y-BIS 2019 Conference Book: Recent Advances in Data Science and Business Analytics

Mimar Sinan Fine Arts University
Department of Statistics,
Fındıklı Campus
http://ybis2019.msgsu.edu.tr

September, 25 - 28, 2019
Istanbul, TURKEY

MİMAR SİNAN FINE ARTS UNIVERSITY

Business Statistics Industry

isi

TÜRKİYE CUMHURİYET MERKEZ BANKASI

tam faktoring

Proceeding Book of the
y-BIS Conference 2019:

Recent Advances in Data Science and Business Analytics

September 25 – 28, 2019
Mimar Sinan Fine Arts University / Fındıklı Campus

Edited by

Ozan Kocadagli
Ali Erkoc
Bilge Baser
Nihan Acar Denizli
Tahir Ekin
LOC of y-BIS 2019

Web design by
Ali Zafer Dalar

Cover photo by
Aylin Yaman Kocadagli

Cover design by
Ali Mertcan Kose

Damla Ilter

Published by

Mimar Sinan Fine Arts University Publications : 884

ISBN:

978-605-5005-95-5

Serial Number:

eMSGSÜ-FEF-İST-019/09-Kat1

Contents

Part VI. Contributed Papers (Abstract)

Part VII. Contributed Papers (Full)

Part I

Preface

Welcome to y-BIS 2019 Conference: Recent Advances in Data Science and Business Analytics.

On the behalf of the Local Organizing Committee we are pleased to welcome you to y-BIS 2019 Conference: ISBIS Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business Analytics, sponsored by ISBIS (International Society for Business and Industrial Statistics) and Mimar Sinan Fine Arts University. This is the fourth conference arranged by ISBIS/y-BIS where the second one was organized in Istanbul before, 2013 Joint Meeting of Young Business and Industrial Statisticians Meeting.

The purpose of y-BIS 2019 is to bring together young statisticians and professionals working in Academia and in Industry. The conference will offer opportunities to meet each other, to share scientific and professional experiences, and to promote new collaborations and international cooperation. This conference will cover many researches in the academia and business world such as finance, medicine, insurance, energy, etc.

The program covers 10 Keynote Speakers, 8 workshops with 11 speakers, 3 invited and 16 contributed parallel sessions with 70 speakers and one poster session. We would like to thank all the speakers and, in particular the Keynote Speakers, Workshop and Invited Paper Session organizers who helped greatly to improve the scientific program of the conference.

The end of y-BIS 2019, all the presented studies have been published as a full-paper or abstract in the conference book with ISBN under the refereeing procedure and editorial policy of the conference. In addition, it is expected that, after refereeing process the selected papers will be directed to the five special issues of the journals: Applied Stochastic Models in Business and Industry, Istanbul Business Research Journal of Ambient Intelligence and Humanized Computing, Journal of Computational and Applied Mathematics and Turkish Journal of Forecasting.

The program also includes social events which will allow nearly 200 participants to know each other and to get experience of Turkish culture and history in addition to the taste of the Turkish cuisine and hospitality.

The organizers would like to thank to all the institutions that have provided financial support to make this organization possible. Many thanks to Faculty of Sciences and Letters of Mimar Sinan Fine Arts University, ISBIS-International Society for Business and Industrial Statistics, The Central Bank of Turkey and Tam Faktoring. Lastly, I really appreciate Local Organizing and Scientific Program Committees for their efforts performing on y-BIS 2019.

I am looking forward to seeing you in the next scientific events of ISI/ISBIS.

On the behalf of the Local Organizing Committee,

Ozan Kocadagli

(General Chair of y-BIS 2019)

Dear colleagues,

We are excited for your participation in the 2019 y-BIS (Young Business and Industrial Statisticians) Conference on Recent Advances in Data Science and Business Analytics.

y-BIS the Young Statisticians' group in the International Society for Business and Industrial Statistics (ISBIS), was formed in 2008. The purpose of y-BIS is to bring together young researchers and professionals working on business, financial and industrial statistics, to help support their career development.

ISBIS is an association of the International Statistical Institute (ISI) that is dedicated to the promotion of business and industrial statistics worldwide. ISBIS promotes applications, research, and best current practices in business and industrial statistics, facilitates technology transfer, and fosters communications among members and practitioners worldwide. Please visit http://www.isbis-isi.org/index.html for more information.

y-BIS has organized conferences previously in Lisbon (2012), Istanbul (2013) and Hamedan (2017). We are excited to get back to Istanbul for the fourth y-BIS Conference.

We would like to extend our sincere thanks to the organizing committee of y-BIS 2019. They have done a great job putting together a great scientific and social program. The scientific program includes a great mix of keynote speakers, short courses and sessions. We are sure that this will turn out to be a great conference.

Tahir Ekin (2017-2019 y-BIS Chair) and Luca Frigau (2019-2021 y-BIS Chair)

Committees

Ozan Kocadagli (Mimar Sinan Fine Arts University, Istanbul, Turkey)

(General Chair of y-BIS 2019)

The Local Organizing Committee

- Ali Erkoc (Co-chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Bilge Baser (Co-chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Nihan Acar (Co-Chair, Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Ali Zafer Dalar (Giresun University, Giresun, Turkey)
- Berk Kucukaltan (Trakya University, Edirne, Turkey)
- Busenur Kizilaslan (Marmara University, Istanbul, Turkey)
- Coskun Parim (Yildiz Technical University, Istanbul, Turkey)
- Damla Ilter (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Ezgi Özer (Istanbul Okan University, Istanbul,Turkey)
- Metin Yangin (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Neslihan Gokmen (Istanbul Technical University, Istanbul, Turkey)
- Selin Saridas (Mimar Sinan Fine Arts University, Istanbul,Turkey)
- Turgut Ozaltindis (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Zeynep Atli (Mimar Sinan Fine Arts University, Istanbul, Turkey)
- Zeynep Bal (Mimar Sinan Fine Arts University, Istanbul, Turkey)

The International Scientific Program Committee (Referees, Editorial Review Board)

- Alev Bakir (Turkey)
- Ali Shojaie (USA)
- Ali Erkoc (Turkey)
- Arzu Baygul (Turkey)
- Ayfer Ezgi Yilmaz (Turkey)
- Aytac Atac (Germany)
- Babak Zafari (USA)
- Bahadir elmas (Turkey)
- Balaji Raman (India)
- Benay Uzer (Turkey)
- Bilge Baser (Turkey)
- Caterina Liberati (Italy)
- Deniz Inan (Turkey)
- Elif Coker (Turkey)
- Elif Ozge Ozdamar (Turkey)
- Emilie Devijver (France)
- Emre Dunder (Turkey)
- Emre Nadar (Turkey)
- Erkan Sirin (Turkey)
- Esra Akdeniz (Turkey)
- Esra Pamukcu (Turkey)
- Ettore Lanzarone (Italy)
- Francesca Ieva (Italy)
- Fulya Gokalp Yavuz (Turkey)
- Gregor Kastner (Austria)
- Han- Ming Wu Hank (Taiwan)
- Jeff Goldsmith (USA)
- Jitka Hrabakova (Czechia)
- Kathrin Plankensteiner (Austria)
- Kristine Lurz (Germany)
- Laura Lotero- Velez (Colombia)
- Laura Trinchera (France)
- Luca Frigau (Italy)
- Marie Perrot-Dockes (France)
- Meral Yay (Turkey)
- Miguel Angel Ortiz Barrios (Colombia)
- Mustafa Murat Arat (USA)
- Naciye Tuba Yilmaz Soydan (Turkey)
- Nina Senitschnig (Austria)
- Nihan Acar Denizli (Turkey)
- Nuriye Sancar (Cyprus)
- Oguz Akbilgic (USA)
- Olawale Awe (Nigeria)
- Ozan Kocadagli (Turkey)
- Ozlem Deniz Basar (Turkey)
- Ozlem Turksen (Turkey)
- Paulo Canas Rodrigues (Brazil)
- Pedro Delicado (Spain)
- Rahim Mahmoudvand (Iran)
- Ridvan Keskin (Turkey)
- Seda Tolun Tayali (Turkey)
- Shima Mohebbi (USA)
- Tahir Ekin (USA)
- Tevfik Aktekin (USA)
- Tuba Yilmaz Gozbasi (Turkey)
- Ufuk Beyaztas (Turkey)
- Ufuk Yolcu (Turkey)
- Yasmin Said (USA)

Scientific Advisory Committee (Referees, Editorial Review Board)

- Ahmet Mete Cilingirturk(Turkey)
- Arnaldo Guimaraes Batista(Portugal)
- Alp Ustundag(Turkey)
- Ayca Cakmak Pehlivanli(Turkey)
- Aydin Erar(Turkey)
- Aylin Alin(Turkey)
- Ayse Banu Elmadag Bas(Turkey)
- Aysen Apaydin(Turkey)
- Aysen Dener Akkaya(Turkey)
- Baris Asikgil(Turkey)
- Baris Surucu(Turkey)
- Birsen Eygi Erdogan(Turkey)
- Bogdan Bichescu(USA)
- Cagdas Hakan Aladag(Turkey)
- Candan Gurses (Turkey)
- Cengiz Kahraman(Turkey)
- Chanaka Edirisinghe(USA)
- Cigden Aricigil Cilan(Turkey)
- Coskun Ozkan(Turkey)
- David Banks(USA)
- David Rios Insua(Spain)
- Dilek Onkal(UK)
- Erol Egrioglu(Turkey)
- Eylem Deniz(Turkey)
- Fabrizio Ruggeri(Italia)
- Ferhan Cebi(Turkey)
- Funda Hatice Sezgin(Turkey)
- Gerhard-Wilhelm Weber(Poland)
- Gulay Basarir(Turkey)
- Gulhayat Golbasi Simsek(Turkey)
- G. Geoffrey Vining(USA)
- Hamparsum Bozdogan(USA)
- Halis Aygun(Turkey)
- H. Kemal Sezen(Turkey)
- Jan Gertheiss(Germany)
- Martina Vandebroek(Belgium)
- Mehpare Timor(Turkey)
- Miguel Lejeune(USA)
- Mike Galbreth(USA)
- Mujgan Tez(Turkey)
- M. Levend Duransoy(Turkey)
- Nalan Cinemre(Turkey)
- Nalini Ravishanker(USA)
- Necati Aras(Turkey)
- Nurdan Colakoglu(Turkey)
- Ozgur Yeniay(Turkey)
- Refik Soyer(USA)
- Reza Langari(USA)
- Semra Erpolat Tasabat(Turkey)
- Sotiris Bersimis (Greece)
- Sukru Alp Baray(Turkey)
- Turkay Dereli(Turkey)
- Unal Halit Ozden(Turkey)
- Zahir Irani(UK)

Part II

Scientific Program

TIME	25th SEPTEMBER 2019 WEDNESDAY
08:00-09:00	REGISTRATION
09:00-09:30	OPENING CEREMONY
09:30-10:30	**Owl Hall** ***(Session Chair: Gulay BASARIR)*** **KEYNOTE SPEAKER 1 : Aytul ERCIL** (Vispera, Sabancı University) *The irreparable Rise of Artificial Intelligence*
10:30-11:00	COFFEE BREAK (POSTER SESSIONS)
11:00-12:00	**Owl Hall** ***(Session Chair: Semra ERPOLAT TASABAT)*** **KEYNOTE SPEAKER 2: Umut Satır GURBUZ** (IBM) *Preparing to Exist in the Age of Artificial Intelligence*
12:00-13:30	LUNCH
13:30-15:00	**WORKSHOP 1** **Big Data: Introduction to Hadoop big data ecosystem** Erkan SIRIN - ***Room 201*** ***Session Chair: Ali ERKOC*** **Retail Analytics with Dynamic Linear Models Using R** Balaji RAMAN - ***Room 202*** ***Session Chair: Nihan ACAR DENIZLI*** **Visualization with QlikView (How to make dashboards)** Rahim MAHMOUDVAND - ***Room 203*** ***Session Chair: Bilge BASER***
15:00-15:30	COFFEE BREAK
15:30-17:00	**WORKSHOP 2** **Introduction to Apache Spark, Data analysis and Machine Learning with Apache Spark** Erkan SIRIN- ***Room 201*** ***Session Chair: Ufuk BEYAZTAS*** **Introduction to DLM and Kalman filter, Setting up DLM in R using packages astsa, dlm and INLA, Real-life applications** Balaji RAMAN - ***Room 202*** ***Session Chair: Fatih KIZILASLAN*** **Fraud Analystics** Tahir EKIN - ***Room 203*** ***Session Chair: Arzu BAYGUL***
17:00-18:00	**Owl Hall** ***(Session Chair: Eylem DENIZ)*** **KEYNOTE SPEAKER 3: Erkal BIYIKLIOGLU** (Tam Factoring) *Financial Risk and Data Analysis*
18:30	WELCOME RECEPTION
TIME	**26th SEPTEMBER THURSDAY**
09:00-10:00	**Owl Hall** ***(Session Chair: Gulay ILONA TELSIZ)*** **KEYNOTE SPEAKER 4 : Gerhard Wilhelm WEBER** (Poznan University of Technology, Adviser to EURO Conferences) *RMARS under Cross-Polytope Uncertainty - Prediction of Natural Gas Consumption*
10:00-10:30	COFFEE BREAK
10:00-10:30	**POSTER SESSION** **The Examination of Real Estate Prices in Istanbul by Using Hybrid Hierarchical K-Means Clustering** Ilkay TUG, Betul KAN KILINC **Statistical properties and modeling of stable-like word count time series in nation-wide language data** Hayafumi WATANABE
10:30-12:00	**INVITED PAPER SESSION 1** **CLASSIFICATION BASED ALGORITHMS: METHODS AND APPLICATION** - ***Room 201 (Session Chair: Luca FRIGAU)*** **Classification-based Approach for Validating Image Segmentation Algorithms** Luca FRIGAU, Francesco MOLA **Portfolio composition strategy through a P-Spline based clustering approach** Carmela IORIO, Giuseppe PANDOLFO **Network-based Semisupervised Clustering** Giulia CONTU, Claudio CONVERSANO, Luca FRIGAU
10:30-12:00	**INVITED PAPER SESSION 2** **TUPRAS SESSION (Data Mining and Big Data Analytics in Refinery Processes)** - ***Room 202 (Session Chair: Cagla ODABASI)*** **Fault Detection and Diagnosis Methodology in Refineries: A Data-Driven Approach** Cagla ODABASI **Big Data Solutions in Refineries with Heat Exchangers** Ocan SAHIN, Cagla ODABASI
10:30-12:00	**WORKSHOP 3** **Real world applications/cases of transportation analytics-optimization with a potential demo** Tuba YILMAZ GOZBASI (Optiyol, Ozyegin University), Ozan GOZBASI (Optiyol, Bosphorus University) - ***Room 203*** ***Session Chair: Bahadır ELMAS***
12:00-13:00	LUNCH
13:00-13:50	**Owl Hall** ***(Session Chair: Baris ASIKGIL)*** **KEYNOTE SPEAKER 5 : Barıs SURUCU** (METU) *Data Analytics and Machine Learning: Real Life Applications in Various Field*
14:00 - 14:50	**Owl Hall** ***(Session Chair: Caterina LIBERATTI)*** **KEYNOTE SPEAKER 6 : HAMPARSUM BOZDOGAN (University of Tennessee)** *Robust Bayesian Relevance Vector Machines in Regression and Supervised Classification using Information Complexity and Genetic Algorithm (Application in Early Detection of Heart Attack Classification Problem)*
15:00-16:40	**STATISTICS THEORY I** - ***Room 201 (Session Chair: Mahmude Revan OZKALE)*** **A Robust Method for Estimation of Models with Random Effects** Beste Hamiye BEYAZTAS **A Percentile Bootstrap Based Method on Dependent Data: Harrell Davis Quantile Estimator vs NO Quantile Estimator** Gozde NAVRUZ, A. Firat OZDEMIR **Evaluating New Optimization Methods for Two Parameter Ridge Estimator via Genetic Algorithm** Erkut TEKELI, Selahattin KACIRANLAR, Nimet OZBAY **Stochastic Linear Restrictions in Generalized Linear Models** Mahmude Revan OZKALE **The GO estimator: A New Generalization of Lasso** Murat GENC, Mahmude Revan OZKALE
15:00-16:40	**BUSINESS/FINANCE I** - ***Room 202 (Session Chair: Ayca CAKMAK PEHLIVANLI)*** **An Approach for Considering Claim Amount and Varying** Atefeh MORADI,Maryam SHARAFI, Rahim MAHMOUDVAND **Churn Analysis for Factoring: An Application in Turkish Factoring Sector** Enis GUMUSTAS, Huseyin BUDAK **Opportunities in Location Based Customer Analytics** Murat OZTURKMEN

TIME	26th SEPTEMBER THURSDAY
15:00-16:40	**TIME SERIES/MODELING** - *Room 203* *(Session Chair: Rahim MAHMOUDVAND)* **Conditional Autoregressive Model Approach to Generalized Linear Spatial Models by CARBayes** Leyla BAKACAK KARABENLI, Serpil AKTAS ALTUNAY **Highlighting a Mathematical Property of Sample ACF for Time Series Analysis** Rahim MAHMOUDVAND **A New Approach to Econometric Modelling of Monthly Total Air Passengers: A Case Study for Ataturk Airport** Resit CELIK, Hasan Aykut KARABOGA, İbrahim DEMIR **Feature Selection Approaches for Machine Learning Classifiers on Yearly Credit Scoring Data** Damla ILTER, Ozan KOCADAGLI, Nalini RAVISHANKER
15:00-16:40	**FUNCTIONAL DATA ANALYSIS** - *Owl Hall (Session Chair: Gulhayat GOLBASI SIMSEK)* **Investigation of the Electricity Consumption of Provinces of Turkey using Functional Principal Components Analysis** Sumeyye INAL, Gulhayat GOLBASI SIMSEK **On function-on-function regression: Partial least squares approach** Ufuk BEYAZTAS, Han Lin SHANG **Wavelet Regression for Noisy Data** Gokce Nur TASAGIL, Eylem DENIZ **A Functional Data Framework to Analyse the Effect of Quinoa Consumption on Blood Glucose Levels** Nihan ACAR DENIZLI, Pedro DELICADO, Belchin KOSTOV, Diana A. DIAZ RIZOLLO, Antoni SISO, Ramon GOMIS
16:40-17:00	COFFEE BREAK
17:00-18:40	**BUSINESS/FINANCE II** - *Room 201 (Session Chair: Ipek DEVECI KOCAKOC)* **Predicting Business Survival From Their Websites** Desamparados BLAZQUEZ, Lisa CROSATO, Josep DOMENECH, Caterina LIBEATI **Fast Fault Finding Methods in Smart Manufacturing Lines with Augmented Reality Applications** Adem KAYAR, Fatih OZTURK, Ozkan KAYACAN **A Customer Segmentation Model Proposal for Hospitals: LRFM-V** Ipek DEVECI KOCAKOC, Pınar OZKAN
17:00-18:40	**APPLIED STATISTICS I** - *Room 202* *(Session Chair: Gul INAN)* **The Effect of Weights on Multi-rater Weighted Kappa Coefficients** Ayfer Ezgi YILMAZ **Probabilistic Structural Equation Modeling Approach to Investigate the Relationships Between Passenger Perceived Value, Image, Trust, Satisfaction and Loyalty** Tugay KARADAG, Gulhayat GOLBASI SIMSEK **Two Structural Equation Modelling Approaches for Cloud Use in Software Development** Erhan PISIRIR, Cuneyt SEVGI, Oumout CHOUSEINOGLOU, Erkan UCAR **Joint Modeling the Frequency and Duration of Physical Activity from a Lifestyle Intervention Trial** Gul INAN, Juned SIDDIQUE
17:00-18:40	**BIOSTATISTICS / BIOINFORMATICS** - *Room 203* *(Session Chair: Candan GURSES)* **Analyzing the Competition of HIV-1 Phenotypes with a Quantum Computation Perspective** Bilge BASER **HIV-1 Prostease Cleavage Site Prediction with Generating Dataset Using a New Encoding Scheme Based on Physicochemical Properties** Metin YANGIN, Ayca CAKMAK PEHLIVANLI, Bilge BASER **Time-Frequency Analysis of EEG Signals: Visual Identification of Epileptic Patterns** Ezgi OZER, Arnaldo Guimaraes BATISTA, Ozan KOCADAGLI
TIME	**27th SEPTEMBER FRIDAY**
09:00-10:00	**Owl Hall** *(Session Chair: Tahir EKIN)* **KEYNOTE SPEAKER 7 : Sotiris BERSIMIS (University of Piraeus)** *Using Data Analytics for Fraud Detection in Health Care: Applications and Some Results*
10:00-10:30	COFFEE BREAK
10:30-12:00	**INVITED PAPER SESSION 3** **RECENT ADVANCES ON FUNCTIONAL DATA ANALYSIS** - *Room 201 (Session Chair: Nihan ACAR DENIZLI)* **From Multivariate to Functional Classification** José Luis TORRECILLA **Functional Linear Model for Monitoring and Prediction of Profiles** Alessia PINI **Depth-Based Functional Time Series Forecasting** Antonio ELIAS, Raúl JIMENEZ
10:30-12:00	**WORKSHOP 4** **Innovation in Germany (with potential emphasis on internet of things, Supply Chain Analytics)** Aytac ATAC (Supply Chain Wizard) - ***Room 202*** ***Session Chair: Ozge CAGCAG YOLCU***
10:30-12:00	**APPLIED STATISTICS II** - *Room 203* *(Session Chair: Elif COKER)* **Serial Mediation Model of Leader Member Interaction in Work Values and Job Satisfaction** Meral YAY, Mine AFACAN FINDIKLI, Ali Mertcan KOSE **Finding the Determinants of National Problem Perceptions of Turkish Citizens** Ipek DEVECI KARAKOC, Ozlem KIREN GURLER **Analysis of Data Comparing the use of Different Social Media for Scientific Research Across Different Countries of the World** Fatima HARIS **Analysis of the Science Scores of Turkish Students in PISA 2015 via Multilevel Models** Elif COKER **A StarCraft 2 Player Skill Modeling** Natasa A. CIROVIC, Zoran Z. CIROVIC
12:00-13:30	LUNCH
13:30-14:30	**Owl Hall** *(Session Chair: Birsen EYGI ERDOGAN)* **KEYNOTE SPEAKER 8 : Bahar KINAY ERGUNEY** (CISCO) *Big Data and IoE*

14:30-15:00	COFFEE BREAK
15:00-16:40	**CLUSTERING / CLASSIFICATION** - *Room 201* *(Session Chair: Nurdan COLAKOGLU)* **Hierarchically Built Trees with Probability of Placing Clusters** Nebahat BOZKUS, Stuart BARBER **Comparison of Internal Validity Indices According to Distance Measurements in Clustering Analysis** Derya ALKIN, Aydin KARAKOCA, Ibrahim DEMIR **Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases** Nurdan COLAKOGLU, Berke AKKAYA **An Application of XGBOOST on Diabetes Dataset** Gulcin YANGIN, Elif Ozge OZDAMAR **How Does Resampling Affect the Classification Performance of Support Vector Machines on Imbalanced Churn Data** Serra CELIK , Seda TOLUN

TIME	27th SEPTEMBER FRIDAY
15:00-16:40	**STATISTICS THEORY II** - *Room 202 (Session Chair: Berk KUCUKALTAN)* **Stress-Strength Reliability Estimation of Series System with Cold Standby Redundancy at System and Component Levels** Gulce CURAN, Fatih KIZILASLAN **Statistical Inference of Consecutive k-out-of-n System in Stress-Strength Setup Based on Two Parameter Proportional Hazard Rate Family** Duygu DEMIRAY, Fatih KIZILASLAN **Approximation of Continuous Random Variables for the Evaluation of the Reliability Parameter of Complex Stress-Strength Models** Alessandro BARBIERO **Chaos Control in Chaotic Dynamical Systems Via Auto-tuning Hamilton Energy Feedback** Atike Reza AHRABI, Hamid Reza KOBRAVI
15:00-16:40	**WORKSHOP 5** **Medical Analytics / Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with** Arzu BAYGUL, Cagdas AKTAN and Neslihan GOKMEN - ***Room 203*** ***Session Chair: Esra AKDENIZ***
16:40-17:10	COFFEE BREAK
17:10-18:50	**REGRESSION/FUZZY MODELING** - *Room 201 (Session Chair: Ozlem TURKSEN)* **A Seemingly Unrelated Regression Modeling for Extraction Process in Green Chemistry** Ozlem TURKSEN, Serhan TUNCEL, Nilufer VURAL **The Effect of WoE Transformation on Credit Scoring By Using Logistic Regression** Zeynep BAL, M. Aydin ERAR **Bivariate Intuitionistic Fuzzy Time Series Prediction Model** Ozge CAGCAG YOLCU, Erol EGRIOGLU, Eren BAS, Ufuk YOLCU **Nonlinear Neural Network for Cardinality Constraint Portfolio Optimization Problem: Sector-wise analysis of ISE-all Shares** Ilgım YAMAN, Turkay ERBAY DALKILIC **Statistical and Fuzzy Modeling of Extraction Process in Green Chemistry** Nilufer VURAL, Ozlem TURKSEN
17:10-18:50	**BUSINESS/FINANCE III** - *Room 202* *(Session Chair: Esra N. KILCI)* **Bitcoin Cash: Returns Distributions and Dissimilarity Analysis** Muhammad SHERAZ, Vasile PREDA, Silvia DEDU **Do Confidence Indicators Have Impact on Macro-financial Indicators? Analysis on the Financial Services and Real Sector Confidence Indexes:** Esra N. KILCI **Granger-Causality-Based Portfolio Selection In The Moroccan Stock Market** Abdelhamid Hamidi ALAOUI
17:10-18:50	**STATISTICS THEORY III** - *Room 203* *(Session Chair: Nursel KOYUNCU)* **Multivariate Skew Laplace Normal Distribution: Properties and Applications** Fatma Zehra DOGRU, Olcay ARSLAN **Fitting lognormal distribution to actuarial data** Mahdi MAHDIZADEH, Ehsan ZAMANZADE **Gamma and Inverse Gaussian Distributions in Fitting Parametric Shared Frailty Models with Missing Data** Nursel KOYUNCU, Marthin PIUS, Nihal ATA TUTKUN
20:00	GALA DINNER
TIME	**28th SEPTEMBER SATURDAY**
08:45-09:30	**Owl Hall** *(Session Chair: Elif Ozge OZDAMAR)* **KEYNOTE SPEAKER 9 : Selim DELILOGLU (Data Analyst at Telecommunication Sector)** *Fundamental Skills for Data Science & Business Analytics*
09:30-10:15	**Owl Hall** *(Session Chair: Meral YAY)* **KEYNOTE SPEAKER 10 : Ayse OZMEN** *Multi-objective Sparse Regression Models for short- and long-term Natural Gas Demand Prediction*
10:15-10:30	COFFEE BREAK
10:30-12:00	**WORKSHOP 6** **Hands-on Introduction Course in R(Acquiring data from different sources on command line and R, data pre-processing, a map package to map one of the up-to-date data (potentially with 2019 Turkish local election data), SQL in R** Fulya GOKALP (METU) - **Owl Hall** **Session Chair: Deniz INAN**
10:30-12:00	**BUSINESS/FINANCE IV** - *Room 201* *(Session Chair: Mujgan TEZ)* **Risk-based Fraud Analysis for Bank Loans With Autonomous Machine Learning** Yunus Emre GUNDOGMUS, Mert NUHUZ, Mujgan TEZ **Bivariate Credibility Premiums Distinguishing Between Two Claims Types in Third Party Liability Insurance** Pervin BAYLAN, Serdar KURT, Neslihan DEMIREL, Jeffrey S. PAI **Methods for Optimum Establishment of Government-imposed global budget caps -Monitoring Pharmaceutical Expenditures using SPM** Nika ELISAVET **Prediction of Claim Probability in the Presence of Excess Zeros** Aslihan SENTURK ACAR
10:30-12:00	**OUTLIER DETECTION** - *Room 202* *(Session Chair: Erkan SIRIN)* **Detection and Handling Outliers in Longitudinal Data: Can Wavelets Decomposition Be a Solution?** Marwa BENGHOUL, Berna YAZICI, Ahmet SEZER **Outlier Detection on Big Data** Erkan SIRIN, Hacer KARACAN **Identification of Vehicle Warranty Data and Anomaly Detection by Means of Machine Learning Methods** Halil Ibrahim CELENLI, Esin OZKAN
10:30-12:00	**OPTIMIZATION/DECISION MAKING** - *Room 203* *(Session Chair: Semra ERPOLAT TASABAT)* **Alternative Subway Project Selection with TOPSIS Method Using Different Weighting Techniques** Nihan YUCEL, Semra ERPOLAT TASABAT **Recycle Project with RFM Analysis** Esra AKCA, Semra ERPOLAT TASABAT **Inferences About Development Levels of Countries with Data Envelopment Analysis** Semra ERPOLAT TASABAT

12:00-12:30	CLOSING
14:00-17:30	SOCIAL EVENT (Bosphorus Boat Tour)

Part III

Keynote Lectures

Robust Bayesian Relevance Vector Machines in Regression and Supervised Classification Using Information Complexity and the Genetic Algorithm

Hamparsum Bozdogan, Ph.D.
McKenzie Professor
The University of Tennessee, U.S.A.
bozdogan@utk.edu

Support Vector Machines (*SVMs*) have been popular kernel methods in regression and classification applications. However, *SVMs* suffer from a number of limitations. In this talk, we propose a new and novel model selection in *Bayesian Relevance Vector Machines* (*BRVMs*) in regression and classification problems (Tipping, 2001). *BRVM* is a sparse kernel technique, which is an improvement over the *SVMs* from the Bayesian learning perspective, while avoiding the limitations that exist in *SVMs*. Unresolved model selection issues in *BRVM* regression and classification include:*choosing the optimal form of the kernel function* among a portfolio of kernel choices for a substantive data set; the *parameters of the kernel function*; and the *subset selection of the best predictor variables* in regression and classification.

We introduce novel statistical modeling techniques based on the *information-theoretic measure of complexity called ICOMP criterion* developed by Bozdogan (1990, 1994,2000-2019) as the fitness function hybridized with the *genetic algorithm* (*GA*) as our *optimizer* to perform the model selection. *ICOMP* allows the identification of the best fitting kernel function or functions among a large portfolio of kernel functions. It measures both the *lack-of-fit* (*LOF*) and the *complexity* of the *BRVM* models. The *genetic algorithm* (*GA*) enables the rapid computation of models that would otherwise be impossible in a reasonable amount of time for subset selection of best predictor variables for high-dimensional data.

We illustrate the advantages of this new approach on simulated and on real benchmark data sets in regression and classification problems including the classification of cardiac imaging of diseased aortic tissues for early detection of the cause of heart attack.

As a conclusion, we discuss how to robustify *BRVMs* using general distributional models along with smooth and flexible priors to enforce a stronger sparsity in the model to achieve further *Occam's Razor* in regression and classification problems.

Keywords: Bayesian Relevance Vector Machines; Model Selection; ICOMP and Genetic Algorithm.

Multi-objective Sparse Regression Models for short- and long-term Natural Gas Demand Prediction

Ayşe ÖZMEN [*1]
[1] Department of Mathematics and Statistics, University of Calgary, Canada

In this study, as an innovative contribution to NG demand forecasting for short and long term, new Conic Multivariate Adaptive Regression Splines (CMARS) [2,3] and SINDy (Sparse Identification of Nonlinear Dynamics) algorithm of LASSO [1] are applied as multi-criteria nonlinear models in prediction of daily, weekly and yearly natural gas consumption of residential users in Ankara City; the gas is distributed by the Başkentgaz DSO company. Also, Tikhonov Regularization (TR) and LASSO are applied as multi-criteria linear models. Then, the results of these methods are compared with the results of exciting model Linear Regression (LR) and Artificial Neural Network (ANN).

This study is based on a real dataset which is divided into training and testing data sets. The training set involves daily input variables of the period 2004–2009. The test set has inputs from 2010 to 2013. Predictor variables are given to the model, including daily meteorological data, previous-day daily gas consumption data, the number of residential users and other supplementary inputs. Here, daily minimum and maximum temperatures and HDD values are used together to obtain natural gas forecasting models with our tools. Here, each sovereign model has the capability to provide the daily consumption of Ankara, for all four seasons and without creating separate models for winter and summer terms. In this study, as sparse regression-based methods, LASSO, SINDy and CMARS models are obtained for demand forecasting and they are compared to regression-based methods LR, TR and ANN. It is observed that CMARS significantly outperforms the other approaches based on MAPE values for all the forecasting intervals. Regarding the maxAPE values, CMARS still outperforms LASSO (in linear and nonlinear cases), TR and LR.

Keywords: Natural Gas Consumption, CMARS, Conic Quadratic Programming, LASSO

References

[1] Brunton, S. L., Proctor, J. L., and Kutz, J. N (2016). Discovering Governing Equations from Data: Sparse Identification of Nonlinear Dynamical Systems. Proceedings of the National Academy of Sciences, 113 (15), 3932-3937.

[2] Taylan P., Weber G.-W., and Beck A. (2007). New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. Optimization, 56, 5-6.

[3] Weber, G.-W., Batmaz, I., Köksal, G., Taylan P., and Yerlikaya- Özkurt, F. (2012). CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimisation. Inverse Problems in Science and Engineering (IPSE), 20, 371-400.

RMARS under Cross-Polytope Uncertainty – Prediction of Natural Gas Consumption

Gerhard-Wilhelm Weber [*1], Ayşe Özmen [2] and Yuriy Zinchenko [3]
[1] Faculty of Engineering Management, Poznan University of Technology, Poznan, Poland, and IAM, METU, Ankara, Turkey
[2] Mathematics and Statistics, University of Calgary
[3] Math and Stat, University of Calgary

Multivariate Adaptive Regression Spline (MARS) is a modern methodology of data mining, statistical learning and estimation theory that is essential in both regression and classification. In recent years, MARS is applied in various areas of science, technology, finance, and engineering. It is a form of flexible non-parametric regression analysis capable of modeling complex data. There, it is supposed that the input data are known exactly and equal to some nominal values to construct a model. However, both output and input data include noise in real life. Solutions to optimization problems may present significant sensitivity to perturbations in the parameters of the problem. So, optimization affected by parameter uncertainty is a focus of the mathematical programming and a need to handle uncertain data when optimization results are combined within real-life applications. As a result, in inverse problems of modeling, solutions to the optimization problems involved in MARS can represent a remarkable sensitivity with respect to perturbations in the parameters which base on the data, and a computed solution can be highly infeasible, suboptimal, or both. Under this motivation, we have included the existence of uncertainty into MARS and robustified it through robust optimization which is proposed to cope with data and, hence, parametric uncertainty. We have represented Robust MARS (RMARS) under polyhedral uncertainty.In our previous studies, although we had small data sets for our applications, the uncertainty matrices for the input data had a huge size since vertices were too many to handle. Consequently, we had no enough computer capacity to solve our problems for those uncertainty matrices. To overcome this difficulty, we obtained different weak RMARS (WRMARS) models for all sample values (observations) applying a combinatorial approach and solved them by using MOSEK program. Indeed, we have a tradeoff between tractability and robustification. In this presentation, we present a more robust model using cross-polytope and demonstrate its performance with the application of Natural Gas consumption prediction. Applying robustification in MARS, we aim to reduce the estimation variance.

Keywords: Robust Optimization; Machine Learning; Mathematical Programming; Energy Sector

References

[1] Özmen, A., Batmaz, I., and Weber, G.-W. (2014). Precipitation Modeling by Polyhedral RCMARS and Comparison with MARS and CMARS, Environmental Modeling and Assessment 19 (82) 425–435.
[2] Özmen, A., Weber, G.-W., Çavuşoğlu, Z., and Defterli, Ö. (2013). The new robust conic GPLM method with an Application to Finance: prediction of credit default, Journal of Global Optimization (JOGO) 56 (2) 233-249.
[3] Özmen, A., Yilmaz, Y., and Weber, G.-W. (2018). Natural Gas Consumption Forecast with MARS and CMARS Models for Residential Users, Energy Economics 70, 357–381.

[*]Corresponding author: gerhard-wilhelm.weber@put.poznan.pl

Data Analytics and Machine Learning: Real Life Applications in Various Fields

Barış Sürücü

Department of Statistics, Orta Doğu Teknik Üniversitesi

Data analytics and the concept of data science are of great concern in many fields. Rapidly developing technology has significantly increased the amount of data and hence the importance of data analytics. In this talk, we will talk about some basics about data analytics, machine learning and some real-life applications, such as modelling in industry 4.0, finance, technology, health, economy fields as well as decision support systems developed for government institutions and private companies. We will discuss how we deal with data preparation and how we model the data by using machine learning algorithms. Different types of machine learning algorithms will also be discussed in this context.

Keywords: Machine Learning; Statistical Methods; Industry 4.0; Decision Support Systems

References

[1] Hastie, T; Tibshirani, R; Friedman, J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag New York.

[2] Mitchell, T. M (2017). Machine Learning, Cloudtail India.

*Barış Sürücü: sbaris@metu.edu.tr

The Imperable Rise of Artificial Intelligence and How it Effects our Lives

Aytul Ercil [*1]

[1] Vispera Information Technologies A.S.

In this talk we will summarize the state of the art in Artificial Intelligence (AI) technology, the reasons behind the recent tremendeous increase in the applications of AI technologies. We will talk about the different stages of Artificial Intelligence, namely Artificial Narrow Intelligence, Artificial General Intelligence, Artificial Super Intelligence and the components of Artificial Intelligence Systems.

We will give examples of AI technology in various industries and discuss the effects of AI technologis on our business life, economy, labour force and the society. We will end with discussing legal, ethical and security issues

Keywords: Artificial Intelligence; Image Processing; Machine Learning

[*]Corresponding author email: a.ercil@vispera.co

Fundamental Skills for Data Science & Business Analytics

Selim Deliloglu

While the data science and business analytics fields are constantly developing, it has become a necessity for the experts working in these fields to constantly develop themselves and to shape their skills according to the needs. Projects in these areas need technical skills as well as non-technical skills to guarantee the success of the projects. In the speech, critical technical and non-technical skills and why these skills are so important to the success of the projects will be emphasized. Finally, skills can be used at each stage of a possible project scenario in the telecommunication sector will be discussed.

Financial Risk and Data

Erkal Biyiklioglu

This presentation mainly focused on a real life application of the data analytics for credit risk management in factoring business. Even there are many studies and reports on how useful data science can be applied in real life, most real life application experiences are not common as academic studies. This report underlines the importance of centralized data aggregation and sharing in finance (banking) sector for data science. Biggest problem of real life applications of data analytics is lack of full data set, most of the time companies have only data generated by themselves. In this report, it is explained that how credit risk data has been collected and shared by K.K.B. (Kredi Kayıt Bürosu) and its usage for very low NPL (non-profit loans) in factoring business.

Also in this report, the human side of the data analytics in business is another subject. How the main ingredients of data science such as leadership, empathy, communication, curiosity should be supported in the analytics teams is explained by example. Human factor in data science application in business is underlined.

Keywords: Credit Risk in Factoring, Data Analytics in risk management, Data science real life application, Human factor in data analytics.

References: Tam Faktoring KIOS system algorithm.

erkalbiyiklioglu@tamfaktoring.com.tr

Preparing to Exist in The Age of Artifical Intelligence

Umut ŞATIR GÜRBÜZ

In this session we will review the latest trends in the Artificial Intelligence, Data Science and Machine Learning areas, discuss the reason behind the increasing importance of data and it's timing, learn about the business applications of these technologies or disciplines in various industries, the list of skills that we need to invest our time on, and the roadmap to prepare ourselves for this competitive job market.

Keywords: Data Science, Artificial Intelligence, Machine Learning

Big Data and IoE

Bahar Kınay Erguney*
*CISCO, TURKEY.

This presentation is to be able to explain what is IoT (IoE) and how IoT will cause and collect data in Cloud or Data Centers and also how to apply Big Data in Organizations. Big Data does require a paradigm shift, and knowing the realities will help organizations move forward[1]. These eight realities that your organization should know and take into consideration are culture, the people within organization, Big Data is at everywhere, Big Data Engineers, security, privacy, government, amount of data. So to be able to feed this Big Data, the most important partner is IoT. What is IoT? The Internet of Things (IoT)is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment[2]. IoT is historically used in Cambridge University; the famous and well-known coffee pot at Cambridge University demonstrated the first use of a public webcam. Rather than navigate several flights of stairs to see if their coffee had finished brewing, researchers set up a camera that sent images of the coffee pot to their computers over the internet.

Now IoT is also mentioned as New Gold or New Petrol. If we talk about the architecture and steps of IoT, then we should know about the steps; data detection, network label, data storage, evaluation and implementation. Although Big Data and IoT came into existence separately, with this relationship we cannot spare them anymore as the will continue to feed each other hereupon.

Internet of Things started to be popular in 2015s with the electronic exposition. Although the reason seems like satisfying the needs of people, the reality of this popularity is coming from costs. According to Intel, the global market for IoT technology will reach $6.2 trillion by 2025. Businesses and manufacturing will account for 40% of that. Another 30% will come from healthcare, which will rely heavily on IoT.

In the world, there is a huge investment to IoT for all sectors, such as smart cities, smart offices, smart homes, manufacturing, health etc. As an example, Cisco systems has many product and solutions; such as TelePrecense, WebeX Teams and Webex to help people and organizations to connect people from all over the world with a monitor or phone.

So what organizations should do with Big Data and IoT. First of all, they should connect all devices to Internet and make them Smart. Then the correct network infrastructure should be positioned. And with this network, they should collect data to data centers or cloud. The last but most important step is to make correct analysis with collected data and make them useful information for their organizations.

Keywords: Big Data; Internet of Things; IoT, IoE, IoP

References

[1] van Rijmenam, Mark (2014). Think Bigger, Devoloping a Succesful Big Data Strategy for your Business, 271-356.
[2]https://www.cisco.com/c/dam/en/us/products/collateral/se/internet-of-things/at-a-glance-c45-731471.pdf?dtid=osscdc000283

*Bahar Kinay Erguney: baharkinay@gmail.com

Part IV

Short Courses

Visualization with QlikView (how to make dashboard)

Rahim Mahmoudvand [*1]

[1] Bu-Ali Sina University, Hamedan, Iran

The use of visualization to present information is not a new phenomenon. Simple forms of visualization have used in maps, scientific drawings, and data plots for over a thousand years. However, the recent emphasis on visualizations started after 1970 with releasing statistical and business intelligence (BI) software. In this new era, organizations have too much data with their work and prefer to use a fast technique to see what happen in their work.

A dashboard is a type of graphical user interface, which often provides at-a-glance views of key performance indicators relevant to a particular objective or business process. In other usage, "dashboard" is another name for "progress report" or "report." There are several methods to design a dashboard.

In this course, I use QlikView, which is BI software and can be downloaded via https://www.qlik.com/us/try-or-buy/download-qlikview. I show participants how they can construct dashboard-using QlikView. To do this, I use three data files including birth, death and marriage-divorce and show them how they can construct a simple dashboard. This simple dashboard is indeed an alternative of a report around 150 pages and this might be very amazing for participants.

Keywords: Dashboard; Object; Chart; Text; Input; List; Button

References

[1] Pover, K. (2016). Mastering QlikView Data Visualization. Packt Publishing Ltd, UK.
[2] Tukey, J. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.
[3] Beniger, J.R., Robyn, D.L. (1978). Quantitative Graphics in Statistics: A Brief History. The American Statistician Vol. 32, No. 1, pp. 1-11

[*]Corresponding author: r.mahmodvand@gmail.com

Fraud Analytics

Tahir Ekin

Texas State University
tahirekin@txstate.edu

Fraud has been around since the early days of commerce, continuously evolving and adapting to changing times. Fraud instances can be seen in a wide range of domains such as finance, credit card, telecommunications, insurance and healthcare. For instance, in healthcare, overpayments are estimated to correspond up to 10 percent of expenditures. This short course aims to present how data and statistical methods are used for fraud assessment. The fraud data and types will be introduced followed by a discussion of the data pre-processing techniques. Next, the course will cover the use of unsupervised and supervised methods. Interested attendees can refer to Bolton and Hand (2002), Ekin et al. (2018) for related overviews and Ekin (2019) for a comprehensive discussion of health care fraud analytics. The course is at an introductory overview level, and aims to provide both practical and technical insights. No statistical background is required, but it would be helpful if the attendees know the basics of probability and descriptive statistics.

Keywords: Fraud; Financial Fraud; Health Care Fraud; Statistics; Analytics; Sampling.

References

[1] Ekin, T. (2019). Statistics and Health Care Fraud: How to Save Billions. CRC Press.

[2] Ekin, T., Ieva, F., Ruggeri, F., & Soyer, R. (2018). Statistical medical fraud assessment: exposition to an emerging field. International Statistical Review, 86(3), 379-402.

[3] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 235-249.

Dynamic Linear Models (DLM) using R

Balaji Raman

Cogitaas AVA

Applications of Dynamic Linear Models (DLM) are widespread. These models are regularly used in domains like econometrics, marketing, finance, social sciences and retail. This workshop will focus on implementing DLMs in R using two packages dlm and R-INLA. We will discuss examples ranging from simple univariate time series to hierarchical time series.

Real World Applications/Cases of Transportation Analytics-Optimization

Tuba Yılmaz Gözbaşı[1,2], Ozan Gözbaşı[2]

[1] Ozyegin University
[2] Optiyol

In this talk, we will talk about transportation optimization solutions and different use cases: (i) pickup/delivery route optimization, (ii) periodic/static route optimization, (iii) intercity transportation optimization, (iv) passenger shuttle routing. We will summarize the main inputs and outputs for each of these use cases, and provide case studies from practical applications.
We will discuss the advances in transportation optimization through integrating driver preferences, business strategies, variability of demand profiles and traffic patterns to create both efficient and practical route plans in an autonomous fashion through unique algorithms and learning mechanism from the field data.

Introduction to Apache Spark, Data Analysis and Machine Learning with Apache Spark

Erkan Şirin[*1]

[1] Ph.D. Student, Institute of Information, Gazi University, Ankara, Turkey
erkan.sirin@gazi.edu.tr

In this workshop, we will understand the basic concept of Apache Spark. We will do some practice on some Spark APIs like dataframe and machine learning.

Apache Spark is a fast and general-purpose cluster computing framework[1, 2]. Spark is the successor of MapReduce. The advantage over MapReduce is that it makes good use of memory. This advantage is also a disadvantage for Spark. Because if Spark is not well configured, the most common error is a memory error. Having advanced APIs, dozens of publications, books, and documentation Spark is the brightest project of big data processing. It doesn't store data but connects, reads and writes almost any data store.

Apache Spark has different components and APIs for different needs. Initially, Spark came out with the Resilient Distributed Dataset (RDD) API. After a while, developers realized that it was difficult for users to write code with RDD API and that the written applications could not demonstrate Spark's potential performance. Therefore, the high-level dataframe/dataset API was developed. Now new developments continue through this API. Spark switched to the high-level API not only in the dataframe/dataset but also in the machine learning. Thanks to Spark SQL, SQL queries can be performed using Spark on large data sets. In fact, it is possible to use Spark, even with SQL, without knowing almost any code.

One of Spark's most outstanding aspects is that it is possible to train machine learning models on large data sets by successfully running distributed machine learning algorithms. The Spark MLlib library has a widely used algorithm, from regression to classification and from there to clustering.

Keywords: Spark; MLlib; RDD; Dataframe; SparkSQL; Distributed Computing

References

[1]Foundation, A.S. Apache Spark. 2019; Available from: https://spark.apache.org/docs/latest/.

[2]Zaharia, M., et al., Fast and interactive analytics over Hadoop data with Spark. USENIX; login, 2012. 37(4): p. 45-51.

Big Data: Introduction to Hadoop Big Data Ecosystem

Erkan Şirin*[1]

[1] Ph.D. Student, Institute of Information, Gazi University, Ankara, Turkey
erkan.sirin@gazi.edu.tr

In this workshop, we will understand the basic concept of big data and Hadoop ecosystem. We will do some practice on basic Hadoop components like HDFS, YARN, and related projects like Hive and Kafka. We install on our computer Virtualbox and Cloudera Quickstart VM. We have Spark installed in VM we also install Spark out of VM on our host operating system.We will discuss why big data exist, the definition of big data, use cases, how to handle big data opportunities and hardships that big data brings.Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model[1]. Hadoop has its core modules and related projects. Core modules are Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, Hadoop MapReduce. Recently Hadoop Ozone and Submarine have added to Hadoop. There are plenty of Hadoop related projects like Spark, Hive, Ambari, HBase, etc.HDFS is a distributed java based file system, fault-tolerant and allows to efficiently write, read and store data that has big data characteristics in an affordable way. It has two main services namenode and datanode. Namenode manages metadata while datanode is responsible to manipulate data itself. Namenode doesn't store any real data but knows every single detail of it. HDFS is a core Hadoop module.YARN manages cluster resources. What is called a source is actually a container that consists of processor and memory. Prior to YARN, resource management at Hadoop-1 was done by MapReduce. With Hadoop2, YARN has begun managing resources and saved HDFS operations from being sentenced to MapReduce. YARN has three main services; resource manager, node manager, and application master. Resource manager schedules resources among user and application and is located master node. Node managers are run every slave node and can be considered representative of resource manager. Application master is created per application and dies when the job is done.Apache Hive enables HDFS operations to be executed using SQL like language called HiveQL[2, 3]. In this way, it makes the more rooted and widespread SQL capabilities used in the big data world. Hive is widely used as a data warehouse. Hive uses a schema for data stored in HDFS, enabling database-like operations. However, it cannot be considered a relational world-class database. Apache Kafka is a distributed streaming platform[4]. Kafka is widely used, especially in real-time data processing applications. Many applications and services that need to exchange data with each other can achieve this exchange without any complexity by using Kafka. Kafka is very similar to the enterprise message queue. It can store messages in a fault-tolerant manner without losing a limited time.

Keywords: Hadoop; Spark, Hive; Kafka

[1]Foundation, A.S. Apache Hadoop. 2019 [cited 2019 2019-09-22]; Available from: https://hadoop.apache.org/.
[2]Thusoo, A., et al., Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment, 2009. 2(2): p. 1626-1629.
[3]Foundation, A.S. Apache Hive. 2019; Available from: https://hive.apache.org/.
[4]Foundation, A.S. Apache Kafka. 2019; Available from: https://kafka.apache.org.

Innovation in Germany
Industry 4.0 Case Examples

Aytac Atac, Ph.D.

Vice President EMEA & APAC, Supply Chain Wizard

The forth industry revolution has arrived, very fast and for some unexpectedly. Germany has been one of the countries that is defining and driving the technological enhancements within Industry 4.0. With this workshop, the systematical approach of Germany will be analyzed and some practical case examples will be presented that demonstrates the visible benefits of industrial transformation.

For many industrial applications, automation has been defined in the last decade. This has formed more automatic and self-driven systems. However, this also formed silos within organizations: islands systems that are not communicating with each other.

With Industry 4.0, the silos systems are now being connected. The connection opens up endless possibilities, as now there is enormous amount of data that can be correlated, further analyzed and even be used to form smart decisions. Internet of Things (IoT) plays an important role at this automation, as well as concepts like Machine Learning and Artificial Intelligence.

This workshop aims to present the structured approach German is taking in defining the next industrial revolution as well as the practical benefits of concepts like IoT and Machine Learning.

Keywords: Industrial Revolution; Industry 4.0; Internet of Things, Machine Learning; Supply Chain

Medical Analytics/Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with potential discussion of cancer molecular)

Arzu Baygul[1], Cagdas Aktan[2], Neslihan Gokmen[3]

[1] Department of Medicine, Koc University
[2] Department of Medical Biology, Beykent University
[3] Department of Basic Sciences, Istanbul Technical University

Statistics helps us to organize, understand and make sense of data that we collect from the study. We use statistics in collecting data, describing, analyzing data, interpreting and drawing conclusions from data and for predicting future situations. Biostatistics is where Medicine and Statistics meet. In other words, Biostatistics is a statistical approach, that is used in health data.

The aim of this study is doing a short introduction to statistics and biostatistics, to reveal differences between statistics and biostatistics. Discussing about the areas where biostatistics is used. One another important topic of this study is interaction points between biostatistician and clinician.

Biostatistics is an intersection of medicine and mathematics. The methods in biostatistics are specified for medical data, which are obtained from vital things and nature. So biostatistics helps us to understand the medicine, reasons and results of diseases. The generally used areas of biostatistics are planning and evaluating healthcare services, finding new diagnostic methods, exploring reasons of medical problems by clinical researches, discovering epidemiological characteristics of populations. Biostatistical approach should be used at study design, sample size calculation, ethics committee application, analyzing and interpreting the results.

Clinical research is a branch of healthcare science that determines the safety and efficacy of medications, devices, diagnostic products and treatment regimens intended mostly for human use. These may be used for prevention, treatment, diagnosis or for relieving the symptoms of a disease. The first step to conduct clinical research is to obtain approval from the ethics committee. The main responsibility of a research ethics committee is to protect potential participants in the research, but it must also take into account potential risks and benefits for the community in which the research will be conducted. The application file should contain detailed research project report, researcher resumés, informed consent forms and budget of the study. In order to prepare a detailed project report, it is necessary to determine the research design. Clinical research design is divided into two as experimental and observational.

Bioinformatics is an interdisciplinary field that develops methods, techniques for collection, classification, organizing, storing and analyzing biological data. It integrates mathematics, statistics, engineering, computer science, and medical science to analyze and interpret the data. There are many bioinformatics tools. Some of these tools help researchers in the comparison of genomic data. At a more integrative level, they helps analyze and catalogue the biological pathways and networks such as gene-gene, protein-peptide interactions that are an important part of systems biology. In structural biology, they helps in the simulation and modeling of nucleic acids, drug and protein structures as well as molecular interactions. Moreover, at clinical level bioinformatics may be changed diagnostics, therapeutics and prognosis strategies, may be precisely tailored to patient-specific needs for personalized medicine.

Keywords: biostatistics; clinical research; bioinformatics

Hands-on Introduction Course in R

Fulya Gökalp Yavuz [*1]

[1] Middle East Technical University, Department of Statistics, Ankara, Turkey

There is a common sense on the importance of Big Data and Data Science topics in both 'Academia' and 'Industry'. R is one of the interactive application tools which handles the data manipulation, statistical analyses and visualization under the scope of these contemporary topics. One of the challenges for Data Science is acquiring the larger data sets from different sources. In this course, we build a bridge between the command line and R to get the data sets larger in size. After covering some data pre-processing, a map package will be used to map one of the up-to-date data sets. Lastly, we will cover some basics for SQL in R. All subjects covered in this course will be examined through examples to explore the program.

For the mentees, a laptop with command line (UNIX/LINUX) included, R and RStudio installed will provide the ease of participation, learning and application. This course welcomes all of you who want to add some more to your skills on LINUX and R programming.

Please provide the following tools before attending the course:

Visit the R website: https://www.r-project.org and download R. RStudio is a free, open-source, integrated development environment (IDE) for R. To learn more about RStudio and download a copy, see http: //www.rstudio.org.

Please install "cygwin" to use LINUX distribution on Windows from the following link:
http://www.cygwin.com.

Use the setup program and you are ready to use LINUX on your Windows. Apple users may use "iTerm" located on their IOS.

[*]Corresponding author: fgokalp@metu.edu.tr

Part V

Invited Sessions

Classification-based Approach for Validating Image Segmentation Algorithms

Luca Frigau*[1], Francesco Mola[1],Giulia Contu[1]

[1] Department of Economics and Business Sciences

In computer vision, image segmentation is a process that partitions an image into different objects of interest. Despite it is a trivial activity (almost always) to the human vision, in image processing, image understanding and artificial intelligence that is one of the most demanding problems [1]. In literature several image segmentation algorithms have been developed [2], which follow both global and local approach. Nonetheless, to assess and consequently to compare their outputs is still challenging.

Specifically, the image segmentation algorithms can be compared considering three factors: reliability, viability and validity [3]. Among these three factors, validity is the most challenging one, since validating the output of a segmentation algorithm is very important, both for users of algorithms and for algorithm developers.

We propose an approach for validating the image segmentation algorithms that ranks the performances of two or more outputs obtained from different image segmentation algorithms, consequently being able to define the best one.

This is a classification-based approach that ranks the outputs of different segmentation algorithms by performing machine learning classifiers. Furthermore, it takes into accounting for both the computational complexity of the validation experiment and for the robustness of its results: Fisher consistent estimates are obtained with sample of pixels of extremely-reduced size by using a subsampling approach.

This image validation approach has been tested on several real images differing each other in terms of shape, color and texture.

Keywords: Image Validation; Subsampling, Machine Learning

References

[1] David G. Lowe. Distinctive image features from scale-invariant key- points. International Journal of Computer Vision, 60(2):91–110, Nov 2004. ISSN 1573-1405. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[2] HP Narkhede. Review of image segmentation techniques. International Jour- nal of Science and Modern Engineering, 1(8):54–61, 2013.

[3] Jayaram K Udupa, Vicki R Leblanc, Ying Zhuge, Celina Imielinska, Hilary Schmidt, Leanne M Currie, Bruce E Hirsch, and James Woodburn. A frame- work for evaluating image segmentation algorithms. Computerized Medical Imaging and Graphics, 30(2):75–87, 2006.

*Corresponding author: frigau@unica.it

Portfolio Composition Strategy through a P-Spline Based Clustering Approach

Carmela Iorio[1], Giuseppe Pandolfo[1]

[1] Department of Industrial Engineering, University of Naples Federico II, Italy
carmela.iorio@unina.it, giuseppe.pandolfo@unina.it

Portfolio management and asset selection are important issues in the financial domain. Portfolio composition is concerned with the problem of making portfolio decisions by selecting the securities to include in the portfolio and the amount to invest in each security. Modern portfolio theory aims at building an optimized portfolio by selecting stocks with the highest expected return for a given level of risk, which is measured by the standard deviation of returns. Modern portfolio theory also suggests to consider assets in a diversified portfolio that have correlations of returns less than one with each other because in this way it can decrease portfolio the risk without sacrificing the return. The simplicity of portfolio composition by using modern portfolio theory have attracted significant attention both in academia and in practice. Over the last years, thanks to the recent interests in a financial context, both statistical and machine learning techniques have been used in order to both building and selecting portfolios. Amongst others, many proposals are based on clustering the financial series as preliminary steps for portfolio selection. Recently, a new clustering method for time series applications has been proposed by exploiting the properties of the P-splines approach. This semi-parametric tool has several advantages, i.e. it facilitates the removal of noise from time series and it allows for a reduction of the dimensionality of the clustering task ensuring a computational time saving.
In this paper, we propose to use this clustering approach on financial data with the aim of providing a strategy to build a portfolio of stocks able to support the investment choices of the portfolio managers. Our proposal works directly on time series without any heavy pre-processing step, it does not require the well-known conditions of stationarity and invertibility of time series. In a few words, we propose to cluster the series of prices and then to build a portfolio on selected stocks coming from the achieved partitions. From the modern portfolio theory point of view, examples on real financial time series show that our strategy is useful to support the investment decisions of financial practitioners.
Keywords: Time Series; P-Spline; Cluster Analysis; Portfolio Selection.

References

[1] C. Iorio, G. Frasso, A. D'Ambrosio, and R. Siciliano, "A P-spline based clustering approach for portfolio selection." Expert Systems with Applications, Elsevier, Pergamon Press Ltd., United Kingdom, 95 (2018), pp. 88-103.

[2] G. Pandolfo, C. Iorio, and A. D'Ambrosio, ``Depth-based portfolio selection". In A. Abbruzzo, E. Brentari, M. Chiodi, D. Piacentino (Eds.), Book of Short Papers SIS 2018, Pearson, 2018.

[3] C. Iorio, G. Frasso, A. D'Ambrosio, and R. Siciliano, "Parsimonious time series clustering using p-splines". Expert Systems with Applications, Elsevier, Pergamon Press Ltd., United Kingdom, 52 (2016), pp. 26-38.

Network-based Semisupervised Clustering

Giulia Contu*[1] , Luca Frigau[1], Claudio Conversano[1]

[1] Department of Economics and Business Sciences

Semisupervised learning [1] recently emerged as a new challenge: it uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. In this framework, semi-supervised clustering ([2], chap. 20) concerns the application of clustering methods to partially labeled data or to data with other types of outcome measures. Most of the approaches to semisupervised clustering focused on the extension of standard clustering methods to the semisupervised setting. Some methods concern the situations when clusters are associated with a given outcome variable that acts as a "noisy surrogate". This variable is considered a good proxy of the unknown clustering structure if it is used in combination with the other variables in the case conventional clustering methods may fail in identifying suitable clusters. In this framework, the term "suitable" denotes consistency between the identified clusters and the levels of the outcome variable.

