International Conference

APPLIED STATISTICS

2015

ABSTRACTS and PROGRAM

September 20 - 23, 2015
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2015

International Conference

APPLIED STATISTICS 2015

ABSTRACTS and PROGRAM

2015

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2015

Organized by

Statistical Society of Slovenia

Supported by

NIB

IBMI

Alarix

RESULT

Generali

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)

INTERNATIONAL Conference Applied Statistics (2015; Ribno)
Abstracts and program / International Conference Applied Statistics 2015, Ribno (Bled), Slovenia organized by Statistical Society of Slovenia ; [edited by Lara Lusa and Janez Stare]. - Ljubljana : Statistical Society of Slovenia, 2015

ISBN 978-961-93547-4-2

1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
281080576

Scientific Program Committee

Janez Stare (Chair), Slovenia
Jacques Billiet, Belgium
Matevž Bren, Slovenia
Anuška Ferligoj, Slovenia
Dario Gregori, Italy
Irena Križman, Slovenia
Stanislaw Mejza, Poland
Jože Rovan, Slovenia
Vasja Vehovar, Slovenia
Vladimir Batagelj, Slovenia
Andrej Blejec, Slovenia
Maurizio Brizzi, Italy
Herwig Friedl, Austria
Katarina Košmelj, Slovenia
Lara Lusa, Slovenia
Mihael Perman, Slovenia
Tamas Rudas, Hungary

Organizing Committee

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

Published by: Statistical Society of Slovenia
Vožarski pot 12
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: 150

ABSTRACTS and PROGRAM

PROGRAM

		Hall 1	Hall 2
Sunday	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	Statistical Modeling	
	11.40 – 12.00	Break	
	12.00 – 13.20	Measurement	Statistical Applications
	13.20 – 15.00	Lunch	
	15.00 – 16.20	Modeling and Simulation	Statistical Applications
	16.20 – 16.40	Break	
	16.40 – 17.40	Measurement and Design of Experiments	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	Statistical Applications
	13.20 – 14.30	Lunch	
	14.30	Excursion	
Wednesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Network Analysis	Statistical Applications
	11.40 – 12.00	Break	
	12.00 – 13.20	Economics and Econometrics	Social Science Methodology
	13.20 – 13.30	Closing of the Conference	

15.00–18.00 **Workshop** (Hall 1)

1. **Research Data Management planning: problems and solutions**
   Irena Vipavc Brvar and Sonja Bezjak

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: Janez Stare*

1. **Dynamic treatment regimes as a prediction problem**
   Elja Arjas

10.00–10.20 **Break**

10.20–11.40 **Statistical Modeling** (Hall 1) *Chair: Elja Arjas*

1. **Potential parenthood and career progression of men and women - a simultaneous hazards approach**
   Martin Biewen and Stefanie Seifert

2. **Analysing survival of groups that outlive the population – the French Olympians study**
   Maja Pohar Perme, Juliana Antero-Jacquemin, Aurelien Latouche and Jean-François Toussaint

3. **Combining cross-sectional and panel information to disentangle attrition models in panels**
   Ulrich Pötter and Holger Quellenberg

4. **How to handle underreporting utilizing statistical models**
   Herwig Friedl

11.40–12.00 **Break**

12.00–13.20 **Measurement** (Hall 1) *Chair: Nataša Kejžar*

1. **Visibility graph analysis of seismicity: application to real and synthetic sequences**
   Luciano Telesca, Michele Lovallo, Alejandro Ramirez-Rojas and Tony A. Stabile

2. **Agreement among physicians assigning ICF categories based on medical documentation in a prosthetics and orthotics outpatient clinic**
   Gaj Vidmar, Helena Burger, Barbara Kavčič, Nataša Bizovičar, Pavel Ptyushkin and Bernard Francq

3. **Weighting methodology for a composite indicator of well-being**
   Jože Rovan, Kaja Malešič and Lea Bregar

4. **Outcome-based measures of the Rasch model fit**
   Gregor Sočan

12.00–13.20 **Statistical Applications** (Hall 2) *Chair: Klemen Pavlič*

1. **Evaluation of phenotypic variation of wheat hybrids for resistance to Fusarium head blight**
   Maria Surma, Tadeusz Adamski, Halina Wiśniewska, Karolina Krystkowiak, Anetta Kuczyńska, Zygmunt Kaczmarek and Stanislaw F. Mejza

2. **Multivariate approach for selection of SSD lines in grain legumes**
   Tadeusz Adamski, Maria Surma, Zygmunt Kaczmarek, Wojciech Świecicki, Paweł Barzyk, Anetta Kuczyńska and Karolina Krystkowiak

3. **X bar control chart for non-normal symmetric distributions**
   Kristina Veljkovic

4. **Control charts in anticoagulant treatment**
   Susana Martins, Lino Costa and Pedro Oliveira

13.20–15.00 **Lunch**

15.00–16.20 **Modeling and Simulation** (Hall 1) *Chair: Herwig Friedl*

1. **The distribution of time delay concerning breakdown point**
   Biljana C. Popović, Vidosav Lj. Marković, Aleksandar P. Jovanović and Milica M. Skamagkoulis

2. **Comparing partitions**
   Marjan Cugmas and Anuška Ferligoj

3. **Discrete kernels and their application to sport matches results**
   Blanka Sediva

4. **Model selection and model averaging on multiply-imputed data sets in a child growth study: a comparison**
   Khuneswari Gopal Pillay, John H. McColl, Charlotte Wright and The Gateshead Millenium Study Core Team

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

1. **Historical time series of Gross Domestic Product**
   Jaroslav Sixta and Martina Simkova

2. **Regional input-output analysis**
   Kristyna Vltavska and Jaroslav Sixta

3. **How to estimate utilities in Health Economics**
   Günal Bilek

4. **Determining factors that affects housing prices in Turkey by Bayesian network**
   Tuba Koç, Mehmet Ali Cengiz, Haydar Koç and Efehan Ulaş

16.20–16.40 **Break**

16.40–17.40 **Measurement and Design of Experiments** (Hall 1) *Chair: Katarina Košmelj*

1. **The journalistic dimension of statistics. Reporting human development and its quantification.**
   Alessandro Martinisi

2. **D-optimality of experimental designs**
   Bronislaw Ceranka and Małgorzata Graczyk

3. **Pilot study: current experiences with mixed mode including web mode in opinion survey (official statistics)**
   Marta Arnež, Mateja Zgonec and Nino Zajc

4. **Effect of supporting variables on imputation of missing data**
   Yucel Tandogdu

16.40–17.40 **Statistical Applications** (Hall 2) *Chair: Jože Rovan*

1. **Clustering analysis of the European forest sector production**
   Michael D. Burnard, Monika Cerinšek, Andreja Kutnar and Boris Horvat

2. **Fuzzy knowledge representation in Bayesian networks**
   Duygu İçen and Derya Ersel

3. **Design and application of phase-II joint monitoring of location and scale based on percentile modified rank scores**
   Amitava Mukherjee and Rudra Sen

4. **Microsimulation as a tool for the evaluation of personal taxation and social security in Finland**
   Antti Liski and Erkki Liski

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Mihael Perman*

1. **Networks: between measures and data**
   Ernst C. Wit

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics** (Hall 1) *Chair: Ernst C. Wit*

1. **Never fit Sequence - The design and analysis of multi-period clinical trials**
   Hans-Ulrich P Hockey
2. **Comparison of selection methods of genotypes based on non-replicated breeding experiments**
   Stanislaw F. Mejza and Iwona Mejza
3. **Reliability of measurements with continuous outcomes applied to heart rate variability parameters**
   Nataša Kejžar, Breda Podjaveršek and Fajko F. Bajrović
4. **Comparing net survival distributions with the log-rank type test**
   Klemen Pavlič and Maja Pohar Perme

11.40–12.00 **Break**

12.00–13.20 **Biostatistics and Bioinformatics** (Hall 1) *Chair: Maja Pohar Perme*

1. **Statistical approach for the development, prediction, and validation of a simple risk score: application to a neurocritical care study**
   Jay Mandrekar
2. **Challenges in accurate prediction of rare events with penalized likelihood methods**
   Georg Heinze, Angelika Geroldinger, Rok Blagus and Lara Lusa
3. **Accurate prediction of rare events with Firth's penalized likelihood approach**
   Angelika Geroldinger, Daniela Dunkler, Rainer Puhr, Rok Blagus, Lara Lusa and Georg Heinze
4. **Penalized logistic regression with rare events: preliminary results**
   Lara Lusa, Rok Blagus, Angelika Geroldinger and Georg Heinze

12.00–13.20 **Statistical Applications** (Hall 2) *Chair: Stanislaw F. Mejza*

1. **The use of R in Official Statistics**
   Jerneja Pikelj
2. **High-quality low-risk Public Use Files: an empirical evaluation of open-source anonymization software**
   Sebastian Kočar
3. **Determination of the profile of social media usage habits by the Latent Class Analysis**
   Haydar Koç, Mehmet Ali Cengiz, Tuba Koç and Efehan Ulaş

4. **Higher order moments of order statistics from Lindley distribution and associated inference**
   Khalaf S. Sultan and W.S. AL-Thubyani

13.20–14.30 **Lunch**

14.30 **Excursion**

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

1. **A few notes on centrality and flow**
   Steve Borgatti

10.00–10.20 **Break**

10.20–11.40 **Network Analysis** (Hall 1) *Chair: Steve Borgatti*

1. **Use of SAOM for modelling of network structure stability**
   Luka Kronegger, Marjan Cugmas and Anuška Ferligoj

2. **Multilevel blockmodeling**
   Aleš Žiberna

3. **Using semantics to recommend collaboration across domains in biomedicine**
   Dimitar Hristovski, Andrej Kastrin and Thomas C. Rindflesch

4. **Methodological challenges of measuring and analysing the Slovenian Twitter community**
   Ana Slavec and Vladimir Batagelj

10.20–11.40 **Statistical Applications** (Hall 2) *Chair: Lara Lusa*

1. **Can clustering based on predictors' dynamics improve the accuracy of financial distress classifier?**
   Lukas Sobisek and Maria Stachova

2. **Trend component estimation**
   Tomáš Toupal, Patrice Marek and František Vávra

3. **GraphGibbs algorithm - A Gibbs sampling method for motif finding in DNA with initial graph representation of sequences**
   Živa Stepančič

4. **Analysis of quantal models**
   Emel Cankaya, Nick Fieller and Mutlu Kaya

11.40–12.00 **Break**

12.00–13.20 **Economics and Econometrics** (Hall 1) *Chair: Aleš Žiberna*

1. **A simple randomization test of spatial correlation in the presence of common factors**
   Giovanni Millo

2. **Long-run regional equilibria in a large motor insurance market**
   Giovanni Millo and Antonio Salera

3. **Does government fiscal spending impact the quality of the environment?**
   Žiga Kotnik and Damjan Škulj

4. **Estimating technical and scale efficiencies in airline industry: a non-parametric DEA approach**
   Burak Keskin, Efehan Ulaş and Enes Filiz

12.00–13.00 **Social Science Methodology** (Hall 2) *Chair: Luka Kronegger*

1. **Testing hypotheses on crime seriousness perceptions**
   Nace Čebulj and Matevž Bren

2. **Sample size determination for testing two group variances of non-inferiority/superiority and equivalence with cost considerations**
   Wei-ming Luh and Jiin-huarng Guo

3. **The importance of balanced study designs in comparative studies**
   Ana Kolar

13.20–13.30 **Closing of the Conference** (Hall 1) *Chair: Andrej Blejec*

ABSTRACTS

Workshop

Research Data Management planning: problems and solutions

Irena Vipavc Brvar and Sonja Bezjak

Slovenian Social Science Data Archives, University of Ljubljana, Ljubljana, Slovenia
Irena.Vipavc@fdv.uni-lj.si, sonja.bezjak@fdv.uni-lj.si

Funding bodies increasingly require data, which were created in publicly funded research, to be open to the fullest possible extent. Furthermore authors are required by scientific journals to make data, which underpin publications promptly available to the editors at the time of submission and later on to readers without undue qualifications.
To provide access to high quality data to a wider research community, it is necessary to start proper research data management planning early, which enables avoiding significant problems in the data creation and curation phases.
Topics covered at the workshop address basic questions related to Research Data Management for open data, which include preparing a Research Data Management (RDM) plan, licensing data and intellectual property, metadata and contextual description (documentation), ethical and legal aspects of sharing sensitive or confidential data, anonymizing research data for reuse, data archiving and long-term preservation, and data security and storage.
Workshop is a combination of lectures and hands-on: with the emphasis on research data management on-line tool.

