2009) panel cointegration test with Multiple Structural Breaks, which has the assumption of cross-sectional dependence, parameter heterogeneity and structural break. Thus, in order to estimate long-term relationships, the panel cointegration model estimator DOLS, which is the assumption of cross-sectional dependence and parameter heterogeneity, was preferred, and the model without cointegration relation was estimated with the AMG estimator. Stata 14 and Gauss 10 software were used during the analyzes.

2.1 Testing the Cross-sectional Dependency

The cross-sectional dependence, defined as the interaction between the units that make up the panel (eg households, firms, countries, etc.), can be regarded as the equivalent of the series correlation in time series. This can occur in behavioral interactions between individuals, consumers in a community, or firms working in the same sector. It can also be caused by unobservable common factors or common shocks that are very common in macroeconomics. As with the correlation problem observed in time series, cross-sectional dependence leads to loss of productivity in the OLS estimator and may result in the invalidation of traditional t tests and F tests using standard variance covariance estimators, and in some cases even inaccuracy. For this reason, it is emphasized that it is wise to test cross-sectional dependence before starting the analysis (Baltagi and Kao, 2012: 137). In this section, Pesaran CD test is introduced to investigate cross-sectional dependence.

Table 3. Pesaran CD Test Results

Variables	CD-Test Stat.	Correlation Coefficient	Prob.
Unemployment Rate	32.65	0.232	0.000*
Youth Unemployment Rate	34.76	0.247	0.000*
GDP	103.62	0.735	0.000*

Table 3 shows the results of the Pesaran CD Test for measuring cross-sectional dependence for all variables examined.

According to these results, the null hypothesis, which states that there is no dependency between units' errors in all three variables at 5% significance level, was rejected and it was found that there was a cross-sectional dependency problem in the panel. Cross-sectional dependence is usually; country, region, city etc. This problem is expected to be encountered when working with units. The mean correlation coefficients of the variables were 23% for the unemployment series, 24% for the youth unemployment series and 73% for the GDP series. Therefore, first generation panel unit root tests are inadequate in analyzing these variables since they do not consider the cross-sectional dependence problem. Therefore, in this study, second generation panel unit root tests which take the cross-sectional dependence problem into consideration were preferred.

2.2 Panel Unit Root and Stationary Tests

Unit root presence testing has become a common practice among researchers in time series analysis. However, it is stated that unit root research in panel datasets is more recent in the literature than time series. In the literature, panel unit root tests are divided into two groups according to cross-sectional dependence. The first generation panel unit root tests work with the assumption that there is no cross-sectional dependence, while the second generation panel unit root tests work with the presence of cross-sectional dependence (Chen and Lu, 2003:343).

The assumption that there is no cross-sectional dependence when analyzing with panel data is seen as a very strict restriction in applied research. For this reason, second generation panel unit root tests were developed considering cross-sectional dependence (Levin, Lin and Chu, 2002: 14).

CIPS panel unit root test developed by Pesaran (2007) is based on the logic of modeling cross-sectional dependence through factors. In his study, which uses the horizontal cross-sectional mean of the individual series forming the sections as a tool variable for the factors not observed in the model, he suggested that this approach eliminates the cross-sectional dependence. It is suggested that ADF regression is extended with the horizontal cross-sectional mean and delayed values of the series and when this first regression difference is taken, it eliminates the cross-sectional dependence. When traditional unit root tests are used, it can be concluded that the series is not stationary if there is a structural break in the series. For this case, Lee and Strazicich (2003) found that the element that disrupts the stationary of the time series is in fact caused by structural breaks, thus concluding that the time series is actually stationary. To solve this problem, Carrion-i Silvestre et al. (2005) developed the PANKPSS test. This test has the null hypothesis that the panel is stationary and has a test statistic that allows for the presence of multiple structural breaks, taking into account the problem of cross-sectional dependence. Two different characteristics are considered depending on the fixed term and / or structural breaks that occur in the trend. The statistical model is flexible enough to allow the number of fractures and times to differ between sections (Carrion-i-Silvestre vd., 2005: 162). Table 4

shows the results of panel unit root and stationarity analysis for all three macroeconomic variables.

Table 4. Panel Unit Root and Stationary Tests

Unemployment Rate				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	4.295	1.0000	5.385	1.0000
PANKPSS	-	-	4.251	0.000
Δ Unemployment Rate				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	-13.697	0.0000	-13.697	0.0000
PANKPSS	-	-	-0.512	0.697

GDP				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	2.500	0.994	2.723	0.9970
PANKPSS	-	-	14.243	0.000
Δ GDP				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	-18.089	0.000	-17.167	0.0000
PANKPSS	-	-	-0.503	0.692

Youth Unemployment Rate				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	2.790	0.997	2.685	0.9960
PANKPSS	-	-	11.244	0.000
Δ Youth Unemployment Rate				
Equation	Constant		Constant and Trend	
Unit Root Tests	Stat.	Prob.	Stat.	Prob.
CIPS	-19.165	0.0000	-18.802	0.0000
PANKPSS	-	-	-0.212	0.584

Table 4 shows the CIPS and PANKPSS test results of the unemployment variable. According to these results, it was found that the unemployment variable was not stationary at the level. However, when the difference of unemployment variable from the first order is taken, it has been found that this variable becomes stationary. Table 4 (part two) shows the CIPS and PANKPSS test results of the GDP. According to these results, it was found that the GDP variable was not stationary at the level. However, when the difference of GDP from the first order is taken, it has been found that this variable becomes stationary. Table 4(part three) shows the CIPS and PANKPSS test results of the youth unemployment variable. According to these results, it was found that the unemployment variable was not stationary at the level. However, when the difference of youth unemployment variable from the first order is taken, it has been found that this variable becomes stationary. In this case, it is seen that all three variables are not stationary at the level, but the first difference of these variables is stationary.

2.3 Testing the Parameter Heterogeneity

Another important concept in panel data is parameter heterogeneity. When working with panel data, first of all, it must be tested whether the regression parameters of the sections forming the panel are specific to the relevant section. If there is heterogeneity in the regression parameters, tests and / or estimation methods that take this heterogeneity into consideration should be preferred. Otherwise, choosing prediction methods that ignore such

heterogeneity may lead to inconsistent or meaningless estimation of parameters (Tatoğlu, 2018: 51).

In this study, Swamy S test was used to test parameter heterogeneity. The first studies for testing homogeneity in panel data analysis studies were made by Swamy (1970). In Swamy's S test, the null hypothesis states that the parameters of the sample units examined are homogeneous (Tatoğlu, 2017: 247). Table 5 shows the results of the parameter heterogeneity test.

Table 5. Swamy S Test for Parameter Heterogeneity

Models	Parameter Heterogeneity Test
Model 1 (Unemployment Rate-GDP)	Chi-Square Stat. 28578.56
	Chi-Square Prob. 0.0000*
Model 2 (GDP-Unemployment Rate)	Chi-Square Stat. 9.300
	Chi-Square Prob. 0.0000*
Model 3 (Youth Unemployment Rate-GDP)	Chi-Square Stat. 25197.41
	Chi-Square Prob. 0.0000*
Model 4 (GDP-Youth Unemployment Rate)	Chi-Square Stat. 7.100
	Chi-Square Prob. 0.0000*

According to the test results, the null hypothesis that the parameters of 5% significance level was homogeneous was rejected. Parameter heterogeneity is observed in these models. Therefore, second generation panel cointegration tests that take into account parameter heterogeneity should be selected.

2.4 Panel Cointegration Test With Sutstructural Breaks

Panel cointegration test with structural break developed by Westerlund (2009) considers cross-sectional dependence, parameter heterogeneity and multiple structural breaks. This test allows up to three structural breaks in the model. The test also includes a constant or a constant and a break in trend. Table 6 shows the results of the Westerlund Panel Cointegration Test with Multiple Structural Breaks for the four models examined.

Table 6. Westerlund (2009) Panel Cointegration Test with Multiple Structural Breaks

Models	Coef.	Bootstrap Prob.
Model 1 (Unemployment Rate-GDP)	1.412	0.004*
Model 2 (GDP-Unemployment Rate)	10.553	0.180
Model 3 (Youth Unemployment Rate-GDP)	6.501	0.840
Model 4 (GDP-Youth Unemployment Rate)	7.405	0.880

According to the results shown in Table 6, the null hypothesis stating that there is a cointegration relationship at 5% significance level is rejected. Thus, considering the breaks, there is no evidence of the cointegration relationship in the Unemployment Rate-GDP model.

The null hypothesis, which states that there is a cointegration relationship at 5% significance level according to the calculated probability for the second model (GDP-Unemployment Rate), could not be rejected. In this case, there is a cointegration relationship between the two variables and it is necessary to estimate the cointegration model in order not to lose this long-term synchronous relationship structure. According to the results of cointegration analysis for the third and fourth models, the null hypothesis stating that there is a cointegration relationship at 5% significance level could not be rejected. Thus, a long-term relationship was found between the variables of the third and fourth models. Therefore, three of these models (2,3,4) should be estimated within the framework of panel cointegration model.

2.5 Dynamic Ordinary Least Squares (DOLS) Estimator

In the literature, there are many estimation methods used for estimating the panel cointegration model for the variables that have been determined to have cointegration relations between them.

Dynamic Least Squares (DOLS) estimator is among the estimators that can be preferred in the presence of parameter heterogeneity and cross-sectional dependence in estimating the long-term relationship between the cointegrating variables. In this estimator, firstly, the cointegration model is estimated for each section and in the next step these estimation results are combined for the whole panel with the Pesaran and Smith Mean Group (MG) approach.

Table 7. Long Term Parameters (DOLS Estimator)

Models	Coef.	t Stat.
Model 2 (GDP-Unemployment Rate)	-0.168	-31.34*
Model 3 (Youth Unemployment Rate-GDP)	-2.281	-29.32*
Model 4 (GDP-Youth Unemployment Rate)	-0.165	-28.38*

For the GDP-Unemployment Rate model (2), the t-statistic calculated at 5% significance level is greater than the critical value, so that the long-term parameter is statistically significant. According to this result, when unemployment increased by 1%, GDP displayed an average decrease of approximately 0.17% in the whole panel. Therefore, it can be said that the increase in unemployment causes a decrease in GDP for the examined countries.

According to the results of the second model, long-term coefficient was found to be statistically significant and there was an inverse relationship between youth unemployment and GDP. This situation is in harmony with the economic literature. 1% increase in GDP variable reduces youth unemployment by 2.28% on average.

According to the results of the panel-based cointegration model, a statistically significant negative correlation was found between GDP and Youth Unemployment Rate variables. The 1% increase in youth unemployment has a 0.16% reduction in GDP.

2.6 Augmented Mean Group (AMG) Estimator

There are many estimators in the literature in order to estimate panel data models. This estimator, developed by Eberhardt and Teal (2010) and Bond and Eberhardt (2009) in this section, is used especially for estimating panel time series. It is resistant to cross-sectional dependence, parameter homogeneity and non-stationary variables. At this stage, the first model, which was not found to be cointegration regression, was estimated with the AMG estimator. Table 8 shows the panel-based regression results.

Table 8. Regression Coefficients (AMG Estimator)

Model 1 (Unemployment Rate-GDP)		
Variables	Coef.	Prob.
Slope	-1.062	0.001**
Common Factor	1.069	0.000*
Chi-Square	-	0.0009*

According to the results shown in Table 8, the model (chi-square probability) is a statistically significant and reliable regression. According to the estimation results, both variables had a statistically significant effect on the dependent variable. The 1% increase in GDP has a 1.06% reduction in unemployment rate.

3. Conclusion

The four models examined in this study were tested with the Westerlund (2009) panel cointegration test with Multiple Structural Breaks, which investigated the cointegration relationship under the assumption of cross-sectional dependence, parameter heterogeneity and structural breaks. With this test, the cointegration relationship could not be reached only in the Unemployment Rate-GDP model of the four models examined, and the cointegration relationship was found in the other three models. As a result of the analyzes, 1% percentage point increase in GDP decreased the unemployment rate by 1.06%, while %1 percentage point increase in GDP reduced the youth unemployment rate by 2.29%. Thus, it was found that youth unemployment is almost two and a half times more sensitive to economic growth than unemployment. At the same time, 1% increase in the unemployment rate reduces GDP by 0.17%, while 1% increase in youth unemployment reduces GDP by 0.16%. The conclusion from these rates is that youth unemployment is a very fragile structure and if it is reduced by countries with the right policies, it will affect economic growth as much as unemployment.

References

Baltagi, B. H. , Feng, Q. and Kao, C. (2012). “A Lagrange Multiplier Test for Cross-Sectional Dependence in a Fixed Effects Panel Data Model”, *Journal of Econometrics*, vol. 1, pp. 164-177.

Carrion-i-Silvestre, J. , Castro, T. and López-Bazo, E. (2005). “Breaking The Panels: An Application to the GDP Per Capita”, *The Econometrics Journal*, vol.2, pp.159-175.

Chen, J. ve Lu, W. (2003). “Panel Unit Root Tests of Firm Size and Its Growth.*Applied Economics Letters*”, vol.10, pp.343-345.

Levin, A. , C.F. Lin and C-S.J. Chu (2002). “Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties”, *Journal of Econometrics*, vol.108 pp.1–24.

Tatođlu, F. Y. (2017). “Panel Zaman Serileri Analizi: Stata Uygulamalı”, Istanbul: Beta

Tatođlu, F.Y. (2018). “Avrupa Ülkelerinde Okun Yasasının Çok Boyutlu Panel Veri Modelleri İle Analizi”, *Yönetim ve Çalışma Dergisi*, vol.1, pp. 43-56.

Westerlund, J. and Basher, S. (2009). “Panel Cointegration and The Monetary Exchange Rate Model. *Economic Modelling*”, vol.26, pp.506-513.

O-65 The Modelling of Earthquake Events Based on Bivariate Extreme Value Theory

Gamze Özel^{1*}

¹*Department of Statistics, Hacettepe University, Turkey, gamzeozl@hacettepe.edu.tr*

Abstract – This study provides estimation for the probability of such extreme events where the mainshock and the largest aftershocks exceed certain thresholds. We analyze the earthquake data from the North Anatolian Fault Zone (NAFZ) in Turkey during 1965–2018 and show that the two approaches provide unifying results.

Keywords – *Extreme value theory, Generalized Pareto Distribution, Generalized Extreme Value Distribution, Quantile Estimation, Risk Measures, Maximum Likelihood Estimation*

1. Introduction

Extreme value theory (EVT) originated in the 20th century. It is a unique statistical discipline providing a way to “predict the unpredictable” (Embrechts et al., 1997). The aim of EVT is to quantify the behaviour of sequences of random variables and stochastic processes at unusually high or low levels. It provides the solid fundamentals needed for the statistical modelling of such events and the computation of extreme risk measures. In many fields of modern science, engineering and insurance, EVT is well established (Embrechts et al. (1999); Reiss and Thomas (1997)). The statistical analysis of extremes is also key to many of the risk management problems related to seismology.

Nature can be seen as an unpredictable phenomenon and for this reason EVT is very useful in modelling and predicting events such as flooding, temperature, rainfall, even pollution. Furthermore, the theory has been extended and is used beyond environmental data. It can be applied for the seismological studies. In seismology the theory of extreme values may be applied to a finite series of observed earthquake magnitudes within a given geographic region covering prescribed periods of the seismicity file. The mathematics behind Gumbel's theory as applied to maximum earthquake magnitudes has been reported by many investigators (Yegulalp and Kuo, 1974; Burton, 1979; Burton, 1981; Kijko, 1984; AI Abbasi and Fahmi, 1985). The development of earthquake hazard assessment in Turkey has a substantial history and it has been produced considerable progress and innovation because Turkey has frequently suffered from major damaging earthquakes since the year 2000 BC. The aim of this study is to evaluate the North Anatolia region of Turkey with regard to seismicity and earthquake hazard using available data on $M_s \geq 5.0$ earthquakes that occurred between 1900 and 2018. Additionally, the study also aimed to determine the parameters associated with earthquake hazards.

Multivariate extreme statistics has been exhibited to be a powerful tool for inference on multidimensional risk factors. Examples of applications can be found in Hann and Ronde (1998), Ledford and Twan (1997), Ponn et al. (2004), among others. The goal of this paper is

to provide a statistical analysis for the joint event of an extreme main shock and extreme aftershocks. In this paper we consider estimating the tail probability of an extreme earthquake event where the mainshock magnitude X and the largest aftershock magnitude Y both exceed certain thresholds. We approach the problems from two directions. On one hand, based on the well-known stochastic rules for aftershocks, we propose a joint parametric model for (X, Y) , estimate the model using (censored) maximum likelihood, and from the model, calculate the desired probabilities. On the other hand, we use non-parametric methods from bivariate extreme value analysis to extrapolate tail probabilities.

2. Material and Methods

2.1 Material

The NAFZ is one of the Earth's major continental strike-slip fault zones that shapes the neotectonic evolution of Turkey and the eastern Mediterranean region. It is characterized by strong seismic activity, posing a high risk to densely populated areas. Despite its importance for the general understanding of large strike-slip fault systems, Eurasian tectonics and earthquake cycle modelling, the deep structure of the NAFZ remains largely enigmatic. In this study, the NAFZ is selected as an area of investigation since it provides an important natural laboratory for understanding earthquake mechanics and fault behavior over multiple earthquake cycles due to its long and extensive historical record of large earthquakes (Ambraseys and Finkel, 1987, 1991, 1995; Ambraseys, 2002). In this paper, the earthquake data which were occurred in the area coordinated $39.00^\circ - 42.00^\circ$ North latitudes and $30.00^\circ - 40.00^\circ$ East longitudes between 1900-2018 years and whose magnitudes equal 5 or higher were used. We use the window algorithm proposed in [15] as follows. For each shock with magnitude $X \geq 5$, we scan the window within distance $L(X)$ and time $T(X)$. If a larger shock exists, we move on to that shock and perform the same scan. If not, then the shock is labelled as the mainshock and all shocks within the specified window are pronounced as its aftershocks

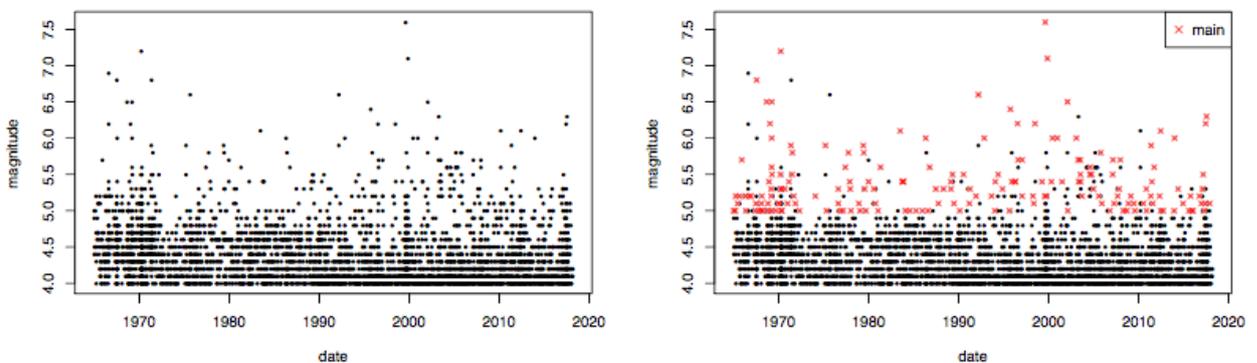


Figure 1: Shocks and labelled mainshocks in the NAFZ during 1965–2018.

2.2 Methods

It is agreed in the literature that the distribution of aftershocks in space, time and magnitude can be characterized by stochastic laws, see Utsu (1971, 1972, 1973) for a summary with detailed empirical studies. In this section, we propose a simple parametric model for the joint magnitudes of the mainshock and the largest aftershock based on these relationships. This derivation is similar to that in Vere-Jones et al. (2006). Throughout the paper, we denote the magnitude of a main shock with X and that of the largest aftershock with Y . We estimate via two approaches the probability of $P(X > s, Y > t)$, for large values of s and t . The first approach uses a parametric model based on a series of well-known stochastic laws that describe the empirical relationships of the aftershocks and the main shock. In the second approach, we apply bivariate extreme value theory to estimate the joint tail. Both methods are applied to the extreme earthquake events in the NAFZ, the region where the 1999 Izmit earthquake occurred.

Let X_A denote the magnitude of an aftershock. Given the mainshock $X = m_0$, we assume that the aftershocks sequence follows a non-homogeneous Poisson process with intensity function (3). Multivariate extreme statistics has been exhibited to be a powerful tool for inference on multidimensional risk factors. Recall that the goal is to estimate the probability: $P(X > t, Y > s)$. To this end, we assume that the joint distribution of (X, Y) is in the max domain of a bivariate extreme distribution introduced in Haan and Resnick (1977). This is a common condition in multivariate tail analysis and includes distributions with various types of copulas.

3. Conclusion

First, we estimate the tail probability defined in (8) for the ten largest earthquakes (mainshocks) in the NAFZ since 1965. As shown in the fifth and sixth column of Table 1, the estimates by two approaches are surprisingly close to each other, which supports the reasonability of the results. We emphasize that the two approaches only share one common assumption, that is, the marginal distribution of X . The distribution of Y and the dependence between the (X, Y) are modelled separately.

Table 1. Tail probability estimation for the ten largest earthquakes in the NAFZ since 1965.

	date	mainshock	largest aftershock	parametric probability	non-parametric probability	location
1	1999-08-17	7.6	5.8	0.00265	0.00257	İzmit
2	1970-03-28	7.2	5.6	0.00618	0.00580	Gediz
3	1999-11-12	7.1	5.2	0.00815	0.00805	Düzce
4	1967-07-22	6.8	5.4	0.01413	0.01356	Mudurnu
5	1992-03-13	6.6	5.9	0.01429	0.01464	Erzincan
6	2002-02-03	6.5	5.8	0.01785	0.01827	Afyon
7	1969-03-28	6.5	4.9	0.02927	0.02745	Alaşehir
8	1968-09-03	6.5	4.6	0.03092	0.03051	Bartın
9	1995-10-01	6.4	5.0	0.03437	0.03296	Dinar
10	2017-07-20	6.3	5.1	0.03938	0.03747	Mugla Province

Next we obtain the level curves of (X, Y) for the tail probability sequence as shown in Figure 2. The points (x,y) on each curve represents such that $P(X > x, Y > y) = p$ for the given probability level.

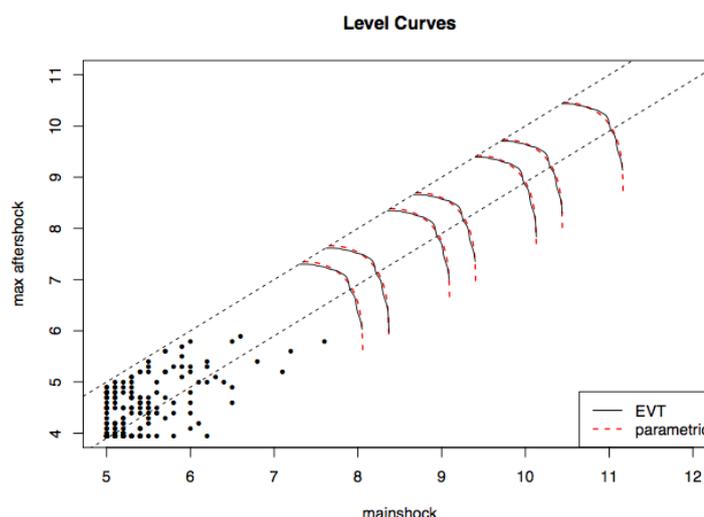


Figure 2: Predicted level curve for (X, Y) based on the parametric model (red dotted) and extreme value analysis (black solid), with existing observations.

References

- Haan, L., and Ronde, J., Sea and wind: multivariate extremes at work. *Extremes*, 1(1):7, 1998. [\[1\]](#)
- Haan, L., Resnick, S.I., Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 40(4):317–337, 1977.
- Ledford, A.W., Tawn, J.A. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2): 475–499, 1997. [\[2\]](#)
- Embrechts P., Klüppelberg C., Mikosch T. (1997): *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag. 645.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1999). *Modelling Extremal Events for Insurance and Finance. Applications of Mathematics*. Springer. 2nd ed.
- Poon, S.-H., Rockinger, M., Tawn, J., Extreme-value dependence in financial markets: diagnostics, models and financial implications. *Review of Financial Studies*, 17:581 – 610, 2004. [\[3\]](#)

Reiss, R. D., Thomas, M. (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhauser Verlag, Basel.

Utsu, T., Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(3):129–195, 1970. ^[1]_[SEP]

Utsu, T., Aftershocks and earthquake statistics (2): Further investigation of aftershocks and other earthquake sequences based on a new classification of earthquake sequences. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(4):197–266, 1971. ^[1]_[SEP]

Utsu, T., Aftershocks and earthquake statistics (3): Analyses of the distribution of earthquakes in magnitude, time and space with special consideration to clustering characteristics of earthquake occurrence (1). *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(5): 379–441, 1972. ^[1]_[SEP]

Vere-Jones, D., Murakami, J., Christophersen. A., A further note on Bath’s law. In *The 4th International Workshop on Statistical Seismology (Statsei4)*, pages 611–0011, 2006. ^[1]_[SEP]

O-66 Healthcare Tourism Demand and an Empirical Analysis for Istanbul

Bilgesu Bayir¹, Erkan Isikli^{2*}

¹Industrial Engineering Department, Istanbul Technical University, Turkey,
bayirb@itu.edu.tr

²Industrial Engineering Department, Istanbul Technical University, Turkey, isiklie@itu.edu.tr

Abstract – Healthcare and long waiting lists in developed countries. More people are traveling nowadays to developing countries to receive quality healthcare at lower costs. All the decision making and planning in any industry eventually rely on reliable forecasts. Thus, building a model that accurately reflects the nature and attributes of healthcare tourism demand has a significant impact on the host country’s economy. This study attempts to incorporate various time series techniques to provide accurate forecasts for international demand for healthcare tourism to a particular destination, namely Istanbul. Since there is a paucity of research specifically focusing on the empirical investigation of healthcare tourism, studies on the modelling and forecasting of tourism demand since 2010, in general, are perused and four main methods (Holt-Winters, ETS, STL, and SARIMA) are fitted based on the results of the literature review. In addition, various combinations of these models are formed to improve forecast accuracy. SARIMA is found to perform usually better among the four main models and the combination approach that utilizes constrained linear regression is found to have the best overall accuracy in terms of Maximum Absolute Percentage Error and Root Mean Squared Error.

Keywords – Healthcare tourism demand, service systems, time series forecasting, combination forecasts

1. Introduction

The Medical Tourism Association defines healthcare tourism as people going on a trip to a different country with the purpose of getting medical, surgical, or dental care. Even though it is mainly focused on health improvement and wellness, healthcare tourism is not only about medical procedures. After the completion of medical procedures, healthcare tourists and their families look for friendly natives, tourist amenities, affordable cost of living, authentic experiences, and some souvenirs to bring home (Bookman and Bookman, 2007). Thus, tourism demand of a country is fully or partially affected by its healthcare tourism demand. The main purpose of healthcare tourism is to improve one’s health and wellness through a wide range of healthcare activities rather than leisure or pleasure. Today, considerably many people, especially those from developed countries, are traveling for healthcare. Horowitz et al. (2007) claim that even though people have always gone abroad to sanitariums or spas for health improvement, making trips to other countries for medical treatments or surgeries was not common until 1980’s (Genc, 2012). Bookman and Bookman (2007) state that as early as 1989, an OECD report indicated the competitive edge that this sector offered to developing

countries, regarding the skills in medicine and capital availability and their great amount of labor supply. Woodward et al. (2002) state that, according to the World Health Organization, healthcare tourism is an increasing trend which has great economic significance (Bookman and Bookman, 2007). A growing number of people are nowadays going abroad for more affordable healthcare services that are often offered together with complementary tourism services such as accommodation in a hotel and city visits (Crooks et al., 2011; Hudson and Li, 2017) and it has become more and more widespread for the residents of industrialized countries to travel distant developing countries for benefiting from their more affordable and good quality healthcare services and facilities (Hudson and Li, 2017). The tourism industry in Turkey is one of the most revenue generating sectors. Healthcare tourism has been growing as a special part of it with recent initiatives of the government, lately advancements in healthcare services and facilities, and private investments. Therefore, demand for accurate healthcare tourism forecasts is of particular importance due to the significant contribution of this sector to the national economy. The primary objective of this study is to examine whether combined models of time series can improve the forecasts of international demand for healthcare tourism, in this case to Istanbul, in terms of accuracy. The research also provides a comparison of the performances of time series forecasting techniques and can be easily extended to cover other destinations as well.

2. Background

Stemming from the globalization in healthcare and growth of tourism, healthcare tourism is today among the fastest-growing industries (Heung et al., 2011; Fernando and Khei, 2015). Motivation to do research in this area and attempts to understand its nature have been significantly increasing and the topic is becoming more popular every day. According to SRI International (2010), there are three essential factors causing the healthcare tourism demand to increase:

failure of traditional medical systems, making governments, healthcare providers, and consumers look for more cost-effective and prevention-centered rather than solution-to-existing-problem centered options,

aging population of the world,

accelerating globalization, making consumers use the Internet to get informed about alternative health paradigms and the strong effect of famous advocates of wellness (Hudson and Li, 2017).

Healthcare tourism commonly includes travelling for elective health interventions such as replacement of hip/knee, spinal surgeries, and dental procedures, whereas in the category of disease treatment, there are medical interventions ranging from health screening tests and health check-ups to open heart surgeries and treatments of cancer (Crooks et al. (2011); Rezvani and DeMicco, 2017). Bookman and Bookman (2007) classify healthcare tourism

services into three categories: Invasive, diagnostic, and lifestyle. Invasive procedures are conducted by specialized people for patients who have non-infectious diseases. Dental procedures keep on being the most popular kind of invasive procedure. Its popularity majorly results from the fast treatment and even faster recovery, which allows patients spend time and energy on their holidays. Another reason for dental procedures' popularity is the fact that insurance policies rarely cover the expenses. For a similar reason, cosmetic surgeries are also a popular kind of healthcare tourism services (Bookman and Bookman, 2007).

In addition to providing patients with the opportunity to receive good quality healthcare at lower prices, healthcare tourism creates economic opportunities for the destination country as well. Bookman and Bookman (2007) state that healthcare tourism is firstly associated with economic development. The sales of healthcare tourism services and goods are an example of international trade, which contributes to the economy of a country in several ways. According to Lordan (2013), due to both healthcare and tourism expenditures made by healthcare tourists, healthcare tourism results in increased revenues for the country of destination. Moreover, healthcare tourists are most likely to bring a companion along with them, creating even more tourism revenues (Lordan, 2013) and they not only buy healthcare services, but also tourism goods and services such as transportation, accommodation, food and beverage, souvenirs, entertainment, and sightseeing activities. According to the World Bank, outside the hotel, tourists spend up to nearly two times of the amount they spend in the hotel, and tourism services are thus very important (Bookman and Bookman, 2007).

Peters and Sauer (2011) assert that healthcare tourism provides advantages for developing countries by creating an opportunity for local entrepreneurship, increasing the activities of tourism, improving the infrastructure of healthcare, creating more career opportunities, and increasing the healthcare service exports of the country (Fernando and Khei, 2015). With the help of increased revenues gained through healthcare tourism, developing countries can make investments to create further revenues and employment. These extra revenues generated can then be used to develop healthcare facilities and to invest in more advanced technologies (Lordan, 2013). As a result of the economic development that healthcare tourism brings about, Dollar and Kray (2004) claim that healthcare tourism decreases poverty in developing countries (Bookman and Bookman, 2007).

