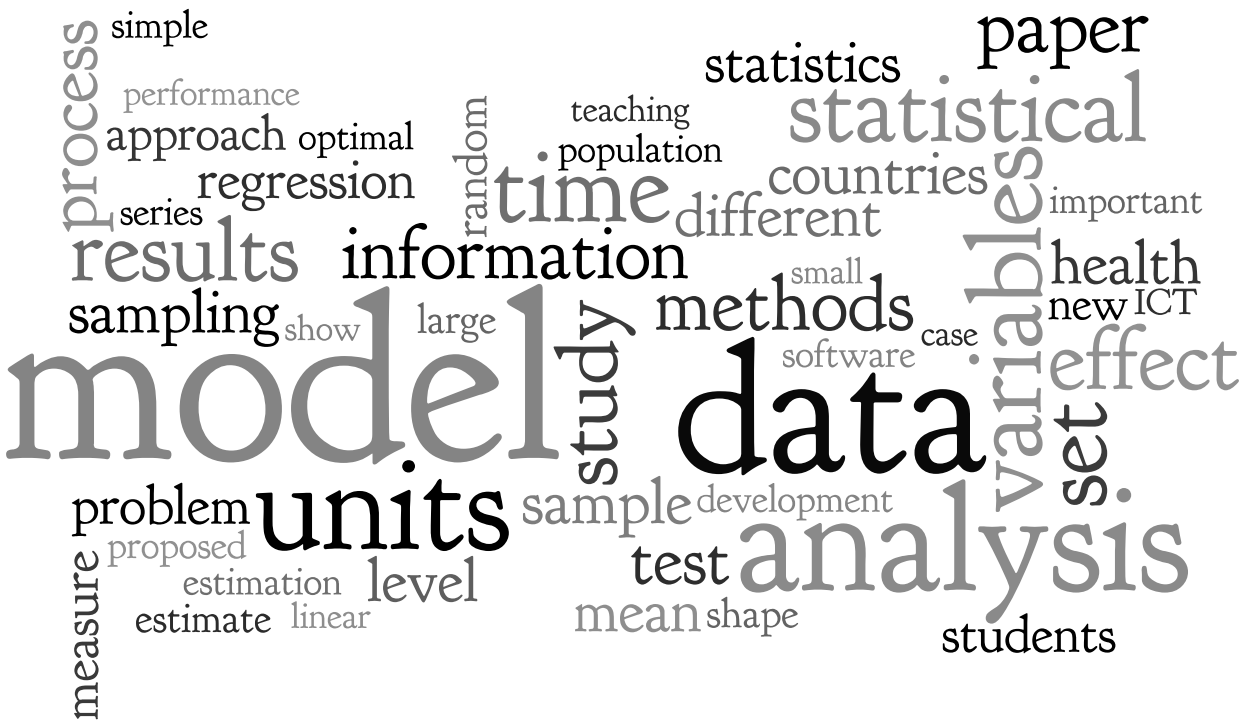
International Conference

APPLIED STATISTICS

2014

ABSTRACTS and PROGRAM



September 21 - 24, 2014
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2014

International Conference

APPLIED STATISTICS 2014

ABSTRACTS and PROGRAM

2014

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2014

Organized by

Statistical Society of Slovenia

Supported by

Generali SpA

Result

Alarix

NIB

IBMI

The word cloud on the cover was generated using www.wordle.net. The source text included the abstracts of the talks; the fifty most common words were displayed, and greater prominence was given to words that appeared more frequently.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

311(082)(0.034.2)

INTERNATIONAL Conference Applied Statistics (2014; Ribno)
Abstracts and program [Elektronski vir]/ International Conference Applied Statistics 2014, September 21 - 24, 2014, Ribno (Bled), Slovenia organized by Statistical Society of Slovenia ; [edited by Lara Lusa and Janez Stare]. - El. knjiga. - Ljubljana : Statistical Society of Slovenia, 2014

Način dostopa (URL): http://conferences.nib.si/AS2014/AS2014-Abstracts.pdf
ISBN 978-961-93547-2-8

1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
275377152

Scientific Program Committee

Janez Stare (Chair), Slovenia
Vladimir Batagelj, Slovenia
Maurizio Brizzi, Italy
Anuška Ferligoj, Slovenia
Dario Gregori, Italy
Dagmar Krebs, Germany
Lara Lusa, Slovenia
Mihael Perman, Slovenia
Tamas Rudas, Hungary
Albert Satorra, Spain
Hans Waege, Belgium
Tomaž Banovec, Slovenia
Jaak Billiet, Belgium
Brendan Bunting, Northern Ireland
Herwig Friedl, Austria
Katarina Košmelj, Slovenia
Irena Križman, Slovenia
Stanislaw Mejza, Poland
Jože Rovan, Slovenia
Willem E. Saris, The Netherlands
Vasja Vehovar, Slovenia

Organizing Committee

Andrej Blejec (Chair)
Lara Lusa
Bogdan Grmek
Irena Vipavc Brvar

Published by:	Statistical Society of Slovenia Litostrojska cesta 54 1000 Ljubljana, Slovenia
Edited by:	Lara Lusa and Janez Stare
Printed by:	Statistical Office of the Republic of Slovenia, Ljubljana
Produced using:	generbook R package
Circulation:	Available on the Internet

ABSTRACTS and PROGRAM

PROGRAM

		Hall 1	Hall 2
Sunday	14.00 – 15.00	Registration	
	15.00 – 18.00	Workshop	
	18.00 – 19.00	Registration	
	19.00	Reception	
Monday	8.00 – 9.00	Registration	
	9.00 – 9.10	Opening of the Conference	
	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Network Analysis	
	11.40 – 12.00	Break	
	12.00 – 13.20	Social Science Methodology	Statistical Applications
	13.20 – 15.00	Lunch	
	15.00 – 16.20	Mathematical Statistics	Statistical Applications
	16.20 – 16.40	Break	
	16.40 – 17.40	Education and Statistical Software	Statistical Applications
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Biostatistics and Bioinformatics	
	11.40 – 12.00	Break	
	12.00 – 13.20	Biostatistics and Bioinformatics	Bibliometry
	13.20 – 14.30	Lunch	
	14.30	Excursion	
Wednesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.20	Econometrics	Statistical Applications
	11.20 – 12.00	Break	
	12.00 – 13.20	Statistical Applications	Modeling and Simulation
	13.20 – 13.30	Closing of the Conference	

14.00–15.00 **Registration**

15.00–18.00 **Workshop** (Hall 1)

1. **Development and deployment of statistical web applications using R and Shiny**
   Lara Lusa and Črt Ahlin

18.00–19.00 **Registration**

19.00 **Reception**

8.00–9.00 **Registration**

9.00–9.10 **Opening of the Conference** (Hall 1) *Chair: Andrej Blejec*

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Anuška Ferligoj*

1. **Statistical models for interaction dynamics**
   Carter T. Butts

10.00–10.20 **Break**

10.20–11.40 **Network Analysis** (Hall 1) *Chair: Carter T. Butts*

1. **Generalized blockmodeling of sparse networks**
   Aleš Žiberna
2. **Are actor non-response treatments able to reveal network structure and central actors in valued networks?**
   Anja Žnidaršič, Patrick Doreian and Anuška Ferligoj
3. **Supervised link prediction for literature-based discovery: preliminary results**
   Andrej Kastrin and Dimitar Hristovski
4. **Evolution of university friendship networks: examining the influence of group size**
   Valentina Sokolovska and Aleksandar Tomasevic

11.40–12.00 **Break**

12.00–13.20 **Social Science Methodology** (Hall 1) *Chair: Damjan Škulj*

1. **Factor analysis and structural equation modeling in oncology research**
   Veronica Distefano, Sandra De Iaco, Monica Palma and Alessandra Spennato
2. **Psychometric assessment of the Slovenian translation of the Measure of Process of Care (MPOC-20)**
   Gaj Vidmar, Gregor Sočan, Katja Groleger Sršen and Anton Zupan
3. **Robustness of multivariate methods to different ways of computing factors in exploratory factor analysis**
   Jerneja Šifrer, Anja Žnidaršič and Matevž Bren
4. **Combating climate change in industry**
   Žiga Kotnik, Maja Klun and Damjan Škulj

12.00–13.20 **Statistical Applications** (Hall 2) *Chair: Katarina Košmelj*

1. **Tree-based methods for (life) insurance: on the interplay between statistics and actuarial mathematics**
   Walter Olbricht

2. **Predictive modeling for private health insurance claim severity in Turkey**
   Aslıhan Şentürk Acar and Uğur Karabey

3. **Discrimination of lactic acid bacteria by infrared spectroscopy and linear discriminant analysis**
   Cecile Levasseur-Garcia, Adeline Blanc and Hélène Tormo

4. **Identification of historical polymers using Near-Infrared Spectroscopy**
   Vilma Šuštar, Jana Kolar, Lara Lusa and Dušan Koleša

13.20–15.00 **Lunch**

15.00–16.20 **Mathematical Statistics** (Hall 1) *Chair: Mihael Perman*

1. **Estimating parameters using Ranked Set Sampling**
   Marija Minić and Marko Obradović

2. **Tests of normality and their sensitivity against particular alternatives**
   Zoran Vidović, Bojana Milošević, Marko Obradović and Konstantin Ilijević

3. **On suitability of negative binomial marginals and geometric counting sequence in some applications of Combined INAR(p) model**
   Aleksandar S. Nastić, Miroslav M. Ristić and Predrag M. Popović

4. **Ridit and exponential type scores for estimating the kappa statistics**
   Ayfer Ezgi Yılmaz and Serpil Aktas

15.00–16.20 **Statistical Applications** (Hall 2) *Chair: Gaj Vidmar*

1. **A geostatistical approach for radon risk prediction**
   Claudia Cappello, Sandra De Iaco, Monica Palma and Daniela Pellegrino

2. **Soil radon analysis through geostatistical tools implemented in a GIS**
   Alessandra Spennato, Sandra De Iaco, Veronica Distefano, Monica Palma and Sabrina Maggio

3. **Geostatistical methods in time series analysis: an application on environmental data**
   Sabrina Maggio, Monica Palma and Daniela Pellegrino

4. **Spatial analysis of soil's properties: a case study of Camgazi region**
   Hakan Basbozkurt, Ayse Basbozkurt, Adnan Karaibrahimoglu and Taskın Oztas

16.20–16.40 **Break**

16.40–17.40 **Education and Statistical Software** (Hall 1) *Chair: Andrej Blejec*

1. **Dealing with missing values in an educational effectiveness survey**
   Boštjan Mohorič and Matevž Bren

2. **Model of usefulness of SPSS for students of economics and business: differences between undergraduate and postgraduate students**
   Urban Šebjan and Polona Tominc

3. **Dynamic graph generation and data analysis of complex data: a web-application based on R and shiny**
   Črt Ahlin, Daša Stupica, Franc Strle and Lara Lusa

16.40–17.40 **Statistical Applications** (Hall 2) *Chair: Nataša Kejžar*

1. **Validity assessment in the mixed methods research from the view of mixed methods experts**
   Joca Zurc

2. **Measurement of efficiency using Data Envelopment Analysis: an application in European Airline industry**
   Burak Keskin and Yucel Turker

3. **An application of Liu-Type logistic estimators to the population data of Sweden**
   Adnan Karaibrahimoglu and Yasin Asar

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Janez Stare*

1. **Information theory and statistics: an overview**
   Daniel Commenges

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics** (Hall 1) *Chair: Daniel Commenges*

1. **Augmenting backward elimination by a standardized change-in-estimate criterion for variable selection in statistical models**
   Georg Heinze, Max Plischke, Karen Leffondré and Daniela Dunkler

2. **The non-parametric approach to analysing time to event with missing at risk information**
   Maja Pohar Perme and Tomaž Štupnik

3. **Mortality model of a human age-structured population based on an interval type-2 fuzzy logic**
   Andrzej Szymański and Agnieszka Rossa

4. **Image segmentation using a spatially regularized mixture model: application to lesion segmentation in stroke**
   Brice Ozenne, Fabien Subtil, Leif Ostergaard and Delphine Maucort-Boulch

11.40–12.00 **Break**

12.00–13.20 **Biostatistics and Bioinformatics** (Hall 1) *Chair: Georg Heinze*

1. **The ROC curve and their indexes based on bootstrap sampling**
   Pattaraporn Duriyakornkul and Pedro Oliveira

2. **Identification of tissue at risk after ischemic stroke**
   Ceren Tozlu, Brice Ozenne, Leif Ostergaard and Delphine Maucort-Boulch

3. **Reconstructing reasons for unsatisfactory status of Western Capercaillie (Tetrao urogallus) in Croatia through time and space**
   Andreja Radović and Robert Spanić

4. **Particle-based filtering applied to medical time series**
   Mohamed M. Shakandli

12.00–13.00 **Bibliometry** (Hall 2) *Chair: Vlado Batagelj*

1. **On standardization of the Activity Index**
   Nataša Kejžar and Janez Stare

2. **Co-authorship structures of researchers in scientific disciplines in time**
   Marjan Cugmas, Anuška Ferligoj and Luka Kronegger

3. **Stability of co-authorship blockmodeling structure in time**
   Luka Kronegger, Anuška Ferligoj and Marjan Cugmas

4. **A comparison of methods for capturing the treatment effect of an observational study for programme evaluation**
   Thanawit Bunsit

13.20–14.30 **Lunch**

14.30 **Excursion**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Andrej Blejec*

1. **Bayesian uncertainty analysis for complex physical systems modelled by computer simulators**
   Michael Goldstein

10.00–10.20 **Break**

10.20–11.20 **Econometrics** (Hall 1) *Chair: Michael Goldstein*

1. **An estimate of the degree of interconnectedness between countries: a lasso approach**
   Davide Fiaschi and Angela Parenti
2. **Spatial clubs in European regions**
   Davide Fiaschi, Lisa Gianmoena and Angela Parenti
3. **Bivariate regression model for count data based on the generalised Poisson distribution**

   Vera Hofer and Johannes Leitner
4. **Local finance and the demand for property-casualty insurance**
   Giovanni Millo and Pietro Millossovich

10.20–11.20 **Statistical Applications** (Hall 2) *Chair: Lara Lusa*

1. **A multilevel analysis on the importance of work in several European Union's countries**
   Laura Asandului and Roxana Otilia Sonia Hritcu
2. **Multivariate stochastic volatility estimation using Particle filter**
   Jian Wang
3. **Risk measurement of the future annuity prices: effects of different interest rate models**
   Ugur Karabey and Sule Sahin