Invited Lecture

Dynamic treatment regimes as a prediction problem

Elja Arjas

University of Helsinki, University of Oslo, Finland
elja.arjas@helsinki.fi

Adequate medical treatment of many diseases, including different types of cancer, involves a sequence of treatment assignments over time.
For an optimal allocation, each assignment of a new treatment should be allowed to depend adaptively on how the patient responded to the ones that were administered previously. The task of establishing a close-to-optimal dynamic treatment regime of this type for patients with different individual characteristics, based on the accrued follow-up information, offers several challenges which are both conceptual and technical. The purpose of this talk is to consider their solution from a Bayesian perspective, by combining tools from constrained nonparametric modeling of stochastic processes and their inference, and the consequent predictive distributions, with dynamic programming for establishing the optimum. While the ideas underlying this approach are general, practical considerations will often restrict their direct applicability, and require that certain simplifying assumptions in the modeling are employed. An example based on HIV data is considered as an illustration.

Statistical Modeling

Potential parenthood and career progression of men and women - a simultaneous hazards approach

Martin Biewen[1] *and Stefanie Seifert*[2]

[1]University of Tuebingen, IZA Bonn, Tuebingen, Germany
[2]University of Tuebingen, Tuebingen, Germany
martin.biewen@uni-tuebingen.de, stefanie.seifert@uni-tuebingen.de

We analyze individual career courses of men and women in Germany. Our particular focus is on the association between career transitions and individual fertility. In contrast to most of the literature we focus on potential rather than realized fertility. Measuring career status in terms of the number of subordinates directly supervised by a given person, we focus on career transitions of men and women before the birth of a first child. Our analysis is based on the ALWA data set from the Institute for Employment Research, Nuremberg. It contains the life histories of more than 10,400 individuals. To model the parenthood hazard simultaneously with the career transition hazards we estimate a multivariate proportional hazard model with competing risks, in which the parenthood hazard enters as an explanatory variable in the career transition equations. As explanatory variables we consider socioeconomic variables as well as regional information and effects of duration dependence and lagged duration dependence. In order to address the aspect of unobserved individual characteristics, we also consider a fixed effects panel data model. Our results suggest that the effects on the timing of first birth are relatively similar for women and men, but gender differences are prevalent in the estimated coefficients of the career hazard equations. For our main variable of interest, the results suggest that the parenthood hazard is significantly negatively related to women's horizontal job mobility. For men, the results suggest a significant positive association of the parenthood hazard with upward transitions and insignificant effects for downward and horizontal transitions. These results persist if we allow for a correlation of parenthood hazards with unobserved individual characteristics such as time- or spell-constant personal preferences. In one specification, we also find a significant (but smaller) negative effect on female upward career mobility.

Analysing survival of groups that outlive the population – the French Olympians study

Maja Pohar Perme[1], *Juliana Antero-Jacquemin*[2], *Aurelien Latouche*[3] *and Jean-François Toussaint*[2]

[1]IBMI, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia

[2]Institut de Recherche bioMédicale et d'Epidemiologie du Sport, INSEP and Université Paris Descartes, Sorbonne Paris Cité, Paris, France

[3]Conservatoire National des Arts et Métiers, Paris, France

maja.pohar@mf.uni-lj.si

When comparing the survival experience of a group of patients to the general population, we usually assume that their hazard due to the disease is added to the hazard they have as members of the general population. However, this may not always be a realistic assumption, on the contrary, we may deal with subgroups that survive better than the general population. In this work, we investigate the properties of several approaches and comment on the interpretation of their results. The motivation for this presentation is the study of survival of French olympic athletes, which is based on a database that contains all athletes who competed from the year 1912 onwards and includes their survival status, cause of death information and several other variables of interest. In addition to exploring the data using standardized mortality rates we investigate two approaches. First, we analyse the data using the transformation approach which has been proposed in the relative survival field and transforms each individual's time with respect to the population mortality distribution of his/her population counterparts. Second, we look at the years lost or saved compared to the general population and the decomposition of this measure with respect to the causes of death. We also explore the usage of pseudo-values in fitting a regression model to this outcome. All results are illustrated with the French Olympians data.

Combining cross-sectional and panel information to disentangle attrition models in panels

Ulrich Pötter and Holger Quellenberg

German Youth Institute, München, Germany
poetter@dji.de, quellenberg@dji.de

Cross-sectional samples are often combined with panel data to provide simultaneously both timely data for social monitoring and detailed information on individual life course trajectories. Additionally, cross-sectional samples can be used as a refreshment for an otherwise dwindling panel population. Apart from these obvious uses, a rather less explored advantage of combined panel and cross-sectional samples lies in its potential to disentangle certain otherwise unidentified forms of panel attrition. Hirano et al. (Econometrica, 2001, 69, 1645-1659) were the first to explore the extent to which combined cross-sectional and panel data help in identifying features of attrition models. Their analysis was based on restrictions on the support and functional form of certain (conditional) distributions. While such an approach is appropriate to theoretically gauge the identifying power of combined cross-sectional and panel data, we will present a bounds approach based on recent developments along the lines of Manski's concept of partially identified models. Our approach is more appropriate to answer practically relevant questions in a constructive way, providing (approximations to) ranges of parameters compatible with certain assumptions on attrition processes.
To check the feasibility of our method, we use a large random sample of German children aged 5 to 8 in 2013 that was conducted as a refreshment sample to panel data on children aged 0-3 in 2009 in the AID:A study of the German Youth Institute. The rather unique feature of this data base is that the sampling frame for the 2013 refreshment sample was identical to the frame used in 2009. Thus, empirical results are not distorted by possible changes in sample frames that must otherwise be discounted.

How to handle underreporting utilizing statistical models

Herwig Friedl

Graz University of Technology, Graz, Austria
hfriedl@tugraz.at

Counts which are based on data from miscellaneous registers are often observed to be too small. Therefore, it is important to have a good estimate of the actual number of cases available. One way is, to model these counts as binomial variables, where both parameters p and n are considered to be unknown. It is known that especially the estimation of n can be very problematic even in the case of a random sample, especially if the empirical mean/variance ratio is smaller than one. Various methods for stabilization - including the use of an alternative beta-binomial model - were proposed to circumvent this problem. We recommend the use of a regression model for n under both, the binomial and beta-binomial model, and calculate the maximum-likelihood estimator of the parameter n.
If you apply this approach on the data from the Austrian crimes register, this will provide useful estimates of actually unknown total counts of crimes, the so-called dark figure. In this connection, p is the probability that such a crime has been reported, and n is its actual incidence. There are similar situations also in the field of epidemiology.

Measurement

Visibility graph analysis of seismicity: application to real and synthetic sequences

Luciano Telesca[1], *Michele Lovallo*[2], *Alejandro Ramirez-Rojas*[3] *and Tony A. Stabile*[4]

[1]Institute of Methodologies for Environmental Analysis, National Research Council, Tito, Italy

[2]ARPAB, Potenza, Italy

[3]Universidad Autonoma Metropolitana, Mexico City, Mexico

[4]Institute of Methodologies for Environmental Analysis, national Research Council, Tito, Italy

luciano.telesca@imaa.cnr.it

The visibility graph (VG) method maps time series into networks allowing the investigation of time dynamics of complex systems focusing on their topological properties. By means of such mapping, the dynamical properties of time series are converted in topological properties of networks; and, vice versa, information about time series can also be deduced analyzing the characteristics of networks. In this paper, the VG method is applied to several sets of seismic sequences to explore the topological features in their time dynamics. Both observational and laboratory-produced seismic sequences are investigated. A relationship between the classical seismological parameters (like the b-value of the Gutenberg-Richter law) and the VG topological parameters is found.

The present study was supported by the Bilateral Project Italy-Mexico "Experimental Stick-slip models of tectonic faults: innovative statistical approaches applied to synthetic seismic sequences", jointly funded by MAECI (Italy) and AMEXCID (Mexico) in the framework of the Bilateral Agreement for Scientific and Technological Cooperation PE 2014-2016.

Agreement among physicians assigning ICF categories based on medical documentation in a prosthetics and orthotics outpatient clinic

Gaj Vidmar[1], *Helena Burger*[1], *Barbara Kavčič*[1], *Nataša Bizovičar*[1], *Pavel Ptyushkin*[2] *and Bernard Francq*[3]

[1]University Rehabilitation Institute, Ljubljana, Slovenia

[2]Swiss Paraplegic Research, Nottwil, Switzerland

[3]Institut de Statistique, Biostatistique et Sciences Actuarielles, UCL, Louvain-la-Neuve, Belgium

gaj.vidmar@ir-rs.si, helena.burger@ir-rs.si, narocanje.dolenjske@terme-krka.si, natasa.bizovicar@ir-rs.si, pavel.ptyushkin@yandex.ru, bernard.g.francq@uclouvain.be

We wanted to verify the applicability of the International Classification of Functioning, Disability and Health (ICF, endorsed by the World Health Organisation) in the field of outpatient rehabilitation, more specifically in the field of prosthetics and orthotics (P&O), by studying inter-coder agreement. Four physicians (two trained and experienced in the use of the ICF, and two only familiar with it) examined the medical documentation of 35 randomly selected patients who had visited a P&O outpatient clinic within a two-month period, in order to apply second-level ICF codes. A 42-item initial list compiled based on the ICF Annex 9 was used, but the physicians were allowed to use additional codes. The codes were assigned without using the qualifiers – the three possibilities were "not applicable" (covering qualifiers 8 and 9), "no impairment" (qualifier 0) and "impairment" (qualifiers 1-4). For environmental factors (chapter e), "barrier" was used instead of "impairment" and "facilitator" could be used in addition. Agreement was assessed using raw agreement and Gwet's AC1 coefficients (unweighted and weighted, for two and multiple raters), including bootstrap-based comparison of dependent coefficients. Codes from all the four ICF chapters (b, s, d, e) were applied by all physicians to all the patients (12 codes per patient on average). There were five codes added by all the four physicians. On average, the physicians applied the largest number of categories from the Activities and Participation (d) chapter. As hypothesised, agreement was the highest among the two ICF-experienced physicians and the lowest among the two less ICF-experienced ones. Overall, the agreement was very high. This supports feasibility of routine ICF use in the tested setting, and our results will be useful for completing the development of the ICF Core Set for P&O.

Weighting methodology for a composite indicator of well-being

Jože Rovan[1], *Kaja Malešič*[2] *and Lea Bregar*[1]

[1]University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia
[2]Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
joze.rovan@ef.uni-lj.si, kaja.malesic@gov.si,
lea.bregar@ef.uni-lj.si

A composite indicator measures multi-dimensional concepts (e.g. competitiveness, e-trade or environmental quality) which cannot be captured by a single indicator. Ideally, a composite indicator should be based on a theoretical framework, which allows individual indicators to be selected, combined and weighted in a manner which reflects the dimensions or structure of the phenomena being measured. Different weighting methodologies have been proposed in the literature for the aggregation of the observed variables and dimensions into composite indicators. Ideal weighting should be transparent, should objectively capture societal valuations, and should produce comparable indicator values across units (Sharpe & Andrew, An Assessment of Weighting Methodologies for Composite Indicators: The Case of the Index of Economic Well-being, CSLS Research Report, 2012). In this paper we focus on the methodology used for the construction of a composite indicator of well-being for Slovenian municipalities. We have defined the concept of well-being by 12 domains for which we have selected the most suitable indicators respecting the underlying concept of well-being. In our case we have chosen the principal components analysis to aggregate basic indicators into a composite indicator of well-being. The advantage of applying a statistical approach is that weights are based on patterns in the data; therefore, no constructor bias in the assignment of weights can be present. However, there are several algorithms for the assignment of weights in case of the principal components analysis. We have applied a methodology of weighting similar to the approach originally proposed by Nicoletti, Scarpetta & Boyland for Summary Indicators of Product Market Regulation. The composite indicator is calculated from weighted basic indicators, while the weights for individual indicators are formed in a way that they are coherent with their impact on the principal component and its variance. The advantages of the proposed methodology are that it eliminates the limitations of negative correlations of basic indicators with principal components and that it allows calculation of partial composite indicators for each domain of well-being. The results based on indicators for 2011 show prevailing higher level of well-being in the western municipalities, while lower well-being is featured in the east of Slovenia.

Outcome-based measures of the Rasch model fit

Gregor Sočan

University of Ljubljana, Dept. of psychology, Ljubljana, Slovenia
gregor.socan@ff.uni-lj.si

The Rasch model is a popular measurement model in educational, health, and behavioural sciences. Since it is a latent trait model, the evaluation of the model fit is a crucial element of the process of constructing measures. Several techniques for the goodness-of-fit evaluation are available, which are based on general statistical principles of model fit evaluation. Unfortunately, these statistics do not have a clear practical interpretation, and the proposed cut-offs often lack a clear rationale. On the other hand, we argue that applied measurement disciplines would benefit from fit measures directly reflecting the impact of misfit on the quality of measurement outcomes. We shall present such a descriptive goodness-of-fit indicator based on simulation and resampling. The proposed approach will be demonstrated using the data from PISA 2012, which have recently been in the focus of a heated debate on the practical effects of model misfit.