3. Methodology

Planning is considered one of the most important functions for policy making at all levels. While making any decision, managers take some type of forecasts into consideration. If managers are to deal with abrupt changes in the level of demand, strikes, seasonal variations and economic fluctuations, reliable estimations of demand and trends become requirements rather than luxury elements (Chambers et al., 1971). Sasser (1976) states that service industries are different from manufacturing industries since they are immediate, which makes it difficult to match supply and demand. Tourism is also traded as a service, which has unique features in the sense that consumption occurs where the product is supplied and thus consumers, namely tourists, have to make a trip to consume the product (Divisekera, 2013).

According to Song and Witt (2000), demand is the primary factor determining the business profitability in tourism. Thus, it is obvious that accurately forecasting demand for tourism is necessary for efficiently planning the tourism business, especially considering the perishable nature of the tourism product. Effective planning of all types of tourism services is largely based on accurate demand estimates (Law and Pine, 2004). For the decision making activities of management, forecasting provides with an essential input. The more the management aims to diminish its dependence on chance and deals with its environment in a scientific way, the more the need for forecasting increases (Makridakis et al., 1998). By making use of the forecasting results, business practitioners and policy makers do not only successfully manage operations and regulations, but also shape their long-term strategies accordingly. Formal scientific techniques that clearly reflect the relationship between travel demand and underlying factors will certainly help tourism decision makers to better understand the demand for travel to a destination. In order to identify the most frequently used techniques for modeling and forecasting tourism demand, studies conducted since 2010 were investigated within the scope of this study. The related literature is mainly based on econometric models, artificial intelligence (AI)-based methods, and time series techniques. None of them have consistently outperformed the others in terms of forecasting accuracy and highest accuracy was observed to change depending on data characteristics.

3.1 Methods for Modeling and Forecasting Tourism Demand

Tourism demand is affected by many factors which are related to the country of destination as well as the country of origin in the tourism activity. Song and Witt (2000) claim that income level and tastes of potential customers, prices of tourism products in the country of destination and alternative destinations, advertisement efforts made, and other geographic, social, political, and cultural factors all affect tourism demand. Song and Witt (2000) state that factors affecting tourism demand should be correctly determined and the model of tourism demand should be properly established for an accurate forecast. Thus, there is a vast literature on the empirical analysis of tourism demand that aims to identify the factors affecting tourism demand and forecasting the future tourist flow (Divisekera, 2013). Causal tourism demand models were reported to generally employ ordinary least squares estimation until 1990's (Song and Witt, 2000). According to Goh and Law (2011), an econometric model of tourism demand aims at explaining and forecasting the future behavior of tourism demand through quantitative relationships between several factors and the amount of demand. During the 1990's, tourism demand studies that employed econometric models were focused more on models of a single equation. However, in order to overcome the drawbacks of these early studies, more complex econometric models have been developed as sets of demand equations (Goh and Law, 2011). Autoregressive Distributed Lag (ADL) Models and Panel Data Models have been especially popular in the last decade for modeling and forecasting tourism demand. On the other hand, the number of AI-based tourism demand models is not quite sufficient (Goh and Law, 2011). However, throughout the last decade, the use of these methods in conjunction with econometric or time series methods has become relatively widespread in the related literature. Even though AI-based methods have conventionally used logic

programming and rule-based systems, their actual use is concentrated on less precise heuristics techniques such as artificial neural networks, fuzzy logic, support vector machines, and genetic algorithms (Song and Li, 2008). Neural Networks have, indeed, been by far the most widely employed AI-based method, followed by Support Vector Regression and Fuzzy Time Series in modeling and forecasting tourism demand. According to Divisekera (2013), a vast majority of the existing studies on the empirical analysis of tourism demand are based on the classical techniques of time series, its variations, and other associated statistical methods. All time series methods are endogenous techniques, implying that they only consider the historical data patterns which are identified and projected into the future to obtain forecasts (Mentzer and Moon, 2005). According to Chopra and Meindl (2016), time series techniques, which assume that historical data are good indicators of future demand, are the simplest methods to implement and a good starting point for forecasting demand. Chopra and Meindl (2016) add that time series forecasting should be adopted when the basic demand pattern does not vary significantly from one year to the next. Box-Jenkins methods such as ARIMA and SARIMA have been most commonly used time series techniques in the related literature. A tendency to employ more sophisticated techniques such as Spectral Analysis and State Space Models has also been observed recently.

3.2 Data and Variables

The data set used in this study was obtained from a private dental clinic that has been serving healthcare tourists for several years. The dental clinic is a part of a large group of clinics that have more than ten branches in several locations in Turkey, Europe, Middle East, and Asia. Healthcare tourists served in the clinic consist of foreign patients coming from other countries as well as Turkish expats who live in other countries and come to Turkey to receive healthcare. The data set consists of monthly foreign patient arrivals to the clinic from January 2013 to March 2019 and its behavior is illustrated in Figure 2 below. The series has an upward trend (indicating a long term increase in foreign patient arrivals) and a seasonal pattern (the arrivals show a similar pattern at the same time of the year). The foreign patient arrivals peak in June, August and December every year. The data set was split into an estimation set (first 60 observations) and a validation set (last 15 observations) for further analysis.

4. Application Results: Modeling the Demand for Healthcare Tourism

Based on the literature review mentioned above, several time series techniques that are appropriate for the characteristics of the data were employed. In this section, a brief explanation of each technique is provided followed by their prediction performances both in the estimation and validation sets.

Time Series Decomposition (STL)

The pattern of a time series can be decomposed into sub patterns in many cases, in such a way to define different components of the series one by one. This type of decomposition often

helps to enhance the accuracy in forecasting through better understanding of the behavior of the series (Makridakis et al., 1998). Decomposition of a time series, therefore, is a useful tool in terms of analyzing the series and accordingly choosing the methods to model it. Here, we employ STL, a seasonal-trend decomposition procedure based on local regression smoothing (LOESS) that was developed by Cleveland et al. (1990). This relatively recent method provides a simple design that is flexible in specifying the amounts of variation in the trend and seasonal components (Cleveland et al., 1990). With an ability to decompose series with missing values, STL provides robust trend and seasonal components that are not distorted by transient, aberrant behavior in the data (Cleveland et al., 1990). Figure 2 below shows the STL decomposition plot created in R for the healthcare tourist arrivals to the dental clinic in the estimation period. Visual inspection of the plot signals for a linear trend and additive seasonality. Thus, we employed STL with additive seasonality. The plot also shows an upward movement in demand in the months March/April, August, and December/January. The seasonal effect appears to be smallest in October. Besides, October data has a very large variation compared to any other month's. The residuals are mean and variance stationary until 2016, but after then they are fluctuating.

Holt-Winters' with Additive Seasonality (HWS)

Exponential smoothing methods generate forecasts that are basically the weighted averages of past observations such that the weights decay exponentially, therefore more recent observations are given larger weight in determining future values than observations in the more distant past (Hyndman and Athanasopoulos, 2008). All the exponential smoothing methods require the estimation of some parameters that take values between 0 and 1 (Makridakis et al., 1998). In general, exponential smoothing methods can be classified as follows:

Simple exponential smoothing: Including only one smoothing parameter (α), this model is appropriate for use when data is stationary. The parameter α controls the weight given to the recent data and can be called the smoothing parameter for the level (L).

Holt's linear method: In addition to α in simple exponential smoothing, Holt's method includes the parameter β that serves as the smoothing parameter for trend. Also known as double exponential smoothing, this method is appropriate to use when the time series exhibits trend, but not seasonality.

Holt-Winters' method: As an extension to Holt's method, a third parameter, γ , is included in this method to smooth seasonality. Thus, it is appropriate to use when the series exhibits both trend and also seasonality.

Since the original series of healthcare tourist arrivals show an upward trend and an additive seasonal pattern, we employed Holt-Winter's method in this study.

Figure 2: Decomposition plot for the period January 2013 to December 2018.

State Space Models for Exponential Smoothing (ETS)

First introduced in late 1970's, State Space Models provide great flexibility in the specification of parameters for level, trend, and seasonal components mainly using two equations – observation equation and state equation (Hyndman et al., 2008). Due to limited space, we are not delving too much into detail explaining the state space models; however, interested users are referred to Aoki (1990) for further reading. State space models for exponential smoothing differ from previously mentioned smoothing methods by introducing an error component in addition to trend and seasonal components. This type of models is generally denoted by ETS(x,y,z) such that x indicates whether the errors are additive (A) or multiplicative (M), y indicates whether the trend is additive (A) or multiplicative (M) or damped, z denotes whether the seasonality is additive (A) or multiplicative (M).

Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA)

As mentioned in the literature review, along with the smoothing methods, ARIMA is one of the most widely used and effective time series modeling approach that aims to describe the behavior of a time series based on autocorrelations (Hyndman and Athanasopoulos, 2018). This type of models assumes that future values of a time series are generated from a linear function of its past observations and some white noise errors. Defined by three terms, they are generally denoted by ARIMA(p,d,q), where d denotes how many times the series should be differenced to make it mean stationary, p denotes the number of autoregressive components, and q denotes the number of moving average components. An ARIMA model can be extended to incorporate seasonal trends along with primary stochastic trend. Known as Seasonal ARIMA (SARIMA), this type of models is denoted by ARIMA(p,d,q) [(P,D,Q)]_s, where D denotes the degree of seasonal differencing, P denotes the number of seasonal autoregressive components, Q denotes the number of seasonal moving average components, and s denotes the seasonal cycle.

Identifying the value of d (and D) is the most crucial first step while fitting an ARIMA-type model. Figure 2 above shows that this series is not mean stationary and exhibits seasonality. Thus, we first take the seasonal difference of the series to decide on the number of parameters to include in the model. Depending on the visual inspection of the related plots, once-seasonally-differenced data does not appear to be mean stationary either. We, then, take the first difference of this new series, which now becomes mean stationary. The significant spikes at lag 3 in the ACF plot and lags 3 and 5 in the PACF plot suggested some non-seasonal autoregressive and moving average components. After trying some combinations, we decided to move on with ARIMA(3,1,1) × [(0,1,0)]₁₂ since it returned the lowest AIC Corrected (AICc) that was even smaller than the one obtained from the auto.arima procedure in R.

Combining Different Time Series Techniques

Combining different forecasting methods is a useful way to improve the overall accuracy. Oh and Morzuch (2005) show that combined models perform better than the worst performing

models and sometimes even outperform the best performing model. Therefore, combining different methods can be useful in decreasing the overall forecast error (Song and Li, 2008). In this study, we combine forecasting models in four different ways: (1) taking the arithmetic mean of the predicted values, (2) taking the geometric mean of the predicted values, (3) fitting a constrained regression model with the original series as a dependent variable and the predicted values as independent variables (Granger and Ramanathan, 1984), (4) taking the weighted average of the predicted values using the procedure proposed by Newbold and Granger (1974). The latter two approaches are not as straightforward as the former two. The third approach incorporates the regression coefficients estimated using the estimation data set to combine the forecasts in the validation data. Here, the combining weights are constrained to sum to one. In contrast, the weights obtained in the fourth approach are estimated using the covariances between the 15-step ahead forecast errors of the four main models. Then, these weights are used to combine the fitted values in the estimation data set.

Evaluation of the Model Performances

To choose the best model among these eight alternatives, we use standard forecasting accuracy measures such as Mean Absolute Percentage Error (MAPE), Maximum Absolute Percentage Error (Max APE), Root Mean Squared Error (RMSE) that are all reported in Table 1 below. Based on Frechtling (2001), none of the models can be said to perform highly accurately both in the estimation and validation sets; however, SARIMA is consistently good and, in general, all of the models perform reasonably well. Among the four main models, SARIMA specifically performs best. The combination approach through constrained linear regression (REGOPT) outperforms each of the four main models as well as the combination models in the validation set. Fitted values of each method are illustrated in Figures 5 and 6. In addition, Figure 7 shows whether the distribution of forecast errors can be assumed normal. Even though it produces normally distributed errors, ETS(M,A,N) seems to fail at modeling the international demand for this clinic when its behavior in the validation data is analyzed. The combination approach with optimal weights (OPTCOMB) performs as well as REGCOMB when only the validation period is considered.

Table 1: Evaluation of the models

Model Accuracy Measure

MAPE

(Estimation) MAPE

(Validation) Max APE

(Estimation) Max APE

(Validation) RMSE

(Estimation) RMSE

(Validation)

HWS	70.36%	13.33%	851.88%	47.40%	180.69	34.43
ETS	25.43%	33.78%	119.30%	86.32%	76.99	83.15
SARIMA	19.10%	11.03%	140.05%	41.48%	78.22	30.11
STL	77.55%	15.17%	1,276.47%	23.32%	224.21	37.88
REGCOMB	39.25%	7.82%	379.33%	19.39%	15.46	21.96
GEOCOMB	28.64%	10.16%	138.18%	29.21%	16.48	29.52
OPTCOMB	313.60%	8.50%	14,459.83%	21.24%	15.95	23.78
AVECOMB	27.09%	10.97%	164.74%	30.76%	16.41	32.11

5. Conclusion

Healthcare tourism industry has been steadily growing in developing countries with better customer relations management practices and more public and private investments. Besides, extreme costs of healthcare and long waiting lists also cause people from developed countries travel to developing countries to receive quality healthcare at lower costs. Healthcare tourism contributes to the economy of the host country since healthcare tourists intend to spend time and money in these countries as a part of their vacation. Like any other service industry, successfully matching supply with demand is a challenging issue in healthcare tourism. Therefore, accurate analyses of the demand for healthcare tourism is of great importance for

all the parties involved in the healthcare tourism supply chain such as healthcare tourists, healthcare facilities, healthcare tourism agencies, airline companies, hotels, and restaurants. Thus, the ability to accurately predict international demand for healthcare tourism will benefit tourism managers and government officials in better planning and decision-making. In this study, first of all, related research in the last decade is investigated and classified. Since the literature that specifically focuses on healthcare tourism is limited, studies on the modeling and forecasting of tourism demand are primarily investigated. Additionally, several time series approaches such as time series decomposition, smoothing, and ARIMA are adopted using a monthly time series data obtained from a dental clinic that has been involved in healthcare tourism for several years. Model performances are evaluated, compared with each other, and improved through a couple of forecast combination methods. In line with previous studies, none of the methods is found to significantly and consistently outperform the others; however, SARIMA and the combination approach using constrained linear regression is found to perform better than the other alternatives. As a possible avenue for future research, social media data can be integrated into the demand model to improve forecasts. Future studies should also consider AI-based methods such as Neural Network Autoregression (NNAR) and Multilayer Perceptrons (MLP) or recently proposed forecasting models such as the Prophet.

Figure 5: Fitted values of the four main methods for the estimation and validation periods

Figure 6: Fitted values of the combination methods for the estimation and validation periods

Figure 7: Normal probability plots for the residuals of the four main methods (on the right) and the residuals of the combination methods (on the left) for the validation period

Acknowledgment

This work was supported by the Research Fund of Istanbul Technical University. Project Number: 42051.

References

- Aoki, M. (1990). *State Space Modeling of Time Series*, 2nd ed. Springer-Verlag, Berlin.
- Bookman, M.Z., Bookman, K.R. (2007). *Medical Tourism in Developing Countries*. New York, NY: Palgrave Macmillan.
- Chambers, J.C., Satinder, K.M., Smith, D.D. (1971). *How to Choose the Right Forecasting Technique*. Retrieved January 11, 2019, from <https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique>.

Chopra, S., Meindl, P. (2016). *Supply Chain Management: Strategy, Planning, and Operation*. Boston, MA: Pearson.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.J. (1990). “STL: A Seasonal-Trend Decomposition Procedure Based on Loess”, *Journal of Official Statistics*, vol.6, no.1, pp.3-33.

Divisekera, S. (2013). “Tourism Demand Models: Concepts and Theories”. In C. A. Tisdell (Ed.), *Handbook of Tourism Economics: Analysis, New Applications and Case Studies* (pp.33-66).

Fernando, Y., Khei, L.H. (2015). “Dive with the Sharks: A Content Analysis of the Medical Tourism Supply Chain”. In M.M.J. Cooper, K. Vafadari, M. Hieda (Eds.), *Current Issues and Emerging Trends in Medical Tourism* (pp.31-43).

Frechtling, D.C. (2001). *Forecasting Tourism Demand: Methods and Strategies*. Oxford: Butterworth-Heinemann.

Genc, R. (2012). “Physical, Psychological, and Social Aspects of QOL Medical Tourism”. In M. Uysal, R. Perdue, & J. Sirgy (Eds.), *Handbook of Tourism and Quality-of-Life Research: Enhancing the Lives of Tourists and Residents of Host Communities* (pp.193-207).

Goh, C., Law, R. (2011). “The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature”, *Journal of Travel & Tourism Marketing*, vol.28, pp.296-317.

Granger, C.W.J., Ramanathan, R. (1984). “Improved Methods of Forecasting”, *Journal of Forecasting*, vol.3, pp.197-204.

Hudson, S., Li, X.R. (2017). “Domestic Medical Tourism: A Neglected Dimension of Medical Tourism Research”. In F.J. DeMicco (Ed.), *Medical Tourism and Wellness: Hospitality Bridging Healthcare (H2H)* (pp.159-181).

Hyndman, R.J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Retrieved from <https://otexts.com/fpp2>.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer-Verlag.

Law, R., Pine, R. (2004). “Tourism Demand Forecasting for the Tourism Industry: A Neural Network Approach.” In *Neural Networks in Business Forecasting* (pp.121-141). IGI Global.

Lordan, G. (2013). “Travelling for Treatment: The Emergence of a Medical Tourist Industry”. In C.A. Tisdell (Ed.), *Handbook of Tourism Economics: Analysis, New Applications and Case Studies* (pp.281-297).

- Makridakis, S., Wheelwright, S.C., Hyndmand, R.J. (1998). *Forecasting Methods and Application*. New York, NY: Wiley.
- Mentzer, J.T., Moon, M.A. (2005). *Sales Forecasting Management: A Demand Management Approach*. London: Sage Publications.
- Rezvani, E., DeMicco, F.J. (2017). “Evaluating the Performance of the Hotels in the Vicinity of the Selected World’s Prominent Hospitals: An Empirical Research Project.” In F.J. DeMicco (Ed.), *Medical Tourism and Wellness: Hospitality Bridging Healthcare (H2H)* (pp.183-223).
- Sasser, W.E. (1976). *Match Supply and Demand in Service Industries*. Retrieved April 11, 2019, from <https://hbr.org/1976/11/match-supply-and-demand-in-service-industries>
- Song, H., Li, G. (2008). “Tourism Demand Modeling and Forecasting: A Review of Recent Research”, *Tourism Management*, vol.29, pp.203-220.
- Song, H., Witt, S.F. (2000). *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches*. Oxford: Pergamon.
- Crooks, V.A., Turner, L., Snyder, J., Johnston, R., Kingsbury, P. (2011). “Promoting Medical Tourism to India: Messages, Images, and the Marketing of International Patient Travel”, *Social Science & Medicine*, vol.72, pp.726-732.
- Dollar, David and Aart Kray , “Trade, Growth, and Poverty”. *The Economic Journal*, vol. 114 (493), 2004, pp. F22-F49.
- Heung, V.C.S., Kucukusta, D., Song, H. (2011). “Medical Tourism Development in Hong Kong: An Assessment of the Barriers”, *Tourism Management*, vol.32, pp.995-1005.
- Horowitz, M.D., Rosensweig, J.A., Jones, C.A. (2007). “Medical Tourism: Globalization of the Healthcare Marketplace”, *Medscape General Medicine*, vol.9, pp.33-39.
- Oh, C.O., Morzuch, B. J. (2005). “Evaluating Time-Series Models to Forecast the Demand for Tourism in Singapore: Comparing Within-Sample and Post-Sample Results”, *Journal of Travel Research*, vol.43, pp.404-413.
- Peters, C.R., Sauer, K.M. (2011). “A Survey of Medical Tourism Service Providers”, *Journal of Marketing Development and Competitiveness*, vol.5, pp.117-126.
- SRI International. (2010). *Spas and the Global Wellness Market: Synergies and Opportunities*. Menlo Park, CA: SRI International.
- Woodward, D.T., Drager, N., Beaglehole, R., Lipson, D.J. (2002). “Globalization, Global Public Goods, and Health”. In N. Drager (Ed.), *Trade in Health Services: Global, Regional, and Country Perspectives* (pp.3-11).

O-67 Some Tests and Comparisons for Homogeneity of Variances

Nilgün Nursu ÖZTÜRK^{1*}, Hamza GAMGAM² and Bülent ALTUNKAYNAK³

¹Department of Statistics, Gazi University, Turkey, mnilgunnursu@gmail.com

²Department of Statistics, Gazi University, Turkey, gamgam@gazi.edu.tr

³Department of Statistics, Gazi University, Turkey, bulenta@gazi.edu.tr

Abstract –The equality of variances is one of the basic assumptions of classical variance analysis. There are many test statistics known in the literature to determine whether this assumption is met or not. The aim of this study is to introduce the Bartlett (B), Hartley (H), Levene (L), Fligner-Killeen (FK) and Bootstrap (FRMD) tests for testing the homogeneity hypothesis of variances and compare these tests. Using Monte Carlo simulation method, these tests were compared in terms of type I error rate and power under various scenarios. With these comparisons, it was aimed to determine the test suitable for testing the homogeneity hypothesis of variances in both normal distribution and some nonnormal distributions.

Keywords – *Simulation, analysis of variance, power, bootstrap*

1. Introduction

Testing of equality of variances is widely used in many areas such as ecology, biology, agricultural production systems, the development of medical and educational methods. The researchers are interested in the equality of variances in the populations of interest for many reasons. For example, the classical F test in analysis of variance is based on the assumption of homogeneity of variances. In the literature, many tests have been developed to test the homogeneity of variances. First, Bartlett (1937) proposed the likelihood ratio test under the assumption of normality. Subsequently, alternative tests were proposed by Hartley (1950), Cochran (1951), Box (1953), Levene (1960), Brown and Forsythe (1974), Lim and Loh (1996), Conever et al.(1981). Later, Jayalath et al. (2017) proposed a test based on mean absolute deviations to test the equality of variances. This test is based on the ratio test proposed by Higgins (2004) for absolute deviations.

This study is organized as follows. In section 2, the tests developed by Bartlett (1937), Levene (1960), Fligner and Killeen (1976), Hartley (1950) and Jayalath et al. (2017) for homogeneity of variances are introduced. In section 3, a simulation study is given to compare these tests in terms of type I error rate and power.

2. Tests for Homogeneity of Variances

Let $\{X_{ij}, i = 1, 2, \dots, k, j = 1, 2, \dots, n_i\}$ represent k independent random samples of size n_i drawn from k distributions each with variance σ_i^2 . The hypotheses for testing the homogeneity of variances are given as follows.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \quad (1)$$

$$H_1 : \sigma_i^2 \neq \sigma_{i'}^2 \text{ for at least one distinct pair of } (i, i'), i, i' = 1, \dots, k \quad (2)$$

In this section, the tests proposed by Bartlett (1937), Levene (1960), Fligner and Killeen (1976), Hartley (1950) and Jayalath et al. (2017) are briefly introduced.

2.1 Bartlett Test

The Bartlett test statistic, shown as B , is defined as follows.

$$B = - \frac{\sum_{i=1}^k (n_i - 1) \ln(s_i^2 / s_p^2)}{1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \left(\frac{1}{n_i - 1} \right) - \frac{1}{\sum_{i=1}^k (n_i - 1)} \right]} \quad (3)$$

where $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ is the sample variance of i th group, $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ is the sample mean of i th group and the pooled variance $s_p^2 = \sum_{i=1}^k (n_i - 1) s_i^2 / \sum_{i=1}^k (n_i - 1)$ is the weighted average of the sample variances. Under H_0 , the B statistic has approximately chi-square distribution with $k - 1$ degree of freedom. H_0 hypothesis is rejected at significance level α if $B > \chi_{k-1, \alpha}^2$ where $\chi_{k-1, \alpha}^2$ is the upper p th percentile of the chi-square distribution with $k - 1$ degrees of freedom.

2.2 Levene Test

The important feature of the Levene test is that it is a robust test when the assumption of normality is not provided. This test is derived from the common F -ratio in one way ANOVA with the observation x_{ij} being replaced by its absolute deviation from its group mean, $z_{ij} = |x_{ij} - \bar{x}_i|$, where \bar{x}_i is the sample mean of the i th group. The Levene test statistic, shown as L , is defined as follows.

$$L = \frac{\sum_{i=1}^k n_i (\bar{z}_i - \bar{z}_{..})^2 / (k - 1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2 / (n - k)} \quad (4)$$

where $\bar{z}_i = \sum_{j=1}^{n_i} z_{ij} / n_i$ is the i th group mean, $\bar{z}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} z_{ij} / n$ is the overall mean of z_{ij} and $n = \sum_{i=1}^k n_i$ is the total sample size.

Under H_0 , the L statistic has approximately F distribution with $k-1$ and $n-k$ degree of freedoms. H_0 hypothesis is rejected at significance level α if $L > F_{k-1, n-k, \alpha}$ where $F_{k-1, n-k, \alpha}$ is the upper α th percentile of the F distribution with $k-1$ and $n-k$ degree of freedoms.

2.3 Fligner-Killeen Test

When the normal distribution assumption is not provided, the use of this test is recommended for the homogeneity of variances. This test jointly ranks the absolute values of $|x_{ij} - m_i|$ and assigns increasing scores $a_{n,i} = \Phi^{-1}\left(\frac{1+(i/(n+1))}{2}\right)$ based on the ranks of all observation where m_i is the median of the i th sample and $\Phi(\cdot)$ is the cumulative distribution function for normal distribution. The Fligner-Killeen statistic, shown as FK , is defined as follows.

$$FK = \frac{\sum_{i=1}^k (\bar{A}_i - \bar{a})^2}{V^2} \quad (5)$$

where \bar{A}_i is the mean score for the i th sample, \bar{a} is the overall mean score of all $a_{n,i}$, and V^2 is the sample variances of all scores. That is $\bar{A}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} a_{n,m_j}$, $n = \sum_{i=1}^k n_i$, $\bar{a} = \frac{1}{n} \sum_{i=1}^n a_{n,i}$ and $V^2 = \frac{1}{n-1} \sum_{i=1}^n (a_{n,i} - \bar{a})^2$ where a_{n,m_j} is the increasing rank score for the i th observation in the j th sample.

Under H_0 , the FK statistic has approximately chi-square distribution with $k-1$ degree of freedom. H_0 hypothesis is rejected at significance level α if $FK > \chi_{k-1, \alpha}^2$ where $\chi_{k-1, \alpha}^2$ is the upper p th percentile of the chi-square distribution with $k-1$ degrees of freedom.

2.4 Hartley Test

Under normal distribution, this test is widely used for the equality of variances when the sample sizes are equal. The Hartley statistic, shown as H , is defined as follows.

$$H = \frac{S_{Max}^2}{S_{Min}^2} \quad (6)$$

where $s_{\max}^2 = \max\{s_1^2, s_2^2, \dots, s_k^2\}$ and $s_{\min}^2 = \min\{s_1^2, s_2^2, \dots, s_k^2\}$. Critical values of this statistic can be obtained from Hartley's table presented in Hartley(1950). H_0 hypothesis is rejected at significance level α if $H > H_{k,n,\alpha}$ where $H_{k,n,\alpha}$ is the critical value for the Hartley's test.

2.5 FRMD Test

Bootstrap method is a resampling method that uses the sample in hand as a surrogate population. One of the aims of this method is to approximate the sampling distribution of the statistic of interest from repeatedly sample(with replacement) large number of times from the sample in hand. While the number of groups is 2, let us define the mean absolute deviation (DEV) statistics around the median for $X = (x_1, x_2, \dots, x_{n_1})$ and $Y = (y_1, y_2, \dots, y_{n_2})$ random variables as follows.

$$Dev(X) = mean\{|X_j - median(X_1, X_2, \dots, X_{n_1})|, j = 1, 2, \dots, n_1\} \quad (7)$$

$$Dev(Y) = mean\{|Y_j - median(Y_1, Y_2, \dots, Y_{n_2})|, j = 1, 2, \dots, n_2\} \quad (8)$$

where $median(x_1, x_2, \dots, x_{n_1})$ is the sample median for the observations x_j and $median(y_1, y_2, \dots, y_{n_2})$ is the sample median for the observation y_j .

The test statistic based on these statistics for testing the equality of two variances is defined as follows.

$$F_{RMD} = \frac{\max\{Dev(X), Dev(Y)\}}{\min\{Dev(X), Dev(Y)\}} \quad (9)$$

The sampling distribution of this test statistic is quite difficult mathematically. Therefore, bootstrap method is proposed for the distribution of this statistic. When $k = 2$, the steps of this proposed method are as follows (Jayalath et al., 2017).

1. Let F_{RMD}^0 be the observation value of the statistic F_{RMD} for the original samples.
2. Since the location parameter of the distribution are known and possibly unequal, we centralize the two samples so that they have equal center location subtracting their corresponding medians, and define the ‘pseudo-samples’ as $\tilde{x}_j = x_j - median(x_1, x_2, \dots, x_{n_1}), j = 1, 2, \dots, x_{n_1}$ and $\tilde{y}_j = y_j - median(y_1, y_2, \dots, y_{n_2}), j = 1, 2, \dots, n_2$.
3. Randomly select two samples of size $[n_1/2]$ with replacement, one from each of \tilde{x} and \tilde{y} , and $n_1 - [n_1/2]$ observations from the combined sample $\{\tilde{x}, \tilde{y}\}$, to get the total of n_1 observations, where $[a]$ is a integer part of a. Combined these three samples to make the first bootstrap sample (say BN_1) for the numerator.
4. Randomly select two samples of size $[n_2/2]$ with replacement, one from each of \tilde{x} and \tilde{y} , and $n_2 - [n_2/2]$ observations from the combined sample $\{\tilde{x}, \tilde{y}\}$, to get the total of n_2 observations, where $[a]$ is a integer part of a. Combined these three samples to make the second bootstrap sample (say BD_1) for the denominator.
5. Calculate the sample DEV’s for both BN_1 and BD_1 separately, and obtain the ratio,

$$F_1^* = \frac{\max\{Dev(BN_1), Dev(BD_1)\}}{\min\{Dev(BN_1), Dev(BD_1)\}} \quad (10)$$

as the first bootstrap F_{RMD} statistic, F_1^* .

6. Repeat steps 2-4 B times and obtain the corresponding bootstrap F_{RMD} statistics $F_1^*, F_2^*, \dots, F_B^*$. 7. The bootstrap sampling distribution $F^* = \{F_1^*, F_2^*, \dots, F_B^*\}$ mimics the null distribution of the test statistic. we reject the hypothesis H_0 at α level if $F_{RMD}^0 > F_{[B(1-\alpha)]}^*$, where $F_{[B(1-\alpha)]}^*$ is the $[B(1-\alpha)]$ th quantile of F^* .

For multi-sample situations, Jayalath et al.(2017) proposed to adopt an approach similar to the Hartley’s F test. In this approach, they only consider the two samples with the largest and smallest values of the sample *Dev*. This approach can be summarized as follows.

Suppose that we have k samples Z_1, Z_2, \dots, Z_k that are random samples obtained from the k populations with sample sizes n_1, n_2, \dots, n_k , respectively. Let us denote the samples with the largest and smallest values of the sample *Dev* as X and Y , respectively, i.e.,

$$X = Z_j, \text{ where } \{j : Dev(Z_j) = \max\{Dev(Z_i), i = 1, 2, \dots, k\}\} \quad (11)$$

$$Y = Z_t, \text{ where } \{t : Dev(Z_t) = \min\{Dev(Z_i), i = 1, 2, \dots, k\}\} \quad (12)$$

Then the test statistic for multi-sample situation is

$$F_{RMD} = \frac{\max\{Dev(Z_i), i = 1, 2, \dots, k\}}{\min\{Dev(Z_i), i = 1, 2, \dots, k\}} = \frac{\max\{Dev(X), Dev(Y)\}}{\min\{Dev(X), Dev(Y)\}} \quad (13)$$

Hence, the aforementioned bootstrap procedure for F_{RMD} test of homogeneity of variances of two population can be applied to the k -sample situations (Jayalath et al., 2017).