11.20–12.00 **Break**

12.00–13.20 **Statistical Applications** (Hall 1) *Chair: Delphine Maucort-Boulch*

1. **Sports predictions made by a statistical model: a Sochi case**
   Slavko Jerič
2. **Analysis of the impact of fatigue on the running technique**
   Melita Hajdinjak and Martin Krašek
3. **Communication and adoption dynamics in new product life cycle: the case of Apple iPod**
   Mariangela Guidolin
4. **Statistical prediction in the production of vulcanized rubber products**
   Melita Hajdinjak and Gregor Dolinar

12.00–13.00 **Modeling and Simulation** (Hall 2) *Chair: Maja Pohar Perme*

1. **Departure from uniform association in square contingency tables**
   Serpil Aktas and Ayfer Ezgi Yılmaz

2. **Estimating dynamic panel data models with random individual effect: Instrumental Variable and GMM approach**
   Johnson T Olajide, Olusanya E. Olubusoye and Iyabode F Oyenuga

3. **Power of tests of heteroscedasticity in non-linear model**
   Iyabode F Oyenuga and Benjamin A Oyejola

13.20–13.30 **Closing of the Conference** (Hall 1)

ABSTRACTS

Workshop

Development and deployment of statistical web applications using R and Shiny

Lara Lusa[1] *and Črt Ahlin*[2]

[1]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
[2]University of Ljubljana, Ljubljana, Slovenia
lara.lusa@mf.uni-lj.si, crt.ahlin@gmail.com

Have you ever dreamt about being able to develop a graphical user interface that would allow (non-statisticians) to easily visualize or explore data? Or that could interactively and intuitively display the results of a complex analysis that you performed? Or that would help your students to get a better understanding of some complex statistical concepts? If you have a basic knowledge of R language it is nowadays rather straightforward to make these wishes come true, using the Shiny web application framework for R (http://shiny.rstudio.com/).
The workshop will introduce the basics needed to build and deploy a web application based on code written in R. The rest of the workshop will be "hands-on". The participants will have the opportunity to explore the code of existing applications and to write their own web applications.
The participants should be familiar with R programming and bring their own laptops, with the latest version or R, R Studio and shiny package installed. A wireless connection to the Internet will be provided. (Help with installation will be provided prior to the beginning of the workshop, if needed)

Invited Lecture

Statistical models for interaction dynamics

Carter T. Butts

Departments of Sociology, Statistics, and EECS and Institute for Mathematical Behavioral Sciences, University of California , Irvine, U.S.A.

buttsc@uci.edu

Patterns of interaction among individuals, organizations, or other entities are often the result of a complex interplay of endogenous behavioral mechanisms and contextual factors. Given observational data on such interactions, how can we identify the specific mechanisms responsible for the dynamics? Likewise, how can we translate candidate mechanisms from behavioral theories into estimable (and ultimately testable) models for social interaction, particularly when many different mechanisms may be at work? In this talk, I will describe one approach to addressing these questions, and discuss how this approach has been used to model social interaction in settings ranging from emergency communication and classroom interaction to email and gang violence. I will also point to some ongoing challenges in this area, and opportunities for further applications and development.

Network Analysis

Generalized blockmodeling of sparse networks

Aleš Žiberna

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
ales.ziberna@gmail.com

The paper starts with an observation that the blockmodeling of relatively sparse binary networks (where we also expect sparse non-null blocks) is problematic. The use of regular equivalence often results in almost all units being classified in the same equivalence class, while using structural equivalence (binary version) only finds very small complete blocks. Two possible ways of blockmodeling such networks within a binary generalized blockmodeling approach are presented. It is also shown that sum of squares (homogeneity) generalized blockmodeling according to structural equivalence is appropriate for this task, although it suffers from "the null block problem". A solution to this problem is suggested that makes the approach even more suitable. All approaches are also applied to an empirical example. My general suggestion is to use either binary blockmodeling according to structural equivalence with different weights for inconsistencies or sum of squares (homogeneity) blockmodeling with null and constrained complete blocks. The second approach is more appropriate when we want complete blocks to have rows and columns of similar densities and differentiate among complete blocks based on densities. If these aspects are not important the first approach is more appropriate as it does in general produce 'cleaner' null blocks.

Are actor non-response treatments able to reveal network structure and central actors in valued networks?

Anja Žnidaršič[1], *Patrick Doreian*[2] *and Anuška Ferligoj*[3]

[1]Faculty of Organizational Sciences, University of Maribor , Maribor, Slovenia

[2]Department of Sociology, University of Pittsburgh and Faculty of Social Sciences, University of Ljubljana, Pittsburgh, United States of America

[3]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

anja.znidarsic@fov.uni-mb.si, pitpat@pitt.edu, anuska.ferligoj@fdv.uni-lj.si

Valued networks are more realistic snap-shot of real world dynamics and relations among entities than simplified binary networks where only information on the presence or absence of ties is recoded. Values on ties represent the strength of ties, for example an intensity of relationship among friends or acquaintances, a quantity of daily contacts or e-mails among employees, a number of common activities among students, etc. The collected network data are likely to be measured with errors regardless the employed measurement level. One source of errors takes the form of actor non-response where outgoing ties are absent for each non-respondent, however incoming ties are available. The actor non-response treatments were already investigated in the case of binary networks. Here, the study was extended to valued networks with six simple actor non-response treatments via simulations. The first treatment is the complete-case approach where beside the row of absent ties also the corresponding column is deleted and the result is a smaller network. A null tie imputation procedure replace all absent ties by zeroes. If the modal value of incoming ties for a non-respondent is used instead of absent tie the procedure is called imputations based on modal values. In the reconstruction procedure an absent outgoing tie from actor i to actor j is replaced by the incoming tie from actor j to actor i. Reconstruction of ties between two non-respondents is not possible, therefore in the simplest case the null tie imputations are used, while the second option is use of imputations based on modal values for ties between non-respondents. The sixth procedure is an imputation of a total mean where valued density of the network is imputed instead of absent ties. The impact of presented non-response treatments on several valued network measures and typical network structures is investigated.

Supervised link prediction for literature-based discovery: preliminary results

Andrej Kastrin[1] *and Dimitar Hristovski*[2]

[1]Faculty of Information Studies, Novo mesto, Slovenia

[2]Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

andrej.kastrin@guest.arnes.si, dimitar.hristovski@mf.uni-lj.si

The growth rate of the scientific literature makes it impossible for the researchers to keep in line with all the relevant information. Novel methods and tools are urgently needed to extract and explore new knowledge in the literature. Literature-based discovery (LBD) is a methodology for automatically generating hypotheses for scientific research by uncovering hidden, previously unknown relationships from existing knowledge. Co-occurrence associations between biomedical concepts are commonly used in LBD to represent domain knowledge. These co-occurrences can be represented as a network that consists of a set of nodes representing the concepts and a set of edges representing their relationships. In this work we propose and evaluate a methodology for supervised link prediction of implicit connections in a network of co-occurring Medical Subject Headings. Specifically, we employ supervised statistical learning to predict implicit links between nodes, using support vector machines, decision trees, k-nearest neighbors, and naïve Bayes. As learning features, we used common neighbors between nodes, Jaccard coefficient, AdamicAdar coefficient, and preferential attachment score. We compared the performance of the classifiers using the area under the ROC curve. The results show high prediction performance, with best score for support vector machines. We also discuss the class imbalance problem, which is inherent in link prediction tasks.

Evolution of university friendship networks: examining the influence of group size

Valentina Sokolovska and Aleksandar Tomasevic

Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia

valentina.sokolovska25@gmail.com, atomashevic@gmail.com

In this paper we describe and analyze the evolution of friendship relations in networks of sociology freshmen from Serbia's two largest universities: University of Belgrade and University of Novi Sad. Due to different admission regulations, these networks differ in size and as a result we observe and analyze initial structural differences in the early stages of network formation. Following previous research in this area and using SIENA software, we analyze the significance of network effects and personal preferences of students in different stages of network evolution. We compare the findings for two groups and examine the influence of group size on significance of those effects.

Social Science Methodology

Factor analysis and structural equation modeling in oncology research

Veronica Distefano, Sandra De Iaco, Monica Palma and Alessandra Spennato

Department of Management, Economics, Mathematics and Statistics, University of Salento, Lecce, Italy

veronica.distefano@unisalento.it, sandra.deiaco@unisalento.it, monica.palma@unisalento.it, alessandra.spennato@unisalento.it

The aim of this paper is to investigate the service quality provided to the patients and the relationship between doctors and long-term cancer patients. Data have been collected during a survey conducted to long-term cancer patients, who follow a therapy at the Hospital Vito Fazzi, in Province of Lecce (located in the Southern region of Puglia, Italy). In particular, factor analysis and structural equation model are used to measure the relations among latent variables related to two aspects of the analyzed issue, such as service quality provided to the patient and the relationship between doctors and long-term cancer patients. The first model describes the perceived service quality provided to the patient, which is influenced by four important factors such as the tangible aspects, the reliability, the empathy (doctor-patient human relations) and the hospital organization. The second model describes the relationship between doctors and long-term cancer patients, which is influenced by three factors, such as the reliability, the empathy (doctor-patient human relations) and the hospital organization. The results are useful to investigate the strategies used to improve the quality service. Moreover, the analysis focuses on highlighting some empirical evidences in health risk through the use of a Geographical Information System (GIS). The advantages of implementing a GIS are related to the possibility to include different demographic databases, relate and analyze them as well as to detect and represent the areas in which there are high mortality rates. This tool, called GIS Cancer Screening, allows to process thematic maps using health data and support public health policies.

Psychometric assessment of the Slovenian translation of the Measure of Process of Care (MPOC-20)

Gaj Vidmar[1], *Gregor Sočan*[2], *Katja Groleger Sršen*[1] *and Anton Zupan*[1]

[1]University Rehabiltation Institute, Ljubljana, Slovenia

[2]Department of Psychology, Faculty of Arts, University of Ljubljana , Ljubljana, Slovenia

gaj.vidmar@ir-rs.si, gregor.socan@ff.uni-lj.si, katja.groleger@ir-rs.si, anton.zupan@ir-rs.si

The Measure of Processes of Care is a questionnaire for parents used to evaluate their experience with health-care providers. It consists of five scales (Enabling and Partnership, Providing General Information, Providing Specific Information about the Child, Coordinated and Comprehensive Care, Respectful and Supportive Care). We applied the Slovenian translation of the 20-item version (MPOC-20) in a large study on rehabilitation of children with chronic illness or disability. Parents of 235 children who were admitted as inpatients or outpatients to six institutions (hospitals, hospital departments or health centres) participated in the study. A combination of quota and random sampling was employed. We assessed internal consistency reliability and one-year stability of the Slovenian MPOC-20 using basic classical-test-theory methods. Concurrent validity was assessed using simple correlation with a related instrument (Client Satisfaction Questionnaire, CSQ-8) and a rating scale of stress and worries reduction. The associations between parental evaluation of processes of care and child, parent and family characteristics were assessed univariately and using multiple linear regression for each scale as well as for a composite standardised score (in the spirit of O'Brien's test). A structural equation model, aimed at the explanation of five latent variables underlying the five MPOC-20 scales, was successfully fitted to the data with minor and substantiated modifications of the original model (one item was assigned to a different scale, residuals were allowed to correlate between five pairs of variables). The characteristics of the child, the family and the treatment explained up to 22% of the latent variables' variances. Overall, the Slovenian translation of the MPOC-20 was found to be a reliable and valid instrument, whereby the availability of a key person was identified as a factor that plays a central role in parental satisfaction.

Robustness of multivariate methods to different ways of computing factors in exploratory factor analysis

Jerneja Šifrer[1], *Anja Žnidaršič*[2] *and Matevž Bren*[1]

[1]Faculty of Criminal Justice and Security, University of Maribor, Ljubljana, Slovenia
[2]Faculty of Organizational Sciences, University of Maribor , Maribor, Slovenia
jerneja.sifrer@fvv.uni-mb.si, anja.znidarsic@fov.uni-mb.si, matevz.bren@fvv.uni-mb.si

The standard research practice in the social sciences is that for a phenomenon, that cannot be measured directly, different aspects of this phenomenon are measured, in other words, one construct (latent variable) is measured with a group of measurable variables. The question that arises is, are these different variables really driven by the same construct (underlying latent variable). To answer this question, exploratory factor analysis is applied to identify groups of variables that account for different dimensions of the construct. Usually the structure of groups of variables is predicted in a research plan and this structure determines the dimensions i.e. factors of the underlying construct. Therefore it is important that researcher understands the structure of a set of measured variables and that he/she can interpret the factors themselves. Software packages for statistical analysis (SPSS, SAS ...) offer different methods for computing factors (regression method, Bartlett method etc.) that give normalized resulting factors. Nevertheless for researchers factors defined (measured) on the same scale as the measured variables are much more useful, in many cases they already facilitate the explanation. Therefore the research practice is to compute factors as an average of the corresponding set of variables or as a weighted average with factor loadings serving as weights. We will (1) present examples of this social sciences research practice, (2) with simulated data we will compare factors computed with different methods and in addition (3) we will perform regression analysis to test its robustness according to factors computed in different ways. We expect high correlation among factors computed via different approaches, and we expect similar results of regression analysis regardless of the applied approach.

Combating climate change in industry

Žiga Kotnik[1], *Maja Klun*[1] *and Damjan Škulj*[2]

[1]Faculty of Administration, University of Ljubljana, Ljubljana, Slovenia
[2]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
ziga.kotnik@fu.uni-lj.si, maja.klun@fu.uni-lj.si,
damjan.skulj@fdv.uni-lj.si

This paper examines the effect of environmental energy taxes and environmental expenditures in industry on greenhouse gas (GHG) emissions in industrial processes. A panel data set of 16 EU countries for the time period 1995-2010 was used. We investigate whether GHG emissions have an effect on environmental energy taxes after a certain time lag. The paper applies multiple regression analysis, fixed and random effects model to take into account the dynamic nature and to properly address the potential endogeneity. The findings reveal that the effect of environmental taxes on GHG emissions is negative. The effect of environmental expenditures in industry on GHG emissions is also negative and even more statistically significant and 2.2 times stronger as the effect of environmental taxes alone. The analysis shows that GHG emissions (in sector energy and sector industrial processes) have an effect on environmental taxes. Higher level of GHG emissions means higher energy taxes after a certain period of time. Consequently, some policy implications may be derived that may help to find the best balance between the level of environmental taxes and expenditures on environmental protection to reduce GHG emissions.
This work has been fully supported by the Croatian Science Foundation under the project number IP-2013-11-8174.