In this paper, we propose a novel approach to semi-supervised clustering that is inspired by the semisupervised clustering associated with an outcome variable approach mentioned above. The proposed approach combines an initialization step with a training step. In the first step, a classification or regression tree is grown considering the outcome variable as response and only one feature at a time and a specific weight, that depends on a variable importance measure arising with the single-feature tree, is assigned to each feature. At the same time, instances are weighted w.r.t. the homogeneity of the terminal node of the tree to which they are assigned. Next, single-feature trees are still learned in the training step. They use weighted instances and select randomly the feature used to grow the "single-feature tree" on the basis of the weights assigned to each feature in the previous iterations. The output of the tree allows to update a proximity matrix that counts how many times pairs of observations, say i and j, have been placed in the same terminal node of the tree. This matrix is the input for the derivation of a network N on which a community detection algorithm is applied. In each iteration, the community detection
algorithm provides a partition of the original set of n instances into k groups. The final (selected) partition is the one characterized by the minimum internal heterogeneity inside groups, the latter evaluated w.r.t. the outcome variable.

Keywords: Tree-based Classifiers; Complex Networks; Community Detection

References

[1] Olivier Chapelle, Alexander Zien and Bernhard Schölkopf (Eds.) Semi-supervised learning MIT Press, 2006.

[2] Charu C. Aggarwal and Chandan K. Reddy Data Clustering: algorithms and applications Chapman & Hall/CRC Press, 2014.

*Corresponding author: giulia.contu@unica.it

From Multivariate to Functional Classification

José Luis Torrecilla [*1]

[1] Department of Mathematics, Universidad Autónoma de Madrid

While classification problems with multivariate data are well understood and optimal rules are known in many cases, the situation is quite different when classifying functional data. Functional data analysis entails some difficulties due to the continuous structure of the functions. In particular, the non-existence of densities of random functions complicates the construction of classification rules. This and other issues are addressed in this talk, and a procedure to derive optimal discrimination rules for binary classification problems of Gaussian processes is presented. The new rules are obtained as limits of sequences of multivariate problems, which are well defined.

Keywords: Functional Data Analysis; Supervised Classification; Optimal Classification Rules; Gaussian Processes; Near-perfect Classification

References

[1] Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. Journal of the American Statistical Association, 113(523), 1210-1218.

[2] Delaigle, A., and Hall, P. (2012). Achieving near perfect classification for functional data. Journal of the Royal Statistical Society: Series B, 74(2), 267-286.

[*]Corresponding author: joseluis.torrecilla@uam.es

Functional Linear Model for Monitoring and Prediction of Profiles

Alessia Pini*[1]

[1]Università Cattolica del Sacro Cuore, Milan, Italy

In many recent industrial and business applications, researchers are provided with functional data, that is, the observation of a quantity varying over a continuous domain (space, time, frequency, ...). The analysis of such data is particularly challenging since functional data belong to an infinite dimensional space. The interest of the statistical literature on this area has faced a rapid recent increase. Inference and prediction through functional linear models is an important topic within this area.

We consider in the present talk a functional linear model, where the functional data are modelled with a set of scalar and/or functional covariates multiplied by functional parameters.

In this context, many of the empirically relevant questions address the effect of either scalar or functional covariates on the functional response. Furthermore, if a covariate has an effect on the response, it is always of interest to identify the portions of the domain where the effect is significantly different from zero. In this talk, an approach to perform inference on such models and select such portions of the domain is presented. To show the flexibility of this approach, we discuss its application to two different real problems: the supervised profile monitoring of signal data in an industrial process [2], and the prediction of bike-mobility flows in a city [3]. We refer to [1], [2] and [3] for the theoretical details of the proposed approach.

Keywords: Functional Data Analysis; Functional Linear Models; Domain Selection.

References

[1] Pini A. and Vantini S. (2017). Interval-wise testing for functional data. Journal of Non parametric Statistics 29(2): 407-424.

[2] Pini A., Vantini S., Colosimo B.M. and Grasso M. (2017) Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. Journal of the Royal Statistical Society C. 67(1): 55-81.

[3] Torti A., Pini A. and Vantini S. (2019). Modeling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan ,Tech. Rep. 2019 MOX, Politecnico di Milano.

*Corresponding author: alessia.pini@unicatt.it

Depth-based Functional Time Series Forecasting

Antonio Elías[1], Raúl Jiménez[1]

[1]Department of Statistics
Carlos III University of Madrid

Functional Time Series are becoming prominent in practice due to a huge progress in the collecting data capabilities. These are time series in which we observe a curve at each time step. However, modelling such a complex data set becomes sometimes hard for practitioners. In this work, we propose an approach for making punctual and band predictions that is completely data-driven, easy to understand and fully interpretable. Both whole curve forecasting and curve segments prediction are addressed and applied to simulated and case-study data.

Keywords: Functional Data; Functional Time Series; Forecasting; Depth Measures.

*Corresponding author: aelias@est-econ.uc3m.es

Fault Detection and Diagnosis Methodology in Refineries: A Data-Driven Approach

Çağla Odabaşı[*1], Ocan Şahin[1]

[1] Tüpraş R&D Center, Izmit, Turkey

Fault detection and diagnosis (FDD) is a popular research area in process operations to enhance manufacturing sustainability and operating safety. Faults and anomalities in the performance of the critical equipments can be detected earlier as well as their root causes using FDD methods; hence, damages or shut-down of process can be prevented and overall process performance increases.

Automated FDD process is composed of four different functional stages[1]. The first step is fault detection, which is monitoring the process conditions and detecting any abnormal situations. Fault diagnosis is the second step in which the root causes of the fault is evaluated. Then, the significance and impact of the fault on system performance are evaluated as third step. Finally, the right action is determined to respond to the fault. Although, all steps are important for process sustainability, fault detection and diagnosis parts are mostly focused in literature. The number of research articles increased tremendously in this field and 995 research articles were published within 20 years according to Sciencedirect search.

In this talk, a general overview about FDD methodology will be given, different methods in the literature will be introduced, their advantages and challenges will be discussed. Future trends will be discussed.

Keywords: Fault Detection and Diagnosis; Machine Learning; Data Mining; Process Monitoring

References

[1] Katipamula, S. and Brambley, M.R. (2005). "Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I," HVAC&R Res., 11(1)1, 3–25.

[*]Corresponding author: cagla.odabasi@tupras.com.tr

Big Data Solutions in Refineries with Heat Exchangers

Ocan Şahin [*1], Çağla Odabaşı[1]
[1] Tüpraş R&D Center, Izmit, Turkey

With recent advances in overall availability, affordability, safety and stability of data recording sensors data collecting and analytics found themselves another domain, chemical processes. Especially, in refineries we are able to say that daily operations generate very large volumes of data with very high frequencies like including intelligence on equipment, performance parameters, and even low frequencies as shift and working schedules, maintenance frequencies, operation costs and purchasing/sale information. We, Tüpraş R&D team of Algorithms and Software Solutions, approached these data with an idea in order to solve a re-occurring problem. That problem was heat exchanger shutdown due to performance loss/fouling and the high operating costs of furnaces because of non-operational or low performing heat exchangers.In each refinery, heating and cooling tasks are done via industrial heat exchangers, taking advantage of the heat transfer between the hot and cold streams respectively. Fuel/electricity consuming external heating/cooling sources satisfy the further heating requirements; making energy consumptions of a regular refinery one of the largest in overall industrial consumption [1][2]. A potential efficiency reduction in the heat exchangers reduces the energy efficiency of the operation and results with more use of fresh energy thus increasing the operation costs significantly [3].Fouling has been widely recognized as the main cause for the degradation of overall performance and efficiency of heat exchangers [4][5]. Accumulating fouling in the heat exchanger not only complicate the overall control and stability of the system, but also result in detrimental capital losses due to higher energy consumption, more frequent maintenance cycles and unsafe operation conditions. Fouling related pressure drop, and the fuel consumed cost an average refinery around US$20.000 per day [6]. Costs associated with the crude oil fouling in the preheat heat exchangers worldwide were estimated to be of the order of US$4 billion per year in 1995 only [7]. Hence, heat exchangers and their fouling monitoring is very important for the overall petroleum refinery economics.Predictive modelling showed great promise in terms of tackling the fouling phenomena in heat exchangers. In this talk I'd like to explain the solution we did, predicted output temperatures of a preheat heat exchanger in vacuum distillation unit combined with fundamental fouling calculations. Resulting set of data used to further estimate potential fouling horizon for future 3 days as a mean and support tool for the plant controllers and engineers.

Keywords: Heat exchanger; Fouling; Shell and Tube; Refinery; Prediction, Estimation, Big Data

References

[1] Haynes, V.O., 1976, Energy use in petroleum refineries (No. ORNL/TM-5433), Oak Ridge National Lab., Tenn., USA.
[2] ESDU, 2000, Heat exchanger fouling in pre-heat train of a crude oil distillation unit, ESDU Data Item 00016, ESDU International plc., London
[3] Pritchard, A. M., 1987, The Economics of Fouling, Fouling Science and Technology, NATO ASI Series E, Kluwer Academic Publishers, vol 145.
[4] Garret-Price, B. A., et al., 1985 Fouling of Heat Exchangers, Characteristics, Costs, Prevention, Control and Removal, Noyes Publications, Part Ridge, NJ USA.
[5] Thacker, P. A., 1980, The Cost of Fouling, Heat Exchanger Plant, Effluent and Water Treatment J.
[6] Yeap B. L., et.al, 2005, Retrofitting crude oil refinery heat exchanger networks to minimize fouling while maximizing heat recovery, Heat transfer engineering, 26.1, 23-34.
[7] Kashani M. N., et.al., 2012, Dynamic crude oil fouling prediction in industrial preheaters using optimized ANN based moving window technique, Chemical Engineering Research and Design, 90.7, 938-949

[*]Corresponding author: ocan.sahin@tupras.com.tr

Part VI

Contributed Papers (Abstract)

Identification of Vehicle Warranty Data and Anomaly Detection by Means of Machine Learning Methods

Halil İbrahim Çelenli[1], Esin Özkan[2]

[1,2] IBSS Consulting, Istanbul, Turkey
[1] halilibrahim.celenli@ibss.com.tr
[2] esin.ozkan@ibss.com.tr

Nowadays, the use of machine learning methods instead of manual methods on the data is increasing. This increase also improve the applied methods. In addition the frequency of the use of machine learning methods in the field of the industry has improved. These methods are also used on vehicle warranty data in predictive maintenance. The warranty data used in the industry is important for companies in terms of cost. A part in a vehicle may cause a fault and if the fault is a continuing error during the warranty period, this will increase the cost.

In this study, the warranty data used in the automotive industry have automatically labeled and the anomaly detection process has performed on the labeled case. We have used 32360 vehicle data together with warranty costs. In addition, gamma distribution and moving average methods are used for labeling. The labeling cases are given as Stable, Issue_New, Issue_Stable, Issue_Decrease and Issue_Increase. These cases show the behavior of the anomaly. For example, Stable; anomaly no as showed, Issue_New; a new anomaly shows the status of. Stable and Issue_New labels are determined using the gamma distributions. The other labels are determined by moving average. Machine learning algorithms are used for anomaly case detection on the labeled data. We are used two different algorithms. The first is a xgboost algorithm based on decision-tree and gradient-boosting. The second is a logistic regression algorithm based on regression. The accuracy is used as an evaluation criterion and results are compared. We have observed xgboost algorithm in the anomaly detection process performed better than logistic regression algorithm.

Keywords: Machine Learning; Anomaly Detection; Gamma Distribution;Warranty Classification; Predictive Maintenance

Predicting Business Survival from their Websites

Desamparados Blazquez[1], Lisa Crosato[2], Josep Domenech[1], CaterinaLiberati[2]

[1] Department of Economics and Social Sciences, Universitat Politécnica de Valéncia
[2] Department of Economics, Management and Statistics, University of Milano-Bicocca
caterina.liberati@unimib.it

The purpose of this paper is to assess to what extent firms' survival can be predicted using just information collected from their corporate websites. The study of firms' survival is a high interest topic due to its implications with stability and growth of the economic system of a country. Generally, such investigations are carried out measuring ratios coming from businesses' financial balance sheets [1]. Unfortunately, those variables are available about two years late with respect to their reference period and that diminishes the significance of the results in a forecasting perspective.

In our work, we employed a complete diff erent set of indicators obtained via web scraping and content analysis techniques of corporate websites in 2016. The online information was collected by accessing the corporate websites with the Wayback Machine of the internet archive [2].

We rely our study on a sample composed by 780 companies established in Spain, where the defaulters count for almost 4% of the total instances. Such an unbalance between the two classes of companies generally aff ects the reliability of the estimates of the default probability when linear models are applied. In order to overcome the limitations of the standard discriminant models, we employed the Kernel Discriminant Analysis [3], which is able to deal with both the non-linear relationship between the response and predictors and the low numbers of defaults. The flexibility of this technique allowed us to design alternative classifiers using diff erent kernel maps. The suitability of the discriminant solutions was then assessed by ranking their error rates.

A comparison among solutions with traditional official indicators and the discriminants estimated with on line predictors is also illustrated, and a further discussion about the results is addressed.

Keywords: Firms' Survival; Default Prediction; Corporate Websites; Kernel Discriminant

References

[1] Lin, S. M., Ansell, J., & Andreeva, G. (2012). Predicting default of a small business using diff erent definitions of financial distress. Journal of the Operational Research Society, 63, 539-548.

[2] Kahle, B. & Gilliat, B. (2016), "Wayback machine", available at: http://archive.org/web/

[3] Mika S., Rätsch G., Weston J., Schölkopf, B. & Müller K.R. (1999) Fisher discriminant analysis with kernels. In: Neural networks for signal processing, vol IX. Proceedings of the 1999 IEEE signal processing society workshop, 41-48

Methods for Optimum Establishment of Government - imposed Global Budget Caps

Nika E.[*1], Dr. Psarakis S.[2] and Dr. Papadaki A.[3]

[1]PhD candidate, Athens University of Economics & Business
[2]Professor in the Department of Statistics, Athens University of Economics and Business
[3]Professor in the Department of Accounting and Finance, Athens University of Economics and Business

One of the most widely known tools for protecting public spending in health care services and pharmaceuticals is the use of government-imposed global budget caps.

However, many concerns have been raised in literature regarding the soundness of this method. The most significant concern is related to the fact that government-imposed global budget caps might constitute a barrier to the universal access of people to medicines and health care services in general. The main justification for these concerns is the insufficient methods used for establishing these budget caps.

In this work, we present an analytical process based on statistical and cost accounting methodologies for estimating government-imposed global budget caps.

Keywords: Budget Caps; Pharmaceuticals; Healthcare

Detection and Handling Outliers in Longitudinal Data: Can Wavelet Decomposition Be a Solution?

Marwa BenGhoul[1], Berna Yazıcı[2], Ahmet Sezer[3]
[1] PhD fellow, Eskişehir Technical University, Faculty of Science, Statistics Department, Eskişehir,Turkey
benghoulmarwa@gmail.com
[2] Professor, Eskişehir Technical University, Faculty of Science, Statistics Department
bbaloglu@eskisehir.edu.tr
[3] Associate Professor, Eskişehir Technical University, Faculty of Science, Statistics Department
a.sezer@eskisehir.edu.tr

Outliers still always a subject of debate: some researches showed that deleting outliers is advantageous for the analysis and others confirmed that keeping them is indispensable as those abnormal observations still data elements and different ways to handle them can be applied. Recently, the wavelet analysis has started to be known as a powerful mathematic tool to decompose a series and provide the frequential and the temporal features. Despite several studies research utilized the wavelets approach to detect outliers and handle them in financial and economic datasets, it still not highly applied on longitudinal data. Therefore, this research aims to improve the accuracy of the outlier detection in longitudinal data by applying the wavelet decomposition and by introducing a new approach to handle them. Indeed, two novel algorithms are developed: the first one consists in applying the wavelets decomposition across subjects and the second one within subjects. In analysis, a historical longitudinal data produced by the AIDS Clinical Trials Group and published by Dr. Hulin Wu is utilized. The results illustrate that the wavelet decomposition, in both algorithms, has a strong capacity to detect and deal with outliers without deleting them. Furthermore, the proposed algorithm within subjects highlights few cases where outliers cannot be addressed (a subject has only one record which is by coincidence an outlier or a subject who has two consecutive outliers) and thereby the final decision will be left to the field's expert to retain or remove those cases.

Keywords: Longitudinal Data; outliers; Wavelet Decomposition; Detection; Handling

References

[1] Al-Khazaleh, A. M. H., Al Wadi, S., and Ababneh, F., Wavelet transform asymmetric winsorized mean in detecting outlier values, Far East Journal of Mathematical Sciences, 96(3), 339–351, 2015.
[2] Grané, A., and Veiga, H., Wavelet-Based Detection of Outliers in Volatility Models, DES - Working Papers. Statistics and Econometrics, WS 090403, Universidad Carlos III de Madrid. Departamento de Estadística, (03), 2009.
[3] Gumedze, F. N., and Chatora, T. D., Detection of outliers in longitudinal count data via overdispersion. Computational Statistics and Data Analysis, 79, 192–202, 2014.
[4] Pollet, T. V., and Van Der Meij, L., To Remove or not to Remove: the Impact of Outlier Handling on Significance Testing in Testosterone Data, Adaptive Human Behavior and Physiology, 3(1), 43–60,2017.

Serial Mediation Model of Leader Member Interaction in Work Values and Job Satisfaction

Meral YAY[1], Mine AFACAN FINDIKLI[2] , Ali Mertcan KÖSE[3]

[1]Mimar Sinan Fine Arts University
meral.yay@msgsu.edu.tr
[2]Beykent University
minefindikli@beykent.edu.tr
[3]Mimar Sinan Fine Arts University
alimertcankose@gmail.com

Nowadays, enterprises state that one of the most important factors in creating a competitive advantage is qualified human resources; they aim to bring their employees, who care about their job satisfaction and happiness, in their institutions for long years. The fact that employees are highly satisfied with their jobs depends on their ability to identify their expectations and needs, their behavioral and cognitive characteristics. An individual's behavioral and cognitive characteristics as well as one of the important factors affecting the expectations and needs are individual's values. Therefore, the correct determination of the values depends on the positive contribution of the employees to their interaction, choices and work outcomes. In this study, by determining the business values of clinical and laboratory workers in the field of assisted reproductive techniques in the health sector, the mediation effect of leader-member interaction is questioned in relation to job values and job satisfaction. The relationships between the variables were analyzed by establishing two different serial mediation models and as a result of the relationship between business values and job satisfaction, it was seen that the mediator variables affecting the leader-member interaction were also related to each other.

Keywords: Work Values; Leader-Member Exchange; Job Satisfaction; Serial Mediation Model

References

[1] G.B. Graen and U. Bien, "Relationship-Based Approach to Leadership: Development of Leader-Member Exchange (LMX) Theory of Leadership over 25 years: Applying a Multi-Level Multi-Domain perspective", Journal, Leadership Quarterly, 1995, pp. 219-247.

[2] A. F. Hayes, Introduction to Mediation, Moderation and Conditional Process Analysis, Guilford Press, New York, 2013.

Outlier Detection on Big Data

Erkan Şirin[1], Hacer Karacan[2]

[1] Ph.D. Student, Institute of Information, Gazi University, Ankara, Turkey
erkan.sirin@gazi.edu.tr
[2] Assoc. Prof., Computer Engineering, Gazi University, Ankara, Turkey

The fraud led costs in health care keeps the important ratio in total healthcare costs. The majority of the papers are focused on the service provider led fraud detection system. These systems are for reimbursement agencies. No paper has been found on providers' internal frauds causing unnecessary costs. Detection methods of health care fraud can be split into two ways [1, 2]: manual and automated. Using the latter, we analyzed material expenses data and offered an unsupervised method to detect unusual episodes which might be fraudulent by using machine learning on big data cluster. We first clustered data in order to get more homogenous sets, then labeled outliers in each cluster according to distance from the cluster center. Our approach labeled 545 episodes as outliers, out of 25.083 episodes.

We validated the results using classification, Spark mllib random forest classifier, and evaluation metrics such as roc[3] and f1[4], which are used to evaluate imbalanced labels. Performed ROC AUC 98.8 % and f1 score 98.1 %. When we investigated results with some SQL queries we found some interesting cases. For example, a patient had 10 episodes but 7 of them were outliers.

Keywords: Healthcare; Big Data; Unsupervised Outlier Detection

References

[1] Li, J., et al., A survey on statistical methods for health care fraud detection. Health care management science, 2008. 11(3): p. 275-287.

[2] Copeland, L., et al., Applying business intelligence concepts to Medicaid claim fraud detection. Journal of Information Systems Applied Research, 2012. 5(1): p. 51.

[3] Fawcett, T., An introduction to ROC analysis. Pattern recognition letters, 2006. 27(8): p. 861-874.

[4] Hossin, M. and M. Sulaiman, A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 2015. 5(2): p. 1.

Joint Modeling the Frequency and Duration of Physical Activity from a Lifestyle Intervention Trial

Gul Inan[1], Juned Siddique[2]

[1] Department of Mathematics, Istanbul Technical University, Istanbul, Turkey
inan@itu.edu.tr
[2] Department of Preventive Medicine, Northwestern University, Chicago, Illinois, USA

World Health Organization guidelines on physical activity (PA) recommend that adults should do at least 150 minutes a week of moderate-intensity or 75 minutes a week of vigorous-intensity aerobic PA. Furthermore, this aerobic activity should be performed in episodes of at least 10 minutes. Unfortunately, many adults do not meet these guidelines and developing behavior interventions to promote physical activity is still an active area of research.

In a lifestyle intervention trial, participants are randomized into one of two activity related intervention arms: an increase PA arm with a goal to increase moderate-to-vigorous PA and a decrease sedentary behavior arm with a goal to decrease leisure-time sedentary screen time. The daily PA behavior changes in terms of frequency and duration of the participants are measured by an accelerometer over 5 weeks to investigate the effect of intervention treatment. The data collected from this intervention trial revealed that the distribution of frequency of daily episodes for the participants is zero-inflated such that there is no exercise on many days and, furthermore, the distribution of the duration of daily (non-zero) episodes for the participants is highly positively skewed.

Motivated by this longitudinal study, here, we are interested in jointly modeling frequency and duration of daily episodes to investigate PA behaviors of the participants over time. In this sense, we propose a multivariate generalized linear mixed-effects model to analyze jointly the zero-inflated frequency and positively skewed duration outcomes. Specifically, the proposed model is a three-part random effects model, where the first two-part is a Poisson hurdle regression model associating probability of having a non-zero episode number on a day with covariates conditional on participant-level random effects via a logit link function and associating expected number of non-zero episodes on that day with covariates conditional on random effects via a log link function. The third part of the proposed model is a Gamma regression model associating expected duration of each non-zero episode on that day with covariates conditional on random effects via a log link function. Bayesian inference is used to estimate the parameters of the proposed model. Lastly, we illustrate the usefulness of the proposed model through the analysis of the motivating PA intervention study data.

Keywords: Count Outcome; Excess Zeros; Hurdle Models; Positive Continuous Outcome

On Function-On-Function Regression: Partial Least Square Approach

Ufuk Beyaztas[*1], Han Lin Shang[2]

[1] Bartin University, Department of Mathematics, Bartin, Turkey
[2] Research School of Finance, Actuarial Studies, and Statistics, Australian National University, Australia

Functional data analysis tools, such as function-on-function regression models, have received considerable attention in various scientific fields because of their observed high-dimensional and complex data structures. Several statistical procedures, including least squares, maximum likelihood, and maximum penalized likelihood, have been proposed to estimate such function-on-function regression models. However, these estimation techniques produce unstable estimates in the case of degenerate functional data or are computationally intensive. To overcome these issues, we proposed a partial least squares approach to estimate the model parameters in the function-on-function regression model. In the proposed method, the B-spline basis functions and generalized cross-validation are utilized to convert discretely observed data into their functional forms and to control the degrees of roughness, respectively. The finite-sample performance of the proposed method was evaluated using several Monte-Carlo simulations and an empirical data analysis. The results produced by our numerical analyses reveal that the proposed method competes favorably with existing estimation techniques and some other available function-on-function regression models, with significantly shorter computational time.

Keywords: Basis Function; Functional Data; Nonparametric Smoothing; NIPALS; SIMPLS

[*]Corresponding author: ubeyaztas@bartin.edu.tr

A Robust Method for Estimation of Models with Random Effects

Beste Hamiye Beyaztas
Bartin University, Department of Mathematics, Bartin, Turkey
bbeyaztas@bartin.edu.tr

Panel data regression models have become a general framework to obtain satisfactory answers to statistical inference problems in different fields such as behavioral, environmental, medical sciences and econometrics. Such models with outliers are very common in real data sets. Outlying data points may distort the structure of the underlying model, and lead to produce biased and inefficient estimates for the model parameters when the generalized least square (GLS) method is used. In this study, we propose a robust estimation approach based on weighted likelihood methodology for linear panel models with random effects. The finite sample performances of the proposed method are investigated by means of Monte-Carlo simulations as well as a real-world example. Our records reveal that the proposed method provides superior performance over the traditional counterpart when the data has outlier(s). Also, it produces competitive results to GLS estimates when no outliers are presented in the data.

Keywords: Panel data; Random effect; Weighted likelihood

Conditional Autoregressive Model Approach to Generalized Linear Spatial Models by CARBayes

Leyla Bakacak Karabenli[1], Serpil Aktaş Altunay[2]

[1] Department of Statistics, Hacettepe University
leylabakacak@hacettepe.edu.tr
[2] Department of Statistics, Hacettepe University

Spatial analysis is a method used to describe spatial patterns on a geographical region. The attributes and location information of the spatial units are used throughout the analysis. Response variable exhibits the spatial autocorrelation structure because of neighboring effects between spatial areal units. Thus, even if explanatory variables are used in the model, the spatial relation might remain in the residuals. Therefore, the assumption of independence by linear model approaches is violated. In such cases, random effects involving spatial relations are included in the model as a part of Bayesian hierarchical model with conditional autoregressive priors for these effects.

The CARBayes package in R programming can be used in order to set up a Bayesian spatial model with conditional autoregressive (CAR) priors for the analysis of spatial areal data. This package is based on the Markov Chain Monte Carlo (MCMC) simulation using a combination of Gibbs sampling and Metropolis Hastings algorithms.

In this study, 81 provinces of Turkey are used as non-overlapping spatial areal units. The response variable is the number of earthquakes whose magnitude is greater than "2.0" and the covariate is the average magnitude of earthquake in each province in 2016. The response variable reveals spatial autocorrelation as a result of Moran's I test. Hence, the residual of generalized linear model exposes the spatial autocorrelation. Based on this, the relation between the number of earthquakes and its magnitude is investigated by the spatial generalized linear model, especially Leroux conditional autoregressive model. After modelling, risk values of each province are calculated based on the ratio of fitted values obtained from the model to the expected number of earthquakes in each province. Then, risk mapping is implemented.

Keywords: Spatial Autocorrelation; CAR Models; MCMC; Generalized Linear Spatial Model

Hierarchically Built Trees with Probability of Placing Clusters

Nebahat Bozkus[1], Stuart Barber[2]

[1]University of Giresun, Turkey
nebahat.bozkus@giresun.edu.tr
[2]University of Leeds, UK

One of the popular questions in hierarchical clustering is how many clusters we have. The available cluster validity indices (CVIs) capture the high percentage of well-separated true components, but their performances deteriorate for indistinct groups or unique data structures. In addition, the available CVIs, or at least the ones we are aware of, find the number of clusters, but they do not show where each cluster needs to be placed on trees. We propose a new algorithm based on non-decimated lifting (NLT) using the 'denoising' abilities of wavelet methods. In our algorithm, we denoise departures from the centroid of each possible cluster (each node) on a tree by the help of lifting 'one coefficient at a time' (LOCAAT) algorithm. To allocate a cluster at a node, we use denoised departures from the centroids for each possible cluster on the tree. If the denoised result of a possible cluster is small enough, we allocate one of clusters at the node of interest on the tree. However, the nature of the NLT algorithm includes some number of repeated LOCAAT algorithm. Thus, denoising the tree by the NLT algorithm finds a clustering pattern for each repetition. We suggest that we can summarize the multi-denoising results by allocating a probability of being a cluster at each node on the tree, and we can place a cluster at a node where the probability of placing a cluster at this node and all its child nodes are greater than a user defined threshold. We compare the performance of our algorithm with some available methods in the literature using some simulated and real data sets.

Keywords: Cluster Validity Indices; Clustering; Lifting; Wavelets

Nonlinear Neural Network for Cardinality Constraint Portfolio Optimization Problem: Sector-wise Analysis of ISE-all Shares

Ilgım YAMAN[*1], Türkan ERBAY DALKILIÇ[2]

[1] Giresun University, Department of Statistics, Giresun, Turkey
[2] Karadeniz Technical University, Department of Statistics and Computer Sciences

Standard portfolio optimization method had proposed by Harry Markowitz in 1952 which is the benchmark problem of finance world [1]. The main purpose of this method while maximizing the expected return minimizing the risk of the portfolio. While portfolio optimization problem known as a quadratic optimization problem, cardinality constrained optimization problem is a mixed-integer quadratic optimization problem. Cardinality constrained portfolio optimization falls into NP-hard class. That makes this algorithm is a complex method and having a solution in polynomial time. Because of that reason, the classic optimization method does not meet the needs of today's financial world. Neural networks were used to solve these problems. The nonlinear neural network proposed by Yan in 2014 which is used to solve the nonlinear fitness functions [2]. Cardinality constraint was added to the Markowitz mean-variance model, which is restricted to the number of assets to be included in the portfolio. Thus, the portfolio optimization model has become a mixed-integer quadratic optimization problem.

In this study, sector-wise analysis is done for the cardinality constraint portfolio optimization problem by using standard deviation as an elimination criterion in order to understand which sector is profitable. Five sectors are examined for the proposed algorithm which are Basic Materials, Consumer Cyclicals, Financials, Industrials & Technology, and Others. For each sector, standard deviations of stocks are taken into consideration. In order to eliminate stock having maximum risk, the standard deviations of each stock are calculated. Stocks are eliminated which have greater standard deviations than the median. Finally, selected stocks are taken into the nonlinear neural network to get proportions of selected stocks. As a result, standard deviation, mean expected return, risks and Sharpe ratio of the selected portfolio are calculated for different sectors.

Keywords: Cardinality Constraint Portfolio Optimization; ISE-all Shares; Sector-wise

References

[1] Markowitz, H. (1952). Portfolio selection. The journal of finance, 7(1), 77-91.

[2] Yan, Y. (2014). A new nonlinear neural network for solving QP problems. International Symposium on Neural Networks, Springer International Publishing, 347-357.

[*]Corresponding author: ilgim.yaman@giresun.edu.tr

Gamma and Inverse Gaussian Distributions in Fitting Parametric Shared Frailty Models with Missing Data

Nursel Koyuncu[1], Marthin Pius[1], Nihal Ata Tutkun[1]

[1]Hacettepe University, Department of Statistics, 06800, Beytepe, Ankara, TURKEY
nkoyuncu@hacettepe.edu.tr, nihalata@hacettepe.edu.tr, martinpius01@gmail.com

Missing data is an important problem for statistical data analysis and must be valuated carefully. In literature many authors focus on missing data for various statistical models. Many imputation methods are proposed to solve missing data issue in modelling. However, there are few studies related missing in survival models.

Survival models are a kind of regression models applied to time-to-event data and commonly used in medical sciences. One of the popular survival model is parametric frailty model with gamma and inverse Gaussian distributions. Frailty models are extensions of the Cox regression model which is the most popular model in survival analysis.

In this study, we compared the effectiveness of gamma and inverse Gaussian distributions under different baseline hazards in fitting shared parametric frailty models with missing data. We first simulated survival data for recurrent event. Samples were simulated with different sizes, censoring rates and missing rates. Parametric shared frailty models with gamma and inverse Gaussian distributions under different baseline hazards were fitted to the data set with and without missing observations. Simulation results indicate superiority of gamma distribution over inverse Gaussian when exponential baseline hazard is assumed regardless of sample size and missing rate. When log-logistic and Weibull baseline hazards are assumed, frailty model with gamma distribution generally appeared to be superior over inverse Gaussian.

Keywords: Frailty; Hazards Function; Missing at Random (MCAR).

References

[1] M. Lee, I Do Ha, and Y. Lee, “Frailty Modeling for Clustered Competing Risks Data with Missing Cause of Failure”,Statistical Methods in Medical Research, 26(1), 2017, pp.356–73.

[2] J. G. Ibrahim, M.H. Chen, S.R. Lipsitz, and A. H. Herring, “Missing-Data Methods for Generalized Linear Models”, Journal of the American Statistical Association, 100:469, 2005, 332-346.

A Functional Data Framework to Analyse the Effect of Quinoa Consumption on Blood Glucose Levels

Nihan Acar-Denizli[*1], Pedro Delicado[2], Belchin Kostov[2,3], Diana A.Díaz-Rizzolo[3], Antoni Sisó[4] and Ramon Gomis[3]

[1]Mimar Sinan Fine Arts University, Istanbul, Turkey
[2]Universitat Politecnica de Catalunya, Barcelona, Spain
[3]IDIBAPS, Barcelona, Spain
[4]CAPSBE, Barcelona, Spain

Functional data analysis deals with observations in the form of functions that are measured on a continuum. With the developments in the technology, it is more common to work with dense data sets. Specially, in health sciences, the analysis of dense data is becoming more important by the use of medical sensors. In this study, 9 patients with pre-diabetes have been followed up during 28 days. During that period, the blood glucose levels of the patients were measured by a sensor which takes records every 15 minutes. First two weeks patients followed a regular diet and the next two weeks they followed a quinoa diet based on products elaborated from quinoa. The main objective of this study is to construct a functional data framework to investigate the eff ect of quinoa consumption for the prevention of type 2 diabetes mellitus. We investigated how monitored glucose levels are eff ected from type of diet and nutrient intake by means of functional linear models. We consider glucose levels as a functional data and use Function on scalar regression (Fosr) models to analyze the eff ect of relevant variables.

Keywords: Functional Data Analysis; Functional Linear Models; Function on Scalar Regression Models.

References

[1] Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J. et al. (2018). refund: Regression with Functional Data R package version 0.1-17.

[*]Corresponding author: nihan.acar@msgsu.edu.tr

Statistical Inference of Consecutive k-out-of-n System in Stress-Strength Setup Based on Two Parameter Proportional Hazard Rate Family

Duygu Demiray*[1], Fatih Kızılaslan[2]

[1]Department of Mathematics, Yeditepe University, Istanbul, Turkey
[2]Department of Statistics, Marmara University, Istanbul, Turkey

A consecutive *k*-out of-*n* :G system is a system which consists of linearly ordered *n* components such that it works if and only if at least *k* consecutive components work. It has been used to model various engineering systems such as the microwave stations of a telecom network, oil pipeline systems and vacuum systems in an electron accelerator[1]. The reliability of consecutive *k*-out of-*n* :G system, $R_{n,k}$ was first considered by Chiang and Niu[1] and Eryılmaz and Demir[1] obtained the reliability of this system in stress-strength setup.

In this study, we consider the estimation problem of this system reliability in stress-strength model. The system is regarded as works only if at least *k* consecutive out of n components exceeds the random stress. It is assumed that strength components $Y_1, \ldots Y_n$ is exposed to a common random stress X when the underlying distributions follow the proportional hazard rate family with two parameters which includes several well-known lifetime distributions such as Kumaraswamy, Weibull(two parameter), Burr-Type XII and so on. The maximum likelihood estimator and the asymptotic confidence interval of $R_{n,k}$ are obtained when the parameters α, β and λ are unknown. The Bayes estimates of $R_{n,k}$ have been developed by using Lindley's approximation and the Markov Chain Monte Carlo method. The highest probability density credible interval is constructed by using Markov Chain Monte Carlo Method. Monte Carlo simulation is performed to compare the proposed reliability estimators.

Keywords: Proportional Hazard Rate Family; Stress-strength Reliability; Multicomponent Reliability; Consecutive *k*-out of-*n* :G system

References

[1] Kuo, W. and Zuo M.J. (2003). Optimal Reliability Modeling, Principles and Applications, Wiley, New York.

[2] Chiang D. T. and Niu S-C. (1981). Reliability of consecutive k-out of-n : F system. IEEE Transactions on Reliability, 30(1), 87–89.

[3] Eryılmaz S. and Demir S. (2007). Success runs in a sequence of exchangeable binary trials. Journal of Statistical Planning and Inference, 137(9), 2954–63.

*Corresponding author: duygu.demiray@yeditepe.edu.tr

Use of Relative Entropy in Contingency Tables

A. Evren[1], B. Sahin[2]

[1]Yildiz Technical University, Faculty of Science and Literature, Department of Statistics, Davutpasa, Esenler, 34210, Istanbul-Turkey
[2]Yildiz Technical University, Institute of Naturel Science, Department of Statistics, Davutpasa, Esenler, 34210, Istanbul-Turkey

There are various information theoretic divergence measures used in determining associations between nominal variables. Among them, Shannon mutual information statistic is appealing, since its sampling properties are well-known. Shannon mutual information is a special case of Kullback-Leibler divergence and, Renyi and Tsallis mutual information. As the envelopes of various tools, Renyi and Tsallis mutual information provides much higher flexi- bility than Shannon mutual information. In this study, large sampling properties of Shannon, Renyi, Tsallis mutual information statistics are considered as well as Pearson, Tschuprow, Sakoda, Cramer, Hellinger, and Bhattacharyya measures. In simulations, the normality of most of the statistics, and the higher positive correlation coefficients between all these tools are observed. Their sampling variabilities are compared. Then by using Renyi and Tsallis mutual information statistics, correlation coefficients are estimated for 8 different contingency tables, and 3 bivariate normal distributions.

Keywords: Contingency Coefficients; Divergence Measures; Renyi Mutual Information; Tsallis Mutual Information

Granger-Causality- Based Portfolio Selection in The Moroccan Stock Market

Abdelhamid Hamidi Alaoui*

Al Akhawayn University, Ifrane, Morocco
a.hamidialaoui@aui.ma

In this paper, we construct a portfolio of 32 companies from the Moroccan stock exchange. Given the market's conditions, for each stock, there is only a "*long position*" or a "*no position*". A network of Granger causalities among the 32 stocks is the basis of the trading decisions. An in-sample of daily returns from 2009 to 2016 is used to develop the network of causalities and an out-of-sample of daily returns from 2017 to 2018 tests the performance of the portfolio. There are four types of decisions: 1) "*no position*" to "*no position*", 2) "*no position*" to "*long position*", 3) "*long position*" to "*no position*", and 4) "*long position*" to "*long position*". Each decision is based on a look-back of how a stock is Granger-caused by other stocks. The performance of the portfolio is then compared to the performance of portfolios that uses ARMA to make the decisions 1) though 4). The results show that the Granger-causality-based portfolio outperforms the ARMA portfolio more than 70% of the times and that its cumulative rate of return over the period between January 2017 and December 2018 is almost 1.8 times the cumulative return of the ARMA portfolio.

Keywords: Granger Causality; Stock Market Networks; Portfolio Selection; Emerging Markets

References

[1] R. Coelho, C. G. Gilmore, B. Lucey, P. Richmond, S. Hutzler. (2007). The evolution of interdependence in world equity markets - Evidence from minimum spanning trees, Physica, A. 376, 455–466.

[2] W.-Q. Huang, X.-T. Zhuang, S. Yao. (2009). A network analysis of the Chinese stock market, Physica, A. 388, 2956–2964.

*Corresponding author: a.hamidialaoui@aui.ma

A Percentile Bootstrap Based Method on Dependent Data: Harrell Davis Quantile Estimator vs NO Quantile Estimator

Gözde Navruz*[1], A. Fırat Özdemir[2]

[1, 2] Dokuz Eylül University

When the observations are measured on two occasions based on the same variable, determining the significant difference between before and after measurements is frequently of interest in applied statistics. For example, the effect of training on triglyceride level can be assessed by the measurements before training and after four weeks of training [1]. The most common strategy for comparing two dependent groups is to use paired t test, which has some restrictive assumptions. Instead of testing the hypothesis that the difference scores have a mean of zero as in paired t test, one can consider a robust measure of location. However, to determine whether the differences occur in the tails of marginal distributions, namely quantiles, sometimes becomes an important issue. With the intent of comparing two dependent groups through different quantiles, the newly proposed NO quantile estimator [2] and Harrell Davis quantile estimator [3] are used in conjunction with a percentile bootstrap approach [4]. To obtain actual type I error rates, a simulation study is conducted by altering the correlation between dependent variables, sample sizes and considered quantile values. According to the results, NO quantile estimator outperforms in many cases studied.

Keywords: NO Quantile Estimator; Two Dependent Groups; Percentile Bootstrap

References

[1] Wilcox, R. R. (2017). Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction, Second Edition, Chapman and Hall/CRC, New York.

[2] Navruz, G. (2019). Analysis of Robust Approaches in the Two Independent Samples Case, PhD Thesis, Dokuz Eylül University.

[3] Harrell, F. E. and Davis, C. E. (1982). A New Distribution-free Quantile Estimator. Biometrika, 69(3), 635-640.

[4] Wilcox, R. R. and Erceg-Hurn, D. M. (2012). Comparing Two Dependent Groups via Quantiles. Journal of Applied Statistics, 39(12), 2655-2664.

*Corresponding author: gnavruz@gmail.com

Fitting Lognormal Distribution to Actuarial Data

M. Mahdizadeh[1], Ehsan Zamanzade[2]

[1] Department of Statistics, Hakim Sabzevari University, P.O. Box 397, Sabzevar, Iran
mahdizadeh.m@live.com
[2] Department of Statistics, University of Isfahan, Isfahan 81746-73441, Iran

Normal distribution is capable of explaining the random variation existing in the data from a variety of fields. A great deal of theory assumes, either explicitly or implicitly, that the probability distribution of data is normal. Many studies, however, show that this assumption is not supported empirically. Models based on this assumption often fail to fit into real-world data satisfactorily. The lognormal (LN) distribution can be a potential alternative model if data are positive and right skewed. It has found application in many scientific disciplines. For example, genesis of the LN distribution arising from biological and pharmacological mechanisms has been discussed. The LN distribution has also been utilized for modeling in biochemistry, medicine and hydrology. In the actuarial context, models with heavy-tailed distributions have been used to provide adequate descriptions of claim size distributions [1].

Parametric inferential methods are sensitive to violation of distributional assumption. Thus, it is critical to develope formal testing procedures for the LN model. Testing distribution assumptions has been one of the major areas of continuing statistical research. As to the LN distribution, two approaches are possible. The first one builds on the definition of the LN distribution, which allows us to reduce the problem of testing lognormality to that of testing normality for the log transformed data. Following this path, one can employ the well-known Shapiro-Wilk or Shapiro-Francia test. There exist other tests constructed based on measures of distance between the empirical distribution function (EDF) and a given distribution function. This class of tests includes Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Kuiper and Watson tests, among others. Zhang and Wu [2] developed three tests of normality based on the likelihood ratio statistic. A second approach for lognormality test is to use the original data. In doing so, we may employ two testing procedures, proposed by Batsidis et al. [3], which are based on the Kullback-Leibler distance.

In this article, we evaluate performance of different test of fit for the LN distribution based on a simulation study. Some practical recommendations are made about the best tests to use. Finally, an application in the context of actuary is discussed.

Keywords: Goodness-of-fit Tests; Transformed Data; Weighted Distribution.

References

[1] P. J. Boland. 2007. “Statistical and probabilistic methods in actuarial science”, CRC Press.

[2] J. Zhang, and Y, Wu. 2005. “Likelihood-ratio tests for normality”. Computational Statistics and Data Analysis, 49, 709-721.

[3] A. Batsidis, P. Economou, and G. Tzavelas. 2016. “Tests of fit for lognormal distribution”, Journal of Statistical Computation and Simulation, 86 (2), 215-235.

Investigation of the Electricity Consumption of Provinces of Turkey using Functional Principal Components Analysis

Sumeyye İNAL[*1], Gülhayat GÖLBAŞI ŞİMŞEK[1]

[1] Yildiz Technical University, Department of Statistics, Istanbul, Turkey

Nowadays, with the availability of more data as a result of the development of technology, classical univariate and multivariate analysis methods used to analyze these data have become insufficient. Functional Data Analysis (FDA) is branch of statistics that analysis data providing information about curves, surfaces or anything else that varies over time. FDA involves transforming data points to continuous functions, which is basically done using Basis Functions and Roughness Penalty Approach. In FDA analysis is done in terms of functions instead of single data points. It provides a richer set of analysis because it is possible to see trends, points of max and min, rate of change and acceleration [1].

The aim of this study is to investigate the effects of temperature and electricity consumption data on the principal component scores of the provinces by using Multivariate Functional Principal Component Analysis, which is a functional data analysis method, and to reveal the structure of variability between functions.

In this study, the temperature and household electricity consumption data of the provinces in Turkey is examined, and functional data is obtained by using the Fourier basis and the Roughness Penalty Approach respectively. Regularized Functional Principal Components Analysis is used with the aim of analyze the variation structure in the data. Average functions, principle component functions and first principle component derivative functions are formed and interpreted. Then the electrical and temperature data is analyzed together by Multivariate Functional Principal Component Analysis and the most effective provinces on the first principle component function have been identified.

Keywords: Functional Data Analysis; Roughness Penalty Approach; Multivariate Functional Principal component analysis

References

[1] Ramsay, J.O. and Silverman B.W. (2005). Functional data analysis, Second Edition, Springer-Verlag, New York.

[2] Keser, İ.K. (2010). The analyse of Aegean Region rainfall data by using functional data analysis. Dokuz Eylül Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi, 25(1), 41-67.

[*]Corresponding author: sumeyye.karatas.inal@gmail.com

Risk-based Fraud Analysis for Bank Loans with Autonomous Machine Learning

Yunus Emre GÜNDOĞMUŞ[*1], Mert NUHUZ[2] and Müjgan TEZ[3]

[1] Marmara University, Faculty of Art & Science, Phase II Statistics Student Istanbul, Turkey.
[2] Marmara University, Faculty of Art & Science, Phase II Statistics Student Istanbul, Turkey.
[3] Marmara University, Faculty of Art & Science, Department of Statistics Istanbul, Turkey.

In this study, we create Classifier model with Supervised learning by using Customer Data and their loan results for customers who applied for loan. We use various data cleaning, feature extraction and feature selection studies were performed on 67 variables containing the customer's financial information. Our scoring model was created using supervised learning and statistical machine learning based on target variable. The risk score was calculated for the customers and a variable cut-off value was determined according to the sample. They were labelled Fraud and Nan-Fraud with our risk scores. The algorithm instead its learning new customer types. It is now possible to continuously develop itself, to analyze data on a monthly basis and to adapt to the conditions of the period. Tested with real data

Keywords: Autonomous Machine Learning; Risk-Based Scoring**;** Fraud Analysis; Feature Selection

References

[1] Gang Kou, Yi Peng, Guoxun Wang (2014), Evaluation of clustering algorithms for financial risk analysis using MCDM methods https://doi.org/10.1016/j.ins.2014.02.137

[2] Jidong Chen, Ye Tao, Haoran Wang, Tao Chen (2015), Big data based fraud risk management at Alibaba https://doi.org/10.1016/j.jfds.2015.03.001

[3] Y. Pandey, "Credit card fraud detection using deep learning" Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, May–Jun. 2017.

[4] N. Malini and Dr. M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection" in 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17).

[*] yemregun@gmail.com

Multivariate Skew Laplace Normal Distribution: Properties and Applications

Fatma Zehra Doğru [*1] and Olcay Arslan[2]

[1] Giresun University
[2] Ankara University

In this study, we proposed a multivariate skew Laplace normal (MSLN) distribution to model both skewness and heavy-tailedness in multivariate data sets. We derived some distributional properties of this distribution. This newly proposed distribution will also be an alternative to the multivariate skew-t-normal (MSTN) distribution introduced by [1]. The MSLN distribution also has the advantage of having less number of parameters for computational tractability according to the MSTN distribution. Then, we used the expectation-maximization (EM) algorithm ([2]) to compute the maximum likelihood (ML) estimators for parameters of interest. We provided a simulation study to show that the EM algorithm works accurately to estimate the parameters of the MSLN distribution.

Keywords: EM Algorithm; ML Estimation; MSLN

References

[1] Lin, T. I., Ho, H. J., & Lee, C. R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution. Statistics and Computing, 24(4), 531-546.

[2] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Ser B. 39:1–38

[*]Corresponding author: fatma.dogru@giresun.edu.tr

Do Confidence Indicators Have Impact on Macro-financial Indicators? Analysis on the Financial Services and Real Sector Confidence Indexes: Evidence from Turkey

Esra N. KILCI*

Assist. Prof. Dr., Istanbul Arel University

Confidence has played a key role in the recovery of macro-financial outlook of Turkey after the 2000-01 Economic crisis until the last few-year period in which there has been felt deterioration in confidence indicators. Therefore, this study analyses the impact of financial services and real sector confidence indexes on some macroeconomic and financial indicators such as industrial production index, inflation, stock market index, foreign exchange rates and interest rates in Turkey for the period of 2012:05-2019:05. In this study, the unit root properties of the series are tested by using ADF and Fourier ADF unit root test and the causality relationships between the series are investigated by employing Fourier Toda Yamamoto causality test. In this way, we appropriately take into consideration multiple structural breaks without a need of the number, form or date of these breaks through the Fourier approach. The results support the impact of confidence indicators on macro-financial indicators as stock market indices and inflation.

Keywords: Financial Sector Confidence Index; Real Sector Confidence Index; Stock Market; Foreign Exchange Rates

References

[1] Acharya, S., Benhabib, J., and Huo, Z. (2017). The Anatomy of Sentiment-driven Fluctuations. Technical Report, National Bureau of Economic Research.

[2] Afshar T., Arabian, G. & Zomorrodian, R. (2007). Stock return, Consumer Confidence, Purchasing Manager's Index and Economic Fluctuations, Journal of Business & Economics Research, 5(8), 97-106.

[3] Akerlof, G. A. And Shiller, R. J. (2009). Animal Spirits: How Human Psychology Drives the Economy and Why It Matters for Global Capitalism. Princeton, NJ: Princeton University Press

*Corresponding author: esra.kilci@gmail.com

Opportunities in Location Based Customer Analytics

Murat Öztürkmen[*1]

[1]PhD Student, Data Scientist

Firms process all kinds of data they can collect from their customers and provide personal recommendations specific to their customers while increasing their profitability. Spatial data such as structural transaction data, image data and audio data have also been an important part of customer data.

As technological developments such as mobile phone applications, sensor data, web site browsing that allow the collection of spatial data increase, spatial data has become an important tool for providing personalized suggestions to customers.

Spatial data not only allows individual customers to be profiled but also the collective behavior of customer segments.

Spatial data analytics, combined with customer analytics, provide the basis for developing new opportunities and techniques for firms. Next location estimation, spatial recommendation systems, spatial segmentation, location based customer segmentation, dynamic and interactive maps are some of these techniques.

In this study, the place of spatial data in customer analytics is discussed with applications with open datasets and the opportunities provided by location-based customer analytics for firms are discussed. In this context, the basic elements of geographic information systems will be discussed first. Secondly, existing applications such as churn analysis, segmentation and personalized recommendation systems which are frequently used in customer analytics will be mentioned. Finally, the existing applications will be expanded spatially by adding customer location information; the real-life habitat of the customer will also be considered as part of customer analytics, and more consistent predictive and prescriptive analyzes will be suggested.

Keywords: Geospatial; Customer Analytics; Data Science; Location Intelligence

[*]Corresponding author: murat.ozturkmen42@gmail.com

The Effect of Weights on Multi-rater Weighted Kappa Coefficients

Ayfer Ezgi Yilmaz

Department of Statistics, Hacettepe University, Ankara, Turkey.

Cohen's kappa and kappa-like coefficients are popular descriptive statistics for measuring agreement between two raters. Weighted versions of kappa coefficients are used to determine the level of agreement between the ordinal classifications of two raters. Similarly, weighted kappa coefficients have been extended to the case of multiple raters. Although there are different coefficients in multi-rater studies, there is no clear agreement on the use of a particular approach. The choice of weights has also a great importance in the agreement studies because the weighting scheme used affects the values of the weighted kappa coefficients. In addition to linear weights which is the most common weighting scheme, we discuss the quadratic, ridit type, and exponential scores for the multi-rater case. In this study, we conduct a Monte Carlo simulation to compare the accuracy of multi-rater weighted kappa coefficients with each other; the effects of different weighting schemes and different table structures on the accuracy of these coefficients are discussed.

Keywords: Inter-rater Reliability; Multi-raters; Weighted Kappa; Weighting Schemes

*Corresponding author: ezgiyilmaz@hacettepe.edu.tr

Evaluating New Optimization Methods for Two Parameter Ridge Estimator via Genetic Algorithm

Erkut Tekeli[1*], Selahattin Kaçıranlar[2], Nimet Özbay[3]

[1]Department of Computer Technologies, Kozan Vocational School, Çukurova University, Adana, Turkey
[2]Department of Statistics, Faculty of Science and Letters, Çukurova University, Adana, Turkey
[3]Department of Statistics, Faculty of Science and Letters, Çukurova University, Adana, Turkey

In linear regression analysis, the ascending use of two parameter estimators becomes noticeable in the existence of multicollinearity. In such biased estimators, two different parameters offer various improvements in two different subjects. At this circumstance, Lipovetsky and Conklin [4] introduced two-parameter ridge estimator and demonstrated that it provides significant benefits. This estimator eliminates the negative effects of multicollinearity and increases the magnitude of coefficient of multiple determination for the linear regression model. This paper focuses on estimating the parameters of the two-parameter ridge estimator simultaneously with some new optimization techniques of genetic algorithm. To observe the approval of new approaches, we analyze a numerical example as well as a Monte Carlo simulation. In these numerical studies, the benefits of the new optimization techniques are demonstrated and a broad comparison with existing methods is performed.

Keywords: Genetic Algorithm; Multicollinearity; Parameter Optimization; Ridge Estimator; Two Parameter Ridge Estimator

References

[1] Ahn, J.J., Byun, H.W., Oh, K.J., Kim, T.Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. Expert Systems with Applications. 39, 8369–8379.

[2] Al-Hassan, Y.M. (2010). Performance a new ridge regression estimator. Journal of the Association of Arab Universities for Basic and Applied Sciences. 9, 23-26.

[3] Holland, J.H. (1975). Adaptation in natural and artificial systems, University of Michigan Press.

[4] Lipovetsky S., Conklin W.M. (2005). Ridge regression in two parameter solution. Applied Stochastic Models in Business and Industry. 21, 525-540.

*Corresponding author: etekeli@cu.edu.tr

Probabilistic Structural Equation Modeling Approach to Investigate the Relationships between Passenger Perceived Value, Image, Trust, Satisfaction and Loyalty

Tugay KARADAG*[1], Gulhayat GOLBASI SIMSEK[1]

[1]Yildiz Technical University, Department of Statistics, Istanbul, TURKEY

Probabilistic structural equation modeling (PSEM) can be regarded as combination of structural equation modeling (SEM) and Bayesian Network (BN) [1]. One can arrive to the BN phase purely from a structure hypothesized from theory, this is done using classical SEM techniques. As we acknowledge the structure as a SEM, the final BN is then called a probabilistic SEM (PSEM). The other side of the spectrum is a completely data-driven approach, assuming no theory driving the structure. That is an explanatory approach and can be useful in scenarios where data exist, but no extensive theory is available. The resulting BN is then also called an explanatory BN (EBN), rather than a PSEM as the structure is not based on theory as is the case with SEM. If we consider the case where the factors are created according to the theory, but the structural paths are learned using data, the resulting BN is then called a semi-PSEM.

The aim of this study is to conduct PSEM to investigate the relationships between the customer satisfaction and customer loyalty considering image, trust, and perceived value in the context of public transportation. In order to explore the relationship between these latent variables, the passenger survey dataset corresponding high speed rail system (HSRS) in Turkey [2] was analyzed. A measurement model of 37 variables for these five latent variables was verified using confirmatory factor analysis (CFA) following explanatory factor analyses (EFA). It was also shown that the 5- factor measurement model was supported by EBN analysis. Latent variable scores obtained by averaging the original 5 point-Likert scale responses of the variables belonging the same latent factor were re-categorized into 5 of classes. This study was also elaborated focusing on the frequency of use by HSRS customers giving attention to the directions of the arc between satisfaction and loyalty as a main concern.

Keywords: Bayesian Network; High Speed Rail System; Loyalty; Probabilistic Structural Equation Modeling; Satisfaction

Acknowledgements: This research is supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under the support programme of 3001 (Project No 114K093).

References

[1] Yoo, K. (2017). Probabilistic SEM: an augmentation to classical Structural Equation Modelling (Doctoral dissertation, University of Pretoria).

[2] Akyıldız Alçura, G., Kuşakcı, Ş., Gölbaşı Şimşek, G., Gürsoy, M. and S. C. Tanrıverdi. 2015. "Impact Score Technique for Analyzing the Service Quality of a High-Speed Rail System." Transportation Research Record: Journal of the Transportation Research Board 2541: 64-72.

*Corresponding author: karadagt@yildiz.edu.tr

Comparison of Internal Validity Indices According to Distance Measurements in Clustering Analysis

Aydın KARAKOCA[*1], İbrahimDEMİR[2] and Derya ALKIN[3]

[1] Necmettin ErbakanUniversity, The Institute of Science, Department of Statistics
[2] Yıldız Technical University, The Institute of Science, Department of Statistics
[3] Yıldız Technical University, The Institute o f Science, Department of Statistics

Cluster analysis has been applied in many different areas. Although it has many application areas, many studies have been produced on the subject of determining the number of clusters. The number of sets may vary depending on the method selected for the same data set. The possibility of creating different clusters for the same data set is problematic in the selection of the number of clusters. Verification of the cluster results is the most important part of the cluster studies. Internal validity indices are the most widely used approach for cluster validity.

In this study, the performance of internal validity indices has been tested and compared according to different distance measurements. The results obtained from a comprehensive simulation study have profoundly showed the index performance differences.

Keywords: Clustering; Internal Validity Indices; Distance Measuruments

Prediction of Claim Probability in the Presence of Excess Zeros

Aslıhan Şentürk Acar[1]
[1] Hacettepe University, Department of Actuarial Sciences, Ankara, Turkey, aslihans@hacettepe.edu.tr

Claim severity and claim frequency are used to estimate expected pure premium for the next policy period in non-life insurance. Generalized linear models (GLM) are widely preferred to model both components due to the easy interpretation and implementation. Since GLMs have some distributional restrictions, more flexible Machine Learning (ML) methods are applied to insurance data in recent years. ML methods use learning algorithms to establish relationship between the response and the predictor variables as an intersection of computer science and statistics.

Due to the insurance policy modifications such as deductible and no claim discount system, excess zero structure is usually observed in claim frequency data [1]. In the presence of excess zeros, probability of claim may be modelled instead of claim frequency since maximum value of claim counts is observed very rarely in the portfolio. Excess zeros create imbalance problem in data that is difficult to deal with. When the data is highly imbalanced, predictions will be biased towards majority class due to the priors [2]. As the important subject is minority class (claim occurrence), we are interested in a model that will provide high accuracy for the minority class. There are different approaches to overcome class imbalance problem. Standard solution is to rebalance classes with resampling methods before training process.

In this study, we are interested in claim occurrence probability when there is excess of zeros. An imbalanced motor insurance data set taken from a Turkish insurance company is used for the case study. Ensemble methods and neural networks are used for the probability prediction as an alternative to logistic regression and resampling methods are used to deal with class imbalance. Predictive performances are compared.

Keywords: Imbalanced Data; Claim Probability; Classification, Non-life Insurance; Machine Learning.

References

[1] Yip, K. C., & Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics, 36(2), 153-163.

[2] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

Stochastic Linear Restrictions in Generalized Linear Models

M. Revan Ozkale

Cukurova University, Faculty of Science and Letters, Department of Statistics, 01330, Adana, TURKEY

The generalized linear models (GLMs) are important tools in the modelling of the response following an exponential family distribution. In estimating such models, iteratively reweighted least squares (IRLS), also known as maximum likelihood, method is used because of being the structure of nonlinear model. However, if additional information about the regression coefficients exists such as theoretical arguments about the subject under study or estimates from previous studies, then methods diff erent from IRLS should be applied. In this study, we will discuss on estimating the GLMs if an additional information which takes the form of stochastic linear restrictions follows normal distribution. In this context, we will utilize from both the mixed estimation and ridge regression approaches. In addition, we will propose a testing procedure to evaluate the compatibility of the sample and additional information. We introduce three methods for obtaining the ridge parameter which follow the ridge trace, asymptotic mean square error and an iterative procedure in the context of GLMs when additional information exists in the form of stochastic linear restriction. The theoretical discussions are evaluated on a data set concerning the introduction of West Nile Virus in the United States in 1999 where the response follows Poisson distribution.

Keywords: Generalized Linear Models; Stochastic Linear Restrictions; Mixed Estimation; Ridge Regression

References

[1] Nyquist H. (1991). Restricted estimation of generalized linear models. Applied Statistics, 40(1), 133-141.