Statistical Applications

Evaluation of phenotypic variation of wheat hybrids for resistance to Fusarium head blight

Maria Surma[1], *Tadeusz Adamski*[1], *Halina Wiśniewska*[1], *Karolina Krystkowiak*[1], *Anetta Kuczyńska*[1], *Zygmunt Kaczmarek*[1] *and Stanislaw F. Mejza*[2]

[1]Institute of Plant Genetics Polish Academy of Sciences, Poznan, Poland

[2]Poznań University of Life Sciences, Poznan, Poland

msur@igr.poznan.pl, tada@igr.poznan.pl, hwis@igr.poznan.pl, kkry@igr.poznan.pl , akuc@igr.poznan.pl, zkac@igr.poznan.pl, smejza@up.poznan.pl

Fusarium head blight (FHB) is a disease of small grain cereals and is caused by some Fusarium species, among others by F. culmorum, F. graminearum and F. avenaceum. The pathogenes affect spikes and kernels that results in reduction of yield and its quality.Wheat cultivars grown in Europe are mostly susceptible to FHB. Selection of wheat genotypes with improved resistance to FHB should be started at the beginning of breeding process, when cross combinations are selected for developing new cultivars. The present study was conducted to evaluate variation in the FHB resistance of wheat hybrids derived from crosses between winter wheat cultivars of various origin and different susceptibility to FHB. Experiment covering several cross combinations was conducted under field conditions in the complete randomized block design. Plants were inoculated with a mixture of conidial suspension containing F. culmorum, F. graminearum and F. avenaceum isolates. Inoculations were performed individually on each plot at the beginning of anthesis, and repeated about 3 days later at full anthesis. After inoculation micro-irrigation during 2 days was used. After harvesting, disease symptoms on kernels was observed and expressed as percentage of Fusarium damaged kernels (FDK) in kernel samples. Additionally, kernel weight per spike and thousand kernel weight were evaluated in inoculated and control plants. The data were statistically processed using uni- and multivariate analyses including analysis of canonical variables, Mahalanobis distance and estimation of contrasts between crosses regarding individual and the complex of analysed traits. Results of statistical elaboration permitted us to select cross combinations the most promising for breeding wheat genotypes more resistant to FHB. Acknowledgments. The study was supported by the National Centre for Research and Development, project No. PBS2/B8/10/2013.

Multivariate approach for selection of SSD lines in grain legumes

Tadeusz Adamski[1], *Maria Surma*[1], *Zygmunt Kaczmarek*[1], *Wojciech Świecicki*[1], *Paweł Barzyk*[2], *Anetta Kuczyńska*[1] *and Karolina Krystkowiak*[1]

[1]Institute of Plant Genetics Polish Academy of Sciences, Poznan, Poland

[2]Poznań Plant Breeders Ltd., Tulce, Poland

tada@igr.poznan.pl, msur@igr.poznan.pl, zkac@igr.poznan.pl, wswi@igr.poznan.pl, pawel.barzyk@phr.pl, akuc@igr.poznan.pl, kkry@igr.poznan.pl

Grain legumes, especially pea and lupins, are important crops in agriculture of Central and North Europe. Because of a high level of protein in their seeds these species constitute an alternative for soya imported from South America. Creation in a short time new cultivars of high yielding and early flowering is a main goal of breeding. In the paper we present preliminary characterization of pea, yellow and leaf-narrowed lupin lines regarding earlinesss, duration of vegetation period and seed yield. Materials for the study were lines attained by single seed descent (SSD) technique connected with in vitro culture of embryos. This technique permitted us to obtain eight generations in pea and six generations in lupins during 3 years. Totally, 160 lines of pea derived from eight cross combinations and 40 lines of yellow and leaf-narrowed lupins derived from two crosses were examined in field experiments conducted in the complete randomized block design with two replications, in which flowering time, date of maturity and seed yield were observed. The data were processed by multivariate analysis of variance and related methods, including analysis of canonical variables and estimation of the Mahalanobis distance. To assess biological progress, the contrasts between SSD lines and parental varieties were estimated for individual traits and all traits treated simultaneously. Based on the results of statistical analyses we select twelve out of 160 SSD lines of pea and three out of 40 SSD lines of lupins, which were characterized by short period of vegetation and relatively high seed yield. Acknowledgements. The study was supported by NATIONAL, MULTI-YEAR PROGRAM 2011-2015 “Improvement of domestic sources of plant protein, their production, economy and feeding technologies”, funded by the Polish Goverment and Polish Ministry of Agriculture and Rural Development.

X bar control chart for non-normal symmetric distributions

Kristina Veljkovic

Faculty of Mathematics, University of Belgrade, Belgrade, Serbia

kristina@matf.bg.ac.rs

In statistical quality control, X bar control chart is extensively used to monitor a change in the process mean. In this paper, X bar control chart for non-normal symmetric distributions is proposed. For chosen Student, Laplace, logistic and uniform distributions of quality characteristic, we calculated theoretical distribution of standardized sample mean and fitted Pearson type II or type VII distributions. Width of control limits and power of X bar control chart were established, giving evidence of goodness of fit of corresponding Pearson distribution to the theoretical distribution of standardized sample mean. For implementation of X bar control chart in practice, numerical example of construction of proposed chart is given.

Control charts in anticoagulant treatment

Susana Martins[1], *Lino Costa*[2] *and Pedro Oliveira*[3]

[1]Algoritmi R& D Centre University of Minho, Braga, Portugal

[2]Algoritmi R& D Centre, Department of Production and Systems University of Minho, Braga, Portugal

[3]EPIUnit, ICBAS, Universidade do Porto, Porto, Portugal

maleafar@gmail.com, lac@dps.uminho.pt, pnoliveira@icbas.up.pt

Atrial fibrillation is characterized by an abnormal functioning of the heart, which can result on accumulation of blood on headset and it will form a clot. This clot can move to other organs and limbs and it can trigger other medical problems. Oral anticoagulant treatment using warfarin is the most used treatment, because through the ingestion of medication the blood will become more fluid and consequently it reduces the possibility of blood clots, avoiding the problems referred to above. A patient who undertakes this treatment should be periodically monitored to ensure maximum effectiveness of the drug and ensure that it is under control. Therefore, control charts are important tools to assess whether or not the coagulation is controlled. The control charts are able to detect possible changes in processes of several areas, particularly in healthcare services, where the application of control charts has increased. There are several types of charts: Shewhart, EWMA and CUSUM. Shewhart charts are traditionally used to control the average of individual observations. Shewhart control charts are simple and very easy to apply and interpret for monitoring the process in real time allowing the rapid detection of changes in the process. However, some authors argue that there are better charts to detect process variations. EWMA and CUSUM charts are an alternative to Shewhart because they are more sensitive to small variations and use the information contained in all sampling, therefore they are more appropriate to study individual observations. The goal of this study is to use control charts to monitor anticoagulant treatment and choose the best chart to help this monitoring. In this work, Shewhart, EWMA, and CUSUM are applied to case studies with data collected from a public hospital. The advantages and drawbacks of each type of control chart are discussed.

Modeling and Simulation

The distribution of time delay concerning breakdown point

Biljana C. Popović[1], *Vidosav Lj. Marković*[2], *Aleksandar P. Jovanović*[2] *and Milica M. Skamagkoulis*[3]

[1]Department of Mathematics, Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
[2]Department of Physics, Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
[3]Department of Mathematics, Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
bipop@pmf.ni.ac.rs, vidosav@pmf.ni.ac.rs, alexandar.bmf@gmail.com, milicap78@yahoo.com

We characterize and investigate the random time delay of electrical breakdown in gases. It is well known that the time delay can be treated as the sum of two random variables: the statistical time delay and the formative time delay. Usually, the last two random variables are considered as independent random variables. But, the fact is that they are dependent. In this paper, we discuss the distribution of the sum of these random variables and characterize the distribution of time delay in general.

Comparing partitions

Marjan Cugmas and Anuška Ferligoj

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

The Rand Index (Rand 1971) is one of the most commonly used indices for measuring the stability of two partitions of one set of units. In some cases, there is a need to compare two partitions obtained on two sets of units, where one set is a subset of another set of units. The merging and splitting of clusters can have different impacts on the value of the indices when comparing two partitions. Therefore, we propose different modified Rand Indices. In addition, we also suggest the adjustment for chance for all proposed indices, which can be obtained by simulations. Some examples of comparing partitions obtained on different sets of units is also given.

Discrete kernels and their application to sport matches results

Blanka Sediva

University of West Bohemia, Pilsen, The Czech Republic
sediva@kma.zcu.cz

This paper deals with estimation of probabilities for cardinal variables. The presented methods are based on discrete kernels and our methods are inspired by kernel approach for kernel density estimation of continuous random variables. A smoothing problem is basic for kernel density estimation, but there are problems with interpretation of the term ”smoothing” for cardinal variables. Entropy and position of estimate between uniform distribution and empirical relative frequencies for quantification of smoothness are used in this paper. The theoretical results are applied to estimates of probabilities of the sport matches results.

Model selection and model averaging on multiply-imputed data sets in a child growth study: a comparison

Khuneswari Gopal Pillay[1], *John H. McColl*[1], *Charlotte Wright*[1] *and The Gateshead Millenium Study Core Team*[2]

[1]University of Glasgow, Glasgow, United Kingdom
[2]Newcastle University, Newcastle, United Kingdom
khuneswari@gmail.com, John.McColl@glasgow.ac.uk, Charlotte.Wright@glasgow.ac.uk,

There are no agreed guidelines for how best to carry out model selection and model averaging with multiply-imputed datasets. The main objective is to compare the performance of model selection and model averaging in the context of a UK longitudinal study (the Gateshead Millennium Study) of 1029 children, where the primary purpose of the analysis is to predict the future weight of individual children. The response is weight measured at 4-5 years, school entry level (30% missing). The covariates are weights measured at five time points from birth to one year (17% missing on average). One auxiliary variable, gestational age at birth, is also available. The performance of three strategies (inclusive, restrictive and non-overlapping) for building imputation and prediction models is discussed. Imputation was carried out using chained equations (via the “norm” method in the R package MICE). Possible combinations of linear multiple regression models for prediction were fitted and the best model was chosen using the model selection criterion AIC_c (corrected AIC). STACK method was used for model selection with multiply-imputed datasets. Non-Bayesian model averaging was explored to average the estimates of multiply-imputed datasets using AIC_c based weights. The performance of STACK and model averaging was compared using mean square error of prediction (MSE (P)) in a 10% cross-validation test. STACK provided better prediction of response values than model averaging. The inclusive strategy for building imputation and prediction models was better than the restrictive and non-overlapping strategies in this study. The presence of highly correlated covariates and response is believed to have led to better prediction. It is concluded that STACK should be used with an inclusive model-building strategy when highly correlated covariates are available to make predictions in the presence of moderate amounts of missing data.

Statistical Applications

Historical time series of Gross Domestic Product

Jaroslav Sixta and Martina Simkova

University of Economics, Prague, Czech Republic
sixta@vse.cz, martina.simkova@vse.cz

Time series of main economic indicators are very much appreciated by skilled economists. Modern official statistics offers incredible amounts of data but with lots of challenges to the users. Data from national accounts are more and more used for analyses mainly due to the fixed methodological background and international comparability. Unfortunately, the key problem of current national accounts statistics is comparability in the sense of practical possibilities of statistical authorities. Complexity of the measurement of economy is preferred to the comparability and this is not very much in line the requirements of analysts and researchers who are usually conservative. Frequent changes of methodology and computation procedures are not very welcomed by the users. Last time it happened in 2014 when the implementation of ESA 2010 broke time series of statistical data. In 2012, we prepared the estimates of sources and uses of gross domestic product for the Czech Republic for the period 1970-1989. Our research data were prepared in ESA 1995 methodology with crucial respect to full compatibility with data published by the Czech Statistical Office from 1990 onwards. Implementation of ESA 2010 in September 2014 caused incomparability that we face. Therefore we updated our estimates and implemented in the time series 1970-1989 main methodological changes containing mainly expenditures on research and development and military expenditures. Our presentation offers comparison and impact on the users of statistical data.

Regional input-output analysis

Kristyna Vltavska and Jaroslav Sixta

University of Economics, Prague, Czech Republic
kristyna.vltavska@vse.cz, sixta@vse.cz

Statistical measurement of economy is currently mainly dependent of national accounts. The official statistics publishes input-output tables of the national economy only. However, regional input-output tables offer interesting and valuable data as well. Their compilation is nevertheless very complicated that is why there are not used widely. Our contribution demonstrates how regional input-output tables can be compiled for regions of the Czech Republic and how regional input-output analysis can be done. Our regional input-output tables are product-by-product tables with 82 rows and columns and they are prepared for all 14 regions (NUTS 3) of the Czech Republic. They are combined from officially published regional accounts and national symmetric input-output tables. The key issue of their construction lies in the definition of statistical unit and its decomposition between regions. Our contribution brings brief description of the methodology and the example of forecasting issues based regional input-output tables. The case of dependency between the regions is deeply discussed and main issues are tackled and explained.