Statistical Applications

Tree-based methods for (life) insurance: on the interplay between statistics and actuarial mathematics

Walter Olbricht

Institute of Mathematics, University of Bayreuth, Bayreuth, Germany
walter.olbricht@uni-bayreuth.de

Actuarial mathematics and statistics share a lot of common ground. However, there are also obvious differences. Data sets in the field of insurance tend to be very large so that sampling aspects and random errors are not of prime concern. On the other hand they are typically heterogeneous so that substructures matter. Furthermore, frequently contextual knowledge (such as shifts in legislation which prompted effects in the data) is available which could - and should - be incorporated. The talk analyses this background and suggests tree-based methods as an interesting statistical tool even for the classical field of life insurance. In particular a "hybrid" approach (using regression trees for a classification situation) is proposed. The main advantage of this approach is its ease of interpretability and its inherent transparency.

Predictive modeling for private health insurance claim severity in Turkey

Aslıhan Şentürk Acar and Uğur Karabey

Hacettepe University, Department of Actuarial Sciences, Ankara, Turkey
aslihans@hacettepe.edu.tr, ukarabey@hacettepe.edu.tr

Private health insurance or medical expense insurance covers the unexpected expenses that incurred through the sickness of the insured. Actuaries try to predict future costs using demographic characteristics of the insured, diagnosis and prior claim information. The aim of this study is to model the private health insurance claim severities with statistical models such as log-transformed linear models, generalized linear models, two-parts models and compare the results.

Discrimination of lactic acid bacteria by infrared spectroscopy and linear discriminant analysis

Cecile Levasseur-Garcia[1], *Adeline Blanc*[1] *and Hélène Tormo*[2]

[1]Université de Toulouse, INP-Ecole d'Ingénieurs de Purpan, LCA (Laboratoire de Chimie Agro-Industrielle)/INRA, UMR 1010 CAI, Toulouse, France

[2]Université de Toulouse, INP-Ecole d'Ingénieurs de Purpan, Département Sciences Agronomiques et Agroalimentaires, Toulouse, France

cecile.levasseur@purpan.fr, adeline.blanc@purpan.fr, helene.tormo@purpan.fr

In goat milk, the most important group of microorganisms are lactic acid bacteria. They are involved in making cheese through fermentation. Their action gives the texture and the taste of the final product. The bacteria of interest in our study are genera Enterococcus and Lactobacillus. The aim of our project is to use a rapid tool to discriminate between them: near infrared spectroscopy. This technology deals with the infrared region of the electromagnetic spectrum and exploits the fact that molecules absorb specific frequencies that are characteristic of their structure. It is widely used in characterization of the structure of molecules, and in quantification of parameters, such as proteins, starch ... One hundred and five bacteria were used in this study: 48 Enterococcus and 57 Lactobacillus. They were cultured in petri dishes, on nutritive elements. Infrared spectra were recorded between 908 and 1684 nm on a GETSPEC 1.7 NIR 356 spectrometer and a reflecting probe. A mathematical pre-treatment was applied on raw spectra: MSC (Multiplicative Signal Correction) in order to minimize noise. We used a quadratic discriminant analysis as supervised classification method (The Unscrambler® X v10.2. Ninety three percent of the bacteria were well classified.

Identification of historical polymers using Near-Infrared Spectroscopy

Vilma Šuštar[1], *Jana Kolar*[2], *Lara Lusa*[3] *and Dušan Koleša*[2]

[1]Faculty of Agriculture and Life Sciences, University of Maribor, Maribor, Slovenia

[2]Karakta d.o.o., Ivančna Gorica, Slovenia

[3]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia

vilma.sustar@um.si, jana.kolar@karakta.eu, lara.lusa@mf.uni-lj.si, dusan.kolesa@karakta.eu

Near-infrared spectroscopy (NIRS) is a spectroscopic method that uses the near-infrared region of the electromagnetic spectrum (from about 800 nm to 2500 nm). NIRS has become a popular method in the analysis of cultural heritage due to the portability of instrumentation, the non-destructive spectral acquisition and the wealth of information that can be extracted from the spectra. Data obtained using NIRS are high dimensional (contain hundreds of variables) and autocorrelated (the amount of light absorbed is similar for similar wave-lengths). In this presentation we will focus on the use of NIRS data for classification purposes, where the aim is to develop a rule that can be used to accurately determine the class membership of new samples. More specifically, we will focus on multiclass problems, where the number of classes is large and on situations where the number of samples in each class is not equal (class-imbalanced problems). Although the problem of imbalance has been well established for the two class case, there is a lack of the research on the topic of multiclass imbalanced data classification. Moreover, few studies used large number of classes. In this presentation, we will present a rule to classify 41 classes of 535 historical and modern polymers using NIRS in the presence of strong class-imbalance. The rule was developed by inspecting 112 models that were obtained by combining 4 spectra pre-processing methods, 4 dimension reduction methods and 7 classification methods.

Mathematical Statistics

Estimating parameters using Ranked Set Sampling

Marija Minić[1] *and Marko Obradović*[2]

[1]Faculty of Agriculture, University of Belgrade , Belgrade, Serbia
[2]Faculty of Mathematics, University of Belgrade , Belgrade, Serbia
minic.m.marija@gmail.com, marcone@matf.bg.ac.rs

Ranked Set Sampling (RSS) is a statistical technique for data collection which has had an increasing popularity during the previous two decades. This method uses quantitative or qualitative information that is cheap to get to obtain a more representative sample before the real, more expensive sampling is done. In this paper we estimate some population parameters on the sample obtained by RSS method. In addition, obtained RSS estimators are compared with common estimators and differences are discussed in detail. Finally, one example of RSS method application on real data is given.

Tests of normality and their sensitivity against particular alternatives

Zoran Vidović[1]*, Bojana Milošević*[2]*, Marko Obradović*[2] *and Konstantin Ilijević*[3]

[1]Teacher's Training Faculty, University of Belgrade, Belgrade, Serbia
[2]Faculty of Mathematics, University of Belgrade , Belgrade, Serbia
[3]Faculty of Chemistry, University of Belgrade, Belgrade, Serbia
zoranvidovic1990@gmail.com, bojana@matf.bg.ac.rs,
marcone@matf.bg.ac.rs, kilijevic@chem.bg.ac.rs

Normal distribution is one of the most important distributions and it is assumed in many statistical tests and procedures. In this paper we compare different normality tests against various types of alternatives distributions. Special attention is given to the cases where tests lead to opposite conclusions.

On suitability of negative binomial marginals and geometric counting sequence in some applications of Combined INAR(p) model

Aleksandar S. Nastić[1], *Miroslav M. Ristić*[1] *and Predrag M. Popović*[2]

[1]Department of Mathematics, Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
[2]Department of Mathematics, Faculty of Civil Engineering and Architecture, University of Niš, Niš, Serbia
anastic78@gmail.com, miristic72@gmail.com, popovicpredrag@yahoo.com

A combined negative binomial integer-valued autoregressive process of order p is defined. Correlation structure and regression properties are presented. Model parameters are estimated using conditional least squares and Yule-Walker methods and the asymptotic distributions of the obtained estimators are derived. Model interpretation is provided, especially focusing on usage of geometric counting sequence and negative binomial marginals and further it is justified by application of the introduced model to certain counting data, where it is compared with some other possible known model solutions.

Ridit and exponential type scores for estimating the kappa statistics

Ayfer Ezgi Yılmaz and Serpil Aktas

Hacettepe University, Ankara, Turkey
ezgiyilmaz@hacettepe.edu.tr, serpilaltunay@gmail.com

Cohen's kappa coefficient is a commonly used method for estimating interrater agreement for nominal and/or ordinal data, thus the agreement is adjusted for that expected by chance. The weighted kappa statistic is used as an agreement index for ordinal data. The weights quantify the degree of discrepancy between the two categories. The choice of this particular set of weights affects the value of Kappa. The common scores are Cicchetti-Allison and Fleiss-Cohen weights. In this article, we discuss the use of the ridit type and exponential scores to compute Kappa statistics in general.

Statistical Applications

A geostatistical approach for radon risk prediction

Claudia Cappello, Sandra De Iaco, Monica Palma and Daniela Pellegrino

Department of Management, Economics, Mathematics and Statistics. University of Salento, Lecce, Italy

claudia.cappello@unisalento.it, sandra.deiaco@unisalento.it, monica.palma@unisalento.it, daniela.pellegrino@unisalento.it

In many environmental sciences, several correlated variables are observed at some locations of the domain of interest, then appropriate modeling and prediction techniques for multivariate spatial data are necessary. This paper aims to highlight the convenience of using multivariate geostatistical methods to study the spatial distribution of radon soil concentration, to map and assess high risk areas. Indeed, this soil gas, due its nature, is known to be carcinogen: many studies have demonstrated that risk of lung cancer increases substantially with the exposure to higher radon concentrations. In analyzing radon concentrations, it is relevant to consider the available data regarding the geology, geomorphology and soil type since this gas is released during the decay of some radioactive elements found in rocks and soil. Thus, the application of multivariate geostatistical techniques, such as indicator-cokriging and indicator kriging for conditional probability analysis, is convenient to classify areas according to radon levels and to provide probability maps of radon risk. Note that in this paper geostatistical bootstrap for quantifying spatial uncertainty has been performed. A case study on sample data concerning radon-222 concentrations, permeability, lithology, fault and polje in Lecce district (Southern Italy) is proposed. Radon risk prediction maps for the probability to exceed certain threshold values, conditioned to specific soil type, can be useful especially for regions with no or only few measurements of soil gas radon.

Soil radon analysis through geostatistical tools implemented in a GIS

Alessandra Spennato, Sandra De Iaco , Veronica Distefano, Monica Palma and Sabrina Maggio

Department of Management, Economics, Mathematics and Statistics, University of Salento, Lecce, Italy

alessandra.spennato@unisalento.it, sandra.deiaco@unisalento.it, veronica.distefano@unisalento.it, monica.palma@unisalento.it, sabrina.maggio@unisalento.it

Radon (Rn) gas has been ranked by the World Health Organization (WHO) as one of most the dangerous natural elements, and it is known to be the second leading cause of lung cancer after smoking. Rn gas concentration depends on many parameters and generally it is difficult to quantify its spatial variation and define the potential of Rn over the study area. Geostatistical techniques provide several tools to study the spatial structure of Rn concentration,while the simultaneous use of thematic maps based on a Geographical Information System (GIS) allows researchers to track the levels of Rn gas concentration. In this paper, after introducing the usefulness of a GIS supported by geostatistical results, a powerful tool for the analysis of the spatial distribution of Rn concentration is presented. Then informative maps, where different layers related to data measured in situ, soil properties, land use and other territorial features can be overlaid, are created. A comparison among the results obtained by using different techniques implemented in an extension module of ArcGIS, i.e. Ordinary Kriging (OK), Lognormal Kriging (LG) and Inverse Distance Weighting (IDW) is provided. Moreover, using Voronoi maps allows the detection of clusters related to different levels of Rn concentration and the geostatistical mapping based on variogram model is useful to identify areas with high risk of Rn pollution over the domain under study. A case study on a data set regarding the soil Rn concentrations is presented.

Geostatistical methods in time series analysis: an application on environmental data

Sabrina Maggio, Monica Palma and Daniela Pellegrino

Department of Management, Economics, Mathematics and Statistics. University of Salento, Lecce, Italy

sabrina.maggio@unisalento.it, monica.palma@unisalento.it, daniela.pellegrino@unisalento.it

Geostatistical techniques are usually applied to analyze spatial relationships among sample data and to prediction spatial phenomena, but the use of these methods is not widespread in the analysis of variables that present temporal evolution. The paper aims to highlight the powerful of geostatistical tools in time series analysis. In this context, the variogram could represents a very useful exploratory tool for assessing stationarity in time series. Moreover, it allows to identify trends and periodicity exhibited by the data. The use of variogram is also convenient to obtain kriging predictions of the variable under study, either for temporal intervals with missing values and in time points after the last available date. A case study concerning temporal evolution of environmental data has been discussed in order to underline the role of geostatistical techniques, such as kriging, in modeling, prediction and reconstruction of time series. In particular, time series of PM10 daily concentrations measured at a monitoring station located in an area of Southern Italy, for the period 2010-2013, have been assessed. Note that this area is characterized by high levels of PM10. After the identification of trend and periodicity, reconstruction of the analyzed time series by estimation of missing values, has been performed. Finally, predictions of PM10 daily concentrations at some unsampled points have been obtained. For interpolation and prediction purposes, a modified version of GSLib kriging routine has been used.

Spatial analysis of soil's properties: a case study of Camgazi region

Hakan Basbozkurt[1], *Ayse Basbozkurt*[2], *Adnan Karaibrahimoglu*[3] *and Taskın Oztas*[4]

[1]Department of Statistics, Selçuk University, Konya, Turkey

[2]Department of Geography, Bingol University, Bingol, Turkey

[3]Medical Education and Informatics Department, Necmettin Erbakan University, Konya, Turkey

[4]Department of Soil, Atatürk University, Erzurum, Turkey

hakan.basbozkurt@gmail.com, ayse.basbozkurt@gmail.com, akara@konya.edu.tr, toztas@atauni.edu.tr

Spatial analysis is dealing with the space based data which is becoming one of the most popular branches of statistics. Tobler's (1970) "all the places are related to, but close ones are more closely related with each other" rule plays a key role when the subject of research is about physical and social issues. Classical regression models are usually insufficient to analyse spatial data where spatial regression models are needed to use to describe the statistical significance of the statistical changes on the spatial data.

In this study, spatial interaction of the soil properties on soil samples which were obtained from the Southeastern Anatolian Region in the range of 0-30 cm will be presented with the help of Geographic Information Systems (GIS) and spatial regression methods. Mainly GeoDa, GS+, SPSS and ArcGIS software programs are used to analyse the spatial data. Three methods; Moran's I index, variogram and variation coefficiency are used to describe the spatial dependency of the soil's properties. According to the spatial dependency, spatial error and spatial lag regression methods are compared to classical regression method to assess the relationship between soil's properties. It is aimed to test and compare each three methods to analyse spatial dependency and regression model of the soil's properties.