[2] Özkale MR. (2009) A stochastic restricted ridge regression estimator. Journal of Multivariate Analysis, 100, 1706-1716.

Acknowledgements
This research is partially supported by Research Fund of C¸ukurova University under Project Number 10707.

*Corresponding author: mrevan@cukurova.edu.tr

The GO estimator: A New Generalization of Lasso

Murat Genç[1], M. Revan Özkale[2]

[1]Faculty of Science and Letters, Department of Statistics, Çukurova University
[2]Faculty of Science and Letters, Department of Statistics, Çukurova University

The lasso (least absolute shrinkage and selection operator) proposed by [1] carries out shrinkage of parameters and variable selection, simultaneously. The lasso has many generalizations such as elastic net, group lasso and fused lasso. We present a new shrinkage and variable selection method named the GO estimator, which is based on double shrinkage on the regression coefficients. The ridge, lasso and elastic net are special cases of the GO estimator. The GO estimator contains the shrunken estimator which causes the GO estimator to yield less biased estimates depending on the shrunken parameter compared to the elastic net by maintaining the feature of variable selection. The new estimator has a grouping property similar to the elastic net. We conduct real data and simulation studies to compare the GO estimator with some methods including the lasso and elastic net [2].

Keywords: Elastic Net; Lasso; Shrinkage; Variable Selection

References

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[2] Zou, H., Hastie, T., (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 2, pp.
301-320.

*Corresponding author: mgenc@cukurova.edu.tr

Bivariate Credibility Premiums Distinguishing Between Two Claims Types in Third Party Liability Insurance

Pervin Baylan[1], Serdar Kurt[1], Neslihan Demirel[1] and Jeffrey S. Pai[2]

[1]Dokuz Eylul University, Department of Statistics
[2]University of Manitoba, Warren Centre for Actuarial Studies and Research

The accurate calculation of the premiums plays a key role to prevent the loss of the insurance sector. The premium assigned to each policyholder is based on the number of claims made deriving from an annual claim frequency on an automobile insurance policy. This premium calculation causes an unfair premium evaluation since a policyholder who had small claim is penalized to have the same premium (or risk) rating as one with a more costly claim. This study deals with estimating priori claims of property damages and body injuries in third party liability insurance. The aim is to propose warranted insurance premiums for different types of claims by using a multidimensional credibility model. Generalized linear mixed models (GLMMs) are performed on the data provided from a Canadian insurance company. The statistical model distinguishing between two different claims types, incorporating a bivariate distribution, is presented in this study. In addition, credibility theory is encompassed within the theory of GLMMs and credibility premiums are estimated by means of the B¨uhlmann-Straub credibility model. The results are obtained by using R software.

Keywords: Credibility Theory; Generalized Linear Mixed Models; Logistic Regression; Premium Rating

References

[1] Gómez-Déniz, E. (2016). Bivariate credibility bonus-malus premiums distinguishing between two types of claims. Insurance: Mathematics and Economics, 70, 117-124.

[2] Kafková, S. and Křivánková, L. (2014). Generalized linear models in vehicle insurance. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(2), 383-388.

[3] Kafková, S. (2015). Bonus-malus systems in vehicle insurance. Procedia Economics and Finance, 23, 216-222.

[4] Nelder, J.A. and Verrall, R.J. (1997). Credibility theory and generalized linear models. Astin Bulletin, 27(1), 71-82.

*Corresponding author: pervin.baylan@deu.edu.tr

Churn Analysis for Factoring: An Application in Turkish Factoring Sector

Enis Gumustas[1], Huseyin Budak[2]

[1]Mimar Sinan FA University, Statistics Dept., Istanbul, Turkey
[2]Tani Pazarlama ve Iletisim Hizmetleri A.S., Istanbul, Turkey

Due to the increasing competitive environment in many areas in recent years, customers can easily turn to alternative services. For this reason, it is very important to predict that customers will turn to another service, especially in sectors such as telecom and banking, which have a membership-based revenue model. As in many sectors, in the factoring sector churn prediction models are being developed which predict customers who plan to move to competitors. According to the prediction results, companies aim to prevent customers from leaving the company by developing various campaigns or diff erent actions related to the customers to be lost and to increase the loyalty of the customer to the company. At this stage, focusing on the right customer is critical in order to reduce campaign costs and increase customer loyalty. In order to identify the correct customer, successful prediction models are being developed by using current classification algorithms. However, it would not be enough to treat customer churn prediction as just a classification model. Additional analyzes are needed to provide information to decision processes such as selecting the targeted customer, determining the types of actions to be taken for the customers, and personalizing the actions according to diff erent customer groups. Therefore, it is necessary to consider customer churn prediction as a holistic customer relationship model, which includes the developed forecasting model, as well as analysis to recognize the customer, such as profiling, segmentation.

In this study, a profiling and risk segmentation study was conducted primarily to identify the customers in diff erent dimensions through a data set containing information such as location, demography, transaction history and intelligence results of customers of a private factoring firm. Then, the customer churn prediction model was developed by adding engineered features to the existing data set. Tree - based methods such as CatBoost, Random Forest, LightGBM, and XGBoost have been used for the prediction model. Furthermore, the methods used were compared over metrics such as accuracy, F1 score, sensitivity and precision.

Keywords: Churn Analysis; Financial Data; Classification; Machine Learning

*Corresponding author: 20172107001@msgsu.edu.tr

Two Structural Equation Modelling Approaches for Cloud Use in Software Development

Erhan Pişirir[*1], Oumout Chouseinoglou[2], Cüneyt Sevgi[3] and Erkan Uçar[4]

[1] Hacettepe University, Department of Statistics
[2] Hacettepe University, Department of Industrial Engineering
[3, 4] Bilkent University, Computer Technology & Information Systems

Structural equation modelling (SEM) is a statistical analysis method that can be used to calculate relationships between variables in complex models. SEM is a preferred method for complex social models because of its ability to calculate effects of normally unmeasurable or unobservable variables on each other. This is made possible by using measurable indicators to understand the effect of unmeasurable (latent) variables [1]. There are two different general approaches to SEM analysis, covariance based (CB)-SEM and Partial Least Squares (PLS)-SEM methods. Technology adoption models are complex social models. They are used to estimate factors that affect users' intention to adopt or use innovations over the previous alternatives. These factors might include personal reasons, environmental factors, or business-related depending on the context of the technology and population in the study.

This study is a technology adoption study specifically focusing on the adoption of cloud computing services in software development projects. To model the use intention, a hybrid conceptual model is developed with the inclusion of a novel variable structure (called Person-Organisation-Project, POP) to two existing theories in literature, namely Technology Acceptance Model (TAM) [2] and Technology-Organisation-Environment (TOE) [3]. A questionnaire is designed and personally administered questionnaire sessions are conducted in 30 different software development organisations. 268 valid observations are collected for statistical analysis. Hypotheses based on the conceptual model are tested with CB-SEM and PLS-SEM methods. Thus, the suggested hybrid model is evaluated and also two SEM approaches are compared in the context of a cloud adoption in software development study. After modifications, the final model is reached. The novel hybrid technology adoption model is validated and several practical conclusions are drawn about the population and cloud use in software development.

Keywords: Structural Equation Modelling; Technology Adoption; Cloud Computing; Software Development.

References

[1] Fornell, C., Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18 (1), 39-50.
[2] Davis, F. (1989). Perceived Usefulness, Perceived East of Use, and User Acceptance of Information Technology. MIS Quarterly, 13(3), 319-340.
[3] Tornatzky, L. G., Fleischer, M., Eveland, J. D. (1990). Technological Innovation as a Process, Lexington Books.

[*]Corresponding author: erhanpisirir@hacettepe.edu.tr

A New Approach to Econometric Modelling of Monthly Total Air Passengers: A Case Study for Atatürk Airport

Reşit Çelik[*1], Hasan Aykut Karaboğa[2], İbrahim Demir[3]

[1] Yıldız Technical University, Faculty of Arts and Sciences, Statistics Department, Istanbul, Turkey, rcelik@yildiz.edu.tr.
[2] Yıldız Technical University, Faculty of Arts and Sciences, Statistics Department, Istanbul, Turkey, karaboga@yildiz.edu.tr.
[3] Yıldız Technical University, Faculty of Arts and Sciences, Statistics Department, Istanbul, Turkey, idemir@yildiz.edu.tr.

Air transportation is a major contributor to the development of countries. But, in order to achieve this, airports should have sufficient capacity, sufficient to meet the needs and be accessible. In 2003, air transport sector has developed tremendously with opening of the domestic lines to the competition. This conversion leads decreasing in ticket prices. Also, the convenience of airway transportation and the speed of transportation have led to large increases in the number of passengers. With the effect of the building new airports and increasing in the domestic connecting flights, international airports, especially Atatürk Airport, started to provide services beyond capacity. In order to prevent overcrowding, it is necessary to know the increase in the number of passengers. This is only possible with correct modeling. In our study, Atatürk Airport's total number of passengers is modeled by multiple linear regression analysis. However, it is observed that some of the basic regression assumptions like homoscedastic residuals was not provided. It has also found that, residuals show a special type of the heteroskedastic distribution named butterfly type distribution. The heteroscedasticity detected by RCEV test which is developed in 2017. It is obvious that modeling errors will cause many economic and social losses. Therefore, the analysis is repeated with the Weighted Least Squares Regression method (WCEV Regression). End of the analysis, it is observed that all assumptions are achieved and butterfly distributed residual problem had solved. Adjusted R2 of the weighted model is found 92.1%. Finally, total air passenger of the Atatürk Airport successfully modelled with a new weighted regression model. In addition to this, we believe that, our model will be useful for new Istanbul Airport's service and capacity planning.

Keywords: Regression; Residual Model; Autocorrelation, Studentized Residuals; Heteroscedasticity

[*]Corresponding author: rcelik@yildiz.edu.tr

Analyzing the Competition of HIV-1 Phenotypes with a Quantum Computation Perspective

Bilge BAŞER*

Mimar Sinan Fine Arts University
Department of Statistics

HIV-1 (Human Immunodeficiency Virus) is a virus widespread in the world that can cause immunodeficiency syndrome or AIDS (Acquired Immune Deficiency Syndrome) if it is not treated. HIV-1 attacks the immune system and spreads through blood and lymph nodes. Like all other viruses, HIV-1 cannot replicate on its own and it needs live cells to reproduce itself. The primary target of HIV-1 is lymphocytes, called CD4+T cells. In an infected person, while the CD4+T cells rapidly decrease over time, HIV-1 is greatly increased. With high viral load and low CD4+T cell number, the body's defense mechanism is broken and becomes apparent to many other infections.

Antiretroviral therapy is still being studied for the infection of the HIV-1, and these treatment methods are based on drug designs that are developed using inhibitors that promote the dynamics which provide the development of the virus. However, there is a need for studies to improve the drugs against the infection because of HIV-1 is frequently undergoing mutation and mutant viruses develop resistance against the treatment in use. During virus replication, about ten billion viruses are originated every day. The tendency of such rapid reproduction and high-order mutation provides diversity and evolutionary success for HIV-1.

Therefore, it is important to model the evolutionary development of the virus. The studies known to be done so far have been carried out based on the rules of classical physics. However, referring to Dawkins [1], games of survival are being played on the molecular level, where the rules of quantum mechanics work. Hence, figuring out the development of the virus from a quantum computation perspective may make a difference in drug design.

For this purpose, the replication and developmental process of HIV-1 are modeled as a game with the players of phenotypes following Harada [2] in this study. Since the HIV-1 game is being played on the molecular level, the behaviors of the virus phenotypes are examined from the perspective of quantum computation. The findings obtained by quantum game theory are compared with the decision of the classical evolutionary game theory.

Keywords: HIV-1 Phenotypes; Evolutionary Game Theory; Quantum Game Theory

References

[1] Dawkins, R. (1976). The Selfish Gene, Oxford University Press, Oxford.

[2] Harada, K. (2013). Self-Destruction Dynamics of HIV-1 Quasi-Species Population in the Presence of Mutagenic Activities. Procedia Computer Science. 1259-1265.

*Corresponding author: bilge.baser@msgsu.edu.tr

Analysis of Data Comparing the Use of Different Social Media for Scientific Research across Different Countries of the World

Fatima R. Haris*

School of Mathematics, Cardiff University, Cardiff, Wales, United Kingdom

The use of social media in scientific research is growing [1-2], enabling researchers to gather data beyond national borders. This provides opportunity for using "big data" to speed up progress in many areas of research including business and health. The aims and objectives of this research is to analyse the use of *Facebook*, *Instagram*, *Twitter*, *WeChat* and *WhatsApp* in scientific research and how they vary between different research areas and between different countries of the world.

The methodology used in the research involved using the *Web of Science (WOS)* to search for Abstract of articles that mention *Facebook*, *Instagram*, *Twitter*, *WeChat* and *WhatsApp*. The search was conducted between April-August 2019 covering the period 1900-2019. The order of number of publications (27 August 2019) that have mentioned these media are as follows: *Facebook* (20,599)>*Twitter* (19,017)>*Instagram* (1,890)>*WeChat* (1,185), *WhatsApp* (1,185). For *Facebook*, USA has the highest number of publications (5,987) followed by England, Australia, China and Spain. USA (5,963) also leads the world with respect to *Twitter* related articles, followed by England, China, Spain and India. With regards to *WhatsApp*, Spain (120) has the highest number of publications (101). This is followed by India, USA, England and Brazil. Articles where *Instagram* is mentioned is highest for USA (532) followed by Australia, Spain, England and Indonesia. China leads in the use of *WeChat* (1,272) followed by USA, Australia, England and Canada.

In all five social media investigated in this study, the *WOS* research area that comes up at the top is computer science. When combining a specific social media with 'health' and 'business', the number of publications related to 'health' is higher compared to 'business' with the exception of *WeChat* which has almost identical numbers for 'business' (96) and 'health' (95) related articles. *WhatsApp* has the highest 'health' related articles, followed by *Facebook* (13.6%) and *Instagram* (13.5%). With respect to 'business', *WeChat* has the highest percentage (8%) followed by *Facebook* (6.5%) and *Instagram* (6.5%). Detailed discussion of the variation in the use of social media in different areas of research will be discussed.

Keywords: Research; Facebook; Instagram; Twitter; WeChat; WhatsApp

References

[1] Rowlands, I., Nicholas, D., Russell, B., Canty, N. and Watkinson, A., (2011). Social media use in the research workflow. Learned Publishing, 24(3), 183-195.

[2] Watson, J., (2018). Social Media Use in Cancer Care. In Seminars in oncology nursing 34 (2), 126-131.

*Corresponding author: harisf@cardiff.ac.uk

Finding the Determinants of National Problem Perceptions of Turkish Citizens

Özlem Kiren Gürler*[1] İpek Deveci Kocakoç[1]

[1] Dokuz Eylül Üniversitesi, İİBF, Ekonometri Bölümü

There are many problems facing people and societies in the globalizing world. Significant changes and transformations bring many problems. Similar socio-economic developments cause countries to face the same problems. When individual countries are examined, global security and economic problems are encountered. Security, economy, and poverty are in the top three of countries' most important problems [1]. As can be seen from the results of most public and private surveys, the top three problems in Turkey are security, economy, and education [2].

In this study, we aim to reveal the demographic and socio-economic factors that lead people to the perception of "the most important problem in the country" by both econometric and classification tree models. Importance, relation, and pressure of social factors such as traditions, social environment, religion, political opinion, appearance, etc. are taken as independent variables besides demographic variables.

9719 citizens' responses in the "Life Satisfaction Survey 2018" is used as the data set. This survey is conducted by Turkish Statistical Institute and measures the general happiness perception of the individual, the social values, the general satisfaction from main life areas and the satisfaction from public services. In the microdata set; there is data about various subjects such as happiness, life satisfaction level, satisfaction from main life areas, satisfaction from the services of education, health, public security, judicial, transportation, municipality/provincial administration, environment security, level of hope, perception of social pressure [2].

Keywords: National Problems; Turkey; Multinomial Logit; Classification Tree; Social Factors

References

[1]http://www.tr.undp.org/content/turkey/tr/home/library/human_development/_nsani-geli_me-endeksleri-ve-goestergeleri--2018-statistiksel-gue.html

[2] http://www.turkstat.gov.tr/MicroVeri/YMA_2018/english/index.html.

*Corresponding author: ozlem.kiren@deu.edu.tr

Approximation of Continuous Random Variables for The Evaluation of the Reliability Parameter of Complex Stress-strength Models

Alessandro Barbiero[1]
[1]Department of Economics, Management and Quantitative Methods, via Conservatorio 7, Universitá degli Studi di Milano, Italy

A stress-strength model consists of an item, a component, or a system with an intrinsic random strength that is subject to a random stress during its functioning, so that it works until the strength is greater than the stress. The probability of this event occurring is called the reliability parameter. Since stress and strength are often functions of elementary stochastic factors and the form of these functions is usually very complex, it descends that finding their exact statistical distribution, and then the value of the reliability parameter, is at least cumbersome if not actually impossible. It is standard practice to carry out Monte Carlo simulations in order to find this value numerically. A convenient alternative solution to this impasse comprises discretization, i.e., substituting the probability density functions of the continuous random variables with the probability mass functions of properly chosen approximating discrete random variables. Thus, an approximate value of the reliability parameter can be recovered by enumeration. Many discretization methods have been proposed in the literature so far, which may diff er from one another in their ultimate scope and range of applicability [1, 2, 3]. In this work, we will revise and further refine these techniques and apply them to the context of complex stress-strength models. A comparative study will empirically investigate the performance of these methods by considering several well-known engineering problems and give some practical advice on their mindful use.

Keywords: Approximation; Discretization; Monte Carlo Simulation; Reliability Parameter; Stress-strength Model

References

[1] Roy D., Dasgupta T. (2001). A discretizing approach for evaluating reliability of complex systems under stress-strength model. IEEE Transactions on Reliability, 50(2), 145-150.

[2] Barbiero A. (2012). A general discretization procedure for reliability computation in stress-strength models. Mathematics and Computers in Simulation, 82, 1667-1676.

[3] Drezner Z., Zerom D. (2016). A simple and eff ective discretization of a continuous random variable. Communications in Statistics-Simulation and Computation, 45(10), 3798-3810.

*Corresponding author: alessandro.barbiero@unimi.it

A Customer Segmentation Model Proposal for Hospitals: LRFM-V

İpek Deveci Kocakoç*[1], Pınar Özkan[2]

[1] Dokuz Eylül Üniversitesi, İİBF, Ekonometri Bölümü
[2] Dokuz Eylül Üniversitesi, İİBF, İşletme Bölümü

RFM analysis, which is used for easy and efficient analysis and classification of large amounts of customer data, is a popular technique used to perform customer segmentation by examining how recently, frequently and in monetary value customers shop.

Due to its variable structure, which is relatively easy to use and can be processed in harmony with big data, RFM analysis, which is popular and very good for many industries such as retailing, banking and telecommunication, needs to be overcome to be used for service and customer segmentation in the health sector. Therefore, by adding L (length) to the model, the model has become more useful for the health sector [1-4].

In this study, an LRFM analysis was carried out using patient data covering 20 months to develop a CRM strategy for patients receiving services from a private hospital operating in Izmir. However, a new model (LRFM-V) was proposed to consider the fact that the results of the analysis did not meet the needs of the hospital. The proposed model is calculated by R codes. Clustering analysis was carried out using the data obtained from the analysis of the new model and the characteristics identified for each cluster and marketing strategies were proposed in accordance with these characteristics.

Keywords: RFM Analysis; LRFM Analysis; Market Segmentation; CRM; Service Variety; Healthcare Sector

References

[1] Chang, H. H. and Tsay, S. F. (2004). Integrating of SOM and K-man in Data Mining Clustering: An Empirical Study of CRM and Profitability Evaluation. Journal of Information Management, 11(4), 161–203.

[2] Shih, Y.Y. and Liu, C.Y. (2003). A Method for Customer Lifetime Value Ranking: Combining the Analytic Hierarchy Process and Clustering Analysis. Database Marketing & Customer Strategy Management,11(2):159–172.

[3] Wei, J. T., Lin, S. Y., Weng, C. C., and Wu, H. H. (2012). A Case Study of Applying LRFM Model in Market Segmentation of A Children's Dental Clinic. Expert Systems with Applications, 39(5), 5529-5533.

[4] Wu, H.H., Lin, S.Y., and Liu C.W. (2014). Analyzing Patients' Values by Applying Cluster Analysis and LFRM Model in a Pediatric Dental Clinic in Taiwan. The Scientific World Journal, 2014:1-7.

*Corresponding author: ipek.deveci@deu.edu.tr

The Effect of WoE Transformation on Credit Scoring by using Logistic Regression

Zeynep Bal[*1], M. Aydın Erar[1]

[1] Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey

Credit is one of the most considerable item of the bank's balance sheet so the credit risk has an important position among the risk branches. In this context, determining credit decision to consumers is crucial for credit institutions. The models used in credit scoring are assisted to organizations on the sidelines of making decision to extend loans to consumers.

In this study, the analysis were performed on the data set containing the credit card applications. Within the scope of Basel II, the data set must contain at least one economic cycle, a minimum of 5 years for retail loans [1]. In this regard, model development sample covers the time intervals of 06 /2010 - 05/2015 and the minimum 5 years requirement is provided.

In the analysis conducted during the credit scoring, one of the most important Bank's data was used within the framework of the acceptances signed within the scope of data protection and usage. According to the confidentiality agreement, only the results of the analysis are included in the study.

In the analysis, the raw data set is first separated into training and test data sets at different time periods in terms of ensuring model validation. After that, the logistic regression is applied to both raw and transformed variables with WoE (Weight of Evidence) for the estimation of probability of default.

In presence of the categorical variables, all variables are transformed with WoE technique to make them independent of scale instead of adding dummy variables. In this way, the predictive success of raw and transformed logistic regressions were compared each other and the impact of the transformation in predictive power of the variables was measured.

All the analysis was executed by SAS Enterprise Miner, SAS Enterprise Guide and R studio programs. As a result, the transformed logistic regression gives better results for each data sets against logistic regression with raw variables.

Keywords: Credit Scoring; Logistic Regression; Weight of Evidence; Dummy Variables; Model Validation

References

[1] TBB Çalışma Grubu. (2006). Kredi Riski Modelleri. Bankacılar Dergisi (57).

[*]Corresponding author: zeynep.bal@msgsu.edu.tr

Highlighting a Mathematical Property of Sample ACF for Time Series Analysis

Rahim Mahmoudvand[1]

[1]Bu-Ali Sina University, Hameda, Iran

It is around a century that sample autocorrelation function (ACF) has been introduced and used as a standard tools in time series analysis. A vast literature can be found on the statistical properties of the sample ACF. It has been shown that the sum of the sample autocorrelation over the lags 1 to T − 1 is -0.5 for all time series of length T ([1]-[5]). However, this property has not been deeply discussed in the literature. To the researchers and students it must appear mysterious that standard textbooks on time series analysis fail to acknowledge this property. This paper, consider this property and try to follow up the drawback that might be seen with the application of sample ACF in practice.

Keywords: Time series analysis; Autocorrelation Function; Stationarity

References

[1] Hassani, H. (2009). Sum of the sample autocorrelation function. Random Oper. Stoch. Equ., 17, 125–130.

[2] Anderson, O. D. (1975). Bounding sums for the autocorrelations of moving average processes. Biometrika, 62(3), 706–707.

[3] Anderson, O. D. (1983). On the use of autocorrelation in forecasting: Discussion of Aucamp and Eckardt's article in vol. 21, no. 1 of this journal. Technological Forecasting and Social Change, 24(4), 343–349.

[4] Anderson, O. D. (1990). Small-sample autocorrelation structure for long-memory time series. Journal of the Operational Research Society, 41(8), 735–754.

[5] Anderson, O. D. (1995). More eff ective time-series analysis and forecasting. Journal of computational and applied mathematics, 64(1-2), 117–147.

*Corresponding author: r.mahmodvand@gmail.com

An Approach for Considering Claim Amount and Varying Deductibles in Designing Bonus-Malus Systems

Atefeh Moradi [1], Maryam Sharafi[2], Rahim Mahmoudvand [3]

[1]Atefehmoradi324@gmail.com
[2] mmaryamsharafi@gmail.com
[3]r.mahmodvand@gmail.com

Determining fair premium is very important for both insured and insurer. There is a vast literature on this topic. Bonus-Malus System (BMS) is a common approach for adjusting premiums in automobile insurance. BMS consider historical claim of policyholders and usually give a discount in the premium on no claim and increase the premium if there is a claim in the previous year. There are many different types of BMS. Despite of the importance of severity of loss for ratemaking in automobile insurance, most of BMSs use only the frequency of claim. In this case, BMS penalize equally the policyholders with different claims severity but with the same number of claims. In this paper, in addition to the number of claims, the amount of claim is also considered. Specifically, the premium relativities are softened and the policyholders who are in the malus zone are subject to per claim deductibles. The deductibles are found using the indifference principle. Finally, we analyze a real data set using the idea of this paper.

Keywords: Bonus-malus System; Varying Deductible; Indifference Principle; Allocation Principle.

References

[1] J. Lemaire, “Bonus-malus systems in automobile insurance”, Springer science & business media, 2012.

[2] J. Lemaire, H. Zi, “High deductibles instead of Bonus-Malus: can it work?”, ASTIN Bulletin, The Journal of the IAA, 1994,24(1), pp.75-86.

[3] J. Holtan, “Bonus Made Easy 1”, ASTIN Bulletin, The Journal of the IAA, 1994, 24(1), pp. 61-74.

[4] M. Denuit, X. Maréchal, S. Pitrebois, and J-F. Walhin, “Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems”, John Wiley & Sons, 2007.

[5] O. Ragulina, “Bonus-malus systems with different claim types and varying deductibles”, arXiv preprint arXiv: 170700917. 2017.

[6] S. Pitrebois, J-F. Walhin, M. Denuit, “Bonus-malus systems with varying deductibles”, ASTIN Bulletin, The Journal of the IAA, 2005, 35(1), pp. 261-74.

Hiv-1 Protease Cleavage Site Prediction with Generating Dataset Using a New Encoding Scheme Based on Physicochemical Properties

Metin YANGIN[1], Ayça ÇAKMAK PEHLİVANLI[1], Bilge BAŞER[1]

[1] Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey

AIDS occurs as a result of HIV, weakening the defense mechanism of the human immune system against infection. HIV/AIDS remains one of the world's most important public health problems, especially in countries with low incomes. HIV infection is an infection that causes immunity to the host cell, resulting in a wide clinical picture ranging from retrovirus-mediated carrier to severe fad and fatal diseases.

First, HIV-1 protease data consists of eight amino acid sequences and peptides called octamers. Amino acids have many different physicochemical properties such as hydrophobicity, polarity, molecular weight. Physicochemical properties of amino acids are frequently used in the classification of proteins [1]. As the peptides are composed of amino acids, data on the 566 physicochemical properties of amino acids have been compiled. Assignment of value to missing values was evaluated with maximum expectation and Jackknife methods.

Then, the data set was created by combining HIV-1 protease data generated by Rögnvaldsson et al. with a new encoding approach [2]. These data consist of the eight amino acid sequence called octamer and the state of cleavage site. The data set is completed by subtracting the repetitive and different class variables from the octamers in the data set. One of the most important characteristics that distinguishes this study from previous studies is that there are 544 features in previous studies and 566 features have been used in this study. Machine learning methods are utilised to cleavage site (cleavage, noncleavage) prediction for this data set.

In this work, it is possible to look accuracy, however probability excess is more useful in imbalanced dataset. The results obtained in the classification algorithms are examined in terms of precision, specificity, sensitivity, accuracy, AUC and probability excess values. The best result for probability excess was obtained with random forest method with 0.77 value.

Key words: Classification Algorithms; HIV-1 Protease; Cleavage Site Prediction, Peptide; Physicochemical Properties.

References

[1] Gonzalez, M.W. ve Kann, M.G. (2012). Chapter 4: Protein Interactions and Disease, PLoS Comput Biol, 8(12), e1002819.

[2] Rivas, J.D.L. ve Fontanillo, C. (2010). Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks, PLoS Computational Biology, 6(6).

Wavelet Regression for Noisy Data

Gökçe Nur Taşağıl * and Eylem Deniz[1]

[1] Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey

This study emphasizes the statistical evaluation of wavelet transform and thresholds. It investigates how the thresholding with different wavelet types affect the model. The main purpose of thresholding is to reduce the noise in the original data and to ensure its smoothness. By means of reverse conversion made as a result of thresholding, the statistical analysis can be handled over a data set containing important details instead of original one. In analysis, a benchmark data set, which includes are daily average temperature, wind speed, humidity and pressure variables between 2018-2019 recorded by the German meteorological service, were used. From the plotting line graphs of these variables, it can be seen that the graphs drawn for each variable consist of noisy data. For this reason, all the variables were subject to discrete wavelet transform, and then the wavelet coefficients obtained from this transformation were applied to universal thresholding and cross validated thresholding methods, respectively.

According to analysis results, it can be said that thresholding methods provide more suitable data sets for the regression analysis by means of inverse transformation. In addition, the estimated regression models over the wavelet coefficients obtained from different wavelet types and thresholding methods were compared with respect to RMSE and Akaike Information Criterion using k -fold cross validation.

Keywords: Wavelet Transform; Thresholding; Regression; k-fold Cross Validation

References

[1] Nason, G.P. (2008). Wavelet Methods in Statistics, Springer.

[2] Vidakovic, B. (1999). Statistical Modeling by Wavelet, John Wiley & Sons Inc.

[3] Morettin, P., Pinheiro, A. and Vidakovic, B. (2017). Wavelet in Functional Data Analysis, Springer.

[4] Fryzlewicz, P. (2010). Wavelet Methods. Wires Computational Statistics. Volume (2), 654-667.

*Corresponding author: gokcenurtasagil@gmail.com

An Application of XGBoost on Diabetes Data

Yangın, Gülçin [*1], Ozdamar, E. Ozge[2]

[1,2] Mimar Sinan Fine Arts University

XGBoost; (Extreme Gradient Boosting) is a machine learning algorithm based on decision trees and Gradient Boosting, proposed by Tianqi Chen and Carlos Guestrin in 2016. It has been mostly applied to a wide range of applications using an accelerated frequency; such as energy, health and finance. XGBoost provides many advantages in terms of speed and performance compared to many machine learning algorithms.
In this study, the performance of XGBoost was compared with other decision tree algorithms over a benchmark dataset that consists of people with hypertension and/or diabetes. In analysis, the following algorithms are handled by using R program: Decision Tree, Random Forest, Gradient Boosting and XGboost algorithms. According to analysis results, the classification achievements of algorithms were discussed in detail.

Keywords: XGBoost; Decision Trees; Diabetes

References

[1] Chen Tianqi and Guestrin Carlos (2016). XGBoost: A Scalable Tree Boosting System. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794, San Francisco, California, USA

[*]Corresponding author: gulcn_yangin@gmail.com

Analysis of the Science Scores of Turkish Students in PISA 2015 via Multilevel Models

Gül TİMOÇİN [1], Elif ÜNAL ÇOKER *[2]

[1] DenizBank, Istanbul, Turkey
[2] Mimar Sinan Fine Arts University, Department of Statistics, Istanbul, Turkey

In order to reach a developed society, we should first focus on education. The more developed the education level of a society, the more hopeful the future of that society is. For this reason, it is extremely important to measure the success of the education system. From previous studies, it is known that the success of students is affected, among other things, by individual psychologies and personalities, schools' environment and teachers' support. In accordance with these parameters, the aim of this study is to develop a model so as to detect variables having impact on Turkish students' science scores, by using the latest data made available by the Program for International Student Assessment (PISA) in 2015.

The science competency score, which is determined as the focused area in PISA 2015, is used as a dependent variable in the modeling process. This score is calculated as the average of the students' level of science scores. Factor analysis is applied to the independent variables for the purpose of data reduction. After obtaining the factors, two level multilevel regression models are developed to examine the students' science scores. "Science equipment status school" is the variable having the largest effect on science score. The other effective variables on science score can be listed as the following: material obstacles, absence of teachers, other equipment, gender, age, possession of a computer, study desk, internet and love of science work etc. If the results are interpreted, it can be clearly seen that encouraging students for doing scientific research is enormously important. As a result, it can argued that it would be useful for parents to encourage their children both to study at home and do scientific research, and also to teach them the importance of asking questions.

Keywords: Multilevel Regression Models; PISA; Science Competency Score

*Corresponding Author: elif.coker@msgsu.edu.tr

Part VII

Contributed Papers (Full)

Chaos Control in Chaotic Dynamical Systems Via Auto-tuning Hamilton Energy Feedback

Atike Reza Ahrabi
Department of Electrical Engineering,
Islamic Azad University, Mashhad
Branch, Mashhad , Iran
r.ahrabi@ymail.com

Hamid Reza Kobravi
Research Center of Biomedical
Islamic Azad University, Mashhad
Branch, Mashhad , Iran
hkobravi@mshdiau.ac.ir

Abstract

Recently in some articles, energy-based feedback control is introduced merely as an approach to suppress the chaos. But in this study ,we have demonstrated that an energy-based feedback controller is capable of changing a chaotic dynamics to the other chaotic dynamics. By using energy feedback can also convert a chaos dynamics to another chaos dynamics, and the use of energy feedback should not be limited to suppress the chaos. The importance of the issue relates to some practical applications of chaos to chaos control. A short study has been done on Lorenz chaotic system and has been shown that by combining an energy feedback control with a fuzzy self-regulating gain system, chaos to chaos control will be also possible.

Keywords: Hamilton energy; Chaos control; Fuzzy logic system

1. Introduction

Some disorders and their treatment as epileptic seizure, can be possible by converting an unwilling chaotic behavior of the brain activity into another chaotic behavior. It shows the necessity of applying the chaos to chaos control approaches for controlling the biological systems. But, designing such control strategies needs to present some quantitates can describe the dynamics of a chaotic system. Since the changes of the system's dynamics over time can be presented by the Hamilton energy function, this has attracted attention of many researchers in recent years [1,2]. The Hamiltonian energy has recently been used to the chaos control of dynamical systems [1]. The researchers have suggested a method that by which using the energy modulation and negative feedback of the Hamilton function, it has the ability to suppress the chaos in periodic oscillators and hyper-chaotic systems [1]. However, based on the energy feedback control scheme,in this study has been shown that by using the online regulation of energy feedback gain , convert--ing chaotic dynamics into chaotic dynamics will be possible. Design of the chaos to chaos control approaches In control of biological systems, this can be so important in terms of the practical point of view [3]. For example, it has been shown that during epileptic seizures' episodes, there are some bifurcations to states system [4,5]. It is also shown that geometric properties of basin of attraction for the EEG signal change after epileptic seizures [6]. Hence, it can be assumed that the treatment of such disorders is possible by converting an

undesirable chaotic dynamic into some other chaotic dynamical regime. Due to the fact that energy changes can be expressive of the changes in the pattern of the basin of attraction in the state space, hence, it seemed energy-based feedback controller can also be effective in chaos to chaos controlling. So, in this study the potential of the mentioned control strategy for chaos to chaos control has been shown.

2. Methods And Materials

2.1. Lorenz Chaotic System

The studied system in this research is Lorenz chaotic system and it is based on a nonlinear and chaotic behavior. the third-order continuous-time Lorenz system, described by [2]

$$\begin{aligned} \dot{x_1} &= \sigma(x_2 - x_1) \\ \dot{x_2} &= \beta x_1 - x_2 - x_1 x_3 \\ \dot{x_3} &= -\rho x_3 + x_1 x_2 \end{aligned} \quad (1)$$

Where σ, β and ρ are system parameters which in this study we used $\sigma = 10$, $\beta = \frac{8}{3}$, $\rho = 28$.

2.2. The Proposed Control Strategy

2.2.1. Hamilton Energy

By Helmholtz's theorem, we break down any velocity vector field $f(x)$ into a sum of two vector fields ,one divergence-free vector $f_c(x)$ that calculates rotational tensor of $f(x)$ and the other one is, one gradient vector field f_d which carries its divergence [2]

$$f(x) = f_c(x) + f_d(x) \quad (2)$$

Where $f(x)$ is general autonomous dynamical system. Eq.2 can be used to specify the energy of system. the generalized Hamiltonian form for dynamical equations that plays the role of the system's total energy, are

$$\frac{dX}{dt} = [J(x) + R(x)]\nabla H \quad (3)$$

Where ∇H, $J(x)$ and $R(x)$ are the gradient vector of a smooth energy function $H(x)$, a skew-symmetric matrix and asymmetric matrix, respectively.

According to Eq. (2), in our case study, Lorenz chaotic system, for f_d, the responsible part for divergence of the field and f_c, the part that does not cont- ribute to it, we have [2] :

$$f_c = \begin{pmatrix} \sigma y \\ \rho x + xz \\ xy \end{pmatrix} \qquad f_d = \begin{pmatrix} -\sigma x \\ -y \\ -\beta z \end{pmatrix} \quad (4)$$

According to Eq. (3), the Hamilton equation of Lornz system is as follows [2] :

$$H_{Lorenz}(x.y.z) = \frac{1}{2}(-\frac{\rho}{\sigma}x^2 + y^2 + z^2) \quad (5)$$

And hamilton energy derivative ïs:

$$\dot{H}_{Lorenz}(x.y.z) = \rho x^2 - y^2 - \beta z^2 \quad (6)$$

The negative feedback that is proportional to the Hamiltonian function has been applied to energy as follows:

$$\dot{H} = \nabla H^T f_d(r)\nabla H - \mathrm{kH} \quad (7)$$

Where k is Hamilton's Gain.

Our proposal is that if by using an online self-regulating system, the feedback gain control is set up, the ability to control with Hamilton's feedback will not be bounded to the chaos suppression. A Takagi-Sugeno fuzzy logic system in the control system has been used as an online regulator controller. The input of the fuzzy logic system has been the amount of distance of states from the origin and the speed of its changes. The amount of the

gain of energy feedback has been determined by the output of the fuzzy system. Fuzzy rules have been extracted so that the minimum distance of states from the origin and the speed of their variations from the origin stay limited. By doing so, by changing the parameters of the fuzzy system, the stability of the control system is also guaranteed.

3. Results and Discussion

We appraised the usefulness of the proposed control strategy for different initial conditions. Figure1. shows the membership functions assigned to fuzzy system inputs, and Table 1 shows the rules used for fuzzy logic system. When the controller was activated, the system behavior changed after a transitory period. For example, in the presence of a controller and without the controller for the initial conditions (0.1, 0.1 and 0.1), the trajectories in the state space of Lorenz chaotic system is shown in Figure2. The values of Lyapunov spectrums were calculated by using Wolf's well-known method for investigating chaos dynamic changes in the System.

Table 1. Fuzzy rules extracted for Takagi-Sugeno fuzzy logic system

Edot / *E*	*N*	*Z*	*P*
N	**0.01**	**0.7**	**0.2**
Z	**1**	**0.01**	**0.5**
P	**0.5**	**1**	**0.7**

We use for initial conditions [7], the Averaged Lyapunov spectrums as follows:

$$\bar{\lambda} = \frac{1}{i}\sum_{n=1}^{i} \lambda_n \qquad (8)$$

Where the Lyapunov exponent in the n-th initial condition is λ_n and number of initial conditions is i.

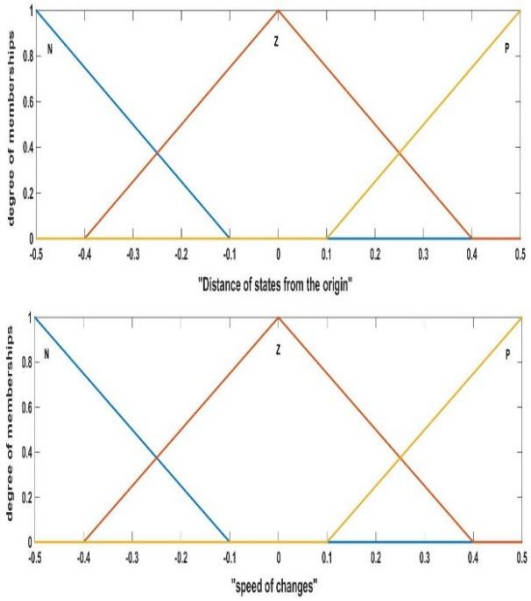

Figure1. Fuzzy membership functions assigned to fuzzy system inputs

The Table 2. shows the results of calculating the Averaged Lyapunov Exponents without a controller and with the presence of a controller.

Table 2. The values of the computed averaged Lyapunov exponents calculated before applying the controller and after apply--ing the controller ($\sigma = 10,\ \beta = \frac{8}{3}$ ، $\rho = 28$)

Uncontrolled system	Controlled System
$\overline{\lambda_x}$ = 0.05044	$\overline{\lambda_x}$ = 0.17634
$\overline{\lambda_y}$ =0.05922	$\overline{\lambda_y}$ = 0.15970
$\overline{\lambda_z}$ =0.04232	$\overline{\lambda_z}$ = 0.15642

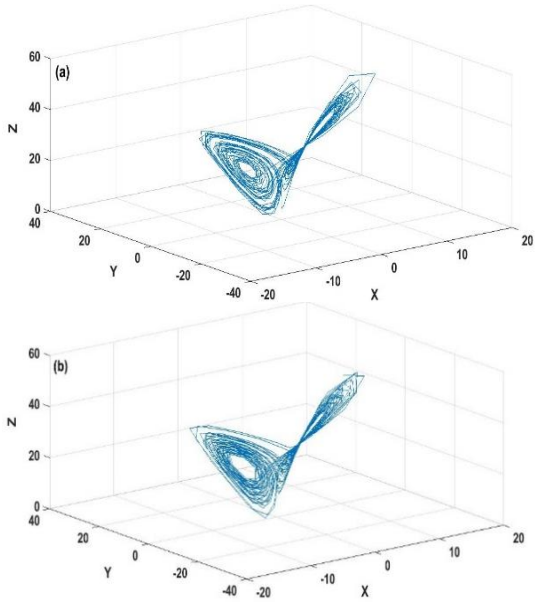

Figure 2. Trajectories in the state space of the Lorenz system in the presence of controller (a) and without the presence of controller (b), with initial conditions: (0.1, 0.1 and 0.1)

According to the results (Table 2), the difference in Lyapunov spectrums in the uncontrolled mode and with the presence of the controller mode were:

$$\begin{aligned} |\overline{\lambda_{x-noco}} - \overline{\lambda_{x-wco}}| &= 0.1259 \\ |\overline{\lambda_{y-noco}} - \overline{\lambda_{y-wco}}| &= 0.10048 \\ |\overline{\lambda_{z-noco}} - \overline{\lambda_{z-wco}}| &= 0.114 \end{aligned} \quad (6)$$

The observed variations of the Lyapunov spectrum means that using the controller gave rise to changing the system dynamics and pattern of the system attractor in the state space. In fact, through exerting the controller the system trajectories gradually entered a new attractor showing the system dynamics converted to some new dynamics.

4. Conclusion

These results indicate that the presence of the fuzzy self-regulating system has made it possible to convert its chaotic dynamics by maintaining the control system's stability. Therefore, it can be claimed that by the use of online optimization mechanisms for controlling the energy feedback, chaos to chaos control can be possible and in addition the application of such control attitude should not be limited to suppress chaos. It can open up more horizons for researchers in the field of chaos control in the future.

References

[1] J. Ma, F. Wu, W. Jin, P. Zhou and T. Hayat, Chaos, AIP, USA, 2017,pp. 1-9.

[2] Sarasola C, Torrealdea FJ, D'Anjou A, Moujahid A, Graña M, Ping Zhou and Tasawar H, Phys Rev E Stat Nonlin Soft Matter Phys, 2004.pp. 1-12.

[3] H.Kobravi and A.Erfanian, Chaos, AIP,USA, 2009.pp. 1-11.

[4] Lashkari S, Sheikhani A, Hashemi Golpayegan M, Moghimi A, Kobravi H, J. Int. Clin. Neurosci. , 2018,pp. 1-15.

[5] K. Lehnertz and C. E. Elger,clin Neurophysiol, 1995.

[6] K. Lehnertz and C. E. Elger,Phys Rev Lett, 1998, pp. 1-19.

[7] A.M. Lopez Jimineza ,C. Camacho Martinez Vara De Rey and A.R. Garcia Toress, Discrete Dynamics in Nature and Society, 2002, pp. 1-41.

Bivariate Intuitionistic Fuzzy Time Series Prediction Model

Ozge Cacgcag Yolcu
Giresun University
ozgecagcag@yahoo.com

Erol Egrioglu
Giresun University
erol.egrioglu@giresun.edu.tr

Eren Bas
Giresun University
eren.bas@giresun.edu.tr

Ufuk Yolcu
Giresun University
varyansx@hotmail.com

Abstract

As it known a time series can be affected by another time series, but almost all the fuzzy inference systems widely used for time series prediction have the univariate structure. The usage of information included by both main and auxiliary time series will improve the prediction performance of fuzzy inference system. This study represents a bivariate intuitionistic fuzzy time series prediction model (B-IFTS-PM) and its some definitions. B-IFTS-PM utilizes pi-sigma artificial neural network (PS-ANN) to determine the relationship between inputs which consist of memberships, non-memberships and crisp observations of both time series, and outputs which consist of the crisp observations of main time series. The training of the PS-ANN is realized by particle swarm optimization. The memberships and non-memberships are obtained via intuitionistic fuzzy C-means clustering. To evaluate the performance of B-IFTS-PM, some stock exchange time series have been analysed. The results indicate that the B-IFTS-PM model, has outstanding prediction performance.

Keywords: Intuitionistic fuzzy time series; bivariate time series; particle swarm optimization, pi-sigma artificial neural network; prediction

1. Introduction

In the time series prediction literature, it can be mentioned two kinds of inference systems (ISs) as probabilistic and non-probabilistic. Statistical-based ones are involved in probabilistic ISs, and fuzzy and computational ISs which are based on fuzzy logic and neural networks, respectively are part of non-probabilistic ISs.

While different kinds of fuzzy inference systems (FISs) [1-3] based on fuzzy sets [4] have been generally used in engineering problems, especially in recent years, they have been widely used for time series prediction. However, most of them are not designed for time series prediction and ignore the dependency structure of the time series observations. Moreover, various FISs [5-13] have been proposed to predict time series from a wide variety of areas.

The other fuzzy sets-based approaches to predict time series are fuzzy time series methods (FTSMs) [14-26].

Furthermore, while the most of the FISs and FTSMs available in the time series prediction only used membership values in the analysis process, just few of them which are based on intuitionistic fuzz sets take into account their memberships and non-membership degrees [27].

From a different viewpoint, a known fact is that fluctuation of a time series can be affected

by fluctuation of some other time series, but almost all the fuzzy inference systems have the univariate structure. From this viewpoint, the usage of information included by both main and auxiliary time series will improve the prediction performance of fuzzy inference system. Moreover, while the most of the fuzzy inference systems used in the time series prediction only include membership values in the analysis process, just few of them which are based on intuitionistic fuzz sets take into account their memberships and non-membership degrees.

In this study related definitions of bivariate intuitionistic fuzzy time series prediction model have been specified and a new prediction model based on these definitions has been proposed. B-IFTS-PM utilizes PS-ANN. PS-ANN to determine the relationship between inputs which consist of memberships, non-memberships and crisp observations of both main and auxiliary time series, and output which consist of the crisp observations of main time series. The optimal weights and biases of PS-ANN are specified by particle swarm optimization (PSO). The memberships and non-memberships are obtained via intuitionistic fuzzy C-means (I-FCM) clustering algorithm. To bring into open performance of the proposed method, some stock exchange time series have been used in the implementation.

The rest of the paper is organized as follow: In the second section, the definitions of the proposed bivariate intuitionistic fuzzy time series prediction model is specified and the prediction process of the proposed B-IFTS-PM is given by an algorithm step-by-step. The third section presents some implementation, and their results including analysis of various stock exchange time series. Finally, the last section involves detailed discussion and future expectations.

2. Bivariate Intuitionistic Fuzzy Time Series Prediction Model

There are lots of FISs and FTSMs using for time series prediction. In recent years, researchers have aimed to progress the prediction performance of their approaches by using non-membership values as well as membership values, in their analysis process. At this point, however, it should be mentioned another issue that needs to be address. Almost all of the available fuzzy-based prediction models compose of univariate structure. However, while the fluctuation of a time series is trying to explain, and modelling to them, using additional information composed of another time series fluctuation will improve prediction performance.

In this study, it is aimed that the prediction performance is improved by using a bivariate intuitionistic fuzzy time series prediction model. For this purpose, firstly, related definitions of bivariate intuitionistic fuzzy time series prediction model can be specified as in (1).

Definition 1. Bivariate intuitionistic fuzzy time series

Let $X_t = \{X_{1,t}, X_{2,t}\}$ is a bivariate time series with real observations. $A_1, A_2, \dots, A_c$ are intuitionistic fuzzy sets on universal set. Bivariate intuitionistic fuzzy time series (BIF_t) is a bivariate time series with real observations, membership and non-membership values for each fuzzy set.

$$BIF_t = \begin{Bmatrix} X_{1,t}, X_{2,t}, \\ \mu_{A_1}(t), \mu_{A_2}(t), \dots, \mu_{A_c}(t), \\ \nu_{A_1}(t), \nu_{A_2}(t), \dots, \nu_{A_c}(t) \end{Bmatrix} \quad (1)$$

Where $\mu_{A_j}(t)$, $\nu_{A_j}(t)$ are membership and non-membership values of t^{th} observation to j^{th} intuitionistic fuzzy set and these values can be obtained intuitionistic fuzzy C-means or other intuitionistic fuzzy clustering methods.

Cluster algorithm is applied for both time series observations and a cluster have p elements in its center.

Definition 2. First order bivariate intuitionistic fuzzy time series prediction model

Let BIF_t be an intuitionistic fuzzy time series that is defined for $X_t = \{X_{1,t}, X_{2,t}\}$. $A_1, A_2, \dots, A_c$ are intuitionistic fuzzy sets on universal set. $\mu_{A_j}(t)$ and $\nu_{A_j}(t)$ are membership and non-membership values of t^{th} observation to j^{th} intuitionistic fuzzy set. In this case First order bivariate intuitionistic fuzzy time series prediction model can be given as:

$$X_t = G\begin{pmatrix} X_{1,t-1}, X_{2,t-1}, \\ \mu_{A_1}(t-1), \dots, \mu_{A_c}(t-1), \\ \nu_{A_1}(t-1), \dots, \nu_{A_c}(t-1) \end{pmatrix} + \varepsilon_t \qquad (2)$$

Where G is a linear or non-linear function and it can be estimated by using linear or non-linear regression analysis, artificial neural network, etc. ε_t is error term and it is a random variable with zero mean.

Definition 3. High order bivariate intuitionistic fuzzy time series prediction model

Let BIF_t be an intuitionistic fuzzy time series that is defined for$X_t = \{X_{1,t}, X_{2,t}\}$. $A_1, A_2, \dots, A_c$ are intuitionistic fuzzy sets on universal set. $\mu_{A_j}(t)$ and $\nu_{A_j}(t)$ are membership and non-membership values of t^{th} observation to j^{th} intuitionistic fuzzy set. In this case p^{th} order bivariate intuitionistic fuzzy time series prediction model can be given as:

$$X_t = G\begin{pmatrix} X_{1,t-1}, \dots, X_{1,t-p}, \\ X_{2,t-1}, \dots, X_{2,t-p}, \\ \mu_{A_1}(t-1), \dots, \mu_{A_c}(t-1), \\ \nu_{A_1}(t-1), \dots, \nu_{A_c}(t-1) \end{pmatrix} + \varepsilon_t \qquad (3)$$

Algorithm: The working principle of B-IFTS-PM.

Step 1. *Determine the parameters of prediction process.*

pn	: # particle of swarm
c_1	: Cognitive coefficient
c_2	: Social coefficient
$maxitr$	: Maximum iteration number
w	: Inertia weight
p	: Model order for membership and non-membership values
q	: Model order for lagged variables part of time series
K	: Order of PS-ANN
$ntest$	: Length of test set
c	: # intuitionistic fuzzy cluster
n	: # observations of time series

Step 2. *Apply I-FCM.*

To get cluster centers ($L_l, l = 1,2,\cdots,c$) and corresponding membership ($U_1 = [u_{lt}]$) and non-membership ($U_2 = [v_{lt}],, l = 1,2, \cdots, c\ ;\ , t = 1,2,\cdots, n - ntest$) values of training data, I-FCM is performed.

Step 3. *Create the inputs (M) and targets (T) of PS-ANN.*

The inputs of PS-ANN can be given as a vector (M) as follow:

$$M = [U_{1p}\ U_{2p}\ y_{t-1} \cdots\ y_{t-q}\ x_{t-1} \cdots\ x_{t-q}] \qquad (4)$$

where U_{1p}, and U_{2p} are constituted the merged memberships and non-memberships by using minimum $t - norm$ considering the model order q. The targets of the system can be presented as follow:

$$T = [y_t] \qquad (5)$$

Step 4. *Randomly generate initial positions and velocities of each particle.*

The Positions and the velocities of particles are randomly generated from distribution of $Uniform(-1, 1)$. Because # inputs of PS-ANN is $(2q + 2c)$, # positions is $d = (2q + 2c)K + K$. The structure of a particle of swarm can be shown as in Figure 1.

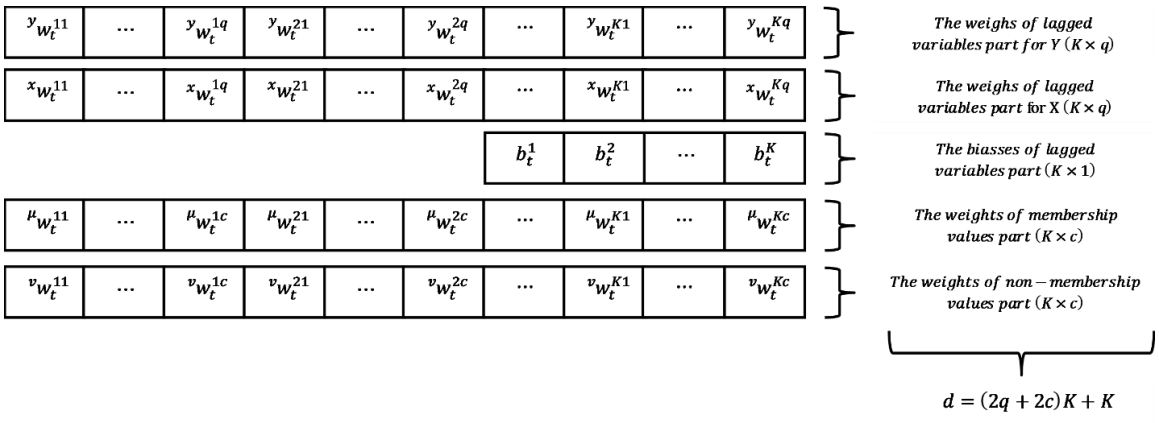

Figure 1. The positions of a particle

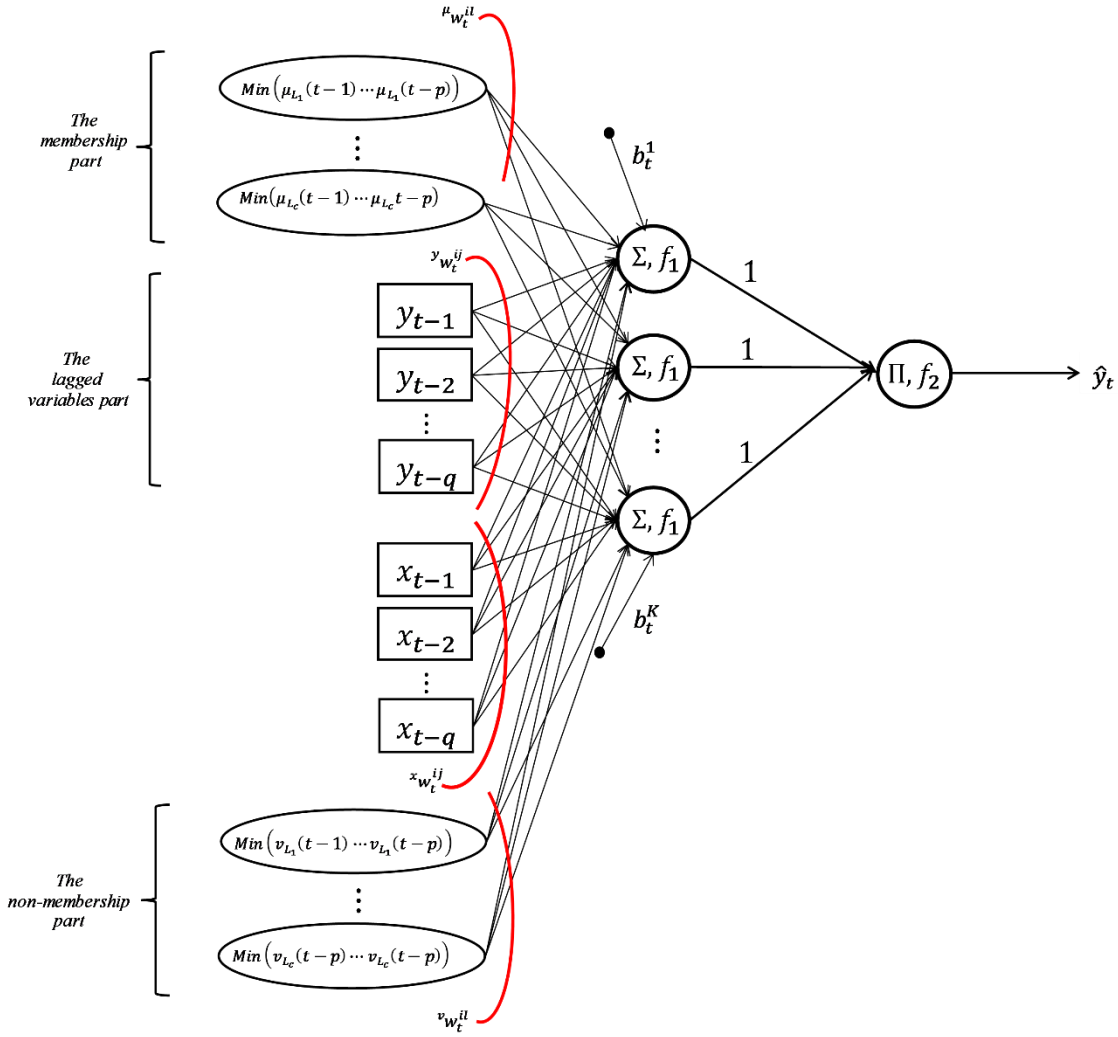

Figure 2. The graphical structure of B-IFTS-PM

Step 5. *Calculate the outputs of PS-ANN.*

Through performing PS-ANN, the outputs are obtained by taking advantage of positions of each particle. This transaction can be explained by followed equations.

$$ {}^{h}o_t^i = f_1\left(\sum_{j=1}^{q}\left({}^{y}w_t^{ij}y_{t-j} + {}^{x}w_t^{ij}x_{t-j}\right) + \sum_{l=1}^{c} {}^{\mu}w_t^{il} min\left(\mu_{L_l}(t-k)\right) + \sum_{l=1}^{c} {}^{v}w_t^{il} min\left(v_{L_l}(t-k)\right)\right), k = 1,2,\dots,p\,;\ t = q+1,\dots,n-ntest \quad (6) $$

$$ o_{PS-ANN_t} = \hat{y}_t = f_2\left(\prod_{i=1}^{K} {}^{h}o_t^i\right) = \frac{1}{1+\exp\left(-\prod_{j=1}^{K} {}^{h}o_t^i\right)} \quad (7) $$

where, $f_1(x) = x$ function is a linear activation function.

Step 6. *Calculate the evaluation function.*

For each particle, root mean square error (RMSE) is calculated as evaluated i.e. fitness function value. B-IFTS-PM with K order and $2q + 2c$ inputs can be seen in Figure 2.

$$ RMSE = \sqrt{\sum_{t=q+1}^{n-ntest}(Target_t - Output_t)^2} \quad (8) $$

Step 7. *Determine $Pbest$ and $Gbest$.*

$$ Pbest = \left[Pb_{r,s}\right];\ r = 1,2.,,,.pn; s = 1,2,\dots,d \quad (9) $$

$$ Gbest = [Pg_s];\ s = 1,2,\dots,d \quad (10) $$

Step 8. *Update the positions and the velocities.*

$$ v_{rs}^{itr+1} = w.v_{rs}^{itr} + c_1.rand_1^{itr}.(Pbest_{rs}^{itr} - P_{rs}^{itr}) + c_2.rand_2^{itr}.(gbest_s^{itr} - P_{rs}^{itr}) \quad (11) $$

$$ p_{r,s}^{itr+1} = p_{r,s}^{itr} + v_{r,s}^{itr+1} \quad (12) $$

Step 9. *Check the stopping criteria.*

If the number of repetitions reach $maxitr$ then stop the process, or else repeat from Step 5 to Step 9 until a predetermined $maxitr$ is reached. When $maxitr$ is reached the optimum values of weights and biases are specified. And then the training of PS-ANN is completed.