How to estimate utilities in Health Economics

Günal Bilek

Bitlis Eren University, Bitlis, Turkey
gunal-34@hotmail.com

This study aims to provide information on health economics and show how utilities and choice probabilities of health states are calculated in Health Economics by using two approaches: the Maximum Likelihood Estimation and a Bayesian Approach.
There are two data sets to work on. The first data set called the TTO data which were derived from a $Asthma\ Quality\ of\ Life\ Questionnaire$ $(AQLQ)$ have a continuous dependent variable stating the utility values of subjects. A one-way error components random effects model was fitted to the TTO data:

$$U_{ij} = 1 - x_{ij}\beta + \delta_i + e_{ij}$$

where U_{ij} is the utility for health state ij measured by respondent i, x_{ij} is the vector of dummy variables representing the health state ij, β is the vector of the parameters to be estimated from the data, δ_i is the i-th random effect and e_{ij} is the usual error
The other data set which were derived from a $Discrete\ Choice\ Experiment$ (DCE) have a binary dependent variable showing subjects' preferences between two health states.

$$P_{ifirst} = 1 - \Phi\Big(\frac{x_{isecond} - x_{ifirst}}{\sqrt{2\sigma^2}}\Big)$$

where Φ represents the cumulative distribution function of the standard normal distribution, x_{ifirst} and $x_{isecond}$ are the vectors of the dummy variables of the first and second health states respectively that are given to subject i, P_{ifirst} is the probability that subject i chooses the first health state and $P_{isecond}$ is the probability that subject i chooses the second health state.
To analyse the two data sets together, the likelihoods of the two data sets were obtained and combined into a single likelihood function and maximised using optim command in R. As an alternative, a Bayesian approach was used and Markov chain Monte Carlo (MCMC) was used to obtain the parameters in WinBUGS.

Determining factors that affects housing prices in Turkey by Bayesian network

Tuba Koç[1], *Mehmet Ali Cengiz*[2], *Haydar Koç*[1] *and Efehan Ulaş*[1]

[1]Department of Statistics, Çankırı Karatekin University , Çankırı, Turkey

[2]Ondokuz May?s University department of statistics, Samsun, Turkey

tubakoc@karatekin.edu.tr, macengiz@omu.edu.tr, haydarkoc@karatekin.edu.tr, efehanulas@karatekin.edu.tr

Developments in the housing sector are the variables that are part of daily life. The need for housing is one of the basic needs for people which comes after the physiological needs. Therefore, housing is not only an economic value is also kind of entity that socio-psychological characteristics. Nowadays, both housing sales and prices are in increasing trend. There are many factors that affects housing prices such as housing type, square meter, location, heating type, building age, number of rooms, and so on. In this study, 50055 house that are selling in Turkey is considered. The impact of different properties of those houses on the house prices are determined by using Tree Augmented Naive Bayes which is used among Bayesian learning algorithms and prior information. Bayesian networks are graphical models that describe the conditional independence and causality relations through graphs.

Measurement and Design of Experiments

The journalistic dimension of statistics. Reporting human development and its quantification.

Alessandro Martinisi

University of Leeds, Leeds, United Kingdom
alessandro.martinisi@icloud.com

This paper argues that in addition to the traditional dimensions of official statistics: the political, the economical, the social and the environmental as pointed out by Michael Ward, there is another dimension no less important: the journalistic dimension, which plays a pivotal role in the "sense-making" and dissemination of statistical information to the general audience. The case of human development and its controversial quantification and, more broadly, the question of development of society, are of vital concern to human beings, and the issue of its rightful reporting is at the core of the journalistic practises worldwide. In this regard the transition of official data from statistical bodies to newsrooms cannot be overlooked, especially in the context of global journalism. In fact, various ideas like those of social justice, equality, poverty, human rights, tolerance etc. evolved or got emphasised at different times and places out of the attempts of different societies to develop themselves. In our times, for instance, it has become commonplace to look at the problem of collective human development from the global perspective. A perspective adopted indeed by both statisticians and journalists in their common quest for a conventional "truth". By highlighting that all the approaches to the problem of human development suffer from the drawback that they are not based on any general framework of reference which leads to a comprehensive system of values and have some ad hoc element in it, this paper intends to describe meanings and boundaries of the journalistic dimension of statistics and at the same time to provoke a discussion about the possibility of a "mindful" approach in reporting official statistics.

D-optimality of experimental designs

Bronislaw Ceranka and Małgorzata Graczyk

Poznań University of Life Sciences, Poznań, Poland
bronicer@up.poznan.pl, magra@up.poznan.pl

In the paper, we consider the problems related to the spring balance weighing designs. There are designs of experiments, in that the result of experiment we can describe as linear combination of unknown measurements of objects with factors of this combination equal to zero or one. In such designs, the assumptions connected with experimental errors play very important role. Here, we have been working under assumption that the errors are uncorrelated and they have the same variances. We consider spring balance weighing designs from the point of view of optimality criteria. We study D-optimal designs, i.e. the designs in that that the determinant of the inverse of the information matrix for the design is minimal. We give the lowest bound of the inverse of the information matrix for the design and necessary and sufficient conditions to this bound to be attained. Moreover, some issues linked with the problem of indicating the construction methods of D-optimal experimental design are presented. Here, the topic is focus on the determining the D-optimal spring balance weighing design on the base of the set of incidence matrices of known block designs: balanced incomplete block designs and partially balanced incomplete block designs.

Pilot study: current experiences with mixed mode including web mode in opinion survey (official statistics)

Marta Arnež , Mateja Zgonec and Nino Zajc

Statistical Office of the Republic of Slovenia (SURS), Ljubljana, Slovenia
Marta.Arnez@gov.si, Mateja.Zgonec@gov.si, nino.zajc@gov.si

In the fourth quarter of 2014 the Statistical Office of the Republic of Slovenia (SURS) conducted a pilot survey with the introduction of the web mode, as part of the mixed mode, in data collection - Consumer Survey, as a part of the internal SURS project.
The Consumer Survey is a monthly opinion survey, which is regularly conducted by Computer Assisted Telephone Interviewing (CATI) at SURS. In a short questionnaire, respondents are asked about their opinion and their expectation of household financial situation, general economic situation in the country, unemployment, savings and intentions regarding purchases.
The main goal of introducing the web mode is to test the possibility to improve coverage and decrease non-response. Since the regular survey gives a smaller share of younger and urban people than in the population, we expected that we will cover this part of the population by offering the web mode. Additionally, we need to adjust the data collection mechanism to the respondents (to better respond to respondents' needs by offering them various data collection modes).
The aim of the mixed mode pilot Consumer Survey was to test the data collection system using sequential mixed-mode data collection (WEB-CATI), procedures, technical solutions, analysing the collected data and examining the optimum data collection mode for the Consumer Survey. We were also interested in comparing results, for example, between the different modes. The mixed mode pilot Consumer Survey was conducted in the same period as the regular survey (parallel run).
The presentation will focus on differences between results of the regular survey and results of the mixed mode pilot survey.

Effect of supporting variables on imputation of missing data

Yucel Tandogdu

Department of Mathematics, Eastern Mediterranean University, Famagusta, Mersin 10, Turkey
yucel.tandogdu@emu.edu.tr

The problem of missing data, regardless of the reason behind it has been the focus of research, as how to impute the missing values. Many studies on imputation of missing data are based on determining the pattern of missing values and devising methodologies along those lines. In this study the imputation process is based on available values of the variable of interest or response variable, as well as available data from variables that are closely associated with the variable of interest, named as supporting variables. Level of correlation between the variable of interest and the supporting variable is important. Variables poorly correlated with the variable for which missing values are to be imputed should not be taken into account. Consider a n $\times$ p data set representing n measurements of a variable for p different time or space intervals, with values missing at random. Determine the variables related with the process involving the variable with the response variable. Establish the empirical relationship between the supporting variables and the variable with missing data. Conversion of the supporting variables to the same unit as the response variable will help attain homogeneity. In the next step, taking a column from the response variable and forming tuples with supporting variables, multivariate regression is used to estimate the missing values in the response variable. Process is repeated for each column, till all missing values are imputed. Imputation using this approach is found to produce satisfactory results, based on very low error levels achieved.

Statistical Applications

Clustering analysis of the European forest sector production

Michael D. Burnard[1], *Monika Cerinšek*[2], *Andreja Kutnar*[1] *and Boris Horvat*[2]

[1]Andrej Marušič Institute, University of Primorska, Koper, Slovenia

[2]Abelium R&D d.o.o., Ljubljana, Slovenia

michael.burnard@iam.upr.si, monika@abelium.eu, andreja.kutnar@upr.si, boris@abelium.eu

Wood is still a very important part of many products available on the market. In order to analyse Europe's forest sector production we performed an exploratory analysis of three data sets and compared hierarchical clusterings of European countries for each data set. For the first clustering we used detailed production quantities and values from 2008-2013 made available by the Food and Agriculture Organization of the United Nations. For the second clustering we used the production data that is derived from the Eurostat Prodcom database and covers approximately 4000 products, 205 of which are forest sector products. For the third clustering we used labour force data (from the Eurostat employment database) that includes employed persons and hours worked in forestry and logging, the manufacture of wood and wood products, the manufacture of paper and paper products, and furniture manufacturing.
This analysis provided an opportunity to identify possible areas of expansion for specific countries. The combination of labour force and production data provides an estimate of the relative efficiency of the forest products sector in European countries. The purpose of this analysis is also to prepare the data for monitoring of logistics in wood value chain.

Fuzzy knowledge representation in Bayesian networks

Duygu İçen and Derya Ersel

Hacettepe University, Ankara, Turkey

duyguicn@hacettepe.edu.tr, dtektas@hacettepe.edu.tr

Bayesian Networks (BNs) are graphical models representing joint probability distributions of a set of random variables and reasoning probabilistic relationships among these variables. They also represent knowledge about uncertain domain by using data and expert opinion. In some situations, it is hard to express knowledge in BNs because of the ambiguity depending on lack of data information and expressing expert knowledge. The use of fuzzy methods in BNs is a solution to these problems. The problem of lack of data information can be solved by using fuzzy probability theory to calculate probabilities in BNs. Besides, fuzzy linguistic expressions can be used to figure out the problem of expressing expert knowledge in BNs. In this study, we propose to define variables in BNs with fuzzy linguistic expressions and calculate marginal and conditional fuzzy probabilities with Buckley's confidence interval approach. This method provides to indicate uncertainty and represent knowledge better in BNs.

Design and application of phase-II joint monitoring of location and scale based on percentile modified rank scores

Amitava Mukherjee and Rudra Sen

XLRI- Xavier School of Management, Jamshedpur, India
amitmukh2@yahoo.co.in

this article we propose a class of Phase-II distribution-free (nonparametric) control charts for simultaneously monitoring the location and scale parameters of a univariate unknown continuous process distribution. The present work is the generalization of the popular Shewhart-Lepage chart for simultaneously monitoring of location and scale parameters and is based on the concept of percentile modifications of rank scores. Depending on various strategies of percentile modifications, we realize a class of percentile modified Shewhart- Lepage (PMSL) chart. We discuss implementation strategies and post signal follow-up procedures of the proposed class of charts. A Monte-Carlo study reveals that more often a chart has lower out-of-control (OOC) average run length (ARL). Moreover, percentile modification is proved to be more beneficial when there are possible bias in Phase-I sample. We illustrate the use of charts with a recent data on Vancouver city call centre service quality monitoring.

Microsimulation as a tool for the evaluation of personal taxation and social security in Finland

Antti Liski[1] *and Erkki Liski*[2]

[1]Statistics Finland, Helsinki, Finland

[2]University of Tampere, Tampere, Finland

antti.liski@stat.fi, erkki.liski@uta.fi

We consider microsimulation model as a tool to predict and analyze the budgetary effects of social and fiscal policies in a future period using data from a previous period that is adjusted to approximate the population in the future period. The model needs to produce a good approximation of taxes and benefits payable in a particular period. We have run our experiments using the microsimulation model SISU (2013) which was developed during 2011-2013 at Statistics Finland in co-operation with the Research Department of the Social Insurance Institution of Finland.

We estimate the population total of the study variables – taxes and benefits - for the population of Finland. In addition to the study variables, certain auxiliary variables are available. We assume the following auxiliary information: (i) The population totals of auxiliary variables, and (ii) the values of auxiliary variables are given for every sample unit. We use the calibration estimator of Deville and Särndal (1992) which is a weighted sum over the sample units. Reweighting is carried out by minimizing a given distance measure under the calibration constraints. We distinguish three types of inputs that can affect the study variable: the calibration, policy and model variables. Income taxation, for example, is a policy variable consisting of a set of rules concerning legislation on the taxation and social security. Variables like age, sex, employment status, education etc. are typical control variables.