Education and Statistical Software

Dealing with missing values in an educational effectiveness survey

Boštjan Mohorič[1] *and Matevž Bren*[2]

[1]Elementary school Davorin Jenko, Cerklje na Gorenjskem, Slovenia

[2]University of Maribor and Institute of Mathematics, Physics and Mechanic, Ljubljana, Slovenia

bostjan.mohoric1@guest.arnes.si, matevz.bren@fvv.uni-mb.si

Throughout the last years a lot of effort has been invested in the quality of education in Slovenia, consequently education-quality projects address the entire vertical of education from kindergarten to university. Two projects 'Design and implementation of a system of quality assurance of educational institutions' (KVIZ) and 'Quality for the Future of Education' (KzP) of quality assurance in elementary schools are primarily developing, both based on self-evaluation and ISO 9001.
In our experimental study we intend to examine whether primary schools involved in the education-quality projects perform better than the schools not involved in these projects. Therefore we joined 50 schools involved in KVIZ or KzP into the experimental group and in the control group adequate pairs of schools not involved in KVIZ or KzP. We gathered and compared the data on achievements of pupils at national assessment, sports achievements, bullying in schools, and teacher satisfaction for all of these schools. In this presentation we will give focus on missing values and uncompleted survey results.
The questionnaire was divided in three parts. In the first part we were asking class teachers from 6th to 9th grade about perception of bullying, dealing with bullying and what kind of prevention measures are used in their schools. In the second part we were asking them about pupils' achievements at competitions of knowledge and in the third part about their job satisfaction. In the second part of the questionnaire we were faced with many answered items. In this presentation we will show what kind of tests and analyses were used for missing values and how much the analyses results depend on whether or not multiple imputation of missing values was performed. We will also provide possible reasons for not answered items and give directions to improve items response rates in our survey.

Model of usefulness of SPSS for students of economics and business: differences between undergraduate and postgraduate students

Urban Šebjan and Polona Tominc

Faculty of Economics and Business, University of Maribor, Maribor, Slovenia
urban.sebjan@uni-mb.si, polona.tominc@uni-mb.si

Nowadays, expertise in statistical software solutions represents a significant competitive advantage, both for individuals as well as for organizations, since databases are becoming increasingly large and complex. Usually we gain expertise in the implementation of statistical software during the educational processes at university. For students to perceive the usefulness of statistical software in the educational process, appropriate support from the teaching staff is needed. Learning support usually enables student to understand and perceive the usefulness of several statistical software solutions, that include the use of statistical methods for database analysis. In this paper a conceptual model of the impact of learning support on the perceived use of SPSS statistical software (Statistical Package for the Social Sciences) for students of the University of Maribor Faculty of Economics and Business is presented. The authors address the research question of whether there are significant differences between students of undergraduate and postgraduate programs in the relationship between constructs within the conceptual model. We have developed a conceptual model derived from the TAM model (Technology Acceptance Model). The structural equation modeling (SEM) was used. The partial least squares method and corresponding t-value allowed the authors to confirm the proposed relations and statistical significance difference, and to validate the conceptual model. The conceptual model showed a relationship between learning support and the following included constructs: perceived usefulness of SPSS, ease of use, and purpose of use. Based on a sample of 300 students and with help of SmartPLS software, we found that there are no statistically significant differences between students of undergraduate and postgraduate programs regarding relationship between constructs within the conceptual model. In the context of the individual constructs, however, we found that there are statistically significant differences between students at undergraduate and postgraduate levels. The level of perceived usefulness of SPSS by postgraduate students is higher as compared to students of undergraduate study. These findings are expected, because postgraduate students on general know more about the demands regarding quantitative aspects of study, and are as well aware of the need for additional knowledge which is required by companies and therefore perceive the usefulness of SPSS at higher level. At the same time, they also possess more knowledge obtained from their previous studies and gained by the study process in the past.

Dynamic graph generation and data analysis of complex data: a web-application based on R and shiny

Črt Ahlin[1], *Daša Stupica*[2], *Franc Strle*[2] *and Lara Lusa*[3]

[1]University of Ljubljana, Ljubljana, Slovenia

[2]Department of Infectious Diseases, University Medical Center, Ljubljana, Slovenia

[3]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia

crt.ahlin@gmail.com, cerar.dasa@googlemail.com, infek030@kclj.si, lara.lusa@mf.uni-lj.si

R statistical environment includes facilities for data display and analysis that are extremely flexible. It recently became rather straightforward to create interactive web applications and interactive graphics based on code written in R, using the `shiny` package. We illustrate, through an example, the feasibility of developing a user friendly web application that incorporates a variety of interactive graphical displays and tools for the analysis of complex data.

The `medplot` application was developed to help medical researchers explore and analyse longitudinal data, where numerous variables are recorded for each patient over time. Several interactive graphical displays allow an easy exploration of the data. The analyses tools evaluate the association of the variables with other characteristics of the patients, taking into account the multiple testing problem, the repeated measurements and the possibility of non-linear associations between the covariates and the outcomes.

The application can be used by users that are not familiar with R or other statistical programs. It can be used through a web-browser and it does not require the installation of any program. Template spread sheets for data preparation are also provided, together with example data from a clinical study including patients with erythema migrans, where the variables are the presence and intensity of numerous symptoms recorded over time.

Statistical Applications

Validity assessment in the mixed methods research from the view of mixed methods experts

Joca Zurc

Faculty of Health Care Jesenice, Jesenice, Slovenia

joca.zurc@guest.arnes.si

The mixed methods emphasize strengths and minimize weakness of quantitative or qualitative approach in single research study, and well correspondent to the many today's complex research questions in social sciences. Despite the importance, prevalence and popularity of mixed methods research, there has been lack of attention given on quality criteria development for mixed methods research. The purpose of this study was to identify the need, types and approaches of validity assessment criteria in mixed methods research thorough view of mixed methods experts. The data was collected with the structured interviews, in which participate 9 experts in mixed methods methodology, key speakers and workshops providers at the 1st Mixed Methods International Research Association Inaugural Conference 2014. The interviews duration was between 12 and 29 minutes. The experts agree that validity as a general construct is very important in all research, emphasis caution regarding terminology, justification of the mixed methods research, importance of complexity of the mixed methods research contexts and cultural variations. The experts agree there is a lack of validity criteria in the mixed methods research, although the field developed some validity assessment frameworks recently. The set of standardized criteria of mixed methods research will be useful for novice researcher and stakeholders, but difficult to develop because the mixed methods has many different models and depends on the context of particular research. Our study showed the importance of validity criteria in the mixed methods research, lack of recognized and structured criteria of validity assessment and close connection between the criteria and study phenomenon of the mixed methods research. The further studies are needed for empirically testing the existed theoretical models of mixed methods criteria, and to select and prioritized the most important set of validity criteria, which can be use to assess the study with quantitative and qualitative methodology integration.

Measurement of efficiency using Data Envelopment Analysis: an application in European Airline industry

Burak Keskin[1] *and Yucel Turker*[2]

[1]Cankiri Karatekin University, Cankiri, Turkey

[2]Eskisehir Osmangazi University, Cankiri, Turkey

burakkeskiin@gmail.com, yturker@ogu.edu.tr

The aim of this study is to present an application of Data Envelopment Analysis (DEA) to measure of efficiency 11 European airline companies which are members of Star Alliance organization. For that purpose number of aircrafts, number of airports served and number of employee were specified as input parameters; number of annual passenger and sales revenue were specified as output parameters. And then, companies were ranked based on their efficiency scores that were produced by Data Envelopment Analysis. Finally, some recommendations were given on how to improve low efficiency scores for inefficient companies.

An application of Liu-Type logistic estimators to the population data of Sweden

Adnan Karaibrahimoglu and Yasin Asar

Necmettin Erbakan University, Konya, Turkey

akara@konya.edu.tr, yasar@konya.edu.tr

Logistic regression methods have become an integral component of the data analysis concerned with the discrete binary or dichotomous outcome variable with many kinds of explanatory variables. It is known that the problem of multicollinearity affects the maximum likelihood estimator (MLE) negatively. Its variance is inflated and the estimations of the regression coefficients become unstable. In this study, we tried to get rid of this problem by using Liu-type logistic estimators with optimal shrinkage parameter and some existing ridge parameters. Thus, we designed a simulation study to evaluate the ridge regression parameters and determine the performances of the estimators. Moreover, we tried to model the data concerning the municipalities of Sweden. Since there exists a multicollinearity problem on the data, we used Liu-type logistic estimators in order to decrease the mean squared error (MSE) and overcome the multicollinearity problem.

Invited Lecture

Information theory and statistics: an overview

Daniel Commenges

Institut de Santé Publique, d'épidémiologie et de développement (ISPED), Université de Bordeaux and Institut National de la Santé de la Recherche Médicale (INSERM), Bordeaux, France

daniel.commenges@isped.u-bordeaux2.fr

We give an overview of the role of information theory in statistics, and particularly in biostatistics. We recall the basic quantities in information theory; entropy, cross-entropy, conditional entropy, mutual information and Kullback-Leibler risk. Then we examine the role of information theory in estimation theory. Then the basic quantities are extended to estimators, leading to criteria for estimator selection, such as Akaike criterion and its extensions. Finally we investigate the use of these concepts in Bayesian theory.

Biostatistics and Bioinformatics

Augmenting backward elimination by a standardized change-in-estimate criterion for variable selection in statistical models

Georg Heinze[1], *Max Plischke*[1], *Karen Leffondré*[2] *and Daniela Dunkler*[1]

[1]Medical University of Vienna, Vienna, Austria

[2]University of Bordeaux Segalen, Bordeaux, France

georg.heinze@meduniwien.ac.at, plischke@gmail.com, Karen.Leffondre@isped.u-bordeaux2.fr, daniela.dunkler@meduniwien.ac.at

Typical statistical modeling situations in prognostic or etiologic research often involve a large number of potential explanatory variables. Selecting among them the most suitable ones in an objective and practical manner is usually a non-trivial task.
We briefly revisit the purposeful variable selection procedure suggested by Hosmer and Lemeshow in their textbooks on binary and survival data analysis which combines significance and change-in-estimate criteria for variable selection and critically discuss the change-in-estimate criterion. We show that using a significance-based threshold for the change-in-estimate criterion reduces to a simple significance-based selection of variables, as if the change-in-estimate criterion is not considered at all. The change-in-estimate criterion may still be useful if in a specific situation the reported selected model should be sufficiently close to a model with all potential predictors. We propose to standardize the change-in-estimate criterion for this purpose and incorporate the standardized criterion in an augmented backward elimination (ABE) algorithm.
In a simulation study, where we compared ABE and backward elimination (BE) in linear, logistic and Cox regression, ABE tended to select larger models than BE, and often approximated the model with all potential predictors up to negligible differences in point estimates of the regression coefficients. On average, regression coefficients after ABE were closer to the coefficients of correctly specified models than after BE. In summary, we propose augmented backward elimination as a reproducible variable selection algorithm that gives the analyst more flexibility in adopting model selection to a specific statistical modeling situation.

The non-parametric approach to analysing time to event with missing at risk information

Maja Pohar Perme[1] *and Tomaž Štupnik*[2]

[1]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
[2]Department of Thoracic Surgery, University Medical Center, Ljubljana, Slovenia
maja.pohar@mf.uni-lj.si, tomaz.stupnik@kclj.si

When analysing time to disease recurrence in diseases with benign nature, patients are only seen at recurrences. When the average time to disease recurrence is long enough in comparison to the expected survival of the patients, several patients may die in the course of the study, but we shall have no record of that. Hence, the at risk number is wrong and the analysis of the time to recurrence biased. Under the assumption that the expected survival of an individual is not influenced by the disease itself, we try to reduce this bias by using the general population mortality tables. We focus on the non-parametric approach to this question and try to estimate the usual quantities of interest by correcting the at risk information. Our results are supported by simulations and real data examples.

Mortality model of a human age-structured population based on an interval type-2 fuzzy logic

Andrzej Szymański and Agnieszka Rossa

Institute of Statistics and Demography, University of Lodz, Lodz, Poland
anszyman@math.uni.lodz.pl, agrossa@uni.lodz.pl

Determination of the best mortality model is one of the basic problems in demographic forecasting. One of the popular mortality models is the so-called Lee-Carter stochastic mortality model (LC) for age-specific mortality rates mx(t), where x represents an age group and t – a calendar year. Koissi and Shapiro (2006) have formulated a fuzzy version of the LC model (FLC), in which mortality rates are assumed to be fuzzy numbers with the symmetric triangular membership function.
To make inference and forecast more precise and elegant an improved fuzzy LC model by Szymański and Rossa has been proposed, termed Extended Fuzzy Lee-Carter model (EFLC). In this approach the Banach algebra of fuzzy numbers introduced by Kosiński et al.(2004) – called OFN-algebra – has been applied. Formal details concerning EFLC model have been presented at the conference Applied Statistics in 2012.
The EFLC as well as LC model have been applied by the authors to forecast age-specific mortality rates in Poland. It has been shown that – in general – EFLC performed better than LC regarding predictive accuracy measures. However, for some cases results were equivocal. Therefore, we have decided to improve EFLC by means of the so-called interval type-2 fuzzy logic.
The concept of a type-2 fuzzy set was introduced by Zadeh (1975) as an extension of the concept of an ordinary fuzzy set, called a type-1 fuzzy set. Type-2 fuzzy sets are very useful especially in determining the unknown membership function.
In the presentation we will introduce the EFLC model with type-1 fuzzy sets replaced by interval type-2 fuzzy sets. The predictive accuracy of mortality forecasting by means of the proposed model will be also studied.
The research was supported for both authors by a grant from the National Science Center under contract DEC-2011/01/B/HS4/02882.

Image segmentation using a spatially regularized mixture model: application to lesion segmentation in stroke

Brice Ozenne[1], *Fabien Subtil*[1], *Leif Ostergaard*[2] *and Delphine Maucort-Boulch*[1]

[1]Service de Biostatistique, Hospices Civils de Lyon and Université Lyon I, Lyon, France

[2]Department of Clinical Medicine - Center for Functionally Integrative Neuroscience, Aarhus, Denmark

brice.ozenne@chu-lyon.fr, fabien.subtil@chu-lyon.fr, leif@cfin.dk, delphine.maucort-boulch@chu-lyon.fr

Lesion segmentation is a crucial but difficult task in medical imaging. Automated segmentation algorithms have been proposed based on intensity histogram analysis. Recent developments have shown that integration of spatial information enhance automatic image segmentation. However spatial modelisation is currently limited to short range scale dealing with small artefacts. A broader scale approach is then required for patients with white matter hyper-intensities which can be confounded with Stroke or Alzheimer Disease, leading to elongated shape artefacts. We developed a robust unsupervised algorithm using the flexibility of finite mixture models to incorporate spatial information and handle multiparametric modelisation. We extend spatial modelisation at the regional scale using a multi-order neighborhood potential. Estimation of the finite mixture model uses an EM algorithm with mean field approximation for the spatial model. A validation study was performed on simulated images including noise and moderate range artefacts. Short range regularization was able to remove most of the noise with 95.7% of good classification compared to 91.1% for the non spatial model. Furthermore, regional regularization was able to remove 100% of image artefacts, compared to 34.5% removed with short range regularization. Combining both regularizations leads to excellent results (99.9% of correct classification). Finally we apply the segmentation algorithm to 5 stroke patients with white matter hyper-intensities. Using a multivariate approach with short range and regional regularization succeeds in removing most of the artefacts, improving agreement with physician segmentation compared to non-spatial or short range regularized model. In conclusion, the proposed algorithm has shown good performances in lesion segmentation even with noisy images with artefacts.