3. Simulation Study

In this section, a simulation study was conducted to compare the tests introduced in the second section in terms of type I error rate and power. To generate data, the normal (2,2), uniform and gamma (1,4) distributions were used. In this study, while the number of group(k) is 3, the sample sizes(n) are taken as 5,7,10,15,20 and 25. In the simulation study, the number of repetitions is 10000, and the number of bootstrap repetitions for the F_{RMD} test is 1000. For the performance comparison of tests, the experimental type I error rates under H_0 hypothesis were calculated by dividing the number of rejections to 10000. Experimental power values under H_1 hypothesis were also calculated by dividing the number of rejections to 10000. Experimental type I error rates are given in Table 1 and the power results are given in Table 2.

Table 1. When nominal type I error is 0.05 and $k = 3$, the experimental type I error rates of the Bartlett (B), Levene (L), Fligner-Killeen (FK), Hartley (H) and F_{RMD} (RMD) tests.

n	B	L	FK	H	RMD	B	L	FK	H	RMD	B	L	FK	H	RMD
	Normal					Uniform					Gamma (1,4)				
5,5,5	0.050	0.004	0.000	0.050	0.004	0.019	0.002	0.000	0.024	0.003	0.238	0.015	0.000	0.221	0.012
7,7,7	0.054	0.011	0.007	0.050	0.008	0.011	0.006	0.004	0.013	0.004	0.277	0.026	0.021	0.278	0.024
10,10,10	0.047	0.029	0.028	0.051	0.012	0.006	0.020	0.020	0.008	0.005	0.322	0.040	0.061	0.316	0.029
15,15,15	0.048	0.035	0.025	0.049	0.012	0.003	0.026	0.023	0.004	0.006	0.365	0.041	0.077	0.356	0.030
20,20,20	0.048	0.036	0.036	0.050	0.013	0.002	0.027	0.024	0.002	0.007	0.386	0.047	0.085	0.379	0.033
25,25,25	0.050	0.037	0.038	0.048	0.014	0.001	0.029	0.026	0.001	0.010	0.391	0.048	0.082	0.390	0.037

In the simulation results for normal distribution given in Table 1, it is seen that the experimental type I error rates of the B and H tests are very close to the nominal type I error rate. But, the other tests did not give good results in terms of type I error rate. Not all tests for uniform distribution gave results close to nominal type I error rate. In the simulation results related to this distribution, the B and H tests for (5,5,5) and (7,7,7) sample sizes gave slightly better experimental type I error rates compared to other tests. However, the L and FK tests for medium sample volumes gave slightly better results than other tests. In the Gamma (1,4) distribution, it is seen that the F_{RMD} and L tests give better experimental type I error rates for all sample sizes.

Simulation results for the normal distribution given in Table 2 show that the B and H tests yield better results in terms of experimental power values than other tests for all variance values and sample sizes. When the variances for uniform distribution are (1: 1.5: 2.5), (1: 2: 3), (1: 2.5: 3.5), the B and H tests for small sample sizes are more powerful compared to other tests. When the variances are (1: 3: 4) and (1: 3.5: 4.5), the B and H tests yielded better results in terms of experimental power than other tests as in normal distribution. In the simulation results for the gamma distribution given in this table, it is seen that the B and H tests give better results in terms of power values compared to other tests for all variance values and sample sizes considered.

Table 2. When nominal type I error is 0.05 and $k = 3$, the experimental power values of the Bartlett(B), Levene(L), Fligner-Killeen(FK), Hartley(H) and F_{RMD} (RMD) tests.

σ^2	n	B	L	FK	H	RMD	B	L	FK	H	RMD	B	L	FK	H	RMD
		Normal					Uniform					Gamma (1,4)				
1:1.5:2.5	5,5,5	0.097	0.012	0.000	0.092	0.008	0.047	0.007	0.000	0.048	0.005	0.282	0.024	0.000	0.277	0.016
	7,7,7	0.138	0.038	0.021	0.126	0.016	0.055	0.031	0.023	0.051	0.014	0.361	0.043	0.032	0.347	0.040
	10,10,10	0.199	0.118	0.096	0.181	0.045	0.083	0.148	0.126	0.075	0.037	0.436	0.090	0.130	0.419	0.059
	15,15,15	0.293	0.182	0.156	0.284	0.083	0.163	0.230	0.234	0.153	0.085	0.511	0.110	0.151	0.509	0.073
	20,20,20	0.404	0.285	0.255	0.390	0.125	0.296	0.403	0.424	0.329	0.160	0.572	0.160	0.229	0.565	0.090
	25,25,25	0.493	0.369	0.342	0.487	0.180	0.428	0.530	0.556	0.420	0.240	0.626	0.187	0.274	0.617	0.111

1:2:3	5,5,5	0.115	0.012	0.000	0.111	0.010	0.054	0.008	0.000	0.057	0.007	0.319	0.024	0.001	0.293	0.012
	7,7,7	0.165	0.046	0.028	0.167	0.027	0.072	0.033	0.027	0.076	0.020	0.392	0.048	0.038	0.381	0.043
	10,10,10	0.253	0.140	0.128	0.243	0.065	0.117	0.187	0.166	0.135	0.055	0.478	0.100	0.160	0.472	0.074
	15,15,15	0.405	0.236	0.206	0.401	0.125	0.276	0.323	0.313	0.291	0.140	0.572	0.140	0.192	0.570	0.102
	20,20,20	0.535	0.382	0.356	0.533	0.196	0.479	0.545	0.546	0.504	0.247	0.633	0.194	0.297	0.634	0.125
	25,25,25	0.654	0.500	0.459	0.658	0.275	0.662	0.685	0.708	0.680	0.390	0.695	0.253	0.362	0.687	0.160
1:2.5:3.5	5,5,5	0.138	0.014	0.000	0.133	0.012	0.073	0.011	0.000	0.073	0.008	0.337	0.025	0.000	0.323	0.020
	7,7,7	0.203	0.057	0.027	0.214	0.037	0.101	0.048	0.031	0.106	0.025	0.430	0.057	0.042	0.415	0.053
	10,10,10	0.323	0.176	0.169	0.332	0.076	0.194	0.228	0.203	0.210	0.080	0.525	0.118	0.179	0.525	0.092
	15,15,15	0.520	0.310	0.278	0.531	0.170	0.430	0.424	0.412	0.472	0.212	0.623	0.161	0.239	0.622	0.134
	20,20,20	0.678	0.492	0.473	0.675	0.275	0.687	0.679	0.668	0.712	0.381	0.705	0.234	0.371	0.700	0.170
	25,25,25	0.790	0.622	0.587	0.800	0.400	0.855	0.824	0.832	0.865	0.532	0.766	0.321	0.448	0.757	0.212
1:3:4	5,5,5	0.162	0.017	0.000	0.157	0.015	0.084	0.010	0.000	0.093	0.009	0.366	0.027	0.001	0.347	0.023
	7,7,7	0.255	0.058	0.035	0.256	0.046	0.132	0.056	0.035	0.150	0.032	0.452	0.064	0.050	0.453	0.064
	10,10,10	0.404	0.212	0.198	0.413	0.107	0.290	0.284	0.256	0.300	0.107	0.568	0.135	0.209	0.563	0.110
	15,15,15	0.623	0.390	0.350	0.644	0.222	0.599	0.528	0.506	0.631	0.277	0.683	0.189	0.282	0.683	0.165
	20,20,20	0.789	0.596	0.562	0.795	0.365	0.843	0.786	0.784	0.862	0.477	0.756	0.295	0.434	0.759	0.220
	25,25,25	0.886	0.739	0.712	0.893	0.500	0.952	0.909	0.899	0.957	0.660	0.815	0.375	0.530	0.820	0.275
1:3.5:4.5	5,5,5	0.189	0.020	0.000	0.181	0.014	0.102	0.012	0.000	0.114	0.012	0.386	0.031	0.002	0.363	0.027
	7,7,7	0.296	0.071	0.041	0.316	0.051	0.177	0.072	0.042	0.193	0.046	0.496	0.071	0.056	0.484	0.067
	10,10,10	0.480	0.243	0.233	0.496	0.137	0.376	0.336	0.316	0.412	0.144	0.600	0.147	0.242	0.607	0.127
	15,15,15	0.719	0.457	0.416	0.737	0.290	0.752	0.623	0.587	0.775	0.367	0.732	0.225	0.333	0.724	0.205
	20,20,20	0.869	0.681	0.655	0.878	0.440	0.935	0.868	0.850	0.940	0.585	0.804	0.334	0.494	0.809	0.265
	25,25,25	0.940	0.822	0.797	0.945	0.580	0.988	0.957	0.947	0.988	0.750	0.859	0.451	0.615	0.869	0.330

4. Conclusion

According to the experimental type I error rates in Table 1, the B and H tests for normal distribution gave better results, whereas none of the tests gave good results for uniform distribution. However, the F_{RMD} , L and FK tests gave generally better results than the other tests for gamma distribution. The experimental power results given in Table 2 showed that the B and H tests for normal distribution are more powerfully than other tests. On the other hand, the B, H and FK tests generally provided better results for gamma and uniform distributions.

References

- Jayalath, K.P., Ng, H.K.T., Manage, A.B. & Riggs, K.E. (2017). “Improved tests for homogeneity of variances”, Communications in Statistics - Simulation and Computation, vol.469, pp.7423-7446.
- Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Test”, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, vol.160, pp.268-282.
- Levene, H. (1960). “Robust Tests for Equality of Variances”, In Contributions to Probability

and Statistics: Essays in honor of Harold Hotelling, Palo Alto:Stanford University Press, pp.278-292.

Hartley, H. O. (1950). “The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance”, *Biometrika*, vol.37, pp.308-312.

Fligner, M.A. & Killen, T.J. (1976). “Distribution-free two-sample tests for scale”, *J. Amer. Statistical Assoc.* vol.71, pp.210-213

Box, G. E. P. (1953). “Non-Normality and Tests on Variances”, *Biometrika*, vol.40 pp.318-335.

Brown, M. B., Forsythe, A. B. (1974). “Robust Tests for the Equality of Variances”, *Journal of the American Statistical Association*, vol.69, pp.364-367.

Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). “A comparative study of tests for homogeneity of variances”, with applications to the outer continental shelf bidding data. *Technometrics* vol.23, pp.351–361.

Lim, T.-S., Loh, W.-Y. (1996). “A comparison of tests of equality of variances”, *Computational Statistics & Data Analysis* vol.22, pp.287–301.

Higgins, J. (2004). “An Introduction to Modern Nonparametric Statistics”. Brooks/Cole:Duxbury advanced series.

Cochran, W. G. (1951). “Testing a Linear Relation among Variances”, *Biometrics*, vol.7, pp.17-32.

O-69 A Metaheuristic Algorithm for In-Plant Milk-Run System

Islam ALTIN¹, Aydin SIPAHIOGLU²

¹*Department of Industrial Engineering, Eskisehir Osmangazi University, TURKEY,
ialtin@ogu.edu.tr*

²*Department of Industrial Engineering, Eskisehir Osmangazi University, TURKEY,
asipahi@ogu.edu.tr*

Abstract – Milk-run, a cyclic material delivering system, aims to increase the efficiency of transportation and supply chain based on lean logistics perspective. There are two kinds of milk-run systems as supplier and in-plant milk-run system in the literature. In-plant milk-run system that has growing appeals with Industry 4.0 concept, is applied to manage process of delivering materials from warehouse to assembly stations in plants. This system is implemented using Automated Guided Vehicles (AGV), which provide automated materials handling in plant. However, a challenging problem arises in determining milk-run routes and periods allowed to be different for each AGV. Since this problem is quite difficult to handle with exact solution methods, Iterated Local Search algorithm with dynamic penalty function is developed, in this study. Dynamic penalty function increases efficiency of Iterated Local Search algorithm with regard to avoiding local optima by keeping the properties of the last two solutions in memory. Moreover, it prevents exceeding vehicle capacities in obtained solution. In order to evaluate the performance of the proposed algorithm, test problems with different scales derived from the literature are used. The computational results show that the suggested algorithm is efficient to obtain both milk-run routes and periods for each AGV, in a reasonable computational time.

Keywords – *Logistics, In-Plant Milk-Run System, Automated Guided Vehicles, Iterated Local Search Algorithm.*

1. Introduction

Material handling system is one of the most important issues that should be well managed for companies to survive in competition environments. Because, some studies have revealed that a well operated material handling system can reduce plant's operating cost by 15-30% (Sule, 1994). Moreover, efficient material handling system ensures on time delivery, which is a key factor in the implementation of a lean logistics. It is a logistics application in lean manufacturing environment and provides the delivery of the products at the right time to the right place, effectively (Patel, 2013). Lean logistics activities consist of 3 parts such as in-bound, out-bound and in-plant logistics (Baudin, 2005). Just in time material supply should be provided to assembly stations in-plant logistics. Otherwise, it causes inventory holding cost or stopping assembly lines due to the parts shortage (Satoglu and Sipahioğlu, 2018).

In-plant milk-run system is applied to manage the process of in-plant logistics. It ensures delivering materials from the warehouse to the assembly stations in cyclic manner. Material demands of assembly lines are satisfied from a single warehouse in this system. Since it provides goods in small lot sizes with shorter durations, it is vital to determine milk-run routes and periods simultaneously. Implementation of this system to the factory, mass production is made, is appropriate because many different materials should be delivered to the assembly stations at certain periods (Alnahhal et al., 2014).

Choosing the right material handling equipment is essential to improve facility utilization, increase efficiency of material flow and reduce waste in plant (Kilic et al., 2012). There is growing interest using automated material handling vehicles especially with the Industry 4.0 concept. In this sense, in-plant milk-run system can be applied using Automated Guided Vehicles (AGV). AGVs maintain logistics activities using predefined paths for supplying goods in cyclic manner.

Our research aims at developing effective solution method for obtaining milk-run routes and periods simultaneously in-plant. Iterated local search algorithm is applied to overcome this challenging problem in a reasonable computation time. Structure of the iterated local search is enhanced by proposed dynamic penalty function and Simulated Annealing algorithm. The key contribution of this work is solution representation that it provides all information about milk-run routes and periods and exploring the solution space, effectively. Another contribution is that proposed approach achieves high quality solutions at a reasonable computation time even for large-scale problems.

The remainder of this paper has the following structure. In Section 2, problem is described and proposed method is presented in detail. In Section 3, computational results of test problems are discussed. Finally, conclusion and future work are presented in Section 4.

2. Problem Definition and Proposed Method

Fundamental problem encountered in the operation of in-plant milk-run system is determination of milk-run routes and periods for AGVs. Routes and periods of the AGVs must be determined to satisfy demands of the assembly stations to prevent delay in manufacturing within facility. At the same time, it is aimed to minimize total distance of routes. There are some parameters and assumptions to consider when solving this problem. These are;

- Each route starts and ends at the depot.
- Each vehicle is operated at most only one route.
- Each customer is visited only once.
- Homogenous vehicles are used.
- Demands of each stations are known but varies with regard to milk-run period.

- Distances between stations are known and used as a distance matrix.
- Unloading time at each stations are 0.5 minute.
- Speed of each vehicle is 10 meters per minute.

It is necessary to determine milk-run routes and periods for each AGV in order to operate in-plant milk-run system. However, a challenging problem arises in determining milk-run routes and periods, simultaneously. This problem is difficult to handle by exact solution method in a reasonable computation time. Therefore, we have developed Iterated Local Search (ILS) algorithm, which is a metaheuristic method to solve this problem easily.

The quality of solutions obtained by generic local search methods depends on initial solution and that's why these methods stuck into local optima generally. Besides, the initial solution is generated randomly in multi-start local search approaches. To improve this weakness of local search algorithms, ILS algorithm has been proposed by Martin et al. (1991) and generalized by Stutzle (1999), Lourenco et al. (2003). ILS algorithm prevents getting stuck into local optima and allows perturbing local best solutions to use them as initial solutions. There are two different local search algorithms in the structure of ILS. The first algorithm is applied to obtain initial solution. On the other hand, the second algorithm is executed iteratively to improve solution quality. Furthermore, the outstanding mechanism of ILS algorithm is perturbation method. This method makes a major change in the current solution to move from one region of the search space to better one. Accordingly, reverse operator has been used as a perturbation strategy. Pseudo code of ILS is given below.

Pseudo Code of Iterated Local Search Algorithm

Generate initial solution (s_0)

Get another solution (s_*) by applying a local search algorithm (s_0)

Repeat

 Generate different solution (s') by using Perturbation function (s_*)

 Get another solution (s'_*) by applying a local search on the perturbed solution (s')

 Accept solution (s'_*) considering accepting criteria

Until Stopping criteria

The best solution achieved

In order to apply metaheuristic algorithm effectively, an appropriate solution representation structure should be used. It should be easy to generate new and feasible solutions from a current solution. Proposed solution representation is shown in Figure 1. It consists of positive and zero numbers. Positive numbers represent delivery customers and zero numbers indicate vehicles. So, this solution representation length is related to the number of customers and vehicles.

8	2	5	0	1	9	7	0	6	4	3
---	---	---	---	---	---	---	---	---	---	---

Figure 1. An example of solution representation.

The main advantage of the proposed ILS algorithm is the suggested solution representation. This representation provides all information about milk-run routes and periods that are vital to operate in-plant milk-run system, efficiently. Referring to Figure 1, it is clear that there are 9 customers and 3 vehicles. Milk-run routes for each vehicle are as follows; first route is 0-8-2-5-0, second route is 0-1-9-7-0 and third route is 0-6-4-3-0. On the other hand, milk-run periods are calculated taking into account trip time of route and unloading time at each station.

We have applied Simulated Annealing (SA) algorithm as a local search approach due to its simplicity and efficiency. SA algorithm has strengthened the structure of the ILS with regard to allowing the acceptance of non-improving solutions. In addition, dynamic penalty function has been developed in order to obtain high quality solutions with SA algorithm. Proposed dynamic penalty function (1) is given below with definitions.

$Z = \text{Objective Function Value}$

$p = \text{The Longest Distance in Distance Matrix}$

$k = \text{Penalty Coefficient}$

$$k = \begin{cases} 2, & \text{If the vehicle capacity is exceeded during the last two iterations} \\ 1, & \text{If the vehicle capacity is exceeded only the last iteration} \\ 0, & \text{If the vehicle capacity is not exceeded only the last iteration} \end{cases}$$

$$Z' = Z + (k * p) \quad (1)$$

Vehicle capacity is allowed to be exceeded in the last two successive iterations with dynamic penalty function. Therefore, this approach makes exploring the solution space, extensively.

Another attempt to achieve high quality solutions is using different neighbor search operators in SA. Four different search operators have been used such as swap, insert, 3-change and 4-change. These operators have been used probabilistically with regard to the scale of the problem. For small scale problems (number of customer ≤ 35), only swap and insert operators have been used with equal probability. On the other hand, for medium (number of customer $\in (35, 70]$) and large scale problems (number of customer > 70), all operators have been used with equal probability.

Pseudo code of the SA algorithm applied with the proposed dynamic penalty function and different neighborhood procedures is given below.

Pseudo Code of Proposed Simulated Annealing Algorithm

Create initial solution, generate current and the best solution from the initial solution:

$$s_0, f(s_0); s_{best} = s_0; f(s_{best}) = f(s_0); s = s_0; f(s) = f(s_0)$$

Setting initial temperature, cooling coefficient, the longest distance in distance matrix:

$$T = T_0; \alpha = 0,99; p = \max(c_{ij})$$

Repeat

Repeat

Generate s' neighbor solution randomly (using neighbor search operators)

Calculate $f(s')$ with dynamic penalty function; $f(s') = f(s') + (k * p)$

If the vehicle capacity is exceeded during the last two iterations, then; $k=2$

If the vehicle capacity is exceeded only the last iteration, then; $k=1$

If the vehicle capacity is not exceeded the last iteration, then; $k=0$

$$f(s') = f(s') + (k * p)$$

$$\Delta = f(s') - f(s);$$

If $\Delta \leq 0$

Accept neighbor solution ($s = s'$); Otherwise generate random number $u = (0,1)$

If $u < e^{-\Delta/T}$, accept neighbor solution ($s = s'$);

If $f(s') < f(s_{best})$, $s_{best} = s'$

Until (Equilibrium Condition)

Update temperature $T = T * \alpha$

Until (Stopping Criteria Satisfied)

The Best Solution Found s_{best} , $f(s_{best})$

The performance of the proposed ILS algorithm is shown on some test problems presented by Hoff et al. (2009). Obtained results are discussed in following section.

3. Computational Results

In order to use test problems presented by Hoff et al. (2009), adapting the test problems to the nature of in-plant milk-run system, some changes have been made. Pickup demands of the assembly stations have been removed and delivery demands have been used as the amount of material required by each assembly stations in 10 minutes' period. Therefore, the amount of materials required by each station varies with regard to milk-run period.

In the proposed ILS algorithm, number of iteration has been used as a stopping criteria. Regardless of the problem size, number of iterations for all test problems has been determined as 5. Acceptance criteria, another criterion of the ILS, has been applied, deterministically. In the SA applied as a local search algorithm, number of iteration has been used as a stopping

criteria and geometric cooling schedule has been used. Parameter values of two different SA algorithm used in the structure of the ILS for different size of problems are presented in Table 1.

Table 1. Simulated Annealing Parameter Values.

	Small scale problems		Medium scale problems		Large scale problems	
	First SA	Second SA	First SA	Second SA	First SA	Second SA
Initial temperature (T_0)	100	100	150	150	200	200
Final temperature (T_F)	0.1	0.1	0.1	0.1	0.1	0.1
Number of iteration (I)	15	50	20	75	25	100
Cooling rate (α)	0.99	0.99	0.99	0.99	0.99	0.99

The developed ILS algorithm for the in-plant milk-run system was coded in Python 3.7 and run on an Intel Core i5 3.1 GHz PC with 4 GB memory. Algorithm has run 5 times for each test problem and the best obtained computational results have presented on Table 2.

Table 2. Numerical Results of the Proposed Algorithm on the Generated Instances.

Test Problem	Assembly Stations	Number of AGVs Used	Total Distance	Computation Times
E-016-03	15	4	282.94	16.87
E-021-04	20	5	385.29	26.32
E-022-04	21	5	437.95	20.56
E-026-08	25	9	626.19	21.98
E-030-03	29	6	659.96	25.81
E-036-11	35	10	666.27	64.09
E-041-14	40	13	804.65	78.05
E-045-04	44	8	921.87	89.08
E-051-05	50	8	651.54	72.80
E-072-04	71	6	324.81	113.09
E-076-07	75	11	896.33	115.74
E-101-10	100	14	1268.76	126.07

Numerical results of the different scale instances are summarized in Table 2 in terms of number of AGVs used, total travelled distance and computation times. The first two columns of Table 2 contain information about test problems, such as instance name and the number of assembly stations. Third column shows the number of AGVs used in test problems. Next column indicates the obtained total travelled distance and the last column includes computation times in seconds. In these test problems, demands of the assembly stations have been satisfied by AGVs without exceeding their capacity and the total distance has been minimized. It is seen that the proposed ILS algorithm is able to produce solutions for even large-scale problems at a reasonable computation time.

As an example, detail solution report of a test problem is given in Table 3.

Table 3. Milk-Run Routes and Periods for the Selected Test Problem.

Test Problem	AGV Capacity	Number of AGVs Used	Milk-Run Routes	Milk-Run Periods
E-016-03	90	4	0-4-13-14-0	8.37
			0-6-7-8-0	7.89
			0-11-2-3-1-0	9.10
			0-5-9-10-15-12-0	10.41

From Table 3, it can be understood that there are 4 AGVs to satisfy demands of 15 assembly stations in cyclic manner. First AGV supplies material 4, 13, 14th customers every 8.37 minutes, second AGV supplies material 6, 7, 8th customers every 7.89 minutes, third AGV supplies material 11, 2, 3, 1st customers every 9.10 minutes and fourth AGV supplies material 5, 9, 10, 15, 12th customers every 10.41 minutes. Of course, these periods can be rounded upper value.

Sensitivity analysis has also been performed on capacity of AGVs for E-021-04 instance. As the capacity of AGV is decreased, it is observed that the duration of milk-run periods is shortened and the number of AGV used is increased. On the other hand, larger AGV capacity results in longer milk-run periods. However, too long milk-run periods are inconsistent to the in-plant milk-run principle. Because, in-plant milk-run strategy is applied for supplying material to stations with short period. Consequently, decision maker can identify milk run periods and number of AGV, by using this approach.

4. Conclusion

In this paper a metaheuristic algorithm based on iterated local search was suggested to obtain milk-run routes and periods for AGVs, simultaneously. The proposed approach allows achieving different milk-run periods for each AGV considering trip time and unloading time at each station. Generic structure of iterated local search has been strengthened by using simulated annealing algorithm with dynamic penalty function and various neighbor search operators. In addition, a well-designed solution representation has been proposed. The main advantageous of the proposed approach is to allow obtaining milk run periods and number of AGV's to be used. Suggested algorithm was tested over a set of test problems derived from the literature. Computational results show the success of the developed algorithm with regard to solution time even for large scale instances. In future studies, buffer stock constraints can be examined.

References

- Alnahhal, M., Ridwan, A., Noche, B. (2014). “In-plant milk run decision problems”, In 2014 International Conference on Logistics Operations Management, IEEE, Rabat, Morocco.
- Hoff, A., Gribkovskaia, I., Laporte, G., & Løkketangen, A. (2009). “Lasso solution strategies for the vehicle routing problem with pickups and deliveries”, *European Journal of Operational Research*, vol. 192, no. 3, pp. 755-766.
- Kilic, H.S., Durmusoglu, M.B., Baskak, M. (2012). “Classification and modeling for in-plant milk-run distribution systems”, *The International Journal of Advanced Manufacturing Technology*, vol. 62, no. 9-12, pp. 1135–1146.
- Lourenco, H. R., Martin, O. C., Stutzle, T. (2003). “Iterated local search”, F. Glover and G. Kochenberger (eds.), *Kluwer Academic Publishers, International Series in Operations Research & Management Science*, pp. 321-353.
- Martin, O., Otto, S. W., & Felten, E. W. (1991). “Large-step markov chains for the traveling salesman problem”, *Oregon Graduate Institute of Science and Technology, Department of Computer Science and Engineering*, vol. 5, no. 3, pp. 299-326.
- Patel, M. B. (2013). “Optimization approach of vehicle routing by a milk-run material supply system”, *International Journal for Scientific Research & Development*, vol. 1, no. 6, pp. 1357-1360.
- Satoglu, S. I., Sipahioglu, A. (2018). “An assignment based modelling approach for the inventory routing problem of material supply systems of the assembly lines”, *Sigma Journal of Engineering and Natural Sciences*, vol. 36, no. 1, pp. 161-177.
- Stutzle, T. (1999). “Local search algorithms for combinatorial problems: Analysis, algorithms and new applications”, PhD thesis, DISKI—Dissertationen zur Kunstliken Intelligenz., Sankt Augustin, Germany.
- Sule, D. R. (1994). *Manufacturing Facilities: Location, Planning and Design*, 2nd ed. PWS Publishing Company, Boston, US.

O-70 A New Risk Assessment for the Right-Skewed Processes

Melis Zeybek^{1*} and Onur Köksoy²

¹*Department of Statistics, Ege University, İzmir, melis.zeybek@ege.edu.tr*

²*Department of Statistics, Ege University, İzmir, onur.koksoy@ege.edu.tr*

Abstract – Poorly operated production units and incomplete designed products cause major incidents involving monetary and social losses. Quality experts commonly utilize loss functions to develop new methodologies for quality improvement. The widespread use of loss functions in industrial applications has increased their popularity among statisticians and engineers due to their different loss-handling features. They mainly emphasize the importance of being on desired target with a small variation to reduce all types of manufacturing costs. This paper presents a new member of the inverted probability loss family. We propose a loss function by inverting the density first introduced by Zeybek and Köksoy (2018) for the responses under gamma noise effect. The proposed loss function has a right-skewed structure along with the range of $(-\infty, \infty)$.

Keywords – *asymmetric quality loss, risk assessment, inverted probability loss family, right-skewed processes.*

1. Introduction

Traditionally, loss functions have an undisputed background in the development of quality engineering. Taguchi claimed that in much industrial production, there is a need to produce an outcome on target. Therefore, quality engineering needs to start with an understanding of quality costs. According to Taguchi (1986), the damage caused by weakly operated production resources and undesirably designed products should be evaluated as economic and communal losses. Moreover, even if the product performance were to lie within the pre-defined limits, a loss may still be inevitable. Losses can be considered from two different perspectives: the corporation view (for example, the party responsible for meeting the costs relating to returned products, reworking, scrap, and repair) and the customer view (for example, customer dissatisfaction due to a poor product performance). Thus, the quality loss concept is referred as “loss to society.” Some useful references for the Taguchi’s robust design methodology include Yang et al. (2012), Köksal et al. (2013), and Low et al. (2016).

Quality experts commonly utilize loss functions to develop new methodologies for quality improvement. Taguchi’s novel philosophy popularized the quadratic loss function in industrial applications. Even though the basic idea behind Taguchi’s quadratic loss function (Taguchi, 1986 and Taguchi et al., 1989) is sound, it has some difficulties in application. For example, it is unbounded and therefore continuously penalizes the off-target process measurements, rather than attaining an upper limit. Therefore, it might be inefficient and may

result in unsatisfactory product performance. Based on these shortcomings, Spiring (1993) introduced the inverted normal loss function (INLF) to provide more reasonable economic loss assessments. Subsequently, new modified versions have appeared in papers by Sun et al. (1996), Drain and Gough (1996), Pan (2007), and Köksoy and Fan (2012). Moreover, Spiring and Yeung (1998) introduced a class of inverted probability loss functions (IPLF). Then, Leung and Spiring (2002, 2004) discussed the properties of the IPLF class, including both symmetric and asymmetric loss concepts.

2. Overview of IPLF family

The idea of utilizing the inverted normal density as a loss function was an important step and has led to the use of other inversions to generate loss functions with different loss-handling features. Spiring and Yeung (1998) and Leung and Spiring (2002, 2004) studied the IPLF family and discussed its properties. This general class has a number of fundamental criteria – for example, that the loss must be non-negative, can take value of zero only at target, increases monotonically, and reaches a quantifiable maximum near the lower specification limit (LSL) and/or upper specification limit (USL).

In the IPLF context, let $f(y)$ be a pdf possessing a unique maximum at $y = \tau$, where τ is the target value. Let $\pi(y, \tau)$ be the form of this pdf in terms of y and τ , and $m = \sup_{y \in Y} f(y) = f(\tau)$ is the maximum value of $\pi(y, \tau)$. The general formulation of the IPLF takes the following form,

$$L(y) = K \left(1 - \frac{\pi(y, \tau)}{m} \right) \quad (1)$$

where $\frac{\pi(y, \tau)}{m}$ is referred to the loss inversion ratio (LIR) and K is a constant which represents the maximum loss incurred when the target is not attained. When the process characteristic y is on the target, LIR has a minimum value of one, and the loss function therefore takes the values between zero and K . The key point here is that the target must be at the modal point of the pdf to be inverted (Spiring and Yeung, 1998; Leung and Spiring, 2002, 2004).

The INLF was first proposed by Spiring (1993) and takes its place in the IPLF class. The shape of an inverted normal density looks almost like a quadratic loss function; however, it provides a bounded loss assessment feature. The INLF has the following form,

$$L_{IN}(y) = K \left(1 - \exp \left(-\frac{1}{2} \frac{(y - \tau)^2}{\sigma_L^2} \right) \right) \quad (2)$$

where σ_L^2 is the scale parameter. Figure 1(a) illustrates the shape of the INLF when $\tau = 0$ for various values of σ_L . As seen from Figure 1(a), the INLF has a symmetric shape around the target and has a smooth tendency to the maximum loss. When the value of σ_L decreases, the loss tends to reach its maximum more slowly and large deviations from the target can also be tolerated.