We ask how we can do calibration estimation when the correct values of control variables are not known for sample units. This is the case when the effects of alternative policies in a future period are estimated. A crucial step is to identify control variables, important from a reweighting perspective, in being able to capture economic and demographic change over time.

Invited Lecture

Networks: between measures and data

Ernst C. Wit

Johann Bernoulli Institute, University of Groningen, Groningen, The Netherlands

e.c.wit@rug.nl

Most statististical models are defined as a probability measure on some observable outcome. Clearly, this definition is rarely helpful directly to analyze real data. In fact, modern data tend to be rather complex. For example, genomic data comes from large monitoring systems with no prior screening. Longitudinal psychiatric studies measure patients on a large number of symptoms. However, in most of these systems, these interactions are rather the structured and the actual set of relationships, therefore, tends to be sparse. A graph is one possible way to describe complex relationships between many actors, such as for example genes and psychiatric symptoms.
Graphical models present an appealing and insightful way to describe graph-based dependencies between the random variables. Although potentially still interesting, the main aim of inference is not the precise estimation of the parameters in the graphical model, but the underlying structure of the graph. Combining graphical models with exponential random graph models is an interesting new way to model the underlying topology of such non-observed graphs.

Biostatistics and Bioinformatics

Never fit Sequence - The design and analysis of multi-period clinical trials

Hans-Ulrich P Hockey

Biometrics Matters Ltd, Hamilton 3216, New Zealand
hans@biometricsmatters.com

Examples are given of the design and analysis of multi-period studies, typically cross-overs, in early phase studies in the pharmaceutical industry. The aim is to explore the usefulness of the inclusion of the between-subject sequence term typical in the linear models used for analysis and the effect this has on design. Starting with the simple two-period, two-treatment, two-sequence design, various multi-period studies are introduced and alternative designs and analyses considered. The usefulness for inference of different designs and analyses using models with and without sequence are contrasted. It is hoped to be able to show that omitting the sequence term from analysis models allows a broader range of designs, more useful models, and can give more information with the same experimental resources.

Comparison of selection methods of genotypes based on non-replicated breeding experiments

Stanislaw F. Mejza and Iwona Mejza

Poznan University of Life Sciences, Poznan, Poland
smejza@up.poznan.pl, imejza@up.poznan.pl

This presentation deals with selection problems in the early stages of a breeding program. During the improvement process, it is not possible to use an experimental design that satisfies the requirement of replicating the treatments, because of the large number of genotypes involved, the limited quantity of seed and the low availability of resources. Hence unreplicated designs are used. To control the real or potential heterogeneity of experimental units, control (check) plots are included in the trial. The selection of genotypes for further breeding is one of the important problems of this methodology. Hence, in the literature, there are many selection methods (discussed in the presentation) of using the information resulting from check plots. Each of them is appropriate for some specific structure of soil fertility. The problem here is that we do not know what kind of soil structure occurs in a given experiment. Hence we cannot say which of the existing methods is appropriate for a given experimental situation. It is impossible to compare analytically those methods. Hence we propose to compare known from a literature six selection methods with the method proposed by the authors. The last method (proposed by authors) is based on the theory of response surface. We compare the rankings of genotypes obtained by the all methods on the basis of the experiment with spring barley. We can see that the rankings are completely different. In the paper we recommend the selection methods based on the response surface methodology using check plots.

Reliability of measurements with continuous outcomes applied to heart rate variability parameters

Nataša Kejžar[1], *Breda Podjaveršek*[2] *and Fajko F. Bajrović*[1]

[1]University of Ljubljana, Faculty of Medicine, Ljubljana, Slovenia
[2]University Clinical Centre, Ljubljana, Slovenia
natasa.kejzar@mf.uni-lj.si, breda.podjaversek@kclj.si, fajko.bajrovic@mf.uni-lj.si

Evaluation (of quality) of measurements is important for assessing the measurement instruments and for how to deal further with data measured in this way. In medicine (sports and some other fields) heart rate variability (HRV) is quite extensively used for assessment of autonomic cardiovascular control.
HRV is known to be influenced by many factors (blood pressure, stress, breathing etc). The repeatability (reliability) of HRV is therefore limited. In the example of healthy adults, where HRV was measured for 40 consecutive minutes we explore/evaluate and describe the methodological and practical issues in reliability of HRV parameters. We further focus on the reproducibility of these parameters under the changing condition (i.e. focused thoughts during the second part of the measurements).

Comparing net survival distributions with the log-rank type test

Klemen Pavlič and Maja Pohar Perme

University of Ljubljana, Faculty of Medicine, Institute for Biostatistics and Medical Informatics, Ljubljana, Slovenia
klemen.pavlic@mf.uni-lj.si, maja.pohar@mf.uni-lj.si

In survival analysis, the log-rank test is the most commonly used test to compare survival distributions between groups. One of its basic assumptions is that the hazard function is homogeneous within each group, if this assumption is violated, stratified log-rank test may be used instead. In this work, we study the properties of the recently proposed log-rank type test in the framework of net survival. In particular, we are interested in its relation to the tests of regression coefficients and in the difference between the stratified and non-stratified version of the test. We study the properties of both versions with simulations, comment on their interpretation and present guidelines for their usage. We also present an R function which provides an efficient algorithm despite the computational intensity of the test. This function is included in the relsurv package for relative survival.

Biostatistics and Bioinformatics

Statistical approach for the development, prediction, and validation of a simple risk score: application to a neurocritical care study

Jay Mandrekar

Mayo Clinic, Rochester, United States of America
mandrekar.jay@mayo.edu

Patients admitted to neurocritical care units often have devastating neurologic conditions and are likely candidates for organ donation after cardiac death. Given the uncertainty in the time to death, improving the prediction of time to death after WLSM based on pre-WLSM clinical factors is crucial to have a positive impact on the rates of organ donation. In the first part of the presentation, we will discuss how we arrived at a pool of factors associated with earlier time to death using a retrospective database. Next, we will discuss the validation of these identified factors done via a multicenter prospective study using logistic regression and ROC analysis.

Challenges in accurate prediction of rare events with penalized likelihood methods

Georg Heinze[1], *Angelika Geroldinger*[1], *Rok Blagus*[2] *and Lara Lusa*[2]

[1]CeMSIIS, Medical University of Vienna, Vienna, Austria

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

georg.heinze@meduniwien.ac.at, angelika.geroldinger@meduniwien.ac.at, rok.blagus@mf.uni-lj.si, lara.lusa@mf.uni-lj.si

Logistic regression is one of the most commonly used statistical methods to estimate prognostic models that relate a binary outcome (with levels 'event' and 'non-event') to a number of binary, categorical or continuous explanatory variables. A low prevalence of events, encountered frequently in clinical or epidemiological studies, but also in other fields of empirical research, causes underestimation and instability of estimates of the event probability in subjects who are likely to experience the rare event. This happens because the analysis is heavily influenced by the subjects without events. This effect is even more pronounced when the number of explanatory variables approaches or exceeds the number of outcome events. Recently, penalized likelihood regression (PLR) methods have become popular for analyses with high-dimensional explanatory variable spaces. PLR methods shrink the estimates of regression coefficients towards zero in order to decrease their mean squared error. While this also decreases the overall mean squared error of predicted event probabilities, in the rare events situation (RES) poor predictions for the subjects which are at high risk for an event are still encountered. We define inner, middle and outer layers of model building with PLR, corresponding to estimation, tuning and validation of PLR models, respectively. We discuss some possible modifications of PLR in each of these layers to improve the accuracy of rare events prediction. These modifications are based on putting more weight on the subjects with events, modifying the type of penalty, using tuning criteria better adapting to the RES, and focusing on the RES when selecting validation criteria. These modifications will be exemplified on a real-world diagnostic study aiming at predicting bacteraemia by the values of laboratory parameters of blood samples.

Accurate prediction of rare events with Firth's penalized likelihood approach

Angelika Geroldinger[1], *Daniela Dunkler*[1], *Rainer Puhr*[2], *Rok Blagus*[3], *Lara Lusa*[3] *and Georg Heinze*[1]

[1]Center for Medical Statistics, Informatics and Intelligent Systems; Medical University of Vienna, Vienna, Austria

[2]The Kirby Institute for Infection and Immunity in Society, University of New South Wales , Kensington, Austria

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

angelika.geroldinger@meduniwien.ac.at, daniela.dunkler@meduniwien.ac.at, rpuhr@kirby.unsw.edu.au, rok.blagus@mf.uni-lj.si, lara.lusa@mf.uni-lj.si, georg.heinze@meduniwien.ac.at

David Firth's penalized likelihood approach (Firth, Biometrika, 1993) removes the first-order bias term of maximum likelihood estimates of regression coefficients. Instead of correcting bias after estimation, Firth's method prevents bias by penalizing the likelihood function utilizing the Jeffreys invariant prior.

In logistic regression, rare events (and/or a large number of explanatory variables) often result in the problem of separation, also termed 'monotone likelihood'. If separation occurs, one of the explanatory variables or a combination of them (almost) perfectly predicts the outcome variable. As a consequence, corresponding maximum likelihood estimates are infinite. Firth's penalization allows to compute reliable finite regression coefficients even in the situation of separation. As this penalization is also available in standard statistical software like SAS (simply add '/ firth' to the model statement of proc logistic) or R (eg. brglm or logistf package), Firth's penalization has become a popular approach for logistic regression analyses.

This presentation discusses advantages and limitations of Firth type penalization. We focus on accurate prediction of rare events with a particular interest in the situation of high-dimensional explanatory variable spaces. In these $p \gg n$ situations, Firth's penalization breaks down, resulting in instable estimates. However, there are penalized likelihood methods such as ridge regression or LASSO which can handle $p \gg n$ situations. Combining ridge regression or LASSO with the Firth type penalization might be a way to extend Firth's approach to the high-dimensional setting. For the low-dimensional setting a combination of ridge regression and Firth type penalization was already proposed by Shen and Gao (Journal of Data Science, 2008). We review this and other modifications of Firth's type penalty proposed in literature with regard to their applicability to the prediction of rare events.

Penalized logistic regression with rare events: preliminary results

Lara Lusa[1], *Rok Blagus*[1], *Angelika Geroldinger*[2] *and Georg Heinze*[2]

[1]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
[2]CeMSIIS, Medical University of Vienna, Vienna, Austria
lara.lusa@mf.uni-lj.si, rok.blagus@mf.uni-lj.si,
angelika.geroldinger@meduniwien.ac.at,
georg.heinze@meduniwien.ac.at

In binary class prediction problems the aim is to estimate the most likely class membership of the samples (event or non-event), rather than the probability for an event. Penalized logistic regression (PLR) models are becoming increasingly popular tools for the estimation of predictive models when the outcome is binary and many covariates are available; they are widely used when the number of variables is larger than the number of samples (high-dimensional data). PLR models are efficiently implemented in two R packages (glmnet and penalized), which include lasso (l1) and ridge (l2) penalties. In this presentation we focus on the use of PLR models for rare events and high-dimensional data. We present a selected series of simulations in which we show that PLR models are sensitive to rare events, producing poor predictions for the rare events and classifying most samples as non-events. We present some preliminary simulation results in which we explore the performance of some of our proposals aimed at reducing the rare event bias by using a marginal event rate of 0.5 in the training set. The explored methods include upsizing, downsizing, multiple downsizing and data reweighing. We also address the problem of tuning of the complexity parameter in the rare event situation and discuss the use of the two R packages in this context.

Statistical Applications

The use of R in Official Statistics

Jerneja Pikelj

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
jerneja.pikelj@gov.si

Without doubt R is one of the most used software for statistical computing and graphics. It includes a wide set of packages implementing various functions, specifically formed for solving problems that occur in different fields of statistics.
The paper has two practical purposes. The first one is to analyze how successfully R can be used for data analysis in surveys carried out by the Statistical Office of the Republic of Slovenia. To achieve this goal, we analyzed the data of the Monthly Statistical Survey on Earnings Paid by Legal Persons. The second purpose is to analyze how the assumption on the non-response mechanism that occurs in the sample impacts the estimated values of the unknown statistics in the survey. Depending on these assumptions, different approaches to adjust the problem caused by unit non-response are presented. The paper focuses on different stages of the statistical process, ranging from data importing and exporting into R, survey sampling, data manipulation with large datasets, weight calculation and aggregation to some visualization of the results at the end.