Biostatistics and Bioinformatics

The ROC curve and their indexes based on bootstrap sampling

Pattaraporn Duriyakornkul and Pedro Oliveira

University of Porto, Porto, Portugal
pum_k@yahoo.com, pnoliveira@icbas.up.pt

ROC curve is perhaps the most used approach for the evaluation and comparison of diagnostic systems. Some indexes have been used to describe the accuracy of ROC curves such as the area under the curve (AUC), the confidence interval such as the vertical averaging, the confidence region and the optimum cutoff point. We studied two well-known confidence bands, namely, the Local Confidence Regions and the Simultaneous Joint Confidence Regions. In addition, we introduced the Bootstrap Band and the Bootstrap Percentile Band. Both of them are the confidence bands of Sensitivity and Specificity based on the bootstrapping technique. Lastly, we studied five indexes to provide the optimum cutoff point namely: 1-The point closest to ideal cutoff point (0,1), 2-Youden index, 3-Misclassification rate, 4-Minimax and 5-Prevalence matching with their confidence interval and percentile confidence interval based on bootstrap samples. Since the ROC curve is a graphically representation between the Sensitivity and Specificity, for all possible thresholds, in order to provide the confidence interval of the ROC curve, we have to create the confidence interval for both values along the threshold. To provide all of the bands and indexes we mainly used R programming. Comparisons between the different approaches for the definition of the confidence bands will be presented based on real and simulated data; at the same time, comparisons will be made for the five studied indexes. The confidence bands based on the bootstrap percentile bands seem to provide the narrower bands as opposed to simultaneous joint confidence regions; on the other hand, with respect to the indexes, although the point closest to ideal cutoff point (0,1),Youden index and misclassification present similar results, the Youden index seems to perform better.

Identification of tissue at risk after ischemic stroke

Ceren Tozlu[1], *Brice Ozenne*[1], *Leif Ostergaard*[2] *and Delphine Maucort-Boulch*[1]

[1]Service de Biostatistique, Hospices Civils de Lyon and Université Lyon I, Lyon, France

[2]Department of Clinical Medicine - Center for Functionally Integrative Neuroscience, Aarhus, Denmark

ceren_ial07@hotmail.com, brice.ozenne@chu-lyon.fr, leif@cfin.dk, delphine.maucort-boulch@chu-lyon.fr

Prediction of tissue infarction risk is a required information to help clinician to identify stroke patients for treatment. The Magnetic Resonance Imaging (MRI) is an effective tool which provides measurements that characterize the tissue status with different sequences (T2 FLAIR, diffusion and perfusion). Using parameters extracted from these MRI sequences, several classification methods have been proposed to predict the final outcome of tissue. The aim of our study is to compare the ability of five classification methods which have shown good performances in the literature to predict the brain tissue outcomes under two categories. (healthy or necrosis) The classification methods were used with 8 MRI parameters and applied on 5 patients. Logistic Regression (LR), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest (RF) and Adaboost were applied and compared in terms of Area Under ROC Curve value, sensitivity (Se), specificity (Sp), proportion of correctly classified (Pcc), Jaccard and accuracy of predicted infarction volume was assessed. Different AUC results were found among the methods where the highest value is obtained by Random Forest (0.994) and the lowest by Logistic Regression (0.914). Considering Se, Sp, Pcc and Jaccard, the RF was also found to perform best especially in terms of Sp and Jaccard. LR, SVM and ANN are systematically overestimated the infarction volume (LR =333.639 mL, SVM=214.734 mL, ANN=340.537 mL whereas RF was correctly calibrated (95.069 mL) vs the observed infarction volume is 95.175 mL). Contrary to previous studies on experimental animal data, we found that there were different performances among these five classification methods particularly for Random Forest which had excellent performance in terms of all evaluation criteria. SVM and ADA methods have also shown good performances after RF and the worst performer was the Logistic Regression. To confirm these results, future work will be performed on a larger data sample. Adjusting the threshold decision value for necrosis might enable better volume prediction for LR, SVM and ANN methods.

Reconstructing reasons for unsatisfactory status of Western Capercaillie (Tetrao urogallus) in Croatia through time and space

Andreja Radović[1] *and Robert Spanić*[2]

[1]Faculty of Science, University of Zagreb, Zagreb, Croatia
[2]Institute for Research and Development of Sustainable Ecosystem, Zagreb, Croatia
andreja.radovic@biol.pmf.hr, robert.spanic@ires.hr

The Western Capercaillie (Tetrao urogallus) is a bird species with wast distribution across the western hemisphere but unsatisfactory conservation status in Croatia. During the last century significant decline of population occured.The Red Book of Birds of Croatia classified status of the national population as endangered. We tried to reconstruct the main driving forces that caused population collaps in the last decades. We collected all available information on species from monitoring program such as: lek distribution on the part of the distribution range, number of males and females at leks, hunting practice, hunting period and similar. Knowing the species connectedness with spatial distribution of some plant species, we evaluated the Croatian territory for suitability of preferred species, Vaccinium mirtillus. In that process, we used diverse spatial environment variables like land cover, climate variables, acidity (ph) of the soils as well as variables prepared by authors, like Shannon-Wienner Index of habitat diversity. Technique used was Ecological Niche Factor Analysis (ENFA), method that uses presence-only data to describe the marginality and specialisation factors of the species. Marginality describes species selectivnes of specific habitat characteristics in regards to available habitats in the area and specialisations reveals niche breath. The valorisation of Croatian territory for Vaccinium mirtillus revealed a very good match with historical distribution of Western Capercaillie in the country. Furthermore spatial evaluation reveiled habitat fragmentation as the most probable reason for creating smaller ”islands” of sub populations in the past that could not communicate ammong each other. This eventually led to the local extinctions. Additionally, we evaluated hunting practices through the time that disabled population to recover. In the light of our findings we evaluated new plans of hunting organisations for enlargement of game production in the area of remaining active laks in the country.

Particle-based filtering applied to medical time series

Mohamed M. Shakandli

Sheffield University, Sheffield, United Kingdom

MMShakandli1@sheffield.ac.uk

This talk concerns the set-up and application of particle filtering to medical time series. Considering count time series (such as number of asthma patients recorded over time) we discuss and propose non-linear and non-Gaussian state space models, in particular dynamic generalized linear models (DGLMs). Inference and forecasting is achieved by employing sequential Monte Carlo methods, also known as particle filters. These are simulation based methods used for tracking and forecasting dynamical systems subject to both process and observation noise in non-linear and non-Gaussian models. They have far-reaching and powerful applications in time series analysis problems involving state space models. In this talk we propose the DGLMs as a modelling framework suitable to asthma data and we will discuss its practical implementation. We will specifically address the important problem of estimating static parameters, such as autoregressive coefficients and variance components, within the context of recursive Bayesian estimation. In our formulation we discuss Poisson and negative binomial type responses and we highlight implementation issues.

Bibliometry

On standardization of the Activity Index

Nataša Kejžar and Janez Stare

Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
natasa.kejzar@mf.uni-lj.si, janez.stare@mf.uni-lj.si

Relative Specialization Index (RSI) was introduced as a simple transformation of the Activity Index (AI), the aim of this transformation being standardization of AI, and therefore more straightforward interpretation. RSI is believed to have values between -1 and 1, with -1 meaning no activity of the country (institution) in a certain scientific field, and 1 meaning that the country is only active in the given field. While it is obvious from the definition of RSI that it can never be 1, it is less obvious, and essentially unknown, that its upper limit can be quite far from 1, depending on the scientific field. This is a consequence of the fact that AI has different upper limits for different scientific fields. This means that comparisons of RSIs, or AIs, across fields can be misleading. We therefore believe that RSI should not be used at all. We also show how an appropriate standardization of AI can be achieved.

Co-authorship structures of researchers in scientific disciplines in time

Marjan Cugmas, Anuška Ferligoj and Luka Kronegger

Faculty of Social Sciences, University of Ljubljana , Ljubljana, Slovenia
marjan.cugmas@fdv.uni-lj.si, anuska.ferligoj@fdv.uni-lj.si,
luka.kronegger@fdv.uni-lj.si

The scientific collaboration networks of Slovenian researchers at the level of scientific disciplines during two ten-year time periods (from 1991 to 2000 and from 2001 to 2010) are studied. The collaboration is defined as co-authorship of one or more published outputs that Slovenian Research Agency (ARRS) evaluates as scientific works.
The analysis is based on the work of Kronegger et al (2011), who studied the co-authorship networks of four scientific disciplines in Slovenia and proposed the most typical form of collaboration structure that consists of multi core - semi-periphery - periphery. Based on their study, we applied pre-specified blockmodeling to the most of the scientific disciplines (we excluded those disciplines with too small number of the researchers in them). In the presentation we will discuss the problem of determining the optimal number of clusters and the local optimization problem to determine structurally equivalent clusters. We will graphically present the dynamics of the collaborating groups over time obtained by the blockmodeling procedure.
To measure the transitions between two time periods or the stability of obtained clusterings we used Adopted Rand Index. To explain the obtained transitions by the characteristics of the co-authorship networks (e.g., the number of researchers in the discipline, average number of co-authors of the researchers in the discipline, the rate of change of the number of researchers, the rate of change of density of the network), the characteristics of the obtained block structures (e.g., the number of core clusters, the proportion of the periphery, the number of bridging cores) and the characteristics of the disciplines will be used.

Stability of co-authorship blockmodeling structure in time

Luka Kronegger, Anuška Ferligoj and Marjan Cugmas

Faculty of Social Sciences, University of Ljubljana , Ljubljana, Slovenia
luka.kronegger@fdv.uni-lj.si, anuska.ferligoj@fdv.uni-lj.si, marjan.cugmas@fdv.uni-lj.si

Co-authorship as form of scientific collaboration presents the major interaction mechanism between actors at the micro-level of individual scientists. Wide range of mechanisms fostering collaboration produce different structures within general network. The dynamic nature of co-authorship networks presents an interesting problem when trying to analyze the properties of established, emerging and dissolving groups of co-authoring researchers in time.
To analyze the properties of structure dynamics we used blockmodeling method (structural equivalence) with following of individual researchers through time (Kronegger et. al 2011), and stochastic actor based modeling of network dynamics (Siena). In Siena we used two approaches to modeling the effect of structural equivalence to formation of ties within the network: i) including the "balance effect" (Ripley et. al 2013) which is included in predefined set of Siena effects and ii) including the information on structural equivalence on dyadic level using dissimilarity matrix as explanatory variable.
In our research we observed and compared collaborative structures in complete longitudinal co-authorship networks for selected disciplines. Dataset gathered from national bibliographic system COBISS, spanning from 1996 to 2010, was split into three consecutive five-year intervals.

A comparison of methods for capturing the treatment effect of an observational study for programme evaluation

Thanawit Bunsit

Thaksin University, Songkhla, Thailand

thanawit.b@tsu.ac.th

This paper aims to compare the estimation of the impact of microfinance programme on happiness and wellbeing by comparing different types of econometric models for capturing the genuine impacts. The outcome for programme evaluation includes self-reported happiness and subjective wellbeing indicators such as life domain satisfaction, positive and negative affects.

Different types of methods were employed and compared the results. The first method was the matching estimators including the bias-corrected matching estimator and the matching with bias-corrected and robust variance estimators. For the matching with bias-corrected and robust variance estimators, the treatment effect on subjective and psychological wellbeing was assessed considering the following covariates or matching variables: (a) borrower's gender; (b) borrower's age; (c) borrower's age squared; (d) borrower's education attainment; (e) borrower's health status; and (f) locality (living in the remote area or in the village centre).

This study also estimated the treatment effect using propensity score matching with nonparametric regression. Two matching approaches were used in order to estimate the impact of participation in the microfinance programme on wellbeing: the kernel estimator approach and the local linear regression estimator approach. Several methods have been developed to estimate a kernel function. In this study four types of kernel functions were used including the tricube kernel (TRI), the Gaussian kernel (GAU), the rectangular kernel (REC) and the Epanechnikov kernel (EPA).

By comparing different types of methods for estimating the treatment effect, it was found that using methods with adjusted mean difference showed better results of the counterfactuals representing different treatment effects of interest than the unadjusted mean difference methods. The results from this study can be applied to all programme evaluation especially the government intervention programme.

Invited Lecture

Bayesian uncertainty analysis for complex physical systems modelled by computer simulators

Michael Goldstein

The University of Durham, Durham, United Kingdom
michael.goldstein@durham.ac.uk

Most large and complex physical systems are studied by mathematical models, implemented as high dimensional computer simulators. While all such cases differ in physical description, each analysis of a physical system based on a computer simulator involves the same underlying sources of uncertainty. There is a growing field of study which aims to quantify and synthesise all of the uncertainties involved in relating models to physical systems, within the framework of Bayesian statistics, and to use the resultant uncertainty specification to address problems of forecasting and decision making based on the application of these methods. This talk will give an overview of aspects of this emerging methodology, with particular emphasis on Bayesian emulation, structural discrepancy modelling and iterative history matching. The methodology will be illustrated with examples of current areas of practical application.

Econometrics

An estimate of the degree of interconnectedness between countries: a lasso approach

Davide Fiaschi and Angela Parenti

University of Pisa, Pisa, Italy

davide.fiaschi@unipi.it, aparenti@ec.unipi.it

This paper provides a methodology to estimate the degree of economic interconnectedness across different regions, and applies such methodology to a sample of countries in the period 1960-2008. The first step consists in the estimate of a panel of volatility growth rates in each period for each country, under the assumption that country growth rates follow an autoregressive process; in the second step, via a generalized variance decomposition analysis, this panel is used to estimate the *connectedness matrix* (called *adjacent matrix* in network literature), which measures the interconnectedness between regional income, conditioned to how many periods ahead we allow the shocks to propagate across regions.
The estimated connectedness appears very heterogeneous and not symmetric; the own connectedness is not very relevant (at most 14% of the variance is due to shocks arising and remaining in the same country, i.e. true idiosyncratic component of regional shocks); there exists a periphery of countries with high interconnectedness inward and outward; and, finally, the countries with the highest negative net value of connectedness are in the core of Western countries (this means that they are net source of shocks for the other countries). Finally, the comparison of this connectedness matrix with some spatial matrices generally used in spatial econometrics reveals that an exogenous spatial matrix, as, e.g., a first-order or a second-order matrix, are far from representing the actual interconnectedness between countries.