3. Implementations

To bring into open the performance of the B-IFTS-PM, daily Istanbul Stock Exchange Market Index (IEX) observed in the period of first 5 months of the years 2009, 2010, 2011, 2012, and 2013 is predicted. While this time series constitute the main time series i.e. requested to predict, EURO/TL and USD/TL exchange rates form the auxiliary time series observed in same period.

In the implementation, the size of test set is taken as 7 and 15. # cluster is changed from 3 to 10. Moreover, p, q, and K are selected by changing from 2 to 5. Comparing the performance of B-IFTS-PM is realized by using the results obtained from six different prediction tools given follow:

ARIMA : Autoregressive moving average
ES : Exponential smoothing
MLP-ANN : Multilayer perceptron ANN
SC : Song and Chissom's FTS method [17]
FF-T1 : Type 1 fuzzy function approach [28]
F-TSN : Fuzzy time series network [26]
ITSFIS : Intuitionistic time series fuzzy inference systems [27]

The results of the methods for the best cases are summarized in Tables 1 and 2.

When the results are considered, it is clearly seen that B-IFTS-PM has the best prediction performance for both case which use EURO/TL, and USD/TL exchange rates, except the results obtained for IEX2010-test size 15-

Table 1. Obtained RMSE values from test sets of IEX – auxiliary time series: EURO/TL and USD/TL

	IEX2009		**IEX2010**		**IEX2011**		**IEX2012**		**IEX2013**		**RMSE's Average**		
	Test Size		**Test Size**		**Test Size**		**Test Size**		**Test Size**		**Test Size**		
	7	**15**	**7**	**15**	**7**	**15**	**7**	**15**	**7**	**15**	**7**	**15**	**Overall**
ARIMA	345	540	1221	1612	1058	1130	651	621	1362	1269	927	1034	981
ES	345	540	1208	1612	1057	1130	651	621	1362	1269	925	1034	980
MLP-ANN	325	525	1077	1603	920	1096	775	783	1315	1233	882	1048	965
SC	1402	1754	1128	1742	1396	1360	1292	1047	1450	1931	1334	1567	1450
FF-T1	446	534	1180	1852	1083	1146	1034	1038	1512	1279	1051	1170	1110
FTS-N	267	514	1050	1357	765	917	590	582	786	1208	692	916	804
I-TSFIS	166	1046	250	251	817	384	277	228	451	1106	392	603	498
B-IFTS-PM[1]	**97**	**152**	**231**	775	**164**	**203**	**133**	**120**	**260**	**323**	***177***	***315***	***246***
B-IFTS-PM[2]	**71**	**159**	304	1096	**105**	**193**	**108**	**131**	**238**	**247**	***165***	***365***	***265***

B-IFTS-PM[1]: auxiliary time series: EURO/TL **B-IFTS-PM[2]**: auxiliary time series: USD/TL

4. Conclusion and Discussion

It is obviously said that fluctuation of a time series may be caused by fluctuation of some other time series as well as own past observations. However, almost all the fuzzy inference systems have the univariate structure. It is a fact that the usage of information included by both main and auxiliary time series will enhance the prediction performance of FISs. Furthermore, the most of FISs available in the time series prediction literature have taken advantage of only membership values in the analysis process. Recent years, although some IFISs which consider memberships and non-membership degrees together have been put forward, the number of them is very few and they have first order prediction model.

In this study related definitions of bivariate intuitionistic fuzzy time series prediction model have been given and B-IFTS-PM has been introduced as a new prediction model. B-IFTS-PM make use of PS-ANN, IFCM and PSO in its prediction process. On the basis of implementations' results, it is clearly said that the proposed B-IFTS-PM has outstanding prediction performance for corresponding time series when we evaluate the results obtained from six benchmark prediction tools.

In the future studies, an ensemble approach in which PS-ANNs will be used for each part or group of inputs, separately can be designed.

References

[1] Mamdani, E.H. (1974). Application of fuzzy algorithms for control of simple dynamic plant. *Proc. Inst. Electr. Eng.*, 121 (1585).

[2] Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.*, SMC-15, 116–132.

[3] Jang, J.S.R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.*, 23, 665–685.

[4] Zadeh, L.A. (1965). Fuzzy sets. *Inf. Control*, 8, 338–353.

[5] Şişman-Yılmaz, N.A and Alpaslan, F.N., Jain, L. (2004). ANFIS_unfold_edin_time For multivariate time series forecasting. *Neurocomputing*, 61, 139–168.

[6] Firat, M., Turan, M.E. and Yurdusev, M.A. (2009). Comparative analysis of fuzzy inference systems for water consumption time series prediction. *J. Hydrol.*, 374, 235–241.

[7] Yurdusev, M.A. and Firat, M. (2009). Adaptive neuro fuzzy inference system approach for municipal water consumption modeling: An application to Izmir. *Turkey J. Hydrol.*, 365, 225–234.

[8] Cheng, C.-H., Wei, L.-Y., Chen, Y.-S., 2009. Fusion ANFIS models based on multi-stock volatility causality for TAIEX forecasting. Neurocomputing 72, 3462–3468.

[9] Atsalakis, G.S. and Valavanis, K.P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Syst. Appl.*, 36, 10696–10707.

[10] Li, K., Su, H. and Chu, J. (2011). Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study. *Energy Build.*, 43, 2893–2899.

[11] Azadeh, A., Asadzadeh, S.M., Saberi, M., Nadimi, V., Tajvidi, A. and Sheikalishahi, M. (2011). A neuro-fuzzy-stochastic frontier analysis approach for long-term natural gas consumption forecasting and behavior analysis: The cases of Bahrain, Saudi Arabia, Syria, and UAE. *Appl. Energy*, 88, 3850–3859.

[12] Chen, Y.-S., Cheng, C.-H. and Tsai, W.-L. (2014). Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting. *Appl. Intell.*, 41, 327–347.

[13] Sarıca, B., Egrioglu, E. and Asıkgil, B. (2018). A new hybrid method for time series forecasting: AR–ANFIS. *Neural Comput. Appl.*, 29, 749-760.

[14] Chen, S.-M. (1996). Forecasting enrollments based on fuzzy time series. *Fuzzy Sets Syst.*, 81, 311–319.

[15] Chen, S.-M. (2002). Forecasting enrollments based on high-order fuzzy time series. *Cybern. Syst.*, 33, 1–16.

[16] Song, Q. and Chissom, B.S. (1993). Fuzzy time series and its models. *Fuzzy Sets Syst.*, 54, 269–277.

[17] Song, Q. and Chissom, B.S. (1993). Forecasting enrollments with fuzzy time series – part I. *Fuzzy Sets Syst.*, 54, 1–9.

[18] Song, Q. and Chissorn, B.S. (1994). Forecasting enrollments with fuzzy time series-part II. *Fuzzy Sets Syst.*, 62, 1–8.

[19] Huarng, K.-H. (2001). Effective lengths of intervals to improve forecasting in fuzzy time series. *Fuzzy Sets Syst.*, 123, 387–394.

[20] Egrioglu, E., Aladag, C.H., Yolcu, U., Uslu, V.R. and Basaran, M.A. (2010). Finding an optimal interval length in high order fuzzy time series. *Expert Syst. Appl.*, 37, 5052-5055.

[21] Egrioglu, E., Aladag, C.H., Basaran, M.A., Yolcu, U. and Uslu, V.R. (2011). A new approach based on the optimization of the length of intervals in fuzzy time series. *J. Intell. Fuzzy Syst.*, 22, 15–19.

[22] Huarng, K.-H. and Yu, T.H.-K. (2006). Ratio-based lengths of intervals to improve fuzzy time series forecasting. IEEE Trans. *Syst. Man, Cybern. Part B*, 36, 328–340.

[23] Yolcu, U., Egrioglu, E., Uslu, V.R., Basaran, M.A. and Aladag, C.H. (2009). A new approach for determining the length of intervals for fuzzy time series. *Appl. Soft Comput.*, J. 9, 647-651.

[24] Alpaslan, F. and Cagcag, O. (2012). A seasonal fuzzy time series forecasting method based on gustafson-kessel fuzzy clustering. *J. Soc. Econ. Stat.*, 1, 1–13.

[25] Tak, N., Evren, A.A., Tez, M. and Egrioglu, E. (2018). Recurrent type-1 fuzzy functions approach for time series forecasting. *Appl. Intell.*, 48, 68–77.

[26] Bas, E., Egrioglu, E., Aladag, C.H. and Yolcu, U. (2015). Fuzzy-time-series network used to forecast linear and nonlinear time series. Appl. Intell. 43, 343-355.

[27] Egrioglu, E., Bas, E., Cagcag Yolcu, O. and Yolcu, U. (2019), Intuitionistic time series fuzzy inference system, *Engineering Applications of Artificial Intelligence*, 82, 175-783.

Stress-Strength Reliability Estimation of Series System with Cold Standby Redundancy at System and Component Levels

Gülce Cüran
Yeditepe University
gulceulupinar@gmail.com

Fatih Kızılaslan
Marmara University
fatih.kizilaslan@marmara.edu.tr

Abstract

In this paper, we consider the stress-strength reliability of the series system with cold standby redundancy at the component and system levels. Classical and Bayesian approaches are studied in order to obtain the reliability estimates when the underlying components (stress, strength and standby components) follow the exponential distribution with different parameters. Bayesian estimation for the reliability of these systems is obtained by using Markov Chain Monte Carlo (MCMC) method when the priors have gamma distribution. The asymptotic confidence and the highest probability density credible intervals are also derived. Monte Carlo simulations are performed to compare the performances of the obtained point and interval estimations for the reliability of these systems.

Keywords: Stress-strength reliability; multicomponent reliability; cold standby.

1. Introduction

In the simplest form, the stress–strength model describes the reliability of a component or system in terms of random variables. In this case, the reliability is defined as $P(X < Y)$ where X is the random stress experienced by the system and Y is the random strength of the system available to overcome the stress. The system fails if the stress exceeds the strength. This main idea was introduced by Birnbaum [2] and developed by Birnbaum and McCarty [3]. Estimation of reliability for a system such as simple system or series system or parallel system or multicomponent system etc. has been extensively discussed by many authors in the literature. Some recent contributions on the topic can be found in the following papers Basirat et al. [1], Kizilaslan [7], Rasethuntsa and Nadar [10], Çetinkaya and Genç [5].

The lifetime of system can be enhanced by introducing some redundancies to it. In the literature, two types redundacies are used, i.e., active and standby redundacies. In the standby redundancy, standby components function only after the failure of its corresponding original component. It is generally called "cold" standby redundancy. The effectiveness of adding cold standby redundancy to any system has been investigated by many authors, see, for example, Shen and Xie [11], Kumar [8], Boland and El-Neweihi [4], Zhao et al. [13].

Standby redundacy can be applied at system or component levels. Generally, stochastic comparison of the systems with standby redundancy at component level versus system level are studied in the literature. The lifetime of the system after standby redundancy at system level is given as $T_S = \phi(T_1, \ldots, T_n) + \phi(T_1^*, \ldots, T_n^*)$, where T_i and T_i^* are the lifetime of component i and standby component i, respectively and $\phi(T_1, \ldots, T_n)$ is the lifetime of coherent system with component lifetimes $T_1, \ldots, T_n$. The lifetime of the system after

standby redundancy at component level is given as $T_C = \phi(T_1 + T_1^*, \ldots, T_n + T_n^*)$.

In this study, we consider the stress-strength reliability for the series system when the standby components are applied at system or component levels. It is assumed that all the system components and standby components are independent but not identically distributed random variables belonging to the exponential distribution. The rest of this paper is organized as follows. In Section 2, we present our model and derive some results about it. In Section 3 and 4, we obtain maximum likelihood estimate (MLE) and Bayesian estimate of stress-strength reliability for the series system with system and component redundancies, respectively. In Section 5, a simulation study is performed to compare the obtained estimates by using Monte Carlo simulations.

2. Model Description

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random variables that have exponential distributions with parameters α and β, respectively. The random variables $Y_1, \ldots, Y_n$ are supposed to represent the standby components. Standby redundancy can be applied at system or component levels. If the standby redundancy is applied at component level, then the lifetime of series system is $Z_i = X_i + Y_i,\ i = 1, \ldots, n$ for each component. In this case, the cumulative distribution function (cdf) and probability density function (pdf) of this system are given by

$$F_{Z_i}(z) = \int_0^z F_x(z-y) f_y(y) dy = 1 + \frac{\alpha e^{-\beta z} - \beta e^{-\alpha z}}{\beta - \alpha}$$

and

$$f_{Z_i}(z) = \frac{\alpha\beta}{\beta-\alpha}(e^{-\alpha z} - e^{-\beta z}), \quad (1)$$

for $\alpha \neq \beta$. Let T represents the stress that is experienced by each strength component Z_i, $i = 1, \ldots, n$. Assume that T has exponential distribution with parameter θ, denoted by $T \sim \text{Exp}(\theta)$. Then, the series system reliability is given as

$$R_{Comp} = P\left(\min_{1 \le i \le n} Z_i > T\right) = \left(\frac{\theta(\alpha+\beta+\theta)}{(\beta+\theta)(\alpha+\theta)}\right)^n \quad (2)$$

If the standby redundancy is applied at system level, then the lifetime of series system is obtained by taking minimum of the lifetimes of components, i.e. $Z_{(1)} = X_{(1)} + Y_{(1)}$ where $X_{(1)} \sim \text{Exp}(n\alpha)$ and $Y_{(1)} \sim \text{Exp}(n\beta)$. In this case, the cdf and pdf of $Z_{(1)}$ are given by

$$F_{Z_{(1)}}(z) = 1 + \frac{\alpha e^{-\beta n z} - \beta e^{-\alpha n z}}{\beta - \alpha},$$

and

$$f_{Z_{(1)}}(z) = \frac{n\alpha\beta}{\beta-\alpha}(e^{-\alpha n z} - e^{-\beta n z}). \quad (3)$$

Then, the series system reliability at system level is given as

$$R_{System} = P(Z_{(1)} > T) = \frac{n\theta(\beta+\alpha)+\theta^2}{(\alpha n+\theta)(\beta n+\theta)}, \quad (4)$$

where T is the stress component and $T \sim \text{Exp}(\theta)$.

3. Estimation of R_{Comp}

In this section, we consider estimations of the reliability of series system when standby redundancy is applied at component level.

3.1 MLE of R_{Comp}

Suppose that there exists an m sample with standby components in the series system. The

strength observations are represented as $Z_{i1},\dots Z_{in}$ and stress T_i for $i=1,\dots,m$. The log-likelihood function of this sample is given by

$$\ell(\alpha,\beta,\theta;\mathbf{z},\mathbf{t}) = nm(\ln(\alpha\beta)-\ln(\beta-\alpha)) + m\ln\theta + \sum_{i=1}^{m}\sum_{j=1}^{n}\ln\left(e^{-\alpha z_{ij}}-e^{-\beta z_{ij}}\right)-\theta\sum_{i=1}^{m}t_i.$$

The MLE of θ is given by $\hat{\theta}=m/\sum_{i=1}^{m}t_i$, and ML estimates of α and β, say $\hat{\alpha}$ and $\hat{\beta}$, are the solutions of the following nonlinear equation system

$$nm\frac{1}{\alpha}+\left(\frac{1}{\beta-\alpha}\right)-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{z_{ij}}{1-e^{-(\beta-\alpha)z_{ij}}}=0,$$

$$nm\left(\frac{1}{\beta}-\frac{1}{\beta-\alpha}\right)+\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{z_{ij}}{1-e^{-(\alpha-\beta)z_{ij}}}=0.$$

Then, $\hat{\alpha}$ and $\hat{\beta}$ is obtained by using numerical methods. Therefore, the ML estimate of R_{Comp}, say $\hat{R}_{Comp}$, is obtained from (2) by using the invariance property of MLE

$$\hat{R}_{Comp}=\left(\frac{\hat{\theta}(\hat{\alpha}+\hat{\beta}+\hat{\theta})}{(\hat{\beta}+\hat{\theta})(\hat{\alpha}+\hat{\theta})}\right)^{n}. \qquad (5)$$

3.2 Asymptotic distribution and confidence interval for R_{Comp}

The observed information matrix of $\boldsymbol{\tau}=(\alpha,\beta,\theta)$ is given as

$$J(\boldsymbol{\tau}) = -\begin{pmatrix}\frac{\partial^2\ell}{\partial\alpha^2} & \frac{\partial^2\ell}{\partial\alpha\partial\beta} & \frac{\partial^2\ell}{\partial\alpha\partial\theta}\\ \frac{\partial^2\ell}{\partial\beta\partial\alpha} & \frac{\partial^2\ell}{\partial\beta^2} & \frac{\partial^2\ell}{\partial\beta\partial\theta}\\ \frac{\partial^2\ell}{\partial\theta\partial\alpha} & \frac{\partial^2\ell}{\partial\theta\partial\beta} & \frac{\partial^2\ell}{\partial\theta^2}\end{pmatrix} = \begin{pmatrix}J_{11} & J_{12} & J_{13}\\ J_{21} & J_{22} & J_{23}\\ J_{31} & J_{32} & J_{33}\end{pmatrix}$$

Since $J_{13}=J_{31}=J_{23}=J_{32}=0$, other entries of this matrix are

$$J_{11}=nm\left(\frac{1}{\alpha^2}-\frac{1}{(\beta-\alpha)^2}\right)+z^*,$$

$$J_{12}=J_{21}=\frac{nm}{(\beta-\alpha)^2}-z^*,$$

$$J_{22}=nm\left(\frac{1}{\beta^2}-\frac{1}{(\beta-\alpha)^2}\right)+z^*,$$

$$J_{33}=\frac{m}{\theta^2},$$

where

$$z^*=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{z_{ij}^2e^{-z_{ij}(\beta-\alpha)}}{(1-e^{-z_{ij}(\beta-\alpha)})^2}$$

The expectations of the entries of the observed information matrix can not be obtained analytically. Hence, the Fisher information matrix $I(\boldsymbol{\tau})=E(J(\boldsymbol{\tau}))$ can be obtained by using numerical methods.

The MLE of R_{Comp} is asymptotically normal with mean R_{Comp} and asymptotic variance

$$\sigma^2_{R_{Comp}}=\sum_{j=1}^{3}\sum_{i=1}^{3}\frac{\partial R_{Comp}}{\partial\theta_i}\frac{\partial R_{Comp}}{\partial\theta_j}I_{ij}^{-1},$$

where I_{ij}^{-1} is the $(i,j)^{th}$ element of the inverse of $I(\boldsymbol{\tau})$, see Rao [9]. Afterwards,

$$\sigma^2_{R_{Comp}}=\left(\frac{\partial R_{Comp}}{\partial\alpha}\right)^2I_{11}^{-1} +2\frac{\partial R_{Comp}}{\partial\alpha}\frac{\partial R_{Comp}}{\partial\beta}I_{12}^{-1}+\left(\frac{\partial R_{Comp}}{\partial\beta}\right)^2I_{22}^{-1},$$

Note that $I(\boldsymbol{\tau})$ can be replaced by $J(\boldsymbol{\tau})$ when $I(\boldsymbol{\tau})$ is not obtained. Therefore, an asymptotic $100(1-\gamma)\%$ confidence interval of R_{Comp} is given by

$$R_{Comp} \in \left(\hat{R}_{Comp} \pm z_{\gamma/2}\hat{\sigma}_{R_{Comp}}\right),$$

where $z_{\gamma/2}$ is the upper $\gamma/2$th quantile of the standard normal distribution and $\hat{\sigma}_{R_{Comp}}$ is the value of $\sigma_{R_{Comp}}$ at the MLE of the parameters.

3.3 Bayes Estimation of R_{Comp}

Let α, β and θ be random variables that have independent gamma prior distributions with parameters (a_i, b_i), $i = 1,2,3$, respectively. A gamma random variable X with parameters (a_i, b_i) is given as

$$f(x) = \frac{b_i}{\Gamma(a_i)} x^{a_i-1} e^{-xb_i}, \quad x > 0, \quad a_i, b_i > 0.$$

The joint posterior density function of α, β and θ is

$$\pi(\alpha, \beta, \theta|\mathbf{z}, \mathbf{t}) = I(\mathbf{z}, \mathbf{t})\alpha^{nm+a_1-1}\beta^{nm+a_2-1}(\beta - \alpha)^{-nm}\theta^{a_3+m-1}e^{-\alpha b_1-\beta b_2-\theta\left(b_3+\sum_{i=1}^{m} t_i\right)}z_{\alpha,\beta} \quad (6)$$

where $I(\mathbf{z}, \mathbf{t})$ is the normalizing constant and written by

$$\frac{I(\mathbf{z}, \mathbf{t})^{-1}}{\Gamma(a_3 + m)}\left(b_3 + \sum_{i=1}^{m} t_i\right)^{a_3+m} = \int_0^\infty \int_0^\infty \left(\frac{\alpha\beta}{\beta-\alpha}\right)^{nm} \alpha^{a_1-1}\beta^{a_2-1}e^{-\alpha b_1-\beta b_2} z_{\alpha,\beta} d\alpha d\beta$$

where

$$z_{\alpha,\beta} = e^{\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \ln\left(e^{-\alpha z_{ij}} - e^{-\beta z_{ij}}\right)\right)}.$$

The Bayes estimator of R_{Comp} using the SE loss function is

$$\hat{R}_{Comp}^{Bayes} = \int_0^\infty \int_0^\infty \int_0^\infty R_{Comp}\pi(\alpha, \beta, \theta|\mathbf{z}, \mathbf{t}) d\alpha d\beta d\theta. \quad (7)$$

This integral can not be computed analytically so some approximation methods are needed. In order to obtain the Bayes estimator of R_{Comp}, we use MCMC method.

3.3.1 MCMC Method

The joint posterior density function of α, β and θ is given in (6). Then, the marginal posterior density functions are given respectively as

$$\theta|\mathbf{t} \sim Gamma\left(m + a_3, b_3 + \sum_{i=1}^{m} t_i\right),$$

$$\pi(\alpha|\beta, \mathbf{z}) \propto \alpha^{nm+a_1-1}(\beta - \alpha)^{-nm}e^{-\alpha b_1}z_{\alpha,\beta}$$

and

$$\pi(\beta|\alpha, \mathbf{z}) \propto \beta^{nm+a_2-1}(\beta - \alpha)^{-nm}e^{-\beta b_2}z_{\alpha,\beta}.$$

Then, samples of θ can be easily generated by using a gamma distribution. Since the posterior distribution of α and β cannot be reduced analytically to a well-known distribution, it is not possible to sample directly by standard methods. We use the Metropolis-Hasting algorithm with the normal proposal distribution to generate a random sample from the posterior density of α and β in our implementation. The hybrid Metropolis-Hastings and Gibbs sampling algorithm, suggested by Tierney [12], is used to solve our problem. This algorithm combines the Metropolis-Hastings scheme with the Gibbs sampling scheme under the Gaussian proposal distribution.

Step 1: Start with initial guess $\alpha^{(0)}, \beta^{(0)}$.

Step 2: Set $i = 1$.

Step 3: Generate $\theta^{(i)}$ from $Gamma(m + a_3, b_3 + \sum_{i=1}^{m} t_i)$.

Step 4: Generate $\alpha^{(i)}$ from $\pi(\alpha|\beta, \mathbf{z})$ using the Metropolis-Hastings algorithm with the proposal distribution $q_1(\alpha) \equiv N(\alpha^{(i-1)}, 1)$ as follows.

(a) Let $v = \alpha^{(i-1)}$.

(b) Generate w from the proposal distribution q.

(c) Let

$$p(v,w) = \min\left\{1, \frac{\pi(w|\beta^{(i)},\mathbf{z})q(v)}{\pi(v|\beta^{(i)},\mathbf{z})q(w)}\right\}.$$

(d) Generate u from $Uniform(0,1)$. If $u \le p(v,w)$, then accept the proposal and set $\alpha^{(i)} = w$; otherwise, set $\alpha^{(i)} = v$.

Step 5: Similarly, $\beta^{(i)}$ is generated from $\pi(\beta|\alpha,\mathbf{z})$ using the Metropolis-Hastings algorithm with the proposal distribution $q_2(\beta) \equiv N(\beta^{(i-1)},1)$.

Step 6: Compute the $R_{Comp}^{(i)}$ at $(\alpha^{(i)},\beta^{(i)},\theta^{(i)})$.

Step 7: Set $i = i+1$.

Step 8: Repeat Steps 2 through -7, N times and obtain the posterior sample $R_{Comp}^{(i)}$, $i = 1,\dots,N$.

These samples are used to compute the Bayes estimate and to construct the HPD credible interval for R_{Comp}. Then, the Bayes estimate of R_{Comp} under a SE loss function is given by

$$\hat{R}_{Comp}^{Bayes} = \frac{1}{N-M}\sum_{i=M+1}^{N-M} R_{Comp}^{(i)}, \qquad (8)$$

where M is the burn-in period. The HPD $100(1-\gamma)\%$ credible interval of R_{Comp} is also obtained by using the method of Chen and Shao [6].

4. Estimation of R_{System}

In this section, we consider estimations of the reliability of series system when standby redundancy is applied at system level.

4.1 MLE of R_{System}

Suppose that there exists an m sample with standby components in the series system. Let $Z_{i(1)}$, $i = 1,\dots,m$ represent the strength components and T_i, $i = 1,\dots,m$ be stress components. The log-likelihood function is

$$\ell = m[\ln(n\alpha\beta\theta) - \ln(\alpha-\beta)] + \sum_{i=1}^{m}\ln(e^{-\beta n z_i} - e^{-\alpha n z_i}) - \theta\sum_{i=1}^{m} t_i.$$

Then, the MLE of θ is given by $\hat{\theta} = m/\sum_{i=1}^{m} t_i$, and ML estimates of α and β, say $\hat{\alpha}$ and $\hat{\beta}$, are the solutions of the following nonlinear equation system

$$m\left(\frac{1}{\alpha} + \frac{1}{\beta-\alpha}\right) - \sum_{i=1}^{m}\frac{n z_i}{1-e^{-z_i(\beta-\alpha)n}} = 0,$$

$$m\left(\frac{1}{\beta} - \frac{1}{\beta-\alpha}\right) - \sum_{i=1}^{m}\frac{n z_i}{1-e^{-n z_i(\alpha-\beta)}} = 0.$$

Then, the MLE of R_{System}, say $\hat{R}_{System}$, is obtained from (4) by using the invariance property of MLE

$$\hat{R}_{System} = \frac{\hat{\theta}[\hat{\theta}+n(\hat{\alpha}+\hat{\beta})]}{(\hat{\beta}n+\hat{\theta})(\hat{\alpha}n+\hat{\theta})}. \qquad (9)$$

4.2 Asymptotic Distribution And Confidence İnterval For R_{System}

The entries of the observed information matrix of $\boldsymbol{\tau} = (\alpha,\beta,\theta)$ is given as $J_{13} = J_{31} = J_{23} = J_{32} = 0$, and

$$J_{11} = m\left(\frac{1}{\alpha^2} - \frac{1}{(\beta-\alpha)^2}\right) + z^{**},$$

$$J_{12} = J_{21} = \frac{m}{(\beta-\alpha)^2} - z^{**},$$

$$J_{22} = m\left(\frac{1}{\beta^2} - \frac{1}{(\beta-\alpha)^2}\right) + z^{**},$$

where

$$z^{**} = \sum_{i=1}^{m}\frac{n^2 z_i^2 e^{-n z_i(\beta-\alpha)}}{(1-e^{-n z_i(\beta-\alpha)})^2}.$$

The expectations of the entries of the observed information matrix cannot be obtained analytically. Therefore, the Fisher Information matrix $I(\boldsymbol{\tau}) = E(J(\boldsymbol{\tau}))$ can be obtained by using numerical methods. The MLE of R_{System} is asymptotically normal with mean R_{System} and asymptotic variance

$$\sigma^2_{R_{System}} = \sum_{j=1}^{3}\sum_{i=1}^{3} \frac{\partial R_{System}}{\partial \theta_i}\frac{\partial R_{System}}{\partial \theta_j} I^{-1}_{ij},$$

where I^{-1}_{ij} is the $(i,j)^{th}$ element of the inverse of $I(\boldsymbol{\tau})$, see Rao [9]. Afterwards,

$$\sigma^2_{R_{System}} = \left(\frac{\partial R_{System}}{\partial \alpha}\right)^2 I^{-1}_{11} + 2\frac{\partial R_{System}}{\partial \alpha}\frac{\partial R_{System}}{\partial \beta} I^{-1}_{12} + \left(\frac{\partial R_{System}}{\partial \beta}\right)^2 I^{-1}_{22}$$

Therefore, an asymptotic $100(1-\gamma)\%$ confidence interval of R_{System} is given by

$$R_{System} \in \left(\hat{R}_{System} \pm z_{\gamma/2}\hat{\sigma}_{R_{System}}\right),$$

where $z_{\gamma/2}$ is the upper $\gamma/2$th quantile of the standard normal distribution and $\hat{\sigma}_{R_{System}}$ is the value of $\sigma_{R_{System}}$ at the MLE of the parameters.

4.3 Bayesian Estimation of R_{System}

In this section, we assume that α, β and θ be random variables that have independent gamma prior distributions with parameters (a_i, b_i), $i = 1,2,3$, respectively. The joint posterior density function of α, β and θ

$$\pi(\alpha,\beta,\theta|\mathbf{z},\mathbf{t}) \propto \alpha^{m+a_1-1}\beta^{m+a_2-1}(\beta-\alpha)^{-m}\theta^{a_3+m-1}e^{-\alpha b_1-\beta b_2-\theta\left(b_3+\sum_{i=1}^{m} t_i\right)} e^{\left(\sum_{i=1}^{m}\ln\left(e^{-\alpha n z_i}-e^{-\beta n z_i}\right)\right)}. \quad (10)$$

The Bayes estimator of R_{System} using SE loss function is

$$\hat{R}^{Bayes}_{System} = \int_0^\infty \int_0^\infty \int_0^\infty R_{System}\pi(\alpha,\beta,\theta|\mathbf{z},\mathbf{t})d\alpha d\beta d\theta. \quad (11)$$

Since the integral in (11) can not be computed analytically, we need some approximation methods to obtain the Bayes estimator of R_{System}. We use MCMC method for this purpose.

4.3.1 MCMC Method

The joint posterior density function of α, β and θ is given in (10). Then, the marginal posterior density functions are given respectively as

$$\theta|\mathbf{t} \sim Gamma(m+a_3, b_3+\textstyle\sum_{i=1}^{m} t_i),$$

$$\pi(\alpha|\beta,\mathbf{z}) \propto \alpha^{m+a_1-1}(\beta-\alpha)^{-m}e^{-\alpha b_1} e^{\left(\sum_{i=1}^{m}\ln\left(e^{-\alpha n z_{ij}}-e^{-\beta n z_{ij}}\right)\right)},$$

and

$$\pi(\beta|\alpha,\mathbf{z}) \propto \beta^{m+a_2-1}(\beta-\alpha)^{-m}e^{-\beta b_2} e^{\left(\sum_{i=1}^{m}\ln\left(e^{-\alpha n z_{ij}}-e^{-\beta n z_{ij}}\right)\right)}.$$

Therefore, samples of θ can be easily generated by using a gamma distribution. Since the posterior distribution of α and β cannot be reduced analytically to a well-known distribution, it is not possible to sample directly by standard methods. We can obtain Bayes estimate and HPD credible interval of R_{System}using the hybrid Metropolis-Hastings and Gibbs sampling algorithm similar. Since this procedure is similar to in Section 3.3.1, it is omitted.

5. Simulation Study

In this section, numerical results are presented for the series systems when standby redundancy is applied at system or component levels. The performances of the estimates are compared by using mean square error (MSE) and estimated risk (ER). The ER of θ is given by $ER(\theta) = \frac{1}{N}\sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2$ under the SE loss function where $\hat{\theta}$ is estimate of θ. The average confidence lengths and coverage probabilities (cp) provide a comparison between the asymptotic confidence and HPD credible intervals. The coverage probability of a confidence interval is the proportion of time that the interval contains the parameter of interest. All results are based on 2500 replications.

The obtained results for the series system with component and system levels are listed in Tables 1 and 2, respectively. In the tables, sample sizes for the stress and strength components are chosen as 10(10)40 in every case. For the series system with component and system levels, the parameters are chosen as $(\alpha, \beta, \theta) = (6,2,3)$ when the strength samples are $n = 3,5$ and $n = 2,4$, respectively. Depending on these samples the ML and Bayesian estimates of the stress-strength reliability, their average MSE and ER are listed. Moreover, 95% asymptotic confidence and HPD credible intervals of the stress-strength reliability with the coverage probabilities (cps) are presented.

In the MCMC case, we run two MCMC chains with fairly different initial values and generate 5000 iterations for each chain. In order to reduce the effect of the starting distribution, the first 2500 results of each sequence are discarded. This process is called as burn-in. So as to cut off the dependence between the results in the Markov chain, using thinning, save every d^{th} draw of the chain. While applying MCMC method, we produce the estimates by use of every 5^{th} sampled values after the discarding procedure.

In Tables 1-2, the average MSE and ERs generally decrease for all estimates when the sample size increases. The average lengths of the intervals also decrease as the sample size increases.

From Table 1, it is observed that the Bayes estimates of R_{Comp} have generally smaller errors than that of ML estimates. Moreover, these errors are close to each other as the sample size increases. The average lengths of the HPD credible interval of R_{Comp} are generally smaller than those belonging to the asymptotic confidence intervals.

From Table 2, the Bayes estimates of R_{System} have smaller errors than that of ML estimates. Moreover, these errors are close to each other as the sample size increases. The average lengths of the HPD credible interval of R_{System} are generally smaller than those belonging to the asymptotic confidence intervals.

Table 1. Estimates and confidence intervals of R_{Comp}

$(\alpha, \beta, \theta) = (6,2,3)$						
n	m	R_{Comp}	$\hat{R}_{Comp}$	$\hat{R}_{Comp}^{MCMC}$	ACI of R_{Comp}	$HPDCI$ of R_{Comp}
	10	0.3944	0.4010	0.3684	(0.1411,0.6609)	(0.1685,0.5748)

			0.0180	0.0070	0.5198/0.9172	0.4063/0.9720
	20		0.3985	0.3739	(0.2071,0.5899)	(0.2113,0.5416)
			0.0094	0.0053	0.3827/0.9328	0.3303/0.9612
	30		0.3996	0.3785	(0.2411,0.5582)	(0.2361,0.5250)
			0.0067	0.0044	0.3171/0.9284	0.2889/0.9580
	40		0.3986	0.3814	(0.2602,0.5371)	(0.2523,0.5139)
			0.0050	0.0036	0.2769/0.9412	0.2616/0.9624
	10	0.2121	0.2343	0.2033	(0.0052,0.4634)	(0.0492,0.3778)
			0.0162	0.0048	0.4582/0.8928	0.3286/0.9556
	20		0.2252	0.2063	(0.0581,0.3923)	(0.0758,0.3504)
			0.0081	0.0038	0.3342/0.9200	0.2745/0.9524
	30		0.2202	0.2084	(0.0829,0.3576)	(0.0910,0.3359)
			0.0053	0.0032	0.2747/0.9344	0.2449/0.9584
	40		0.2188	0.2092	(0.0992,0.3385)	(0.0870,0.3344)
			0.0039	0.0026	0.2393/0.9384	0.2474/0.9684
$(\alpha, \beta, \theta) = (1.5,2.5,10)$						
	10	0.9485	0.9296	0.9285	(0.8426,1.0166)	(0.8549,0.9832)
			0.0022	0.0011	0.1740/0.9384	0.1284/0.9896
	20		0.9356	0.9341	(0.8776,0.9936)	(0.8805,0.9771)
			0.0010	0.0007	0.1160/0.9568	0.0966/0.9852
	30		0.9393	0.9375	(0.8942,0.9844)	(0.8937,0.9737)
			0.0006	0.0004	0.0902/0.9688	0.0800/0.9804
	40		0.9421	0.9402	(0.9047,0.9795)	(0.9028,0.9718)
			0.0004	0.0003	0.0748/0.9548	0.0689/0.9820
	10	0.8997	0.8758	0.8682	(0.7438,1.0078)	(0.7461,0.9621)

			0.0050	0.0030	0.2640/0.9240	0.2160/0.9888
	20		0.8863	0.8796	(0.7979,0.9748)	(0.7918,0.9517)
			0.0021	0.0017	0.1769/0.9496	0.1599/0.9812
	30		0.8895	0.8840	(0.8186,0.9604)	(0.8108,0.9456)
			0.0013	0.0011	0.1418/0.9520	0.1348/0.9804
	40		0.8927	0.8853	(0.8327,0.9527)	(0.8119,0.9446)
			0.0010	0.0011	0.1200/0.9444	0.1328/0.9792

Notes: The first row represents the average estimates and the second row represents corresponding MSE or ERs for the point estimates. The first row represents a %95 confidence interval and the second row represents their lengths and cps for the interval estimates.

Table 2. Estimates and confidence intervals of R_{System}

$(\alpha, \beta, \theta) = (6,2,3)$						
n	m	R_{System}	$\hat{R}_{System}$	$\hat{R}_{System}^{MCMC}$	ACI of R_{System}	$HPDCI$ of R_{System}
	10	0.4286	0.4345	0.4155	(0.2252,0.6438)	(0.2456,0.5938)
			0.0108	0.0044	0.4186/0.9316	0.3482/0.9852
	20		0.4301	0.4199	(0.2799,0.5803)	(0.2860,0.5580)
			0.0060	0.0035	0.3004/0.9360	0.2720/0.9688
	30		0.4288	0.4202	(0.3053,0.5522)	(0.3062,0.5374)
			0.0038	0.0026	0.2469/0.9464	0.2311/0.9748
	40		0.4282	0.4207	(0.3210,0.5354)	(0.3195,0.5246)
			0.0029	0.0021	0.2144/0.9456	0.2051/0.9652
	10	0.3007	0.3080	0.2946	(0.1315,0.4844)	(0.1554,0.4481)
			0.0078	0.0030	0.3530/0.9440	0.2927/0.9896
	20		0.3060	0.2993	(0.1800,0.4320)	(0.1888,0.4171)

			0.0041	0.0024	0.2521/0.9388	0.2283/0.9732
	30		0.3052	0.3002	(0.2021,0.4084)	(0.2057,0.3996)
			0.0027	0.0018	0.2063/0.9532	0.1939/0.9756
	40		0.3019	0.2979	(0.2130,0.3907)	(0.2144,0.3853)
			0.0020	0.0015	0.1776/0.9440	0.1709/0.9676
$(\alpha, \beta, \theta) = (1.5,2.5,10)$						
	10	0.9231	0.8980	0.8977	(0.7832,1.0129)	(0.8037,0.9738)
			0.0035	0.0018	0.2297/0.9492	0.1701/0.9892
	20		0.9035	0.9035	(0.8225,0.9845)	(0.8328,0.9642)
			0.0019	0.0012	0.1620/0.9616	0.1314/0.9852
	30		0.9098	0.9091	(0.8460,0.9735)	(0.8510,0.9603)
			0.0012	0.0008	0.1274/0.9536	0.1093/0.9812
	40		0.9109	0.9107	(0.8559,0.9658)	(0.8593,0.9564)
			0.0009	0.0006	0.1099/0.9584	0.0971/0.9748
	10	0.8125	0.7843	0.7805	(0.6059,0.9627)	(0.6284,0.9172)
			0.0083	0.0044	0.3568/0.9436	0.2887/0.9856
	20		0.7920	0.7895	(0.6638,0.9202)	(0.6719,0.8978)
			0.0043	0.0028	0.2564/0.9452	0.2259/0.9792
	30		0.7934	0.7923	(0.6881,0.8987)	(0.6921,0.8861)
			0.0032	0.0022	0.2106/0.9424	0.1940/0.9748
	40		0.7978	0.7963	(0.7070,0.8887)	(0.7077,0.8797)
			0.0023	0.0017	0.1817/0.9420	0.1720/0.9716

Notes: The first row represents the average estimates and the second row represents corresponding MSE or ERs for the point estimates. The first row represents a %95 confidence interval and the second row represents their lengths and cps for the interval estimates.

6. Conclusions

In this paper, we have considered ML and Bayesian estimations of series system with cold standby redundancy at system and component levels while components which belong to exponential distribution, exposed to a common random stress that follows exponential distribution.

References

[1] Basirat, M. Baratpour, S. and Ahmadi, J. (2015). Statistical inferences for stress-strength in the proportional hazard models based on progressive Type-II censored samples. *Journal of Statistical Computation and Simulation*, 85(3), 431–449.

[2] Birnbaum, Z.W. (1956). On a use of Mann-Whitney statistics, in *Proc. 3rd Berkeley Symposium on Mathematical Statistics and Probability*, 1, 13–17.

[3] Birnbaum, Z.W. and McCarty, B.C. (1958). A distribution-free upper confidence bounds for $Pr(Y < X)$ based on independent samples of X and Y. *The Annals of Mathematical Statistics*, 29(2), 558–562.

[4] Boland, P.J. and El-Neweihi, E. (1995). Component redundancy vs system redundancy in the hazard rate ordering. *IEEE Transactions on Reliability*, 44(4), 614–619.

[5] Çetinkaya, Ç. and Genç, A.İ. (2019). Stress-strength reliability estimation under the standard two-sided power distribution. *Applied Mathematical Modelling*, 65, 72–88.

[6] Chen, M.H. and Shao, Q.M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1), 69–92.

[7] Kızılaslan, F. (2018). Classical and Bayesian estimation of reliability in a multicomponent stress-strength model based on a general class of inverse exponentiated distributions. *Statistical Papers*, 59(3), 1161–1192.

[8] Kumar, C.S. (1995). Standby redundancy at system and component levels-A comparison. *Microelectron Reliability*, 35(4), 751–752.

[9] Rao, C.R. (1965). Linear statistical inference and its applications, Wiley, New York.

[10] Rasethuntsa, T.R. and Nadar, M. (2018). Stress-strength reliability of a non-identical-component-strengths system based on upper record values from the family of Kumaraswamy generalized distributions. *Statistics*, 52(3), 684–716.

[11] Shen, K. and Xie, M. (1991). The effectiveness of adding standby redundancy at system and component levels. *IEEE Transactions on Reliability*, 40(1), 53–55.

[12] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4), 1701–1728.

[13] Zhao, P. Zhang, Y. and Li, L. (2015). Redundancy allocation at component level versus system level. *European Journal of Operational Research*, 241(2), 402–411.

A StarCraft 2 Player Skill Modeling

Zoran Ćirović
School of Electrical and Computer Engineering of Applied Studies
zoran.cirovic@viser.edu.rs

Nataša Ćirović
School of Electrical Engineering
University of Belgrade
natasa@etf.bg.ac.rs

Abstract

With increased popularity and monetization potential of e-sports, the player skill modeling in strategy games has long been a subject of interest for researchers. In this paper we use publicly available video game telemetry data from StarCraft 2 to determine the characteristics of players that can best describe their skill level. We create player skills models by selecting relatively small number of statistical features with SFS algorithm, that have higher importance to discriminate between skill levels, for several different classifiers with acceptable accuracy. The analysis of importance of selected features to discriminate between levels is conducted. By applying ReliefF algorithm for ranking we can obtain significance of features for separation of skill levels, one vs. one. Also, we compare distribution for one level for specific feature with distribution for all other levels, one vs. all, by applying KL divergence.

Keywords: Player skill modeling; Feature selection; Feature ranking

1. Introduction

With the ever-increasing popularity of e-sports and having in mind that the potential of further monetization in the next few years can reach up to $3 billion [1], the ability to predict the parameters that can establish the level ofacquired skills in gaming is increasingly important. The significance of player modeling for various types of games and prediction that can establish the level of acquired skills is highly important for the gain in user experience and interaction, [1], [2].

For real-time strategy (RTS) games the motor skills with a keyboard and mouse are an integral component of the game, since the players do not play in turn. One such game is StarCraft 2, that supports semi-professional and professional players, [3]. StarCraft 2 is a fast-paced strategy game where a player controls many soldiers, vehicles and alien creatures in order to defeat the opposing team's army. The game is competitive and assigns players a quantified skill level.

Great popularity, as well as profitability, of gaming yields the increased research of the player modeling, [2], [4], [5]. Some of the research directions belong to the cognitive science [4], [6]. Others are in the domain of machine learning, [7], [8], [9]. Chen et al. [10] researched classification applying various clustering techniques and dimension reduction, over the dataset in [4].

With the increased amount of information available today, the problem of handling big datasets appears. This is done by using feature selection techniques, [11], [12], in order to obtain a feature subset that can accurately discriminate between two or more classes.

In our research we use the video game telemetry data from StarCraft 2 from the publicly available dataset collected by Thompson et al. [4]. The goal of our research

is to model the player skills by selecting relatively small number of features that have higher importance to discriminate between skill levels, for several different classifiers with acceptable accuracy.

2. Feature selection

Feature selection process is used to reduce the dimensionality of feature space, remove redundant, irrelevant or noisy data. Some of the benefits for the applications are simplification and more efficient usage of the model, both in long-term training and in testing or exploitation phase, emphasizing generalization of features, increasing robustness, reducing the influence of excessive overfitting, [11], [12].

The straightforward algorithm is to test all possible subsets of features thus determining the one that gives the best results for the classifier. This is the process of complete search, which is not feasible in practice and it can be specific to the used classifier. The alternative is to use a combination of some feature selection techniques, i.e. determining a subset of the complete set of features by applying a metric that defines the quality of the selected subset. There are several classifications of feature selection methods. Based on the number of variables considered:

- *Univariate methods,* variable ranking - considers the input variables one by one;
- *Multivariate methods,* variable subset selection - considers entire groups of variables jointly (e.g. greedy search algorithms).

Based on the usage of the classification model in the feature selection process:

- *Filter:* evaluate quality of selected features, independent from the classification algorithm that will use them;
- *Wrapper:* require application of a classifier, which should be trained on a given feature subset, to evaluate the quality;
- *Embedded methods:* the feature selection method is built in the model itself.

2.1. Feature Selection Based on Ranking

Feature selection based on ranking methods can be used as a method for individual independent feature selection, based on a chosen criterion. A rank is assigned to each feature and the ones with the best rank are selected. Some of the criterion that can be used for ranking are T-test, Bhattacharyya, Kullback-Leibler - KL divergence. KL divergence, also known as relative entropy, filters out those features whose expression distributions are random. It is used to calculate how much information is lost when one distribution is approximated with another [13].

Compared to the other ranking methods, Relief based algorithms include interaction among features and may capture local dependencies which other methods miss. These algorithms use the concept of nearest neighbors to derive feature statistics that account for interactions. Relief calculates a proxy statistic for each feature that can be used to estimate feature ‘quality’ or ‘relevance’ to the target concept, i.e. feature weights. One interpretation of the Relief weight estimate W[A] of feature A is approximation of the following difference of probabilities, [14], [15]:

W[A] = P(different value of A | nearest instance from different class) – P(different value of A | nearest instance from same class)

The ReliefF variant of the algorithm finds the weights of features in multiclass problems, penalizing the features that give different values to neighbors of the same class, and rewards the ones that give different values to neighbors of different classes. ReliefF handles the multi class problems by selecting equal number of instances from all different classes and normalizes their contribution with their prior probabilities. In this way the algorithm is able to estimate the ability of features to separate each pair of classes regardless of which two classes are closest to each other.

2.2. Sequential feature selection search By ranking the significance of individual features can be determined, but not the significance of a subset of features, thus neglecting to consider the correlation that unavoidably exists among the various features and influences the classification capabilities.

Sequential feature selection search algorithms are a family of greedy search algorithms, which are of multivariate type. Sequential forward selection – SFS is iterative and starts with an empty set. In each iterative step the best feature is chosen based on the applied criterion. The process is repeated as many times as the number of features we want to select or get appropriate accuracy. Alternatively, we can start with the full set, and iteratively throw out the worst features (Sequential Backward Selection - SBS). Another method is Sequential Floating Forward Search – SFFS which uses both directions of search. Class separability measures are used as criterion for measuring the discrimination effectiveness of feature vectors in SFS algorithms. Some of the measures that can be used in feature selection procedures are divergence, scatter matrices and Bhattacharyya distance [13]. We implement SFS using the Bhattacharyya distance.

3. Data Set

For obtaining the experimental results the dataset was collected by Thompson et al. [4], from 3,340 players of RTS video game - StarCraft 2, across 7 distinct levels of skill called leagues (Bronze, Silver, Gold, Platinum, Diamond, Masters, Professional). Every record contains 18 different features. The data is gathered by remote measurement without affecting the players in any way. The features are [4]:

1. Age: Age of each player (integer);
2. HoursPerWeek: Reported hours spent playing per week (integer);
3. TotalHours: Reported total hours spent playing (integer);
4. APM: Action per minute (continuous);
5. SelectByHotkeys: Number of unit or building selections made using hotkeys per timestamp (continuous);
6. AssignToHotkeys: Number of units or buildings assigned to hotkeys per timestamp (continuous);
7. UniqueHotkeys: Number of unique hotkeys used per timestamp (continuous);
8. MinimapAttacks: Number of attack actions on minimap per timestamp (continuous);
9. MinimapRightClicks: number of right-clicks on minimap per timestamp (continuous);
10. NumberOfPACs: Number of PACs per timestamp (continuous);
11. GapBetweenPACs: Mean duration in milliseconds between PACs (continuous);
12. ActionLatency: Mean latency from the onset of a PACs to their first action in milliseconds (continuous);
13. ActionsInPAC: Mean number of actions within each PAC (continuous);
14. TotalMapExplored: The number of 24x24 game coordinate grids viewed by the player per timestamp (continuous);
15. WorkersMade: Number of SCVs, drones, and probes trained per timestamp (continuous)

16. UniqueUnitsMade: Unique unites made per timestamp (continuous);

17. ComplexUnitsMade: Number of ghosts, infestors, and high templars trained per timestamp (continuous);

18. ComplexAbilitiesUsed: Abilities requiring specific targeting instructions used per timestamp (continuous).

further processing. In order to achieve equal contribution of all features in classification, normalization of input vectors should be implemented [13]. If the expected feature distribution is standard Gaussian, and the variance is not too small, standardization should be implemented instead. If statistical data is formed in the preprocessing phase, then framing of input vectors is performed. In framing overlapping is conducted, in order to include the sudden changes in data. Next, feature selection is performed. Lastly, four different type classifiers are created, where the selected features are used in two separate phases: (i) in the training phase – for modeling activities, and (ii) in the testing phase. Since the dataset is not big enough, the procedure of cross validation is applied.

Analysis of the selected features is conducted in the second part of our research, which includes: (i) importance of the selected features regarding their discriminative properties, one vs. one, (ii) probabilistic distribution functions of the selected features are analyzed, one vs. all.

4.1. Statistical Features

A collection of statistical functions used for creation of statistical features is shown in Table 1 [16], [17].

Table 1. Set of analyzed features

	Description
μ	Mean value
σ	Standard deviation
min	Minimum value
max	Maximum value
iqr	Interquartile Range
sk	Skewness
k	Kurtosis

5. Results

We created several ML models for Video Game Player Skill Modeling based on the available dataset. Investigated models are based on the following classifiers: k-NN, decision tree, SVM and QDA. For all classifiers and all experiments 3-fold cross-validation was used.

Specific versions of classifiers that we used in our experiments are:

- k-NN classifier for = 3 (for other values of k the results did not vary significantly);
- Tree with the maximum number of branches per node being 7 (for other values the results did not vary significantly);
- multiclass SVM is based on (7.6)/2 =21 binary SVM one-vs-one models, for = 7 being the number of classes, and the kernel function is linear.

QDA Quadratic discriminant analysis uses covariance matrix, assuming different mean values and variance for each class, thus the full covariance matrix and vector of mean values are used.

5.1. Modeling Original Features

Firstly, we used the 18 original input features, after normalization, from the dataset. The error rate for original features for specific classifiers is shown in Table 2.

Table 2. Error rates for original features

	k-NN	Tree	SVM	QDA
Err [%]	68.23	63.42	58.58	66.17

Obviously, the error rate for model based on original features was not satisfactory.

5.2. Modeling Statistical Features

Statistical data processing is realized by grouping normalized input data in frames of length 30, with overlapping between frames. We used statistical functions shown in Table 1 for creation of statistical features. Thus, we got total of 18·7 = 126 statistical features. Next, we investigated 7 statistical groups separately and jointly with the same classifiers. The error rates are shown in Table 3.

Table 3. Error rates for statistical features

Err [%]	k-NN	Tree	SVM	QDA
μ	8.34	9.41	2.06	6.05
σ	24.45	40.3	8.58	25.71
min	7.1	25.47	10.47	42.57
max	5.78	22.82	9.48	25.30
iqr	30.19	36.37	9.65	15.42
sk	4.23	55.45	34.61	39.64
k	11.03	59.29	40.43	49.20
all	5.98	8.75	1.93	5.60

Obviously, the selection of optimal features depends on the applied model and significant features can be found in different statistical groups, as expected. Thus, we will select a subset of features with a sequential feature selection search algorithm.

5.3. Modeling with SFS

Using Sequential Forward Selection – SFS with Bhattacharyya distance as class separability measure, we get subsets of M features for each skill level. Final set of selected features, N, is a union of selected features for each level. Results for different M are shown in Table 4.

Table 4. Error rates classification for the best N features applying SFS

	Err [%]			
M; N	k-NN	QDA	Tree	SVM
3;6	4.1	32.55	14.53	9.13
5; 10	3.31	25.56	14.21	5.61
7; 14	3.35	17.57	13.27	2.61
10;19	3.8	12.87	13.37	1.19
15;25	3.06	6.37	8.26	0.23

These results show significant increase in accuracy for specific classifiers. We choose the subset of N = 19 features, (M = 10), for further analysis for the player skills modeling. Smaller number of selected features would result in smaller accuracy, and larger number could increase the complexity of the player model. The selected features are shown in the Table 5.

Table 5. Subset of 19 selected features

Statistical function	Feature description
μ	Age, ActionPerMinute, SelectByHotkeys, AssignToHotkeys, UniqueHotkeys, MinimapAttacks, MinimapRightClicks, NumberOfPACs, GapBetweenPAC

σ	UniqueHotkeys
min	Age, AssignToHotkeys, MinimapAttacks
max	HoursPerWeek, TotalHours, UniqueHotkeys, NumberOfPACs
sk	HoursPerWeek
k	TotalHours

5. 4. Feature Discriminative Importance For Two Levels

Feature importance for player skill model must take into account certain dependency between the features and the multiclass nature of the model. By applying the ReliefF algorithm we have determined the feature discriminative importance. The same feature can be highly important to discriminate between two levels, but not so important to discriminate between other two levels.

In order to illustrate this, in Figure 1 we show feature importance for discrimination between Level 1 and 2 (lower skills), between Level 6 and 7 (higher skills) and between Level 2 and 6 (lower and higher skills), based on the rank position of the ReliefF algorithm, applied to the two observed levels.

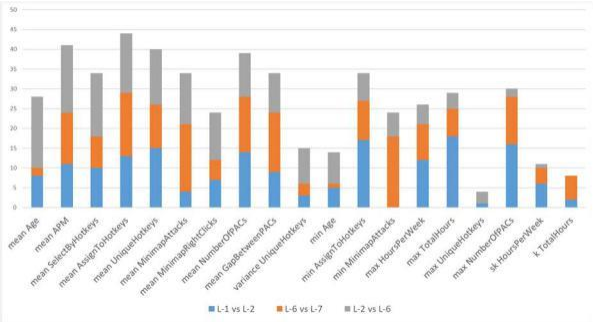

Figure 1. Comparative representation of discriminative importance of features between levels based on ReliefF

Based on similar diagrams we can interpret features regarding their discriminative importance between specific skill levels. For example, features related to Age, i.e. mean Age and min Age, are more important in discriminating between lower levels and between lower and higher skill levels but are less important to discriminate between higher levels. On the other hand, feature such as mean UniqueHotkeys is important for all 3 observed discriminations.

5.5. Feature Discriminative Importance for All Levels

With the aim to visualize and perceive the selected features, we present distribution functions for three features. For every feature we show 7 distributions, one for each skill level. Every graphic shows the distribution of the feature for the observed level (in blue), compared to the distribution of the feature for all other 6 levels (in red).

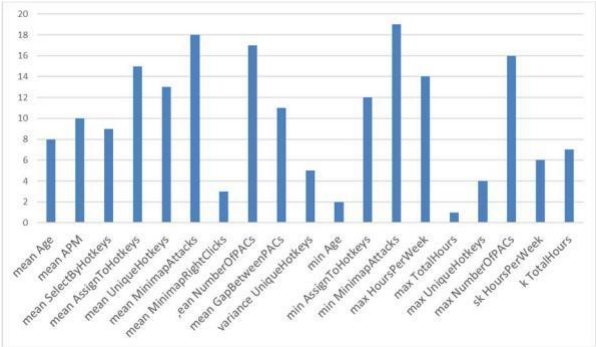

Figure 2. Discriminative importance of features for all levels based on ReliefF

5.6. Feature Distribution Analysis

With the aim to visualize and perceive the selected features, we present distribution functions for three features. For every feature we show 7 distributions, one for each skill level. Every graphic shows the distribution of the feature for the observed level (in blue), compared to the distribution of the feature for all other 6 levels (in red).

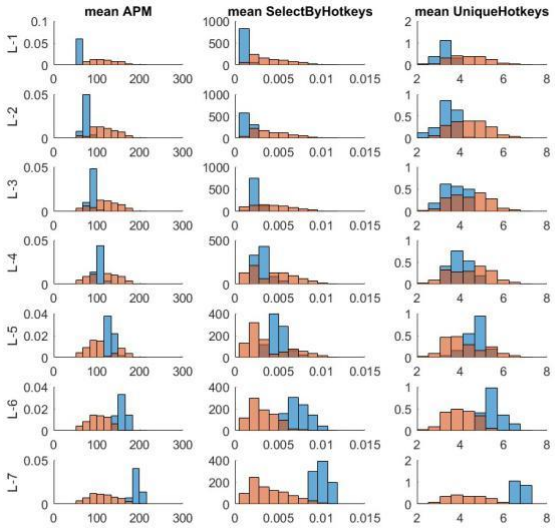

Figure 3. Distributions for the observed level compared to the distribution for all other levels, for three features

Obviously, there are certain differences but also certain similarities when comparing the two distribution on the same figure. In order to compare two distributions, we analyze the discriminative importance of features for specific skill levels by applying KL divergence. In this way we can measure how different are the compared distributions. Figure 4 shows the values of KL divergence for 3 features, for each skill level.

6. Conclusions

We investigated dataset gathered from players of one RTS game, to determine the characteristics of players that can best describe their skill level. We created four different classifiers: k-NN, QDA, Tree, SVM.

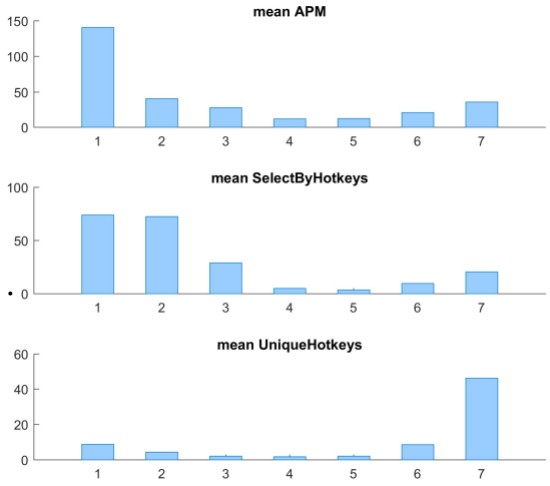

Figure 4. KL divergence for three features across all levels

We started the research with the original dataset. Achieved accuracy was not satisfactory, Table 2. Next, we created a new feature set by introducing 7 statistical functions. We conducted experiments for each statistical group of features and all features jointly, getting higher accuracy, Table 3. Based on the achieved results we concluded that there is a need to implement a sequential feature selection algorithm. By applying the SFS algorithm for every level, we created a union of 10 best features for each level, obtaining the set of 19 different features. Achieved results show satisfactory accuracy, Table 4.

Additionally, analysis of selected features was conducted regarding the importance of features to discriminate between levels. It is shown that by applying the ReliefF algorithm for ranking, feature significance for separation of skill levels, one vs. one, can be obtained. We created probability distribution function for specific feature for each level and compared it to the distribution of the feature for the other 6 levels, Figure 1. Also, we applied KL divergence to get the difference of the two distributions for each level, one vs. all.

In this way it was shown that by applying classifier player skills modeling can be performed with good enough quality. The comparative statistical analysis of the features, one vs. one and one vs. all, provides additional information about the features.

The procedure of feature selection is efficient, considers the nature of the features and it can be implemented to similar applications, e.g. skill acquirement.

References

[1] Merwin, C. D., Sugiyama, M., Mubayi, P., Hari, T., Terry, H.P., Duval, A., "The World of Games eSports From Wild West to Mainstream," The Goldman Sachs Groups, Inc., New York, NY, USA, Rep. Oct., 2018.