High-quality low-risk Public Use Files: an empirical evaluation of open-source anonymization software

Sebastian Kočar

Social Science Data Archives, Ljubljana, Slovenia
sebastian.kocar@fdv.uni-lj.si

Since there has been growing demand for Public Use Files microdata, data protection/anonymization field has had to responds to those requirements. Nowadays there are numerous methods and techniques, which are applied in data protection software. However, to produce high-quality low-risk Public Use Files microdata, the emphasis should not be placed on disclosure risk only, but on data utility as well. This paper discusses applicability of data protection software with various anonymization methods applied; from balanced sampling approaches to anonymization and synthetic microdata generation to data reduction and perturbation techniques such as global coding, local suppression, micro-aggregation etc. Evaluation is done by comparing the effect on both the disclosure risk and the information loss as an indicator of data quality and research applicability. Having in mind that social science data archives as services providing access to Public Use Files have limited funds, only open-source anonymization software is evaluated. Balanced sampling is executed using R software and its packages sdcMicro, bethel and sampling. The Cornell Anonymization Toolkit and μ-ARGUS are used to perform data reduction and perturbation, while R synthpop package generates synthetic data. The information loss is assessed using different methods, such as direct comparison, comparison of contingency tables, comparisons of mean square, mean absolute, and mean variation etc. At the end of the anonymization process, the disclosure risk (local, global) of final Public Use Files is measured. The empirical evaluation is done processing two original detailed databases, the first one from the the social science field (Slovenian Public Opinion Survey) and the second one from the official statistics field (Labour Force Survey). Finally, the results of the evaluation are presented from three different perspectives: disclosure risk, data utility and usability of certain open-source anonymization software.

Determination of the profile of social media usage habits by the Latent Class Analysis

Haydar Koç[1], *Mehmet Ali Cengiz*[2], *Tuba Koç*[1] *and Efehan Ulaş*[1]

[1]Department of Statistics, Çankırı Karatekin University, Çankırı, Turkey
[2]Department of Statistics, Ondokuz Mayis University, Samsun, Turkey
haydarkoc@karatekin.edu.tr, macengiz@omu.edu.tr,
tubakoc@karatekin.edu.tr, efehanulas@karatekin.edu.tr

Nowadays, social media has become irreplaceable for many people. This platform which consists of items that ease and accelerate our lives is a total that provides opportunity mutual sharing to their users and enables them to form media content personally or as a group and gathers up digital media and technologies. Although social media tools draw great interest generally by youth, they are used by people of all ages. In general people have made a habit of these tools because they consider them as a without tiring activity in their free time. In this study we investigated social media usage habit of people by latent class analysis. The data that we used in the study is a survey data which was obtained by face to face interview to 1065 people whose age interval is 16-67 and who live in Samsun, which is a city in the North of Turkey. In the study it was determined that the individuals who live in this city can be classed into three classes according to their social media usage activities.

Higher order moments of order statistics from Lindley distribution and associated inference

Khalaf S. Sultan and W.S. AL-Thubyani

King Saud University, Riyadh 11451, Saudi Arabia
ksultan@ksu.edu.sa , statistic14@hotmail.com

In this paper, we derive the exact explicit expressions for the single, double (product), triple and quadruple moments of order statistics from Lindley distribution. Then, we use these moments to obtain the best linear unbiased estimates (BLUEs) of the location and scale parameters based on Type-II right censored samples. Also, we use these results to determine the mean, variance, and coefficients of skewness and kurtosis of certain linear functions of order statistics to develop Edgeworth approximate confidence intervals of the location and scale Lindley parameters. In addition, we present some numerical illustrations through Monte Carlo simulations. Finally, we apply the findings of the paper to some real data set.

Invited Lecture

A few notes on centrality and flow

Steve Borgatti

University of Kentucky, United States of America
steve.borgatti@gmail.com

Centrality is often described as measuring the "importance" of a node in a network. Dozens of measures have been proposed in the literature, most of which can be characterized quite directly in terms of a node's involvement in the walk structure of a network (e.g., the average distance from the node to all other nodes; the number of walks of all lengths emanating from a node, weighted inversely by their length). Under a general rubric of social capital, measures of centrality are often used to predict node outcomes, such as career success. Unfortunately, almost none of these measures are explicitly derived from any kind of model of what is happening in the network and how these processes lead to the outcome state of each node. It is true that it has been shown that some centrality measures implicitly contain (or are consistent with) underlying models of network flows. For example, the formula for betweenness centrality yields precisely the expected values of the number of times that a token flowing through the network will reach a given node, given that the token takes only shortest paths, can only be in one place at one time, and chooses randomly between equally short paths. The problem is, this implicit model virtually never makes sense in the empirical settings where these measures are employed. For example, in the management field, we commonly study flows of information among members of an organization. Yet information does not travel only along shortest paths. Indeed, information does not try to get anywhere in particular, and no mind is guiding it, through intermediaries, to a specific target. It also does not so much "move" from node to node as "copy" – i.e., it can be in more than place at a time. There are several other models that are implicit in the set of extant centrality measures, but it is safe to say that they too are inappropriate for the vast majority of empirical settings social scientists wish to model. The objective of this paper is to explore new measures that are based on more plausible flow processes.

Network Analysis

Use of SAOM for modelling of network structure stability

Luka Kronegger, Marjan Cugmas and Anuška Ferligoj

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

In presented analysis we made a step closer towards combining two methodological approachies of social network analysis: well established method of blockmodeling based on works of Lorrain & White (1971) and Doreian et al. (2005), and relatively new approaches to stochastic actor based modeling of network dynamincs presented by Snijders (2001, 2005) and Snijders et al. (2007, 2010). Combination of two methods offers a great opportunity for analysis of influence of individual and group characteristics on the dynamics of emergent structure in the network. Presented analysis is performed on data of collaboration networks of Slovenian researchers measured in period from 1996-2010, sliced in to two consecutive 10-year time spans.

Multilevel blockmodeling

Aleš Žiberna

University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
ales.ziberna@fdv.uni-lj.si

In this work, different approaches to blockmodeling of multilevel network data will be presented. Multilevel network data consist of networks that are measured on at least two levels (e.g. between organizations and people) and information on ties between these levels (e.g. information on which people are members of which organizations). Several approaches will be considered: a) separate analysis of the levels; b) transforming all networks to one level and blockmodeling on this level using information from both/all levels; c) truly multilevel approach, where both/all levels and ties between them are modeled at the same time. Advantages and disadvantages of these approaches will be discussed. Most of the approaches will be supported by examples.

Using semantics to recommend collaboration across domains in biomedicine

Dimitar Hristovski[1], *Andrej Kastrin*[2] *and Thomas C. Rindflesch*[3]

[1]University of Ljubljana, Ljubljana, Slovenia

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

[3]National Library of Medicine, Bethesda, MD, USA

dimitar.hristovski@mf.uni-lj.si, andrej.kastrin@guest.arnes.si, trindflesch@mail.nih.gov

We propose a novel method for using semantics to recommend research collaboration across domains in biomedicine. To begin, we construct a large network representing authors, their expertise, current collaborations, and general biomedical knowledge. This network is derived from the MEDLINE bibliographic database along with semantic relations extracted from it using the SemRep natural language processing system. Finally, within the literature-based discovery paradigm, we recommend novel collaborations, which include not only pairs of authors, but also novel topics for collaboration as well as an explanation of the motivation for the collaboration.

Methodological challenges of measuring and analysing the Slovenian Twitter community

Ana Slavec[1] *and Vladimir Batagelj*[2]

[1]Faculty of Social Sciences, UL, Ljubljana, Slovenia

[2]Faculty of Mathematics and Physics, UL, Ljubljana, Slovenia

ana.slavec@gmail.com, vladimir.batagelj@fmf.uni-lj.si

Slovenia has a vibrant Twitter community – according to the former SiTweet website there were more than 20.000 Twitter users in 2012. However, there is no straightforward way to measure the activity and get network data as there is no complete list of Slovenian Twitter users. One of the previous attempts to generate a network of Slovenian Twitter users relied on geolocation (Penko 2014) but many users don't reveal their location in tweets. Moreover, the community is not limited only to people geographically located in Slovenia but also Slovenians living abroad. Thus, this approach omits certain groups of users.

Another possible approach is to apply user-generated lists of Twitter users. The most complete lists we were able to find are the four Slovenian Twitter Users lists created by the user @mn3njalnik which we used to generate a network using the NodeXL Excel module. Due to size restrictions we limited only the most frequent users and generated their network at different points in time. The data is analysed and visualized using the Pajek program. We looked into various characteristics of the network: number of vertices and lines, network density, input and output degrees, longest shortest path. We also analysed components and important subnetworks. Furthermore, we observed changes in network characteristics in time.

Statistical Applications

Can clustering based on predictors' dynamics improve the accuracy of financial distress classifier?

Lukas Sobisek[1] *and Maria Stachova*[2]

[1]University of Economics, Prague, Czech Republic

[2]Matej Bel University, Banska Bystrica, Slovakia

lukas.sobisek@vse.cz, maria.stachova@umb.sk

Financial distress is the situation in which company cannot pay or has difficulty to pay off its financial obligation. The main objective of our contribution is to study relationship between financial distress and financial indicators of selected companies. To maintain this we use a data set that consists of quantitative characteristics, e.g. financial ratios of Slovak companies collected over five year time period, namely years 2009-2013. The data thus can be transformed in multi-dimensional data sets, called panel data. We assume that knowledge about past values and trends in financial indicators can help us to predict future status of financial health. We use dynamics of time-varying predictors (financial indicators) to identify clusters of companies. The companies within a cluster are homogeneous in the trajectories of selected financial indicators. We estimate classification models fitted for all subjects and for each cluster separately and finally we compare the predictive accuracy of classifiers.

Trend component estimation

Tomáš Toupal, Patrice Marek and František Vávra

University of West Bohemia, Pilsen, Czech Republic

ttoupal@kma.zcu.cz, patrke@kma.zcu.cz, vavra@kma.zcu.cz

This paper deals with the problem of the trend component estimation particularly for the economic time series. In the real life situations it may be used in many applications, especially in macroeconomic (Gross Domestic Product, Inflation, Unemployment etc.), microeconomic (company profit and other indications), statistics, stock markets etc. The trend estimation of time series has been discussed extensively in a literature from different perspectives and there will be presented one possible method using orthonormal system generated by Gram-Schmidt orthonormalization process from some available linearly independent sequence of time series or functions in an inner product space. These results are then applied to the data collections of the balance of payments of the Czech Republic, particularly to its foreign trade part.

GraphGibbs algorithm - A Gibbs sampling method for motif finding in DNA with initial graph representation of sequences

Živa Stepančič

University of Ljubljana, Ljubljana, Slovenia
ziva.stepancic@gmail.com

Finding short patterns with residue variation in a set of sequences is still an open problem in genetics, since motif finding techniques on DNA and protein sequences are inconclusive on real data sets and their performance varies on DNA sequences of different species. In this talk we present an approach to search for possible motifs in DNA sequences in connection to Gibbs sampling method. Starting points in the search space are partly determined via graphical representation of input sequences opposed to completely random initial points with the standard Gibbs sampling. In addition, information from graphical representation helps us determine the distribution of motif sites in the set. Our algorithm was evaluated on generated as well as on real data sets by using several statistics, such as sensitivity, positive predictive value, specificity, performance and correlation coefficient.

Analysis of quantal models

Emel Cankaya[1], *Nick Fieller*[2] *and Mutlu Kaya*[1]

[1]Department of Statistics, Sinop University, Sinop, Turkey
[2]School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
ecankaya@sinop.edu.tr, nick.fieller@sheffield.ac.uk,
mutlu.alt@gmail.com

Analysis of quantal models is a particular aspect of the general problem of investigating multimodality. The distinction is that the spacings between modes are integral multiples of some unspecified fundamental unit and that the number of modes is not defined. Such semistructured models arise in a wide variety of contexts such as biology, cosmology, archaeology and molecular physics. This presentation presents a brief review of their history and development followed by an outline of the statistical methods available for their analysis, including optimality properties under particular parametric assumptions and an account of a Bayesian approach to the problem. The techniques are illustrated on problems from developmental biology and archaeology.

Economics and Econometrics

A simple randomization test of spatial correlation in the presence of common factors

Giovanni Millo

Group Insurance Research, Generali SpA, Trieste, Italy
giovanni.millo@generali.com

A randomization test is proposed for detecting spatial dependence in panel models with cross-sectional dependence induced by an unobserved common factor structure. Spatial dependence is related to the position of observations in space while cross-sectional dependence is generally not; yet spatial correlation tests have power against both. Permuting the pairs of neighbouring observations in the proximity matrix yields a simple spatial dependence test which is robust to the presence of non-spatial cross-sectional correlation and, unlike some alternatives, can accommodate short and unbalanced panels. The proposed procedure is evaluated through Monte Carlo simulation and illustrated by application to recent research on technology spillovers.