The introduction of the general IPLF class allows the generation of asymmetric loss functions. Spiring and Yeung (1998) suggested the inverted gamma loss function (IGLF), which evaluates the losses for positive values only. The right-skewed shape of the IGLF provides practitioners with a realistic loss assessment where the drifts of process to different sides of the target result in different kinds of economic loss. The IGLF is defined as follows:

$$L_{IG}(y) = K \left(1 - \left(\frac{y \exp\left(1 - \frac{y}{\tau}\right)}{\tau} \right)^{\alpha_L - 1} \right) \quad (3)$$

where α_L is the shape parameter. Figure 1(b) illustrates the IGLF for various values of α_L when $\tau = 5$. As the process drifts to the left-hand side of the target, the loss reaches its maximum more rapidly. Thus, the right-hand side of the IGLF can be more tolerable than the other side. Additionally, the larger α_L gives a more sensitive loss function.

Köksoy, Ergen and Zeybek (2019) suggested the inverted Wald loss function (IWLF) as a new member of the inverted probability loss family, which evaluates the losses for positive values only. The right-skewed shape of the IWLF is defined as follows:

$$L_{IW}(y) = K \left(1 - \frac{\sqrt{\tau^3}}{\sqrt{y^3}} \exp\left(-\frac{(\lambda_L - 3\tau)(y - \tau)}{2\tau^2} - \frac{\lambda_L}{2} \left(\frac{1}{y} - \frac{1}{\tau} \right) \right) \right) \quad (4)$$

where λ_L is the scale parameter, and this allows practitioners to customize the IWLF in order to reflect the losses associated with deviations from the target. Since the Wald pdf has a right-skewed shape, the arm on the right side of the IWLF has a wider form than its counterpart on the left side. Therefore, an equal drifting of the process from either side of the target does not assign the same loss. The small λ_L opens up the arms of the IWLF around τ , so that the IWLF reaches its maximum with a slower increment.

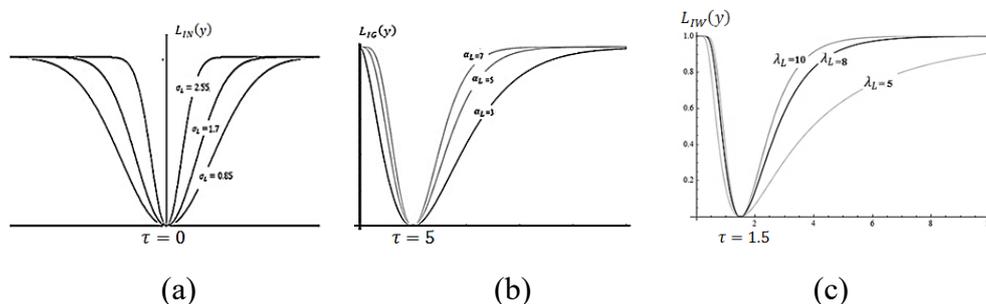


Fig. 1. The illustrations of some members of IPLF: (a) INLF when $\sigma_L = 2.55, 1.7$ and 0.85 ; (b) IGLF when $\alpha_L = 3, 5$ and 7 ; (c) IWLF when $\lambda_L = 5, 8$ and $10, \tau = 1.5$.

While the loss function is a useful tool to reflect the economic loss incurred from any departures from the target, the risk function provides an average economic loss, and it can be determined by the expected loss as follows:

$$EL(y) = \int_{-\infty}^{+\infty} f_Y(y)L(y) dy$$

where $f_Y(y)$ is the pdf for a given process and $L(y)$ is the loss function of interest. Thus, this approach makes the process pdf more important in selecting the most appropriate IPLF member to determine the true process loss.

3. The effect of gamma noise to the response distribution

The distribution of data plays an important role for response surface optimization. Generally, the usual assumption behind the robust design is that the data follows a normal distribution or there is no major contamination in data. However, the belief that 'the normality assumption is a robust assumption' may not be true and there are many cases where it does not apply in real-world problems. In many industrial experiments, the responses are Poisson (count data), exponential or gamma (time-to-failure data), or Bernoulli (defective/non-defective).

The treatment for the noise factors during the experiment is also a topic of discussion in the literature. As mentioned in Bingham and Nair (2012), the settings of noise factors in the robust design of experiments should be chosen judiciously and the correct levels depend upon the true mean and the true variance of a distribution in manufacturing and operating/use conditions. However, their true distributions are rarely known by the practitioners. Fluctuations in the noise factors may not be well described by any symmetric probability distribution, such as the normal, in the real life problems, as noted by Abraham *et al* (2009), if the noise factors are non-normal, then the distribution of responses is affected by its structure. In this case, we encounter a problem of non-normal responses. The unrealistic assumptions about noises cause inaccurate results in quality improvement studies, so the success of a robust design heavily relies on describing the accurate distributional structures for the noise factors.

Zeybek and K oksoy (2018) investigated the effects of a gamma distributed noise on the response and also on the optimization. They assumed that there are k control factors and one noise factor affect the response,

$$Y(x, Z) = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^k \beta_{jj} x_j^2 + \sum_{j < t}^k \sum \beta_{jt} x_j x_t + \gamma Z + \sum_{j=1}^k \delta_j x_j Z + \varepsilon \quad (5)$$

Where $\varepsilon \sim NID(0, \sigma_\varepsilon^2)$ and the noise variable follows a gamma distribution with a density

$$f_Z(z) = \frac{1}{\Gamma(\alpha)\beta^\alpha} z^{\alpha-1} \exp\left(-\frac{z}{\beta}\right),$$

where $\alpha > 0$ and $\beta > 0$. And they proposed the following theorem provides a density function for the given response in Equation (5).

Theorem of Zeybek and K oksoy (2018):

Under $\varepsilon \sim NID(0, \sigma_\varepsilon^2)$ and $Z \sim G(\alpha, \beta)$, the density function of the response in Equation (5) is as follows:

$$f_Y(y(x, z)) = \frac{\sigma_\varepsilon^{\alpha-1}}{\sqrt{2\pi}\beta^\alpha(\gamma + \sum_{j=1}^k \delta_j x_j)^\alpha} D_{-\alpha} \left(\frac{\sigma_\varepsilon \left(\frac{1}{\beta} - \frac{(y(x, z) - \mu_x)(\gamma + \sum_{j=1}^k \delta_j x_j)}{\sigma_\varepsilon^2} \right)}{(\gamma + \sum_{j=1}^k \delta_j x_j)} \right) \times \exp \left(-\frac{(y(x, z) - \mu_x)^2}{2\sigma_\varepsilon^2} + \frac{\sigma_\varepsilon^2 \left(\frac{1}{\beta} - \frac{(y(x, z) - \mu_x)(\gamma + \sum_{j=1}^k \delta_j x_j)}{\sigma_\varepsilon^2} \right)^2}{4(\gamma + \sum_{j=1}^k \delta_j x_j)^2} \right) \quad (6)$$

where $(\gamma + \sum_{j=1}^k \delta_j x_j) > 0$, $\alpha = 1, 2, 3, \dots$, $D_{-\alpha}(\cdot)$ is a parabolic cylinder function and,

$$\mu_x = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^k \beta_{jj} x_j^2 + \sum_{j < t}^k \sum \beta_{jt} x_j x_t \quad (7)$$

The graph of $f_Y(y(x, z))$ is presented in Figure 2 when $\varepsilon \sim NID(0, 4)$ and $V(Y(x, Z)) = 20, 9.34, 5.77$ for $\alpha = 1, 3, 9$ respectively. From Figure 2, it appears that there is an opposite relationship between $V(Y(x, Z))$ and α . In other words, when α is increased, the variability is goes down and accordingly at the higher values of α reduces the skewness in the response. The most conspicuous property is that the proposed density in Equation (5) has a right-skewed shape just like gamma density; however, it also takes negative values. In addition, the skewness of a distribution depends only on the shape parameter and it approaches to the normal density when α is large (e.g., as seen from Figure 2, when $\alpha = 9$, the density curve may be approximated by a normal distribution). The distributional structure of the proposed pdf can be explained by an effective dominance of the right-skewed noise factor on the response.

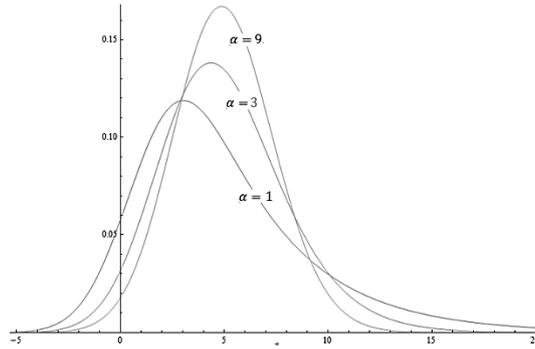


Fig. 2. The overlaid graphs of the proposed pdf $f_Y(y(x, z))$ for $\alpha = 1, 3$ and 9 .

4. Discussion on a new loss function

Then the main motivation of this study, the reversed version of the pdf of Zeybek and Köksoy (2018) can be used as a loss function, and, *it may be have superior ability than IGLF in terms of quantifying the losses of negative values for quality characteristic for right-skewed processes.*

Figure 3 illustrate the revised version of the pdf of Zeybek and Köksoy (2018).

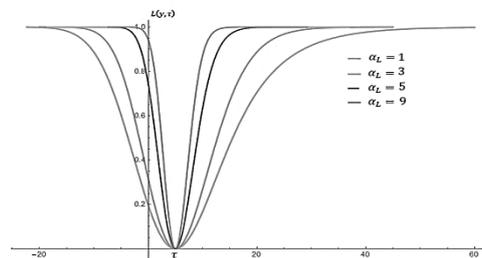


Fig 3. The revised version of the $f_Y(y(x, z))$ of Zeybek and Köksoy (2018).

It is clear that, the proposed loss function:

- ✓ Asymmetric – right skewed around τ
- ✓ Evaluate the losses when quality characteristic has *positive and negative* values.
- ✓ Larger α_L gives a more sensitive loss function.

In fact, both the revised version of the pdf of Zeybek and Köksoy (2018) and IGLF have right-skewed structure, but the main difference that the revised version of the pdf of Zeybek and Köksoy (2018) has ability to measure loss for negative values.

Suppose that the process has a right-skewed distribution along with the range $(-\infty, \infty)$ such that the pdf of Zeybek and Köksoy (2018) as discussed in Section 3. The illustrations of IGLF

and the revised version of the pdf of Zeybek and Köksoy (2018) are given in Figure 4 in terms of quantifying the losses for the mentioned process distribution.

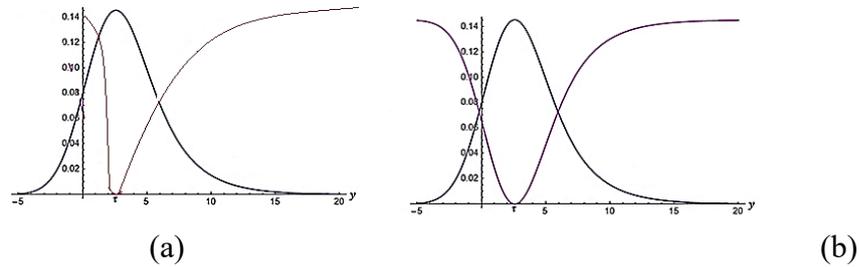


Fig. 4. The illustrations of (a) IGLF and (b) the revised version of the pdf of Zeybek and Köksoy (2018).

It is clear that from Fig. (a), the IGLF is not suitable for the process distribution, since the losses of negative values for quality characteristic cannot be quantified. However, the proposed asymmetric loss given in Fig. (b), has an appropriate definition range for the process.

5. Conclusion

In this study, in fact, we aim to reach a loss function that is skewed to the right but can also measure negative values. In the first findings we obtained that the revised version of the pdf of Zeybek and Köksoy (2018) is an appropriate probability distribution for our purpose. And, we agreed that using the technique of Spiring and Yeung (1998), Leung and Spiring (2002, 2004) to obtain the exact formulation of the inversion version of $f_Y(y(x, z))$. So, the next stage of this study is to obtain the exact formulation of the proposed loss function and make a comparison study of the proposed loss function via INLF and IGLF under various process distributions.

References

- Abraham, A.A., Robinson, T.J. and Anderson-Cook, C.M. (2009). “A graphical approach for assessing optimal operating conditions in robust design”, *Quality Technology and Quantitative Management*, 6(3): 235-253.
- Bingham, D. and Nair, V.N. (2012). “Noise variable settings in robust design experiments”, *Technometrics*, 3: 388-397.
- Drain, D.C., Gough, A.M. (1996). “Applications of the upside-down normal loss function”, *IEEE Transactions on Semiconductor Manufacturing*, 9(1):143–145.

- Köksal, G., Taşeli, A., Dolgun, L.E., Batmaz, I. (2013) ‘The effect of inspection error on quality and producer losses: the case of nominal-the-best type quality characteristic and rework’, *European Journal of Industrial Engineering*, 7(4):497 – 528.
- Köksoy, O., Fan, S.S. (2012). “An upside-down normal loss function-based method for quality improvement”, *Engineering Optimization*, 44(8): 935–945.
- Köksoy O., Ergen P., Zeybek M. (2019). “A new right-skewed loss function in process risk assessment”, *European Journal of Industrial Engineering*, 13:(4), 536-551.
- Leung B.P.K., Spiring, F.A. (2002) “The inverted beta loss function: properties and applications”, *IIE Transactions*, 34(12):1101-1109.
- Leung, B.P.K., Spiring, F.A. (2004) “Some properties of the family of inverted probability loss functions”, *Quality Technology and Quantitative Management*, 1(1):125-147.
- Low, S.N., Kamaruddin, S., Azid, I.H.A. (2016). “An integrated simulation with design on experiment approach for shop floor improvement solution selections”, *European Journal of Industrial Engineering*, 10(4):479 – 498.
- Spiring, F.A. (1993). “The reflected normal loss function”, *Canadian Journal of Statistics*, 3: 321-330.
- Spiring, F.A., Yeung, A.S. (1998) ‘A general class of loss functions with industrial applications’, *Journal of Quality Technology*, 3(2):52-162.
- Sun, F, Laramee, J. and Ramberg, J. (1996). “On Spiring’s inverted normal loss function”, *Canadian Journal of Statistics*, 2: 214-249.
- Pan, J.N.(2007). “A new loss function-based method for evaluating manufacturing and environmental risks”, *International Journal of Quality and Reliability Management*, 24(8): 861-887.
- Taguchi, G. (1986) *Introduction to Quality Engineering: Designing Quality into Products and Processes*, Kraus, White Plains, NewYork.
- Taguchi, G., Elsayed, E.A. and Hsiang, T. (1989) *Quality Engineering in Production Systems*, McGraw-Hill, NewYork.
- Yang, T., Shen, Y.A., Cho, C., Lin, Y.R. (2012) ‘The use of a simulation, a hybrid Taguchi, and dual response surface methods in the automated material handling system tool-to-tool strategy for a 300-mm fab’, *European Journal of Industrial Engineering*, Vol. 6, No. 3, pp.281 – 300.
- Zeybek, M. and Köksoy, O. (2018). The effects of gamma noise on quality improvement. *Communications In Statistics-Simulation And Computation*.

O-71 Analysis of Internal Migration in İzmir with a Binary Logit Model

TUBA İLHAN^{1*} , ŞENAY ÜÇDOĞRUK BİRECİKLİ²

^{1*} Econometrics, Social Sciences Institute, Dokuz Eylül University, Turkey,
tubailhan476@gmail.com

² Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylül University,
Turkey, s.ucdogruk@deu.edu.tr

Abstract- With the help of the Household Labor Force Survey data for 2014-2017, it has aimed to determine the internal migration to İzmir. In the Household Labor Force Survey, a data set was created based on the question “Which of these settlements you have previously resided? ”. A second model was established by adding İzmir's labor force participation rate and GDP per capita data to the data. The wage income of the individual has been added to the data set by deducting the price effect and taking the logarithm. Binary logit model has used for analysis. A second model was established by adding İzmir's labor force participation rate and GDP per capita rate to the data. According to the results of the binary logit model, which gives the possibility of migration to İzmir, migration decreased from 2014 to 2017. There has a linear relationship between education and age. Another finding is that married individuals are 16% more likely to migrate than unmarried individuals. According to the variables added to Model 2, the increase in İzmir's per capita gross domestic product increased the probability of migration to İzmir by 38%. Labor force participation rate leads to an increase of 1% on the probability of immigration to İzmir.

Keywords – *Internal Migration, Binary Logit Model*

1. Introduction

Migration is simply the way a person moves from one's place of residence to another. Migration has aroused curiosity and has been researched in many fields of science since it is an action that has existed from the past to the present and continues to exist. For this reason, the definition of the concept of migration differs according to each field and its definition is diversified. For this reason, the definition of the concept of migration differs according to each field and its definition is diversified. The census are utilized when dealing with the concept of internal migration in Turkey. From 1927 to 1950s, internal migration was done as a result of mechanization and industrialization in agriculture in order to search for jobs in cities. The migrations after the 2000s were mostly due to the need for educated and qualified labor force (Bostan, 2017: 9). Işık (1999), in his study, it was seen that the migrations made to İzmir with the data of 1970-1990 were from the provinces close to İzmir, but this situation

changed and left its place to Southeast, East and Central Anatolia. In the study of Işık (2009), it is concluded that İzmir province continues to be the province with the highest net migration after İstanbul with the data of 1995-2000 period. When the migration to İzmir province is taken into consideration at the regional level, there has been a significant increase in the migrations received from the Southeastern Anatolia Region. Işık (2017) also has used the population data for the years 2010-2015 to observe the change in İzmir's position among the provinces. According to the findings of the study, it has been concluded that İzmir's population mobility has increased and the migration volume has increased however, it has been concluded that it has become a province giving as much migration as it receives migration.

2. Materials and Methods

In the "Introduction" section, the definition of migration is mentioned first. We then discussed the internal migration in Turkey and Izmir. The phenomenon of internal migration in İzmir will be dealt with from the studies on İzmir. Then, the method used in the study, theoretical framework, model outputs will be mentioned.

2.1 Data and method

This study's data set were made by the Turkey Statistical Institute Household Labor Force Survey of the 2014-2017 year were created by pulling of the data from Izmir data. Based on the question " Which of these settlements will you count as a place of residence? " included in the questionnaire, individuals with residency abroad were removed from the data of İzmir province. With the data set, the dependent variable has migrated and has not migrated, and a binary logit model with independent variables such as household size, gender, age, education, marital status, working status, sector and year has been established. By using the Classification of Economic Activities, sector codes are divided into five groups as agriculture, mining, manufacturing, construction and services. Then, in order to see the relationship between individual migration and this data set, the model was re-established by adding İzmir's per capita gross domestic product and labor force participation rate. In the second model, the year variable is considered as a trend variable so that it does not reflect multiple linear connection problems.

2.2. Theoretical framework

In addition to quantitative data, we may need to add qualitative variables to the model. Variables such as gender, education, and marital status of individuals are qualitative variables and most of the qualitative variables consist of binary options such as gender (female-male), education (illiterate), marital status (married-single). The model types in which the dependent variable decides between these two preferences and explains their preferences are binary preference models. Logit models are subjected to logit regression and designed specifically for binary dependent variables, but it is a regression model that can be linearized by

appropriate transformations (Astar, 2009: 48). Logit model, which is one of the binary preference models, is a special case of the generalized linear model created under certain conditions and the model is derived from the logistic distribution function (Gujarati, 1999: 555). Logistic distribution function;

$$P_i = F(Z_i) = F(\beta_0 + \beta_1 X_i) = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_i)}} \quad (1)$$

P_i, Information about X_i descriptive variable is data (constant) i. expresses the possibility of an individual making a particular choice. ‘e’ is the natural logarithm base and is equal to 2,718. Here, the variable Z_i takes values between -∞ and + ∞. The probability value as the variable Z_i changes in this range it will take values between 0-1 in (P_i). P_i is not linear according to both independent variables and parameters (Gujarati, 2001: 554-559). Since the dependent variable will not be defined when probability takes 1 or 0, parameters cannot be directly estimated by the least squares method to determine the function (Wang ve Chaman, 2003: 217). In order to estimate the model, the relationship between Z_i and P_i must be linearized.

2.3. Empirical Findings

According to Table 1, the number of individuals who migrated in the dependent variable used in the two-state logit model was 34.409 while the number of individuals who did not migrate was 32.193.

Table 1: Distribution of Dependent Variable

Dependent variables used in binary logit model		
Migrated		34 409
Not migrated		32 193
Total		66 602

Table 2: Descriptive Statistics

Variable		Mean	Standard Deviation	Variable		Mean	Standard Deviation
Wage income		0.2929	0.4551	marital status	Single (B.C.)	0.2337	0.4231
Household size		3.1970	1.3352		Married	0.6503	0.4768
Sex	Male	0.4793	0.4995		Divorced	0.0432	0.2034
	Woman (B.C.)	0.5206	0.4995		Wife is dead	0.0727	0.2596
Age condition	15-19 age (B.C.)	0.0879	0.2832	Working status	Reference week worked (B.C.)	0.4440	0.4968
	20-24 age	0.0727	0.2597		Reference week not worked	0.5559	0.4968

	25-29 age	0.1399	0.3469	Sector	Agriculture (B.C.)		0.0669	0.2498	
	30-34 age	0.0809	0.2727		Mine		0.0012	0.035	
	35-39 age	0.0937	0.2914		Production		0.0986	0.2982	
	40-44 age	0.1004	0.3005		Build		0.0342	0.1817	
	45-49 age	0.0977	0.2969		Service		0.2565	0.4367	
	50-54 age	0.0870	0.2819		Not working		0.5424	0.4981	
	55-59 age	0.0922	0.2894		Year	2014 (B.C.)		0.2527	0.4346
	60-64 age	0.0788	0.2694			2015		0.2500	0.4330
	Over age 65	0.0683	0.2522			2016		0.2494	0.4326
				2017		0.2477	0.4317		
Education	School has not finish	0.1166	0.3210						
	Primary school graduate (B.C.)	0.3454	0.4755		Min.	Max	Average	Standard Deviation	
	secondary school graduate	0.1734	0.3786	Gross domestic product of İzmir per capita	3.439	3.807	3.6025	0.1360	
	High school graduate	0.1947	0.3960	Labor force participation rate of İzmir	53.5	55.4	54.3951	0.9029	
	Graduate of College or Faculty	0.1542	0.3611	Trend			2.4921	1.1184	
	Graduate or Ph.D.	0.0154	0.1232						

B.C. : Base Class

According to descriptive statistics in Table 2, while the average household size was 3.19, approximately 48% of the respondents were male and 52% were female. When we look at the age groups in the study, it is seen that 14% to 25-29 years of age are at most. Other age groups have similar average values. In the case of education, it is seen that the individuals with the most primary school graduation are 34%. This is followed by high school graduates with 19%. 65% of the subjects subject to the study are married, 43% are divorced and 23% are unmarried individuals. 44% of the individuals worked during the reference week and 56% did not. As the sector, 25% of individuals work in the service sector and then 9% in the manufacturing sector. While the gross domestic product per capita for the 66,602 people included in the survey is 36 thousand TL, the labor force participation rate of these people is 54 percent.

Table 3: Marginal Effects of Binary Logit Model

Variable	dy /dx	p> z	Variable	dy /dx	p> z
Wage income	0.0820	0.000	Working status		
Household size*	0.0362	0.000	Reference week not working *	0.0852	0.000
Sex			Sector		
Male*	-0.0010	0.832	Mine*	0.3589	0.000
Age			Production*	0.3926	0.000
20-24 age*	0.1494	0.000	Build*	0.4295	0.000
25-29 age*	0.2934	0.000	Service*	0.4427	0.000
30-34 age*	0.1978	0.000	Not working*	0.4677	0.000
35-39 age*	0.2332	0.000	Year		
40-44 age*	0.2265	0.000	2015*	-0.0159	0.008
45-49 age*	0.2919	0.000	2016*	-0.0108	0.071
50-54 age*	0.3059	0.000	2017*	0.0018	0.764
55-59 age*	0.3444	0.000	Gross domestic product of İzmir per capita		1.527***
60-64 age*	0.3269	0.000	Labor force participation rate of İzmir		0.040**
Over age 65*	0.3378	0.000	trend		-0.207***
Education					
School has not finish*	0.1105	0.000			
secondary school graduate*	0.0390	0.000			
High school graduate*	0.0712	0.000			
Graduate of College or Faculty*	0.1952	0.000			
Graduate or Ph.D.*	0.2949	0.000			
Marital status					
Married*	0.1567	0.000			
Divorced*	0.0858	0.000			
Wife died*	0.0877	0.000			

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Table 3 shows the marginal impact results of the logit model, which gives the possibility of migration to Izmir. The outputs of the two models are shown in a single table. Model results have been interpreted according to marginal effects.

3. Conclusion

When the educational status of individuals is examined, it is seen that the probability of migration is high in educated individuals. Individuals with graduate or doctorate degrees are 29% more likely to migrate than primary school graduates and 19% more likely to graduate from college or faculty. When we evaluate the age and education level jointly, individuals who are university graduates, graduate or doctorate graduates fall in the 25-29 age range. There is a linear relationship between education and the possibility of migration. According to this relationship, individuals' tendency to migrate will increase with the educational level. Individuals with high levels of education are relatively more active, intelligent, more alert and adaptable to changing opportunities (Sahota, 1968: 225). It is seen that the increase of İzmir's per capita gross domestic product increases the probability of migration to İzmir by 38%. Labor force participation rate leads to an increase of 1% on the probability of immigration to İzmir. When we look at the year variable, which is the trend variable in Model 2, the probability of emigration to İzmir decreases by 5% compared to years. According to TÜİK data, İzmir is a province receiving immigration and increasing the rate of receiving immigration. However, there was a decrease in the rate of net migration as it began to emigrate as well as receive migration. As it is seen in the studies discussed, İzmir province is known to have a large portion of the immigrant population from the eastern provinces. Due to the weak education system in the eastern provinces, this means unskilled labor for İzmir. It is also seen in the model results that Izmir is preferred by educated individuals. The city of İzmir will be able to turn internal migration into favor when it creates new job opportunities by combining the unqualified labor force and educated individuals on a common ground.

Kaynakça

- Astar, M. (2009). The Validation of Taylor Rule in OECD Countries with Logit Models (Unpublished Master's Thesis).
- Bostan, H. (2017). Social Structure Changes Caused by Internal Migration in Turkey, Problems and Solutions Caused. *Journal of Geography*. 35(2017): 1-16.
- Gujarati, D. N. (1999). *Basic Econometrics*. New York: McGraw-Hill Companies.
- Gujarati, D. N. (2001). *Basic Econometrics*. New York: McGraw-Hill Companies.
- Işık, Ş. (1999). Geographical Dimensions of Migrations to Izmir. *Turkish Journal of Geography*.34: 383-405.
- Işık, Ş. (2009). Migrations to İzmir in 1995-2000 Period. *Turkish Journal of Geography*.52: 9-16.
- Işık, Ş. (2017). How has Izmir's Position in Inter-Provincial Migrations Changed at the Beginning of the 21st Century. *Aegean Geography Journal*. .26(1):1-19.

Sahota, G. S. (1968). An Economic Analysis of Internal Migration in Brazil. *Journal of Political Economy*, 76(2), 218-245.

Wang, G. C. ve Jain, C. L. (2003). *Regression Analysis: Modeling & Forecasting*. Institute of Business Forec.

O-73 M- Estimation Use Pearson Type IV Distribution Weight Function in Robust Regression

Yasin BÜYÜKKÖR^{1*}, Hatem ÇOBAN² and Ali Kemal ŞEHİRLİOĞLU³

^{1*}*Faculty of Economic and Administrative Sciences, Karamanoglu Mehmetbey University, Turkey, yasinbuyukkor@hotmail.com*

²*Faculty of Economic and Administrative Sciences, Dokuz Eylul University, Turkey, hatem.coban@deu.edu.tr*

³*Faculty of Economic and Administrative Sciences, Dokuz Eylul University, Turkey, kemal.sehirli@deu.edu.tr*

Abstract – In many regression applications, the distribution of errors is considered normal and the Least Squares (OLS) method is used to estimate parameters. However, in practice, even if the distribution of errors is assumed to be normal, residuals are not generally normally distributed. If the data contains outliers or there are observations which suspected to be outlier, the assumption of normality is violated and parameter estimates are biased. Many researchers use robust methods when such problems occur. One of these methods is M-estimators. Traditional M-estimators can easily be used when the data is symmetrical. However, traditional M-estimators can not achieve a good solution if the data has skewness and excess kurtosis. The differential equation of the Pearson Type IV (PIV) distribution, which provides a better solution for both symmetric and asymmetric distributions, can be used as the Influence Function (IF). The commonly used M-estimators do not take into account the skewness and kurtosis measures of the data, while the PIV distribution used in the study contains the skewness and kurtosis parameters. In this study, it is shown that Pearson differential equation can be used as Influence Function. By using the probability density function of PIV distribution, the Objective Function, Influence Function and Weight functions are obtained. For parameter estimation, Iteratively Re-Weighted Least Squares Estimation (IRWLS) method is used and parameter estimations are made on many real data sets. In addition, simulation studies with different scenarios are performed. The method used is compared with the performance of other M-estimators.

Keywords – M- Estimation, Robust Regression, Pearson Type IV Distribution, Iteratively Re-Weighted Least Squares

1. Introduction

In many social sciences studies, researchers make the assumption that the data set has a Gaussian (Normal) distribution. However, when data contains extreme values (outliers) or has skewness, the distribution of the data set is generally different from the normal distribution. When the assumption of normality does not meet, OLS estimators of regression coefficient will be biased (Hampel et al., 1986). The anomalies (skewness and excess kurtosis) in the data set can be caused by many reasons. Measurement and recording errors are the most important

reasons. However, in some cases, natural observation of the data set may also act as outliers and cause skewness or kurtosis. If the analysis of these observation is not performed carefully, the observations of the data set may be excluded from the analysis. Using OLS method, especially when estimating parameters in regression analysis, will result biased parameter estimates. Therefore, many researchers use robust estimating methods when data has anomalies. The most popular robust estimation method is M-estimators developed by Huber (1964). M- Estimators are an extended and robust version of the Maximum Likelihood (ML) estimation method.

In this study, M-estimators theory will be introduced in section 2 and most commonly used robust regression estimators Huber M- Estimators and Tukey’s Bisquare (Biweight) M- Estimators will be introduced. Section 3 Pearson Differential Equation and it’s objective, influence and weight functions will be introduced. In the last section we generated data sets which all have different scenarios and compared them bias and MSE criteria.