[2] Bakkes S.C.J., Spronck, P.H.M., van Lankveld, G., "Player behavioral modelling for video games", *Entertainment Comp.*, vol. 3, 2012, pp. 71-79.

[3] Adil, K., Jiang, F., Liu, S., Jifara, W., Tian, Z., Fu, Y., "State-of-the-art and open challenges in RTS game-AI and Starcraft", *Int. J. Adv. Comp. Sci. Appl.*, vol. 8, no. 12, 2017, pp.16-24.

[4] Thompson, J.J., Blair, M.R., Chen, L., Henrey, A.J., "Video Game Telemetry as a Critical Tool in the Study of Complex Skill Learning", *PLOS ONE*, vol. 8, no. 9, Sep. 2013, e75129.

[5] Yannakakis, G.N., Togelius, J., "Artifcial Intelligence and Games", Berlin, Germany, Springer, 2018.

[6] Thompson, J.J., McColeman, C.M., Stepanova, E.R., Blair, M.R., "Using Video Game Telemetry Data to Research Motor Chunking, Action Latencies, and Complex Cognitive- Motor Skill Learning", *Top. Cogn. Sci*., vol. 9, 2017, pp. 467-484.

[7] Ravari, Y.N., Bakkes, S., Spronck, P., "StarCraft Winner Prediction", Proc. 12th AAAI Conf. Artif. Intell. Interact. Dig. Entert., AIIDE 2016, Burlingame, California, USA, pp. 2-8, Oct. 8-12, 2016.

[8] Avontuur, T., Spronck, P., Van Zaanen, M., "Player skill modeling in starcraft II," Proc. 9th AAAI Conf. Artif. Intell. Interact. Dig. Entert., AIIDE 2013, Boston, USA, pp. 2-8, Oct. 14-18, 2013.

[9] Aung, M., Bonometti, V., Drachen, A., Cowling, P., Kokkinakis A.V., Yoder C., Wade A., "Predicting Skill Learning in a Large, Longitudinal MOBA Dataset", 2018 *IEEE Conf. Comp*. Intell. Games CIG, Maastricht, 2018, pp. 1-7

[10] Chen, P., Qi, Z., Pan, Y., Cheng, S., "Multivariate and Categorical Analysis of Gaming Statistics," *Proc. 18th Int. Conf. Network-Based Infor. Syst.* Taipei, Taiwan, pp. 286-293, Sep. 2-4, 2015.

[11] Guyon, I., Elisseeff, A., "An Introduction to Variable and Feature Selection", *J. Mach. Learn. Res.*, vol. 3, 2003, pp 1157-1182.

[12] G. Chandrashekar, F. Sahin, "A survey on feature selection methods", *Comp. El. Eng*, vol. 40, no. 1, 2014, pp. 16-28.

[13] Theodoridis S., Koutroumbas K., "Pattern Recognition", *4th ed., Academic Press*, 2009.

[14] Robnik-Šikonja M., Kononenko I., "Comprehensible interpretation of relief's estimates", Machine Learning: Proc. XVIII Int. Conf. on Mach. Lear., Williamstown, MA, USA, San Francisco, 2001, pp. 433-440.

[15] Kononenko I., Šimec E., Robnik-Šikonja R., "Overcoming the myopia of inductive learning algorithms with relieff", *Appl. Intell.* Vol. 7(1), 1997, pp. 39-55.

[16] B. Esmael, A. Arnaout, R. K. Fruhwirth, G. Thonhauser, "A Statistical Feature-based Approach for Operations Recognition in Drilling Time", *Series. Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 5, pp. 454-461, 2015.

[17] J. P. Verma, A-S. G. Abdel-Salam, Testing Statistical Assumptions in Research, Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2019.

A Seemingly Unrelated Regression Modeling for Extraction Process in Green Chemistry

Özlem Türkşen
Ankara University
Faculty of Science
Statistics Dept.
Ankara, Turkey
turksen@ankara.edu.tr

Serhan Tunçel
Ankara University
Faculty of Science
Statistics Dept.
Ankara, Turkey
serhantuncel3@gmail.com

Nilüfer Vural
Ankara University
Faculty of Engineering
Chemical Engineering Dept.
Ankara, Turkey
nvural@science.ankara.edu.tr

Abstract

The crucial step in multi-response experiment is to model the responses simultaneously with considering correlation structure of the responses. The ordinary least squares (OLS) is the main approach where there is no dependency among the response variables. However, some of the multi-response experiments may have correlated responses. In this case, seemingly unrelated regression (SUR) analysis should be used to model the correlated responses in which the correlation of responses is included through covariance matrix.

In this study, a multi-response experimental data set was used to apply the SUR modeling for extraction process of grape seeds phenolic compounds in green chemistry. The main aim of the study is obtaining predicted response functions with considering the correlation structure between them. The analysis was done by using R programming. It was seen from the analysis results that the SUR produces more precise model parameter estimates then the OLS when responses were correlated.

Keywords: Seemingly unrelated regression (SUR); correlated responses; extraction process

1. Introduction

Most of the experiments in extraction process have multiple responses in order to model the experimental data set, the experimenter fits a regression model, e.g. second order polynomial model, to each response by using ordinary least squares (OLS). The detailed information about modeling with OLS method can be seen in the studies of [1] and [2]. However sometimes, it is inevitable that multi-responses can be correlated. When the response variables are correlated in a multi-response problem, seemingly unrelated regression (SUR) can be very useful. The SUR model, originally developed by [3], is a set of different regression models having correlations with each other. The SUR has been used for various fields such as engineering, econometrics, biometrics, and statistical quality control. There have been some studies about the SUR application for modeling of the correlated responses in multi-response experiments, e.g. [4], [5], [6] and [7]. Besides different approaches for estimation of covariance structure can be seen in the studies of [8], [9], [10] and [11].

In this study, a multi-response data set about grape seeds phenolic compounds extraction

was used to apply the SUR modeling. The data was composed with four input variables, EtOH concentration (X_1), extraction time (X_2), solvent: solid ratio (X_3) and extraction temperature (X_4) and two responses, total phenolic content-TPC (Y_1) and total antioxidant activity-TAA (Y_2) [12]. The main aim of the study is obtaining predicted functions of Y_1 and Y_2 with considering the correlation structure between them.

The paper was organized as follows. Section 2 contains a brief description about the SUR modeling for a multi-response experimental data set. In Section 3, a real data set about grape seeds phenolic compounds extraction was used to apply the SUR modeling with comparison results. Finally, conclusion was given in Section 4.

2. Seemingly Unrelated Regression

The basic seemingly unrelated regression (SUR) model assumes that there are r independent variables $Y_{i1}, Y_{i2}, ..., Y_{ir}$ available, for each individual observation i, as below

$$Y_{ij} = \sum_{t=1}^{k_j} X_{ijt}\beta_j + \varepsilon_{ij} \qquad (1)$$

where Y_{ij} is the ith observation on the jth response variable which is to be explained by the ith response function, X_{ijt} is the ith observation on tth input variable appearing in the jth response function, β_{ij} is the coefficient associated with X_{ijt} at each observation and ε_{ij} is the ith value of the random error component associated with jth response function, $i = 1, 2, ..., n$; $j = 1, 2, ..., r$; $t = 1, 2, ..., k_j$. These r different linear regression models with cross-correlation can be defined as

$$\underset{\sim}{Y}_j = \underset{\sim}{X}_j \underset{\sim}{\beta}_j + \underset{\sim}{\varepsilon}_j, \; j = 1, 2, ..., r \qquad (2)$$

in which $\underset{\sim}{Y}_j : n \times 1$, $\underset{\sim}{X}_j : n \times k_j$ is of rank k_j, $\beta_j : k_j \times 1$ and $\underset{\sim}{\varepsilon}_j = n \times 1$. For these r models, there are assumptions [13]:

(i) $E(\underset{\sim}{\varepsilon}_j) = 0$

(ii) $Var(\underset{\sim}{\varepsilon}_j) = \sigma_{jj} I_n$

(iii) $Cov(\underset{\sim}{\varepsilon}_j, \underset{\sim}{\varepsilon}_l) = \sigma_{jl} I_n = \sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}} r_{jl} I_n \quad j, l = 1, 2, ..., r, \; j \neq l$

The SUR model, given Equation (2) can be expressed in a compact from as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (3)$$

where $\underset{\sim}{Y} = [\underset{\sim}{Y}_1 : \underset{\sim}{Y}_2 : ... : \underset{\sim}{Y}_r']'$, $\underset{\sim}{\beta} = [\underset{\sim}{\beta}_1' : \underset{\sim}{\beta}_2' : .... : \underset{\sim}{\beta}_r']'$, $\underset{\sim}{\varepsilon} = [\underset{\sim}{\varepsilon}_1' : \underset{\sim}{\varepsilon}_2' : .... : \underset{\sim}{\varepsilon}_r']'$ and X is the black-diagonal matrix, $diag(\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_r)$. Compactly, the assumptions can be written as

(i) $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}_s$

(ii) $Var(\underset{\sim}{\varepsilon}) = \psi = \sum \otimes I_p$

where $\otimes$ donates Kronecker product operator, $\sum = (\sigma_{ij}) = \sum_d \Lambda \sum_d : r \times r$ positive definite matrix, $\sum_d = diag(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, ..., \sqrt{\sigma_{pp}}) : r \times r$ diagonal matrix with positive diagonal elements, $\Lambda = (r_{ij}) : r \times r$ positive definite matrix, and $s = nr$.

The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}'(\sum^{-1} \otimes \boldsymbol{I_n})\boldsymbol{X}]^{-1}\boldsymbol{X}'(\sum^{-1} \otimes \boldsymbol{I_n})\boldsymbol{Y} \qquad (4)$$

In this study, covariance matrix estimator ($\hat{\sum} = \hat{\sum}_d \hat{\Lambda} \hat{\sum}_d$) is used to estimate the model parameters since the $\sum$ is unknown. To apply the SUR modeling, individual responses are initially modeled by using the OLS approach.

Then, the SUR is applied with considering the covariance matrix of the OLS residuals as the estimate of covariance matrix given in Equation (4).

3. Application

In this section, a real data set, from study of [12], was used to illustrate the SUR modeling procedure. The multi-response experimental data set was about the ultrasound assisted extraction of polyphenolics from grape seed and composed with four input variables, EtOH concentration (X_1), extraction time (X_2), solvent: solid ratio (X_3) and extraction temperature (X_4) and two responses, total phenolic content-TPC (Y_1) and total antioxidant activity-TAA (Y_2) [12]. The main aim of the study is obtaining predicted functions of Y_1 and Y_2 with considering the correlation structure between them.

The experiment was conducted in a central composite design (CCD) with 30 runs and given in Table 1.

Table 1. Experimental data set for extraction process [12].

No	X_1(%)	X_2(min)	X_3(mL/g)	X_4(°C)	Y_1(TPC)	Y_2(TAA)
1	-1	-1	-1	-1	3.67	28.18
2	+1	-1	-1	-1	8.10	53.11
3	-1	+1	-1	-1	9.53	55.66
4	+1	+1	-1	-1	10.99	58.70
5	-1	-1	+1	-1	10.62	51.82
6	+1	-1	+1	-1	17.31	75.94
7	-1	+1	+1	-1	8.75	47.88
8	+1	+1	+1	-1	21.50	88.97
9	-1	-1	-1	+1	10.91	56.20
10	+1	-1	-1	+1	13.15	67.41
11	-1	+1	-1	+1	5.03	36.87
12	+1	+1	-1	+1	13.27	67.28
13	-1	-1	+1	+1	16.44	55.82
14	+1	-1	+1	+1	22.37	82.33
15	-1	+1	+1	+1	9.44	57.73
16	+1	+1	+1	+1	32.43	95.83
17	-2	0	0	0	5.13	38.82
18	+2	0	0	0	12.71	66.29
19	0	-2	0	0	5.12	15.32
20	0	+2	0	0	14.35	73.82
21	0	0	-2	0	8.82	46.84
22	0	0	+2	0	20.69	80.00
23	0	0	0	-2	9.05	54.94
24	0	0	0	+2	24.86	89.28
25	0	0	0	0	17.91	79.90
26	0	0	0	0	17.51	79.16
27	0	0	0	0	10.91	63.85
28	0	0	0	0	11.27	65.56
29	0	0	0	0	17.46	78.32
30	0	0	0	0	18.67	82.85

It should be noted here that the input variables were given with coded levels in Table 1. The coded levels of the input variables and corresponding real values were presented in Table 2.

Table 2. The coded levels and real values of input variables

	Coded Levels				
Input Variables	-2	-1	0	1	2
X_1:EtOH concentration (%)	0	25	50	75	100
X_2:Extraction time (min)	0	10	20	30	40
X_3:Solvent/solid ratio (mL/g)	4.5	13	21.5	30	38.5
X_4:Extraction temperature (°C)	20	30	40	50	60

According to the multi-response experimental data set, given in Table 1, the correlation between the Y_1 and Y_2 are calculated and obtained as 0.9062. So, the predicted response function should be obtained with considering the correlation of Y_1 and Y_2.

In order to obtain the SUR model of the data set, given in Table 1, firstly the OLS parameter estimates for the responses, Y_1 and Y_2, were calculated. The covariance matrix for residuals of the OLS predicted response model was computed and considered as the estimate of covariance matrix, $\hat{\Sigma}$, given as

$$\hat{\Sigma} = \begin{bmatrix} 11.64530 & 20.18903 \\ 20.18903 & 88.00718 \end{bmatrix}$$

Then, the SUR parameter estimates were obtained according to the Equation (4). The parameter estimates for responses, Y_1 and Y_2, using the OLS and the SUR were reproduced in Table 3 and Table 4, respectively.

Table 3. Parameter estimates (Standard error) of Y_1 using the OLS and the SUR

Parameters	OLS		SUR	
	Parameter Estimates (St. error)	Pr(>|t|)	Parameter Estimates (St. error)	Pr(>|t|)
intercept	14.7519 (0.9012)	0.0001	14.7519 (0.8043)	0.0001
X_1	3.3288 (0.7822)	0.0002	3.3288 (0.6966)	0.0001
X_2	1.1179 (0.7822)	0.1664	1.1179 (0.6966)	0.1222
X_3	3.6646 (0.7822)	0.0001	3.6646 (0.6966)	0.0001
X_4	2.6746 (0.7822)	0.0023	2.6746 (0.6966)	0.0008
X_1^2	-1.4412 (0.7140)	0.0554	-1.4412 (0.6359)	0.0331
$X_1 X_3$	1.9994 (0.9580)	0.0482	1.0930 (0.6395)	0.1009

Table 4. Parameter estimates (Standard error) of Y_2 using the OLS and the SUR

Parameters	OLS		SUR	
	Parameter Estimates (St. error)	Pr(>|t|)	Parameter Estimates (St. error)	Pr(>|t|)
intercept	72.3756 (3.0409)	0.0001	70.1241 (2.5026)	0.0001
X_1	10.5979 (2.1503)	0.0001	10.5979 (1.9149)	0.0001
X_2	6.4629 (2.1503)	0.0063	6.4629 (1.9149)	0.0026
X_3	8.3013 (2.1503)	0.0007	8.3013 (1.9149)	0.0002
X_4	5.3288 (2.1503)	0.0210	5.3288 (1.9149)	0.0106
X_1^2	-4.7641 (1.9751)	0.0242	-4.4827 (1.7542)	0.0177
X_2^2	-6.7604 (1.9751)	0.0023	-4.2274 (1.3186)	0.0039

It can be seen from Tables 3-4 that the SUR parameter are more precise (smaller std. error) than those obtained with the OLS. Therefore, the SUR estimates are more efficient in the sense that they have smaller variances. This will lead to a more precise estimate of the optimum operating conditions on the extraction process of grape seeds phenolic compounds.

In addition, the predicted response models were compared according to the model performance metrics, e.g. coefficient of determination (R^2), mean absolute percentage error (*MAPE*), root mean square error (*RMSE*) and mean absolute error (*MAE*). The results are presented in Table 5.

Table 5. Performance metrics of the predicted responses according to the OLS and the SUR

Performance Metrics	Modeling Method	Predicted Responses	
		$\hat{Y}_1$	$\hat{Y}_2$
R^2	OLS	0.7272	0.7519
	SUR	0.7165	0.7341
MAPE	OLS	27.5844	16.5519
	SUR	26.4665	17.8153
RMSE	OLS	3.3552	9.2235
	SUR	3.4198	9.5476
MAE	OLS	2.7666	8.2309
	SUR	2.7289	8.1934

It can be said from Table 5 that performance metrics of the SUR and the OLS predicted responses, $\hat{Y}_1$ and $\hat{Y}_2$, are quite close even though the R^2 and the *RMSE* metrics of the OLS method are slightly better than the SUR method for both $\hat{Y}_1$ and $\hat{Y}_2$.

4. Conclusion

In this study, it was aimed to present that the SUR modeling approach is more proper than the OLS approach when the responses are

correlated in the multi-response experiments. For this purpose, a real data set about extraction process of grape seeds was used. It was seen from the analysis results that SUR parameter estimates have smaller variances than the OLS parameter estimates which affects the statistical inference of parameter estimates. For future work, it is planned to use different covariance matrix estimates which represent the correlation structure of the responses more precisely.

References

[1] Khuri, A. I. and Cornell, M. (1996). Response Surfaces, Marcel Dekker, Inc. New-York.

[2] Khuri, A. I. and Mukhopadhyay, S. (2010). Response Surface Methodology. *WIREs Computational Statistics*, 2, 128-149.

[3] Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298), 348-368.

[4] Shah, H. K., Montgomery, D.C. and Carlyle, W. M. (2004). Response Surface Modeling and Optimization in Multi-response Experiments Using Seemingly Unrelated Regressions. *Quality Engineering*, 16(3), 387–397.

[5] Peterson, J. J., Quesada, G. M. and Castillo, E.D. (2009). A Bayesian Reliability Approach to Multiple Response Optimization With Seemingly Unrelated Regression Models. *Quality Technology & Quantitative Management*, 6(4), 353–369.

[6] Hejazi, T. H., Esfahani, M. S., Mahootchi, M. (2015). Optimization of Degree of Conformance in Multi-response-Multistage
Systems with a Simulation-based Metaheuristic. *Quality Reliability Engineering International*, 31(4), 645–658.

[7] Fogliatto, F. S. and Albin, S. L. (2000). Variance of Predicted Response as an Optimization Criterion in Multi-response Experiments. *Quality Engineering*, 12(4), 523-533.

[8] Matsuura, S. and Kurata, H. (2019). Covariance Matrix Estimation In A Seemingly Unrelated Regression Model Under Stein's Loss. *Statistical Methods & Applications*, 1-21.

[9] Kurata, H. and Matsuura, S. (2016). Best Equivariant Estimator of Regression Coefficients in a Seemingly Unrelated
Regression Model with Known Correlation Matrix. *Annals of the Institute of Statistical Mathematics*, 68(4), 705–723.

[10] Rana, S. and Mastak Al Amin, M. (2015). An Alternative Method of Estimation of SUR Model. *American Journal of Theoretical and Applied Statistics*, 4(3), 150-155.

[11] Ghazal, G. A. and Hegazy, S. A. (2015) The Two Feasible Seemingly Unrelated Regression Estimator. *International Journal of Scientific & Technology Research*, 4(4), 247-253.

[12] Vural, N., Cavuldak, Ö. A. and Anlı, R. E. (2018) Multi Response Optimization of Polyphenol Extraction Conditions From Grape Seeds by Using Ultrasound Assisted Extraction (UAE). *Separation Science and Technology*, 53(10), 1540-1551.

[13] Türkşen, Ö. (2011). Fuzzy and Heuristic Approach to the Solution of Multi Response Surface Problems, PhD. Thesis, Ankara.

Statistical and Fuzzy Modeling of Extraction Process in Green Chemistry

Nilüfer Vural
Ankara University
Faculty of Engineering
Chemical Engineering Dept.
Ankara, Turkey
nvural@science.ankara.edu.tr

Özlem Türkşen
Ankara University
Faculty of Science
Statistics Dept.
Ankara, Turkey
turksen@ankara.edu.tr

Abstract

In this study, a replicated response measured (RRM) data set about grape seeds phenolic compounds extraction is used to apply the statistical and fuzzy regression analysis in green chemistry. The grape seeds phenolic compounds ultrasound assisted extraction (UAE) can be considered as an application field of green chemistry. The RRM data set was composed with four input variables, EtOH concentration (X_1), extraction time (X_2), solvent: solid ratio (X_3) and extraction temperature (X_4) and two responses, total phenolic content-TPC (Y_1) and total antioxidant activity-TAA (Y_2), The Y_1 and Y_2 have five replicates. The median of five replicates were used for statistical regression analysis. In order to apply fuzzy regression analysis replicated responses were represented as fuzzy numbers, e.g. triangular, trapezoidal, pentagonal. Descriptive statis tics of the replication response measures are used to transform the replications to the fuzzy numbers. The modeling performance of predicted responses were compared by using several performance metrics. It is seen from the analysis results that soft computing modeling tool can be used alternatively for modeling stage of extraction process.

Keywords: Replicated response measures; fuzzy regression analysis; extraction process

1. Introduction

Extraction is very important for chemical industry applications and analytical purposes. Studies in this field are mostly concentrated in the production of polyphenolic compounds with bioactive properties used in the field of food, medicine and cosmetics. The grape seed can be used as a source of natural food additive, in which showing extremely rich source of phenolics [1, 2]. At this point, the extraction technique being studied is environmentally friendly. For this purpose, green chemical (microwave, ultrasound, high pressure, stressed electric field, ohmic and super critical fluid extraction, etc.) methods, which are developed as an alternative to traditional extraction methods are aimed at increasing the efficiency as well as decreasing solvent usage and extraction time [3]. A basic problem, at the present stage of the green chemistry processes, is how to manage the cognitive process while taking into account its intrinsic features of uncertainty, including imprecision and vagueness [4]. In fact, uncertainty stems from

a deficiency of information. This has both theoretical and practical implications in Chemical Technology. In fact, real-engineering problems are the prime source of motivation for this management to be considered. In the experimental studies of engineering field, researchers need to analyse the variability of the responses in general. For this purpose, the experimental data set is composed with replicated response measures. One the main step of analysis for replicated response measured (RRM) data set can be considered as mode- ling stage with minimum error [5]. It is possible to model the data set by using statistical computational methods, e.g. statis- tical regression analysis, and soft computing based methods, e.g. fuzzy regression analysis, in chemometric studies. The RRM data set should satisfy some modeling assumptions to apply statistical computati- onal methods. However, satisfying statistical modeling assumptions seems hard in many of the chemometric studies. In this case, the soft computing based methods are proper to model the RRM data set [6].

This study was presenting a flexible modeling approaches for replicated response measured data set. In this study, the replicated response measured data was modeled by using statistical and fuzzy regression. The concept of fuzzy numbers is one of the fuzzy tools used in engineering applications in recent years. It is particularly useful in representing and capturing subjectivity and uncertainty. Fuzzy numbers are basically a fuzzy set with normality, convexity and continuity per piece. The concept of fuzzy numbers and fuzzy arithmetic was introduced by Zadeh [7]. In particular, the pervasive expansion of the original theory of fuzzy sets in the fields of computer, mathematics, and engineering has provided fruitful ideas and new tools to chemometric methodology. Different fuzzy numbers and range-valued fuzzy numbers have been used in the literature depending on the nature of the uncertainty and the problems in various applications.

Trapezoidal fuzzy numbers are a popular form of fuzzy numbers [8]. Interval numbers and triangular fuzzy number [9] are special cases of trapezoidal fuzzy number. The definition of heuristic fuzzy number has been proposed and this idea has been used to introduce intuitionistic triangular fuzzy numbers and heuristic trapezoidal fuzzy numbers[10,11]. The pentagonal fuzzy number was first proposed by [12] and [13]. Another field where the fuzzy approach provides fruitful results is regression analysis [14,15,16]. Fuzziness may affect the regression model assumed for analyzing the data (i.e. its parameters, may be thought of as fuzzy sets, in particular, fuzzy numbers), as well as the observed data. In the latter case, three situations can be envisaged: (a) fuzzy response, crisp explanatory variables; (b) crisp response, fuzzy explanatory variables; and (c) fuzzy response and explanatory variables. This study, take into consideration situation (a).

The manuscript was organized as follows in Section 2, brief description about trans-forming replicated response measures to fuzzy numbers is given. Section 3 contains modeling of replicated response measures multi-response data set by using statistical and fuzzy modeling approaches. Section 4 presents an application of a real data set about grape seeds phenolic compounds extraction. Finally, conclusion was given in Section 4.

2. Describing Replicated Response Measures As Fuzzy Numbers

An experimental design in which *r* response variables have *t* replicated values for each observed n independent units was expressed as $y_i = [y_{i1}\ y_{i2}...y_{ir}]$; $i = 1; 2; ..; n$; with *p* crisp input variables, $X = [X_1X_2...X_p]$.

The experimental design can be seen in Table 1.

Table 1. An experimental design for multi responses with replicated response measures

No	Input levels				Responses											
	X_1	X_2	…	X_p	$Y^{(1)}$				$Y^{(2)}$				…	$Y^{(r)}$		
1	x_{11}	x_{12}	…	x_{1p}	$y_{11}^{(1)}$	$y_{12}^{(1)}$	…	$y_{1t}^{(1)}$	$y_{11}^{(2)}$	$y_{12}^{(2)}$	…	$y_{1t}^{(2)}$	…	$y_{11}^{(r)}$	$y_{12}^{(r)}$	… $y_{1t}^{(r)}$
2	x_{21}	x_{22}	…	x_{2p}	$y_{21}^{(1)}$	$y_{22}^{(1)}$	…	$y_{2t}^{(1)}$	$y_{21}^{(2)}$	$y_{22}^{(2)}$	…	$y_{2t}^{(2)}$	…	$y_{21}^{(r)}$	$y_{22}^{(r)}$	… $y_{2t}^{(r)}$
…	……				…				…				..	……		
n	x_{n1}	x_{n2}	…	x_{np}	$y_{n1}^{(1)}$	$y_{n2}^{(1)}$	…	$y_{nt}^{(1)}$	$y_{n1}^{(2)}$	$y_{n2}^{(2)}$	…	$y_{nt}^{(2)}$	…	$y_{n1}^{(r)}$	$y_{n2}^{(r)}$	… $y_{nt}^{(r)}$

In order to transform the replicated measures to the fuzzy numbers (FNs), descriptive statistics of replicated response measures were calculated for each unit. The experi-mental design with fuzzy response values, $\tilde{y}_i$, $i=1,2,...,n$,;was given in Table 2.

Table 2. A multi-response experimental design with fuzzy responses

	Input variables				Responces			
No	X_1	X_2	…	X_p	$\sim^{(1)}$ Y	$\sim^{(2)}$ Y	…	$\sim^{(r)}$ Y
1	x_{11}	x_{12}	…	x_{1p}	$\sim^{(1)} y_1$	$\sim^{(2)} y_1$	…	$\sim^{(r)} y_1$
2	x_{21}	x_{22}	…	x_{2p}	$\sim^{(1)} y_2$	$\sim^{(2)} y_2$	…	$\sim^{(r)} y_2$
…	…				…	…	.	..
n	x_{n1}	x_{n2}	…	x_{np}	$\sim^{(1)} y_n$	$\sim^{(2)} y_n$	…	$\sim^{(r)} y_n$

In this study, the replicated response measu-res used for the ultrasound assisted extrac- tion (UAE) process were considered as tri angular fuzzy numbers (TFNs), trapezoidal fuzzy numbers (TrFNs,) and pentagonal fuzzy numbers (PFNs). The fuzzification of the replicated response measures was achieved by using descriptive statistic of replicated measures.

In this study, two different fuzzification formula, Formula-A and Formula-B, were used for fuzzification procedure. Let consi-der $y_i=(y_{i1},y_{i2},y_{i3},...,y_{it})$ be the replicated values of *rth* response and belong to its observation, $i=1,2,...,n$. A PFN is presen ted with five components, donated as $\tilde{y}_i=(y_i^l,y_i^m,y_i^c,y_i^n,y_i^u),\ i=1,2,...,n$.

Here, y_i^l and y_i^m were the smallest posibble values, y_i^c was the most promising value, y_i^n and y_i^u were the largest posibble values.

i. Fuzzification Formula-A

The first formula to generate the pentagonal fuzzy value was defined as

$$\tilde{y}_i=(y_{i(0.25)}-1.5*IQR_i,\ y_{i(0.25)},\ y_{i(0.50)},\ y_{i(0.75)},\ y_{i(0.75)}+1.5*IQR_i),\ i=1,2,...,n \quad (1)$$

in which $y_{i(0.25)}$ was the first quartile of replicated response measures, $y_{i(0.50)}$ was the second quartile (median) of replicated response measures, $y_{i(0.75)}$ was the third quartile of replicated response measures res, $IQRi$ was the inter quartile range of replicated response measures, calculated as $IQRi = y_{i(0,75)} - y_{i(0.25)}$ for each unit i, $i=1,2,...,n$ [17].

According to the Eq.(1), trapezodial and triangular fuzzy response values are defined as

$$\tilde{y}_i = (y_{i(0.25)} - 1.5 * IQR_i, y_{i(0.25)}, y_{i(0.75)}, y_{i(0.75)} + 1.5 * IQR_i) \quad (2)$$

and

$$\tilde{y}_i = (y_{i(0.25)} - 1.5 * IQR_i, y_{i(0.50)}, y_{i(0.75)} + 1.5 * IQR_i) \quad (3)$$

respectively.

ii. **Fuzzification Formula-B**

The second formula is generate the penta gonal fuzzy response value was defined as

$$\tilde{y}_i = (y_{i(\min)}, y_{i(0.25)}, y_{i(0.50)}, y_{i(0.75)}, y_{i(\max)}), \quad i = 1, 2, ..., n \quad (4)$$

where $y_{i(\min)}$ and $y_{i(\max)}$ are the minimum and maximum values of the replicated for *ith* observation. According to the Eq.(4), trapezodial and triangular fuzzy response values are defined as

$$\tilde{y}_i = (y_{i(\min)}, y_{i(0.25)}, y_{i(0.75)}, y_{i(\max)}), \quad i = 1, 2, ..., n \quad (5)$$

and

$$\tilde{y}_i = (y_{i(\min)}, y_{i(0.50)}, y_{i(\max)}), \ i = 1, 2, ..., n \quad (6)$$

respectively.

3. Modeling Through Regression Analysis

3.1. Statistical Regression Analysis

One of the main aim in a multi-response experiment with replicated response measures is modeling of the responses as a functions of the input variables simultenously with minimum error. General formula for multi response model can be written in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (7)$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_1 : \boldsymbol{Y}_2 : ... : \boldsymbol{Y}_r']'$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1' : \boldsymbol{\beta}_2' : .... : \boldsymbol{\beta}_r']'$, $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1' : \boldsymbol{\varepsilon}_2' : .... : \boldsymbol{\varepsilon}_r']'$, and $\mathbf{X}$ is the block diagonal matrix, $\boldsymbol{diag}(\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_r)$ [18].

The model parameters are estimated by using ordinary least squares (OLS) approach as a common way of statistical regression analysis. The estimates of model parameters are obtained as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (8)$$

However, modeling the data set with replicated response measures, is not simple since the classical modeling assumptions can be violated and replicated measures of responses may cause uncertainty. In this case, it is necessary to represent the replicated measures without loosing much information at each experimental unit [18].

3.2. Fuzzy Regression Analysis

$$\tilde{\boldsymbol{Y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \quad (9)$$

in which the observed response values, model coefficients and errors are PFNs, TrFNs and TFNs denoted as

(i) $\tilde{\boldsymbol{Y}}^{P} = (\boldsymbol{Y}^{l}, \boldsymbol{Y}^{m}, \boldsymbol{Y}^{c}, \boldsymbol{Y}^{n}, \boldsymbol{Y}^{u})$, $\tilde{\boldsymbol{\beta}}^{P} = (\boldsymbol{\beta}^{l}, \boldsymbol{\beta}^{m}, \boldsymbol{\beta}^{c}, \boldsymbol{\beta}^{n}, \boldsymbol{\beta}^{u})$, $\tilde{\boldsymbol{\varepsilon}}^{P} = (\boldsymbol{\varepsilon}^{l}, \boldsymbol{\varepsilon}^{m}, \boldsymbol{\varepsilon}^{c}, \boldsymbol{\varepsilon}^{n}, \boldsymbol{\varepsilon}^{u})$,

(ii) $\tilde{\boldsymbol{Y}}^{Tr} = (\boldsymbol{Y}^{l}, \boldsymbol{Y}^{m}, \boldsymbol{Y}^{n}, \boldsymbol{Y}^{u})$, $\tilde{\boldsymbol{\beta}}^{P} = (\boldsymbol{\beta}^{l}, \boldsymbol{\beta}^{m}, \boldsymbol{\beta}^{n}, \boldsymbol{\beta}^{u})$, $\tilde{\boldsymbol{\varepsilon}}^{P} = (\boldsymbol{\varepsilon}^{l}, \boldsymbol{\varepsilon}^{m}, \boldsymbol{\varepsilon}^{n}, \boldsymbol{\varepsilon}^{u})$, and

(iii) $\tilde{\boldsymbol{Y}}^{T} = (\boldsymbol{Y}^{l}, \boldsymbol{Y}^{c}, \boldsymbol{Y}^{u})$, $\tilde{\boldsymbol{\beta}}^{P} = (\boldsymbol{\beta}^{l}, \boldsymbol{\beta}^{c}, \boldsymbol{\beta}^{u})$, $\tilde{\boldsymbol{\varepsilon}}^{P} = (\boldsymbol{\varepsilon}^{l}, \boldsymbol{\varepsilon}^{c}, \boldsymbol{\varepsilon}^{u})$,

respectively. The fitted fuzzy response model can be written as

$$\hat{\tilde{\boldsymbol{Y}}} = \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}} \quad (10)$$

where $\hat{\tilde{\boldsymbol{Y}}}$ and $\hat{\tilde{\boldsymbol{\beta}}}$ are predicted fuzzy response values and fuzzy model coefficients. The estimaters of fuzzy model coefficient vector was calculated by optimizing the following

least squares problem with respect to Diamond distance metric [19].

$$\min_{\tilde{\beta}} \emptyset\left(\tilde{\beta}\right) = \sum_{k=1}^{r} d\left(\tilde{Y}_k, \hat{\tilde{Y}}_k\right) \quad (11)$$

According to the operation properties of PFNs and assuming that the $(X'X)$ is nonsingular, the elements of fuzzy model coefficient vector are obtained as

$$\hat{\beta}^{L} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{L}$$
$$\hat{\beta}^{m} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{m}$$
$$\hat{\boldsymbol{\beta}}^{c} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{c} \quad (12)$$
$$\hat{\boldsymbol{\beta}}^{n} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{n}$$
$$\hat{\beta}^{u} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{u}$$

It is possible to obtain trapezodial and triangular fuzzy model coefficients easily by using the Eq. (12)

4. Application

The RRM data set about grape seeds phenolic compounds extraction was used to apply the statistical and fuzzy regression analysis in green chemistry. In the scope of the study, the RRM data was taken from the study of [2].The RRM data set was composed with two responses, total phenolic content-TPC (mg GAE/g, Y_1) and total antioxidant activity-TAA ((% inh, Y_2), and four input variables, EtOH concentration (v/v%, X_1), extraction time (min, X_2), solvent: solid ratio (mL/g, X_3) and extraction temperature (X_4). The real values, corresponding to the coded levels of independent variables,were given in Table 3.

Response Surface Methodology (RSM) was used to determine the optimal conditions for the UAE extraction of polyphenolics from grape seed. Five level and four independent variables central composite rotatable design (CCRD-Box Wilson-star orthogonal design) with 30 runs was used. The experimental results, Y_1 and Y_2, with five replicated response measures for extraction process were given in Table 4. The detailed information about the experiment is given in the study of [2]. Vural et al (2018) data set was considered as replicated response measu red data set to illustrate the proposed fuzzy modeling approach with FNs and statistical regression modelling.

In this paper, it is assumed that the responses are uncorelated and are represented by second order polinomial functions. All the calculations were done in MatLab R2013a.
For the data set given in Table 4, statistical regression analysis was performed using the ordinary least squares (OLS) approach and the model parameters were estimated. The results of ANOVA of OLS regression analysis were given in Table 5. The design of multi-response experiment with fuzzy observed responses were shown in Table 6 and Table 7. Fuzzification-A were used in Tables 6(a)-7(a) were used and Fuzification-B responses were shown in Table6(b)-7(b). The predicted fuzzy regression models and OLS model for the first response of data set in Table 8-(a) and, the second response in Table 8-(b) was shown. It should be noted that the observed and predicted response values are triangular fuzzy numbers in fuzzification models. The results of root mean square error (RMSE) and mean absolute error (MAE) criteria were shown in Tables 9 (a)-(b).

Table 3. The coded and actual levels of independent variable

	Coded levels				
Independent variables	***-α (-2)***	***-1***	***0***	***+1***	***+α (+2)***
X_1: EtOH concentration (%)	0	25	50	75	100
X_2: Extraction time (min)	0	10	20	30	40
X_3: Solvent/solid ratio (mL/g)	4.5	13	21.5	30	38.5
X_4: Extraction temperature (°C)	20	30	40	50	60

Table 4. Experimental design with five replicated response measures

No	X_1	X_2	X_3	X_4	Y_1 Rep1 Rep2 Rep3 Rep4 Rep5	Y_1 Median	Y_2 Rep1 Rep2 Rep3 Rep4 Rep5	Y_2 Median
	(%)	(min)	(mL/g)	(^oC	(mgGAE/g)		(% inh)	
1	-1(25)	-1(10)	-1(13)	-1(30)	3.43, 3.67, 3.91, 3.54, 4.01	3.67	27.94, 28.18, 28.42, 27.87, 28.31	28.18
2	+1(75)	-1(10)	-1(13)	-1(30)	8.03, 8.10, 8.17, 7.91, 8.37	8.10	51.29, 53.11, 54.93, 52.33, 55.69	53.11
3	-1(25)	+1(30)	-1(13)	-1(30)	9.15, 9.53, 9.91, 9.48,10.12	9.53	54.03, 55.66, 57.29, 54.45, 57.65	55.66
4	+1(75)	+1(30)	-1(13)	-1(30)	10.54, 10.99, 11.44, 10.48, 11.35	10.99	56.35, 58.70, 61.05, 57.87, 62.11	58.70
5	-1(25)	-1(10)	+1(30)	-1(30)	10.45, 10.62, 10.79, 10.39, 10.73	10.73	50.80, 51.82, 52.84, 51.45, 53.29	51.82
..	……							
28	0(50)	0(20)	0(21.5)	0(40)	11.15, 11.27, 11.39, 11.19, 11.43	11.27	65.24, 65.56, 65.88, 65.38, 66.00	65.56
29	0(50)	0(20)	0(21.5)	0(40)	17.21, 17.46, 17.71, 17.58, 17.87	17.58	76.59, 78.32, 80.05, 77.17, 80.53	78.32
30	0(50)	0(20)	0(21.5)	0(40)	18.42, 18.67, 18.92, 19.02, 18.97	18.92	78.95, 82.85, 86.75, 78.89, 86.65	82.85

Table 5. ANOVA results for TPC and TAA using quadratic statistical regression analysis

	TPC						**TAA**					
	Coefficient	Adj SS	df	Adj MS	F value	P value Prob > F	Coefficient	Adj SS	df	Adj MS	F value	P value Prob > F
Model	14.87	932.03	6	155.34	10.22	0.000	72.38	7797.2	6	1299.53	11.71	0.000
X_1	3.36	270.55	1	270.55	17.80	0.000	10.60	2695.6	1	2695.6	24.29	0.000
X_2	1.17	33.09	1	33.09	2.18	0.154	6.46	1002.5	1	1002.5	9.03	0.006
X_3	3.70	329.15	1	329.15	21.65	0.000	8.30	1653.9	1	1653.9	14.90	0.001
X_4	2.66	169.50	1	169.50	11.15	0.003	5.33	681.5	1	681.5	6.14	0.021
X_1^2	-1.48	63.00	1	63.00	4.14	0.053	-4.76	645.6	1	645.6	5.82	0.024
X_2^2							-6.76	1300	1	1300	11.72	0.002
X_1X_3	2.04	66.75	1	66.75	4.39	0.047						
Error		349.66	23	15.20				2552.2	23	110.97		
Lack of Fit		280.44	18	15.58	1.13	0.491		2224.8	18	123.60	1.89	0.249
Pure Error		69.22	5	13.84				327.4	5	65.47		
Total		1281.69	29					10349.4	29			
R^2		0.7272						0.7534				
R^2 Adj		0.6560						0.6891				

P values represents the statisticaly significant terms ($p<0.05$)

Table 6(a). The data set with fuzzy valued responses for TPC (Fuzzification-A)

No	X_1	X_2	X_3	X_4	$\tilde{\mathbf{Y}}_1$		
					$\tilde{Y}_1^P$	$\tilde{Y}_1^{Tr}$	$\tilde{Y}_1^T$
1	-1	-1	-1	-1	(2.88, 3.51, 3.67, 3.94, 4.57)	(2.88, 3.51, 3.94, 4.57)	(2.88, 3.67, 4.57)
2	1	-1	-1	-1	(7.67, 8.00, 8.10, 8.22, 8.55)	(7.67, 8.00, 8.22, 8.55)	(7.67, 8.10, 8.55)
3	-1	1	-1	-1	(8.55, 9.40, 9.53, 9.96, 10.81)	(8.55, 9.40, 9.96, 10.81)	(8.55, 9.53,10.81)
..							
29	0	0	0	0	(16.87, 17.40, 17.58, 17.75, 18.28)	(16.87, 17.40, 17.75, 18.28)	(16.87, 17.58, 18.28)
30	0	0	0	0	(18.05, 18.61, 18.92, 18.98, 19.55)	(18.05, 18.61, 18.98, 19.55)	(18.05, 18.92, 19.55)

Table 6(b). The data set with fuzzy valued responses for TPC (Fuzzification-B)

No	X_1	X_2	X_3	X_4	$\tilde{\mathbf{Y}}_1$		
					$\tilde{Y}_1^P$	$\tilde{Y}_1^{Tr}$	$\tilde{Y}_1^T$
1	-1	-1	-1	-1	(3.43, 3.51, 3.67, 3.94, 4.01)	(3.43, 3.51, 3.94, 4.01)	(3.43, 3.67, 4.01)
2	1	-1	-1	-1	(7.91, 8.00, 8.10, 8.22, 8.37)	(7.91, 8.00, 8.22, 8.37)	(7.91, 8.10, 8.37)
3	-1	1	-1	-1	(9.15, 9.40, 9.53, 9.96, 10.12)	(9.15, 9.40, 9.96, 10.12)	(9.15, 9.53, 10.12)
..							
29	0	0	0	0	(17.21, 17.40, 17.58, 17.75, 17.87)	(17.21, 17.40, 17.75, 17.87)	(17.21, 17.58, 17.87)
30	0	0	0	0	(18.42, 18.61, 18.92, 18.98, 19.02)	(18.42, 18.61, 18.98, 19.02)	(18.42, 18.92, 19.02)

Table 7(a). The data set with fuzzy valued responses for TAA (Fuzzification-A)

No	X_1	X_2	X_3	X_4	$\tilde{Y}_2$		
					$\tilde{Y}_2^P$	$\tilde{Y}_2^{Tr}$	$\tilde{Y}_2^T$
1	-1	-1	-1	-1	(27.30, 27.92, 28.18, 28.34, 28.96)	(27.30, 27.92, 28.34, 28.96)	(27.30, 28.18, 28.96)
2	1	-1	-1	-1	(47.50, 52.07, 53.11, 55.12, 59.70)	(47.50, 52.07, 55.12, 59.70)	(47.50, 53.11, 59.70)
3	-1	1	-1	-1	(49.79, 54.35, 55.66, 57.38, 61.93)	(49.79, 54.35, 57.38, 61.93)	(49.79, 55.66, 61.93)
…							
29	0	0	0	0	(72.31, 77.03, 78.32, 80.17, 84.89)	(72.31, 77.03, 80.17, 84.89)	(72.31, 78.32, 84.89)
30	0	0	0	0	(67.33, 78.94, 82.85, 86.68, 98.29)	(67.33, 78.94, 86.68, 98.29)	(67.33, 82.85, 98.29)

Table 7(b). The data set with fuzzy valued responses for TAA (Fuzzification-B)

No	X_1	X_2	X_3	X_4	$\tilde{Y}_2$		
					$\tilde{Y}_2^P$	$\tilde{Y}_2^{Tr}$	$\tilde{Y}_2^T$
1	-1	-1	-1	-1	(27.87, 27.92, 28.18, 28.34, 28.42)	(27.87, 27.92, 28.34, 28.42)	(27.87, 28.18, 28.42)
2	1	-1	-1	-1	(51.29, 52.07, 53.11, 55.12, 55.69)	(51.29, 52.07, 55.12, 55.69)	(51.29, 53.11, 55.69)
3	-1	1	-1	-1	(54.03, 54.35, 55.66, 57.38, 57.65)	(54.03, 54.35, 57.38, 57.65)	(54.03, 55.66, 57.65)
..							
29	0	0	0	0	(76.59, 77.03, 78.32, 80.17, 80.53)	(76.59, 77.03, 80.17, 80.53)	(76.59, 78.32, 80.53)
30	0	0	0	0	(78.89, 78.94, 82.85, 86.68, 86.75)	(78.89, 78.94, 86.68, 86.75)	(78.89, 82.85, 86.75)

Table 8 (a). The predicted fuzzy regression models for the TPC ($\widetilde{Y}_1$)

OLS	Y_1	Y_1= 14.866 +3.358 X_1+1.174 X_2+3.703 X_3+ 2.658 X_4+ 2.043 X_1X_3-1.479 X_1^2
Fuzzification A	$\widetilde{Y}_1^T$	$\widetilde{Y}_1^T$= (13.125, 14.859, 16.587) + (2.255, 3.362, 4.415)X_1 +(-0.096, 1.179, 2.338) X_2+(2.656, 3.698, 4.707) X_3+ (1.472, 2.662, 3.915)X_4+(0.688, 2.049, 3.411) X_1X_3 + (-2.337, -1.478, -0.564) X_1^2
	$\widetilde{Y}_1^{Tr}$	$\widetilde{Y}_1^{Tr}$== (13.125, 14.423, 15.289, 16.587) + (2.255, 3.065, 3.605, 4.415)X_1 +(-0.096, 0.817, 1.425, 2.338) X_2+(2.656, 3.425, 3.938, 4.707) X_3+ (1.472,2.388, 2.999, 3.915)X_4+(0.688, 1.709, 2.390, 3.411) X_1X_3 + (-2.337, -1.672, -1.229, -0.564) X_1^2
	$\widetilde{Y}_1^P$	$\widetilde{Y}_1^P$=(13.125, 14.423, 14.859, 15.289, 16.587) + (2.255, 3.065, 3.362, 3.605, 4.415)X_1 +(-0.096, 0.817, 1.179, 1.425, 2.338) X_2+(2.656, 3.425, 3.699, 3.938, 4.707) X_3+ (1.472, 2.388, 2.662, 2.999 3.915)X_4+ (0.688, 1.709, 2.049, 2.390, 3.411) X_1X_3 + (-2.337, -1.672, -1.478, -1.229, -0.564) X_1^2
Fuzzification A	$\widetilde{Y}_1^T$	$\widetilde{Y}_1^T$=(14.186, 14.859, 15.452) + (2.946, 3.362, 3.719)X_1 +(0.659, 1.179, 1.571) X_2+(3.309, 3.698, 4.039) X_3+ (2.274, 2.662, 3.114)X_4+(1.550, 2.049, 2.516) X_1X_3 + (-1.769, -1.478, -1.114) X_1^2
	$\widetilde{Y}_1^{Tr}$	$\widetilde{Y}_1^{Tr}$=(14.186, 14.423, 15.289, 15.452) + (2.946, 3.065, 3.605, 3.719)X_1 +(0.659, 0.817, 1.425, 1.571) X_2+(3.309, 3.425, 3.938, 4.039) X_3+ (2.274, 2.388, 2.999, 3.114)X_4+(1.550, 1.709, 2.390, 2.516) X_1X_3 + (-1.769,-1.672, -1.229, -1.114) X_1^2
	$\widetilde{Y}_1^P$	$\widetilde{Y}_1^P$=(14.186, 14.423, 14.859, 15.289, 15.452) + (2.946, 3.065, 3.362, 3.605, 3.719)X_1 +(0.659, 0.817, 1.179, 1.425, 1.572) X_2+(3.309, 3.425, 3.699, 3.938, 4.039) X_3+ (2.274, 2.388, 2.662, 2.999, 3.114)X_4+ (1.550, 1.709, 2.049, 2.390, 2.516) X_1X_3 + (-1.769, -1.672, -1.478, -1.229, -1.114) X_1^2

Table 8(b). The predicted fuzzy regression models for the TAA ($\widetilde{Y}_2$)

OLS	Y_2	Y_2= 72.38 +10.60 X_1+6.46 X_2+8.30 X_3+ 5.33 X_4 - 4.76 X_1^2-6.76 X_2^2
Fuzzification A	$\widetilde{Y}_2^T$	$\widetilde{Y}_2^T$ = (67.796, 72.376,77.105)+(7.026,10.598, 14.239)X_1+(3.232, 6.463, 9.721)X_2+(5.244, 8.301, 11.139) X_3+(2.396, 5.329, 8.255) X_4 +(-7.345,-4.764, -2.035) X_1^2 +(-8.958, -6.760, -4.463) X_2^2
	$\widetilde{Y}_2^{Tr}$	$\widetilde{Y}_2^{Tr}$ =(67.796, 71.287, 73.614, 77.105)+(7.026, 9.731, 11.534, 14.239)X_1+(3.232, 5.665, 7.288, 9.721) X_2+(5.244, 7.455, 8.929, 11.139) X_3+(2.396, 4.593, 6.058, 8.255) X_4 +(-7.345, -5.354, -4.026, -2.035) X_1^2 +(-8.958, -7.272, -6.149, -4.463) X_2^2
	$\widetilde{Y}_2^P$	$\widetilde{Y}_2^P$=(67.796, 71.287, 72.376, 73.614, 77.105)+(7.026, 9.731, 10.598, 11.534, 14.239)X_1+(3.232, 5.665, 7.288, 9.721) X_2+(5.244, 7.455, 8.929, 11.139) X_3+(2.396, 4.593, 6.058, 8.255) X_4 +(-7.345, -5.354, -4.026, -2.035) X_1^2 +(-8.958, -7.272, -6.149, -4.463) X_2^2
Fuzzification B	$\widetilde{Y}_2^T$	$\widetilde{Y}_2^T$= (70.989, 72.376, 73.888)+(9.410,10.598, 11.835)X_1+(5.378, 6.463, 7.576)X_2+(7.215, 8.301, 9.211) X_3+(4.344, 5.329, 6.306) X_4 +(-5.556,-4.764, -3.856) X_1^2 +(-7.445, -6.760, -6.000) X_2^2
	$\widetilde{Y}_2^{Tr}$	$\widetilde{Y}_2^{Tr}$ =(70.989, 71.287, 73.614, 73.888)+(9.410, 9.731, 11.534, 11.835)X_1+(5.378, 5.665, 7.288, 7.576)X_2+ (7.215, 7.455, 8.929, 9.211) X_3+(4.344, 4.593, 6.058, 6.306) X_4 +(-5.556, -5.354, -4.026, -3.856) X_1^2 + (-7.445, -7.272, -6.149, -6.000) X_2^2
	$\widetilde{Y}_2^P$	$\widetilde{Y}_2^P$ = (70.989, 71.287, 72.376,73.614, 73.888)+(9.410, 9.731, 10.598, 11.534, 11.835)X_1+(5.378, 5.665,6.463, 7.288, 7.576)X_2+ (7.215, 7.455, 8.301, 8.929, 9.211) X_3+(4.344, 4.593,5.329, 6.058, 6.306) X_4 +(-5.556, -5.354,-4.764, -4.026, -3.856) X_1^2 + (-7.445, -7.272, -6.760, -6.149, -6.000) X_2^2

Table 9(a).RMSE for the TPC ($\widetilde{Y}_1$)andTAA ($\widetilde{Y}_2$)

	RMSE			
OLS	Y_1	3.8987	Y_2	10.5342
Fuzzification A	$\widetilde{Y}_1^T$	4.7081	$\widetilde{Y}_2^T$	12.8397
	$\widetilde{Y}_1^{Tr}$	4.5519	$\widetilde{Y}_2^{Tr}$	12.4962
	$\widetilde{Y}_1^P$	4.4934	$\widetilde{Y}_2^P$	12.2946
Fuzzification A	$\widetilde{Y}_1^T$	3.5865	$\widetilde{Y}_2^T$	9.6670
	$\widetilde{Y}_1^{Tr}$	3.7070	$\widetilde{Y}_2^{Tr}$	10.1291
	$\widetilde{Y}_1^P$	3.6349	$\widetilde{Y}_2^P$	9.8794

Table 9 (b). MAE for the TPC ($\widetilde{Y}_1$)and TAA($\widetilde{Y}_2$)

	MAE			
OLS	Y_1	2.8397	Y_2	8.2309
Fuzzification A	$\widetilde{Y}_1^T$	3.7990	$\widetilde{Y}_2^T$	10.4921
	$\widetilde{Y}_1^{Tr}$	3.6818	$\widetilde{Y}_2^{Tr}$	10.2257
	$\widetilde{Y}_1^P$	3.0118	$\widetilde{Y}_2^P$	10.0928
Fuzzification B	$\widetilde{Y}_1^T$	2.9791	$\widetilde{Y}_2^T$	8.4568
	$\widetilde{Y}_1^{Tr}$	3.0668	$\widetilde{Y}_2^{Tr}$	8.6991
	$\widetilde{Y}_1^P$	3.0118	$\widetilde{Y}_2^P$	8.5663

5. Conclusion

In this study, replicated response measured data set about grape seeds phenolic compounds extraction was modeled by using statistical and fuzzy regression analysis. In order to apply the fuzzy modeling replicated response measures were transformed to fuzzy numbers, e.g. pentagonal, trapezoidal, triangular. The main aim of the study was to present the applicability of different type fuzzy numbers for the replicated data set. It was seen from the analysis results that riangular fuzzy numbers, composed with order statistics, were more proper for fuzzy modeling of five replicated response measured data set.

References

[1] Shi, J., Yu, J., Pohorly, J.E. (2003) Polyphenolics in grape seeds biochemistry and functionality. *Journal of Medicinal Food*.; 6:291-299.

[2] Vural, N., Cavuldak, Ö.A., Anlı, R.E. (2018) Multi response optimisation of polyphenol extraction conditions from grape seeds by using ultrasound assisted extraction (UAE). *Separation Science and Technology*. 53(10), 1540-1551.

[3] Azmir, J., Zaidul, I., Rahman, M., Sharif, K., Mohamed, A., Sahena, F., Omar, A. (2013) Techniques for extraction of bioactive compounds from plant materials: a review. *J Food Eng.* 117 (4), 426-436.

[4] Coppi, R., Gil, M. A., Kiers, H.A.L. (2006) The fuzzy approach to statistical analysis. *Computational Statistics & Data Analysis.* 51(1), 1-14.

[5] Khuri, A.I. and Cornell, M. (1996) Response Surfaces, Marcel Dekker Inc., New York.

[6] Türkşen, Ö. and Güler, N. (2015) Comparison of fuzzy logic based models for the multi-response surface problems with replicated response measures. *Applied Soft Computing*, 37, 887-89.

[7] Zadeh, L. A. (1975) The concept of a Linguistic variable and applications to approximate reasoning part-I, II, III. *Information Science.* 8 (3),199–249. doi:10.1016/0020-0255(75)90036-5.

[8] Giachetti, R.E., Robert E.Y., Analysis of the error in the standard approximation used for multiplication of triangular and trapezoidal fuzzy numbers and the development of a new approximation, Fuzzy Sets and Systems, 91, No.1 (1997), 1-13.

[9] Pathinathan, T. and Ajay M. (2018) Interval-Valued Pentagonal Fuzzy Numbers. *International Journal of Pure and Applied Mathematics* , 119 (9), 177-187

[10] Abdullah, L., Kwan Ismail, W., Hamming. (2012), Distance in Intuitionistic Fuzzy Sets and Interval-valued Intuitionistic Fuzzy Sets: A Comparative Analysis Hamming Distance in Intuitionistic Fuzzy Sets and Interval-valued Intuitionistic Fuzzy Sets: A Comparative Analysis, 1, No. 1 7-11.

[11] Li, J., Wenyi Z., and Ping G.,Interval-valued intuitionistic trape- zoidal fuzzy number and its application, Systems, Man and Cybernetics (SMC), IEEE International Conference, San Diego, USA 2014.

[12] Kumar, R., Pathinathan, T (2015), Sieving out the Poor using Fuzzy decision Making Tools, *Indian Journal of Science and Technology*, 8, No.22 1-16.

[13] Kumar, R., Pathinathan, T. (2015), Sieving out the poor using Fuzzy Decision Making Tools with reference to Nalanda District, Bihar, In- dia, *Interantional Conference on Convergence Technology*, 5, (1) 890-891.

[14] Näther, W., (2000). On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. Metrika 51, 201–221.

[15] Guo, P., Tanaka, H. Dual models for possibilistic regression analysis. *Computational Statistics and Data Analysis* 51(1), 253–266 (2006).

[16] Coppi, R.,D'Urso, P., P., Giordani,P.,Santoro, A. (2006) Least squares estimation of a linear regression model with LR fuzzy response. *Computational Statistics & Data Analysis*, 51(1), 267-286.

[17] Türkşen, Ö. (2019) A nonlinear modeling with linear fuzzy numbers for replicated response measures, *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/ 03610918 .2019 .1634813.

[18] Türkşen, Ö., and N. Güler. (2015) Comparison of fuzzy logic based models for the multi-response surface problems with replicated response measures. *Applied Soft Computing* 37,887-896.

[19] P. Diamond. (1998) Fuzzy least squares, *Inf. Sci.* 46, 141–157.

Risk-based Fraud Analysis for Bank Loans with Autonomous Machine Learning

Yunus Emre Gundogmus
Marmara University
yemregun@gmail.com

Mert Nuhuz
Marmara University
mertnuhuz@gmail.com

Mujgan Tez
Marmara University
mtez@marun.edu.tr

Abstract

In this study, we create Classifier model with Supervised learning by using Customer Data and their loan results for customers who applied for loan. We use various data cleaning, feature extraction and feature selection studies that are performed on 67 variables containing the customer's financial information. Our scoring model was created using supervised learning and statistical machine learning based on target variable. The risk score was calculated for the customers and a variable cut-off value was determined accordingly to the sample. They were labeled Fraud and non-fraud with our risk scores. The algorithm instead its learning new customer types. It is now possible to continuously self-develop and analyze data on a monthly basis and to adapt to the conditions of the period. Tested with real data

Keywords: Autonomous Machine Learning; Risk-Based Scoring; Fraud Analysis; Feature Selection;

1. Introduction

Machine learning has recently made great strides in many application areas, fueling a growing demand for machine learning systems that can be used effectively by novices in machine learning.

Correspondingly, a growing number of commercial enterprises aim to satisfy this demand (e.g., BigML.com, Wise.io, SkyTree.com, RapidMiner.com, Dato.com, Prediction.io, DataRobot.com, Microsofts Azure Machine Learning [8], Googles Prediction API [9], and Amazon Machine Learning). At its core, every effective machine learning service needs to solve the fundamental problems especially in the field of finance especially for institutions that give credit to individuals, or the change of financial macroeconomic variables (inflation, interest rates). To solve this problem, new models are installed with performance times (customers separated by time), the model is constantly renewed, but this is a very costly and long task.

1.1. Related Works Overview

In other studies (DataRobot.com, H20, Sklearn's AutoML), automl has been described as a method that gives you the most appropriate model by automatically selecting variables and hyper parameters when given the data set. Here, automl actively renews its data set according to the current data and modeling it with the renewed data set, it will be described as a structure that can adapt to the current portfolio for you.

2. Auto-ML Architectural

The AutoML infrastructure has a unique pipeline in this work. There are 3 different data sets depending on the problem. To set up an automl infrastructure, data sets must first be created. In this part, we have to prepare 3 different data sets for our problem. These

Base: Financial data and fraud by people with improved performance (credit repayment completed) (our target variable) We use this data set when training the first model and then do not use it again. We continue to use the Main data set to train the model.

Main: This is the data set that will be modeled and will be updated monthly. It includes Customers' financial data and fraud. (Constantly updated every month, different from Base data set.)

Monthly: This data set is reset monthly. And there are new customers who have applied within the month. It is constantly updated according to the applications received during the day. In this data set, the Financial values of the Customers, Estimated Value (Fraud / Non-Fraud) and Actual value (Did he / she paid the loan at the end of the month?) Actual Value in Data is taken by the company at the end of the month. If the person has paid the loan, it is labeled as Non-Fraud, or if it has not.