Spatial clubs in European regions

Davide Fiaschi[1], *Lisa Gianmoena*[2] *and Angela Parenti*[1]

[1]University of Pisa, Pisa, Italy

[2]IMT Institute for Advanced Studies, Lucca, Italy

davide.fiaschi@unipi.it, lisa.gianmoena@imtlucca.it, aparenti@ec.unipi.it

The aim of this paper is to identify the possible presence of spatial clubs in European regions. First, we propose a new methodology based on nonparametric approach to identify spatial clubs, and we apply it to a sample of 257 European regions in the period 1991-2008. Secondly, we identify the contributions of spatial spillovers versus cross-region heterogeneity in formation of those spatial clubs. In particular, 1. the Moran scatter plot of per worker GDP in 2008 highlights the presence of three spatial clubs: one populated by regions belonging to the former Eastern Bloc countries, one by regions of PIGS countries (Portugal, Italy, Greece and Spain) and the last one by regions of other EU countries (notably Germany, France, UK and Northern Europe countries). The dynamic extension of the Moran scatter plot, which consists in the non parametric estimate of the joint dynamics for the period 1991-2008 shows that in the long-run the convergence should happen only to two spatial clubs: with Eastern regions converging to PIGS regions. 2. Spatial spillovers are present across European regions, but their contribution to the emergence of spatial clubs is very low as compared to regional (unobserved) characteristics. A note of caution on our findings derives from the missing inclusion of some crucial characteristics of regions in the analysis (the most important the regional levels of human capital).

Bivariate regression model for count data based on the generalised Poisson distribution

Vera Hofer and Johannes Leitner

University of Graz, Graz, Austria
vera.hofer@uni-graz.at, johannes.leitner@uni-graz.at

A regression model for bivariate count data is introduced. The most popular approach to construct a bivariate distribution is trivariate reduction. Using three independent Poisson distributed variables X_1, X_2, X_3, a bivariate distribution (Y_1, Y_2) with Poisson marginals is derived as $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$. The disadvantage of this approach is that the covariance can only be positive.
The bivariate count data model presented here is based on the Sarmanov distributions. The marginals are chosen to be generalised Poisson distributions. Even though the derivation of the distribution is demanding, a closed form of the likelihood function is obtained using the Lambert-W-function. Thus common maximum likelihood parameter estimation without simulation can be carried out. Our model yields less complicated formulas than copula based models and parameter estimation is less time-consuming.
The model is applied to artificial data and a large real dataset on health care demand. Its performance is compared to alternative models presented in the literature. Our model not only turns out to be superior to the other models in many settings. In addition, it gives insights into influences on the variance of the response variables.

Local finance and the demand for property-casualty insurance

Giovanni Millo[1] *and Pietro Millossovich*[2]

[1]Generali R&D, Trieste, Italy

[2]Cass Business School, London, United Kingdom

giovanni.millo@generali.com, Pietro.Millossovich.1@city.ac.uk

Interest rates affect both non-life insurance supply and demand, possibly in opposite directions. Insurers issue contingent debt contracts and invest the funds until they are needed to pay the claims, so interest rates are a source of revenue for the insurers and a cost for the insured. Therefore a negative effect on demand should be expected: yet the theory is inconclusive and empirical evidence is scant. This paper focuses on analysing, from both a theoretical and empirical point of view, the role of the cost of financing on the insurance decision. We first develop a formal intertemporal model of optimal insurance when there is a spread between lending and borrowing rates. We study the resulting insurance demand and provide conditions for it to be inversely related to the cost of financing. Then, we test the implications of the formal model on an actual data set. A problem of aggregate models of insurance consumption is the unavailability of price data, which hinders the estimation of supply and demand equations. In general, only total revenues are observable. To isolate the effect on demand, we resort to a new observational context: a panel of Italian provinces over five years. At this level, the insured face local borrowing conditions while the insurers' returns are uniform. We bring evidence that demand for non-life insurance is in fact decreasing with the interest rate on borrowing. This result is robust across a number of specifications. Spatial econometric techniques are employed to ensure consistent inference. We conclude that credit conditions are a significant driver of non-life insurance development, and an important limiting factor in the particular case of Southern Italy.

Statistical Applications

A multilevel analysis on the importance of work in several European Union's countries

Laura Asandului and Roxana Otilia Sonia Hritcu

"Alexandru Ioan Cuza" University, Iasi, Romania
lasandului@gmail.com, hritcu.otilia@gmail.com

The purpose of the paper was to analyse the individuals' opinion regarding the level of the importance of work from several countries in the European Union. People living within the same country may be more similar to each other than people living in other countries; they share a similar lifestyle, same social factors, and health care availability, which may have a collective influence over and above individual circumstances.
The claimed level for the importance of work may be determined by factors that define the job (degree of independence at work) and demographic characteristics such as age, gender, marital status or education level.
The method used is multilevel analysis. Multilevel models assume that data are hierarchical, with the response variable measured at the lowest level, and explanatory variables measured at all existing levels. The analyzed data is a subsample of the World Values Survey Wave 6 database, a global study of what people value in life that was carried out between 2010 and 2012. We studied if, on average, the countries differ in their inhabitants' opinion on the importance of their work and we analyzed the relationship between the respondents' individual characteristics (gender, age, marital status, education level, social class, employment status), the specific job characteristics (business sector, degree of independence at work,) and importance of work with focus on Central and Eastern Europe.

Multivariate stochastic volatility estimation using Particle filter

Jian Wang

The University of Sheffield, Sheffield, United Kingdom
jwang33@shef.ac.uk

This talk considers a modelling framework for multivariate volatility in financial time series. The talk will briefly review particle filtering or sequential Monte Carlo methods. An overview of the multivariate volatility modelling literature will be given. As most financial returns exhibit heavy tails and skewness, we are considering a model for the returns based on the skew-t distribution, while the volatility is assumed to follow a Wishart autoregressive process. We define a new type of Wishart autoregressive process and highlight some of its properties and some of its advantages. Particle filter based inference for this model is discussed and a novel approach of estimating static parameters is provided. The proposed methodology is illustrated with two data sets consisting of asset returns of the FTSE-100 stock exchange.

Risk measurement of the future annuity prices: effects of different interest rate models

Ugur Karabey and Sule Sahin

Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey
ukarabey@hacettepe.edu.tr, sule@hacettepe.edu.tr

The purpose of this study is to analyse the effect of different interest rate models on the prices and the risk measures of the annuity portfolios. The main concern of the risk managers of the insurance companies that provide annuities is measurement and management of financial risks. Mortality and interest rates are two main components of the financial risk. Thus, we need models for future mortality rates and future interest rates. In this study mortality rates are modelled by the Cairns-Blake-Dowd model (Cairns et al., Journal of Risk and Insurance, 2006, 73, 687–718). On the other hand, interest rates effect the future price of the bonds that annuity providers buy in order to make annuity payments and they are also used for discounting annuity values for valuation. As for the interest rate models we use the Cox-Ingersoll-Ross model (Cox et al., Econometrica, 1985, 54, 385–407) and the yield curve model proposed by Sahin et al. (Annals of Actuarial Science, 2014, 8, 99-130) which is constructed based on the UK nominal, real and implied inflation spot zero-coupon rates simultaneously for actuarial applications. Furthermore, after calculation of the total risk capital we determine factor risk contributions such as interest rates and mortality rates.

Statistical Applications

Sports predictions made by a statistical model: a Sochi case

Slavko Jerič

RTV Slovenija, Rakek, Slovenia
slavko.jeric@rtvslo.si

Sport is very unpredictable. Just ask bookmakers, who generate large profits from sports. Everyone would like to know who will win specific sporting event. Many sports media predict medallists prior Olympic Games. These predictions are usually drawn up by experts covering a particular sport. But their selection is the result of subjective decisions. Recently some statistical companies predict solely on the basis of statistical models. One of them is Infostrada Sports. Dutch company is working with the International Olympic Committee since 2000.
Infostrada Sports firstly announced Olympic medallists for Games in London 2012. They correctly predicted a third of Olympic champions and a half of medallists. Infostrada Sports has updated its statistical model ahead of the Winter Games in Sochi 2014. Their work was based on a statistical model which took account of three factors: performance (the better an athlete's performance in an event, the more points he/she is awarded in the model), time (results in winter sports from Vancouver 2010 onwards have been assessed but more recent results are given a higher weighting. A World Championship gold in 2013 weighs far more than similar success in Vancouver) and importance of the event (A World Championship weighs more than a World Cup event).
At least four media predicted medallists for every event at Sochi Games prior the event. The analysis shows that the accuracy of statistical forecasts made by Infostrada Sports is virtually same as the media predictions.

Analysis of the impact of fatigue on the running technique

Melita Hajdinjak[1] *and Martin Krašek*[2]

[1]Faculty of electrical engineering, University of Ljubljana, Ljubljana, Slovenia

[2]Faculty of sport, University of Ljubljana, Ljubljana, Slovenia

melita.hajdinjak@fe.uni-lj.si, martin.krasek@triatlonklub-lj.si

A junior triathlet was video recorded while running 3 km as fast as possible, once after a normal running warming up (isolated run) and once right after a quality bike workout (transition run). The analysis focused on 10 control points—the first was set at 100 m distance after the start, and the last at 100 m distance before the end of run. Several running parameters were measured, such as speed of running, stride length and frequency, duration of different phases of the running gait cycle, several angles (upper torso, ankles, hips, knees) as well as the oscillation of the centre of gravity. Where applicable, each leg was observed separately. The aim of the study is to detect possible differences between the biomechanics of both runs.

We have performed several statistical tests on the data. First, regression analysis of different types has been done to estimate the relationships between the distance or control point and single running parameters as dependent variables. For the parameters that could not be modelled satisfactorily randomness tests have been performed. Differences among pairs of regression models for both runs (isolated run and transition run) and several running parameters, including the upper torso forward lean and the oscillation of the centre of gravity, have been detected. Second, in order to see if there are significant differences in single running parameters for both types of runs the single factor repeated measures ANOVA has been used. Significantly different values of some running parameters, including running speed, stride length and stride frequency, have been observed.

The observed differences indicate the impact of fatigue on the running technique, and show which specific abilities and conditions a triathlet needs to control muscle fatigue during a triathlon race more efficiently.

Communication and adoption dynamics in new product life cycle: the case of Apple iPod

Mariangela Guidolin

Department of Statistical Sciences, University of Padua, Padua, Italy
guidolin@stat.unipd.it

The ability to forecast new product growth is especially important for innovative firms that compete in the marketplace. Innovation diffusion models are used in business statistics and quantitative marketing to describe and forecast the growth dynamics of new products and technologies when launched into markets. These models help understand the mechanisms that generate and enhance the penetration of an innovation into a socio-economic system. The most famous diffusion model is that by Bass, BM, which forms the basis of several extensions of it. One of these, the Guseo-Guidolin model, GGM, assumes a time-dependent market potential, which is generated by a communication process among consumers. This model has turned out a useful improvement of the Bass model in different industrial sectors (e.g. pharmaceutical, energy). In this paper, we propose a new application of the GGM to the yearly sales data of the iPod, the portable media player designed and marketed by Apple inc., highlighting the forecasting upgrade obtained with respect to the simple Bass model. Moreover, we show that the GGM not only allows a separation between communication and adoption dynamics in new product life cycle, but also permits a temporal allocation of these two phases, based on easy to compute location indexes (mode, median, and mean). Interestingly, we see that in the case of the iPod the phase of adoption precedes that of communication, underlying the crucial role played by the group of early adopters in determining the success of the product. From a statistical point of view, the estimation of the GGM only requires sales data and is performed with standard Nonlinear Least Squares techniques through the Levemberg-Marquardt algorithm.

Statistical prediction in the production of vulcanized rubber products

Melita Hajdinjak and Gregor Dolinar

Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
melita.hajdinjak@fe.uni-lj.si, gregor.dolinar@fe.uni-lj.si

We will present how statistical methods have been integrated to evaluate and improve the quality of technological processes in the production of various vulcanized rubber products. We concentrate on the prediction of production-parameter values.
Using standard image analysis methods we automatically determine several graphical properties of the cross-section images of vulcanized rubber products. These include not only the area and the perimeter of the cross section, of its convex hull and its bounding box, but also more shape-describing properties such as eccentricity, the Euler number, extent and solidity. The graphical properties help us grouping already manufactured rubber products in 4 clusters such that products in the same cluster are more similar (in a chosen sense) to each other than to those in other clusters. The graphical as well as additional non-graphical properties of already manufactured products such as specific weight, hardness and tensile strength of the used rubbers are then used to build several regression models (one set of models for each cluster). The predicted variables are the values of different technological parameters such as generator powers, speed of the injector, pull speed and various temperatures.

Modeling and Simulation

Departure from uniform association in square contingency tables

Serpil Aktas and Ayfer Ezgi Yılmaz

Hacettepe University, Ankara, Turkey
serpilaltunay@gmail.com, ires

Square contingency tables where there is one to one correspondence between the row and column variables are RxR tables. For the analysis of two way square contingency tables with ordered categories, Goodman (1979) considered some association models, for example, the uniform association model, which is a generalization of the independence model and takes the order restrictions into account in the analysis of contingency tables. For this model all local odds-ratios are considered to be a uniform local association for all cells in the square table. In this paper, a measure of departure from uniform model is proposed using the local odds ratios. This measure allows us to test the several RxR tables from the uniform association model.