2. M- Estimators

The OLS estimating method is obtained by minimizing the likelihood function under the assumption that the distribution of errors is normal. Based on this idea, M-estimators using different distributions or functions when the distribution of errors is different from the normal distribution (skewed, long-tailed, excess kurtosis, etc.). M- estimators and OLS are Maximum Likelihood type estimators (Stuart, 2011; Rousseeuw and Leroy, 1987; Andersen 2008). The ML estimate of the regression parameter vector θ , can be written as

$$\prod f(\varepsilon_i) = \prod f(y_i - x_i^T \theta) \tag{1}$$

$f(\varepsilon)$ is the probability density function (pdf) of errors. If the errors has normal distribution one can minimize $\sum r_i^2 = \sum (y_i - x_i^T \theta)^2$. However, if the distribution of errors differs from the normal distribution, the function that should be minimized or maximized changed. M-estimators can be showed that

$$\sum \rho(\varepsilon_i) = \sum \rho(y_i - x_i^T \theta) \tag{2}$$

$$\rho(\varepsilon) = -\ln(f(\varepsilon)) \tag{3}$$

where $\rho(\varepsilon)$ is objective function and continuous and differentiable. The function which should be minimize

$$\min \sum \rho(r_i) = \min \sum \rho\left(\frac{\varepsilon_i}{s}\right) \tag{4}$$

s is the estimation of standart deviation and can be written as $s = \frac{MAD}{0.6745} = \frac{\text{medyan}|r_i - \text{medyan}(r_i)|}{0.6745}$ (Draper and Smith 2014). MAD is Median Absolute Deviation and 0.6745 is used if the sample actually from normal distribution (Hogg, 1979). For the estimating regression parameters, taking partial derivatives of $\rho(\varepsilon_i)$ with respect to β and setting them zero, one can easily obtain the system of normal equations

$$\sum x_j \psi \left(\frac{y_i - x_i^T \theta}{s} \right) = 0. \quad (5)$$

$\psi(r)$ is the Influence Function (IF) and $\psi(r) = \partial \rho(r) / \partial r$. The weight function is (Beaton and Tukey, 1974)

$$w(r) = \frac{\psi(r)}{r}. \quad (6)$$

2.1 Huber M- Estimator

Most popular M- estimator is Huber (1964) M- Estimator and objective, influence and weight functions defined as follows respectively,

$$\rho(r) = \begin{cases} \frac{1}{2}r^2 & ,|r| < k \\ k|r| - \frac{1}{2}k^2 & ,|r| \geq k \end{cases}, \quad \psi(r) = \begin{cases} r & ,|r| < k \\ k \text{sign}(r) & ,|r| \geq k \end{cases}, \quad w(r) = \begin{cases} 1 & ,|r| < k \\ \frac{k}{|r|} & ,|r| \geq k \end{cases} \quad (7,8,9)$$

where k is tuning constant and usually default value 1.345 (2 MAD) for efficiency under normal distribution. Huber M- estimator act like normal distribution in the middle of distribution and double exponential in the tails.

2.2 Tukey’s Bisquare M- Estimator

Tukey’s Bisquare objective, influence and weight function defined as follows

$$\rho(r) = \begin{cases} \frac{k^2}{6} \left\{ 1 - \left[1 - \left(\frac{r}{k} \right)^2 \right]^3 \right\} & ,|r| < k \\ \frac{k^2}{6} & ,|r| \geq k \end{cases}, \quad \psi(r) = \begin{cases} r \left[1 - \left(\frac{r}{k} \right)^2 \right]^2 & ,|r| < k \\ 0 & ,|r| \geq k \end{cases}, \quad (10,11,12)$$

$$w(r) = \begin{cases} \left[1 - \left(\frac{r}{k} \right)^2 \right]^2 & ,|r| < k \\ 0 & ,|r| \geq k \end{cases}$$

where k is tuning constant and usually default value 4.685 (7 MAD) for efficiency under normal distribution (Beaton and Tukey, 1974).

3 Pearson Distribution System

A Pearson Density $f(x)$ is defined to be any valid solution to the differential equation (Pearson, 1895):

$$\frac{1}{f(x)} \frac{df(x)}{dx} = \frac{d \ln f(x)}{dx} = \frac{f'(x)}{f(x)} = \frac{x - a}{b_0 + b_1 x + b_2 x^2} \quad (13)$$

The solutions to this equation define a family, which is called the *Pearson Distribution Family*. It is assumed that the function $H(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \dots$ in the denominator of

the differential equation is extended to the Maclaurin series. The main reason for this assumption allows the use of the method of moments. In the equation, the product of cross-multiplying and multiplied by both sides x^n consecutive moment equation is obtained as follows;

$$a\mu'_n + nb_0\mu'_{n-1} + (n+1)b_1\mu'_n + (n+2)b_2\mu'_{n+1} + (n+3)b_3\mu'_{n+2} + \dots \quad (14)$$

If consecutive moment equation is used for $H(x) = b_0 + b_1x + b_2x^2$ and $n = 0, 1, 2, 3, 4$, the parameters can be obtained in origin moments. See more details Şehirlioğlu & Dündar, 2014. The relationship between the moments with respect to the origin and the central moments is used and when the average of the dispersion is shifted to the origin;

$$b_1 - a = 0, \quad b_0 + (3b_2 + 1)\mu_2 = 0, \quad (15)$$

$$(3b_1 - a)\mu_2 + (4b_2 + 1)\mu_3 = 0, \quad 3b_0\mu_2 + (4b_1 - a)\mu_3 + (5b_2 + 1)\mu_4 = 0 \quad (16)$$

is obtained (Şehirlioğlu & Dündar, 2014). If $Q = 10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3$, $Q' = 10\beta_2 - 12\beta_1 - 18$ and $\beta_1 = \mu_3^2/\mu_2^3$ (skewness), $\beta_2 = \mu_4/\mu_2^2$ (kurtosis) a, b_0, b_1, b_2 parameters as follows

$$b_1 = a = -\frac{\mu_3(\mu_4 + 3\mu_2^2)}{Q} = -\frac{\sigma\sqrt{\beta_1}(\beta_1 + 3)}{Q'} \quad (17)$$

$$b_0 = -\frac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{Q} = -\frac{\sigma^2(4\beta_2 - 3\beta_1)}{Q'} \quad (18)$$

$$b_2 = -\frac{(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)}{Q} = -\frac{(2\beta_2 - 3\beta_1 - 6)}{Q'} \quad (19)$$

It is the $H(x)$ function in the denominator that determines the distribution types in Pearson system. When the roots of this function are folded or single, they form the transition types and, if they have two roots (real or complex), they constitute the main types. The main types are Type-I when the roots are real and different signed, and Type-VI if the roots are real and the same sign, and Type-IV if the roots are complex. The differential equation in equation (1.1) has different types. More details Friori and Zenga (2009), Xi et al. (2012), Nagahara (2008), Elderton, (1953).

3.1 Pearson Type IV Distribution

Pearson Type IV (PIV) distribution has the following pdf;

$$f(x) = K[(x+r)^2 + s^2]^{-m} \exp\left[-v \arctan\left(\frac{x+r}{s}\right)\right], \quad -\infty < x < \infty \quad (20)$$

where K is the constant to be sure $f(x)$ is a pdf, $m = 1/2b_2$, $v = c/b_2$, $c = a + r/s$, $r = \text{real}(r_1)$, $s = \text{imag}(r_1)$ and r_1 and r_2 root of differential equation and complex. If considering similarity between IF and Pearson Differential Equation;

$$\frac{d\rho(x)}{dx} = \frac{d(-\ln(f(x)))}{dx}, \quad \psi(x) = -\frac{f'(x)}{f(x)} \quad (21,22)$$

Equation (21) shows that the Pearson differential equation can be used as a influence function (Dzhun, 2011). If constant K removed from the equation (20) objective, influence and weight function of PIV distribution (Wisniewski, 2014)

$$f(x) \propto [(x+r)^2 + s^2]^{-m} \exp\left(-v \arctan\left(\frac{x+r}{s}\right)\right) \quad (23)$$

$$\rho(x) = -\ln f(x) = m \ln((x+r)^2 + s^2) + v \arctan\left(\frac{x+r}{s}\right) \quad (24)$$

$$\psi(x) = \frac{2m(x+r) + vs^2}{(x+r)^2 + s^2}, \quad w(x) = \frac{2m(x+r) + vs^2}{x((x+r)^2 + s^2)} \quad (25)$$

is obtained. As can be seen that Huber and Tukey’s M- estimator doesn’t contain skewness and kurtosis parameter. Figure 1-3 shows different skewness and kurtosis values of PIV distribution objective, influence and weight functions. If the data differs from normal (skewed and excess kurtosis) PIV distribution can be used for estimating regression parameters.

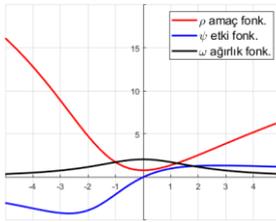


Figure 1. $\beta_1 = 0.8, \beta_2 = 6$

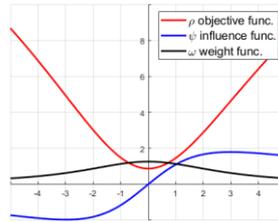


Figure 2. $\beta_1 = 0.25, \beta_2 = 4$

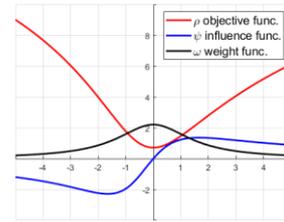


Figure 3. $\beta_1 = 2, \beta_2 = 8$

4 Simulation Study

In this section we consider different data sets to comparison of M-estimating methods. These methods are:

1. Ordinary Least Squares
2. Huber M- Estimator
3. Tukey’s Bisquare M- Estimator
4. Pearson Type IV Weight Function.

Data are generated simple linear regression model:

$$Y = 2 + 2X + \varepsilon$$

$X \sim N(0, 1)$ and true regression parameters $\theta_0 = \theta_1 = 2$. For the error (ε) has following scenarios:

1. $\varepsilon \sim N(0, 1)$, standart normal distribution
2. $\varepsilon \sim N(1, 1) + 0.2N(2, 1)$, contaminated normal distribution with different mean (Khan, 2016)
3. $\varepsilon \sim \Gamma(1, 0.5)$, gamma distribution with skewness and kurtosis (Mohebbi et al, 2007)
4. $\varepsilon \sim N(0, 1) + 0.2\chi^2(5)$ normal distribution with Chi Squared contamination for skewness and kurtosis (Xu and Chen, 2018)

In each scenario we choosed skewness greater than 0 ($\beta_1 > 0$) and excess kurtosis ($\beta_2 > 3$). Also 1000 replications were simulated and used sample sizes $n=30$ and $n=100$. In each scenario we generated samples and choose which can be suitable for PIV. For the comparison of regression coefficients Mean Squared Error (MSE) and Absolute Bias criteria used. Two following criterias:

$$MSE = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\theta}_i - \theta)^T (\hat{\theta}_i - \theta), \quad Bias = \left| \frac{\sum_{i=1}^n \hat{\theta}_i}{1000} - \theta \right| \quad (32)$$

For the parameter estimation Iteratively Re- Weighted Least Squares (IRWLS) estimation method was used. The steps of IRWLS: see more details Andersen (2008). According to the Scenario 1 results (Table 1), if errors normally distributed OLS estimates achieves best performance. As the sample size increases value of both MSE and Bias decrease. In scenario 2 (Table 2), two normal distribution different mean and same standart deviation (contaminated normal), with the increases skewness and kurtosis, PIV has better performance among others. In scenario 3 (Table 3), if errors has Gamma distribution with $\alpha = 1$ and $\gamma = 0.5$, due to increasing skewness and excess kurtosis PIV has better both sample size. In last scenario (Table 4), 80 percent of data has standart normal distribution and 20 percent of data has Chi Squared distribution (contaminated distribution), also skewness and excess kurtosis PIV has better both sample size.

Table 1. Scenario 1

Sample Size	Method	Mean θ_0	Mean θ_1	Bias θ_0	Bias θ_1	MSE
n=30	OLS	2,0165	2,0018	0,1534	0,1258	0,0617
	Huber	2,0167	2,0019	0,1559	0,1359	0,0669
	Tukey	2,0171	2,0019	0,1610	0,1474	0,0743
	Type IV	1,9608	2,0025	0,2080	0,1602	0,1092
n=100	OLS	2,0009	2,0031	0,0780	0,0704	0,0175
	Huber	1,9992	2,0041	0,0812	0,0740	0,0191
	Tukey	1,9983	2,0049	0,0840	0,0767	0,0204
	Type IV	1,9805	2,0065	0,1652	0,0925	0,0573

Table 2. Scenario 2

Sample Size	Method	Mean θ_0	Mean θ_1	Bias θ_0	Bias θ_1	MSE
n=30	OLS	2,4146	1,9983	0,4160	0,1694	0,2479
	Huber	2,3826	1,9966	0,3859	0,1809	0,2327
	Tukey	2,3642	1,9953	0,3694	0,1934	0,2311
	Type IV	2,2852	1,9991	0,3291	0,1972	0,2245
n=100	OLS	2,4627	1,9984	0,4630	0,2335	0,3335
	Huber	2,2585	1,9984	0,2803	0,2058	0,1748
	Tukey	2,1737	1,9977	0,2283	0,1983	0,1403
	Type IV	2,0069	1,9981	0,2174	0,1980	0,1397

Table 3. Scenario 3

Sample Size	Method	Mean θ_0	Mean θ_1	Bias θ_0	Bias θ_1	MSE
n=30	OLS	2,5061	1,9974	0,5061	0,1405	0,2913
	Huber	2,4683	1,9973	0,4683	0,1135	0,2446
	Tukey	2,4458	1,9972	0,4458	0,0952	0,2193
	Type IV	2,3446	2,0007	0,3446	0,0831	0,1393
n=100	OLS	2,5007	1,9972	0,5007	0,0452	0,2563
	Huber	2,4583	1,9978	0,4583	0,0363	0,2143
	Tukey	2,4502	1,9979	0,4502	0,0359	0,2068
	Type IV	2,2790	1,9991	0,2790	0,0295	0,0811

Table 4. Scenario 4

Sample Size	Method	Mean	Mean	Bias	Bias	MSE
		θ_0	θ_1	θ_0	θ_1	
n=30	OLS	2,7331	1,9959	0,7336	0,4047	0,8879
	Huber	2,5631	1,9982	0,5644	0,2981	0,5108
	Tukey	2,4566	2,0040	0,4603	0,2450	0,3540
	Type IV	2,2067	2,0097	0,2968	0,2399	0,2383
n=100	OLS	2,7413	1,9982	0,7413	0,1899	0,6223
	Huber	2,5008	2,0012	0,5008	0,1351	0,2926
	Tukey	2,4155	2,0028	0,4155	0,1180	0,2087
	Type IV	2,0090	2,0045	0,1290	0,1109	0,0452

3. Conclusion

In this study, compared of performance of different estimation methods. With using Pearson Differential Equation as a Influence function, Objective, Influence and Weight function of PIV distribution obtained. If data has skewness and excess kurtosis PIV distribution can be used for estimating regression parameters.

Acknowledgment

The authors would like to thank congress organizers.

References

- Andersen, R. (2008). *Modern methods for robust regression* (No. 152). Sage.
- Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2), 147-185
- Draper, N. R., & Smith, H. (2014). *Applied regression analysis*(Vol. 326). John Wiley & Sons.
- Dzhun', I. V. (2011). Method for diagnostics of mathematical models in theoretical astronomy and astrometry. *Kinematics and Physics of Celestial Bodies*, 27, 260-264.
- Elderton, W. P. (1953). *Frequency Curves and Correlation* Cambridge University. *New York*.
- Fiori, A. M., & Zenga, M. (2009). Karl Pearson and the origin of kurtosis. *International Statistical Review*, 77(1), 40-50.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence functions. Wiley, New York

Hogg, R. V. (1979). Statistical robustness: One view of its use in applications today. *The American Statistician*, 33(3), 108-115.

Huber, Peter J. Robust Estimation of a Location Parameter. *Ann. Math. Statist.* 35 (1964), no. 1, 73--101. doi:10.1214/aoms/1177703732.

Khan, Sajjad Ahmad. (2016). A Comparative Study of Three Improved Robust Regression Procedures. *Pakistan Journal of Statistics.* 32. 425 - 441.

Mohebbi, M., Nourijelyani, K., & Zeraati, H. (2007). A simulation study on robust alternatives of least squares regression. *Journal of Applied Sciences*, 7(22), 3469-3476.

Nagahara, Y. (2008). A method of calculating the downside risk by multivariate nonnormal distributions. *Asia-Pacific Financial Markets*, 15(3-4), 175-184.

Pearson, K. (1895). X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186), 343-414.

Pearson, K. (1901). XI. Mathematical contributions to the theory of evolution.—X. Supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 197(287-299), 443-459

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection* (Vol. 1). New York: Wiley.

Stuart, C. (2011). Robust regression. *Department of Mathematical Sciences, Durham University*, 169.

Şehirlioğlu, A.K. (2011). Pearson Dağılım Ailesi (Yayınlanmamış Ders Notları). İzmir: Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi.

Wiśniewski, Z. (2014). M-estimation with probabilistic models of geodetic observations. *Journal of Geodesy*, 88(10), 941-957.

Xu, X., & Chen, X. (2018). A practical method of robust estimation in case of asymmetry. *Journal of Statistical Theory and Practice*, 12(2), 370-396..

O-76 Quasi-Maximum Likelihood Estimator based on Moyal Distribution for Censored Data

Ismail YENILMEZ¹, İlhan USTA², Yeliz MERT KANTAR^{3*}

¹Department of Statistics, Eskisehir Technical University, Turkey,
ismailyenilmez@eskisehir.edu.tr

² Department of Statistics, Eskisehir Technical University, Turkey, iusta@eskisehir.edu.tr

³ Department of Statistics, Eskisehir Technical University, Turkey, ymert@eskisehir.edu.tr

Abstract – The censored data, in which the observed value of some variable is partially known, is related to data-gathering mechanism. In the context of regression analysis, while the ordinary least squares (OLS) is the simplest and most commonly used estimator for full data, Tobit estimator is one of the well-accepted estimation methods for censored data. Also, it is known that OLS method produce a *biased and inconsistent* estimator in the case of censored data. Tobit is the maximum likelihood (ML) estimation under a normal error distribution and it is consistent and effective if the assumption of normality is maintained. However, many examples can be presented where the assumption of normality is not satisfied. At this stage, one of the various alternative methods is the quasi-maximum likelihood estimators (Q-ML). In this study, the Q-ML based on Moyal Distribution (MD) is introduced for censored data. The simulation results show that if the errors are not normal, Q-ML based on MD has a smaller bias and mean square error (MSE) value than classical Tobit and OLS.

Keywords – Censored data, Tobit, Quasi-Maximum Likelihood estimator, Moyal Distribution

1. Introduction

In general, limited dependent variable (LDV) is defined as a dependent variable whose range of values is substantively restricted by Wooldridge (2013). Binary dependent variables, discrete response variables, positive variables, corner solution response, nonnegative integer values can be seen as LDV. The limitation of the variables within the definition ranges due to certain stochastic selection mechanisms and the frequent use of qualitative variables with dummy variables in econometric models expand the use of models with limited dependent variables. Maddala (1983) classifies the models with limited dependent variables into three different categories (Truncated regression, Censored regression and Dummy endogenous models).

According to Davidson and MacKinnon (1999), if there is loss of information in the dependent variable, data is censored data. If there is loss of information for both dependent

and independent variables, truncated data arises. Generally, there is no systematic limitation for the censored data set since censorship occurs due to the nature of the data or data-gathering mechanism. This study deals with censored data and estimation of regression model in the case of censored data.

In the case of censored data, various estimators and their assumptions are presented by Tobin (1958), Lee and Trost (1978), Maddala (1983), Paarsch (1984), Mroz (1987), Cohen (1991), Breen (1996), Sigelman and Zeng (1999), Park (2003) and Greene (2011). It is known that the ordinary least squares (OLS) yields biased and inconsistent estimates at different levels of censoring. Tobit is the maximum likelihood (ML) method under the assumption that the errors are based on the normal distribution. Tobit is known to be consistent and effective if the assumption of normality is satisfied. However, in practice, error term is often observed to be non-normal distribution in many real life examples. In such cases, new estimation methods that are less sensitive to the assumption of normality have been proposed in the literature. At this stage, one of several alternative methods is the Quasi-Maximum Likelihood (Q-ML) estimators. Q-ML estimators, also known as partial adaptable (PAE) estimators, have recently been used by many researchers, McDonald and Xu (1996), Caudill (2012), Lewis and McDonald (2014), McDonald and Nguyen (2015), for regression model in the case of censored data. Q-MLEs based on some flexible distributions (generalized- t , Box-Tiao and exponential generalized beta of the second kind) have been compared by McDonald and Xu (1996) with other estimators for censored regression in case of skewed and leptokurtic error distributions. Q-MLE based on location-scale mixture of normal distribution has been introduced by Caudill (2012). Lewis and McDonald (2014) add new distributions (skewed generalized error, inverse hyperbolic sine etc.) to distributions used in McDonald and Xu (1996) and present a more comprehensive study. The PAE approach is extended to accommodate possible heteroscedasticity as well as non-normality by McDonald and Nguyen (2015). In addition, Martínez-Flórez, et al. (2013) introduce an extension of Tobit model called as the Alpha-power Tobit Model. Karlsson and Laitila (2014) suggest a finite mixture of Tobit models for estimation of regression models with a censored response variable. Yenilmez and Kantar (2017) and Yenilmez et al. (2018) have proposed PAEs for censored regression using generalized normal distribution and generalized logistic distribution, respectively.

The superiority of Tobit's estimator over the others is clearly visible if assumptions are met. However, as stated in the main motivation of the study, the violation of certain assumptions, such as the non-normal errors, give rise to modifications and extensions of the estimation method. In this study, the Q-ML based on Moyal distribution (MD) is introduced for censored regression model. As in other studies in the literature, the proposed new modification for censored regression is compared with the OLS and Tobit.

The paper is organized as follows: Section 2 presents OLS and Tobit estimators. Section 3 briefly reviews Moyal distribution and general framework of Q-ML and also introduces Q-

ML based on Moyal distribution (QMLMO) for the censored dependent variable. Section 4 contains the simulation study carried out under the scope of the study. Finally, the obtained results are presented in conclusion section.

2. Conventional Estimation of Censored Regression

The censored regression model is defined as

$$Y_i = \max(c, X_i\beta + \varepsilon_i) \quad (1)$$

where Y_i is the observed value of the dependent variable, c is the censoring point (in practice, this point is usually taken as 0) and error term ε_i is generally assumed to be normal distributed ($\varepsilon_i \sim N(0, \sigma^2)$).

OLS estimation procedure can be conducted for censored data. All data (uncensored and censored) are taken into account. The OLS is protect the integrity of the data in the analyses, however, it is showed by Greene (1981), Arabmazar and Schmidt (1982) that OLS yields biased and inconsistent estimates for the parameters of the censored regression model.

The ML estimation of the parameters in Eq. 1 gives the Tobit estimates. While the assumptions are met, the consistency of the Tobit estimator is shown by Amemiya (1973). Non-normality of error terms, one of the violations in the assumptions of an OLS causes inconsistency. The Tobit estimator is based on the maximization of the log-likelihood function. For lower (left) censorship, log-likelihood function is defined as

$$\ell(\beta, \sigma) = \sum_{Y_i \leq c} \ln \left(1 - \Phi \left(\frac{c - X_i\beta}{\sigma} \right) \right) + \sum_{Y_i > c} \ln \left(\frac{1}{\sigma} \phi \left(\frac{Y_i - X_i\beta}{\sigma} \right) \right) \quad (2)$$

where β and σ are unknown regression and distributional parameters, respectively. $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cumulative distribution function (CDF) and probability density function (PDF) corresponding to standard normal random variable, respectively. Alternative estimators are used in cases where assumptions of distribution of error terms are not satisfied. For this purpose, the Moyal Distribution (MD) and Q-ML estimator based on MD are presented in the next section.

3. Quasi-Maximum Likelihood Estimation

Quasi-Maximum likelihood (Q-ML) or partially adaptive estimation (PAE) is a type of estimation procedure that maximizes the likelihood function of a sample of size n form error distribution ($f(\cdot)$) to estimate both regression parameters (β) and distribution parameters (θ). The estimates of Q-ML are obtained by minimizing the following log-likelihood function:

$$\ell(\beta; \theta) = \sum_{i=0}^n \ln f(Y_i - X_i\beta; \theta) \quad (3)$$

where $f(\cdot)$ is a flexible PDF. For censored regression, Q-ML procedure can be defined based on following likelihood function

$$\ell(\beta; \theta) = \sum_{Y_i \leq c} \ln F(c - X_i \beta; \theta) + \sum_{Y_i > c} \ln f(Y_i - X_i \beta; \theta) \quad (4)$$

where $f(\cdot)$ and $F(\cdot)$ are a flexible PDF and CDF. MD is used in this study.

3.1 Moyal Distribution

Moyal Distribution (MD) is proposed by Moyal (1955) as an approximation to the Landau distribution. Landau and MD are initially used to model the energy loss of ionizing particles. There are modifications in the statistical studies in the following periods. The MD is a right-skewed distribution. 0 is the mode of MD. These features make the MD attractive for an analysis based on censored data at 0. Therefore, it can be used in the Q-ML procedure for censored regression. The PDF and CDF of standard MD is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x+\exp(-x))}{2}\right\} \quad -\infty < x < \infty \quad (5)$$

$$F(x) = 1 - \frac{1}{\sqrt{\pi}} \gamma\left(\frac{1}{2}, \frac{1}{2} \exp\{-x\}\right) = \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}, \frac{1}{2} \exp\{-x\}\right) \quad -\infty < x < \infty \quad (6)$$

where $\gamma(\alpha, \beta) = \int_0^\beta t^{\alpha-1} e^{-t} dt$ and $\Gamma(\alpha, \beta) = \int_\beta^\infty t^{\alpha-1} e^{-t} dt$ are lower and upper incomplete gamma functions, respectively. PDFs of standard MD are presented in Fig.1.

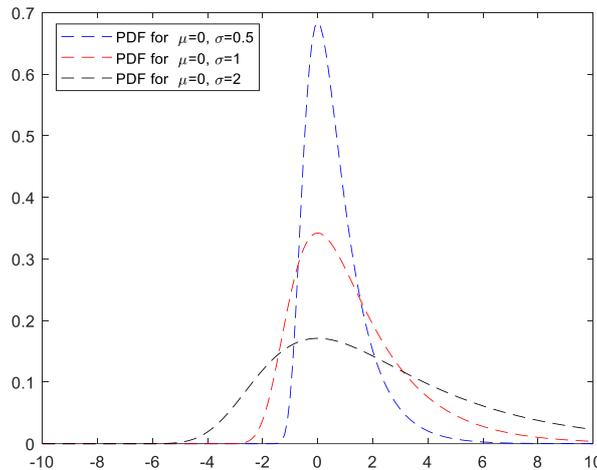


Figure 1. PDFs of MD

3.2 Quasi-Maximum Likelihood Estimation based on Moyal Distribution

If functions in Eq.5-6 are used in Eq.4, a Q-ML based on MD (Q-MLMD) is obtained for the censored data structure. For lower (left) censorship, the log-likelihood function based on the Moyal distribution is defined as

$$\ell(\beta; \theta) = -n_c \frac{\ln(\pi)}{2} + \sum_{Y_i \leq c} \ln \left(\left(\Gamma \left(\frac{1}{2}, \frac{1}{2} \exp\{-(c - X_i\beta)\} \right) \right) \right) - (n - n_c) \frac{\ln(2\pi)}{2} - \frac{1}{2} \sum_{Y_i > c} \ln \left(((Y_i - X_i\beta)) + \exp(-(Y_i - X_i\beta)) \right) \quad (7)$$

where n and n_c are sample of size and number of censored observations, respectively. By minimizing Eq. (7), the obtained estimates are called as Quasi-maximum likelihood estimates.

4. Simulation Results and Application

For comparing relative Bias and mean square error (MSE) of conventional estimators (OLS and Tobit) and proposed Q-MLMD estimator, a simulation study have been conducted under different error distributions (student- t distribution with different degrees of freedom (df) and scale-contaminated mixture-normal distribution with different percentiles). The sample size (n) is taken as 100, 250 and 500. 100000/ n data sets are generated. Censoring point is taken as 0. The linear regression model is defined as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n \quad (8)$$

where β_0 and β_1 is taken as 0 and 1, respectively. x is distributed as uniform distribution ($x_i \sim U(0,1)$). The analysis results of the Student- t distribution with different dfs are provided in Table 1.

Table 1. Simulation results for censored regression under Student- t error distribution with different df

	Slope $b_1 = 1$	$y = b_0 + b_1 x$					
		n=100		n=250		n=500	
		Bias	MSE	Bias	MSE	Bias	MSE
Student- t $df = 1.5$	OLS	1,4280	14,8829	0.4355	0.4444	0.5881	0,4130
	TOBIT	2,9814	37,1600	0.1568	0.7011	0,2187	0,2834
	Q-MLMD	-0.7117	7,0438	-0.1276	0.1702	-0,0798	0,1670
Student- t $df = 3$	OLS	0.4098	0.2992	0.3460	0.1756	0.3508	0.1575
	TOBIT	0.1582	0.2877	0.0048	0.1011	-0.0687	0.0599
	Q-MLMD	0.0895	0.1754	0.1055	0.0688	0.0360	0.0528
Student- t $df = 6$	OLS	0.4858	0.3181	0.3725	0.1569	0.3081	0.1292
	TOBIT	-0.0453	0.3167	-0.4478	0.2164	-0.0583	0.0320
	Q-MLMD	0.3722	0.2991	-0.1727	0.1413	0.0455	0.0243

It is seen from Table 1, according to Bias and MSE, Q-MLMD performs better than conventional estimators for all sample sizes. However, As the degree of freedom increases, the improvement in the Tobit estimator’s performance is remarkable. This is an expected result due to convergence between distributions. If df approaches infinity then the t -distribution approaches the normal distribution. In fact, 30-50 degrees of freedom is sufficient for handling the t -distribution as the normal distribution.

Table 2. Simulation results for censored regression under Student- t with $df=30$ and standard normal distribution

		Slope $b_1 = 1$		$y = b_0 + b_1x$			
		n=100		n=250		n=500	
		Bias	MSE	Bias	MSE	Bias	MSE
Student- t $df=30$	OLS	0.3893	0.1890	0.4018	0.1860	0.3044	0.1091
	TOBIT	0.1129	0.0868	0.1108	0.0420	0.0016	0.0283
	Q-MLMD	0.2383	0.1644	0.1497	0.1292	0.1397	0.0534
SND	OLS	0.4200	0.25711	0.3384	0.1452	0,2906	0,0972
	TOBIT	0.1779	0.1814	0.0255	0.0610	0,0067	0,0276
	Q-MLMD	0.3382	0.2631	0.0964	0.0662	0,1263	0,0352

In addition, in Table 2, the analysis results obtained in case of error distributed as Student- t with $df=30$ and standard normal distribution is presented. So the improvement in the Tobit estimator has become superior in the case of convergence to normal distribution. As sample sizes increase, the obtained estimates are close to the truth value of parameter as seen in Table 2.

The results obtained for the mixture-normal distribution with different percentiles (80%, 50%, 20%) are shown in Table 3. According to these results, the superiority of the proposed estimator Q-MLMD is again striking.

Table 3. Simulation results for censored regression under mixture-normal distribution with different percentiles

	Slope $b_1 = 1$	$y = b_0 + b_1x$					
		n=100		n=250		n=500	
		Bias	MSE	Bias	MSE	Bias	MSE
MixNrm-80%	OLS	0.2195	0.1459	0.1016	0.0536	0.1592	0.0490
	TOBIT	0.0564	0.2011	-0.0366	0.0978	-0.0594	0.0442
	Q-MLMD	0.0306	0.0945	0.0550	0.0399	0.0213	0.0177
MixNrm-50%	OLS	0.3928	0.3545	0.2393	0.1853	0.2258	0.1065
	TOBIT	0.1246	0.4503	0.0594	0.3025	-0.0856	0.0890
	Q-MLMD	0.0063	0.2421	0.0344	0.1073	-0.0091	0.0764
MixNrm-20%	OLS	0.4358	0.6059	0.4278	0.3306	0.4687	0.3081
	TOBIT	0.0998	1,0434	0.0399	0.3781	0.1426	0.2739
	Q-MLMD	0.0855	0.8743	0.0574	0.2032	0.1333	0.2425

5. Conclusion

Conventional estimators (OLS and Tobit) and Q-MLMD are compared with the simulation study. Different distribution assumptions (including non-normality) and sample sizes are used for censored regression within the framework of simulation. In contrast to Tobit's inconsistent results in the case of non-normal error distribution and the inconsistent and biased results of OLS, the success of the proposed estimator Q-MLMD is remarkable. Q-MLMD is a good alternative that can be used for censored regression in violation of the assumption. As a further study, this study will be expanded with Beta-Moyal distribution proposed by Carneiro et al. (2014), Beta Moyal-Slash distribution introduced by Genç et al. (2014) and Generalized log-Moyal distribution proposed by Bhati and Ravi (2018).