After these are prepared, there will be 3 different pipelines in AutoML Architecture. These are shown in the figures below:

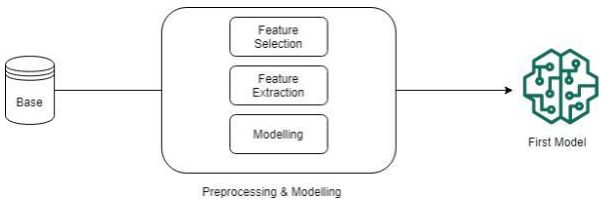

Figure 1. First Pipeline and Creating Base Model

In this study to create the first model, the first step was to combine data from different data sources. In this section, customers' previous application information, KKB information (US based), and the information at the time of the application was used.

The data was then multiplexed because there was more than one data for a single customer in different sources after the data was combined. As a result, we had 174 variables and about 8 million lines of data. [6] The first variable screening was performed by performing 2 different tests on this multiplexed data. The first of these tests was to examine the occupancy rate on a variable basis. In this test, if more than %20 of the variable is null, this variable is excluded from the modeling. The reason is that filling in financial data is very risky. In the content of this test, we found that meaningless if %85 of the categorical variable has same valued. And this variable was excluded from modeling.

Then, aggregate function was used in python to solve the multiplexing problem. This section looks at the modes of values as most customers multiply the same values. And the mode of the variables was taken as the singularized value of that customer.

In addition, a business expert with a business perspective for feature selection was examined for the explanations of the variables. In addition, many different tests were applied to the variables. These;

Variance Control for Continuous Variables: It is meaningless if the variance of the variable is too close to 0.

Gini Coefficient for Variables: If the gini coefficient is low, the variable is meaningless.

Pre-modeling of the variables and obtaining the Importance scores: In this test, firstly, a supervised modeling is performed with RandomForest algorithm with the target variable. Then the variables that the algorithm gives us are the Importance coefficients. We can understand whether the variable is significant according to these coefficients. The factor determining this coefficient is how much the variable explains the target variable on a variance basis.

After all these feature selection, the variables were finally eliminated with the algorithm Boruta [3]. And modeling was performed with the remaining variables. After that, the modeling phase will be started. The modeling section and the procedures are described in the Supervised Learning at Section 2.2

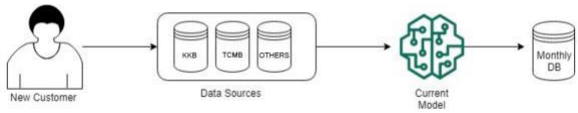

Figure 2. Daily Pipeline

The financial information of the individuals is obtained from the institutions where many banks such as KKB [7] cooperate and share their data. Apart from that person's age, marital status, spouse information is obtained from the system of the republic of turkey.

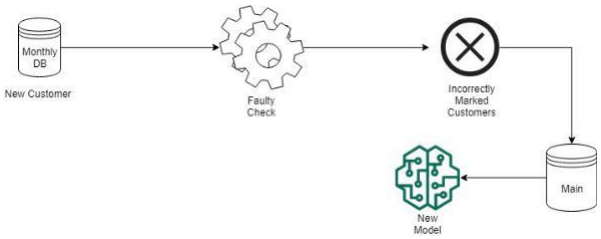

Figure 3. Self-learning and self-development Pipeline

In the self-learning and self-development section, it is checked whether the people who apply for a monthly payment of the loan at the end of the month. People who do not pay their credit here often tend to fraud. The process we do here is sending a label to the customer at the time of application. If we predict the customer that we will fraud, the person is already unable to receive credit, and we cannot track him / her. However, if we predict the person as non-fraud, we observe the person at the end of the month. If the person pays the first loan, it is OK for us, and we observe that our model makes an accurate estimate. However, if the person does not pay his credit, tells the model that made a false prediction and takes this person among the mistakes. And at the end of the month we add the financial data of these errors to the main data set and do another automatic modeling. After this process, our new and current model is activated if desired.

However, the point here is that if the errors are fed continuously, the model will over fit after one stage. In order to overcome this problem, we prevent over fitting by adding the data of the non-faulty person to the main data set as we call it faulty.

2.1.Supervised Learning

Supervised learning accounts for a lot of research activity in machine learning and many supervised learning techniques have found application in the processing of multimedia

content. The defining characteristic of supervised learning is the availability of annotated training data. The name invokes the idea of a supervisor that instructs the learning system on the labels to associate with training examples. Typically these labels are class labels in classification problems. Supervised learning algorithms induce models from these training data and these models can be used to classify other unlabeled data. [4]

In this study, Popular machine learning algorithms (Logistic Regression, Decision Tree, XGBoost, Gradient Boosting, Random Forest, AdaBoost) were pre-modeled. Then, parameter optimization was performed on Gradient Boosting, Random Forest and XGBoost, which had good scores pre-model scoring, and the best parameters were found. And the modeling process is complete. In this study, Gradient Boosting was the model with the best scores.

3. Proposed Methods

Other automl libraries on the market (Scikit-learn AutoML [11], DataRobot.com [10]) find the best features for the given data set and model many different algorithms with these features. Although we have used the same name as the name, we are constantly updating the base model, which is completely different from them, making it suitable for the conditions of the day and ensuring that the model is always alive and effective.

While they use algorithms in all kinds of data sets, we have to change our infrastructure according to the requirements of the field in which we work. For example, frauds in a lending institution can take 1-3 months to mature, whereas in a different institution it can be 3-6 months. And this requires changing the structure of our controls and the Monthly data set.

4. Data Set

In this section, the data set of the Home Credit Default Risk competition published on the platform of data scientists called Kaggle [5] was used.

Here, the data set was home to a lot of information such as financial reports, old application information, credit card information of customers who applied for a home loan of a bank. The schema of the data sets is given in the Figure 4

There were a total of 9 different data sources in the Data Sources. Their names and distributions were as follows;

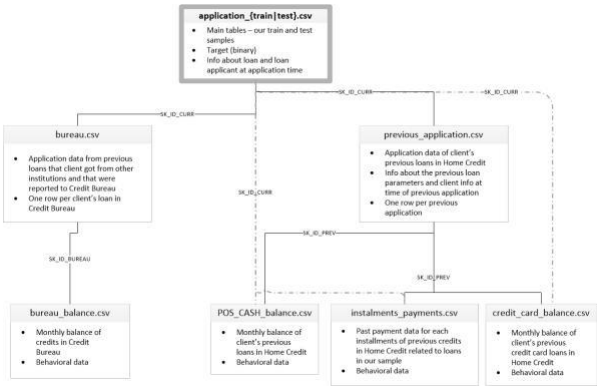

Figure 4. Structure of Data Sources

Application Name train/test

- Static data for all applications. One row represents one loan in our data sample.

Bureau

- The Credit Bureau (for clients who have a loan in our sample). For every loan in our sample, there are as many rows as the number of credits.

Bureau Balance

- Monthly balances of previous credits in Credit Bureau. This table has one row for each month of the past. rows.

Pos Cash Balance

- Monthly balance snapshots of previous POS (point of sales) and cash loans. This table has one row for each month.

Credit Card Balance

- Monthly balance snapshots of the previous credit cards. This table has one row for each month. the table has rows (rows).

Previous Application

- All previous applications for Home Credit loans of clients who have loans in our sample. There is one row for each other.

Installments Payments

- Repayment history for the previously disbursed credits in our sample. There is a) one row for every payment. One row is equivalent to one of the previous installments OR one installment is one of the previous previous payments Home

5. Real World Example

Our aim in this real world example is to determine whether people will pay loans from their financial reports and demographic information. Here, the data of customers who previously fraud is the key point for us. Since our data is labeled, we will use supervised machine learning methods. Our results are shown in Table 1.

Table 1. Results in Supervised Learning Models

Model Name	Precision	F1	Acc.
Logistic Reg.	%62	%54	%66.6
Naive Bayes	%63	%63	%62.9
Decision Tree	%67	%67	%68.2
K-NN	%63	%63	%65.1
AdaBoost	%69	%68	%70.8
Random Forest	%67	%58	%67.7
Gradient Boosting	%70	%69	%71.2
XGBoost	%69	%67	%70.6

6. Conclusion

Aging and overriding of the model is a very expensive and important problem for companies. In this study, a new algorithm has been developed to ensure that the model does not age and is constantly updated. The method can be easily integrated into different sectors. While poor accuracy scores are a major problem in supervised learning based classification studies, this solution aims to improve the model by learning from its own mistakes and adapt it to new customers. As a result of this process, it is planned that companies will make a profit by continuously withdrawing data and eliminating the cost of model renewal. In the experiments conducted in the project, it was observed that the scores improved and the model became more perfect as they learned from their mistakes. In addition, the entire project was developed in the Python scripting language and can be shared and open to partnerships.

References

[1] Gang Kou, Yi Peng, Guoxun Wan (2014).Evaluation of clustering algorithms for financial risk analysis using MCDM methods https://doi.org/10.1016/j.ins.2014.02.137

[2]] Jidong Chen, Ye Tao, Haoran Wang, Tao Chen (2015). Big data based fraud risk management at Alibaba https://doi.org/10.1016/j.jfds.2015.03.001

[3] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki (2010).Boruta A System for Feature Selection Fundamenta Informaticae, DOI 10.3233/FI-2010-288

[4] S Cunningham P., Cord M., Delany S.J. (2008) Supervised Learning. In: Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg

[5] Kaggle, www.kaggle.com

[6] Home Credit Default Risk, kaggle.com/c/home-credit-default-risk

[7] Kredi Kayit Burosu, KKB, www.kkb.com.tr

[8] Azure Machine Learning, https://azure.microsoft.com/en-in/services/machine-learning-service/

[9] Google Prediction API, https://cloud.google.com/prediction/

[10] DataRobot, https://www.datarobot.com/

[11] Sklearn AutoML, https://automl.github.io/auto-sklearn/master/

How Does Resampling Affect the Classification Performance of Support Vector Machines on Imbalanced Churn Data?

Serra Çelik
Informatics Dept.,
Istanbul University
erra.celik@istanbul.edu.tr

Seda Tolun Tayalı
Quantitative Methods Dept.,
School of Business, Istanbul University
stolun@istanbul.edu.tr

Abstract

Churn prediction is an important task for companies. Determining a customer as a probable churner beforehand and keeping her as a result of customer relationship management efforts directly increases the profit of a business. However, churn datasets are imbalanced by nature, which negatively affects the classification performance of algorithms such as the popular Support Vector Machines (SVM). This study handles the classification of imbalanced churn data by applying resampling techniques; Random Under-Sampling, Clustering Based Under-Sampling, Random Over-Sampling, and Synthetic Minority Oversampling Technique as a preprocessing step for to construct a more balanced dataset and investigates their effects on the classification performance of SVM with different kernel functions. The results show that the classification performance of Support Vector Machines improves when resampling is implemented to an imbalanced churn data, especially with Radial Basis Function, and with 5x2 cross validation.

Keywords: under-sampling; over-sampling; churn; binary classification

1. Introduction

In today's competitive business world, building and maintaining successful relationships with customers is an inevitable necessity to survive in a market. Within the context of customer relationship management (CRM), there are several tasks [1] such as; customer profiling, sentiment analysis, churn prediction, and direct marketing and many of these problems are handled by data analytics. The concentration of these problems is understanding the customers and their behavioral patterns.

Customer churn, also known as attrition [2], is an important and expensive problem for businesses. It is explained by the likelihood of customers terminating doing business with a company. The literature has known for a fact that gaining new customers is much more expensive than retaining current ones. Also, within the CRM context, the existing customers are more prone to being in communication with the company and spend more than the new ones. Reichheld and Shefter [3] state from companies' perspective that "*Increasing customer retention rates by 5% increases profits by 25% to 95%".*

In the telecommunications sector, an average customer's monthly spending varies between 20$ and 80$ depending on the country [4]. Customers switching between operators, which is by definition customer churn, is quite common in the telecommunication sector [5].

The annual churn rate in the sector is around 30% in average and acquiring new customers is at least 5 times more expensive than keeping the existing ones [6]. Therefore, losing high number of customers results in high losses for the telecom companies because of lost acquisitions as well as of certain CRM efforts such as reducing the prices to keep the highly potential churners somehow in the company portfolio. This makes churn analysis and consequently the reduction of churn rate a crucial goal for telecom companies.

Customer classification is a good way for realizing churn analysis. Classification is a supervised learning task, where the input data consist of the values each example (customer) takes with respect to the attributes included in the model and the target attribute takes a categorical value refering to the class that the customer belongs to. However, the class imbalance inherited in churn analysis [7-8] as in other CRM datasets -such as fraud detection, response modeling, and credit evaluation- is a reason turning customer classification into a challenging task.

While examining the relation between the data set characteristics and the classification performance, Kwon and Sim [9] mention data imbalance as one of the characteristics that has an effect on the performance of classification algorithms. Imbalance data either distorts the performace of classification or causes overfitting and gives high accuracies.

The literature agrees on two main approaches when dealing with imbalance datasets [10];

1- Data-level approach: Preprocessing steps to balance the classes. Resampling and feature selection techniques are the main data-level approaches.
2- Algorithm-level approach: Modifications of traditional classifiers developed especially for learning from imbalanced datasets such as; one-class learning, cost-sensitive learning, ensemble methods, and recently hybrid approaches.

The literature on churn prediction in the telecommunications sector is more focused on the latter approach, yet there are a few dealing with the effects of the former.

The authors in [11] use both random under- and over-sampling prior to applying several classifying. The study in [12] internally applies four rules generation algorithms based on the rough set theory (RST) with cross validation using six over-sampling techniques on four publicly available dataset. Among the combinations of experiments the ones integrated with over-sampling show better performance as in [13] and [14] that use over-sampling, under-sampling and SMOTE with random forests in the algorithmic level. On the other hand, Verbeke et al [8] examine the effect of over-sampling on the performance of a customer churn prediction model for a telecom dataset and conclude that the dataset structure and the classification technique can change the results completely.

The authors in [15] propose using a genetic programming based approach with Adaboost algorithm and compare results of KNN and random forest for an imbalanced dataset with a 7.3% churner ratio. The study in [16] apply decision trees with Renyi and Tsallis entropies on a dataset that consists 1.96% of the samples as churners.

Ensemble classifers are popular methods in algorithm-level approach and random forest technique is oftenly preferred [17-18]. The study in [19] try to understand the perfomance of negative correlation learning (NCL) ensembles and a multilayer perceptron trained ensembles. Although the focus of the study is on these two techniques, we can easily see that support vector machines (SVM) taken as one of the reference classifiers gives the highest accuracy.

The findings of the study in [20] show that trying to have equally distributed classes is not necessary and an imbalance ratio of 1:3 (minority class: majority class) is a good option for sampling methods. The authors state

that SVM can gain benefits from resampling especially through cost sensitive ones.

There are studies that propose hybrid approaches [21, 22] as well as studies combining feature selection methods either with sampling methods [7, 23] or with ensemble algorithms [24]. The study in [25] handles a huge dataset in the telecom sector. They first applied random forest for feature selection. Then, they proposed an under-sampling approach through clustering and one-sided sampling predetermining the value of k and the imbalance ratio and finally applied decision trees.

The binary classification of imbalanced datasets is a hot topic of data analytics in recent years and churn analysis is one of the application areas. Particularly churn in the telecommunications sector is a specific problem that needs a focused attention since the cost of misclassification is already known. However, the literature does not provide with mature and effective techniques but rather invests in newly developed algorithms [26]. We can safely conclude from the existing research in the field of customer churn prediction that there is not a single model that could give the highest accuracy in all of the cases. Instead, the performance of every algorithm will differ according to the characteristics of the data.

This study examines mainly the effect of a data-level approach on SVM for a binary classification task. The aim is to find the answer to the question "how does resampling affect the classification performance of support vector machines on imbalanced churn data?" for the telecom domain. Section 2 describes the dataset and the methodology followed. Section 3 explains the framework that includes the setup and the evaluation metrics. Section 4 provides the findings and elaborates on them and in the last section, the study concludes with possible research improvements to this study.

2. Modeling Churn with Class Imbalance

This section summarizes the methodology followed in this study.

2.1. Data Set Description

The telecom "churn" dataset is from the UCI Machine Learning Repository [27]. The original dataset has 3333 observations and 21 features in total. Three features -"State", "Area.Code", and "Phone"- are manually eliminated prior to the analyses. The imbalanced ratio is 6:1 [majority class (2850 obs.): minority class (483 obs.)].

2.2. Handling Binary Class Imbalance

The study uses resampling, a data-level approach, to handle the class imbalance problem. For this, the preferred resampling methods are;

Under-sampling techniques

- Random Under-Sampling (RUS): Samples, equal to the number of the minority class or multiples of it, are randomly drawn from the majority class.
- Clustering Based Under-Sampling (CLUSBUS): The training set is divided into groups via clustering techniques. The number of samples to be selected from each cluster belonging to the majority class is calculated and combined with the minority class units. Thus, a new training set is constructed.

Over-sampling techniques

- Random Over-Sampling (ROS): Samples are generated for the minority class so that equal number of majority class or multiples of it is achieved.
- Synthetic Minority Over-Sampling Technique (SMOTE): Samples are generated for the minority class based on the

k-nearest-neighbor method. The parameter k is used to determine the number of samples to be generated for a minority sample.

2.3. Support Vector Machines (SVM) for Classification

Support vector machines is a powerful machine learning method based on structural risk minimization [28], which is proven to show good performance especially for binary classification tasks. Although this performance is not as good for imbalanced classes, motivated by the findings in [20] this study investigates the effect of resampling methods on SVM with different kernels applied.

The customer churn dataset is a binary classification problem, where the customers are coded either as a churner (1) or a non-churner (0). The SVM models with linear, polynomial, sigmoid, and radial basis function kernels are trained once resampling methods are applied and the dataset is partitioned as training and test datasets accordingly.

3. Experimental Framework

The experimental framework of this study is briefly explained in this section. The analyses are implemented in R software. For the SVM parameter optimization, we use grid search with a small size. However, a further fine tuning is not applied in order to prevent the drastic altering of decision boundaries and hence to keep the generalization capability of the models that can resist to minor changes in data.

3.1. Experimental Setup

The study tracks the following setup path:
Step1: The training and test sets are defined in the way that they preserve the specified imbalanced ratio.
Step2: SVM with different kernel functions are trained on the training sets constructed in the previous step.
Step3: Parameters of SVM models are optimized through grid search.
Step4: The final models are selected and used for the evaluation on test sets.

The training and the test sets are split based on 5-fold cross-validation (CV). 5x2 CV method is additionally realized for random under-sampling and random over-sampling.

For the CLUSBUS sampling, Partitioning Around Medoids (PAM) is preferred as the clustering algorithm because of the mixed structure of the dataset features.

3.2. Evaluation Metrics

Evaluation metrics are important when comparing different experimental results. This study uses "Balanced Accuracy", "Sensitivity", and "Lift" measures, which are known to eliminate the drawbacks of traditional metrics for imbalanced classes and expose the correctly classification of the churner class, which is the goal in churn prediction tasks. Table 1 provides the basis for the computation of the evaluation metrics:

Table 1. Confusion matrix

	Classified as	
Actual	Churner	Non-churner
Churner	True Positive (TP)	False Negative (FN)
Non-churner	False Positive (FP)	True Negative (TN)

4. Results and Discussion

To answer our research question, the study includes the analyses conducted with and without resampling techniques. The tables 2-5 show the evaluation metric values of SVM

results with respect to different divisions of training and test sets, imbalanced ratios (IR), and kernel functions. Table 7 shows the best results of each resampling method containing the confusion matrix values, whereas Table 6 provides the base comparison results achieved with SVM without any resampling applied.

The results show that SVM applied with RBF kernel suits better for a churn dataset. RBF generally performs better than SVM with other kernel functions for all the runs in resampling methods.

The random under-sampling method works best with the imbalanced ratios of 3:1 and 4:1 (Table 2 and Table 5). Both RUS and ROS give fairly good results with 5x2 CV applied together with RBF kernel. Yet, the best results obtained with RUS seems better than the ones with ROS (Table 2 and Table 3). The under-sampling methods, RUS and CLUSBUS, give the highest accuracies when the majority class is reduced to three and four times of the minority class. The performance results of ROS are superior than SMOTE's.

Table 2. SVM Results with Random Under-Sampling

		5-fold CV				5x2 fold CV			
IR	**RUS**	**Sig.**	**Lin.**	**RBF**	**Poly.**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.41	0.54	0.50	0.36	0.38	0.47	0.46
	Bal.Acc.	0.67	0.68	0.75	0.72	0.65	0.67	0.72	0.71
	Lift	2.38	2.48	3.29	3.06	2.35	2.49	3.12	3.27
2:1	*Sensitivity*	0.62	0.53	0.54	0.66	0.42	0.46	0.44	0.61
	Bal.Acc.	0.75	0.72	0.74	0.79	0.68	0.70	0.70	0.78
	Lift	3.77	3.23	3.28	4.06	3.04	3.32	3.14	4.30
3:1	*Sensitivity*	0.44	0.44	***0.88***	0.74	0.28	0.48	***0.82***	0.65
	Bal.Acc.	0.65	0.65	***0.90***	0.83	0.62	0.69	***0.87***	0.79
	Lift	2.66	2.66	***5.39***	4.51	3.73	3.37	***5.32***	4.59
4:1	*Sensitivity*	0.45	0.44	0.76	*0.80*	0.48	0.50	***0.81***	0.76
	Bal.Acc.	0.66	0.65	0.84	*0.86*	0.68	0.69	***0.87***	0.84
	Lift	2.77	2.66	4.62	*4.88*	3.51	3.63	***5.90***	5.41

Table 3. SVM Results with Random Over-Sampling

		5-fold CV				5x2 fold CV			
IR	**ROS**	**Sig.**	**Lin.**	**RBF**	**Poly.**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.42	0.43	0.66	0.61	0.41	0.41	*0.70*	0.56
	Bal.Acc.	0.68	0.69	0.79	0.76	0.67	0.67	*0.82*	0.73
	Lift	2.54	2.62	4.06	3.72	2.58	2.56	*5.08*	3.49
2:1	*Sensitivity*	0.53	0.48	0.70	0.61	0.53	0.50	*0.73*	0.66
	Bal.Acc.	0.72	0.70	0.80	0.76	0.71	0.70	*0.83*	0.80
	Lift	3.21	2.97	4.26	3.75	3.29	3.11	*5.29*	4.76
3:1	*Sensitivity*	0.57	0.55	0.80	0.72	0.60	0.54	***0.76***	0.64
	Bal.Acc.	0.72	0.72	0.85	0.82	0.74	0.70	***0.84***	0.79
	Lift	3.48	3.35	4.91	4.43	3.98	3.33	***5.50***	4.99

Table 4. SVM Results with SMOTE 5 fold CV

		5 fold CV			
IR	**SMOTE**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.40	0.46	0.46
	Bal.Acc.	0.67	0.68	0.69	0.70
	Lift	2.40	2.47	2.80	2.84
1.5:1	*Sensitivity*	0.44	0.44	0.57	0.50
	Bal.Acc.	0.65	0.65	0.74	0.71
	Lift	2.66	2.66	3.49	3.08
2:1	*Sensitivity*	0.44	0.44	**0.62**	0.51
	Bal.Acc.	0.65	0.65	**0.77**	0.72
	Lift	2.66	2.66	**3.78**	3.13

The analysis done without any resampling uses the imbalanced ratio of the dataset as it is (6:1). Looking at the confusion matrix values (Table 6), it is obvious that SVM with sigmoid and linear kernel perform extremely poor in terms of detecting the churner class. This is also the reason why some metrics turn out to be non-available. On the other hand, the RBF results and give high sensitivity ratios, referring to the TP rates. An important error to be minimized for churn analysis is the FNs, referring to the actual churners who are predicted as non-churners. Polynomial kernel result seems to be more successful in terms of FN rate.

Comparing the results in Table 6 with the results in Table 7 that summarizes the best performances with resampling methods, we can say that the performance of SVM is improved when resampling is applied. The results in Table 7 are sorted based on the lift value. Overall, the values show that the resampling results improve the classification performance of SVM on imbalanced churn data the most when used with radial basis function (RBF) and with 5x2 CV.

Table 5. SVM Results with CLUSBUS

		5 fold CV			
IR	**CLUSBUS**	**Sig.**	**Lin.**	**RBF**	**Poly.**
1:1	*Sensitivity*	0.39	0.41	0.53	0.49
	Bal.Acc.	0.67	0.68	0.75	0.72
	Lift	2.38	2.48	3.21	3.02
2:1	*Sensitivity*	0.47	0.50	*0.75*	*0.68*
	Bal.Acc.	0.69	0.71	*0.85*	*0.81*
	Lift	2.86	3.06	*4.56*	*4.17*
3:1	*Sensitivity*	0.43	0.46	*0.73*	*0.72*
	Bal.Acc.	0.65	0.67	*0.83*	*0.83*
	Lift	2.64	2.82	*4.45*	*4.43*
4:1	*Sensitivity*	0.47	0.51	***0.86***	0.76
	Bal.Acc.	0.67	0.69	***0.89***	0.84
	Lift	2.87	3.12	***5.26***	4.62

Table 6. SVM Results without Using Resampling

		IR	TN	FP	FN	TP	Sensitivity	Bal. Accuracy	Lift
5-fold CV	**Sigmoid**	6:1	558	109	0	0	NA	NA	NA
	Linear	6:1	558	109	0	0	NA	NA	NA
	RBF	6:1	547	46	11	63	0.85	0.89	5.21
	Polynomial	6:1	549	48	9	61	0.87	0.90	5.33

Table 7. Best SVM Results with Resampling Methods

		IR	TN	FP	FN	TP	Sensitivity	Bal. Accuracy	Lift
5x2 fold CV	**RUS+RBF**	4:1	1412	115	27	113	0.81	0.87	5.90
5x2 fold CV	**ROS+RBF**	3:1	1399	112	37	119	0.76	0.84	5.50
5-fold CV	**RUS+RBF**	3:1	549	43	9	66	0.88	0.90	5.39
5-fold CV	**CLUSBUS+RBF**	4:1	548	48	10	61	0.86	0.89	5.26

5. Conclusion

The purpose of this study is to investigate the effects of resampling techniques on Support Vector Machines for imbalanced customer churn data. We can conclude that resampling techniques, especially random under-sampling improves the classification performance of SVM. Also, support vector machines with RBF yields a better performance than the other kernel functions. The study handles the imbalanced dataset classification from a data-level approach. As we now see how these resampling techniques affect the performance of SVM for a telecom customer churn dataset, other under-resampling techniques could also be investigated. A further improvement to this study would be to formulize the problem as a multi-classification task and see whether resampling improves the classification performance in such setting.

References

[1] Krishna, G. and Ravi V. (2016). Evolutionary computing applied to customer relationship management: A survey. *Engineering Applications of Artificial Intelligence*, 56 (November), 30-59.

[2] Singh, H. and Samalia H.V. (2014). A business intelligence perspective for churn management. *Procedia – Social and Behavioral Sciences*, 109 (January), 51-56.

[3] Reichheld, F. and Schefter P. (2000). E-loyalty: Your secret weapon on the web. *Harvard Business Review*, 78 (July-August), 105-113.

[4] Mattison, R. (2005). The telco churn management handbook, null edition, Lulu.com, XiT Press, Oakwood Hills, Illinois, USA.

[5] Zhang, Y., et al. (2011). Behavior-based telecommunication churn prediction with neural network approach. *Proceedings – 2011 International Symposium on Computer Science and Society, ISCCS 2011*, 307-310.

[6] Lu, J. (2002). Predicting customer churn in the telecommunications industry- An application of survival analysis modeling using SAS. *SAS User Group International (SUG127) Online Proceedings*, 114-127.

[7] Idris, A., Riswan M. and Khan A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Computers and Electrical Engineering*, 38 (6), 1808-1819.

[8] Verbeke, W., et al. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218 (1), 211-229.

[9] Kwon, O. and Sim J.M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40* (5), 1847-1857.

[10] Ali, A., Shamsuddin S.M. and Ralescu A.L. (2015). Classification with class imbalance problem: A review. *International Journal Advances in Soft Computing and its Applications*, 7 (3), 176-204.

[11] Qureshi, S.A., et al. (2013). Telecommunication subscribers' churn prediction model using machine learning. *Eighth International Conference on Digital Information Management (ICDIM 2013)*. IEEE, 131-136.

[12] Amin, A., et al. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, (4), 7940-7957.

[13] Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. *Artificial Intelligence Research*, 6 (2), 93-99.

[14] Hanif, A. and Azhar N. (2017). Resolving class imbalance and feature selection in customer churn dataset. In: *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 82-86.

[15] Idris, A., Khan A. and Lee Y.S. (2012). Genetic programming and adaboosting based churn prediction for telecom. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 1328-1332.

[16] Gajowniczek, K., Ząbkowski T. and Orłowski, A. (2015). Comparison of decision trees with Rényi and Tsallis entropy applied for imbalanced churn dataset. In: *2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE,* 39-44.

[17] De Bock, K.W. and Van Den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, 38 (10), 12293-12301.

[18] Xie, Y. et al. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36 (3), 5445-5449.

[19] Rodan, A. et al. (2015). Negative correlation learning for customer churn prediction: A comparison study. *The Scientific World Journal*, 1-7. http://dx.doi.org/10.1155/2015/473283

[20] Zhu, B. et al. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, 69 (1), 49-65.

[21] Idris, A. and Khan A. (2016). Churn prediction system for telecom using filter–wrapper and ensemble classification. *The Computer Journal*, 60 (3), 410-430.

[22] Ahmed, A.A. and Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, 18 (3), 215-220.

[23] Kim, Y. (2006). Toward a successful CRM: variable selection, sampling, and ensemble. *Decision Support Systems*, 41 (2), 542-553.

[24] Idris, A., Khan, A. and Lee, Y.S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied intelligence*, *39* (3), 659-672.

[25] Li, H. et al. (2016). Supervised massive data analysis for telecommunication customer churn prediction. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom SustainCom), IEEE,* 163-169.

[26] Haixiang, G. et al. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239.

[27] Blake, C.L. and Merz, C.J. (1998). Churn Data Set, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Department of Information and Computer Science, Irvine, CA.

[28] Vapnik, V.N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10* (5), 988-999.

Do Confidence Indicators Have Impact on Macro-financial Indicators? Analysis on the Financial Services and Real Sector Confidence Indexes: Evidence from Turkey

Esra N. KILCI
Istanbul Arel University,
esrakilci@arel.edu.tr

Abstract

Confidence has played a key role in the recovery of macro-financial outlook of Turkey after the 2000-01 Economic crisis period until the last few-years in which there has been felt deterioration in confidence indicators. Therefore, this study analyses the impact of financial services and real sector confidence indexes on some macroeconomic and financial indicators such as industrial production index, inflation, stock market index, foreign exchange rates and interest rates in Turkey for the period of 2012:05-2019:05. In this study, the unit root properties of the series are tested by using ADF and Fourier ADF unit root test and the causality relationships between the series are investigated by employing Fourier Toda Yamamoto causality test. The results support the impact of confidence indicators on macro-financial indicators as stock market indices and inflation.

Keywords: financial sector confidence index; real sector confidence index; stock market; foreign exchange rates

JEL Codes: C10, E70, G40

1. Introduction

Turkish economy, which launched a strong recovery period towards minimizing the adverse effects of 2000-01 Economic crisis and ensuring fiscal and financial stability, has experienced a strong growth period with the contribution of the positive global economic atmosphere in the post-crisis period. It is obvious that confidence in the real economy and financial system has played a major role in this development. Fiscal discipline, structural reforms aimed at achieving macroeconomic stability and tight monetary and fiscal policies have contributed to the significant increase in confidence and stability in Turkish economy. Therefore, the fundamental indicators have shown a a significant improvement in the post-crisis period. Given the importance of confidence for the real economy and financial markets, it is clear that the deterioration tendency of the confidince indicators negatively affects the macroeconomic and financial fundamentals such as economic growth, employment, foreign exchange rates and interest rates by affecting production, spending and investment decisions of economic agents.

The importance of psychological factors as sentiments and expactations was emphasized in the analyses by a wide range of economists such as Keynes (1936) focusing on consumer and investor sentiments, Katona (1951) stating the importance of psychological factors, Cass and Shell (1983), Akerlof and Shiller (2010)

drawing attention to the government's role on manipulating animal spirits, Farmer (2013), De Grauwe and Ji (2016), Bacchetta and Van Wincoop (2013) and Acharya et al (2017).

On the other hand, since confidence cannot be observed or measured directly, any assessment of confidence must rely on indicators which are often partial, qualitative and subject to various interpretations. In this regard, confidence indices, consistent measures of confidence, are implemented based on surveys to measure how economic decision makers respond to the economic developments and in order to express the relationship between expectations and macroeconomic or financial variables. Confidence indicators appear to provide a good picture of major cyclical movements in output; if available in advance of "hard" output data, they may help detect significant acceleration or deceleration in output growth when large changes in confidence are observed (Santero and Westerlund, 1996).

In this study, our question is, whether the confidence indexes, which are simply linked with the sentiments and expectations, help us predict the changes in the macro-financial indicators or not. In this context, we focus on the impact of real sector confidence index and financial services confidence index on some macro-financial indicators. The objective of this study is to evaluate the empirical validity of the real sector and financial services confidence indices in anticipating the evolution of economic activity by considering monthly data from 2012:05 to 2019:05. There are very few studies that concentrate on emerging market economies. To our knowledge, this paper is the first to analyse systematically the relationships from financial services and real sector confidence indexes to macro-financial variables in Turkey by modelling the variables under the Fourier framework. In this way, we appropriately take into consideration multiple structural breaks without a need of the number, form or date of these breaks through the Fourier approach.

The study is organised as follows. After presenting literature review in section 2, the estimation results acquired by employing Fourier Toda Yamamoto causality test are reported in section 3. Finallly, section 4 concludes.

2. Literature Review

Expectations and uncertainty about future economic and financial outlook, which are of key importance particularly according to the behavioral finance theory, are measured through confidence indexes such as consumer confidence index, real sector confidence index and financial services confidence index, which are regularly followed by central banks, governments and other private institutions. Therefore, in the academic literature, there has been an increasing academic literature on their use in monitoring current economic situation and predicting economic developments. On the other hand, it is seen that most of these studies focus on consumer confidence indices and some of these deal with business confidence. Santero and Westerlund (1996), Souleles (2001), Jansen and Nahuis (2003), Utaka (2003), Vuchelen (2004), Ludvigson (2004), Afshar et al (2007), Mariana (2012), Van Aarle and Kappler (2012), Nowzohour and Stracca (2017) focus on confidence indicators and analyze the relationship between confidence indices and some macroeconomic and financial variables. Unfortunately, it is seen there are a few studies in Turkey which analyze the impact of confidence indicators on macro-financial indicators. In this context, Oral (2005), Kandır (2006), Ozsagir (2007), Çelik and Özerkek (2009), Aktas and Akdag (2013), Arisoy (2012) and Iskenderoglu and Akdag (2017) empirically analyzed the relationship between measures of confidence and macro-financial variables.

3. Empirical Analysis

3.1. Data

Confidence based on expectations is one of the key factors determining agents' decisions. In analyzing the impact of confidence indicators on some macroeconomic fundemantals such as indsutrial production index and inflation rate and some domestic volatility indicators of financial variables such as foreign exchange rates, interest rates and stock market indices, we use monthly data belonged to the financial sector confidence index and the real sector confidence index. In Turkey, real sector and financial services confidence surveys have been carried out regularly by the Central Bank of the Republic of Turkey since 2007:05 and 2012:05, respectively. The surveys are carried out each month via face-to-face interviews with 1.000 individuals from all over the country. The individuals are selected via stratified random sampling. The strata are formed based on income, economic activity, education, age and gender. In this study, our data set covers the period 2012:05-2019:05. This sample seems smaller when compared to the samples used in the studies on the other developed economies like U.S.A and some Euro-Area countries including Germany, U.K and France. However, this is the largest data set that can be prepared as there is no financial services confidence data for Turkey before 2012. Despite of this limitation, we think that our data set meets minimum requirements to carry out our analysis.

Table 1. List of Variables

Independent Variables	Measure	Abbreviation	
Confidence Indicator 1	Financial Services Confidence Index	FSCI	
Confidence Indicator 2	Real Sector Confidence Index	RSCI	
Dependent Variables	**Measure**	**Abbreviation**	**Expected Relations-hip**
Production	Industrial Production Index	IPI	(+)
Inflation rate	Consumer Inflation Index	CPI	(+)
Foreign Exchange Rate	USD/TRY (average)	USD	(+)
Stock Market Indıces	BIST100 (average)	BIST	(+)
Interest Rates	Interbank O/N Rate (average)	O/N	(+)

As seen at Table 1, in our analysis, we employ real sector and financial services confidence indices as the independent variables while industrial production index, consumer inflation index, stock market index, interest rates and foreign exchange rates as dependent variables. As measures of real sector and financial services confidence, we benefit from the surveys of the CBRT as we mentioned above.

3.2. Econometric Tests

3.2.1 Stationarity, Cointegration and Causality

In order to assess the forecasting ability of the confidence, at first, we try to get adequate information on the stationarity properties of the variables being used in analysis by employing ADF and Fourier ADF unit root tests. Enders and Jones (2012) propose a new unit-root test by using Fourier function in the deterministic term in a Dickey-Fuller type

regression framework that can complement the Fourier LM and DF-GLS unit root tests. It is seen that this test does have good size and power properties. The Fourier ADF unit root test alllows estimation of multiple structural changes with Fourier functions in testing the stationary of series. Contrary to many other methods, it is not necessary to know the number, form or date of the structural changes.

Table 2. ADF and Fourier ADF Unit Root Test Results (T=85)

Vari-ables	Freq.	MinSSR	Fourier ADF Test-Statistic	ADF Test-Statistic	F-Statistic
FSCI	3	5895.048	-3.87	-3.23	3.24
RSCI	2	954.5519	-3.55	-3.54	1.61
PMI	1	5905.716	-6.24	-1.82	9.42**
INF	5	790.0850	3.82	3.05	3.01
DIFINF	1	773.5279	-6.72	-2.86	2.55
LOG BIST	1	0.250497	-3.91	-2.72	3.77
EUR	5	3.235448	1.33	1.18	1.92
DIFEUR	5	2.979959	-6.39	-8.25	1.01
USD	5	2.384208	1.62	1.21	2.47
DIFUSD	5	2.129045	-5.94	-4.24	1.10
INTRATE	5	140.5274	-0.06	-0.16	2.59
DIFINTRATE	5	115.8366	-5.52	-5.90	1.24
Without a linear trend	ADF Critical Values		Fourier ADF Critical Values (1-2-3-5)		F ADF F-Statistic
1%	-3,51		-4,42 -3,97 -3,77	-3,58	10,35
5%	-2,89		-3,81 -3,27 -3,07	-2,93	7,58
10%	-2,58		-3,49 -2,91 -2,71	-2,60	6,35

Table 2 shows the stationary results of the variables. In the second stage, it is employed Fourier Toda Yamamoto Causality Test proposed by Nazlioglu et al (2016) in order to investigate the causal linkages from FSCI and RSCI to the macro-financial indicators. Since the linkages between the variables have been subjected to gradual shifts and linear specifications are mostly inappropriate to capture the relationships, econometric examinations are not generally direct and simple so traditional procedures which look for sudden shifts become insufficient in capturing gradually emerging structural changes. Therefore, Nazlioglu et al (2016) modify the Toda-Yamamoto (1995) Granger Causality approach by implanting a Fourier approximation to be able to explain gradual or smooth structural shifts. There is no need for a prior knowledge concerning the number, dates and form of breaks when used the Fourier approximation. Their study based on the analysis proposed by Ender and Jones (2016) in which a Fourier approximation is employed by using a limited number of low frequency components in an effort to clarify determination of the form of breaks and estimation of the number and dates of shifts in a VAR framework.

Table 3. Fourier Toda Yamamato Causality Test Results

Relationship	Frequency	Wald-Stat	Asymptotic* p-value	Bootstrap p p-value
FSCI→ BIST	2	17.122	0.017**	0.017
FSCI→ INTRATE	3	2.194	0.334	0.336
FSCI→ USD	3	0.911	0.823	0.828
FSCI→ EUR	3	2.070	0.558	0.563
RSCI→ PMI	1	0.338	0.561	0.575
RSCI→ INF	1	36.676	0.000*	0.003
RSCI→ INTRATE	1	11.609	0.312	0.340

Notes: → denotes to causality. Optimal k (frequency) and p (lag) are determined by Akaike information criterion. Bootstrap p-values are based on 1000 replications. *, **, and *** denote %1, %5, and %10 levels of statistical significance, respectively. Because n>50 in this study, we will take asymptotic p-value in comparison.

Table 3 shows the results of Fourier Toda-Yamamoto Causality Test. The findings supporting the impact of FSCI on BIST and the impact of RSCI on INF, signal that confidence indicators are associated with changes in macro-financial indicators such as inflation and stock market indices.

4. Conclusion

In this study, we look for the answer of this questions: does real sector and financial services confidence indices have impact on industrial production index, inflation rate, stock market index, interest rates and foreign exchange rates. In answering this question, we employ Fourier ADF and ADF unit root tests, Fourier Toda Yamamoto causality test. The analysis has been carried out for the period in which there has been a drop in the confidence indices. Our findings indicate that confidence indicaes have explanatory power on macro-financial outlook of Turkey in the period of 2012-2019. Our findings are in accordance with most of the studies in the literature, which indicate that the confidence indices do have impacts on macro-financial outlook. Therefore, it could be said that FSCI and RSCI can be used to predict some of the macro-financial indicators like stock market indices and inflation rate.

References

Afshar T., Arabian, G. & Zomorrodian, R. (2007). Stock return, Consumer Confidence, Purchasing Manager's Index and Economic Fluctuations, Journal of Business & Economics Research, 5(8), 97-106.

Arisoy, I. (2012). Türkiye Ekonomisinde İktisadi Güven Endeksleri ve Seçilmiş Makro Değişkenler Arasındaki İlişkilerin VAR Analizi, Maliye Dergisi, 16, 304-315.

Bachmann, R., Elstner, S., & Sims, E. R. (2013). Uncertainty and Economic Activity: Evidence from Business Survey Data. American Economic Journal: Macroeconomics, 5 (2). 217-249.

Baumohl, B. (2012). *The Secrets of Economic Indicators: Hidden Clues to Future Economic Trends and Investment Opportunities*. New Jersey: Financial Times Press.

CBRT (2019) Financial Services Statistics and Financial Services Confidence Index.

Dion, D. P. (2006). Does Consumer Confidence Forecast Household Spending? The Euro Area Case. MPRA Papers, No: 911. Available at: https://mpra.ub.uni-muenchen.de/911/ Access Date: 13.05.2019.

Iskenderoglu, O & Akdag, S. (2017). Investigation of the Validity of Financial Services Confidence Index: The Case of Turkey, *Uluslararası Ekonomik Araştırmalar Dergisi*, 3(4), 625-633.

Jansen, W. J. & Nahuis, N. (2003). The Stock Market and Consumer Confidence: European Evidence, *Economics Letters*, 79(1), 89-98.

Kandir, S. Y. (2006), Tüketici Güveni Ve Hisse Senedi Getirileri İlişkisi: IMKB Mali Sektör Şirketleri Üzerinde Bir Uygulama, *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 15(2), 217-230.

Karasoy, H. G (2015). Consumer Confidence Indices and Financial Volatility, CBRT Research Notes in Economics, No: 2015-16, Available at: http://www.tcmb.gov.tr/wps/wcm/connect/4d134ad3-14bc-462b-8acd-fd70c5878441/en1516eng.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-4d134ad3-14bc-462b-8acd-fd70c5878441-m3fw5mC, Access Date: 05.05.2019.

Karasoy, H. G. & Yunculer, C. (2015). The Explanatory Power and the Forecast Performance of Consumer Confidence Indices for Private Consumption Growth in Turkey, CBRT Working Papers, No: 15/19. Available at: http://www.tcmb.gov.tr/wps/wcm/connect/13211994-0096-48be-bef1-61996d585593/wp1519.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-13211994-0096-48be-

bef1-61996d585593-m3fw6dX, Access date: 05.05.2019.

Kuzmanović, M. & Sanfey, P. (2012). Can consumer confidence data predict real variables? Evidence from Croatia, European Bank for Reconstruction and Development, Working Paper: 151. Available at: https://www.ebrd.com/downloads/research/economics/workingpapers/wp0151.pdf, Access date: 07.05.2019.

Ludvigson S. C (2004). Consumer Confidence and Consumer Spending, *Journal of Economic Perspectives*, 18(2):29-50.

Nowzohour, L. & Stracca, L. (2017). More than a Feeling: Confidence, Uncertainty and Macroeconomic Fluctuations, ECB Working Paper Series, No: 2100. Available at: https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2100.en.pdf Access Date: 12.05.2019.

OECD (2019), Business Confidence Index (BCI) (indicator). doi: 10.1787/3092dc4f-en Accessed date: 25.05.2019.

Oral, E. & Ece, D. & Hamsici, T. (2005). Building Up a Real Sector Confidence Index for Turkey. *Central Bank Review*, 1, 23-54.

Ozasagir, A. (2007). Ekonomide Güven Faktörü. *Elektronik Sosyal Bilimler Dergisi*, 6 (20), 46-62.

Souleles, N. S. (2001). Consumer Sentiment: Its Rationality and Usefulness in Forecasting Expenditure - Evidence from the Michigan Micro Data, NBER Working Papers, No: W8410. Available at SSRN: https://ssrn.com/abstract=278761. Access Date: 06.05.2019

Santero, T. and N. Westerlund (1996). Confidence Indicators and Their Relationship to Changes in Economic Activity", OECD Economics Department Working Papers, No. 170, OECDPublishing, Paris, https://doi.org/10.1787/537052766455. Access Date: 10.05.2019.

Utaka, A. (2003). Confidence and the Real Economy - the Japanese Case, Applied Economics 35, 337-342.

Van aarle, B. and Kappler, M. (2012). Economic Sentiment Shocks and Fluctuations in Economic Activity in the Euro Area and the USA, Intereconomics, 47(1), 44-51.

Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases

Assoc. Prof. Dr. Nurdan Çolakoğlu
Istanbul Arel University
nurdancolakoglu@arel.edu.tr

Berke Akkaya
Istanbul Arel University
berkeakkaya@arel.edu.tr

Abstract

In recent years, one of the most common problems in estimation and classification problems has been multi-class classification problems, leading to that several machine learning algorithms have been used to solve such problems.

Today, heart diseases are the cause of the most deaths in the world. Since the early diagnosis of heart diseases plays an important role for the survival of the individual, this study focuses on classification algorithms which are capable of to do multi-class classification like Logistic Regression, Gaussian Naïve Bayes, k-Nearest Neighbors, Support Vector Machines, Multilayer Perceptron, CART, Random Forest, Gradient Boosting Machine, Extreme Gradient Boosting. Those algorithms have been applied to a dataset containing 2126 CTG (Cardiotocogram) reports which are divided into three classes as "Normal", "Suspect" and "Pathological". The classification success of these multi-class classification algorithms has been compared.

Keywords: Multi-class Classification; Early Diagnose; Heart Diseases

1. Introduction

Today, one of the biggest challenges in the healthcare sector is to classify the heart diseases in time, to diagnose early and to start treatment immediately. The techniques used in the diagnosis are improving, the pace of the progress has not been enough. In this study, classification algorithms that can be used in the diagnosis of heart diseases, one of the leading causes of death in the world, are compared.

According to the Turkish Society of Cardiology; heart disease refers to the inability of the heart to send the necessary and sufficient blood to the tissues and organs as a result of a decrease in the performance of the heart or to have any problems in the functioning of the heart [1]. According to World Health Organization (WHO), the causes of 31% of the deaths in the world in 2018, were heart diseases [2]. There are many external factors causing the heart diseases such as smoking, alcohol, stress. Besides, heart diseases may be inherited genetically. The count of the heart beats per minute, the acceleration and deceleration of the heart rate are known important parameters in the diagnosis of heart diseases [3]. For this reason, early diagnosis and classification of the problems in heartbeats and heart rate obtained from CTG reports seems to be very important for the continuity of the individual's life.

There are many publications in the literature on the comparison of classification problems. In these publications, binary classification problems are generally compared, and similar algorithms are used to solve. In this study, several types of multi class classification algorithms like Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Support Vector Machines, Multilayer Perceptron, CART, Random Forest, Gradient Boosting Machine has been used to enrich the possibilities for resolution and compare them to find the best performing algorithm. The CTG data set will be processed by using these algorithms.

2. Literature Review

In the literature, classification algorithms are frequently used in the healthcare sector. Since the heart disease is a very serious factor for human life, there are many studies that used classification techniques for early diagnosis of heart disease. One of the first studies on this area is Cardiovascular Risk Prediction Model article in 2000 by Colombet et al. In this study, Logistic Regression, Multilayer Perceptron and CART algorithms were applied on dataset obtained from INDANA Database. As a result of this study, Multilayer Perceptron is the best predictor leading with 76% in accuracy score, while CART is 69.1% and Logistic Regression is 65.9% [4].

Some of the recent researches on this subject are as follows: In the study of Priyanka and Rami-Kumar in 2017 which is titled as "Usage of data mining techniques in predicting the heart diseases", Naive Bayes and Decision Tree algorithms are used, compared and the most reliable result is the Decision Tree Algorithm with 98.03% accuracy [5]. In 2018, Latha and Jeeva applied Bayes Net, Naive Bayes, Random Forest, C4.5, Multilayer Perceptron, PART algorithms to classify Cleveland heart dataset on UCI machine learning data involving 303 individuals and 14 attributes. The best results are obtained by blending the algorithms with ensemble techniques, and achieves 85.48% accuracy with Naive Bayes, Bayes Net, Random Forest and Multilayer Perceptron approach [6]. In another study, Ul Haq et al. proposes a hybrid heart disease prediction system that works with machine learning algorithms. The data used in this study includes demographic variables such as age and gender, as well as variables related to body functions. The most effective variable properties are determined, and the best results are obtained by using k-Nearest Neighbors, Artificial Neural Network, Support Vector Machines, Naive Bayes and Decision Tree algorithms [7]. In the study conducted by Beyene and Kamat in 2018, variables such as smoking status, alcohol use status, obesity, diabetes and hygiene are discussed in addition to demographic variables. This study, which does not include any variable related to body functions and uses Naive Bayes, Support Vector Machines and J48 algorithms, compares the results obtained across these algorithms, and use the output of the algorithm giving the highest consistency [8].

3. Materials and Methods

3.1. Dataset Description

A dataset provided by UCI Machine Learning Repository, which consists of Measurements of Heart Rate (HR) and Contraction (UC) Features on cardiotocograms classified by medical experts, has been used. To prepare the dataset, 2126 cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured by UCI. The CTGs were also classified by three medical experts and a consensus classification label assigned to each of them. [9]

The variables of the dataset are listed in the Table 1;

Table 1. The variables of the dataset

Name	Description	Values (Per Minute)
LB	Baseline Value, beats of heart (Taken from SisPorto)	Between 106-160
AC	Accelerations (Taken from SisPorto)	Between 0-0,019
FM	Heart Ventricle Movement (Taken from SisPorto)	Between 0-0,481
UC	Heart Ventricle Contractions (Taken from SisPorto)	Between 0-0,015
DL	Light Decelerations	Between 0-0,015
DS	Severe Decelerations	Between 0-0,001
NSP	Class(Dependent Variable)	Normal=1; Suspect=2; Pathological=3

3.2. Classification Algorithms

In this section, classification techniques used in the study will be discussed. Classification is a predictive data mining model. In such models, the dependent variable carries a class value. This class value of the dependent variable depends on the values of the independent variables. It is used in many areas such as pattern recognition, diagnosis of diseases, fraud detection. [10]

i.Support Vector Machines (SVM)

Support Vector Machines (SVM) is a classification algorithm based on statistical learning theory. SVM was initially used to classify binary data. However, it can now be used for the classification of multiclass and non-linear data. The infrastructure of SVM is based on the prediction of the decision function that can distinguish classes from each other [11].

SVMs have been successfully implemented in many areas. SVMs have also been used to classify heart diseases, but most of these uses have been used in binary predictions [12] [13] [14]. The advantages and the disadvantages of SVM has given in Table 2.

Table 2. Support Vector Machines advantages and disadvantages

Support Vector Machines (SVM) [15] [16] [17]	
Advantages	Disadvantages
• Gives good results even if there is not enough information about the data. Also works well with unstructured data. • Solves complex problems with a convenient kernel solution function. • Relatively good scaling of high-dimensional data.	• It is difficult to choose the appropriate kernel solution function. • Training time is long when using large data sets. • It may be difficult to interpret and understand because of problems caused by personal factors and the weights of variables. • The weights of the variables are not constant, thus the contribution of each variable to the output is variant.

3.2.2. Gaussian Naïve Bayes (GNB)

This algorithm originated from Bayes' Theorem and it argues that all properties are independent variables and that the changes in these variables do not affect each other. This algorithm is also useful when classifying large data sets. Using the conditional independence rule, the algorithm assumes that one attribute is independent of the other attributes in the classification. The algorithm is effective in the classification as it computes with the minimum error rate [18] [19] [20]. Table 3 presents the advantages and disadvantages of this algorithm.

Table 3. Gaussian Naïve Bayes advantages and disadvantages

Gaussian Naïve Bayes (GNB) [21] [22]	
Advantages	Disadvantages
• Fast and flexible model gives highly reliable results. • Works well with large data. • There is no need to spend much time for training. • Provides better grading performance by eliminating insignificant specifications.	• Large data records are required to achieve a good result. • Shows lower performance than the other classifiers according to the type of problem.

3.2.3. k-Nearest Neighbors (kNN)

It is the simplest of the classification algorithms. The algorithm takes advantage of previously classified data. Attribute vectors must be created to implement the algorithm. The k parameter in this algorithm specifies the number of neighborhoods. According to the specified k parameter, each data is assigned to the nearest neighbor [23] [24]. Table 4 presents the advantages and the disadvantages of the kNN algorithm.

Table 4. k-Nearest Neighbors advantages and disadvantages

k-Nearest Neighbors (kNN) [25] [26] [27]	
Advantages	Disadvantages
• Easy to implement and understand because it does not include any assumptions. It is also heuristic. • Responds quickly to changes in input during real-time usage. • Can be easily applied to multi-class classification problems.	• As the data set grows, the speed of the algorithm decreases. Also, it becomes difficult to reach the output when the number of variables increases. • Not capable of dealing with missing values and it is affected by outliers. • To work properly, the variable features must be expressed in the same scale.

3.2.4. Multilayer Perceptron (MLP)

This algorithm is an artificial neural network with backward propagation. It consists input layers, hidden layers, and an output layer. Layers, except the input layer, use a non-linear activation function. The activation function provides a curvilinear match between the input and output. This algorithm allows learning by changing the weights of neurons that it assigns to itself [28] [29]. The advantages and the disadvantages of MLP has given in Table 5.

Table 5. Multilayer Perceptron Advantages and Disadvantages

Multilayer Perceptron (MLP) [30]	
Advantages	Disadvantages
• Can be applied to complex non-linear problems. • Works well with large input data. • Provides quick predictions after training. • The same accuracy ratio can be achieved even with smaller data.	• It is not known to what extent each independent variable is affected by the dependent variable. Computations are difficult and time consuming. • The proper functioning of the model depends on the quality of the training data. If the model does not work properly, generalization problems arise.

3.2.5. Random Forest (RF)

Random forest is a simple algorithm that can be used for classification. It is also a preferred decision tree algorithm because it is an algorithm that does not have the problem of overfitting within decision trees. When training the algorithm, it creates many decision trees from the subset of the problem and makes each tree estimate. Classification is made by selecting the most votes among these estimates [31] [32]. Table 6 presents the advantages and the disadvantages of the Random Forest algorithm.

Table 6. Random Forest advantages and disadvantages

Random Forest (RF) [33] [34]	
Advantages	Disadvantages
• Reduces the chance of encountering a classifier that does not perform well due to the relationship between training and testing data. • Extremely flexible and have very high accuracy. Also maintains accuracy even when a large proportion of the data are missing.	• Difficult to understand and interpret visually • Much harder and time-consuming to construct than decision trees. • Requires a lot of computation and the algorithm itself, is less heuristic.

3.2.6. CART (Classification and Regression Trees)

The CART algorithm is a predictive model based on the ability of explaining the result of how one variable affect to other variable values in classification. In this algorithm, which is a decision tree algorithm; each branch expresses the values of the variables; each node contains the result of the prediction. CART trees are formed by information gain that expresses the decrease in entropy, which is the degree of randomness of data [35] [36] [37]. The advantages and the disadvantages of CART has given in Table 7.

Table 7. CART Advantages and Disadvantages

CART (Classification and Regression Trees) [38] [39]	
Advantages	Disadvantages
• Transparent and easy to understand • Assigns specific values to the inputs and output of each decision of the problem, thus, every probability can be evaluated.	• It does not work well if there are smooth limits. • Does not work best if the problem has many un-correlated variables. • It has high variance and it is unstable.

3.2.7. Logistic Regression (LR)

Logistic Regression uses the characteristics of the flows as continuous, discrete or hybrid. It then passes the inputs through a logistic function by combining the inputs linearly. This method, which is widely used, gives more reliable results in large datasets [40] [41]. Table 8 presents the advantages and disadvantages of this algorithm.

Table 8. Logistic Regression advantages and disadvantages

Logistic Regression [42] [43]	
Advantages	Disadvantages
• Has a low variance • Provides probabilities for output • Easy to apply and does not spend much time to train	• Cannot naturally classify multi-class data but can be adapted to multi-class classification with various applications. • Does not work well when there are correlated attributes

3.2.8. Gradient Boosting Machine (GBM)

Gradient Boosting Machine algorithm; a learning method makes predictions by combining outputs from trees. It builds trees by reducing the errors of previous trees. Therefore, as the tree is added to the model, the model becomes more reliable. It consists these parameters; number of trees, depths of the trees and the learning rate [44]. Table 9 presents the advantages and disadvantages of this algorithm.

Table 9. Gradient Boosting Machine Advantages and Disadvantages

Gradient Boosting Machine (GBM) [45] [46]	
Advantages	Disadvantages
• Makes easier to use custom functions because it provides optimization through functions rather than parameters. • Since it is a step-by-step algorithm, it can give good results in large and very different datasets.	• Training takes a long time because the trees are built sequentially. • If the data is noisy, errors occur when estimating or classifying.

3.2.9. Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting, which is called xGBoost, a scalable GBM sub-application, is a preferred algorithm because it makes computation fast and increases the performance of the model [47]. Table 10 presents the advantages and the disadvantages of the xGBoost algorithm.

Table 10. xGBoost advantages and disadvantages

xGBoost (XGB) [48] [49]	
Advantages	Disadvantages
• Can prevent overfitting if the data is clean • Can handle the missing values • Allows a cross-validation at each iteration of the process because of that it optimizes the number of iterations.	• More difficult to understand than other linear algorithms. • If data is noisy, it may overfit.

4. Experiment and Results

In this study, Python's scikit-learn package has been used to create the models and compare the algorithms [50]. Considering the descriptive statistics of the data set in Table 11;

Table 11. Descriptive Statistics of CTG Dataset

	count	mean	std
LB	2126.0	133.303857	9.840844
AC	2126.0	0.003178	0.003866
FM	2126.0	0.009481	0.046666
UC	2126.0	0.004366	0.002946
DL	2126.0	0.001889	0.002960
DS	2126.0	0.000003	0.000057
DP	2126.0	0.000159	0.000590

The mean value of LB is 133.3, so it means heartbeats of every individual in the sample is about 133 beats per minute. As it is seen in the Table 11, the mean of AC, accelerations, is about 0.003 this value may seem meaningless and very small, but for heart diseases it is important that the accelerations of heart should be balanced. Means of FM and UC value, which is Heart Ventricle Movement and Heart Ventricle Contraction, are about 0.009 and 0.004. Heart's Ventricle Movement and Ventricle Contraction are important for the heart to perform its functions. DL and DS values are Light and Severe Decelerations, means of these values are very small but just like the accelerations, these are important for the balanced functioning of heart. Considering the NSP dependent class variable, there are 1655 normal, 295 suspect and 176 pathological individuals of 2126 individuals.

The performance information of the results can be expressed with the confusion matrix in Figure 1. In the confusion matrix, the rows represent total numbers of the actual values of the class variable and the columns represent the total numbers of the predicted values of the model.