Estimating dynamic panel data models with random individual effect: Instrumental Variable and GMM approach

Johnson T Olajide[1], *Olusanya E. Olubusoye*[2] *and Iyabode F Oyenuga*[1]

[1]Mathematics & Statistics Department, The Polytechnic Ibadan, Ibadan, Nigeria
[2]Department of Statistics, University of Ibadan, Ibadan, Nigeria
taiwoolajide2004@yahoo.co.nz, busoye2001@yahoo.com, yaboadebayo2001@yahoo.com

This paper investigates the performance of some Instrumental Variable (IV) Estimators and Generalized Method of Moment (GMM) Estimators of a dynamic panel data model with random individual effect and compare them in terms of their bias and mean square error. The Monte Carlo study were performed to study the impact of sample size on the performance of different estimators and four different generating schemes for the serial correlation of the error term, namely autoregressive of order one (AR(1)) , autoregressive of order two (AR(2)), moving average of order one (MA(1)) and moving average of order two (MA(2)) were considered. The result s of the simulation shows that Anderson –Hsiao Instrumental Variable Estimator in difference form(AH(d)) is found to be the best when the time dimension is small while the one step Arellano-Bond Generalized Method of Moment (ABGMM(1)) perform better when the time dimension is large. The bias of most of the estimators improves as the time dimension increases except in some cases. The effect of serial correlation is minimal using different generating procedures.

Power of tests of heteroscedasticity in non-linear model

Iyabode F Oyenuga[1] *and Benjamin A Oyejola*[2]

[1] Department of Mathematics and Statistics, The Polytechnic, Ibadan, Oyo State, Nigeria

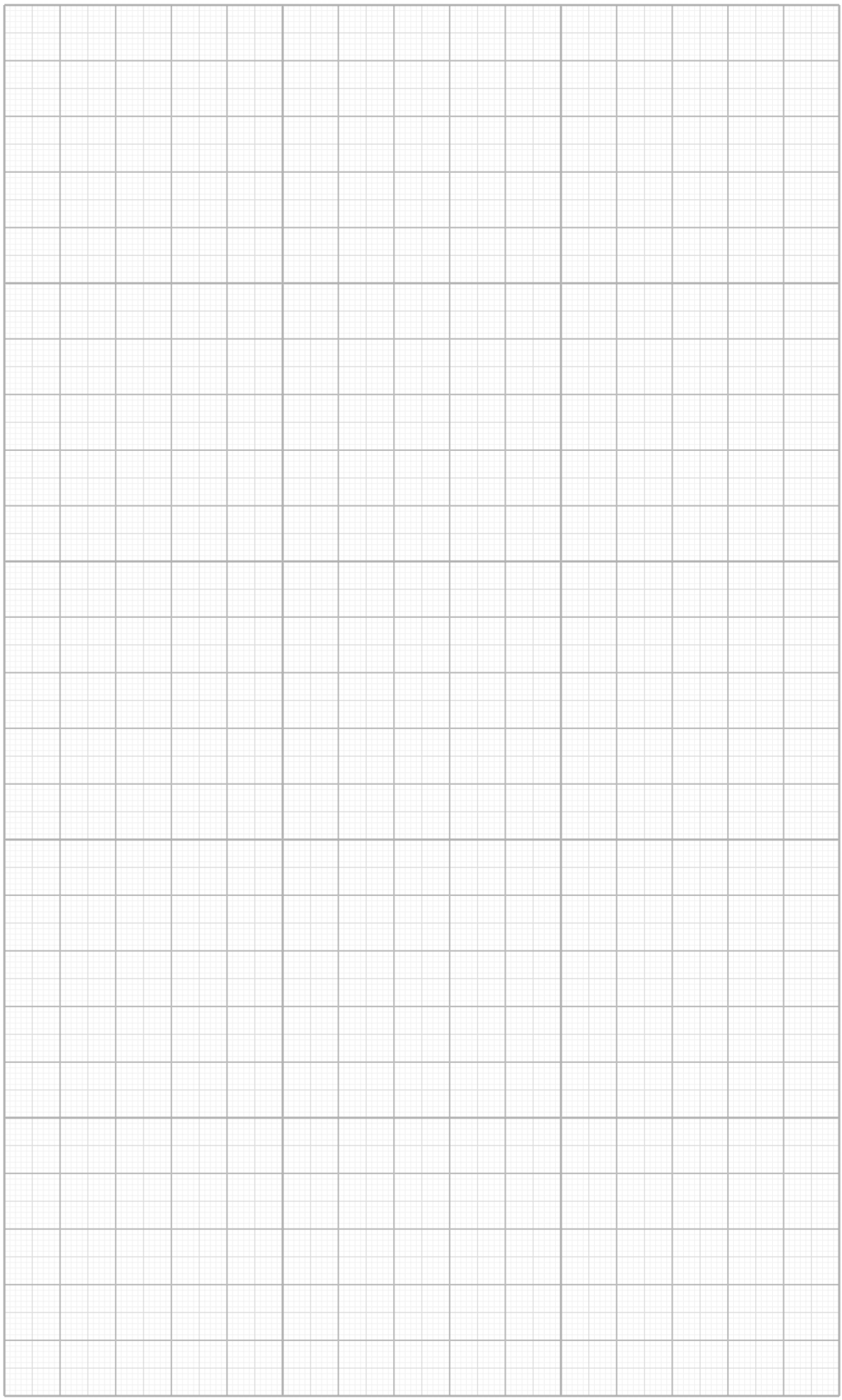
[2]Department of Statistics, University of Ilorin, Kwara State, Nigeria

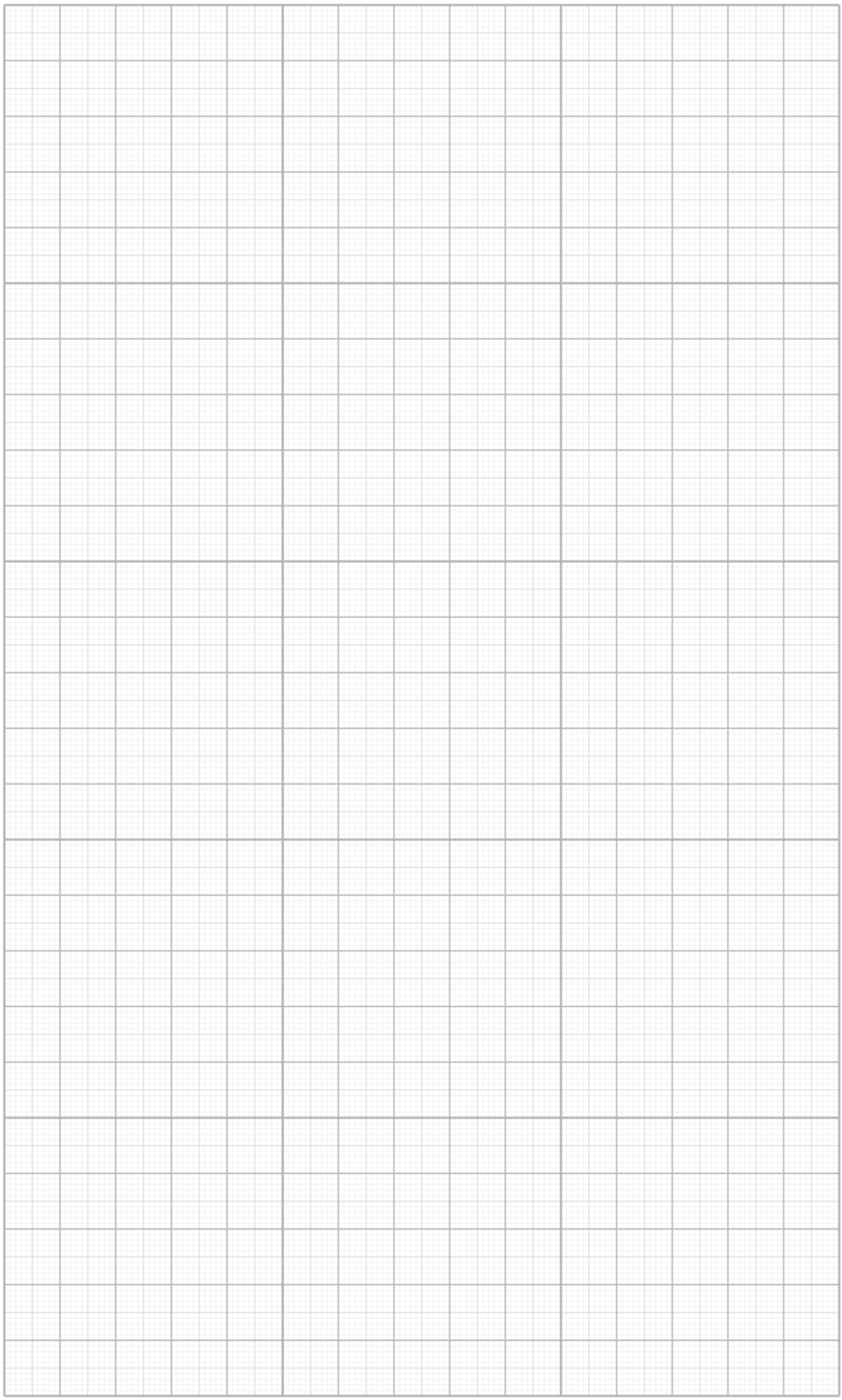
iyaboadebayo2001@yahoo.com, boyejola2003@yahoo.com

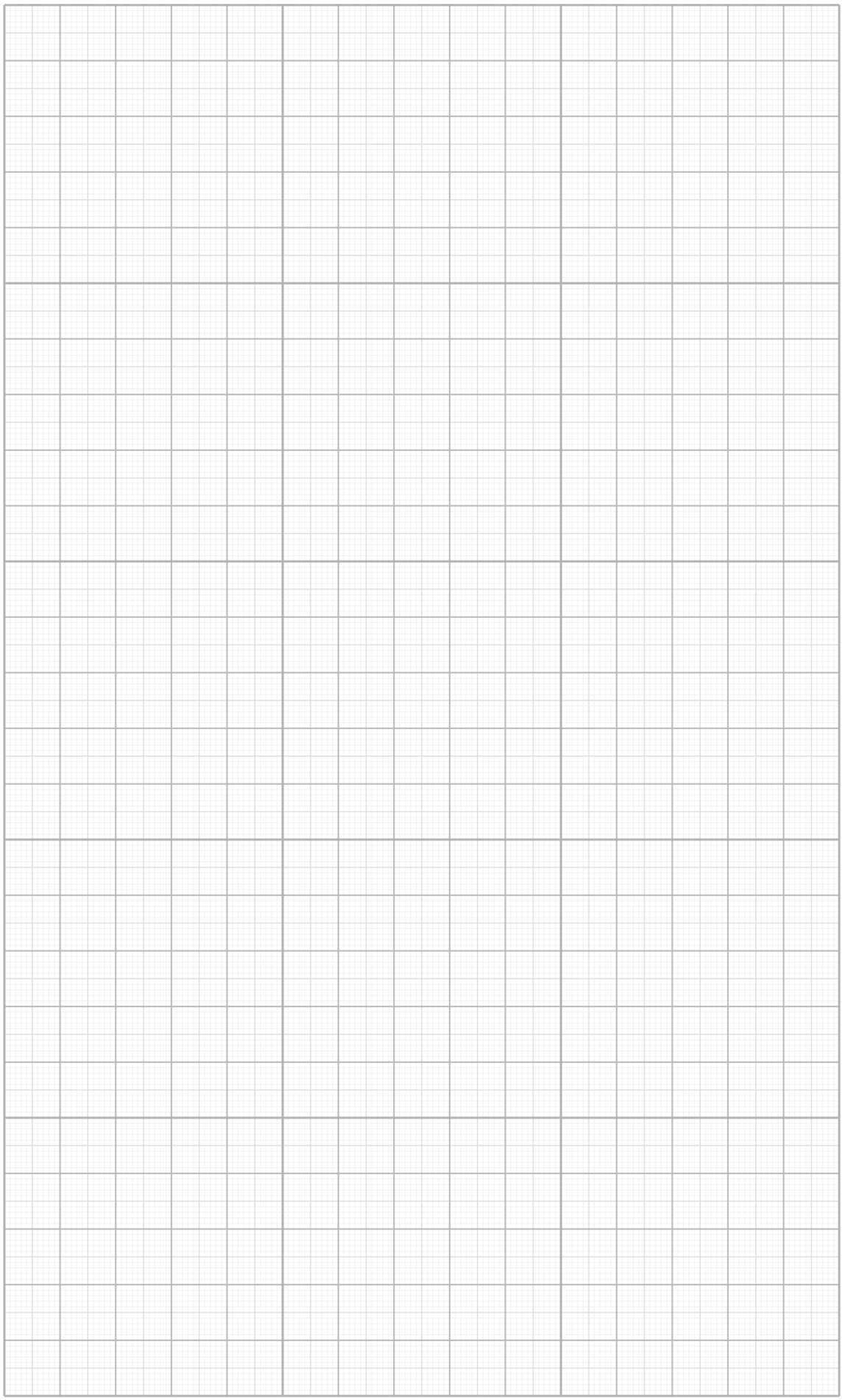
This paper investigates the performances of three different tests of estimating the power of heteroscedasticity in a non-linear model. A bootstrap experiment was performed. The tests used include: Breusch-Pagan, Glejser and Park. The program using R package was developed to carry out this study using Cobb-Douglas production function model. β_1, β_2 and β_3 are the parameters introduced into the error structure and estimated using Newton- Raphson method. Different degrees of heteroscedasticity $\lambda = 0.0$ (homoscedasticity), $\lambda = 0.3$ (weak heteroscedasticity),$\lambda = 0.6$ (mild heteroscedasticity), $\lambda = 0.9$ (strong heteroscedasticity) and $\lambda = 2.0$ (severe heteroscedasticity) with sample sizes 10, 50, 100 and 200 were considered. Bootstrap simulation was based on 1000 iterations and was carried out under normal distribution. The power of test in this paper is measured by the number of times the NR2 statistics exceed the 5% and 10% critical value of the χ^2 distribution for Breusch-Pagan test and t- distribution for both Glejser and Park test with n-1 degree of freedom. We found out that that at small sample, the presence of heteroscedastic is high and as the sample size increases, the heteroscedastic disappear gradually. Finally, comparison of behavior of power of tests show that among the three tests, Breusch –Pagan performed better than the other two tests and Park is the least powerful.