Acknowledgment

This study was supported by Eskisehir Technical University Scientific Research Projects Commission under the grant no: 19ADP093.

References

Amemiya, T. (1973). Regression analysis when the dependent variable is a truncated normal. *Econometrica* 41:997–1016

Arabmazar, A., Schmidt, P. (1982). An investigation of the robustness of the tobit estimator to non-normality. *Econometrica* 50:1055–1069.

Bhati, D. and Ravi, S. (2018). “On generalized log-Moyal distribution: A new heavy tailed size distribution”, *Insurance: Mathematic and Economics* 79, 247-259.

Breen, R. (1996). “Regression Models: Censored, Sample Selected or Truncated Data”, *Quantitative Applications in the Social Sciences*, 7-111.

Cordeiro, G., M., Nobre, J., S., Pescim, R., R., and Ortega, E., M., M. (2012). “The beta Moyal: A useful-skew distribution”, *International Journal of Research and Reviews in Applied Sciences*, 10, 171-192.

Caudill, S. B. (2012). “A Partially Adaptive Estimator for The Censored Regression Model Based on A Mixture of Normal Distributions”, *Stats Methods Appl*, 21:121-137

Cohen, A.C. (1991). *Truncated and Censored Samples: Theory and Applications*. NY: Taylor and Francis Group

Genç, A. A., Korkmaz, M. Ç. and Kuş, C. (2014). “The Beta Moyal-Slash Distribution”, *Journal of Selçuk University Natural and Applied Science*, 3(4): 88-104.

Greene, W. H.(1981).On the asymptotic bias of the ordinary least squares estimator of the tobit model.*Econometrica*49:505-513.

Greene, H.W. (2011). *Econometric Analysis*. Prentise Hall, (7), 1231.

Davidson, R. and MacKinnon J.G. (1999). *Econometric Theory*. USA: Oxford University Press.

Karlsson M., Laitila T. 2014. “Finite mixture modeling of censored regression models”, *Stat Papers* 55: 627

Lee, L. F. and Trost R. P. (1978). “Estimation of Some Limited Dependent Variable Models with Applications to Housing Demand”, *Journal of Econometrics*, 8, 357.

Lewis, R. A. and McDonald, J. B. (2014). “Partially Adaptive Estimation of the Censored Regression Model”, *Economic Reviews*, 33 (7), 732-750.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: CUP

Martínez-Flórez, G., Bolfarine, H. and Gómez, H. W. 2013. “The Alpha-power Tobit Model”, *Communications in Statistics Theory and Methods*, 42:4, 633-643,

McDonald, J. B. and Nguyen H. (2015). “Heteroscedasticity and Distributional Assumptions in The Censored Regression Model”, *Communications in Statistics-Simulation and Computations*, 44, 2151-2168.

McDonald, J. B., Xu Y. J. (1996). “A Comparison of Semi-parametric and Partially Adaptive Estimators of the Censored Regression Model with Possibly Skewed and Leptokurtic Error Distributions”, *Economics Letter*, 51(2), 153-159.

Moyal, J., E. (1955). “Theory of Ionization Fluctuation. The London”, *Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 46, 263-280.

Mroz, T. (1987). “The sensitivity of an empirical model of married women’s hours of work to economic and statistical assumptions”, *Econometrica* 55:765–799.

Paarsch, H. (1984). “A Monte Carlo Comparison of Estimators for Censored Regression Models. *Journal of Econometrics*”, 24, 197-213.

Park, S.Y. (2003). “Unbiasedness or Statistical Efficiency: Comparison between One-stage Tobit of MLE and Two-step Tobit of OLS”, *International Journal of Human Ecology*, 4 (2), 77-86.

Sigelman, L. And Zeng, L. (1999). “Analyzing Censored and Sample-Selected Data with Tobit and Heckit Models”, *The George Washington University*, December 16, WV002-05.

Tobin, J. (1958). “Estimation of Relationships for Limited Dependent Variables”, *Econometrica*, 26, 24-36.

Wooldridge, J. (2013). *Introductory Econometrics: A Modern Approach*, 5th ed. South-Western Cengage Learning.

Yenilmez, I., Kantar, Y.M. (2017). “A Partially Adaptive Estimator for the Censored Regression Model Based on Generalized Normal Distribution”, *Proceedings of the 3rd IRSYSC-2017*, Konya, Turkey.

Yenilmez, I., Kantar, Y.M., Acitas, S. (2018). “Estimation of Censored Regression Model in the case of Non-Normal Error”, *Sigma J. Eng and Nat Sci* 36 (2), 513-521.

O-86 A Golden Ratio Control Chart for Monitoring the Process Mean

Elif KOZAN¹ and Onur KÖKSOY²

¹ Department of Statistics/ Ege University, Turkey, elif.kozan@ege.edu.tr

² Department of Statistics/ Ege University, Turkey, onur.koksoy@ege.edu.tr

Abstract – The monitoring of the process mean is usually accomplished by the quality control charts. This work presents a new control chart based on the well-known “Golden ratio (GR)”. The GR control chart directly incorporates with all the information in the sequence of sample values by plotting the weighted values of the sample via the GR number and median. When the small shifts are important, the Shewart control chart may not be a good option for monitoring the process mean; however, the exponentially weighted moving average (EWMA) control chart is known to be effective in the literature. In this paper, GR control chart is compared with the Shewart and EWMA control charts over an example.

Keywords – golden ratio number, median, shewart control chart, EWMA, CUSUM

1. Introduction

Statistical methods play a vital role in quality improvement in manufacturing and service industries. Statistical process control (SPC) is an effective approach for improving product of a quality and saving production costs. A control chart is used for process monitoring and a detection tool for any out-of-control process situation. Since 1924, when Dr Shewart presented the first control chart, various control charts have been developed and widely applied as a primary tool in SPC. The Shewart’s chart has been widely used monitoring the process mean and it is quite effective in detecting large scale shifts in the location but known as insensitive for small/moderate shifts. The ability of detecting small shifts can be improved by using a chart based on a statistic that incorporates information from past samples in addition to current samples. One such chart is the cumulative sum (cusum) control chart developed by Page (1954, 1961). Some authors, for example, Duncan (1974), Lucas (1976), Hawkins (1981), Lucas and Saccucci (1982a, 1990) stated that the cusum control chart is much more efficient than the usual \bar{x} control chart for detecting smaller variations in the mean. In 1959, EWMA control chart was proposed by Roberts. Cusum and EWMA charts are much more effective than the Shewart chart in detecting small and moderate-sized sustained shifts (Woodall, 2000). Both of these charts are based on an adaptive nature (i.e, the memory) and perform better than the Shewart chart while detecting the smaller shifts in the location.

In this paper, the GR control chart is proposed. The novel Golden ratio number, i.e., 1.618034 and usually symbolized by ϕ , has been used in numerous applications which range from a description of plant growth and crystallographic structure of certain solids to the development of computer algorithms for searching data bases (Dunlap, 1997). Günver et al. (2018)

introduced the weighting coefficient (M_{c_i}), which was calculated by the formula in Equation (1) via the GR number and shown in Figure 1.

$$M_{c_i} = \begin{cases} \text{if } X_i < \text{Med} & \frac{1}{\varphi} + 2 * \frac{i-1}{n-1} \\ \text{else} & 1 + \varphi - 2 * \frac{i-1}{n-1} \end{cases} \quad (1)$$

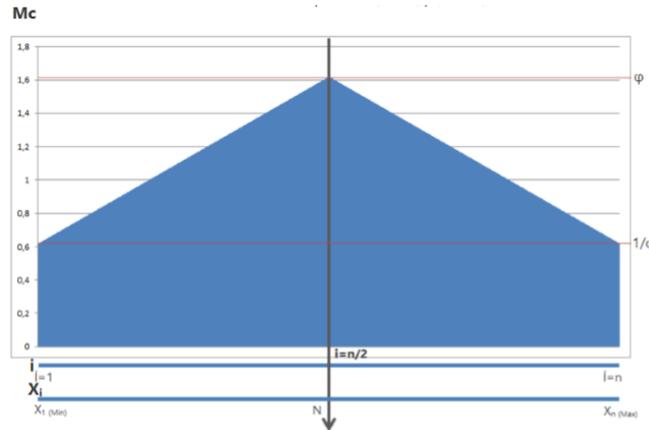


Figure 1. The Mean coefficient mask

The calculated coefficients are used to calculate the weighted distance from median (wdfm) is shown in Table (1). This weighting scheme assumes more weight to values close to the median (Günver et al. (2018)).

Table 1. Wdfm in sorted data

index	Data (X_i)	Wdfm
1	X_1	$M_{c_1} * (X_1 - \text{Med})$
2	X_2	$M_{c_2} * (X_2 - \text{Med})$
3	X_3	$M_{c_3} * (X_3 - \text{Med})$
...
n-2	X_{n-2}	$M_{c_{n-2}} * (X_{n-2} - \text{Med})$
n-1	X_{n-1}	$M_{c_{n-1}} * (X_{n-1} - \text{Med})$
n	X_n	$M_{c_n} * (X_n - \text{Med})$

2. The Proposed Control Chart Based on Golden Ratio for Detecting a Shift in the Process Mean

When calculating the average for \bar{x} chart, all data are given equal weights. However, this may not be an ideal approach for detecting a shift in the process mean. Moreover, the contribution of the extreme values needs to be promoted by lower weights and the other values which close to median needs to be promoted by higher weights. (Günver et al., 2018). And the GR control chart, which is prepared by weighting the distance from the median by the Golden ratio number, is more robust to the outliers. In the GR control chart, consecutive values larger than the wdfm average will be used as the decision rule abbreviated by h.

2.1 Example

In this section, the proposed GR control chart is compared with the \bar{x} and EWMA control charts. To determine the ability of the small shifts by using these control charts, thirty observations in Table (2) are taken into account. These data is borrowed from Montgomery (2013). The first 20 of these observations were drawn at random from a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 1$. And the last 10 observations were drawn from a normal distribution with mean $\mu = 11$ and standard deviation $\sigma = 1$.

Table 2. Data for the example

i	X_i	i	X_i
1	9,45	16	9,37
2	7,99	17	10,62
3	9,29	18	10,31
4	11,66	19	8,52
5	12,16	20	10,84
6	10,18	21	10,9
7	8,04	22	9,33
8	11,46	23	12,29
9	9,2	24	11,5
10	10,34	25	10,6
11	9,03	26	11,08
12	11,47	27	10,38
13	10,51	28	11,62
14	9,4	29	11,31
15	10,08	30	10,52

These observations have been plotted on a Shewhart control chart in Figure 2. The three-sigma control limits on \bar{x} chart has an upper control limit (UCL) 13 and lower control limit (LCL) 7. None of these plotted points are outside the control limits and the \bar{x} control chart has failed to detect the shift in the location.

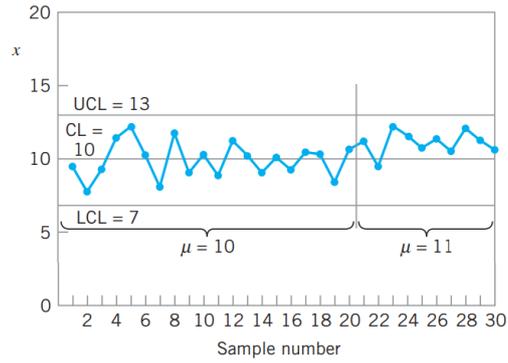


Figure 2. A Shewhart control chart for the data

On the other hand, the center line and control limits for the EWMA control chart are given in Figure (3). The EWMA control chart detecting a shift signal at the observation 28, so the process is out of control.

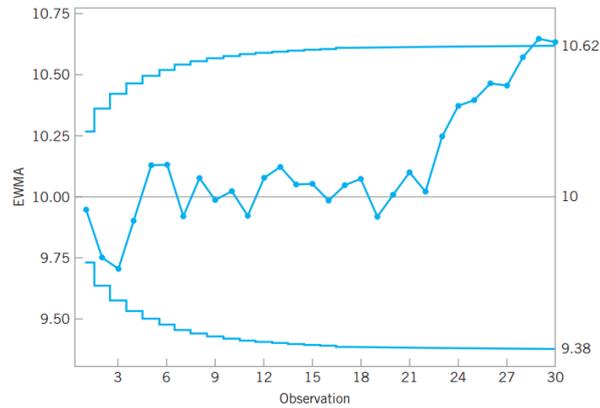


Figure 3. The EWMA control chart for the data

Finally, for the proposed chart, M_{c_i} is calculated by using the golden ratio, and the weighted coefficient mask for the sample of data is obtained in Figure 4. Here the decision rule h is determined as 4.

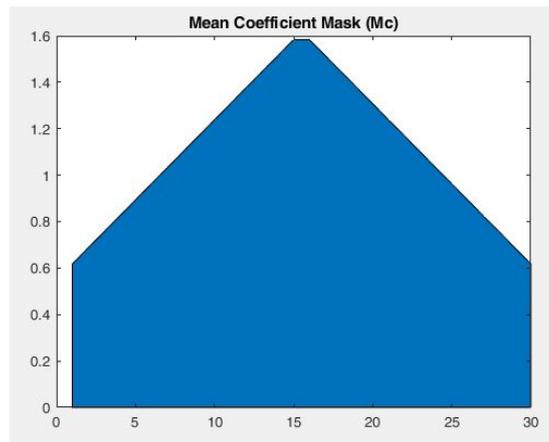


Figure 4. Mean coefficient mask

The next step is to calculate the wdfm of each element. The weighted distances calculated for the sorted values, and then converted to their original order and finally plotted. We conclude that the process was last "in control" at the observation 26th for the decision rule $h=4$ in Figure 5.

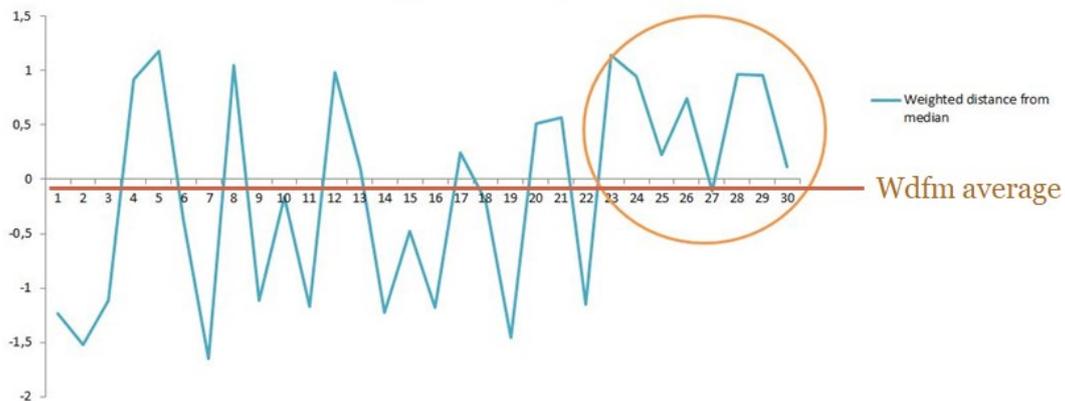


Figure 5. GR control chart

3. Conclusion

This paper presents a new control chart, the GR control chart. In the comparative example, \bar{x} and EWMA charts are examined and some important features of the GR chart are revealed.

As a result based on the given example data, \bar{x} control chart could not detect 1σ shift in location. The EWMA control chart was able to detect the shift at the 28th observation, and the GR control chart can also detect the shift at the 26th observation in the process. The GR control chart seems to be an efficient for detecting a small shift in location compared to the

results of EWMA. However, a simulation study is certainly needed to conclude more complete comments.

References

Duncan, A.J. (1974). *Quality Control and Industrial Statistics*, 4th ed. Wiley, New York, US.

Dunlap, R.A. (1997). *The Golden Ratio and Fibonacci Numbers*, World Scientific Publishing, London, UK.

Günver, M.G., Şenocak, M.Ş., Vehid, S. (2018). “To determine Skewness, mean and deviation with a new approach on continuous data”, *International Journal of Sciences and Research*, vol.74, no.2/1, pp.64-79.

Hawkins, D.M. (1981). “A CUSUM for a scale parameter”, *Journal of Quality Technology*, vol.13, no.4, pp.228-231.

Lucas, J.M. (1976). “The Design and use of V-mask control schemes”, *Journal of Quality Technology*, vol.8, no.1, pp.1-12.

Lucas, J.M., Crosier R.B. (1982a). “Fast initial response for CUSUM quality control schemes: give your CUSUM a head start”, *Technometrics*, vol.24, no.3, pp.199-205.

Lucas, J.M., Saccucci M.S. (1990). “Exponentially weighted moving average control schemes properties and enhancements ”, *Technometrics*, vol.32, no.1, pp.1-12.

Montgomery, D.C. (2013). *Statistical Quality Control*, 7nd ed. Wiley, New York, US.

Page, E.S. (1954). “Continuous inspection schemes”, *Biometrics*, vol.41, no.1, pp.100-115.

Page, E.S. (1961). “Cumulative Sum Control Charts ”, *Technometrics*, vol.3, no.1, pp.1-9.

Roberts, S.W. (1959). “Control chart tests based on geometric moving averages”, *Technometrics*, vol.1, pp.239-250.

Woodall, W.H. (2000). “Controversies and contradictions in Statistical process control”, *Journal of Quality Technology*, vol.32, no.4, pp.341-350.

P-02 Nonparametric Modelling via Wavelet Smoothing

Berna Yazıcı¹, Marwa BenGhoul^{2*}, and Mustafa Çavuş³

¹ Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, Turkey, bbaloglu@eskisehir.edu.tr

² Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, Turkey, benghoulmarwa@gmail.com

³ Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, Turkey, mustafacavus@eskisehir.edu.tr

Abstract – Sometimes, the parametric features of the model can be violated or be too restrictive in some applications. If the parametric model is applied inappropriately, the conclusions from the analysis may be misled. Therefore, the nonparametric models are applied. For the nonparametric features, it is crucial to apply a smoothing approach, the most popular ones are smoothing splines, penalized splines, kernel approaches, and regression polynomial splines. Recently, wavelets decomposition has been known as a powerful mathematical tool applied in various domains such as image processing, audio processing, signal denoising and mentioned as a smoothing approach. Indeed, this paper has an objective to figure out the efficiency of wavelets analysis as a smoothing approach. A longitudinal data produced by the AIDS Clinical Trials Group, sponsored by NIAID/NIH will be used. Different smoothing methods have been used to be compared with the wavelet's decomposition. The results show that the wavelet decomposition demonstrate an impressive capacity as the classic known methods.

Keywords – *wavelets decomposition, smoothing, nonparametric model, longitudinal data*

1. Introduction

Longitudinal data gathers several observations of the same subjects intermittently over time (Müller, 1988; Diggle et al., 2002; Fitzmaurice, Laird and Ware, 2004; Hedeker and Gibbons, 2006). This type of data examines the dynamic measure changes over different timepoints, allow the measurement of the duration of events, studies the relationships and model the differences and the heterogeneity among subjects. Complexity of modelling is a common issue for this type of dataset.

Over the last decades, mixed models have gained a lot of attention in statistical research studies over the traditional analyses due to their flexibility to handle multilevel or clustered data without constraints as equal number of observations or non-missing observations (Howell, 2008; Maxwell, Delaney and Kelley, 2003). Mixed models are current in the design that combines random and fixed effects and widely used in several fields such as pharmaceutical industry and economics.

Mixed effects models can be characterised as parametric and nonparametric; parametric mixed models may take different covariates into account, but they require parametric assumptions that may not be satisfied by non-linear models. Nonparametric mixed models are more flexible to fit longitudinal data and robust against model misspecification, but they involve a few covariates and may be computationally intensive (Maxwell, Delaney and Kelley, 2003). For nonparametric features, smoothing approaches are required. Ruppert, Wand and Carroll (2003) categorized four principal smoothing approaches to nonparametric modelling: smoothing splines (Wahba, 1990; Green and Silverman, 1994; Eubank, 1999); series-based smoothers, including wavelets (Tarter and Lock, 1993; Ogden, 1996); kernel methods (Wand and Jones, 1995; Fan et al., 1996); and regression splines (Friedman, 1991; Stone et al., 1997; Hansen and Kooperberg, 2002). Ruppert, Wand and Carroll (2003) focused on the penalized splines, labelled also as p-splines, pseudo splines, and low-rank spline smoothers in the literature.

It is crucial to highlight that the type of dataset is the reference to choose among the smoothing approaches (Ruppert, Wand and Carroll, 2003).

Despite wavelets analysis was mentioned in Hulin and Jin-Ting (2007) as a smoothing approach, it still not vastly applied notably for longitudinal data. Hence, this research consists in using wavelet decomposition to smooth the data.

The remainder of this dissertation is structured as follows: section two discusses the wavelets background and the dataset. Section three presents the results of this research and the final section sums up this paper.

2. Materials and Methods

2.1 Wavelets background

Historically, since the 18th century, Fourier transform has been known as an important mathematical tool used in the development of telephony, computer science and the audio-visual field. Nevertheless, a major flaw was identified in this tool which is the absence of time information. In fact, Fourier transform gives the information about the number of frequencies contained in the signal but hides the times of the diffusion of these frequencies as if the signal moments are equivalent (Meyer, 1990; Daubechies, 1992). The wavelet was introduced as the alternative approach that breaks down a signal both in time and in frequency. Wavelets have become increasingly a popular tool in various fields such as image processing (digital borrowing, medical X-rays, seismic waves etc.), audio processing (voices, musical notes, etc.) and recently the economic and financial areas (Gençay, Selçuk and Whitcher, 2002). Arneodo, Bacry and Muzy (1995) referred wavelet as a mathematical microscope due to its ability in showing the weak transients and peculiarities in the time series, it utilizes the optics of the microscope, its amplification varies with the scale factor (Struzik, 2001).

Wavelet decomposition is composed by two main functions: *Father function* $\Phi(t)$ defined as $\phi_{j,k}(t) = 2^{-\frac{j}{2}}\phi(2^{-j}t - k)$ and *Mother function* $\psi(t)$ defined as $\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(2^{-j}t - k)$. Decomposing the signal by wavelet provide two components: approximation and details. Approximation coefficients at scale j is presented by $a(j, k) = \int_{-\infty}^{\infty} x(t) \phi_{j,k}(t) dt$ and the detail coefficients at scale j is presented by the function $W(a, b) \cong \int x(t) \psi_{j,k}^*(t) dt = d_{j,k}$.

2.3 Dataset

ACTG388 is a longitudinal Data, produced by AIDS Clinical Trials Group, ACTG 388 study is sponsored by NIAID/NIH. The study enrolled 517 HIV-1 infected patients in three antiretroviral treatments. The data set is only CD4 count data from one of the three treatment arms. Study is described with more details in Fischl, Ribaudo and Collier (2003) and Park and Wu (2006).

3. Results and discussion

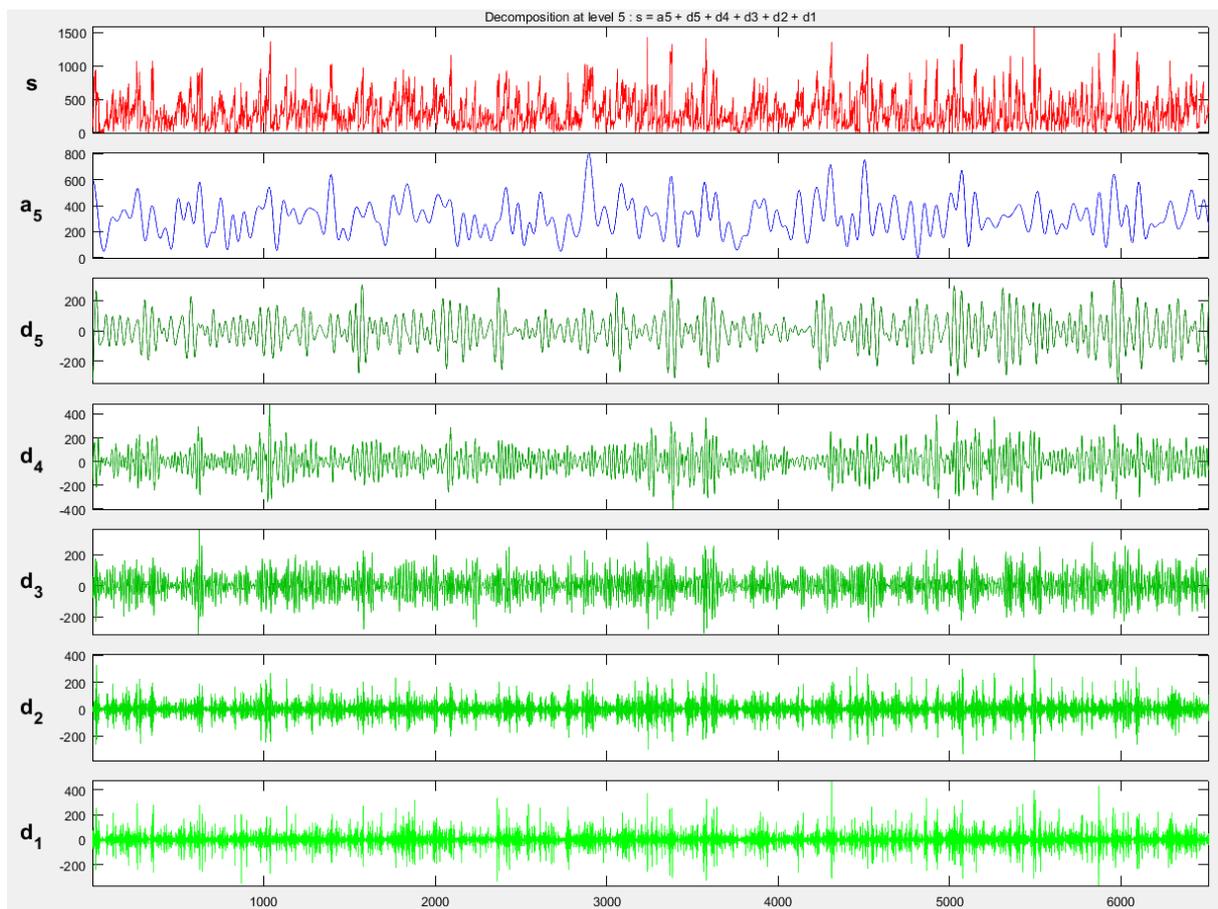


Figure 1. Wavelet Decomposition of CD4 variable

Figure 1 presents the wavelet decomposition. s presents the raw variable CD4, $d1 - d5$ present the detail coefficients, $a5$ defines the approximation coefficients that will be used as

the smoothed CD4. Indeed, the advantage of wavelet analysis is reconstructing the signal or the variable after the decomposition without losing the originality of the data as following:
 $s = d1 + d2 + d3 + d4 + d5 + a5$

Table 1. Comparison of RMSE between smoothing approaches

Smoothing approach*	RMSE (CD4)
Wavelets ¹	0.0448
movmedian	0.0519
movmean	0.0658
lowess	0.0501
rloess	0.0547
rlowess	0.0592
sgolay	0.0488
gaussian	0.0504

***moving**: Moving average (default). A lowpass filter with filter coefficients equal to the reciprocal of the span.
lowess: Local regression using weighted linear least squares and a 1st degree polynomial model
sgolay: Savitzky-Golay filter. A generalized moving average with filter coefficients determined by an unweighted linear least-squares regression and a polynomial model of specified degree (default is 2). The method can accept nonuniform predictor data.
rloess: A robust version of 'lowess' that assigns lower weight to outliers in the regression. The method assigns zero weight to data outside six mean absolute deviations.
rloess: A robust version of 'loess' that assigns lower weight to outliers in the regression. The method assigns zero weight to data outside six mean absolute deviations. (MATLAB website)
¹ The approximation coefficients presented in section 2.1 and presented as a5 in Figure 1 are utilized.

Table 1 compares the Root Mean Square Error (RMSE) associated to different smoothing approaches. RMSE was calculated as following: $RMSE = \sqrt{\text{mean}((\text{Raw variable value} - \text{smoothed variable value})^2)}$.

3. Conclusion

This paper consists in utilizing the approximation coefficients after the wavelet decomposition as smoothed variable. RMSE of different smoothing approaches have been compared, and it has been found that the RMSE related to the wavelet approximation decomposition has the least RMSE. Hence, it can be concluded that utilizing the wavelet decomposition to smooth data show interested results and further research studies are highly recommended to apply it as a smoothing approach.

References

- Arneodo, A., Bacry, E., and Muzy, J.F. (1995). The thermodynamics of fractals revisited with wavelets. *Physica A: Statistical Mechanics and its Applications*, 213(1), 232-275. [https://doi.org/10.1016/0378-4371\(94\)00163-N](https://doi.org/10.1016/0378-4371(94)00163-N).
- Daubechies, I (1992). Ten lectures on wavelets. First edition: Society for Industrial and Applied Mathematics.
- Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Second edition Oxford: Oxford University Press.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. Second edition, New York: Marcel Dekker Inc.
- Fan, J., Gijbels, I., Hu, T.C., and Huang, L.S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6(1), 117–127. Retrieved from <http://www3.stat.sinica.edu.tw/statistica/j6n1/j6n17/j6n17.htm>.
- Fischl, MA., Ribaldo HJ., and Collier, AC. (2003). A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. *Journal of Infectious Diseases*, 188(5), 625-634. <https://doi.org/10.1086/377311>.
- Fitzmaurice, G. M., Laird, N. and Ware, H.J. (2004). *Applied longitudinal analysis*. First edition, New Jersey: John Wiley and Sons. <https://doi.org/10.1198/jasa.2005.s24>.
- Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1), 1-67.
- Gençay, R., Selçuk, F., and Whitcher, B. (2002). *An introduction to wavelets and other filtering methods in finance and economics*. First edition, California: Elsevier, Academic Press. <https://doi.org/10.1016/B978-0-12-279670-8.X5000-9>.
- Green, P. J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. First edition, London: Chapman and Hall/CRC.

- Hansen, M. H., and Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science*, 17(1), 2–51.
- Hedeker, D., and Gibbons, R. (2006). *Longitudinal Data Analysis*. First edition, New Jersey: Wiley-Blackwell.
- Howell, D. C. (2008). The treatment of missing data. *The SAGE Handbook of Social Science Methodology*, 212–226. <https://doi.org/10.4135/9781848607958.n11>.
- Hulin, W., and Jin-Ting, Z. (2007). *Nonparametric regression methods for longitudinal data analysis: Mixed-effects modelling approaches*. First edition, New Jersey: Wiley Series in probability and statistics, 229-270.
- Maxwell, S. E., Delaney, H. D., and Kelley, K. (2003). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Third edition, New York: Routledge.
- Meyer, Y. (1990). *Ondelettes et Opérateurs*, vol. I–III. Paris: Hermann 1990.
- Müller, H.G. (1988). *Nonparametric regression analysis of longitudinal data*. Springer-Verlag. <https://doi.org/10.1007/978-1-4612-5056-2>.
- Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. First edition, Boston: Birkhäuser
- Park, J.G. and Wu, H. (2006). Back fitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data. *Journal of Statistical Planning and Inference*, 136(11), 3760-3782. <https://doi.org/10.1016/j.jspi.2005.03.007>.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. First edition, New York: Cambridge University Press. <https://doi.org/10.1017/CBO9780511755453>.
- Tarter, M.E., and Lock, M.D. (1993). *Model-free Curve Estimation*. First edition: New York, Chapman and Hall/CRC.
- Stone, C.J., Hansen, M. H., Kooperberg, C., and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modelling. *Annals of Statistics*, 25, 1371–1425.
- Struzik, Z. (2001). Wavelet methods in financial time-series processing, *Physica A: Statistical Mechanics and its Applications*, 296 (1), 307-319. [https://doi.org/10.1016/S0378-4371\(01\)00101-7](https://doi.org/10.1016/S0378-4371(01)00101-7).
- Wahba, G. (1990). *Spline Models for Observational Data*. First edition, Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics, <https://doi.org/10.1137/1.9781611970128>.
- Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing*. First edition, New York: Chapman and Hall/CRC.