Long-run regional equilibria in a large motor insurance market

Giovanni Millo and Antonio Salera

Group Insurance Research, Generali SpA, Trieste, Italy
giovanni.millo@generali.com, antonio.salera@generali.com

We propose a long-run approach to the evaluation of economic equilibrium and profitability of the motor insurance market, applying it to highly disaggregated provincial data. We document relatively long spells of profits and losses. Allowing for heterogeneity in the individual elasticities of premiums to losses of each province, we find that premiums and claims are cointegrated. We conclude that the determination of local premiums does generally follow the local cost structure quite closely. From the viewpoint of regulators, our analysis speaks in favour of the need to consider the competitive features of insurance markets in a long-run perspective,.

Does government fiscal spending impact the quality of the environment?

Žiga Kotnik[1] *and Damjan Škulj*[2]

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

The size of the government around the world has been increasing for the last 60 years. The government is a supplier of public goods and corrector of negative environmental external costs. This is done by internalizing the polluters' costs using a polluter pays principle. Besides, in theory and in general, the expansion of government sector is unambiguously related with a rise in social well-being. We investigate the empirical significance between the size of the government, expressed by government fiscal spending, and the quality of the environment taking into account air and climate pollution indicators, water quality indicators and waste management indicators. We have applied standard OLS regression analysis to our panel data (28 EU member states and beyond EU, 1971-2014) and made appropriate robustness checks.
This work has been fully supported by the Croatian Science Foundation under the project number IP-2013-11-8174.

Estimating technical and scale efficiencies in airline industry: a non-parametric DEA approach

Burak Keskin[1], *Efehan Ulaş*[1] *and Enes Filiz*[2]

[1]Cankiri Karatekin University, Cankiri, Turkey
[2]Yildiz Technical University, Istanbul, Turkey
burakkeskiin@gmail.com, ef_ulas@hotmail.com, enesfiliz1987@gmail.com

In this study, Data Envelopment Analysis (DEA) is applied to a dataset of star alliance member airlines for a period 2012-2014 in order to measure technical and scale efficiencies of airlines. For that purpose number of aircrafts, number of airports served and number of employees were specified as input parameters; number of annual passenger and sales revenue were specified as output parameters. First, we calculated the technical, pure technical and scale efficiency scores for individual airlines and highlighted the highest and lowest efficiency scores which produced by DEA. Then, we divide airlines into four regional groups for analytical purposes: America, Europe, Asia and Africa in order to determine the most efficient region. Moreover, it is found that pure technical efficiency contributes more in compared to technical efficiency. The scale efficiency is found to be main source of overall technical efficiency. We observed that decreasing trend in pure technical efficiency whereas an opposite trend is found in scale efficiency.

Social Science Methodology

Testing hypotheses on crime seriousness perceptions

Nace Čebulj and Matevž Bren

Faculty of Criminal Justice and Security, University of Maribor, Ljubljana, Slovenia
nace.cebulj@fvv.uni-mb.si, Matevz.Bren@fvv.uni-mb.si

We want to assess the severity of criminal acts by measuring society perceptions of crime seriousness. Since the prelude of research in this field in 1960s the most significant approaches to measure crime severity have been the magnitude estimation method, scenario-based methods, economic approaches, such as individual's willingness to pay for specific crime control programs, Thurstone's method of scaling, a developmental approach and item response theory scaling. In Slovenia there has not been any research in this field yet. In 2000 Kwan et al. constructed a weighted crime index for Hong Kong. Perceived seriousness of fifteen crime typologies was assessed by the Thurstone's method of paired comparisons. The same method was used in our project. The data were collected via online survey with all possible pairs of (the same) fifteen crimes. For each pair the respondent had to choose the more serious crime between the two in a pair. About 300 Slovenian respondents contributed to over 100 comparisons for each pair of crimes. Finally the crime index was constructed using seriousness weights of the fifteen crimes and Slovenian police data of all crime events in 2008-2013. The weights and the index were then compared to those in Hong Kong. We also wanted to test hypotheses of the following type: "women perceive rape more severely than men". It turned out that in case of our data this is a nontrivial problem, which implies further research.

Sample size determination for testing two group variances of non-inferiority/superiority and equivalence with cost considerations

Wei-ming Luh[1] *and Jiin-huarng Guo*[2]

[1]National Cheng Kung University, Tainan, Taiwan
[2]National Pingtung University, Pingtung, Taiwan
luhwei@mail.ncku.edu.tw, jhguo@mail.nptu.edu.tw

Comparing population variance ratios has many applications and is performed routinely. Such comparison is a classic problem and is interesting to researchers. For example, it may be desired to compare several measuring instruments in terms of precision, treatment variability for bioequivalence of medical research, dispersion equivalence of two groups, etc. In some chronic diseases, investigators feel that the transition from health to disease is marked first by an increased variability in some indicator. Take hypertension as an example, children with a hypertension parent have more variable blood pressure levels than those of control children whose parents are normotensive. Testing two group variances is frequently performed but the sample size determination for the test is rarely discussed. The present study considers two types of statistical test of hypotheses, the non-inferiority/superiority and the equivalence, and discusses a sample size allocation ratio with respect to the statistical power and the sampling cost simultaneously. Approximate sample size formulas are developed for normal Z test based on approximating the percentiles of F distribution with the percentiles of a standard normal distribution and an iterative procedure is employed if the resulting statistical power is under the designated level. Further, we also consider the case of one group size is limited and then to find the required size of other group to achieve the designated power. After the sample size is determined, the present simulation applies the F test to the sample, and then the procedure is validated in terms of Type I errors and statistical power by simulation. Finally, R programs are prepared for practitioners.

The importance of balanced study designs in comparative studies

Ana Kolar

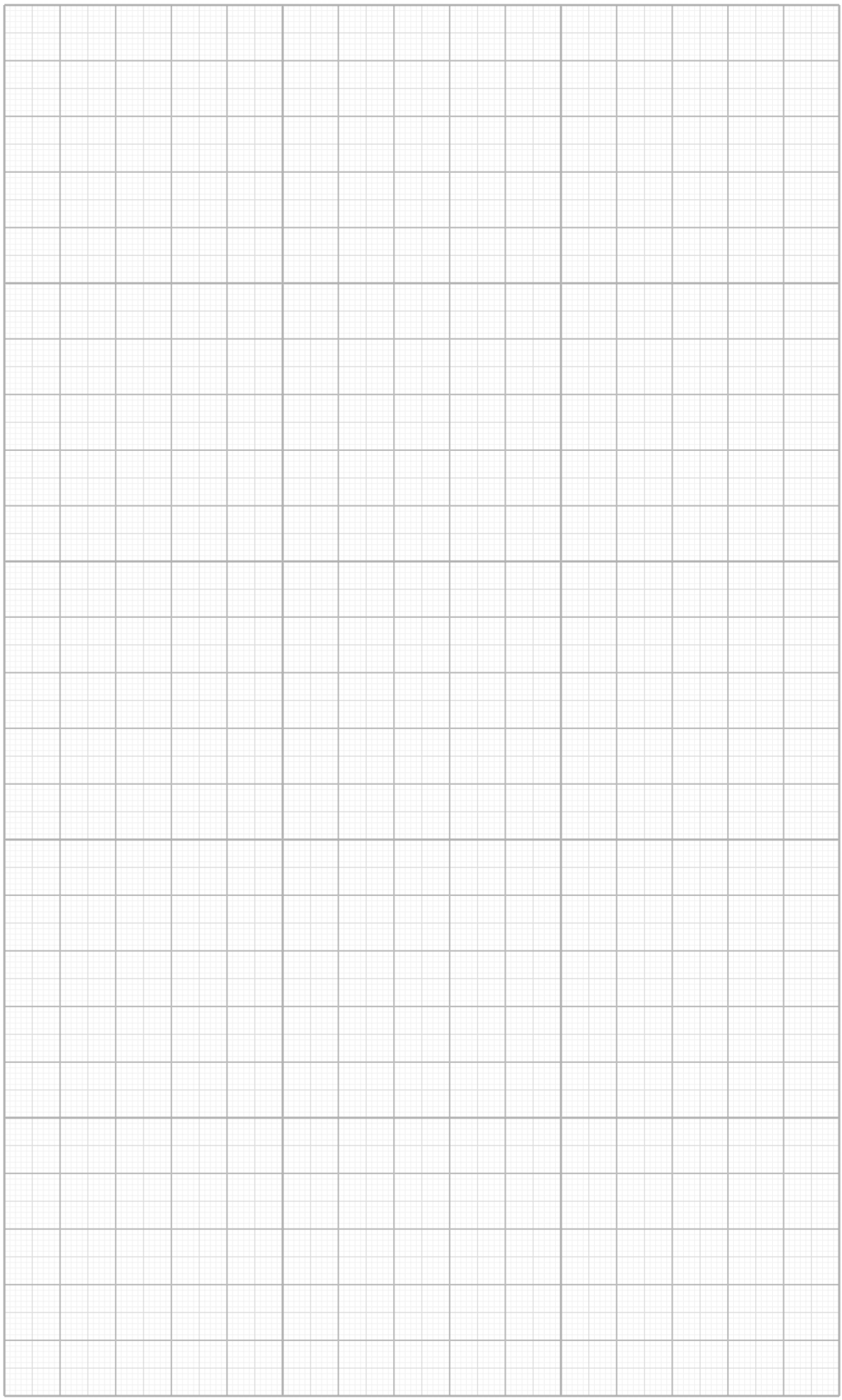
Independent Statistical Consultant, Helsinki, Finland