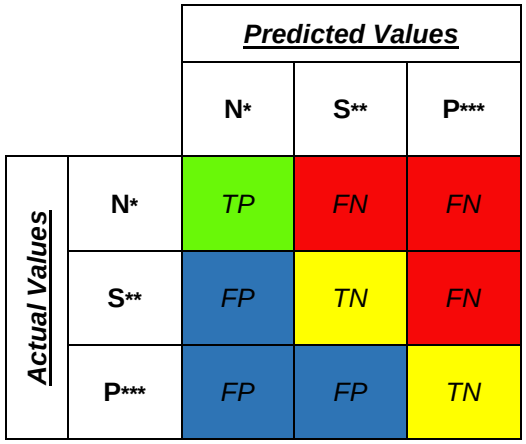

Figure 1. Confusion Matrix for Multi-class

Firstly, data set has been split, 30% for testing and the remaining 70% for the training. In the measurement of the model success based on the confusion matrix, accuracy, precision, recall and F-score of the model can be calculated.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-Score = \frac{2\times \text{Recall} \times Precision}{Recall+Precision} \quad (4)$$

Accuracy (1) is the predictions that the model estimated right. It is the ratio of number of correct predictions and total number of predictions. Precision (2) is calculated as the ratio of predicted correct results to all predicted results. Recall (3) is calculated by the ratio of the predicted correct results to the real correct results. Through these definitions of precision and recall above, the F-Score (4) is the harmonic mean of these values. [51].

Classification results are presented in the Table 12;

Table 12. Classification success of algorithms (%)

Algorithms	*Accuracy*	*Precision*	*Recall*	*F-Score*
SVM	85.7	88.1	94.4	91.1
GNB	85.1	87.6	94.3	91.1
kNN	87.7	89.8	95.2	91.9
MLP	85.3	88.1	93.8	90.3
RF	93.2	95.7	97.3	96.4
CART	84.4	85.2	96.9	90.6
LR	83.7	88.7	90.8	89.7
GBM	90.1	93.4	94	93.7
XGB	91.4	93.2	95.9	94.5

Considering the results, it is seen that RF gives the best result by far which build 6 trees with and accuracy of 93.2%. XGB followed RF with an accuracy of 91.4%. The algorithm closest to RF and XGB is GBM with an accuracy of 90.1%. Then there is kNN, with an accuracy of 87.7% and KNN's k is 6, which means number of neighbors. SVM's accuracy score is 85.7%. SVM followed by MLP, which has 6 hidden layers, has an 85.3% accuracy. GNB has an 85.1% accuracy. One of the least two accuracy is CART which is 84.4% and the other one is LR which has the 83.7% accuracy.

Accuracy and reliability are important in studies conducted in important areas such as healthcare [52]. A different algorithm in a different problem can come to the forefront and give more accurate results due to; the processing of classifier models in different ways, the use of different types of data, and the use of many different classification parameters. Therefore, in general, it is not sure which algorithm is the best [53]. But for this study, the Random Forest algorithm, which is extremely popular in recent years, gave the best results in this study.

5. Conclusion

Multi-class classification algorithms can be applied in many fields and can be used to solve many problems, including the diagnosis of heart diseases. The Random Forest algorithm, which gives the best results in this study, may fall behind other algorithms in another study, because the success rates of the algorithms vary according to the type of problem and the diversity of the data set. However, as seen in many studies, algorithms developed with open source software which are more current than other algorithms give results that are more reliable.

References

[1] "Toplum İçin Bilgiler," 2019. [Online]. Available: https://www.tkd.org.tr/kalp-yetersizligi-calisma-grubu/sayfa/toplum_icin_bilgiler. (10/8/2019)

[2] WHO, "World Health Statistics, 2018," 6 6 2018. [Online]. Available: https://www.tuseb.gov.tr/enstitu/tacese/yuklemeler/istatistik/9789241565585_eng.pdf. (10/8/2019)

[3] Onat, A., Şenocak, M., Örnek E., Gözükara Y., Şurdumavcı G., Karaaslan Y., Özışık U., İşler M., Taşkın V., Tabak F., Öz Ö., Özcan R., "Türkiye'de Erişkinlerde Kalp Hastalığı ve Risk Faktörleri Sıklığı Taraması: 5. Hipertansiyon ve Sigara İçimi *Türk Kardiyoloji Derneği Arastırması,* Vol. 19: 169-171, 1991.

[4] Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC., "Models to Predict Cardiovascular Risk: Comparison of CART, Multilayer Perceptron and Logistic Regression" *Proceedings of the American Medical Informatics Association Symposium*, 2000.

[5] Priyanka, N. and RaviKumar, P. RaviKumar, "Usage of Data Mining Techniques in Predicting the Heart Diseases - Naive Bayes & Decision Tree," *Circuit, Power and Computing Technologies International Conference*, 2017.

[6] Latha, C.B.C., Jeeva, S.C., "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques," *Informatics in Medicine* , 2019.

[7] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S. and Sun R., "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, 2018.

[8] C. Beyene and P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 165-174, 2018.

[9] UCI Machine Learning Repository, "https://archive.ics.uci.edu/ml/datasets/Cardiotocography," 7 9 2010. [Online], (10/8/2019)

[10] Silahtaroğlu G., Veri Madenciliği Kavram ve Algoritmaları, pg. 67, 2016, Papatya Bilim, İstanbul.

[11] Cortes, C., and Vapnik, V., "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[12] Wei-wu, Y., and Hui-he, S., "Application of Support Vector Machines and Least Squares Support Vector Machines to Heart Disease Diagnoses," *Computer Simulation*, vol. 3, 2003.

[13] Çomak, E., Arslan, A. and Türkoğlu, İ., "A Decision Support System Based on Support Vector Machines for Diagnosis of the Heart Valve Diseases," *Computers in Biology and Medicine*, vol. 37, no. 1, pp. 21-27, 2007.

[14] Ghumbre, S., Patil, C., and Ghatol, A., "Heart Disease Diagnosis using Support Vector Machine," *International Conference on Computer Science and Information Technology (ICCSIT')*, Pattaya, 2011.

[15] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," DIW Berlin , Berlin, 2008.

[16] Fedorovici, L.O., and Dragan, F., "A Comparison Between a Neural Network and a SVM and Zernike Moments Based Blob Recognition Modules," *6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania, 2011.

[17] Yu, W. M., Du, T., and Lim, K. B., "Comparison of the Support Vector Machine and Relevant Vector Machine in Regression and Classification

Problems" *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, Kunming, China, 2004.

[18] John, G. H., and Langley, P., "Estimating Continuous Distributions in Bayesian Classifiers," *UAI'95 Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, 1995.

[19] Sebe, N., Lew, M., Cohen, I., Gang, A., and Huang, T., "Emotion Recognition Using a Cauchy Naive Bayes Classifier," *Object Recognition Supported by User Interaction for Service Robots*, Vol. 1, pp. 17-20. IEEE.

[20] Murphy, K. P., *Naive Bayes Classifiers*, University of British Columbia, 18, 60, 2006.

[21] Jadhav, S. D., and Channe, H. P., "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *International Journal of Science and Research* , vol. 1, no. 5, 2016.

[22] McCallum, A., and Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification," 1998.

[23] Cover, T. M., and Heart, P. E., "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory* , vol. 13, no. 1, pp. 21-27, 1967.

[24] S. Patel, "Machine Learning 101-Chapter 4: K Nearest Neighbors Classifier," Medium.com, 17 5 2017. [Online]. Available: https://medium.com/machine-learning-101/k-nearest-neighbors-classifier-1c1ff404d265. (11/8/2019)

[25] Imandoust S. B. and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *International Journal of Engineering Research and Applications*, vol. 3, no. 5, 2013.

[26] Cunningham, P., and Delany, S. J., "k-Nearest Neighbour Classifiers," *Technical Report UCD-CSI-4*, Dublin, 2007.

[27] Genesis, "Pros and Cons of K-Nearest Neighbors," 25 9 2018. [Online]. Available: https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/. (11/8/2019)

[28] Öztemel E., **Yapay Sinir Ağları**, Papatya Yayıncılık, 2003.

[29] Gardnera, M., and Dorlinga, S., "Artificial Neural Networks (the Multilayer Perceptron)—a Review of Applications in the Atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627-2636, 1998.

[30] B. Venkateswaran and G. Ciaburro, Neural Networks with R, Packt Publishing, 2017.

[31] Şimşek, H. K., "Makine Öğrenmesi Dersleri 5b: Random Forest (Sınıflandırma)," Medium.com, 24 3 2018. [Online]. Available: https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-5-bagging-ve-random-forest-2f803cf21e07. (12/8/2019)

[32] Pal, M., "Random forest Classifier for Remote Sensing Classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217-222, 2005.

[33] "Learn Random Forest Using Excel," 27 12 2017. [Online]. Available: https://www.newtechdojo.com/learn-random-forest-using-excel/. (12/8/2019)

[34] Statnikov, A., Wang, L., and Aliferis C. F., "A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-based Cancer classification," *BMC Bioinformatics*, vol. 9, no. 319, 2008.

[35] Kaur, S. and Kaur, H., "Review of Decision Tree Data mining Algorithms: CART and C4.5," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 4, pp. 436-439, 2017.

[36] Lewis, R. J., "An Introduction to Classification and Regression Tree (CART) Analysis," *Annual Meeting of the Society for Academic Emergency Medicine*, San Francisco, 2000.

[37] Silahtaroğlu G., Veri Madenciliği Kavram ve Algoritmaları, 3th Edition Papatya Bilim, p. 82-86, 2016.

[38] Wu, X., and Kumar, V., Top Ten Algorithms in Data Mining, CRC Press, 2009.

[39] Juneja, N., "What Are the Disadvantages of Using a Decision-Tree for Classification?," Quora, 4 4 2018. [Online]. Available: https://www.quora.com/What-are-the-

disadvantages-of-using-a-decision-tree-for-classification. (12/8/2019)

[40] Jothi, N., Rashid, N. A., and Husain, W., "Data Mining in Healthcare – A Review," Procedia Computer Science. 72. 306-313, 2015.

[41] Worth, A. P., and Cronin, M. T., "The Use of Discriminant Analysis, Logistic Regression and Classification Tree Analysis in the Development of Classification Models for Human Health Effects," *Journal of Molecular Structure: THEOCHEM*, vol. 622, no. 1-2, pp. 97-111, 2003.

[42] Tu, J. V., "Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes," *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225-1231, 1996.

[43] Machine Learning Blog, "Logistic Regression 101," Machine Learning Blog, 23 4 2018. [Online]. Available: https://machinelearning-blog.com/2018/04/23/logistic-regression-101. (12/8/2019)

[44] Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[45] Ravanshad, A., "Gradient Boosting vs Random Forest," Medium, 28 4 2018. [Online]. Available: https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80. (12/8/2019)

[46] Rao H., Shi X., Rodrigue A. K., Feng J., Xia Y., Elhoseny, M., Yuan X., Gu L.,, "Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree," *Applied Soft Computing*, vol. 74, pp. 634-642, 2019.

[47] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 2016.

[48] Kumar, N., "Advantages of XGBoost Algorithm in Machine Learning," The Professional Point, 9 3 2019. [Online]. Available: http://theprofessionalspoint.blogspot.com/2019/03/advantages-of-xgboost-algorithm-in.html. (13/8/2019)

[49] Nielsen, D., *Why Does XGBoost Win* "Every" Machine Learning Competition?, *MS Thesis*, Norwegian University of Science and Technology Department of Mathematical Sciences, 2016.

[50] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. , Thirion, B. and Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* , vol. 12, pp. 2825-2830, 2011.

[51] Şeker Ş. E., "F1 Değerlendirme (F1-Scoring)," Bilgisayar Kavramları, 30 10 2010. [Online]. Available: http://bilgisayarkavramlari.sadievrenseker.com/2010/09/30/f1-degerlendirme-f1-scoring/. (13/8/2019)

[52] Zriqat, E., Altamimi, A., and Azzeh, M., "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods," *International Journal of Computer Science and Information Security*, vol. 14, no. 12, pp. 868-879, 2016.

[53] Zhang, M. L. and Zhou, Z. H. "A Review on Multi-Label Learning Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819-1837, 2014.

Recycle Project with RFM Analysis

Semra Erpolat Taşabat
Mimar Sinan Fine Art University
semra.erpolat@msgsu.edu.tr

Esra Akca
Mimar Sinan Fine Art University
20151101073@ogr.msgsu.edu.tr

Abstract

With the advancement of technology and the widespread use of the Internet, the concept of Big Data, which we are often beginning to hear its name, has emerged. Big data analysis analyzes large volumes of data, enabling intelligent decisions. One of these methods is RFM. RFM analysis is an effective and practical marketing model that combines the initials of **R**ecency **F**requency and **M**onetary (RFM) and performs behavioral customer segmentation. In this study, a new model has been proposed by showing the applicability of RFM analysis on the recycling project. Thus, the analysis of big data on the social responsibility project was carried out for the first time with RFM. According to the contribution rate of recycling, it is aimed to establish a profitable relationship between the customer and the company by applying discounts to the customers.

Keywords: RFM, Recycle; Data Analysis; Customer Segmentation

1. Introduction

The main purpose of data analysis is to ensure that the accumulated data is achieved with logical, useful and effective results by performing purposeful analyses. In doing so, it is possible to utilize many different analyses, models and algorithms. One of them is RFM analysis.

RFM, which can be applied to different fields, can also be applied on recycling projects with this study. Recycling is a very important and necessary requirement, especially for the world, which is getting more and more polluted every day. Recycling, which is also of great importance to the production sector, provides benefits at many points, especially money savings and reduction in production costs. Recycling metals provides protection of natural resources, while using raw materials to produce new products requires less energy than energy. It saves money and allows manufacturing businesses to reduce production costs. While managing large data is an important point; there have been very rare studies of recycling in the world as analysis. One of them is the study of hydrogen storage cost analysis by Cassidy Houchins. The purpose of this analysis has been to reduce costs. Another study is the conversion of used oil into valuable diesel by Anthony Kasozi. As a result of this study, it is aimed to reduce the environmental pollution caused by waste oil.

As a result of extensive literature research, it has not been seen that recycling projects have been handled by RFM analysis before. There has been no study in Turkey using any analysis method in which a segment strategy suitable for recycling is developed. For this reason, by segmenting the customers by reaching the right people; The right to special discounts for customers who can contribute to the recycling of metal waste is an incentive method applied for the first time in this study. This method and approach has been studied on the 2019 data collected by a company operating in the production of metal kitchen ware in Istanbul. In the following sections of the study, RFM

will be informed about different RFM models. Finally, the results will be interpreted.

2. Reference Research

2.1. For Recycle

For our future, natural resources need to be preserved. Recycling also means investment in the future and the economy. The following is the study on recycling.

In 2010, the Caught Green Handed campaign was launched in Ohio, USA as one of the "Grab the Prize for Contributing to Recycling" project. Consumers who brought waste to recycling bins were greeted with gift cards from various stores. The few "green handed" at first increased as news spread that prizes were being distributed.

2.2. For RFM Analysis

RFM analysis has an important role in data mining. An important study of RFM analysis is target company's "Pregnancy Prediction Score" application, one of the major U.S. retailers. Target managed to increase its revenues by $23 billion between 2002 and 2010.

3. RFM Model

This concept was first introduced in 1996 by Jan Roelf Bult and Tom Wansbeek as a marketing estimation application of economic models combined with statistical techniques. RFM Mahboubeh Khajvand, part of the decision support system, was used in 2011 to segment customers and find the targeted audience in the most accurate way, a useful, simple and powerful consumer, Customer Relationship Management (CRM) is the application model, he defined.

5. Data Mining and RFM

4.1 Clustering Using RFM

In recent years, some researchers have considered RFM variables in improving clustering. For example, Hosseini (2010) combined the weighted RFM model with K-Means to improve CRM. Clustering is used to find similar customer segments.

4.2. Classification Using RFM

The integration of classification techniques and RFM was used by Olson et al. in 2009 to analyze customers' response suppositions to a particular product promotion. In the same year, Cheng and Chen compared and discussed three data mining techniques: the relative change in customer segmentation between logistics regression, decision trees and neural network algorithms.

4.3. Integrated Approach

The use of RFM analysis in data mining; offers a new three-step approach using a few parameters together. These are classification clustering and product association rule mining. This model is meant to help managers develop exactly better marketing strategies using data mining and RFM analytics.

5. RFM Analysis

RFM analysis can be calculated by multiple methods and formulas. The equation (1) contains the general formulation of RFM.

$$(ProximityScore \times ProximityWeight) + (FrequencyScore \times FrequencyWeight) + (AmountScore \times AmountWeight) \quad (1)$$

"Proximity, Frequency, Amount" in equation (1); date, frequency and amounts are divided into 5 intervals of 20% in itself, the highest 20% of the 5, the highest 2. part 4, 3. part 3, 4. part 2 is calculated by giving 1 point to the lowest or the last lowest or the 20% of the 20%. The weights in the formula underlying this logic differed according to the people and theories. These are listed below, respectively.

5.1. Model (1)

$$(Proximity\ Score \times 100) + (Frequency\ Score \times 10) + (Amount\ Score) \quad (2)$$

As can be seen in Model (1), the most important factors in RFM analysis are proximity, frequency and amount. This is how the RFM score is calculated for each customer [1].

5.2. Model (2)

$$(R \times 3) + (F \times 2) + (M \times 1) \quad (3)$$

Model (2) was proposed by Miglautsch [6].

5.3. Model (3)

$$(R \times 9.9) + (F \times 6.6) + (M \times 3.3) \quad (4)$$

Model (3) was developed by Tsai and Chiu [7], and companies that want to act on a 100-point scale.

5.4. Model (4)

$$(R\ Skor) + (F\ Skor) + (M\ Skor) \quad (5)$$

The basic requirement in this method depends on the company's equal evaluation of the three main features [2].

6. Recycle Project with RFM Analysis

In the analysis, the data set; Turkish-based metal kitchen ware sourcing company includes transactions between 01/01/ 2017 and 31/12/2017. The aims of the analysis identify the customers who can contribute the most to recycling, which is a social responsibility project. Most frequently, customers who shop and pay the most in the most recent date range are potentially the ones with the most tendency to recycle products they've bought in the past. The method proposed in the study is Model (1), Model (2), Model (3), Model (4).

The data has a total of 284 customers and 3 different variables (Recency, Frequency, Moneatry). The Table 1 contains the information of 10 customers as an example.

Table 1. Customer information

CUSTOMER NAME	RECENCY	FREQUENCY	MONEATRY
Customer1	1.10.2017	10	416.938.58 TL
Customer2	1.12.2017	12	187.934.02 TL
Customer3	1.12.2017	12	459.193.52 TL
Customer4	1.12.2017	12	147.629.17 TL
Customer5	1.12.2017	12	356.535.83 TL
Customer6	1.12.2017	9	1.336.913.71 TL
Customer7	1.06.2017	6	35.404.19 TL
Customer8	1.12.2017	12	188.998.24 TL
Customer9	1.12.2017	12	172.308.37 TL
Customer10	1.10.2017	7	124.903.68 TL

The IBM Watson Studio SPSS Modeler was used in the study. It helps to modernize predictive analytics and machine learning processes and accelerate the time it takes to achieve value. Uses advanced analytical capabilities including predictive analytics, interim statistical analysis, predictive modeling, data mining, text analytics, optimization, real-time scoring, and machine learning. These tools help organizations discover models in data and predict what will happen next, beyond knowing what happened in the past. Data has been transferred to IBM

Watson Studio. Modeler Flow has been selected and the analysis has begun.

Figure 1. IBM Watson Studio Modeler Flow

RFM Analysis was integrated into the transferred data and a decision diagram was created. It was requested that the outputs be given in the form of tables and charts.

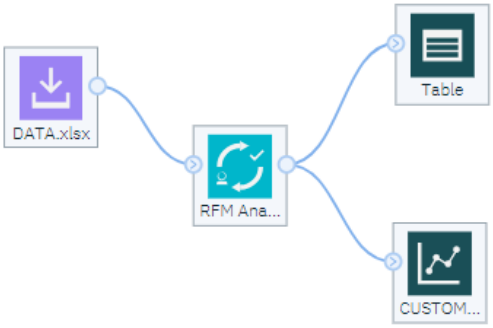

Figure 2. IBM Watson Studio decision stream

According to the models we want to apply RFM Analysis, the relevant weight areas are written on the relevant weights and score calculations are made for each model.

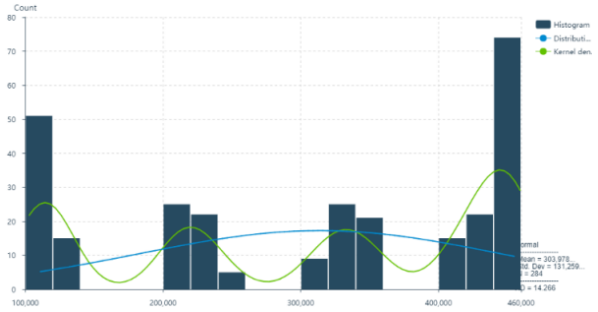

Figure 3. Modeler RFM Score Chart for Model (1)

The average was 303.978, as shown in the chart. The normality distribution is shown in blue and the core density estimate is green.

Figure 4 shows that the size of the boxes varies according to RFM score frequencies for Model (2). While the minimum frequency is of 11, it can be said that the maximum frequency belongs to 27.

Figure 4. Data Refinety Flow RFM Score Chart Model (2).

A graph of the RFM scores of Model (3) is shown in Figure 4. Scores are more divided by weight productivity.

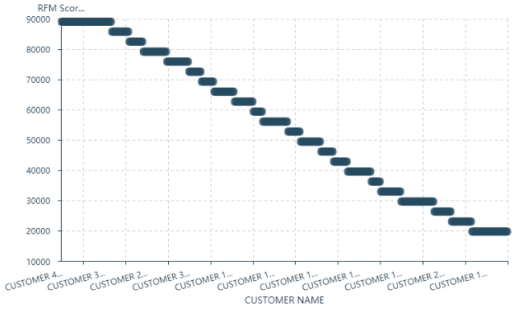

Figure 5. Data Refinety Flow RFM skor chart for Model (3)

Finally, the graph of the Model (4) is shown in Figure 6. Frequency distributions are observed according to RFM scores with a different graph representation. The dimensions of the balls indicate frequency size. The lowest frequency is for the score of 11, while the highest frequency belongs to the 6 score.

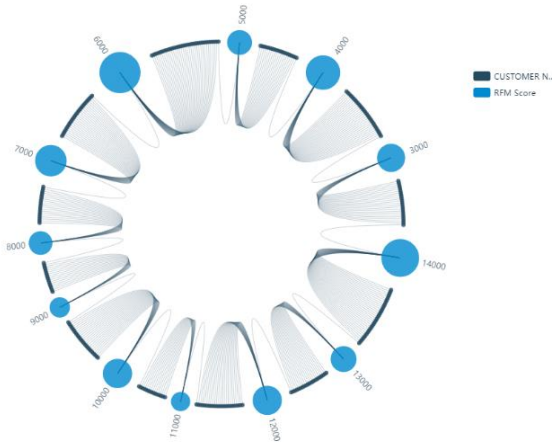

Figure 6. Data Refinety Flow RFM score chart for Model (4)

The score results of the first 8 data are included in Table 2.

Table 2. Scores for 4 different formulas

CST NAME	RFM SKOR-MODEL1	CST NAME	RFM SKOR-MODEL2	CST NAME	RFM SKOR-MODEL3	CST NAME	RFM SKOR-MODEL4
Customer 43	455	Customer 43	27	Customer 43	89.100	Customer 43	14
Customer 6	455	Customer 6	27	Customer 6	89.100	Customer 6	14
Customer 2	455	Customer 2	27	Customer 2	89.100	Customer 2	14
Customer 5	455	Customer 5	27	Customer 5	89.100	Customer 5	14
Customer 64	454	Customer 64	26	Customer 64	85.80	Customer 64	13
Customer 76	454	Customer 76	26	Customer 76	85.80	Customer 76	13
Customer 44	454	Customer 44	26	Customer 44	85.80	Customer 44	13
Customer 83	454	Customer 83	26	Customer 83	85.80	Customer 83	13

What is important here is the ranking of the scores. Although the weights of the formulas are different, when we sort them by values in themselves, it is seen that we reach the same result in all 4 formulas and that the segment distributions of customers are the same. Since these 4 formulas are for the same purpose under the logic of RFM analysis, it is normal for the results to be the same, and if the results were different, it would have been said that the formulas were incorrect or not for the same purpose.

7. Results

Data analysis contains important points not on the size of the data, but on the correct, efficient, purposeful use of the data and what to do with the information.

Analyzing data is very important so that we can make smart decisions and make future decisions. The summary and importance of data mining is the separation of useful information from the accumulated information and transported to the scanning process. Analyzing data in ways appropriate to the company's purpose will bring a lot of return to the companies. Today, data analysis has gained more importance and value with the development of Industry 4.0 and the orientation towards more "human" focus. Companies tend to analyze logically with appropriate techniques.

In this study, in which RFM analysis was included in data mining techniques, customers were segmented according to RFM scale based on customers' purchasing habits. As a result, potential customers have the potential to contribute the most to metal waste recycling. In this way, a useful environmental-customer-vendor relationship has been created by gaining discounts as much as the rate at which customers contribute to recycling. At the same time, sensitivity to the environment has been shown and valued.

References

[1] Bult. J. R., Wansbeek. T. (1995). Optimal selection for direct mail. *Marketing Science*.

[2] Blattberg R.C., Kim BD., Neslin S. A. (2008). Why Database Marketing?. In: Database Marketing. *International Series in Quantitative Marketing*, vol 18. Springer, New York, NY.

[3] Judith A. C. (2002). *Economic Growth versus the Environment*, Palgrave, New York.

[4] Bill Gates. (1999). *Business The Speed of Thought Using a Digital Nervous System*, Warner Books, New York.

[5] Kuo R-J, Ho LM, Hu CM. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, 29(11):1475-93.

[6] Miglautsch JR. (2000). Thoughts on RFM scoring. *J. Database Mark*., 8(1):
67-72.

[7] Tsai CY, Chiu CC. (2004). A purchase-based market segmentation methology. *Expert Syst. Appl*., 27: 265-276.

Inferences About Development Levels of Countries with Data Envelopment Analysis

Semra Erpolat Taşabat
Mimar Sinan Fine Art University
semra.erpolat@msgsu.edu.tr

Abstract

Human development, which was previously calculated only with economic indicators and simple arithmetical formulations, can now be calculated more extensively by taking into account economic indicators as well as education, health and social indicators and with different decision making methods. In this study, human development levels of world countries will be calculated with DEA by using dimensions of Human Development Index (HDI). For this purpose a new model which is named as "weight restricted input-oriented CCR DEA model without input" will be proposed. In the proposed model, these dimensions will be treated as outputs and a dummy input variable will be created. A new approach will be proposed for weight restriction inequalities and weight constraint inequalities for outputs will be added to the model. The results obtained will be evaluated together with HDI values and inferences will be made.

Keywords: Data Envelopment Analysis; Human Development Index; Efficiency

1. Introduction

Human development is a multidimensional concept that deals with the enrichment of the lives of people living in them rather than the economic wealth of countries. Human development or the human development approach is about expanding the richness of human life, rather than simply the richness of the economy in which human beings live. It is an approach that is focused on people and their opportunities and choices [1].

Various criticisms have been brought by the researchers for this calculation of HDI. The biggest criticism has come to the calculation method of HDI. It has been argued that the geometric mean is not an adequate approach to construct such an index. Therefore, different models and approaches have been proposed for the calculation of HDI. Another important criticism concerns the dimensions and indicators used in HDI. It is a common opinion that the dimensions and indicators used in HDI are insufficient. Another area of criticism has been the weight of the dimensions or indicators discussed in HDI. Different approaches and suggestions for these issues have been brought by the researchers.

In this study, a new approach will be proposed for the calculation method of HDI and the weights restriction of dimensions. For this purpose, DEA will be used as calculation method and Type I Assurance Regions (ARI) will be used as weights restriction method. In the following sections, DEA will be discussed first and DEA method proposed within the scope of the study will be elaborated. Then, information about HDI will be given and the calculation of a new HDI will be explained with the help of the proposed model. By evaluating the results obtained, various inferences will be made for the development and effectiveness of world countries.

2. Data Envelopment Analysis

The DEA was developed in 1978 by Charnes, Cooper and Rhodes [2], taking advantage of Farrell's concept of efficiency in 1957 [3]. The DEA compares DMUs that are supposed to be homogeneous. Examining efficiency with DEA requires: Selection of DMUs, selection of inputs and outputs, obtaining data, measurement of relative efficiency, determination of efficient and inefficient DMUs, determination of reference groups for inefficient DMUs, setting targets for efficient DMUs, evaluation of results [4]. It is possible to classify the models used in DEA according to the enveloping method (return assumption) and the distances (orientation) of the inefficient units to the efficient production limit.

In this study, weight restricted input-oriented CCR DEA model without input was preferred. Because, while calculating the human development levels of the world countries, three dimensions of HDI are accepted as output. Since all of the variables in the data set are evaluated as outputs, the input-free approach proposed by Mahlberg ve Obersteiner [5] was applied. According to this approach, a dummy variable that takes the value of 1 for each DMU (all values of 1) is accepted as input. A new model has been developed by adding weight restriction inequalities to this model. In the following subsections, the methodology used in developing the weight restricted input-oriented CCR DEA model without input model is given.

2.1. Weight Unrestricted Input-Oriented CCR DEA Model without Input

The Weight Unrestricted Input-Oriented CCR DEA Model without Input can be constructed as in (1).

$$\begin{aligned} &Max \sum_{r=1}^{s} u_r y_{rk} = Z_0 \\ &\sum_{r=1}^{s} u_r y_{rj} \le 0 \\ &u_r \ge \varepsilon \end{aligned} \tag{1}$$

The meaning of the notations in (1) Z_0 performance score; ε non-archimedean variable ($\varepsilon = 10^{-8}$); x_{ij} i-th input of the j-th DMU ($i = 1,..., m$); y_{rj} r-th output of the j-th DMU ($r = 1,..., s$; $j = 1,..., N$); v_i weight of the i-th input; u_r weight of the r-th output.

2.1. Weight Restricted Input-Oriented CCR DEA Model without Input

There is a large diversity of methods that can be used in incorporating value judgments in DEA. Allen et. al. [6] identified three broad approaches: Direct restrictions on the weights, restricting the virtual inputs and outputs, adjusting the observed input-output levels to capture value judgments.

In this study "Direct restrictions on the weights" method is preferred. The direct restrictions on the weights can be applied in three different ways: Assurance Regions Type I (ARI), Assurance Region Type II (ARII), and Absolute Weights Restriction [6]. In this study, ARI method is used. ARI involves the incorporation of the relative ordering of the weights associated with the input/output attributes into the analysis. There are two forms of this type [7]. In this study, the form given in (2) is adopted.

$$\begin{aligned} &\alpha_n \le \frac{u_m^k}{u_n^k} \le \beta_n \\ &for\ some\ n,m \in \{1,\cdots,I\}, \forall k = 1,\cdots,K \end{aligned} \tag{2}$$

Thus, the Weight Restricted Input-Oriented CCR was created as in DEA Model without Input is as in (3).

$$
\begin{aligned}
&Max \sum_{r=1}^{s} u_r y_{rk} = Z_0 \\
&\sum_{r=1}^{s} u_r y_{rj} \le 0 \\
&\alpha \le \frac{u_r}{u_{r+1}} \le \beta \\
&u_r \ge \varepsilon
\end{aligned} \tag{3}
$$

Here, α and β lower and upper bound of weight relation respectively. The bounds here depend on the scaling of the attributes involved and its choice of values are usually determined based on expert's opinion [8], [9].

3. Application

In this section, the proposed weight restricted input oriented CCR DEA model without input approach will be applied to HDI data.

3.1. Human Development

The first formation of HDI was made by economist Mahbub ul Haq, who argued that economic indicators alone were not sufficient to measure human development. Then, the transformation of human development into an index was realized in 1989 with a project supported by UNDP. The UNDP published a report in 1990, in which the index was computed for each country as a measure of the nation's human development. Since then UNDP has continued publishing a series of annual Human Development Reports (HDRs) [10]. Although it has been subject to various criticisms, the dimensions and indicators of HDI that have been accepted after 2010 and are still valid today.

HDI is a composite index of three dimensions which are a decent standard of living, knowledge and long and healthy life. A decent standard of living dimension which contained Gross National Income (GNI) per capita, create "GNI Index (GNII)". Likewise, life expectancy at birth at the long and healthy life generate "Life Expectancy Index (LEI)". And finally both expected years of schooling and mean years of schooling which is in dimension knowledge create "Education Index (EDI)" [10]. In order to obtain the HDI, the geometric mean of these three basic indexes calculated

3.2. Human Development with DEA

Models with no inputs and only outputs are widely used in performance evaluations [11], [12].

Lovell and Pastor [13], Lovell and et al. [14], Mahlberg and Obersteiner [5], Liu and et al. [11]; Despotis [15, 16] were examined DEA models without input or output. In the Mahlberg and Obersteiner [5]'s model, all the individual indicators are considered as outputs and a dummy input (equal to one) is assumed for all the countries.

3.2. Creating Weight Restriction

In this study, a general weight restriction technique was proposed by modifying this method proposed by Mahlberg and Obersteiner [5]. According to the proposed method, firstly, the minimum and maximum upper limits that can be taken such that the weight of any input or output discussed are not allowed being zero and one are arranged as in (4).

$$[A, C = 1 - B * A] \tag{4}$$

Where; A: Coefficient (according to the number of digits to be used: 0.1, 0.01 or 0.001 etc.), B: Number of inputs or outputs. Then, according to the Type II approach of the ARI method, the upper and lower limits ([α, β]) are arranged as in (5), with large weight denominator and small weight in denominator. The meanings of the coefficients A and B in (5) are the same as in (4) and the meaning of the coefficient D is 10, 100 or 1000 etc. according to the preferred number of digits.

$$[\alpha = A, \beta = D - B] \quad (5)$$

3.2. Model Development

It has been found that the lower and upper bounds for the outputs can be [0.1, 7] or [0.01, 97] or [0.001, 997] with the help of (4) and (5).

According to the proposed weight restriction approach, which weights are to be compared with each other, it has been decided as a result of the evaluation of the opinions in the literature about the significance of the three dimensions of HDI.

GNII is of greater importance than the other two indicators, LEI and EDI opinion is adopted. In this study, it is preferred to use one digits for weight restriction bounds. Therefore, the parameters to be used are: $A = 0.1$, $B = 3$, $C = 0.7$, $D = 10$. The proposed model is in (6).

$$\begin{aligned} &Max\ z_{j0} = w_{LEI} LEI_{j0} + w_{EDI} EDI_{j0} + w_{GNII} GNII_{j0} \\ &w_{LEI} LEI_j + w_{EDI} EDI_j + w_{GNII} GNII_j \le 1 \quad , j \in DMU\ set \\ &\alpha \le \frac{w_{GNII}}{w_{EDI}} \le \beta \\ &\alpha \le \frac{w_{GNII}}{w_{LEI}} \le \beta \\ &A \le w_{LEI}, w_{EDI}, w_{GNII} \le C \end{aligned} \quad (6)$$

3.3. Preparing Data for Analysis

In the study, EDI, LEI and GNII, which are the dimensions of HDI, were used in the HDR data prepared for 2019. First of all, the data was arranged in such a way that there was no blank data for these variables. According to this, there are 189 data for EDI, 192 for LEI and 191 for GNII of 193 world countries in 2019 HDR. Countries with null values for any of the EDI, LEI and GNII were Korea (Democratic People's Rep.), Nauru, Somalia and Tuvalu, and were excluded from the analysis. Thus, the number of DMUs was determined as 189. The structure of the decision matrix is shown in Table 3. To avoid excessive space, the first and last three DMUs are shown in Table 3 [17].

Table 3. Decision matrix for DEA analysis.

DMUs	Input	Outputs		
	Dummy	EDI	LEI	GNII
AFG	1	0.415	0.678	1824
ALB	1	0.745	0.9	11886
DZA	1	0.664	0.866	13802
. . .	. . .	. . .	. . .	. . .
YEM	1	0.349	0.695	1239
ZMB	1	0.58	0.65	3557
ZWE	1	0.558	0.642	1683

4. Findings and Inferences

The development scores and rankings of the world countries obtained as a result of the proposed models (Weight Unrestricted Input-Oriented CCR DEA Model without Input (Model I) and Weight Restricted Input-Oriented CCR DEA Model without Input (Model II)) and the HDI are given in Table 4 together.

When the results of Model I, Model II and HDI of 189 countries in the data set are examined, it is seen that there are countries with the same score. Therefore, the ranking number was 169 for Model I, 180 for Model II and 177 for HDI. Table 5 shows the rankings of Model I and Model II for the first five and the last five countries. The table also includes HDI rankings corresponding to the rankings of Model I and Model II.

Table 4. Comparison of Results

DMUs	Model I Score	HDI Rank	DMUs	Model II Score	HDI Rank
NOR	1.0000	1	NOR	1.0000	1
CHE	1.0000	2	CHE	1.0000	2
AUS	1.0000	3	AUS	1.0000	3
DEU	1.0000	5	DEU	1.0000	5
HKG	1.0000	7	HKG	1.0000	7
SGP	1.0000	8	SGP	1.0000	8
LIE	1.0000	16	LIE	1.0000	16
QAT	1.0000	34	QAT	1.0000	34
JPN	0.9970	17	ISL	0.9948	6
ISL	0.9951	6	IRL	0.9926	4
IRL	0.9926	4	JPN	0.9916	17
SWE	0.9900	7	SWE	0.9900	7
. . .	. . .	. . .	. . .	. . .	. . .
LSO	0.5496	150	LSO	0.5496	150
CIV	0.5373	158	CIV	0.5365	158
TCD	0.5183	174	TCD	0.5103	174
CAF	0.5132	176	SLE	0.4930	172
SLE	0.5030	172	CAF	0.4929	176

The results showed that there was no difference in the first rankings of Model I and Model II, but there were differences in subsequent rankings.

5. Conclusion

In this study, the calculation of human development of world countries by DEA method is discussed. A new model has been proposed for this purpose. This model is named as Weight Restricted Input-Oriented CCR DEA Model without Input. The results of the proposed model were compared with the results of Weight Unrestricted Input-Oriented CCR DEA Model without Input and HDI.

The proposed model has been developed as an alternative to DEA models that can be used in case of weight restriction. In the model, weight restriction was performed according to the Type II approach of the ARI method. Here, the approach proposed by Mahlberg and Obersteiner [x] for the determination of lower and upper bounds has been developed and modified. In the ratio of weights, the opinion that the weight of the input or output, which is considered to be of greater importance in line with the decision makers' opinion, is the denominator and the less important one is the denominator. In DEA method, the determination of the variables that should be taken as input and output in the evaluation of human development is considered as the dimensions of HDI in order to ensure the comparability of the results to HDI values calculated by UNDP. These dimensions are included in the model as outputs. And as input, a dummy variable with all values of 1 is used.

From the results, it was observed that weight restriction inequalities might cause significant differences in the results. Again, restricting the output weights, to not have a value of zero or one, allowed all outputs to be included in the model. Thus, all dimensions were effective in calculating development levels.

References

[1] http://hdr.undp.org/en/humandev. United Nations Development Programme Human Development Reports. (10.08.2019).

[2] Charnes A., Cooper W. W., Rhodes E. (1978). Evaluating Program and Managerial Efficiency: An Asslication of Data Envelopment Analysis to Program Follow Through. *Management Science*, 27, 1978, ss.668-697.

[3] Farrell M. J. (1957). The measurement of productive efficiency. *Journal of The Royal Statistical Society*, 120(3), 253-290.

[4] Taşabat E. S. (2011), Veri Zarflama Analizi, Evrim Kitap Evi, İstanbul.

[5] Mahlberg B, Obersteiner M. (2001). Remeasuring the HDI by data envelopment analysis. *IIASA interim report*, IR-01-069, Luxemburg.

[6] Allen R., Athanassopoulos A., Dyson R.G. and Thanassoulis E. (1997). Weights restrictions and value judgements in Data Envelopment Analysis: Evolution,

development and future directions. *Annals of Operations Research*, 73, 13-34.

[7] Thompson R.G., Langemeier L., Lee C., Lee E. and Thrall R. (1990). The role of multiplier bounds in efficiency analysis with application to Kansas farming. *Journal of Econometrics*, 46, 93-108.

[8] Beasley J. E. (1990). Comparing university departments. *Omega, International Journal of Management Science*, 18, 17-183.

[9] Talaue C. O., Diesta N. A. N., Tapia C. G. (2011). Weights Restriction by Multiple Decision Makers in Data Envelopment Analysis using Fuzzy Programming. *In Proceedings of the 11th Philippine computing science congress*, Philippines.

[10] Taşabat E. S. (2019). A Novel Multicriteria Decision-Making Method Based on Distance, Similarity, and Correlation: DSC TOPSIS. *Mathematical Problems in Engineering*, Volume 2019.

[11] Liu W.B., Zhang D. Q., Meng W., Li X. X. and Xu F. (2011). A Study of DEA models without explicit inputs. *Omega*, 39, 472-480.

[12] Hoarau J.F. and Blancard S. (2013). A new sustainable human development indicator for small island developing states: a reappraisal from data envelopment analysis. *Economic Modelling*, 30, 623-635.

[13] Lovell C. A. K. and Pastor J. T. (1999). Radial DEA models without inputs or without outputs. *European Journal of Operational Research*, 118, 46-51.

[14] Lovell C. A. K., Pastor J. T. and Turner J. A. (1995). Measuring macroeconomic performance in the OECD: A comparison of European and Non-European countries. *European Journal of Operational Research*, 87, 507-518.

[15] Despotis D. K. (2005). Measuring human development via data envelopment analysis: the case of Asia and the Pacific. *Omega*, 33, 385-390.

[16] Despotis D. K. (2005). A reassessment of the human development index via data envelopment analysis. *Journal of the Operational Research Society*, 56, 969-980.

[17] Taşabat E. S. (2019). Efficiency Analysis of World Countries According to Human Development Dimensions. Gece academy, In press.

Alternative Subway Project Selection with TOPSIS Method Using Different Weighting Techniques

Nihan Yücel
Mimar Sinan Fine Arts University
nhnycl@gmail.com

Semra Erpolat Taşabat
Mimar Sinan Fine Arts University
semra.erpolat@msgsu.edu.tr

Abstract

As the population in cities increases, public transportation is inadequate. For this reason, more efficient and faster public transportation technologies have been developed in big cities such as London, New York and Paris at the end of 1800s. Among these technologies, rail systems have the highest passenger capacity. Local governments need to find definitive solutions to the increasing traffic problem. It has been found that this can be solved by the formation of rail system networks. For the 13 alternative metro projects planned in Istanbul, TOPSIS method was proposed by using different weighting methods.

Keywords: MCDM; Railway Systems; TOPSIS

1. Introduction

With the beginning of the industrial revolution in the world, migration from villages to cities began and the population of the cities grew rapidly. With this growth, transportation in cities has become a problem. For the solution, high-capacity rail systems are being used in public transportation. The level of transportation service is low due to the fact that the number of trips in Istanbul is very high but the transportation infrastructure is not developed sufficiently. In the coming years, the new housing areas, new universities, new working areas will increase the demand for travel and transportation services will be insufficient. Road public transport solutions such as buses and minibuses are not effective due to traffic density. The only solution to meet the demand for busy travel in Istanbul is the development of railway networks.

Metro is the most used country in Japan. The Tokyo subway with a population of 37.5 million has an average of 8.7 million passengers per day. Second place comes from New York, the number of passengers per day is 5.5 million. 3.2 million people are traveling in London and 4.1 million in Paris each day. The average number of daily trips for Istanbul is 2.7 million. While the use of rail transport systems is 60% in Tokyo, 31% in New York, 22% in London and 25% in Paris, this rate has made a major breakthrough in Istanbul in recent years and it has increased from 3.6% to 18.7%.

Since rail system investments require large budgets, it is very important to prioritize the lines and projects to be made. It should be noted that resources are limited when trying to meet the needs. Therefore, one of the best and most important projects should be selected. In order to find exact solutions to the

transportation problem in the metropolises where there is traffic density, it should be ensured that the right choice of the rail public transportation system should be selected by analyzing the current problems in the best way.

The study will help the selection of the most efficient project by the addition of qualitative assessment criteria to the decision process besides the use of Multi Criteria Decision Making (MCDM) methods which are used as an alternative to traditional transportation investment evaluation methods.

When the rail system technologies in the world are examined, the explanations of the rail system types in Istanbul are given below according to the passenger capacity, traffic integration status and construction.

In order to choose the best transportation type to be used in the Istanbul Bosphorus Transit, 3 main criteria have been evaluated with 3 sub-criteria and Oncel used 3 different MCDM methods for this purpose [10]. Verma and Dhingra explained how Geographic Information Technologies can be used in the design of the rail system based on the journey request [11]. Monhajeri and Amin have tried to explain how the AHP can be used to select the station. They evaluated the assessment from the technical point of view, in terms of passengers, in terms of architecture and urbanism, and economically [9]. Alkubaisi has made the best tram route selection study. Six alternative pathways have been proposed and a GIS-based system has been used in conjunction with multi-criteria decision making [1].

2. Railway Project Selection

Istanbul is a metropolis with a large population density. Traffic is one of the most important problems of this city connecting the continents. In metropolitan cities like Istanbul, rail systems are the most suitable type of transportation that can solve this problem. The development of the city in the north-south direction and the extension of the distances show that fast and comfortable rail system solutions are needed. The share of the rail system in Istanbul, which has approximately 15 million trips per day, is 18.7%. New projects are planned in order to transfer highway transportation to rail systems. Thus, rail travel ratio in Istanbul is expected to increase. In this study, the best project will be selected between 13 alternatives.

2.1. Method of Research

In this study, the ranking of subway projects planned to be made in urban transportation by using equal, AHP and BWM weighting methods with TOPSIS has been made. “Number of Stations”, “Line Length (km)”, “Approximate Cost (TL)”, “Travel Time” and “Travel Request” are selected as criteria by the decision makers for 13 alternative projects. Weights for the criteria obtained as a result of the application of AHP and BWM were taken as data from the study conducted by Yücel [14].

Table 1. The Alternatives Table.

Line Name	Line Code	Number Of Stations	Line Lenght (km)	Approximate Cost(TL)	Travel Time	Travel Request
MAltunizade-Çamlıca Metro	A1	4	3,6	611540	12	8100
MİstinyeİTÜ-Kağıthane Metro	A2	10	12	4500000	23	35000
MKadıköy-Sultanbeyli Metro	A3	8	19	7600000	28	70000
MKazlıçeşme-Söğütlüçeşme Metro	A4	13	20	6971000	7,5	70000
MSabiha Gökçen Havalimanı-Kurtköy Metro	A5	4	6	2120000	8,5	50000
MSeyrantepe-Alibeyköy Metro	A6	4	7	2199000	8,5	70000
MSultangazi-Arnavutköy Metro	A7	5	15	6050000	16	45000
MVezneciler-Sultangazi Metro	A8	15	17,5	4539109	26,5	45000
M13Sarıgazi-Türkiş blokları Metro	A9	6	5,6	2112500	8	35000
M2Yenikapı-Sefaköy Metro	A10	11	14	2700000	21	70000
M3Kayaşehir-Fenertepe Metro	A11	3	3	2340000	10	70000
TEsenler-Davutpaşa Nostalgic Tram	A12	9	2,2	2950000	20	3000
T5Eyüp-cayrampaşaTram	A13	6	3,1	2950000	20	25000

2.1.1. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

TOPSIS, which is one of the methods used in the decision-making process, is a technique that allows the best choice among the alternatives. TOPSIS is one of the multipurpose decision making (MCDM) methods developed by Hwang and Yoon in 1981 [12]. The word TOPSIS is composed of the initials of the words Technique for Order Preference by Similarity to Ideal Solution.

The TOPSIS method is used to rank alternatives according to certain criteria. The first step of this method is the formation of the decision matrix. Then, the normalized decision matrix is obtained from the decision matrix and this decision matrix is weighted. The distances to the ideal solution and the negative ideal solution are calculated. Finally, the relative scores of each alternative are calculated and sequenced. If these stages are examined respectively [13]:

*Step1.*Create decision matrix.

Step2. Calculate Normalised Matrix

$$\overline{X}_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^{n} X_{ij}^2}} \quad (2.1)$$

Step3. Calculate weighted Normalised Matrix

$$Vij = \overline{X}_{ij} \times Wij \quad (2.2)$$

Step4. Calculate the ideal best and ideal worst value

Step5. Calculate the Euclidean distance from the ideal best

$$S_i^+ = \left[\sum_{j=1}^{m} \left(V_{ij} - V_j^+ \right)^2 \right]^{0.5} \quad (2.3)$$

Step6. Calculate the Euclidean distance from the ideal worst

$$S_i^- = \left[\sum_{j=1}^{m} \left(V_{ij} - V_j^- \right)^2 \right]^{0.5} \quad (2.4)$$

Step7. Calculate Performance Score

$$P_i = \frac{S_i^-}{S_i^+ + S_i^-} \quad (2.5)$$

2.1.2. Weighting

Determining the weight of performance criteria in MCDM methods is another important step. Because, weights given to the criteria can lead to different performance orders. Even small changes in weight values can often change the result significantly. For this reason, the determination of criteria weights is one of the most important stages of decision analysis [8].

In this study "equal and Saaty", which is most used in weighting methods, and "best-worst", which is a remarkable new method, will be studied.

i)Equal weighting method

Assuming that all criteria used to compare alternatives are of equal importance, it is based on equal weighting of criteria. Equally weighted criteria are considered as a control group in order to better understand the importance of criterion weights in MCDM methods.

ii)Saaty method

The Saaty method is one of the binary comparison methods and is also called the Analytic Hierarchy Process (AHP). AHP, introduced by Thomas Saaty (1980), is an effective tool for dealing with complex decision making, and may aid the decision maker to set priorities and make the best decision. By reducing complex decisions to a series of pairwise comparisons, and then synthesizing the results, the AHP helps to capture both subjective and objective aspects of a decision. In addition, the AHP incorporates a useful technique for checking the consistency of the decision maker's evaluations, thus reducing the bias in the decision making process.

iii)Best Worst method (BWM)

The Best Worst method it was developed by Jafar Rezaei (2015) as a multi-criteria decision

making method. In this method, the best (most important) and the worst (least significant) criteria is defined by the decision maker. Bilateral comparisons between these two criteria (best, worst) and other criteria are made. The consistency value of the comparisons is checked for reliability. Various alternative and criterion sets are weighted and final scores are determined and the best alternative is selected. Compared to the AHP method, BWM requires a smaller number of binary comparisons and allows for more consistent comparisons, which helps to achieve more reliable results.

BWM consists of 5 steps. These steps are used to determine the weight of the criteria and to find the scores of alternatives for each criterion [7].

3. Findings and Discussion

As shown in Table 2, different weighting methods were used for the rail system evaluation criteria. It is seen that there are different rankings according to weighting method. A8 was found first for equal weighting. A11 is the first for AHP, while A9 is the first when BWM is used.

While BWM has been created as an alternative for AHP produced similar rankings to AHP. Equal weighting method produced completely different findings compared to other methods.

Table 2. Project Rankings Calculated with TOPSIS: Equal weighing, AHP and BWM

No	Line Name	Line Code	EQUAL	AHP	BWM
1	MAltunizade-Çamlıca Metro	A1	8	10	8
2	MİstinyeİTÜ-Kağıthane Metro	A2	7	12	5
3	MKadıköy-Sultanceyli Metro	A3	10	5	6
4	MKazlıçeşme-Söğütlüçeşme Metro	A4	3	13	9
5	MSabiha Gökçen Havalimanı-Kurtköy Metro	A5	6	4	4
6	MSeyrantepe-Aliceyköy Metro	A6	5	3	3
7	MSultangazi-Arnavutköy Metro	A7	11	6	7
8	MVezneciler-Sultangazi Metro	A8	1	2	10
9	M13Sarıgazi-Türkiş blokları Metro	A9	4	9	1
10	M2Yenikapı-Sefaköy Metro	A10	2	11	11
11	M3Kayaşehir-Fenertepe Metro	A11	9	1	2
12	TEsenler-Davutpaşa Nostalgic Tram	A12	12	8	13
13	T5Eyüp-BayrampaşaTram	A13	13	7	12

4. Results

Today, the efficiency of projects should be taken into consideration in order to make the best use of limited resources. Rail system investments, which local governments spent significant part of their budget, should be evaluated effectively.

In this paper, it is also shown that by using various weighting methods used in TOPSIS project that will be constructed can be selected efficiently.

References

[1] Alkubaisi, M. I. T. (2014). Predefined evaluating criteria to select the best tramway route, Journal of Traffic and Logistics Engineering, 2.3.

[2] Rezaei, J. (2015a). Best-worst multi-criteria decision-making method. Omega (United Kingdom), 53,49-57

[3] Triantaphyllou, E. (2000). Multi-criteria Decision Making Methods, A Comparative Study. ISBN 978-1-4757-3157-6, Springer.

[4] Taşabat, E. S. (2017). A Novel Multi Criteria Decision Making Method Based On Distance, Similarity and Correlation. 10th Internatıonal Statistics Congress, Ankara, Turkey. Special issue.

[5] Saaty, T.L. (2008). Decision making with the analytic hierarchy process. Int. J. Services Sciences. Vol. 1, No. 1.

[6] Saaty, T. L. (2008) "Relative measurement and its generalization in decision making:Why pairwise comparisons are central in mathematics for the measurement of intangible factors, the analytic hierarchy/network process." Review of the Royal Spanish Academy of Sciences, Series A, Mathematics, 102(2), 251–318.

[7] Rezaei, J. (2015b). Best-worst multi-criteria decision-making method: Some properties and a linear model (working paper). Omega (United Kingdom), 23
[8] Taşabat, E.S. (2018). 19th International Symposium on Econometrics, Operations Research and Statistics 2018, Vol 19: 650-670.

[9] Mohajeri, N. and Amin, G. R. (2010) "Railway station site selection using analytical hierarchy process and data envelopment analysis." Computers & Industrial Engineering,1(59), 107- 114.

[10] Öncel, N. (2003) İstanbuldeki trafik probleminin çözümü için en uygun ulaşım alternatifinin seçimi. Yüksek lisans, Galatasaray Üniversitesi, Fen Bilimleri Enstitüsü, Istanbul.

[11] Verma, A. and Dhingra, S. L. (2005) "Optimal Urban Rail Transit Corridor Identification within Integrated Framework Using Geographical Information System." Journal of Urban Planning & Development, 2(131), 98-111.

[12] Hwang, C. L. and Yoon, K. (1981) Multiple Attribute Decision Making: Methods and Application, Springer, NewYork.

[13] Dumanoğlu, S. and Ergül, N. (2010) "İMKB'de İşlem Gören Teknoloji Şirketlerinin Mali Performans Ölçümü", Mufad Journal, Number 48.

[14] Yücel, N. (2019) "The Selection of Railway System Projects with Multi Creteria Decision Making Methods: A Case Study for Istanbul", Mimar Sinan Fine Arts University, Istanbul.

Fast Fault Solving Methods in Smart Manufacturing Lines with Augmented Reality Applications

Adem KAYAR
Department of Industrial Engineering, Faculty of Engineering, Istanbul Ticaret University, Istanbul-Turkey,
akayar@mfd.com.tr

Fatih ÖZTÜRK
Department of Industrial Engineering, Faculty of Engineering, Istanbul Medeniyet University, Istanbul-Turkey,
fatih.ozturk@medeniyet.edu.tr

Özkan KAYACAN
MCS Factory Digitalization End. Bil. Tek. Ltd. Şti., Istanbul-Turkey,
ozkan.kayacan@mfd.com.tr

Abstract

Digitalization and Industry 4.0 are being met with interest in almost every sector. When Industry 4.0 was mentioned for the first time at the Hannover fair in Germany in 2011, no one thought that the process would develop so quickly. We see that Digital Transformation applications, which are developing rapidly all over the world, are now being used in manufacturing lines. It is seen that iot, big data, autonomous robots, simulation, system integration, cybersecurity and augmented reality (AR) applications, which are accepted as the main components of Industry 4.0, are being preferred for different purposes in different processes in the manufacturing lines. Digital transformation applications play an important role especially in manufacturing lines where continuous production is carried out.

One of the most demanded demands is that the process in this type of manufacturing lines will be continued continuously.

It is only possible to adhere to the deadlines given to the customer only if complete and error-free manufacturing takes place.

In manufacturing, both mechanical and industrial automation systems must be correctly designed and implemented to ensure continuity. However, the life span of the equipment used or the unpredictable external factors can cause malfunctions in the manufacturing processes. In such cases, the fastest way to detect and solve the malfunction is the greatest desire of the manufacturing managers. In this article, a real AR application which will provide the solution of a failure in manufacturing processes in the shortest time is discussed.

Keywords: Augmented reality; industry 4.0; fault solving; smart manufacturing

1. Introduction

Companies manufacturing all over the world face serious competition conditions within the globalizing world economy. They have to provide a reduction in production costs, greater production flexibility and more efficient processes in order to become more advantageous than the competition. These are one of the common demands of the enterprises that produce in almost every sector [1].

In order to achieve efficiency in production, the manufacturing line must be operated continuously and failure rates should be minimized. A manufacturing company must

deliver the order to the customer within the deadline. But this will not always happen. Mechanical and automation system failures may occur in the manufacturing line. In particular, devices that are not serviced on time can often cause malfunctions.

The concepts of digitalization and industry 4.0 have come to the fore in almost every sector in recent years. We see that Digital Transformation applications have started to be used in manufacturing lines in recent years. Digital investments in the manufacturing sector have started to increase in recent years.

Smart Manufacturing is a key component of a broader impulse for Industry 4.0, and with the use of cloud systems, the use of data analytics and machine learning, a bridge between digital and physical environments is established through the Internet of Things (IoT) technologies combined with improvements in these digital environments [2].

When Industry 4.0 was first discussed in Germany for the first time in 2011, it was not expected that digitalization would receive such demand in the manufacturing sector in the world.

Germany is not alone in this passion, but is also carrying out a number of other EU-level initiatives [3] and China's Production 2025 initiative in China [4] to digitize and automate production to maintain competitiveness in highly globalized and competitive markets.

Smart Manufacturing (SM) enables companies to perform processes more efficiently, enabling faster demand fulfillment and lowering production costs. For this reason, SM can provide a competitive advantage for manufacturers who successfully implement it. In the process of SM adoption, building blocks play a critical role and can be seen as a prerequisite. Managers of manufacturing companies are interested in learning which building blocks are relevant and can help make their activities smarter and more competitive [5].

The efficient processes of SM systems can scale themselves depending on demand and make the SM highly sensitive to changes in demands [6].

Developments in production systems, production technology and machine tools have been developed for decades due to the introduction of new materials and new products that require new processing techniques, and many corporate strategies to minimize cost increase, quality and reliability, and maximize profit. The greatest impact on the evolution of production systems has been the increased variety of products that motivate the transition from mass production to more flexible, reconfigurable and variable production systems [7, 8].

The main components of Industry 4.0, IoT, big data, autonomous robots, simulation, system integration, cybersecurity and augmented reality (AR) applications have been used in production lines in recent years.

It is observed that digital conversion applications cause significant productivity increases especially in production lines where continuous production is performed.

One of the most demanded demands is the continuous production of the production lines.

It is only possible to deliver the products produced within the deadlines given to the customer only by the realization of complete and error-free manufacturing.

In order to ensure continuity of production, both mechanical and industrial automation systems must be designed and implemented correctly.

However, the life span of the equipment used or the unpredictable external factors can cause malfunctions in the manufacturing processes.

In such cases, the fastest way to detect and solve the malfunction is the greatest desire of the manufacturing managers.

This article describes a real AR application that will provide the solution of a failure in manufacturing processes as soon as possible.

2. Industrial Automation

In order for smart manufacturing to take place, complete and accurate industrial automation systems and a correct project must be realized. Industrial devices used in industrial automation systems must be selected and configured according to the process. PLC, HMI, distributed I / O, servo motor and drive systems are the most widely used devices in industrial automation systems.

2.1. PLC

Siemens S7-1200 CPU 1214 CPU was used in the system we commissioned. The specifications of this CPU are shown in table 1 below [9].

Table 1. The Technical specifications of Siemens S7-1200 CPU1214

Characteristics	CPU 1214C
Variants	
Work memory, integrated	100 KB
Load memory, integrated	4 MB
Memory card	
Digital inputs/outputs, integrated	14/10
Analog inputs, integrated	2
Analog outputs, integrated	0
Process image	1024 bytes for inputs, 1024 bytes for outputs
Expansion by signal board	Max. 1
Expansion by signal modules	Max. 8
Expansion by communication modules	Max. 3

2.2. HMI

Siemens HMI KTP700 Basic operator panel is used in the system we commissioned. The technical specifications of this HMI are shown in table 2 below [10].