An Application of New Stochastic Model Using Generalized Entropy Optimization Methods

Aladdin Shamilov¹, Nihal İnce^{2*}

¹Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey, asamilov@eskisehir.edu.tr

²Department of Statistics, Faculty of Science, Eskisehir Technical University, Turkey, nihalyilmaz@eskisehir.edu.tr

Abstract –In this study, approximate solution of stochastic differential equation (SDE) is used in order to obtain probability density function of solution mentioned SDE at fixed times. Trajectories of SDE allow to obtain approximate random variable according to solutions at fixed times. Probability density functions of approximate random variables is acquired by Generalized Entropy Optimization Distributions (GEOM). Using GEOM is explained oneself by that this method represents more flexible distributions. As application, of described method it is considered reasonable SDE based on the data of Aransas-Wood Buffalo population of whooping cranes from 1939 to 2011 and the results are acquired by using statistical software R and MATLAB. The performances of GEOM are established by Determination Coefficient (R^2), Chi –Square (χ^2), Root Mean Square Error (RMSE) criteria and MaxEnt measure (Entropy). Consequently, obtained distributions using by GEOM can be used for assessment of the biological potential and the performance of biological systems.

Keywords – Generalized Entropy Optimization Distributions, Stochastic Differential Equations, Euler-Maruyama’s method, Probability density function

1. Introduction

Recently, because of wide range applications of stochastic differential equations (SDEs) play a significant role in many departments of science and industry because of their application for modeling stochastic phenomena, e.g., biology, population dynamics, chemistry, epidemiology, medicine, mechanics, microelectronics, economics, and finance, Bayram et al. (2018). Stochastic differential equations are differential equations that have at least one term that is a stochastic process, thus resulting in a solution which is itself a stochastic process. In this paper we consider the general form of one-dimensional SDE with

$$X(t, \omega) = X(0, \omega) + \int_0^t f(s, X(s, \omega))ds + \int_0^t g(s, X(s, \omega))dW(s) \quad (1)$$

and differential form

$$dX(t) = f(t, X)dt + g(t, X)dW(t) \quad (2)$$

for $0 \leq t \leq T$ where f is the drift coefficient, while g is the diffusion coefficient and $X(t, \omega)$ is a stochastic process not a deterministic function. $W(t, \omega) = W(t)$ is the Wiener process or the Brownian motion. $W(t)$ that depends continuously on $t \in [0, T]$ and satisfies the following three conditions.

1. $W(0) = 0$ (with probability 1).
2. For $0 \leq s < t \leq T$ the random variable given by the increment $W(t) - W(s)$ is normally distributed with mean zero and variance $t - s$; equivalently, $W(t) - W(s) \sim \sqrt{t - s}N(0, 1)$, where $N(0, 1)$ denotes a normally distributed random variable with zero mean and unit variance.
3. For $0 \leq s < t < u < v \leq T$ the increments $W(t) - W(s)$ and $W(v) - W(u)$ are independent in Higham, (2001).

Note that it is assumed that the functions f and g are non-anticipating and satisfy the following conditions (c1) and (c2) for some constant $k \geq 0$ of existence and uniqueness theorem of solution of SDE model, Allen (2007).

$$\text{Condition (c1): } |f(t, x) - f(s, y)|^2 \leq k|t - s| + |x - y|^2, s \geq 0, T \geq t, x, y \in \mathbb{R}.$$
$$\text{Condition (c2): } |f(t, x)|^2 \leq k(1 + |x|^2), 0 \leq t \leq T, x \in \mathbb{R}.$$

In many cases analytic solutions are not available for SDE, so we are required to use numerical methods e.g. Euler-Maruyama (EM) method, Milstein method and Runge-Kutta method to approximate the solution. In this study, in order to determine the solution of SDE, EM method will be used.

The present paper is organized as follows. In Section 1, a brief explanation on stochastic differential equations is introduced. In Section 2, EM method and Generalized Entropy Optimization Methods (GEOM) are given. In Section 3, describes a new method to obtain approximate probability density function of solution of SDEs by using GEOM. In Section 4, an application on biologic data is illustrated. And the last section the main results obtained in this study are summarized and some suggestions for further research are given.

2. Materials and Methods

In this section, several auxiliary methods and knowledges are given for developing new method to obtain approximate probability density function for solution of SDE.

2.1 Euler-Maruyama (EM) Method

Stochastic differential equations which admit an explicit solution are the exception from the rule. Therefore numerical techniques for the approximation of the solution to a stochastic differential equation are called for. One of the most important simulation based methods is Euler-Maruyama (EM) method which is recently used for solving stochastic differential equations, Shamilov (2012).

Many SDE systems do not have a (known) analytic solution, so it is necessary to solve these systems numerically: the simplest stochastic numerical approximation is the Euler-Maruyama method, Picchini (2007).

When applied to (2), Euler's method has the form

$$X_{i+1}(\omega) = X_i(\omega) + f(t_i, X_i(\omega))\Delta t + g(t_i, X_i(\omega))\Delta W_i(\omega), \quad X_0(\omega) = X(0, \omega) \quad (3)$$

for $i = 0, 1, 2, \dots, N - 1$ where $X_i(\omega) \approx X(t_i, \omega), t_i = i\Delta t, \Delta t = \frac{T}{N}, \Delta W_i(\omega) = (W(t_i + 1, \omega) - W(t_i, \omega)) \sim N(0, \Delta t)$, and where ω indicates a sample path, Allen (2007).

2.2 Generalized Entropy Optimization Methods (GEOM)

Generalized Entropy Optimization Methods modelling the statistical data in the form of Generalized Entropy Optimization Distributions can be successfully applied in many scientific fields, Kapur and Kesavan (1992). GEOD's can be used in modelling, because by increasing the number and changing the type of characterizing moment vector functions MaxMaxEnt and MaxMinxEnt distributions also can be suitable for estimation, Shamilov (2007, 2008, 2009 and 2010).

Generalized Entropy Optimization Distribution indicated as $(\text{MinMaxEnt})_m$ is closest to a given statistical data and distribution indicated as $(\text{MaxMaxEnt})_m$ is furthest from a given statistical data in the sense of MaxEnt measure. Solving the MinMaxEnt and MaxMaxEnt problems require to find vector functions $(g_0, g^{(1)}(x)), (g_0, g^{(2)}(x))$ where $g_0(x) \equiv 1, g^{(1)} \in K_{0,m}, g^{(2)} \in K_{0,m}$ minimizing and maximizing functional $U(g)$ defined, respectively. It should be noted that $U(g)$ reaches its minimum (maximum) value subject to constraints generated by function $g_0(x)$ and all m -dimensional vector functions $g(x), g \in K_{0,m}$. In other words, minimum (maximum) value of $U(g)$ is least (greatest) value of values $U(g)$ corresponding to $(g_0(x), g), g \in K_{0,m}$. In other words, $(\text{MinMaxEnt})_m ((\text{MaxMaxEnt})_m)$ is distribution giving minimum (maximum) value to functional $U(g)$ along of all distributions generated by $\binom{r}{m}$ number of moment vector functions $(g_0(x), g), g \in K_{0,m}$. Therefore, we denote mentioned distributions in the form of $(\text{MinMaxEnt})_m$ and $(\text{MaxMaxEnt})_m$, Shamilov (2010).

By using the mentioned theoretical statements, the following theorem is proved.

Theorem. Assume that conditions (c1), (c2) of existence and uniqueness of solution theorem for SDE (2) satisfied and solution $X(t)$ of SDE has probability density function (pdf) $\varphi(t, s)$. Besides random variable $\hat{X}(t_i)$ obtained by approximating EM method using given statistical data has pdf $\varphi_i(x)$. Then equality

$$|X(t_i) - \hat{X}(t_i)| \leq E \left(|X(t_i) - \hat{X}(t_i)|^2 \right)^{\frac{1}{2}} < \hat{c}\Delta t \quad (4)$$

holds. Moreover surface obtained by $Z = \varphi_i(x), t \in [t_i, t_{i+1}], i = 0, 1, \dots, N - 1$ represents some approximations to pdf $\varphi(t, x)$ of random variable $X(t)$ of solution SDE (2) when $\Delta t \rightarrow 0$.

3. Application

An application of the developed method we have considered a Aransas-Wood Buffalo population of whooping cranes. These whooping cranes nest in Wood Buffalo National Park in Canada and winter in Aransas National Wildlife Refuge in Texas, Butler et al. (2013). The population size is graphed as black line in Figure 1 over the years 1939–2011.

In order to obtain SDE model fitting on population data, the following steps are realized.

1. The formulas for estimating the parameters are given for SDE model.
2. EM schemes are constructed via estimating the values of parameters of population data.
3. Population data and its approximative EM values from model are calculated for $N=72,144,216$. According to mean square error values, it is chosen randomly approximative values of random variable.
4. Approximative probability density functions (pdfs) of mentioned random variable of solutions of SDE model are constructed via pdfs of mentioned random variable in tables and figures by using GEOM.

In this section, this data is fit to the stochastic differential equation

$$dX(t) = \theta_1 X(t)dt + \sqrt{\theta_2 X(t)}dW(t), \quad X_0 = 18 \quad (5)$$

where $X(t)$ is population size and $\theta = [\theta_1, \theta_2]^T$ is to be determined using the maximum likelihood procedure described in Allen (2007). Then, for this model parameters of $\hat{\theta}_1 = 0.042$ and $\hat{\theta}_2 = 0.691$ is estimated. If these estimated parameters are taken into account in SDE model (5), then

$$dX(t) = 0.042X(t)dt + \sqrt{0.691X(t)}dW(t). \quad X_0 = 18$$

By starting determined SDE model and EM method, population data sample path and its EM trajectories are calculated for each value of N which is selected as 72,144,216. For each sample path of the mean square errors are calculated and given in Table 1.

Table 1. Values of mean square errors of EM trajectories

Value of N	MSE
72	0.8010
144	0.7185
216	0.4749

In the following figure, different colors represents the different EM approximate trajectories for population data.

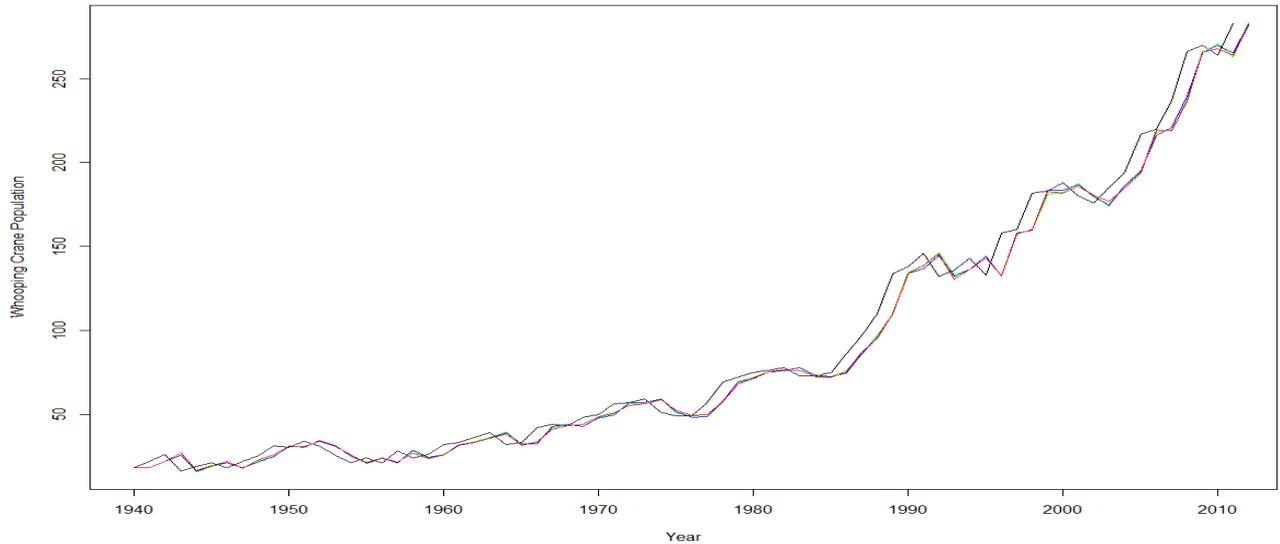


Figure 1. Population data path (black line), sample path of $N=72$ (red line), sample path of $N=144$ (blue line) and sample path of $N=216$ (green line)

According to values of MSE, sample path for $N=216$ is chosen. Then, approximate pdfs of random variables of mentioned path for $\hat{X}(t_{72})$ of solutions of SDE are constructed EM method.

In this study after obtained results, it is shown that $(\text{MinMaxEnt})_m$ and $(\text{MaxMaxEnt})_m$ distributions obtained by Generalized Entropy Optimization Methods (GEOM) is suitable for the assessment of population data in the following form.

- MaxEnt characterizing moments of given moment functions according to data is determined as $g_0(x) = 1, g_1(x) = x, g_2(x) = x^2, g_3(x) = \ln x, g_4(x) = \ln^2(x), g_5(x) = \ln(1 + x^2)$.
- MaxEnt distributions subject to each of MaxEnt characterizing moments is calculated. Hereafter, distributions generated by GEOM corresponding to selected MaxEnt characterizing moments are obtained.
- The performances of distributions generated by GEOM are evaluated by statistical criteria as Determination Coefficient (R^2), Root Mean Square Error (RMSE), Chi-Square (χ^2) and MaxEnt measure (Entropy). The best distribution function can be determined according to the lowest values RMSE, χ^2 , MaxEnt measure and the highest values of R^2 .

Table 2. Evaluation of performance of $(\text{MinMaxEnt})_m, m = 1,2,3,4$ distributions for $\hat{X}(t_{72})$

$(\text{MinMaxEnt})_m$	Entropy	R^2	RMSE	χ^2	Moment Functions
$(\text{MinMaxEnt})_1$	2.1588	0.9782	0.0208	0.0313	$1, \ln^2(x)$
$(\text{MinMaxEnt})_2$	2.1572	0.9809	0.0193	0.0281	$1, x, \ln^2(x)$
$(\text{MinMaxEnt})_3$	2.1572	0.9811	0.0192	0.0279	$1, x, \ln x, \ln(1 + x^2)$
$(\text{MinMaxEnt})_4$	2.1571	0.9809	0.0194	0.0280	$1, x, x^2, \ln^2(x), \ln(1 + x^2)$

Table 3. Evaluation of performance of $(\text{MaxMaxEnt})_m, m = 1,2,3,4$ distributions for $\hat{X}(t_{72})$

$(\text{MaxMaxEnt})_m$	Entropy	R^2	RMSE	χ^2	Moment Functions
$(\text{MaxMaxEnt})_1$	2.2343	0.7766	0.0674	0.1407	$1, x^2$
$(\text{MaxMaxEnt})_2$	2.1640	0.9722	0.0236	0.0380	$1, \ln x, \ln(1 + x^2)$
$(\text{MaxMaxEnt})_3$	2.2162	0.9379	0.0380	0.0605	$1, x^2, \ln x, \ln(1 + x^2)$
$(\text{MaxMaxEnt})_4$	2.1572	0.9808	0.0193	0.0280	$1, x, \ln x, \ln^2(x), \ln(1 + x^2)$

It can be deduced from Table 2,3 that $(\text{MinMaxEnt})_4$ distribution of $\hat{X}(t_{72})$ is more suitable for given estimation data in the sense of R^2 , RMSE, χ^2 criteria and MaxEnt measure. Therefore, it can be explained that $(\text{MinMaxEnt})_m, m = 1,2,3,4$ distributions are regarded as the closest distributions and show better performance than $(\text{MaxMaxEnt})_m$ distributions. These results are also supported by the illustrations in Figure 2 for each data. It can be observed in the all figures that the $(\text{MinMaxEnt})_m$ distributions are more suitable than the $(\text{MaxMaxEnt})_m$ distributions.

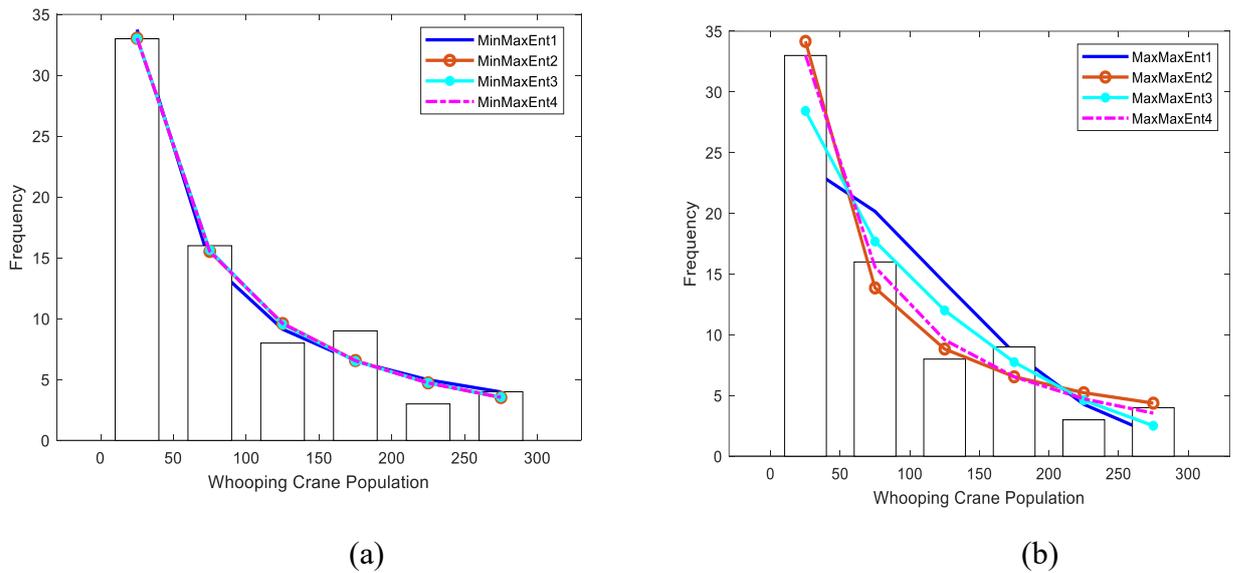


Figure 2. (a) The comparison of the $(\text{MinMaxEnt})_m$ distributions of $\hat{X}(t_{72})$

(b) The comparison of the $(\text{MaxMaxEnt})_m$ distributions of $\hat{X}(t_{72})$

4. Conclusion

In this study after obtained results, it is shown that $(\text{MinMaxEnt})_m$ and $(\text{MaxMaxEnt})_m$ distributions obtained by Generalized Entropy Optimization Methods (GEOM) is suitable for the assessment population data. Approximative probability density functions of random variables of solutions of SDE model are constructed via pdfs of random variables in tables and figures by using GEOM. It is shown that population data and approximative EM values of $\hat{X}(t_{72})$ are fit to selected SDE model. For this data, $(\text{MinMaxEnt})_m$ distributions are compared in terms of modeling population data. The results of the comparison indicate that the obtained $(\text{MinMaxEnt})_4$ distribution gives better result than other $(\text{MinMaxEnt})_m$ and all $(\text{MaxMaxEnt})_m, m = 1, 2, 3, 4$ in the sense of all criteria in application. Finally, the present study may give different and useful insights to biologists and scientists dealing with biological systems.

Acknowledgment

The heading of the Acknowledgment must not be numbered. This paper is supported by Scientific Research Project Office (BAP) in Eskisehir Technical University with project number 16ADP069.

References

Allen, E.J. (2007). Modeling with Itô Stochastic Differential Equations, Springer, USA.

Bayram M., Partal T., Buyukoz Orucova G. (2018). “Numerical methods for simulation of stochastic differential equations”, *Advances in Difference Equations*, vol.17.

Butler, M., Harris, G., Strobel, B. (2013). “Influence of Whooping Crane population dynamics on its recovery and management”. *Biological Conservation*, vol.162. pp.89–99.

Higham, D. (2001). “An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations”, *SIAM Review*, vol.43, pp. 525-546.

Picchini, U. (2007). “SDE Toolbox, Simulation and Estimation of Stochastic Differential Equations with Matlab”.

Kapur J.N., Kesavan H.K. (1992). *Entropy Optimization Principles with Applications*. Academic Press, New York, pp.408.

Shamilov A. (2006). “A Development of Entropy Optimization Methods, *Wseas Transactions on Mathematics*”, vol.5, no.5, pp.568-575.

Shamilov, A. (2007). “Generalized Entropy Optimization Problems and the Existence of Their Solutions”. *Physica A: Statistical Mechanics and its Applications*. vol.382, no.2, pp. 465-472.

Shamilov A. (2009). *Entropy, Information and Entropy Optimization*. Turkey.

Shamilov A. (2010). “Generalized entropy optimization problems with finite moment function sets”, *Journal of Statistics and Management Systems*, vol.13, no.3, pp. 595-603.

Shamilov, A. (2012). *Differential Equations with Theory and Solved Problems*, Nobel Press, Ankara.

Evaluation of User Experience Effects on Ergonomic Behavior with Using Rough-SWARA Method

Sercan Madanlar^{1*} and Şebnem Demirkol Akyol²

¹The Graduate School of Natural and Applied Sciences, Dokuz Eylül University, Turkey,
sercanmadanlar@gmail.com

²Department of Industrial Engineering, Dokuz Eylül University, Turkey,
sebnem.demirkol@deu.edu.tr

Abstract – In product selection decision, user experience about past product usage is an important for evaluating product ergonomics because it recalls good and bad properties of the products. In this study, we develop a participant experiment to evaluate user experience effects on product selection and ergonomic behavior by using Rough-SWARA method. Since it is hard to determine the relative comparison of criteria as percent by participants, this method is suitable for this particular problem. The pruning shear is selected as an investigated product because it is easy to find a participant who has never used that product as a result of urbanization. The experiment is performed with non-expert thirty-two consumers who have evaluated the pruning shears with a total of nine criteria and eleven different types of pruning shears. Participants have ranked the criteria twice which represents the following pair-wise matrices: ranking before experiment and ranking after experiment. Thus, these two matrices are compared according to criteria weight and major changes are observed between these criteria according to product usage experience. Also, this study shows the demonstration of how easy to apply Rough-SWARA method for participants and experts who are evaluating the criteria.

Keywords – *Rough-SWARA, user experience, product ergonomics, weight determination*

1. Introduction

Ergonomics is a branch of science that is occurred from a combination of the two Greek words ergo (work) and nomos (laws). Simply and according to commonly used, ergonomics is a multidisciplinary branch of science that investigates human-machine-environment relationship systematically (Gorner et al., 2015). Ergonomics works for understanding and improving this relationship with the aid of many interdisciplinary fields.

Nowadays, high product variety causes an increase in competition among companies. So, the consumers are investigating the best product to fulfill their requirements. One of the fundamental factor that affects product selection is ergonomics. Product ergonomics became an essential factor for consumers because it affects human-product relationship in many ways. The consumers are evaluated the product ergonomics via checking the properties of the products, asking their friends or user experience. In here, the definition of the user experience handled as a product usage experience of the consumers on the related products. These experienced consumers know or estimates how the products should be for fulfilling the ergonomic requirements because they had used similar types of products.

In this study, we evaluated the user experience effects on ergonomic behavior via participant experiment. Nine criteria were determined for pruning shear evaluation. The pruning shear was selected as an investigated product for evaluating user experience effect that none of the selected participants has a pruning shear usage experience. This condition allows the evaluation of the user experience effects on ergonomic behavior because the evaluation of the criteria was taken from the participants twice: before and after the participant experiment. The evaluation was performed with the Rough-SWARA (Zavadskas et al., 2018) method by the comparison of the before and after criteria ranking matrices.

2. Materials and Methods

The evaluation of the user experience can be performed in many ways. Generally, the questionnaire is used as the main source about an investigation area for gathering information from people. For example, Dempsey et al. (1996) worked on an ergonomic investigation of the letter-carrier satchels. Ten criteria were evaluated for four different types of letter-carrier satchels with the questionnaire. Also, participants evaluated some criteria via past product usage experiences which were specified. Wang et al. (2014) performed a subjective evaluation via the questionnaire for comparison of the normal clothing and the new clothing according to wheelchair user's behavior.

In questionnaire, the criteria can be evaluated in many ways. Likert scale is the most common method for gathering participant evaluation easily. Also, weight determination methods are also used for evaluating the importance of the criteria. AHP (Saaty, 1980), SWARA (Kersulienė et al., 2010), CRITIC (Diakoulaki et al., 1995), SMART (Edwards, 1977) and many other methods are used to determine weight values by researchers. However, most of these methods, the relative comparison between criteria should be determined by experts. This process might be very hard for experts and non-experts participants. Also, the more criteria might be causes an increase of the inconsistency between criteria and participant evaluation. For example, in evaluation of Human Supervisory Control of Unmanned Vehicles (Donmez et al., 2011), the participants evaluated the determined metrics via AHP and RIM method. The authors specified that the participants wanted to give up the experiments if the inconsistency occurs many times.

In this study, we used the Rough-SWARA method for evaluation of the user experience effects on ergonomic behavior because it was found suitable for following features:

- Participants only rank the criteria from best to worst
- It can be performed easily and takes short time to apply
- Relative comparisons between criteria are not determined by participants

2.1 Rough-SWARA method

Rough-SWARA is a weight calculation method that is developed via integrating Rough Set Theory into the SWARA method. According to Zavadskas et al. (2018, p. 99) and Voćkić et al. (2018, p. 217), “In the rough set theory, any vague idea can be represented as a couple of exact concepts based on the lower and upper approximations”. U is the universe that contains all the objects, arbitrary object of U is defined as Y , R is a set of t classes $\{G_1, G_2, \dots, G_t\}$ that cover all the objects in U , $R = \{G_1, G_2, \dots, G_t\}$. If these classes are ordered as $G_1 < G_2 < \dots < G_t$, then $\forall Y \in U, G_q \in R, 1 \leq q \leq t$, by $R(Y)$ meaning the class to which the object belongs. The lower approximation ($\underline{Apr}(G_q)$), upper approximation ($\overline{Apr}(G_q)$) and boundary region ($\overline{Bnd}(G_q)$), of class G_q are defined as (Zavadskas et al., 2018, Voćkić et al., 2018):

$$\underline{Apr}(G_q) = \{Y \in U / R(Y) \leq G_q\}, \quad (1)$$

$$\overline{Apr}(G_q) = \{Y \in U / R(Y) \geq G_q\}, \quad (2)$$

$$Bnd(G_q) = \{Y \in U / R(Y) \neq G_q\} = \{Y \in U / R(Y) > G_q\} \cup \{Y \in U / R(Y) < G_q\} \quad (3)$$

Then G_q can be represented as $(RN(G_q))$, which is determined by its corresponding lower limit

$$\underline{Lim}(G_q) = \frac{1}{M_L} \sum R(Y) | Y \in \underline{Apr}(G_q) \quad (4)$$

$$\overline{Lim}(G_q) = \frac{1}{M_U} \sum R(Y) | Y \in \overline{Apr}(G_q) \quad (5)$$

$$RN(G_q) = [\underline{Lim}(G_q), \overline{Lim}(G_q)] \quad (6)$$

Where M_L , M_U are the numbers of objects that contained in $\underline{Apr}(G_q)$ and $\overline{Apr}(G_q)$ respectively. These six formulas will be used to the Rough SWARA method to calculate RNs from the decision making matrix.

The Rough SWARA methods steps are given below respectively (Zavadskas et al., 2018, Voćkić et al., 2018):

Step 1: Determine criteria for decision-making matrix.

Step 2: The number of participants (k) who should rank the determined criteria (c_n) from best to worst. Then, the pair-wise comparison matrix is obtained.

Step 3: Individual responses are converted to group rough matrix. These calculations are performed with the usage of (1) – (6) equations:

$$RN(C_j) = [c_j^L, c_j^U]_{1 \times m} \quad (7)$$

Step 4: $RN(C_j)$ normalization to obtain the matrix $RN(S_j)$:

$$RN(S_j) = [s_j^L, s_j^U]_{1 \times m} \quad (8)$$

The elements of matrix $RN(S_j)$ are obtained with the following equation:

$$RN(S_j) = \frac{[c_j^L, c_j^U]}{455 \max [c_j^L, c_j^U]} \quad (9)$$

The first element of matrix $RN(S_j)$, i.e. $[s_j^L, s_j^U] = [1.00, 1.00]$, because $j=1$. For other elements $j>1$, the equation (9) can be calculated using the following equation:

$$RN(S_j) = \left[\frac{c_j^L}{\max(c_j^U)}, \frac{c_j^U}{\max(c_j^L)} \right]_{1 \times m} \quad (10)$$

Step 5: Calculate the matrix $RN(K_j)$ with following equation:

$$RN(K_j) = [k_j^L, k_j^U]_{1 \times m} \quad (11)$$

by applying the following equation:

$$RN(K_j) = [s_j^L + 1, s_j^U + 1]_{1 \times m} \quad (12)$$

Step 6: Recalculated weights $RN(Q_j)$ are determined with following equation:

$$RN(Q_j) = [q_j^L, q_j^U]_{1 \times m} \quad (13)$$

The elements of matrix $RN(Q_j)$ with the next equation (14):

$$RN(Q_j) = \left[q_j^L = \begin{cases} 1.00 & j = 1 \\ \frac{q_{j-1}^L}{k_j^U} & j > 1, \end{cases}, q_j^U = \begin{cases} 1.00 & j = 1 \\ \frac{q_{j-1}^U}{k_j^L} & j > 1, \end{cases} \right] \quad (14)$$

Step 7: Relative weight values $RN(W_j)$ calculation with following equation:

$$RN(W_j) = [w_j^L, w_j^U]_{1 \times m} \quad (15)$$

Lastly, individual weight values of criteria are calculated with the last equation:

$$[w_j^L, w_j^U] = \left[\frac{[q_j^L, q_j^U]}{\sum_{j=1}^m [q_j^L, q_j^U]} \right] \quad (16)$$

Thus, the interval values of criteria are calculated.

3. Experiment of User Experience Evaluation

The pruning shear is selected as an investigated product because it is easy to find a participant who has never used those products. The experiment is performed with non-expert thirty-two (15 male and 17 female) participants who have evaluated the pruning shears with a total of nine criteria and eleven different types of pruning shears. However, 3 rankings were eliminated because of the incorrect ranking. So, 29 rankings were evaluated in the participant experiment. The nine criteria are, durability (C1), health (C2), comfort (C3), usability (easy to use) (C4), effort/force (C5), functionality (C6), price/cost (C7), product guarantee (C8), and environmental effect/recyclability (C9). The last three criteria are indirect ergonomic criteria that were evaluated because these are also important for consumers. The experiment was performed with participants who are fourth grade Industrial Engineering students from Dokuz Eylül University. All students took the “IND 4900 Ergonomic Assessment and Job Safety” course and also all students had a background about ergonomic risk assessment. Thus, the evaluation of the pruning shears was performed properly by participants.

The evaluation of the criteria weights is a part of the participant experiment. The experiment consists of usage and evaluation of eleven different pruning shears by each participant. Besides, participants have ranked the criteria twice which represents the following pair-wise matrices: ranking before experiment and ranking after experiment. The participants' ranking is given in appendix 1 that if the participants do not want to change criteria rankings after the experiment, the ranking is transferred to the second matrix. So, we can measure the user experience effects on determined criteria via comparison of the two matrices.

4. Results

After the collection of the participant rankings, the rough-SWARA steps were applied to the pair-wise matrices. Thus the weight of the matrices were calculated and given in table 1.