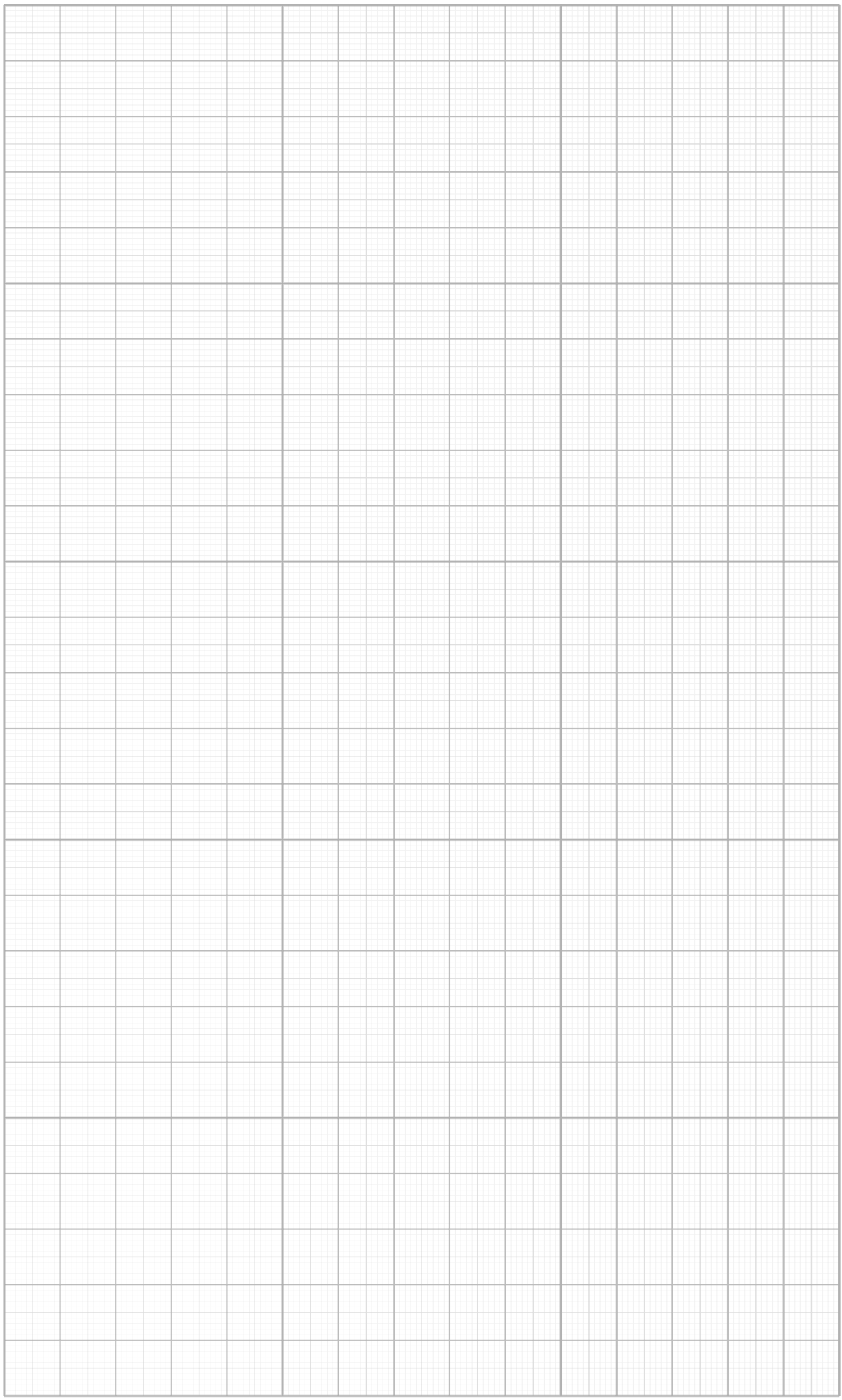
INDEX OF AUTHORS

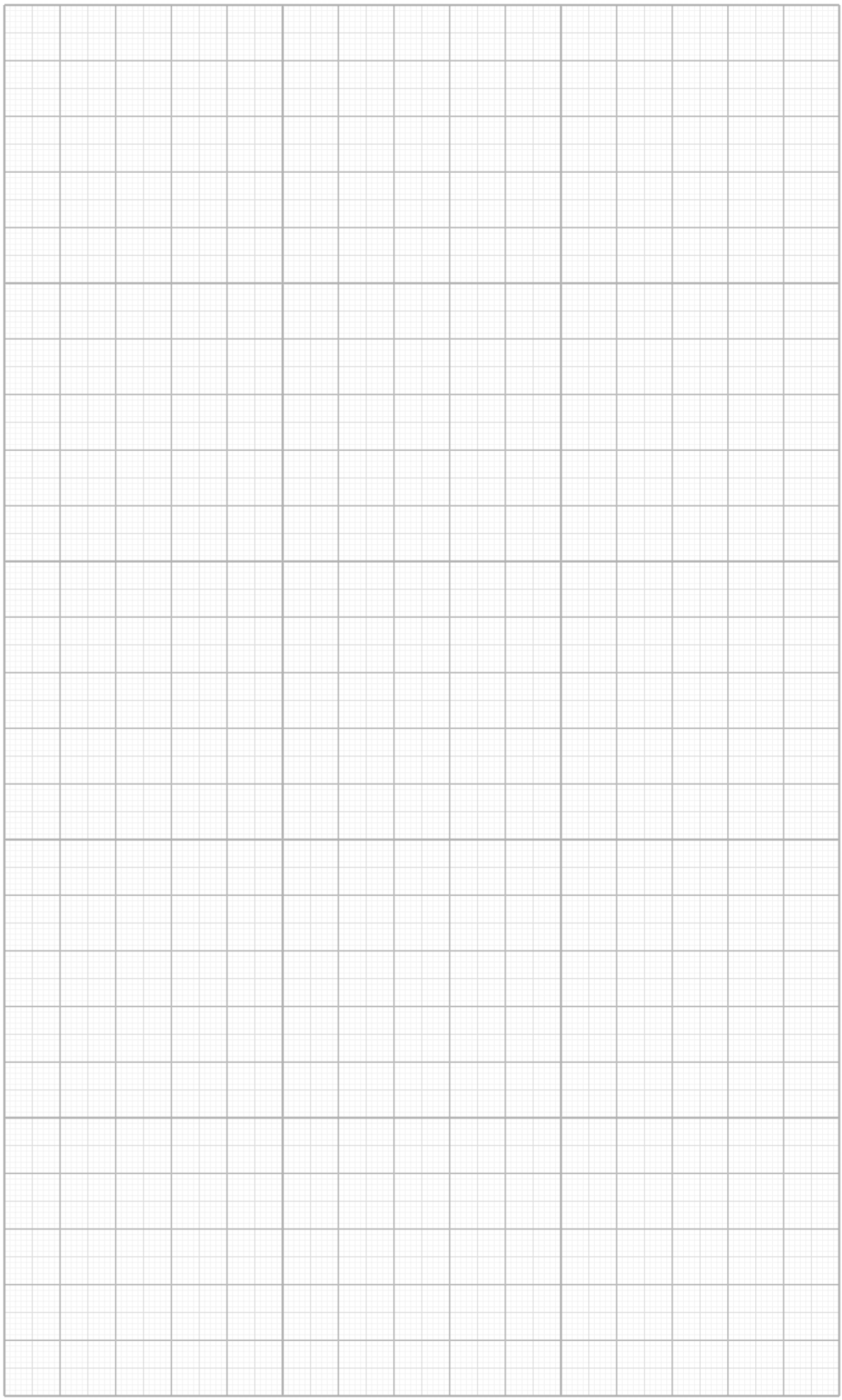
Index of Authors

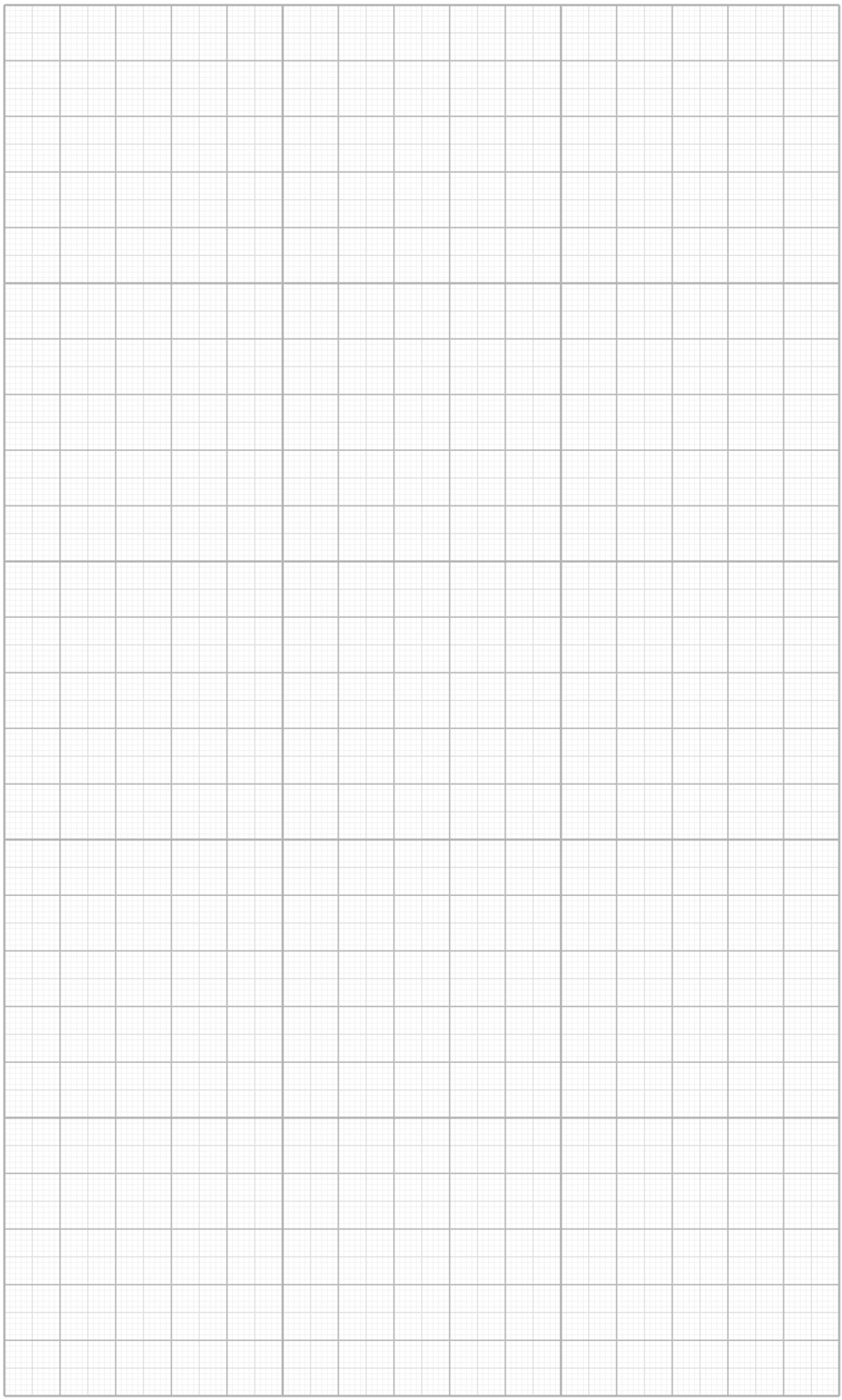


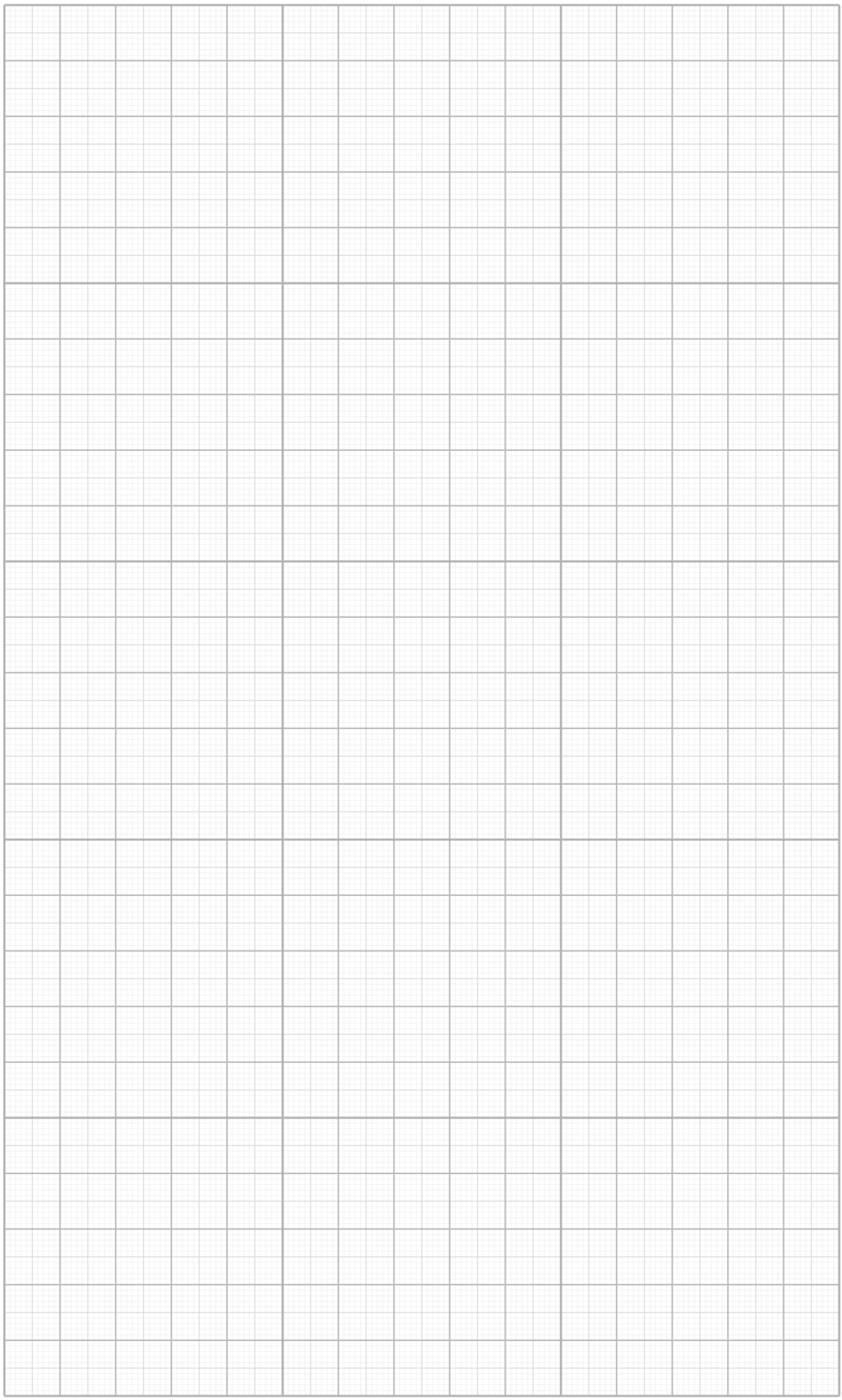












SUPPORTED BY

RESULT