Table 2. The Technical specifications of Siemens HMI KTP700

General information	
Product type designation	KTP700 Basic color PN
Display	
Design of display	TFT widescreen display, LED backlighting
Screen diagonal	7 in
Display width	154.1 mm
Display height	85.9 mm
Number of colors	65 536
Resolution (pixels)	
● Horizontal image resolution	800 Pixel
● Vertical image resolution	480 Pixel

Backlighting	
● MTBF backlighting (at 25 °C)	20 000 h
● Backlight dimmable	Yes
Control elements	
Keyboard fonts	
● Function keys	
— Number of function keys	8
— Number of function keys with LEDs	0
● Keys with LED	No
● System keys	No
● Numeric keyboard	Yes; Onscreen keyboard
● alphanumeric keyboard	Yes; Onscreen keyboard
Touch operation	
● Design as touch screen	Yes
Installation type/mounting	
Mounting position	vertical
Mounting in portrait format possible	Yes
Mounting in landscape format possible	Yes
maximum permissible angle of inclination without external ventilation	35°
Supply voltage	
Type of supply voltage	DC
Rated value (DC)	24 V
permissible range, lower limit (DC)	19.2 V
permissible range, upper limit (DC)	28.8 V
Input current	
Current consumption (rated value)	230 mA
Starting current inrush I^2t	0.2 $A^2 \cdot s$
Power	
Active power input, typ.	5.5 W
Processor	
Processor type	ARM
Memory	
Flash	Yes
RAM	Yes
Memory available for user data	10 Mbyte
Type of output	
Acoustics	
● Buzzer	Yes
● Speaker	No

3. Augmented reality (AR)

Thanks to digitization in the field of cyber-physical system sensors or actuators, technological advances ensure maximum flexibility in all areas along the value chain. Different studies have shown that increasing digitalization affects the sector as well as the sector [11,12].

Digitalization will lead to a redesign of jobs, particularly in the manufacturing industry. This means new digital competencies by changing employee needs. Different technologies such as augmented reality can be used to support employees in building the necessary competencies [13].

In order to integrate digitalization into the company, adjustments are required in all areas from production to human resources (HR). A developing technological trend in this context is Augmented Reality (AR). AR devices are assigned to the digital support systems group. Their use facilitates the work of employees by providing on-demand data specific to the situation and enriching real-time situations with targeted information [14].

4. Augmented Reality Applications Example

The occurrence of a failure in the manufacturing processes hampers the entire production plan. Continuous and error-free manufacturing is often not possible. Sometimes the failure of the devices on the system or the mechanical failures in the manufacturing machines can cause the production to stop.

The downtime in the production which starts with the occurrence of the malfunction causes serious material damages to the enterprise. Sometimes manufacturing companies may experience serious loss of confidence and image towards their customers due to these postures. In such cases, they may face very serious compensation payments to their customers. Worst of all is the risk of losing customers.

Considering such situations, all manufacturing companies demand that the faults that occur during the manufacturing process be solved as soon as possible.

Resolving the defect as soon as possible means minimal manufacturing loss. This means the least loss.

With the augmented reality application, we have made, it is planned to solve a malfunction in the manufacturing enterprises as soon as possible.

The PLC control system, which controls the manufacturing processes, informs the operator of this malfunction by an alarm lamp, buzzer or message when a malfunction occurs.

In this application, the alarm tag that occurs in PLC is shown on the screen of smart glasses. To do this, the alarm tag read in the PLC is saved in the MS SQL data base. A communication is provided between the MS SQL data base and the android operating system of the smart glasses. When a defined alarm is received, this alarm information is displayed on the display of the smart glasses. The display of the smart glasses shows step-by-step scenarios of the alarm that has occurred. The operator applies the scenarios related to the fault resolution shown on the screen of the smart glasses one by one and provides the solution of the fault as soon as possible.

5. Conclusion

The use of AR solutions in manufacturing processes, where digitization is progressing rapidly, offers numerous potentials for the industry. Using these solutions, the fault can be quickly detected and resolved. Many different applications like this can shorten process times, train employees individually and reduce error rates. Including AR technologies in education and further development can result in a significant reduction in costs. This study, which is used to solve faults quickly, shows how more applications of AR technologies can be developed in the industry. Using AR technologies, an innovative learning environment such as quality control practices and digital education can be achieved.

References

[1] Kayar, A. Ayvaz, B. and Öztürk, F. (2018). Akıllı Fabrika, Akıllı Üretim: Endüstri 4.0'a Genel Bakış. 1.International Eurasian Conference on Science, Engineering and Technology (EurasianSciEnTech 2018), 22-23 Nov., 2018; p. 1651-1658. Ankara, 2018.

[2] Tuptuk, N. and Hailes, S. (2018). Security of smart manufacturing systems, Volume 47, April 2018, Pages 93-106

[3] Parliment E. Industry 4.0. 2016, http://www.europarl.europa.eu/RegData/etudes/STUD/2016/570007/IPOL_STU(2016)570007_EN.pdf

[4] Research H. China's 13th Five-Year Plan: made in China 2025 and Industrie 4.0 cooperative opportunities. 2016, http://economists-pick-research.hktdc.com/business-news/article/Research-Articles/China-s-13th-Five-Year-Plan-Made-in-China-2025-and-Industrie-4-0-Cooperative-opportunities/rp/en/1/1X32LK39/1X0A6AZ7.htm.

[5] Mittal, S. Khan, M.A. Romero, D. and Wuest, T. (2019). Building Blocks for Adopting Smart Manufacturing, 47th SME North American Manufacturing Research Conference, Penn State Behrend Erie, Pennsylvania, 2019, Procedia Manufacturing 34 (2019) 978–985

[6] Davis, J. Edgar, T. Porter, J. Bernaden, J. and Sarli, M. (2012). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. Computers & Chemical Engineering 2012;47:145-56.

[7] ElMaraghy, H.A. (2006). Flexible and Reconfigurable Manufacturing Systems Paradigms, Special Issue of the International Journal of Flexible Manufacturing Systems (IJFMS), Vol. 17, No. 4, pp. 261-276.

[8] Wiendahl, H.-P. ElMaraghy, H.A. Nyhuis, P. Zaeh, M. Wiendahl, H.-H. Duffie, N. and Kolakowski, M. (2007). Changeable Manufacturing: Classification, Design, Operation, Keynote Paper, CIRP Annals, Vol. 56/2, pp. 783-809.

[9] Siemens Web Site, "The Technical Specifications of Siemens S7-1200 CPU1214", [Online]. https://mall.industry.siemens.com/mall/en/tr/Catalog/Products/10045649?tree=CatalogTree
as sited on 30/08/2019.

[10] Siemens Web Site, "The Technical Specifications of Siemens HMI KTP700", [Online]. https://support.industry.siemens.com/cs/pd/302298?pdti=td&dl=en&lc=en-WW
as sited on 30/08/2019.

[11] Frey, C. and Osbourne, M. (2013). The Future of Employment: How Susceptible are Jobs to Computerization? University of Oxford, Oxford, (2013).

[12] Spath, D. Ganschar, O. Gerlach, S. Hämmerle, M. Krause, T. and Schlund S. (2013) Studie Produktionsarbeit der Zukunft – Industrie 4.0. Fraunhofer Verlag, Stuttgart, (2013).

[13] Sorko, S.R. and Brunnhofer, M. (2019) Potentials of Augmented Reality in Training, 9th Conference on Learning Factories 2019, Procedia Manufacturing 31 (2019) 85–90

[14] Peddie, J. (2017). Augmented Reality. Where We Will All Live. Springer International Publishing, Cham, (2017).

Time-Frequency Analysis of the EEG Signals: Visual Identification of Epileptic Patterns

Ezgi Özer
Dept. of Statistics
Fac. of Science and Letters
Mimar Sinan Fine Arts University
ezgiozerhs@gmail.com

Ozan Kocadağlı
Dept. of Statistics
Fac. of Science and Letters
Mimar Sinan Fine Arts University
ozan.kocadagli@msgsu.edu.tr

Arnaldo Batista
Dept. of Electrical Eng.
Fac. Of Science and Tech.
Lisbon Nova University
agb@fct.unl.pt

Abstract

This study presents an efficient approach that ensures a visualization of Electroencephalo-gram (EEG) signals for onset and offset detection of epileptic seizures. This visualization is performed both in the time and time-frequency domain. Essentially, this approach is based on the continuous wavelet transform with spectral analysis and time series entropy in the context of the time-frequency signature analysis of EEG signals. Specifically, this framework provides examining non-stationary EEG signals in detail, and reveals the latent components that are the most relevant to the seizure or non-seizure patterns. In addition, the visualization allows the neurologists to figure out seizure dominant patterns in the frequency band. As a result, the proposed visualization mechanism ensures a useful tool for an automatic detection of EEG patterns via very less time-consuming and high efficiency at any frequency level.

Keywords: Epiletic Seizure Detection; Time-Frequency; Wavelet Transforms; Entropy.

1. Introduction

Epilepsy is one of the common neurological disease, occurring by transient abnormal electrical discharge in nerve cells. This disease depends on many trigger factors including genetic, physiologic, brain damage, etc. According to the World Health Organization (WHO) 2019 data, it is estimated that epilepsy, that can be seen in all age groups, affects approximately 70 million people in the worldwide [1]. The epilepsy detection process requires detailed evaluations, including the medical history, blood test, Electroencephalography (EEG), and brain imaging tools such as computed tomography (CT) and magnetic resonance imaging (MRI) scans. Among these tools, EEG is an adequate tool to record the functional and physiological changes in the brain, and to obtain useful information including brain electric activity, possible types of seizures. To present changes in the signals and to obtain the familiar EEG patterns or other brain electric features, visualization is an important method. Generally, the visualization method is based on the spectrum analysis with well-known transformation techniques such as Fourier

transform, Wavelet transform, Hilbert, etc. Specifically, entropy and energy are mostly used to extract EEG patterns in different frequency bands in the spectrum analysis.

2. Motivation and Overview

In the context of epilepsy, identification of EEG patterns helps the neurologists to figure out crucial information during seizure recordings. The EEG signals can have unrelated patterns or the spectral amplitude changes at a different frequency resolution.

In the detection of epileptic seizures, the automated resolution tools are very important to reduce time consuming and false inspection ratio against the classical visual analysis. Generally, these tools take into account some statistical, signal processing methods and AI techniques. Specifically, the signal processing methods are used to extract some useful information from the latent patterns. Wavelet transforms are quite useful to study this latent information, related to epileptic seizures or not in the brain waves [2, 3]. In this context, various wavelet functions provide different window resolutions to reduce irrelevant information, including non-epileptic EEG components and noise. Besides, wavelet transforms and spectral analysis are useful to find reliable spectral patterns using different frequency resolutions [4]. Alternatively, the entropy measure can evaluate uncertainty in patterns exactly [5]. Moreover, during seizure detection, it will be possible to reduce false epileptic case detections due to the similarity of seizures with any of these situations or to find whether there is any relationship between seizure and other any component.

3. Meterials and Methods

3.1 Data

In this study, The EEG data set, recorded by the Children's Hospital Boston, was considered. Basically, this data set includes 916 hours of continuous scalp EEG sampled at 256 Hz with 16 bit-resolution from 23 pediatric patients and the International 10-20 bi-polar system of EEG electrode position [6].

3.2 Wavelet Transform

Wavelet Transform (WT) is used to analyze the non-stationary signals like EEG. WT uses different window sizes and different wavelet functions at different frequencies, allowing the wavelet stretched or compressed [7, 8].

3.3 Spectral Analysis

Spectral analysis is based on analysis of frequency bands using different techniques and gives the variation of energy [4, 9].

3.4 Entropy

Entropy is used to measure the amount of uncertainty or randomness in a statistical sense, in the signal for the pattern from engineering to medicine fields. [10]. In this study, to analyze EEG components, Approximate Entropy (ApEn), Sample Entropy (SampEn), Fuzzy Entropy (FE), Wavelet Entropy (WE), Shannon Entropy (SE), Conditional Entropy (CE), Corrected Conditional Entropy (CCE) and Permutation Entropy (PE) are used [5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

3.6 Analysis

In this analysis, to detect seizure patterns, the following tools are utilized: WT, spectral analysis and various entropy measures. This framework allows exploring time-varying multichannel EEG data sets and provides a visualization in EEG patterns. To obtain this visualization, a toolbox was developed wherein one can select the sampling frequency, wavelet function and entropy's parameters. In this way, it can mainly deal with the components of interest. Also, this method gives the energy resolution in frequency and time domain to detect latent EEG patterns. Moreover, all this information can be obtained from both all channels and just user selected channels. For instance, the following figures give some information about the energy and entropy for FP2-F4 and T8-P8 channels, obtained from one patient's record.

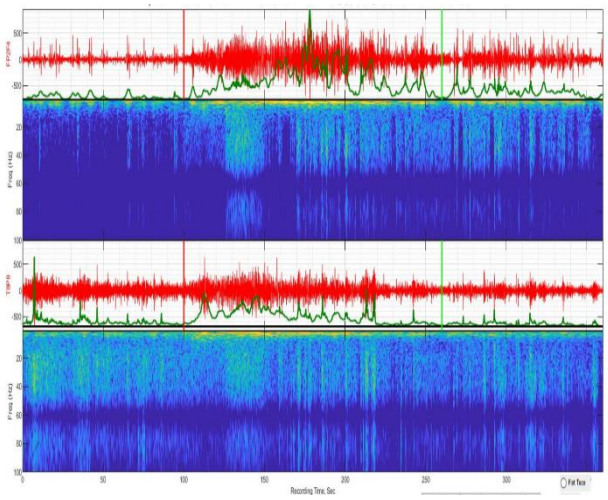

Figure 1. Wavelet-Time Frequency for FP2-F4 and T8-P8

As seen from Figure 1, the spectrum analysis using energy information helps to figure out similar patterns during seizure and in other conditions at different frequency values. During the seizure, it can be seen the high energy density in the low frequency band. Also, at any selected time, the different patterns related to seizure can be seen before and after the seizure.

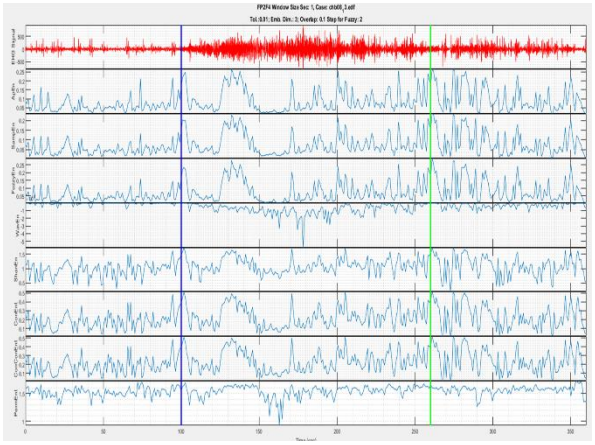

Figure 2. Entropy Plots for FP2-F4

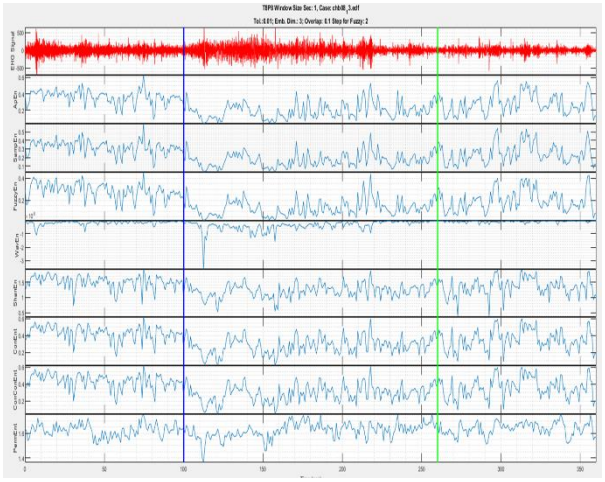

Figure 3. Entropy Plots for T8-P8

As seen from Figure 2 and 3, it can be observed that all the entropy measures produce similar plots for both channels, except of wavelet entropy (WE). Although all the entropy measures are sensitive to the model parameters, WE depends on the wavelet function without controlling any parameter. From analysis, it can be inferred that WE gives

a more reliable performance comparing the others.

4. Conclusion

As a visualization method, it was developed the Epileptic Seizure Scan (ESS) which marks seizure patterns and other EEG components, including artifacts, high frequency oscillations or typically brain waves like delta, theta, alpha, beta and gamma in different frequency bands. It is possible to see the EEG energy variations in the brain before, during and after the seizure and to extract the similar pattern with epileptic seizures and other components. Also, the obtained results for all channels can be mutually compared. And thus, the proposed procedure can help to find seizure patterns in the channels before seizure onset.

5. References

[1] World Healt Organization Report (2019). https://www.who.int/news-room/fact-sheets/detail/epilepsy

[2] Raghu, S., Sriraam, N., Temel, Y., Vasudeva Rao, S., Hegde, A.S., Kubben, P.L. (2019). Performance evaluation of DWT based sigmoid entropy in time and frequency domains for automated detection of epileptic seizures using SVM classifier , *Computers in Biology and Medicine*, 110, 127-143.

[3] Sudalaimani, C., Sivakumaran, N., Elizabeth, T.T., Rominus, V.S. (2019). Automated detection of the preseizure state in EEG signal using neural networks, *Biocybernetics and Biomedical Engineering*, 39,(1), Pages 160-175.

[4] Javed, E., Croce, P., Zappasodi, F.,and Gratta, CD, 2019. Hilbert Spectral Analysis of EEG Data Reveals Spectral Dynamics Associated with Microstates, *Journal of Neuroscience Methods*, 325, 108317.

[5] Luo, H., Qiu, T., Liu, C., and Huang, P., (2019). Research on fatigue driving detection using forehead EEG based on adaptive multi-scale entropy, *Biomedical Signal Processing and Control*, 51, 50-58.

[6] https://physionet.org/content/chbmit/

[7] Daubechies I. Ten Lectures on Wavelets. Philadelphia: SIAM; 1992. p 109-120.

[8] Gao RX, Yan R. (2011). Wavelets: Theory and Applications for Manufacturing. *Springer*.p 49-68.

[9] Kumar, N., Kumar, J. (2016). Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study, Procedia Computer Science, 84 (2016), 70-78.

[10] Gamal, A.E., and Kim, Y.H. (2011). Network Information Theory, Cambridge University Press, UK.

[11] Acharya, U. Rajendra Fujita, H. Sudarshan, Vidya K. Bhat, Shreya Koh, Joel E.W. (2015). Application of entropies for automated diagnosis of epilepsy using EEG signals: A review. *Knowledge-Based Systems*, 88, 85-96.

[12] Arunkumar N., Ramkumar K., Venkatraman V., Abdulhay, E, Fernandes, S.L., Kadry, S., Segal, S. (2017). Classification of focal and non focal EEG using Entropies, *Pattern Recognition Letters*, 94,112-17.

[13] Kang, J., Chen, H., Li, X., Li, X. (2018). Clinical study EEG entropy analysis in autistic children, *Journal of Clinical Neuroscience*, 62, 199-206.

[14] Zhang, T., Chen, M., and Li, M. (2018). Fuzzy distribution entropy and its application in automated seizure detection technique, *Biomedical Signal Processing and Control*, 39, 360-77.

[15] Azami, H., and Escudero, J. (2017). Refined composite multivariate generalized multiscale fuzzy entropy: A tool for complexity analysis of multichannel signals, *Physica A: Statistical Mechanics and its Applications*, 465, 261-76.

[16] Kang, J., Chen, H., Li, X., and Li, X. (2018), Clinical study EEG entropy analysis in autistic children, *Journal of Clinical Neuroscience*, 62, 199-206.

[17] Millan, PC., Garcia-Ferro, MA., Llanes-Estrada, FJ., Riojano, AP., Garcia, EMS. (2018). Shannon Entropy and Particle Decays, *Nuclear Physics B*, 930 (2018), 583-596.

[18] Toranzo, I.V., and Dehesa, J.S. (2019). Exact Shannon entropies for the multidimensional harmonic states, Physica A: Statistical Mechanics and its Applications, 516, 273-279.

[19] Anton, M.U., Karsten, K. (2014). Conditional entropy of ordinal patterns, *Physica D: Nonlinear Phenomena*, 269, 94-102.

[20] Malings, C., Pozzi, M. (2016). Conditional entropy and value of information metrics for optimal sensing in infrastructure systems, *Structural Safety*, 60, 77-90.

[21] Xiaolin, Y., Chong, Z., Longfei, S., Jianbao, Z., Nini, R. (2018). Estimation of the cortico-cortical and brain-heart functional coupling with directed transfer function and corrected conditional entropy, *Biomedical Signal Processing and Control*, 43, 110-116.

Feature Selection Approaches for Machine Learning Classifiers on Yearly Credit Scoring Data

Damla Ilter[1]
Mimar Sinan Fine Arts University
Statistics Dept.
Istanbul, Turkey
damla.ilter@msgsu.edu.tr

Ozan Kocadagli[2*]
Mimar Sinan Fine Arts University
Statistics Dept.
Istanbul, Turkey
ozan.kocadagli@msgsu.edu.tr

Nalini Ravishanker[3]
University of Connecticut
Statistics Dept.
Connecticut, USA
nalini.ravishanker@uconn.edu

Abstract

Credit scoring is one of the efficient methods to measure systematic risk when financing individual customers in the banking sector. While past literature mainly focused on cross-sectional data at a given time, there is increasing interest in credit scoring based on information that a bank has over time. This study describes such analysis on "The Irish Dummy Banks" data, and provides an efficient feature selection procedure based on data over several years, using popular approaches such as: Logistic Regression (LR), Recursive Partitioning (RP), Random Forest (RF), Conditional Inference Trees (CIT), Support Vector Machine (SVM), Least Absolute Shrinkage Selection Operator(LASSO), etc. According to our analysis results, SVM (Radial Basis Function) outperforms the other algorithms with respect to area under the curve (AUC), while do best the Kolmogorov Smirnov (KS) and Gini Index in terms of accuracy and reliability.

Keywords: Credit Scoring; Machine Learning; Classifiers; Feature Selection Methods.

1. Introduction

Credit scoring is a leading problem in financial modeling. In recent years, parallel to the financial crisis in the world economy, as well as many economic factors, the number of non-performing loans increased and thus the importance of credit scoring models increased. Credit scoring is an important research area in the process of measuring and evaluating many situations such as systematic and non-systematic risks, follow-up of payments and legal processes. In general terms, it is aimed to analyze the economic, and financial risks through early warning mechanisms, to determine the international fragilities in advance, and to reduce or eliminate the elements that would pose a risk to the system.

It is obvious that for improving credit scoring process and increasing the accuracy, decision making tools are important to research are in terms of making much more efficient results [1, 2, 3, 4, 5].

In [6, 7, 8, 9, 10, 11] developing nonlinear models and solving algorithms in finance analysts encounter specific problems such as regression decision trees, pattern recognition, classification, clustering, etc. In order to

* Corresponding Author e-mail: ozan.kocadagli@msgsu.edu.tr

handle these kinds of problems efficiently, decision support systems are still being developed. In [12, 13, 14], RF and SVM methodologies accurately pinpoint creditworthiness of the clients involved. SVMs can be used in feature selection methods to emphasize the importance of risk detection. Bayesian techniques have also been successfully applied to neural networks in the context of both regression and classification problems [15, 16, 17].

There is a growing interest in the use of artificial intelligence techniques in statistics to model non-linear multivariate problems. These techniques are capable of supervised or unsupervised learning and finding meaningful solutions without the need to specify the relationships between variables [18, 19]. They are useful for analyzing problems that are either poorly defined or not clearly understood.

In the context of credit scoring of customers, there are numerous applications of classification methods with feature selection algorithms in the banking sector. Credit scoring is one of the major activities in the banking sector and is an efficient method to measure the systematic risk when financing individual customers. For this reason, the banks make a special effort to minimize the systematic risk related to the credit scoring.

In this study, to estimate more robust credit scoring models, various machine learning approaches with feature selection algorithms are considered. Thus, performances of various machine learning classifiers are investigated over a credit scoring data set over several years, instead of only one cross-sectional data for a specified year.

2. Data description

We study the "*Irish Dummy Banks"* data, obtained from Kaggle, see [20]. This is a fictitious lending bank in Ireland which provides funds to potential borrowers and makes profits depending on the risk the bank takes from Lending Club Information. The data set includes records on 465682 customers over seven years (2008-2014). In each year suppose there are a different number of customer for each month. We have the number of loans as well as the number of approved (good) loans and the number of denied (bad) loans are available. Additionally, there is data on 17 features including annual income, employment length, final date, grade, homeownership, id, income, installment, interest payment, interest rate, issue date, loan amount, loan condition, purpose, region, term, and total payment. See [20] for detailed descriptions.

3. The proposed model

This study contains three main stages: (1) data pre-processing, (2) modeling and (3) feature selection (classification). In all of these stages; the optimal parameters are determined by some classification accuracy measure for each algorithm.

To apply the proposed model to the research data, initially it must be pre-processed. There are some data preprocessing methods (without any sequence of operations) that apply to data preparation and cleaning as listed in the following: (i) removing some of the records as well as the unvalued features because of not necessary for these analyses. (ii) integrating data (iii) converting quantitative variables into financial ratios, or the quantitative variables into qualitative variables (iv) eliminating some of the ineffective features (v) data visualization.

Model complexity is a big challenge for the analysts, because controlling many parameters requires extra effort and time. For this reason, it is mostly overlooked by researchers. Given this situation, this study aim to address the issue of complexity. AUC, KS, and Gini Index as a criterion of complexity information was evaluated as an accurate and reliable modeling framework in this study. Consequently, the

percentages of AUC, KS and Gini Index were compared. These results indicate that accuracy is a sensitive measure in determining the model as given in Table 1.

The classification techniques have been used to many applications in Statistical problems. The analysis of such techniques and its significance must be interpreted correctly for evaluating different learning algorithms. There are two types of classification techniques in the literature: One of them is Binary Classification Models and the other is Multi-Class Classification Models. The classification technique used in this study is Binary classification technique.

The following machine learning approaches with feature selection are utilized: LR (Stepwise), LR (with important variables), RP(Basic), RP (Bayesian), RF, Conditional RF, Improve Results of LR using RF, CIT, SVM (Vanilladot), SVM (RBF), SVM (Polydot), SVM (Tanhdot), SVM (Laplacedot), SVM (Besseldot), SVM (Anovadot), SVM (Splinedot), Lasso (Lars), Lasso (GLMnet). Here, SVM is a classification algorithm, while SVM (RBF) is a derivate of SVM. The results are shown in Section 4.

4. Experimental results and discussions

In this analysis, machine learning classifiers were compared to each other using the 10-fold cross-validation with 70% of the data to train and 30% to test. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data. In this study, the different methods mentioned in Section 3 were compared.

The performance of these classifiers were compared to each other according to some criteria such as AUK, KS and Gini Index. From Table 1, we see that SVM (RBF) performs better than other algorithms. Specifically, AUC, KS, and Gini Index are 98%, 90.17%, and 96%, respectively.

Table 1. Comparison of model performances based on "years".

Year	Model	n	AUC	KS	Gini
2008	Lasso (GLMnet)	2393	97.76	89.49	95.52
2009	SVM (RBF)	5281	97.54	89.08	95.08
2010	SVM (RBF)	12537	97.27	87.22	94.54
2011	SVM (RBF)	21721	97.49	88.44	94.98
2012	SVM (RBF)	53367	97.55	89.04	95.01
2013	SVM (RBF)	134755	98	90.17	96
2014	SVM (RBF)	235628	97.67	87.55	95.34

The results of the feature selection methods applied separately for each year are shown in Table 2. For instance, in 2008, the selected feature were “total payment”, “installment” and “region”. It is interesting to note that “total payment” and “installment” are picked as important features for each year.

Table 2. Important variables based on "years" after feature selection.

2008	2009	2010	2011	2012	2013	2014
total payment	annual income	final date	issue date	issue date	issue date	issue date
installment	income	employement length	final date	final date	final date	final date
region	loan amount	annual income	employement length	employement length	home ownership	employement length
	total payment	loan amount	annual income	annual income	annual income	home ownership
	installment	term	loan amount	loan amount	loan amount	annual income
		purpose	term	purpose	term	loan amount
		interest rate	purpose	interest rate	purpose	term
		total payment	interest rate	grade	interest rate	purpose
		installment	total payment	total payment	grade	interest rate
			installment	installment	total payment	grade
					installment	total payment
						installment
						region

5. Discussion and future direction

One of crucial issue in credit scoring applications is that the analysts mostly focus on cross-sectional data sets. For this reason, many latent patterns might be overlooked in

analysis. To figure out these patterns, credit scoring data sets depending on time can be considered.

Basically, this research deals with modelling the credit scoring data sets in the case of time series framework. Actually, this framework allows the analysts to do more efficient temporal analysis with respect to unit of time. Thus, this approach helps to show up many remarkable features that plays important roles on good loan percentage.

The proposed approach provides substantial advantages in terms of feature selection with over methods, using various classifiers and controlling accuracy. The classification algorithms and their derivatives were first applied in order to obtain suitable features with high classification performance. According to the results, the performance of SVM (RBF) is superior to the other algorithms with respect to AUC, KS, and Gini Index. It is also the most accurate and reliable modeling framework over all the years in the data set. Further note that AUC, KS, and Gini Index are good evaluation measures.

The proposed procedure enables finding efficient features, increasing their accuracy, and estimating a model for each year/month to determine the important variables for classification.

Future research will consist of constructing an observed time series of monthly rates of good loans from 2008-2013 (training data) and predicting 2014 (test data) using suitable time-series approaches. Comparison to predicted rates of good loan under each claim: the method of their derivative is the aim of interest.

References

[1] Soui, E. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms, 126, 144–157.

[2] Olaniyan, R. and Maheswaran, M. (2017). Recent Developments in Resource Management in Cloud Computing and Large Computing Clusters. *Research advances in Cloud Computing.* 237-261.

[3] Goh, R. Y. and Lee, L. (2019). Credit scoring: A review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019, 30.

[4] Oreski, S., Oreski, D., and G., O. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with Applications*, 39, 12650–12617.

[5] Xiao, H., Xiao, Z. and Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing*, 43, 73-86.

[6] Bhatia, S., Sharma, P., Burman, R., Hazari, S., and Hande, R. (2017). Credit scoring using machine learning techniques. *Knowledge-Based Systems*, 161(11), 975–8887.

[7] Yufei, X. and Liu, N. (2017). A boosted decision tree approach using Bayesian hyperparameter optimization for credit scoring. *Expert Systems with Applications*, 78:225– 241.

[8] Khasman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), 6233-6239.

[9] Wooldridge, M. J. (2013). Introductory econometrics: A modern approach. *South-Western Cengage Learning* , 584-671.

[10] Ionescu, M. (2008). Use of Neural Networks in Business Process Modeling. *Management Perspectives in the Digital Era*, Politehnica Universit of Bucharest, Romania.

[11] Abdou, H., Pointon, J., and El-Masry, A. (2018). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Science Direct*, 35, 1275–1292.

[12] Roy, A. G. and Urolagin, S. (2019). Credit risk assessment using decision tree and support vector machine-based data analytics. *Springer*.

[13] Belotti, T. and Crook, J. (2009). Support vector machines for credit scoring and the discovery of significant features. *Expert Systems with Applications*, 3302-3308.

[14] Chen, W., Ma, C., and Ma, L. (2009). Mining customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, 36(4), 7611–7616.

[15] MacKey, D. (1992). Bayesian Methods for Adaptive Models. *California Institute of Technology*, Pasadena, California.

[16] Neal, R. (1996). Bayesian Learning for Neural Networks. *Graduate Department of Computer Science, the University of Toronto*.

[17] Ilter, D. and Kocadagli, O. (2019). Credit scoring by artificial neural networks based cross-entropy and fuzzy relations. *Sigma Journal of Engineering and Natural Sciences*.

[18] Ghodselahi, A. and Amirmadhi, A. (2011). Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modelling and Optimization*, 1(3), 243-249.

[19] Abellan, J. and Castellano, J. G. A. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73, 1–10.

[20] Kaggle website: www.kaggle.com/mrferozi/loan-data-for-dummy-bank.

Part VIII

Poster (Abstract)

Statistical Properties and Modeling of Stable-like Word Count Time Series in Nation-wide Language Data

Hayafumi Watanabe*[1]

[1]Kanazawa University College of Science and Engineering

Languages are one of typical complex systems, which have the well known language independent statistical laws such as Zipf's and Heap's. In this study, we investigate the dynamical statistical properties in languages by using massive nation-wide databases related to word usage that has developed in the past 10 years. Especially, we focus on the stability or the slowness of change in the usage of already popular words from the viewpoint of diffusion on a complex system and show that common logarithmic diffusion (i.e. very slow diffusion or change) is approximately observed by some languages or media [1]. The logarithmic diffusion i.e. the diffusion characterized by the logarithmic function, has been extensively studied theoretically in physics, but has hardly been observed empirically. In this study, we find the ultraslow-like diffusion of the time-series of key-word count of already popular words on three different nationwide language databases: (i) newspaper articles (Japanese), (ii) blog articles (Japanese), and (iii) page views of Wikipedia articles (English, French, Chinese, and Japanese). In addition, by mathematical modeling, we show that this diffusion is basically explained by the random walk model with the power law forgetting with the exponent $\beta \approx 0.5$, which is related to the fractional Langevin equation (or the autoregressive fractionally integrated moving average model). From the viewpoint of our model, The exponent β characterises speed of forgetting and $\beta \approx 0.5$ is corresponding to (i)the border between the stationary and the nonstationary and (ii)the right in the middle dynamics between the IID noise for $\beta = 1$ and the normal random walk for $\beta = 0$. Thirdly, we proposed the generative model of the time series of word counts of already popular words. The generative model of the time series of word counts of already popular words, which is a kind of a Poisson process with the Poisson parameter sampled by the above-mentioned the random walk model, can almost reproduce not only the empirical mean squared displacement but also the power spectrum density and the probability density function. In the presentation, we would like to show some practical applications of our results.

Keywords: Languages Statistics; Online Social Media Data; Long Memory Time Series;Poisson Point Process; Statistical Physics; Diffusion; Complex Systems

References

[1] H. Watanabe (2018). Empirical observations of ultraslow diffusion driven by the fractional dynamics in languages. Physical reveiew E, 98, 012308.

*Corresponding author: hayafumi@gmail.com

Part IX

Poster (Full)

The Examination of Real Estate Prices in Istanbul by Using Hybrid Hierarchical K-Means Clustering

Betul Kan-Kilinc
Science Faculty,Department of Statistics Eskisehir Technical University
bkan@eskisehir.edu.tr

Ilkay Tug
Science Faculty,Department of Statistics Eskisehir Technical University
ilkaytug8@gmail.com

Abstract

We present a hybrid approach to combine the merits of the hierarchical clustering and k-means clustering approaches. In this study, the determinants of the real estate prices are examined by hybrid hierarchical k-means clustering on relevant features at 39 different districts in Istanbul. For this purpose, the prices and properties of 312 real estates are recorded from an online web source. The data contains the variables of size in m^2, age, rooms, floor and prices for rent. Additionally, the results from both methods were examined with several graphics and dendrograms. K-means clusters and also the clusters formed by the principal component analysis were extracted using R software.

Keywords: hierarchical clustering; k-means; hybrid

1. Introduction

Clustering is one of the most useful applicationsin data mining process for discovering groups or clusters and identifying interesting distributions and patterns in the underlying data. Partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters is called clustering. The popular common clustering algorithms are k-means, hierarchical; agglomerative, divisive methods, non-hierarchical methods, density base methods, grid based methods, and fuzzy clustering. However, some of these methods have some limitations such as specifying the number of clusters in advance. Another restriction is a mixture of data types (categorical, ordinal, interval) can be complicated for a cluster analysis. For a different number of data type, the distance measure between the observations would be different [3, 4].

K-means is an unsupervised learning algorithm where the user do not assign any label so that the data reveals the underlying structure of data by itself. When the number of variables is excessive, principal component analysis (PCA) is a method of expressing variables with fewer linear components than the number of original variables without correlations [8, 10, 12]. It provides support for clustering by creating shapes around theprime component.

In K-means algorithm being a partitioning method, a random set of observations are chosen as the initial centers. The reason why a hybrid approach was chosen is the final k-means clustering solution is very sensitive to the initial random selection of cluster centers and outliers. The result might be (slightly) different each time you compute k-means. Therefore, a solution to this problem can be to combine the k-means and hierarchical clustering methods [2].

The aim of this study is to identify the districts with common features of real estates

in terms of prices and age, specifically. The other features are number of rooms, size in m^2 and its floor where the real estate is

located. For this purpose, thirtynine districts are considered in Istanbul. The properties of the real estates which for rent are recorded from an online web source. The differences of the districts in Istanbul are examined by a hybrid clustering method. Clustered over districts are observed on various graphics. All the computations are performed by using R software.

2. Clustering Methods and Distance Measurements

Observations that are similar are grouped together when clustering the data. For this aim, many forms can be define to be able to compute the distance between two objects in the data.

2.1. K-Means Clustering

K-means clustering is one of the widely used unsupervised machine learning algorithms. Basically, the data set is partitioned into a set of k clusters (i.e. k groups), where k represents the number of clusters prespecified by the researcher. It classifies observations in multiple clusters, such that observations within the same cluster are as similar as possible (i.e., high intra-class similarity or within-cluster variation), whereas observations from different clusters are as dissimilar as possible (i.e., low inter-class similarity). Also, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster.

The standard k-means algorithm provided by [7], which defines the total within-cluster variation as the sum of squared distances Euclidean distances between observations and the corresponding centroid given in Eq.2.1.

$$HW(c_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (1)$$

where x_i is the observation included by the cluster C_k and μ_k is the mean value of the observations assgined to the cluster C_k.

Each data point x_i is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers μ_k is minimized.

$$wss = \sum_{k=1}^{k} HW(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (2)$$

The total within-cluster sum of square measures the compactness (i.e goodness) of the clustering and we want it to be as small as possible.

One disadvantages of k-means is that the researcher specifies the number of clusters. For determining the optimal clusters, the three most popular methods are defined: Elbow method, Silhouette method or Gap statistic.

2.2. Distance Measures

In cluster analysis, the similartiy measure is based on the distance. There are often used distance measures such as Minkowski distance, Euclidean, squared Euclidean, Chebyshev and Manhattan distance. Selection of the best distance measure is not straightforward and depends on the nature of the dataset. The measure for continous variables is preferably chosen as correlation type or distance [4]. The most common distance measure is known as Euclidean distance that defines the shortest distance between two data points.

The Euclidean distance in two dimensions is the length of the straight line connection to two points x and y. In n dimension the Euclidian distance is defined as:

$$| x - y | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

Euclidean distance is known sensitive to large values or outliers [4]. Also, the Euclidean distance is effective when each variable is scaled or has same unit [6].

3. Data Description

The real estate industry is significantly important for economic activities. It is true not only for economic development but also it plays a decisive role in the world economy [1]. In practice, real estate market provide shelter to other domestic services [5]. This market has impacts on families, firms, banks and governmental buildings and many others. The trading of real estate has three components: commercial (warehouses), industrial (factories, farms) and residential (houses, towns-homes) [9]. The residential form has two options: rent and sale. Understanding the characteristics of sales and rents has undeniable importance for economic conjuncture. In this study, we investigate the decision surrounding of rents.

There are various studies that examine the changes in prices in different locations in different countries [13]. Mikelbank (2004) investigated US suburbs by a hierarchical clustering and presented 10 distinct types of suburbs, specifically [11].

In this study, data collected from an online web platform where users can sell or buy products such as real estate, car, etc. It includes 312 collection of real estate with the following features (name of the variables) given in Table 1.

Table 1. List of variables and descriptive statistics.

Vlaue	Price (tl)	Land(m^c)	Age	Room	Floor
Min	550	75	0	2	0
Mean	1626	112.3	13.38	3	2
Max	4500	150	35	5	25

The city of Istanbul has 39 districts and we collected and recorded 8 observations that are not larger than $150m^2$ for each district. After computing the mean for each district, we then selected the closest observation to the mean in this district to be a representative value of real estate in the corresponding group with its features.

Scaling the data before the analysis is important in distance measuring as the data includes different types of variables. Each variable is scaled with mean 0 and standard deviation 1. However, this in itself presents another problem that is about to reduce the variability (distance) between clusters. This happens because if a particular variable separates observations well then, by definition, it will have a large variance (as the between cluster variability will be high). If this variable is scaled then the separation between clusters will become less. Despite this problem, standardisation is recommended [12, 4].

4. Results

In this section, we implement the cluster algorithms. First, k-means algorithm is performed by specifying k from 1 to 10 clusters. For each cluster, the total within cluster sum of squares (wss) is computed.

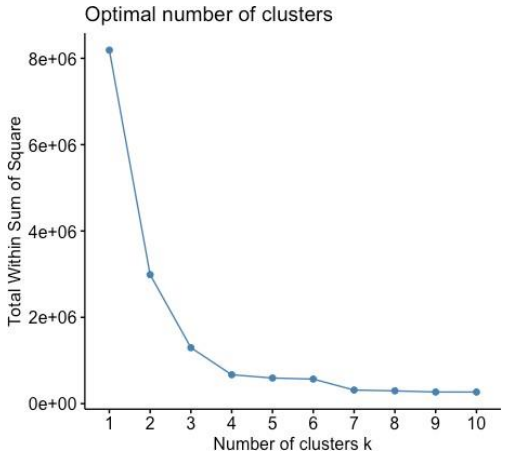

Figure 1. Optimum number of k by k-means clustering

Using the Figure 1 where the number of k is plotted aganist wss, the appropriate number of cluster is considered as 4. The centers of the clusters obtianed from k-means are summarized in Table 2 and represented in Figure 2.

Table 2. Center of clusters after standard k-means clustering

k	Price (tl)	Land(m^2)	Age	Room	Floor
1	0.28236	-0.90383	-0.63521	-1.07502	-0.50577
2	1.00261	0.81983	0.77666	0.89975	0.30596
3	-0.58156	0.28258	0.13609	0.32718	-0.43709
4	-0.92107	-1.02639	-0.94896	-1.07502	0.79301

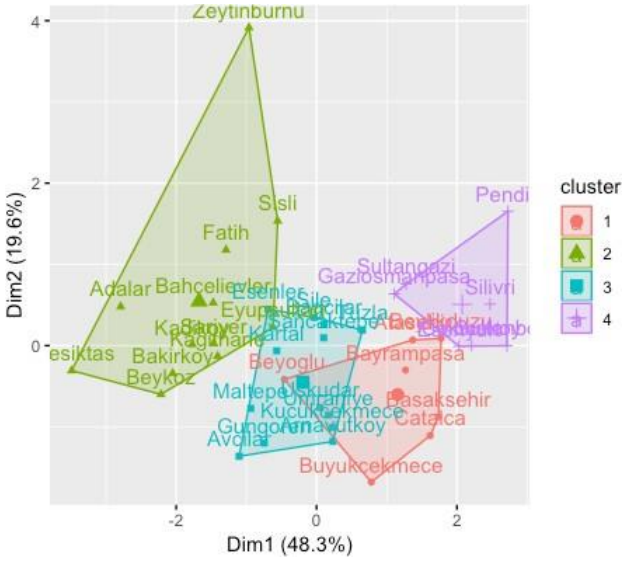

Figure 2. Clusters by standard k-means algorithm

There are potential disadvantages of k-means clustering such as prespefying the clusters, being sensitive to outliers and not providing a tree-based representations of the clusters. However, hierarchical clustering do not require a particular choice of clusters.

Next, hierarchical clustering is performed by specifying 4 clusters and the dendogram is depicted in Figure 3. In this figure, it can be seen that some of the observations are classified in different clusters by means of hierarchical clustering.

In Figure 4, the dendogram obtained from a hybrid approach is performed. In

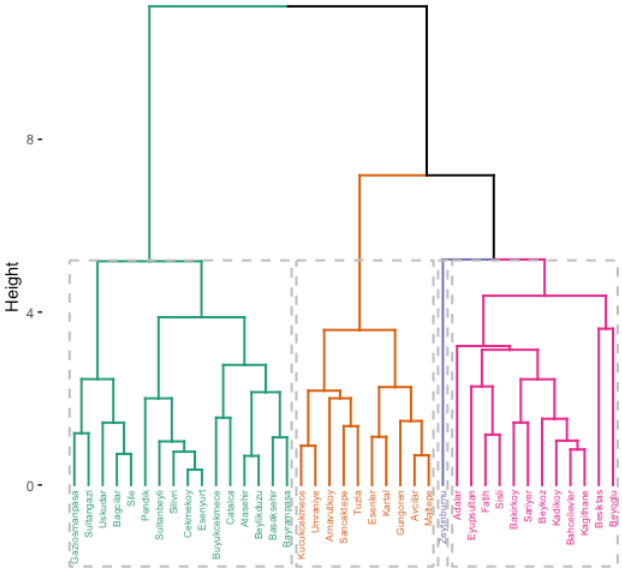

Figure 3. Dendogram by hierarchical clustering

this approach, the center (th mean) of each clusters are computed by hierarchical clustering. Cluster means or centers are computed as the means of the variables in clusters. For the initial selection of cluster centers, the cluster centers obtained from k-means algorithm are used. The difference

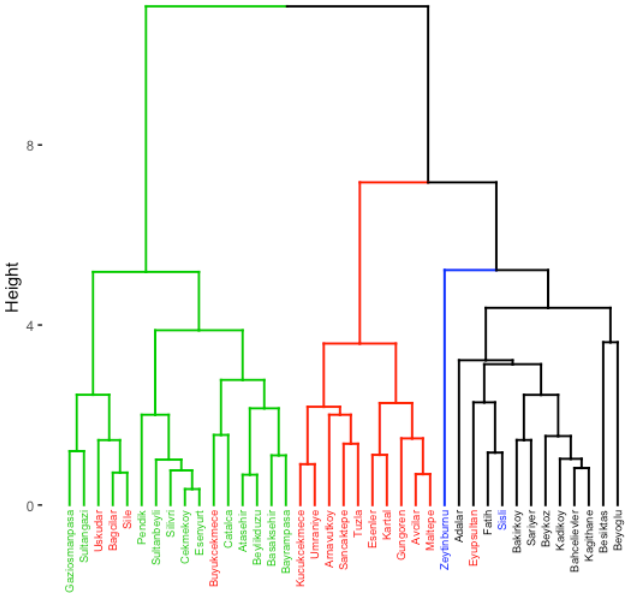

Figure 4. Dendogram by hierarchical clustering with k-means centers

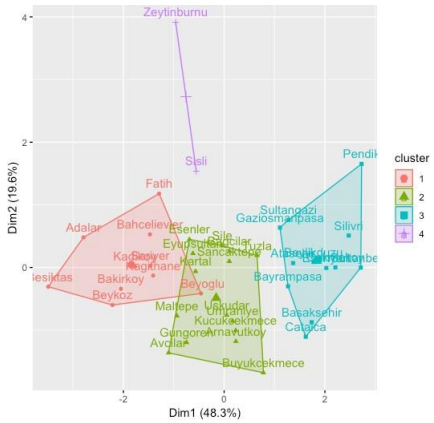

Figure 5. Clusters by hierarchical clustering with k-means centers-hhk

between methods can be easily visualized in the Figure 4. The labels are colored using k-means clusters. Corresponding clusters regarding to the hybrid clusterng approach are represented in Figure 5.

It also can be seen from the Table 3 that, 4 of the observations defined as belonging to cluster 2 by hierarchical clustering has been reclassified to cluster 2; 2 of the observations defined as belonging to clusters 1 have been reclassified to cluster 2 (*Eyupsultan*) and cluster 4 (*Sisli*) in the final solution defined by k-means clustering.

Table 3. Compare the results of hierarchical clustering and hybrid approach

	1	**2**	**3**	**4**
1	10	0	0	0
2	1	10	4	0
3	0	0	12	0
4	1	0	0	1

Here, k-means provides a positive contribution to the selection of the initial cluster means. It can be concluded that hierarchical clustering is improved by k-means algorithm.

5. Conclusion

In this paper, a hybrid clustering method is proposed for real estate data. This method searches for potential initial centroids for k-means clustering and integrates these into hierarchical clustering algorithm.

The data used in the analysis provides a viewpoint for applying a hybrid clustering algorithm. The results provide that hybrid method over performed standard k-means clustering. Apart from that, using the different distance measures is an essential factor for clustering to reveal useful information. Therefore, combining hierarcihal clustering with other methods can be applied for future studies.

References

[1] Case, B., Goetzmann, W.N. and Rouwenshorts, K.G. (2000). Global Real Estate Markets-Cycles and Fundamentals, *NBER Working Paper*, No. 7566.

[2] Chen, B., Tai, P.C., Harrison, R. and Pan, Y. (2005). Novel hybrid hierarchical-k-means clustering method (h-k-means) for microarray analysis. Fourth International IEEE Computer Society Computational Systems Bioinformatics Conference Workshops & Poster Abstracts (CSB 2005 Workshops), Stanford, CA, USA.

[3] Everitt, B.S., Hothorn, T. (2011). An introduction to applied multivariate analysis with R, Springer, NY, USA.

[4] Everitt,B.S, Landau, S., Leese, M., Stahl, D. (2011). Cluster analysis, 5th Edition, Wiley.

[5] Goodhart, C., Hofmann, B. (2008). House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, 24(1), 2008, 180-205.

[6] Hand, D., Mannila, H. & Smyth, P. (2001). Principles of data mining. Cambridge: MIT Press.

[7] Hartigan, J.A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C. Appl.Statist.*, 28,100–108.

[8] Johnson, R.A. and Wichern, D.W. (2007). Applied multivariate statistical analysis, 6th Edition, Pearson Prentice Hall, New Jersey, 430–439 .

[9] Maier, G., Herath, S. (2009) Real Estate Market Efficiency. A Survey of Literature. *SRE - Discussion Papers*, 2009/07.

[10] Manly, B.F.J. (2005), Multivariate statistical methods: A primer, Third Edition, Chapman and Hall.

[11] Mikelbank, B.A. (2004). Spatial Analysis of the Relationship Between Housing Values and Investments in Transportation Infrastructure. *The Annals of Regional Science* 38(4), 705-726.

[12] Rencher, A.C. (2002). Methods of multivariate analysis, Second Edition. Wiley.

[13] Song, Y. and Knaap, G.J. (2004). Measuring the effects of mixed land uses on housing values. *Regional Science and Urban Economics*, 34(6), 663-680, November.

Part X

List of Participants

Contributed/Invited/Poster Speakers		
Name	**Surname**	**Presentation Title**
Abdelhamid Hamidi	ALAOUI	Granger-Causality-Based Portfolio Selection in The Moroccan Stock Market
Adem	KAYAR	Fast Fault Finding Methods in Smart Manufacturing Lines with Augmented Reality Applications
Alessandro	BARBIERO	Approximation of Continuous Random Variables for The Evaluation of The Reliability Parameter of Complex Stress-Strength Models
Alessia	PINI	Functional Linear Model for Monitoring and Prediction of Profiles
Ali Mertcan	KOSE	Serial Mediation Model of Leader Member Interaction in Work Values and Job Satisfaction
Antonio	ELIAS	Depth-based Functional Time Series Forecasting
Aslihan	SENTURK ACAR	Prediction of Claim Probability in the Presence of Excess Zeros
Atefeh	MORADI	An Approach for Considering Claim Amount and Varying Deductibles in Designing Bonus-Malus Systems
Atike Reza	AHRABI	Chaos Control in Chaotic Dynamical Systems Via Auto-tuning Hamilton Energy Feedback
Ayfer Ezgi	YILMAZ	The Effect of Weights on Multi-rater Weighted Kappa Coefficients
Beste Hamiye	BEYAZTAS	A Robust Method for Estimation of Models with Random Effects
Bilge	BASER	Analyzing the Competition of HIV-1 Phenotypes with a Quantum Computation Perspective
Cagla	ODABASI	Fault Detection and Diagnosis Methodology in Refineries: A Data-Driven Approach
Carmela	IORIO	Portfolio Composition Strategy through a P-Spline based Clustering Approach
Caterina	LIBERATI	Predicting Business Survival from Their Websites
Damla	ILTER	Feature Selection Approaches for Machine Learning Classifiers on Yearly Credit Scoring Data
Derya	ALKIN	Comparison of Internal Validity Indices According to Distance Measurements in Clustering Analysis
Duygu	DEMIRAY	Statistical Inference of Consecutive k-out-of-n System in Stress-Strength Setup Based on Two Parameter Proportional Hazard Rate Family
Elif	COKER	Analysis of the Science Scores of Turkish Students in PISA 2015 via Multilevel Models
Enis	GUMUSTAS	Churn Analysis for Factoring: An Application in Turkish Factoring Sector
Erhan	PISIRIR	Two Structural Equation Modelling Approaches for Cloud Use in Software Development
Erkan	SIRIN	Outlier Detection on Big Data
Erkut	TEKELI	Evaluating New Optimization Methods for Two Parameter Ridge Estimator via Genetic Algorithm
Esra	AKCA	Recycle Project With RFM Analysis
Esra N.	KILCI	Do Confidence Indicators Have Impact on Macro-financial Indicators? Analysis on the Financial Services and Real Sector Confidence Indexes: Evidence from Turkey
Ezgi	OZER	Time-Frequency Analysis of EEG Signals: Visual Identification of Epileptic Patterns
Fatima	HARIS	Analysis of Data Comparing the Use of Different Social Media for Scientific Research Across Different Countries of The World
Fatma Zehra	DOGRU	Multivariate Skew Laplace Normal Distribution: Properties and

		Applications
Giulia	CONTU	Network-based Semisupervised Clustering
Gokce Nur	TASAGIL	Wavelet Regression for Noisy Data
Gozde	NAVRUZ	A Percentile Bootstrap Based Method on Dependent Data: Harrell Davis Quantile Estimator vs NO Quantile Estimator
Gul	INAN	Joint Modeling the Frequency and Duration of Physical Activity from a Lifestyle Intervention Trial
Gulcin	YANGIN	An application of XGBOOST on Diabetes Dataset
Gulce	CURAN	Stress-strength reliability estimation of series system with cold standby redundancy at system and component levels
Halil Ibrahim	CELENLI	Identification of Vehicle Warranty Data and Anomaly Detection by Means of Machine Learning Methods
Hasan Aykut	KARABOGA	A New Approach to Econometric Modelling of Monthly Total Air Passengers: A Case Study for Atatürk Airport
Hayafumi	WATANABE	Statistical Properties and Modeling of Stable-Like Word Count Time Series in Nation-Wide Language Data
Ilgim	YAMAN	Nonlinear Neural Network for Cardinality Constraint Portfolio Optimization Problem: Sector-wise analysis of ISE-all Shares
Ilkay	TUG	The Examination of Real Estate Prices in Istanbul by Using Hybrid Hierarchical K-Means Clustering
Ipek Deveci	KOCAKOC	A Customer Segmentation Model Proposal for Hospitals: LRFM-V
José Luis	TORRECILLA	From multivariate to Functional Classification
Leyla	BAKACAK KARABENLI	Conditional Autoregressive Model Approach to Generalized Linear Spatial Models by CARBayes
Luca	FRIGAU	Classification-based Approach for Validating Image Segmentation Algorithms
M.	MAHDIZADEH	Fitting Lognormal Distribution to Actuarial Data
M. Revan	OZKALE	Stochastic Linear Restrictions in Generalized Linear Models
Marwa	BEN GHOUL	Detection and Handling Outliers in Longitudinal Data: Can Wavelets Decomposition Be a Solution?
Metin	YANGIN	Hiv-1 Protease Cleavage Site Prediction with Generating Dataset Using A New Encoding Scheme Based on Physicochemical Properties
Murat	OZTURKMEN	Opportunities in Location Based Customer Analytics
Murat	GENC	The GO estimator: A New Generalization of Lasso
Natasa A.	CIROVIC	A StarCraft 2 Player Skill Modeling
Nebahat	BOZKUS	Hierarchically Built Trees with Probability of Placing Clusters
Nihan	ACAR-DENIZLI	A Functional Data Framework to Analyse the Effect of Quinoa Consumption on Blood Glucose Levels
Nihan	YUCEL	Alternative Subway Project Selection with TOPSIS Method Using Different Weighting Techniques
Nilufer	VURAL	Statistical and Fuzzy Modeling of Extraction Process in Green Chemistry
Nurdan	COLAKOGLU	Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases
Nursel	KOYUNCU	Gamma and Inverse Gaussian Distributions in Fitting Parametric Shared Frailty Models with Missing Data
Ocan	SAHIN	Big Data Solutions in Refineries with Heat Exchangers
Ozge	CAGCAG YOLCU	Bivariate Intuitionistic Fuzzy Time Series Prediction Model
Ozlem	TURKSEN	A Seemingly Unrelated Regression Modeling for Extraction Process in Green Chemistry

Ozlem Kiren	GURLER	Finding the Determinants of National Problem Perceptions of Turkish Citizens
Pervin	BAYLAN	Bivariate Credibility Premiums Distinguishing Between Two Claims Types in Third Party Liability Insurance
Rahim	MAHMOUDVAND	Highlighting a Mathematical Property of Sample ACF for Time Series Analysis
Semra	ERPOLAT TASABAT	Inferences About Development Levels of Countries with Data Envelopment Analysis
Serra	CELIK	How Does Resampling Affect the Classification Performance of Support Vector Machines on Imbalanced Churn Data
Sumeyye	INAL	Investigation of the Electricity Consumption of Provinces of Turkey using Functional Principal Components Analysis
Tugay	KARADAG	Probabilistic Structural Equation Modeling Approach to Investigate the Relationships Between Passenger Perceived Value, Image, Trust, Satisfaction and Loyalty
Ufuk	BEYAZTAS	On Function-on-Function Regression: Partial Least Squares Approach
Yunus Emre	GUNDOGMUS	Risk-based Fraud Analysis for Bank Loans With Autonomous Machine Learning
Zeynep	BAL	The Effect of Woe Transformation on Credit Scoring by Using Logistic Regression

Keynote Speakers	
Name	**Surname**
Ayse	OZMEN
Aytul	ERCIL
Bahar	KINAY ERGUNEY
Baris	SURUCU
Erkal	BIYIKLIOGLU
Gerhard-Wilhelm	WEBER
Hamparsum	BOZDOGAN
Selim	DELILOGLU
Sotiris	BERSIMIS
Umut	SATIR GURBUZ

Short-Course Speakers		
Name	**Surname**	**Title**
Arzu	BAYGUL	Medical Analytics/Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with Potential Discussion of Cancer Molecular)
Aytac	ATAC	Innovation in Germany (with Potential Emphasis on Internet of Things, Supply Chain Analytics)
Balaji	RAMAN	Retail Analytics with Dynamic Linear Models using R (Introduction to DLM and Kalman Filter, Setting up DLM in R using Packages Astsa, dlm and INLA, Real-Life Applications)
Cagdas	AKTAN	Medical Analytics/Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with Potential Discussion of Cancer Molecular)
Erkan	SIRIN	Big Data: Introduction to Hadoop Big Data Ecosystem, Introduction to Apache Spark, Data Analysis and Machine Learning with Apache Spark
Fulya	GOKALP	Hands-on Introduction Course in R(Acquiring Data from Different Sources on Command Line and R, Data pre-Processing, A Map Package to Map One of The up-to-Date Data (Potentially with 2019 Turkish Local Election Data), SQL in R
Neslihan	GOKMEN	Medical Analytics/Bioinformatics (Clinical Research Statistics, Medical Informatics, Validation Studies with Potential Discussion of Cancer Molecular)
Ozan	GOZBASI	Real World Applications/Cases of Transportation Analytics-Optimization with A Potential Demo
Rahim	MAHMOUDVAND	Visualization with QlikView (How to Make Dashboards)
Tahir	EKIN	Fraud Analytics
Tuba	YILMAZ-GOZBASI	Real World Applications/Cases of Transportation Analytics-Optimization with A Potential Demo

Participant only	
Name	**Surname**
Alev	BAKIR
Ayca	CAKMAK PEHLIVANLI
Aydin	ERAR
Baris	ASIKGIL
Batuhan	DOGANAY
Begum	APAYDIN
Berk	KUCUKALTAN
Bengisu	KOYUNCU
Berkay	SENTURK
Birsen	EYGI ERDOGAN
Busenur	CELENK
Busra	SEVINC
Busenur	KIZILASLAN
Cagla	BOLAT
Caner Akin	KECICI
Deniz	INAN
Elif	GULEC
Elif Ozge	OZDAMAR
Erim	HISIM
Esra	AKCA
Esra	AKDENIZ
Eylem	DENIZ HOWE
Fatma	NOYAN
Ferda Esin	GULEL
Gulay	BASARIR
Izem	DISLITAS
Kubra	SENSU
Lisa	CROSATO
Mehmet	AKTUNA
Meral	YAY
Mustafa	YILDIRIM
Mujgan	TEZ
Nihal	ATA TUTKUN
Oguzhan	CIMEN
Omer	SAHIN
Ozge	KORUKCI
Ozlem	TAZEGUL
Ozge	TAZEGUL
Resit	CELIK
Selin	SARIDAS
Sibel	DINC
Suleyman	KIZILTOPRAK
Turan	TOKALAK
Turgut	OZALTINDIS
Zeynep	ATLI
Zoran	CIROVIC

Part XI

Sponsors and Supporting Institutions

Sponsors

Supporting Institutions