Table 2. Results of both matrices criteria weights and its comparison

	Criteria ranking before experiment			Criteria ranking after experiment		
		Lower	Upper		Lower	Upper
Durability	w1	0,056350998	0,19413	w1	0,020280356	0,101963
Health	w2	0,099780625	0,256365	w2	0,163643939	0,313347
Comfort	w3	0,253557848	0,39545	w3	0,10477623	0,257532
Usability (easy to use)	w4	0,167022681	0,321832	w4	0,247671166	0,377778
Effort	w5	0,031416254	0,145713	w5	0,064198094	0,206641
Functionality	w6	0,017338776	0,108827	w6	0,03633078	0,149587
Price/cost	w7	0,00927784	0,074523	w7	0,010852724	0,064684
Product guarantee/warranty	w8	0,004407969	0,040799	w8	0,005330463	0,035202
Environmental effect/recyclability	w9	0,002034736	0,021966	w9	0,002515794	0,018588

These matrices represent the before experiment and after experiment values of consumers' attitude on pruning shears usage. Of the participants, 58.62% changed the benchmark rankings after the pruning shear test, and 41.38% did not change the benchmark ranking. It is clearly seen that the user experience effects the user's ergonomic behavior attitude on pruning shears. After the product usage experience, health, usability (easy to use), effort and functionality criteria weights are significantly increased. On the other hand, durability and comfort criteria significantly decreased. However, these criteria are still founding important by participants.

The last three criteria which are price/cost, product guarantee/warranty, and environmental effect/recyclability are the criteria that the participants did not interest. It is clearly seen that in these products, ergonomic criteria are much more effective in choosing product than indirect

ergonomic products. It is observed that indirect ergonomic criteria became less important against the ergonomic criteria according to the participant's behavior.

5. Conclusion

This study demonstrates the investigation of the user experience effects on ergonomic behavior using Rough-SWARA method via participant experiment. Rough-SWARA method was found suitable for obtaining criteria weights easily in these types of investigations because of the participant only performs criteria ranking from best to worst according to their behavior. In this way, criteria values can be calculated easily from the non-expert participant's rankings. From the perspective of the researchers, it is anticipated that this method might be found more effective to apply than other methods in terms of formulation, interval criteria results, easy to apply/calculate, and consistency.

References

Dempsey, P. G., Ayoub, M. M., Bernard, T. M., Endsley, M. R., Karwowski, W., Lin, J., et al. (1996). “Ergonomic investigation of letter-carrier satchels: Part I. Field study”, *Applied Ergonomics*, vol. 27 (5), pp.303-313.

Diakoulaki. D., Mavrotas, G., Papayannakis. L. (1995). “Determining objective weights in multiple criteria problems: the CRITIC method”, *Computers Ops Res.*, vol. 22(7), pp.763-770.

Donmez, B., Cummings, M. L. (2011). “Metric selection for evaluating human supervisory control of unmanned vehicles”, *International Journal of Intelligent Control and Systems*, vol. 16(2), pp.67-78.

Edwards, W. (1977) “How to use multiattribute utility measurement for social decision making”, *IEEE Trans Syst Man Cybern*, vol. 7, pp.326–340.

Gorner T., Simon M. (2015). “Using the Theory of Technical Systems to Describe the Interaction between Human and Technical Systems within the Ergonomic System”, *Procedia Engineering*, vol. 100, pp.592 – 601.

Kersuliene, V., Zavadskas, E. K., Turskis, Z. (2010). “Selection of Rational Dispute Resolution Method by Applying New Step-Wise Weight Assessment Ratio Analysis (SWARA)”, *Journal of Business Economics and Management*, vol. 11(2), pp.243-258.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Vockic, M., Stojic G., Stevic, B. (2018). “Integrated rough SWARA-ARAS model for selection of electric forklift” *The 2nd International Conference on Management, Engineering and Environment*, pp.216-227.

Wang, Y., Wu, D., Zhao, M., Li, J. (2014). “Evaluation on an ergonomic design of functional clothing for wheelchair users”, *Applied Ergonomics*, vol. 45, pp.550-555.

Zavadskas, E. K., Stevic, Z., Tanacko, I., Prentkocskis, O. (2018). “A novel multicriteria approach-rough step-wise weight assessment ratio analysis method (R-SWARA) and its application in logistics”, *Studies in informatics and control*, vol. 27(1), pp.97-106.

Appendices

Appendix 1

Table 2. Pairwise matrices of criteria rankings

	Criteria ranking before experiment										Criteria ranking after experiment								
	C1	C2	C3	C4	C5	C6	C7	C8	C9		C1	C2	C3	C4	C5	C6	C7	C8	C9
E1	4	5	1	2	6	7	2	9	8	E1	2	3	4	5	6	1	7	8	9
E2	4	3	2	5	6	1	7	8	9	E2	4	3	2	5	6	1	7	8	9
E3	3	7	1	2	4	6	5	8	9	E3	5	3	4	1	2	6	7	8	9
E4	2	6	5	4	3	7	1	8	9	E4	3	1	4	6	5	7	2	8	9
E5	1	6	2	3	4	5	7	9	8	E5	6	5	4	1	2	3	7	8	9
E6	6	2	3	5	2	4	7	9	8	E6	6	2	3	5	2	4	7	9	8
E7	9	1	3	4	7	8	2	5	6	E7	9	1	3	2	4	6	5	7	8
E8	5	4	6	1	8	2	3	7	9	E8	5	4	6	1	8	2	3	7	9
E9	7	3	9	4	2	8	6	5	1	E9	7	3	9	4	2	8	6	5	1
E10	2	1	4	3	6	5	7	8	9	E10	2	1	4	3	6	5	7	8	9
E11	1	4	7	5	3	2	6	8	9	E11	5	2	3	4	1	6	7	8	9
E12	3	4	5	1	6	2	7	8	9	E12	6	3	2	4	1	5	7	8	9
E13	5	2	3	1	7	4	6	9	8	E13	5	2	3	1	7	4	6	9	8
E14	6	2	3	4	1	5	7	8	9	E14	6	2	3	4	1	5	7	8	9
E15	4	7	1	3	2	5	6	8	9	E15	7	2	1	3	4	6	5	8	9
E16	2	3	1	4	5	7	6	8	9	E16	5	3	1	2	4	6	7	8	9
E17	3	2	1	4	5	6	7	8	9	E17	4	5	1	2	3	6	7	8	9
E18	4	7	1	3	2	5	6	8	9	E18	6	4	1	3	2	7	5	8	9
E19	7	4	3	5	6	1	2	9	8	E19	4	6	7	1	2	3	5	9	8
E20	4	5	2	3	6	1	7	9	8	E20	4	5	2	3	6	1	7	9	8
E21	6	2	5	1	3	4	7	8	9	E21	6	2	5	1	3	4	7	8	9
E22	4	6	1	2	3	7	5	8	9	E22	4	6	1	2	3	7	5	8	9
E23	6	5	2	3	1	4	8	7	9	E23	6	5	2	3	1	4	8	7	9
E24	4	3	2	1	5	6	7	8	9	E24	6	4	2	1	3	5	8	7	9
E25	1	2	3	5	8	4	6	7	8	E25	4	1	3	2	5	6	7	8	9
E26	7	1	2	3	5	6	4	9	8	E26	7	1	2	3	5	6	4	9	8
E27	5	1	2	3	6	7	4	8	9	E27	6	1	2	3	4	7	5	8	9
E28	4	6	3	2	7	1	5	8	9	E28	7	4	2	1	3	5	6	8	9
E29	7	4	3	1	2	5	6	8	9	E29	6	1	3	4	2	5	7	8	9

Time Series Analysis of Rice Prices using Box-Jenkins ARIMA Methodology in Hargeisa, Somalia

Abdishakur Ismeal Adam^{1*} and Prof. Dr. Vedide Rezan Uslu¹

^{1*} Department of Statistics, Ondokuz Mayıs University, Turkey, cabsha1994@gmail.com

¹ Department of Statistics, Ondokuz Mayıs University, Turkey, rezzanu@omu.edu.tr

Abstract

The international prices of agricultural commodities have been increasing considerably. This upward trend, which may cause a new food crisis, has attracted the attention of the world. Several explanations for these movements in prices have been provided by analysts, researchers, and development institutions. The main purpose of this study was to determine and get forecasts of rice prices in Hargeisa, Somalia by using Box-Jenkins ARIMA modeling. Rice prices in Hargeisa were examined in order to identify if it is stationary or not. In order to check if it is stationary, we have used time series plot, correlograms and done Augmented Dickey-Fuller test. The results revealed that the data is non-stationary. We have used some approaches such as taking differences to make the data stationary. After getting it stationary we have determined some time series Box-Jenkins models as candidates. After that the determined models were compared with respect to the model accuracy criteria such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Then we have found the best model fitted well to the data set. After doing diagnostic checking we have calculated the forecasts. And all of the results of whole analysis were presented. The outcome of this study can aid both Somalia government and policy makers in making optimal production decisions and in managing overall price risks.

Keywords - Agricultural Commodities, Trend, Box-Jenkins ARIMA modeling, Stationary, Somalia

1. Introduction

Global food prices have recently become an increasingly important international concern with the occurrence of the food price crisis of 2007-2008. The Food and Agriculture Organization (FAO) food price index, which measures the international prices of meat, dairy, cereals, oils and fats, and sugar, climbed from 127 in 2006 to 159 in 2007 to 200 in 2008 (United Nations 2012). Continuing a decade-long increase, global food prices rose 2.7 percent in 2012, reaching levels not seen since the 1960s and 1970s. However, still well below the price spike of 1974. Between 2000 and 2012, the World Bank global food price index increased 104.5 percent, at an average annual rate of 6.5 percent. Food price volatility has increased dramatically since 2006. European Union, (2010). According to the United Nations, Food and Agriculture Organization the standard deviation—or measurement of variation from the

average—for food prices between 1990 and 1999 was 7.7 index points, but it increased to 22.4 index points in the 2000–12 period. FAO, IFAD, WFP (2015),

WFP (2015) reported that Africa is facing its worst food crisis in years, as El Nino rages across the continent; new analysis from Mail & Guardian Africa collating data from the UN, the Famine Early Warning System Network (FEWS Net) and various news agencies reveals that more than 40 million people in Africa are facing food insecurity and some outright starvation. The continent needs at least \$4.5 billion for emergency relief, but just a fraction of that has been raised so far, even as analysis from Oxfam shows that an early response is far cheaper than a late one. Also, the report states that The Horn of Africa particularly Ethiopia and much of southern Africa is in bad straits and the weather is not the only factor at play. A country’s ability to cope depends partly on its public finances and ability to mobilize resources; for some weakening currencies make food imports more expensive, and conflict is making it difficult to move supplies around. World Food Programme, Fews Net, Nasa Fldas (2016),

As Somalia is facing an upward pressure on food prices in the country with deteriorating terms of trade and a larger food import bill to pay. Somalia is totally dependent on imports of sugar, rice, spaghetti, wheat flour and vegetable oil, Somalia does not produce rice and does not have the climatic conditions to grow wheat. Because of converging trends including water scarcity, land degradation, lack of on-farm technological innovation, population growth, and climate change. Barrett, Christopher (1996), Much food supply exists in Somalia and this resulted shock and surprise Food price rates in the country. Hargeisa is the second largest city of Somalia also it’s the second most populous city of whole the country and still there is a high food insecurity and high food price rates. The people most likely to be negatively affected are urban poor, disabled people, elderly people, children and women, the unemployed and IDPs MoNPD (2011).

Time series analysis is a statistical method which is used forecasting for the future based on the events of past and present. It comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data I. Agung (2018). In this report, we applied the principles of Box-Jenkins methodology to secondary data comprised monthly Food prices (particularly Rice price) from November 2013 to April 2016, Hargeisa, Somalia, which was obtained from the Ministry of National Planning and Development (MoNPD) especially Department of Statistics.

The study seeks to model, validate and forecast the Hargeisa monthly food prices and highlight the trend analysis followed by Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) to forecast Hargeisa Food Prices for the next year Apr, 2017. The findings of the study could serve as a guide for a review and help assess the current interventions to curb the high food price rates of Somalia.

2. Materials and Methods

2.1 Box-Jenkins Methodology

Box-Jenkins methodology (named after the statisticians George Box and Gwilym Jenkins) is a statistical procedure that is used to model time series data by using autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models. The Box -

Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. The model is generally referred to as an ARIMA (p, d, q) model where p, d and q are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. The study used secondary monthly data, collected from MoNPD, Hargeisa, Somalia which covered the period of 2 years, from 2013 to 2016. The data were modelled using ARIMA models. The ARIMA models help to fit datasets that have time series structure to describe the trend of food prices and forecast points ahead. It also provides a forecast interval and it is based on a proven model. Box, and Jenkins, (1970)

2.2 The ARIMA Model

The basic processes of the Box–Jenkins ARIMA (p, d, q) model include the autoregressive process, the integrated process, and the moving average process. A dataset Y_t follows ARIMA model if the d^{th} differences $\nabla^d Y_t$ follows a stationary ARMA model. The parameters that help build the ARIMA model are three; p , which determines the AR order; d , denotes the number of differencing required before stationarity, and MA order is given by q . Tebbs, J.M. (2010).

Hence, ARIMA (p, d, q) is represented in a general form according to Tebbs J.M as;

$$\phi(B)(1 - B)^d Y_t = \theta(B)e_t \quad (1)$$

where, the AR and MA characteristic operators are

$$\phi(B) = 1 - \phi_1(B) - \phi_2(B)^2 - \dots - \phi_p(B)^p \quad (2)$$

$$\theta(B) = 1 - \theta_1(B) - \theta_2(B)^2 - \dots - \theta_p(B)^p \quad (3)$$

and
$$(1 - B)^d Y_t = \nabla^d Y_t \quad (4)$$

Where, ϕ is the autoregressive parameter to be estimated; θ is the moving average parameter to be estimated; ∇ , the difference operator; B, the backward shift operator; e_t random process having zero mean and variance. Box and Jenkins proposed the estimation of parameters of ARIMA model, and their approach involves the steps: identification of ARIMA model, model parameter estimation, and model diagnostics. Tebbs, J.M. (2010).

2.3 Unit Root Test

In order to make inferences in time series analysis, it is necessary to determine whether the time series is stationary or not. This study we used the Augmented Dickey Fuller (ADF) test for assessing stationarity of the dataset. The test assumes that y_t follows a randomness in the time series data:

$$Y_t = \rho Y_{t-1} + e_t \quad (5)$$

Where, ρ , the characteristic root of an AR polynomial and e_t is white noise with mean zero and variance σ^2 . The ADF test helps to test the null hypothesis of non-stationarity in the data. This results in the following ADF unit root test: $H_0: \delta = 0$ (non-stationary) versus $H_1: \delta < 0$ (stationary) Dickey, D.A. and Fuller, W.A. (1979).

2.4 Identification of ARIMA Model

There are techniques under ARIMA model identification which estimate the p, q and d values. The autocorrelation function (ACF) and partial autocorrelation function (PACF) help to determine the p, q and d values. The theoretical PACF of ARIMA (p, q, d) process usually show non-zero PACF at first p lags, with remaining lags having zero PACF. The first q lags also report non-zero ACF and the remaining lags having zero ACF for the theoretical ACF. We determine q and p by the total frequency of the significant lags which are not zero for ACF and PACF respectively. If the values of p, d, and q are inaccurately selected, models derived can be inadequate, hence cannot be used for predictions Tebbs, J.M. (2013).

2.5 Estimation of Model Parameters, Model Diagnostic and Validation

If the ARIMA model is identified, then we can estimate the parameters of the model using EViews, Model estimation is followed by model selection, and it is done by considering minimum values of Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Tebbs, J.M. (2013).

$$AIC = -2 \ln(\bar{L}) + 2h \quad (6)$$

$$\text{And } BIC = -2 \ln(\bar{L}) + \ln(n)h \quad (7)$$

where, L is the likelihood value of the likelihood function, h and n are number of parameters to be estimated and number of residuals respectively. For any two competing models, the model with the minimum AIC or BIC will be selected as a better one. After fitting the model, we will be checked whether the model is appropriate or not, or we investigated how the model fits the data. Therefore, the Normality plot and ACF and PACF plots of Residuals are convenient graphical techniques for model diagnostic we used. To validate the model selected, the dataset was modelled using a training set which comprised of data from Nov 2013 to Feb 2017 and validated using a testing sample from Mar 2017 to Dec 2017. The validation measures included mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) will be used. Proietti, T. and Lutkepohl, H. (2011)

3. Results

3.1 Data Handling and Transformation

The data for this study is a Hargeisa monthly rice prices (Nov 2013 – Dec 2017) from Ministry of National Planning and Development, Hargeisa, Somalia. The time series plot below in Figure 1 shows Hargeisa monthly rice prices from November 2013 to December 2017.

The time series plot of our data shows that there is a deterministic trend which provides that the time series is not stationary.

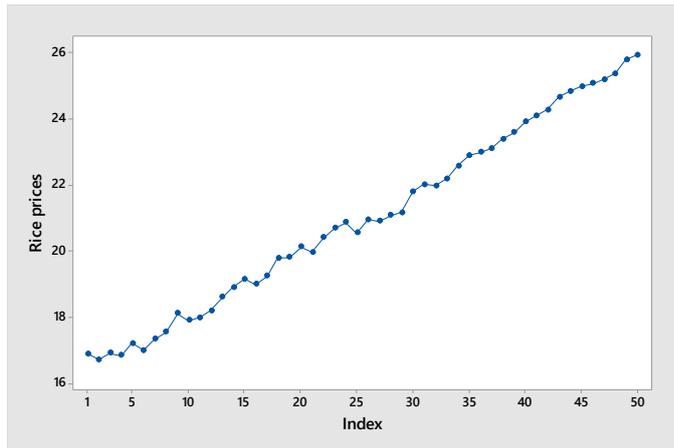


Figure 1. Time series plot of Hargeisa monthly rice prices, Hargeisa, Somalia.

3.2 Testing Stationarity

It is clear there was an increasing trend in our original data which clarifies that the series is non-Stationary. In addition to the graphical approach for testing stationarity, it is also necessary to test with statistical testing. One of these approaches is the Unit Root Test which has been widely used in recent years. Adedia D, Nanga, S, et al (2018).

The Augmented Dickey Fuller (ADF) test were applied for assessing stationarity of the dataset. The results of no differenced data series for ADF tests (constant, constant & trend and none) revealed that the data series in non-stationary since the absolute of the ADF test statistic is less than ADF critical values 1%,5% and 10%, however ADF tests confirmed stationarity after the data series was differenced of the first order as can be seen in Table 1 and Since the data series is non-Stationary we should make the data series stationary before estimation of the best model fit using first difference method.

Table 1: The results of the Unit Root test (ADF)

ADF tests	Differencing Order	ADF 1% critical value	ADF 5% critical value	ADF 10% critical value	ADF test statistic	Decision
No constant and no trend	0	-2.6162	-1.9481	-1.6123	-1.2089	Accept H_0
Constant and no trend	0	-3.5811	-2.9266	-2.6014	1.1879	Accept H_0
Constant and trend	0	-4.1705	-3.5107	-3.1855	-2.1075	Accept H_0
No constant and no trend	1	-2.6198	-1.9486	-1.612	-6.2385	Reject H_0
Constant and no trend	1	-3.5811	-2.9266	-2.6014	-7.7084	Reject H_0
Constant and trend	1	-4.1705	-3.5107	-3.1855	-7.8762	Reject H_0

We recognized below Figure 2 clearly that First-differenced time series is *stationary* (constant mean and approximately constant variance) and now since the data series is stationary, we can study its behavior, calculate the best fit model the data series using the Box-Jenkins Methodology (ARIMA Modelling).

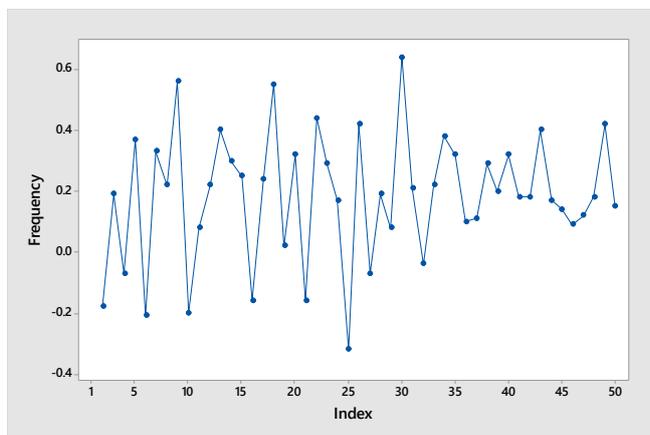


Figure 2: Time series plot of differenced Hargeisa rice prices.

3.3 Model Identification

The output in **Table 1** shows that after the first difference the dataset became stationary. By using the spikes in the ACF and the PACF plot of the differenced data of the first order, we suggest both the q and p values. Figure 2 shows the ACF and PACF plots. The ACF plot has spikes (significant lags) at lags 1, lag 3 and 4, which is the moving average (MA) part to the model and the PACF plot has spikes for lags 1 and 3 which shows the autoregressive part (AR) to the model.

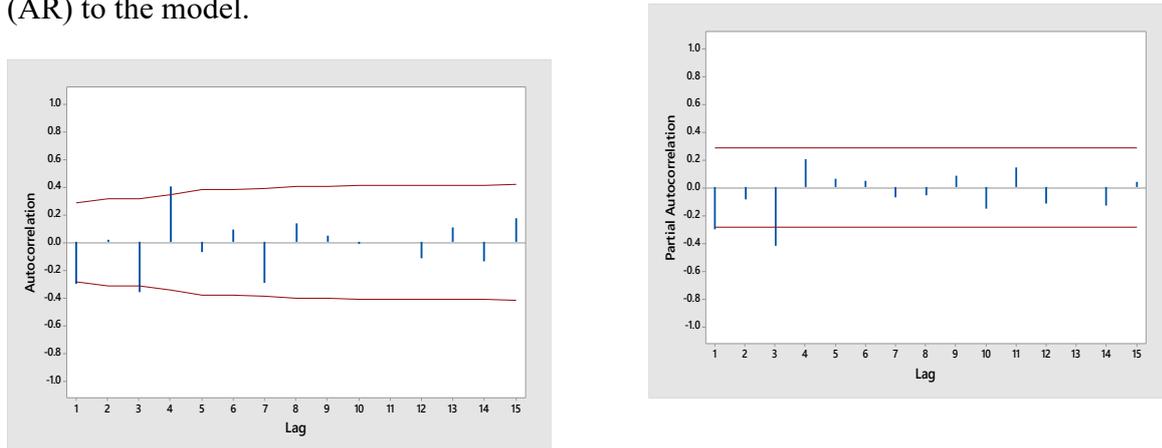


Figure 3: ACF and PACF of first differenced Hargeisa rice prices

Therefore, models were tentatively suggested based on the combination of the significant spikes in both the ACF and PACF plots (Figure 2), and through Box-Jenkins approach the best model was selected as the best. ARIMA (1, 1, 1* (Row 1)) in Table 2 is the best model because it is the model with the least AIC and BIC values. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994).

Table 2. The fitted ARIMA (p, 1, q) models

No	Models	AIC	BIC
1*	ARIMA (1,1,1) *	- 0.8556*	- 0.72631*
2	ARIMA (1,1,3)	0.53888	-0.3234
3	ARIMA (1,1,4)	0.25738	0.001188
4	ARIMA (3,1,1)	0.36045	-0.14051
5	ARIMA (3,1,3)	0.57886	-0.27095
6	ARIMA (3,1,4)	0.41534	-0.06345

3.4 Model Estimation

Once we indicated the best fitted model for the data series, we estimated the parameters of chosen model ARIMA (1, 1, 1) as Table 3 below shows

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.191119	0.014141	13.51519	0.0000
AR(1)	0.378571	0.069416	5.453648	0.0000
MA(1)	-1.34053	0.193358	-6.93289	0.0000

Table 3: Model parameters estimates of ARIMA (1, 1, 1)

therefore, our model will be as following:

$$\hat{y}_t = 0.1911 + 0.3786y_{t-1} - 1.3405\varepsilon_{t-1} \quad (8)$$

3.5 Model Diagnostic

After fitting the model, we should check whether the selected model is appropriate or not, or in other words we study how the model fits our data. Therefore, the *normal probability plot*, ACF and PACF plots of residuals of the selected model is a good convenient graphical technique for model diagnostic.

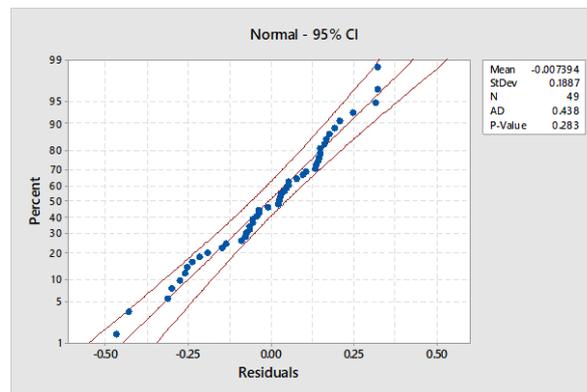


Figure 4: Normal Probability Plot for ARIMA (1, 1, 1) model

According the above figure 4 it indicates that the normal distribution provides an adequate fit for this model since p-value of $0.283 > \alpha = 0.05$ which confirmed the normality of the residuals. The Ljung-Box test also showed that the ARIMA (1, 1, 1) was adequate with p-value = $0.104 > \alpha = 5\%$ and could be used to forecast Hargeisa monthly food prices.

Figure 5 shows the ACF and PACF graphs of residuals obtained from ARIMA (1, 1, 1) model. As we can examine that there is no significant lags and we can conclude that the residuals are white noise (independent and identically distributed) so ARIMA (1, 1, 1) model is the best model for prediction.

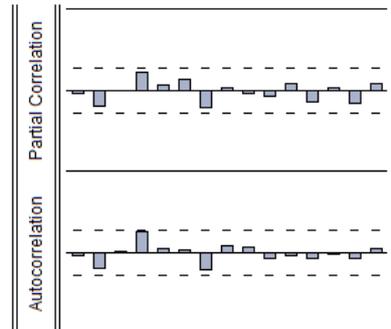


Figure 5: ACF and PACF plots of Residuals for ARIMA (1, 1, 1) model

3.6 Model Validation

The dataset was partitioned as training and testing sample. The training sample contains about 80 % (Nov-2013 to Feb-2017) portion of the dataset for modeling the data. The sample for testing the validity of the model (test sample) contains the remaining portion, 20 % (Mar-2017 to Dec 2017) of the dataset. Since there was an obvious deterministic trend we can also use trend analysis for forecasting, but we determined that ARIMA (1,1,1) is the best model , so we compare both of them based on the estimates of MSE, MAE and MAPE so the model with least estimated is the best model which fit the data series. It is demonstrated in **Table 4** that the ARIMA (1,1,1) has a good predictive ability than Trend Analysis since its estimates is smaller than the other and this confirms that ARIMA (1,1,1) is the best model that can be used for prediction.

Table 4: Model Validation

Model Fit Indexes	ARIMA (1,1,1)	Trend Analysis
MSE	0.1109	0.1437
MAE	0.3126	0.3667
MAPE	1.2439	1.463

Finally **Figure 6** below shows the predicted Hargeisa rice prices which lie within the 95% confidence intervals. The lower confidence limit (LCL) and upper confidence limits (UCL) are also indicated. Hargeisa rice prices are expected to be increase over the time period after December 2017.

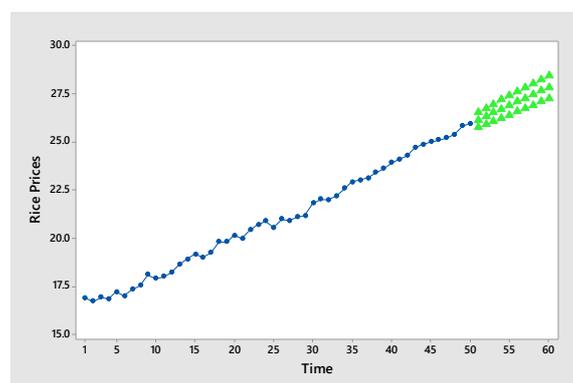


Figure 6: 10 months forecasts of Hargeisa rice price with their lower and upper confidence limits

Our study supports the report by World Food Programme (WFP), that there is a high food prices in all Somalia regions and specially Hargeisa since all rice are imported from other countries. • World Food Programme, Fews Net, Nasa Fldas (2016),

4. Conclusion

This study applied the Box-Jenkins methodology to model the Hargeisa rice prices from Nov 2013 up to Dec 2017 recorded at Ministry of National Planning and Development Hargeisa, Somalia. The time series modeling was employed by first assessing the time plot, ACF and the PACF of the series. The time plot showed increasing trend in rice prices Nov 2013 to Dec 2017. Finally, the appropriate model ARIMA (1, 1, 1) was used to forecast 2018 (12 months) for the Hargeisa rice price. The model adequacy and validation have also shown that ARIMA (1,1,1) is the most appropriate in predicting the rice prices and it was used to forecast data from Jan 2018 to Dec 2018. The forecast values fell within the required 95% confidence interval highlighting the adequacy of the fitted model. The results of the forecasting showed that the Hargeisa rice price rates were steadily increasing and this is due to imports all rice from abroad, For the above findings, there is need to integrate possible actions, efforts, programs and policies from the government, international agencies and local NGOs into Food security plans to achieve a maximum reduction of Food prices in Hargeisa city and whole of the country.

Acknowledgements

The authors would like thank the Somalia Ministry of National Planning and Development (MoNPD) for allowing access to the secondary data.

References

- Adedia D, Nanga, S, et al (2018) “Box-Jenkins’ Methodology in Predicting Maternal Mortality Records from a Public Health Facility in Ghana.”, *Open Journal of Applied Sciences*, 189-202.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). ‘Time series analysis, forecasting and control’ (3rd ed.), New Jersey: Prentice Hall, Englewood Cliffs.
- Box G.E.P. and Jenkins G.M. (1976). ‘Time Series Analysis: Forecasting and Control’. Holden-Day, revised edition, Holden Day, San Francisco.
- Barrett, Christopher (1996), “The geography of food price distributions in low-income countries”; *Journal of Development Studies*, 32(6), pp.830–849.
- Dickey, D.A. and Fuller, W.A. (1979) “Distribution of the Estimators for Autoregressive Time Series with a Unit Root,” *Journal of American Statistical Association*, vol, 74, 427-431.
- European Union, (2010), “Review and Identification of The Agriculture Programme For Somalia”, Nairobi, Kenya.
- FAO, IFAD, WFP (2015), “The State of Food Insecurity in the World, Meeting the 2015 international hunger targets: taking stock of uneven progress”, Rome, Italy.
- Gujarati, Damodar N.: (2002), “Basic Econometrics”, Fourth Edition, McGraw-Hill, New York, USA.
- Joutz, F. L. (1997). “Forecasting CPI food prices: an assessment”. *American journal of agricultural economics*, 1681-1685.
- Maddala GS. 2001. “Introduction to econometrics.”, 3rd ed. John wiley & sons (Asia) pte. Ltd, Singapore.
- Proietti, T. and Lutkepohl, H. (2011)” Does the Box-Cox Transformation Help in Forecasting Macroeconomic Time Series?”, *The University of Sydney Business School, OME Working Paper Series*, 8.
- Phillips, P.C. and Perron, P. (1988) “Testing for a Unit Root in Time Series Regression. *Biometrika*,” vol 75, 335-346.
- Somaliland Government (2011), “Somaliland Food & Water Security strategy.”, MoNPH, Hargeisa, Somalia.
- Tebbs, J.M. (2010) STAT 520 “Forecasting and Time Series”. University of South Carolina, Department of Statistics, 79-246.
- Tebbs, J.M. (2013) STAT 520 Forecasting and Time Series. University of South Carolina. Department of Statistics, 80-306.
- World Food Programme, Fewes Net, Nasa Fldas (2016), “Persistent drought in Somalia leads to major food security crisis”, Mogadishu, Somalia.