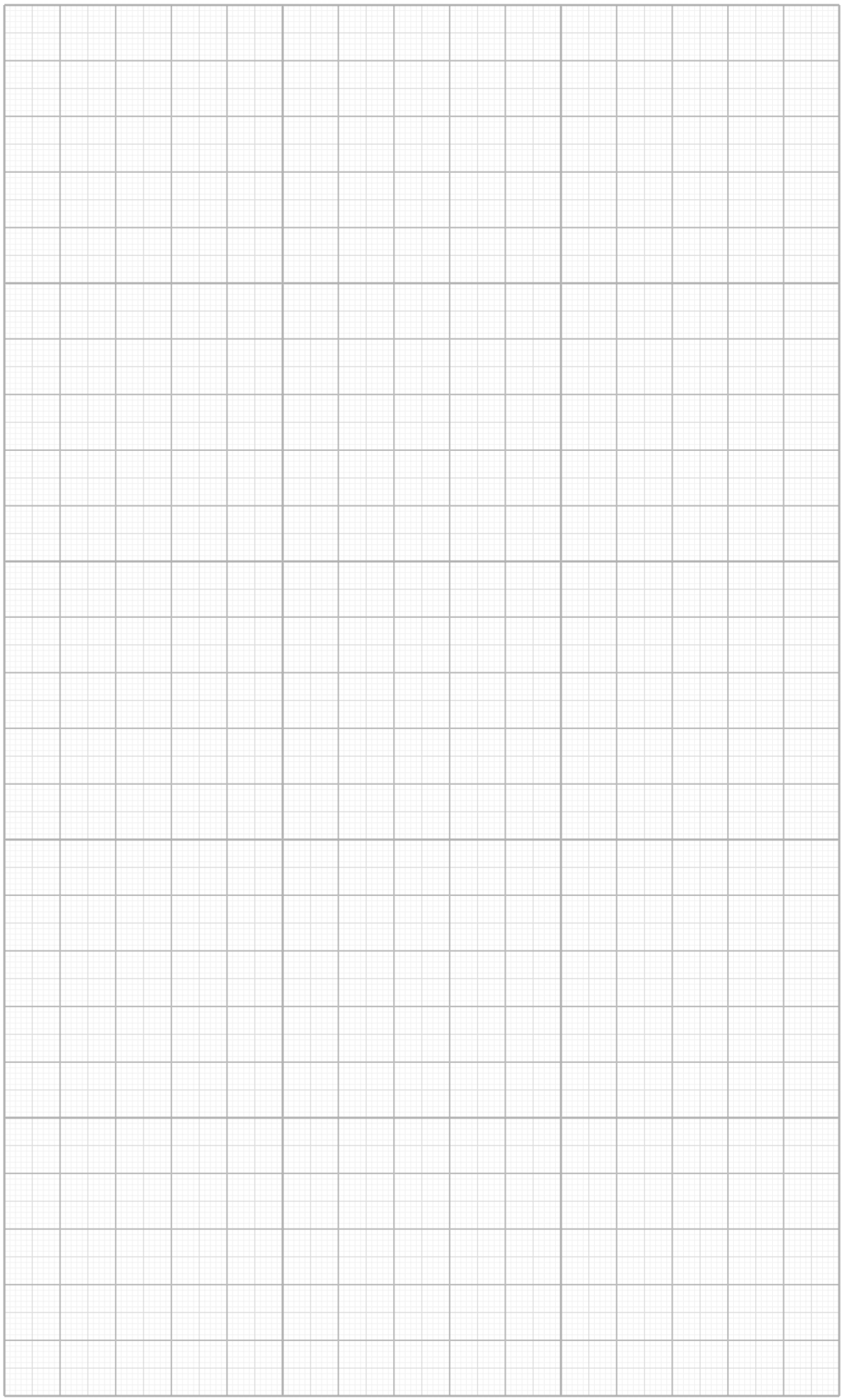
ana.kolar@tutanota.com

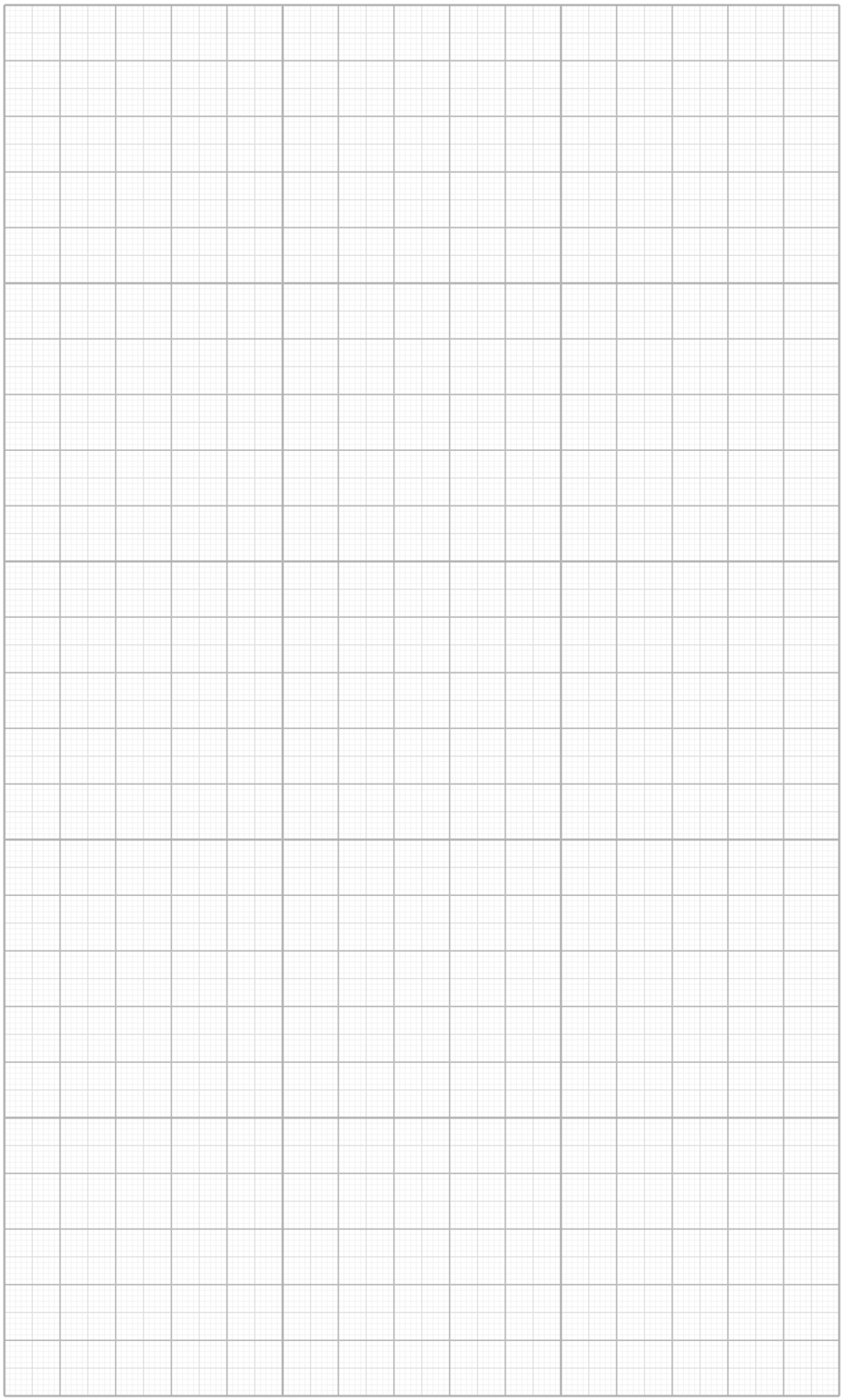
Comparative studies in social sciences research are numerous. However, the awareness that without complete randomised experiment one rarely obtains data capable to answer causal questions, is low. In a completely randomised experiment with two groups, the groups are balanced in observed and unobserved baseline covariates. Hence, the study design is fully balanced and as such satisfies the data requirements to obtain reliable causal effect estimates. Study designs that cannot be fully balanced, can be balanced partly on an observed set of covariates that are of specific interest to a researcher. In this sense, the two groups are comparable only based on this set of covariates. Hence, only estimates of associations, that are conditional on the set of selected covariates, can be obtained. This article discusses the importance of study designs in obtaining reliable statistics in comparative studies. It presents the difference between fully balanced study designs that enable causal effects estimation and so called partly balanced study designs that enable estimation of associations. All of that within the framework of propensity score methods – the most complete set of tools for balancing study designs via matching, stratification and weighting.

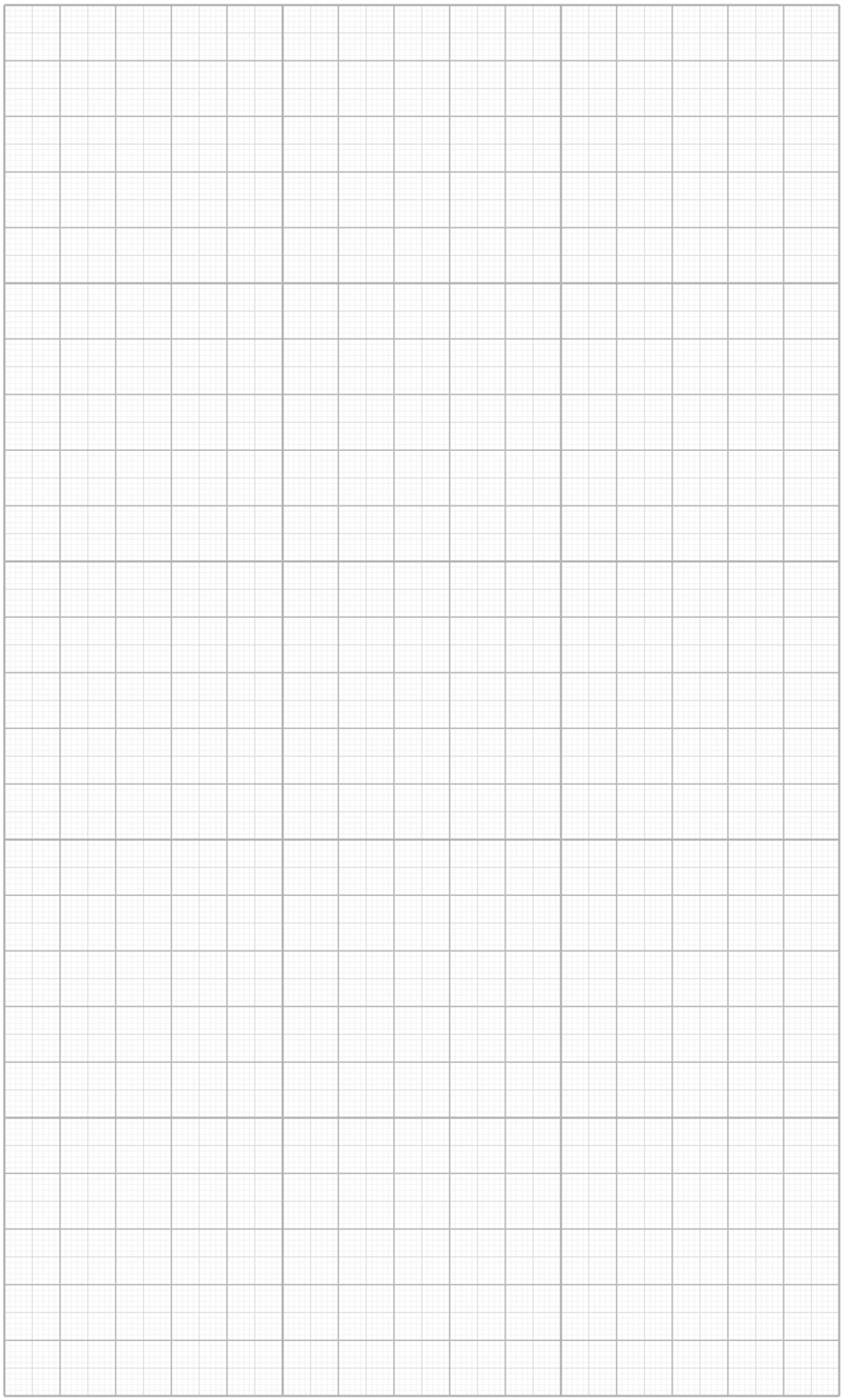
INDEX OF AUTHORS

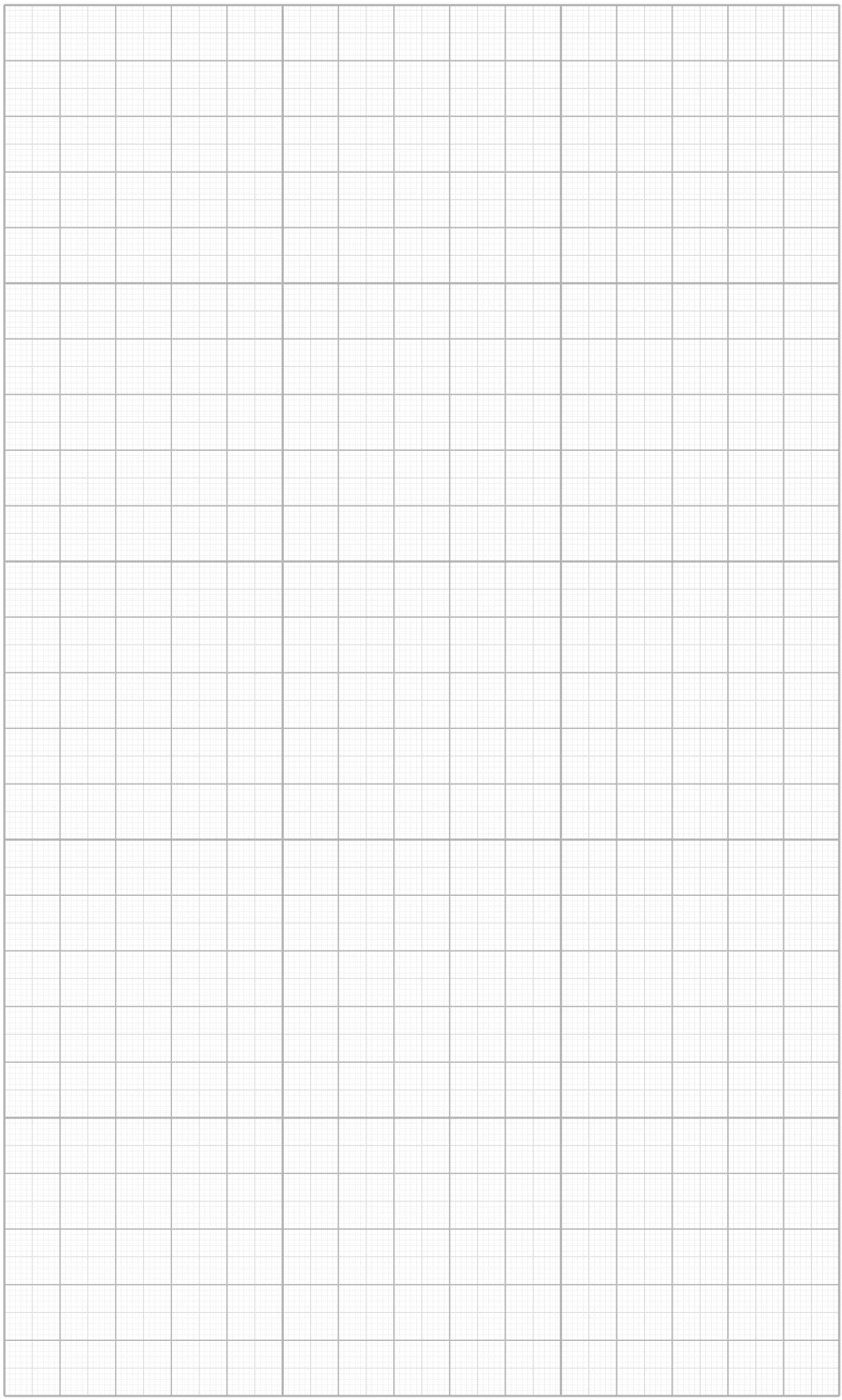
Index of Authors

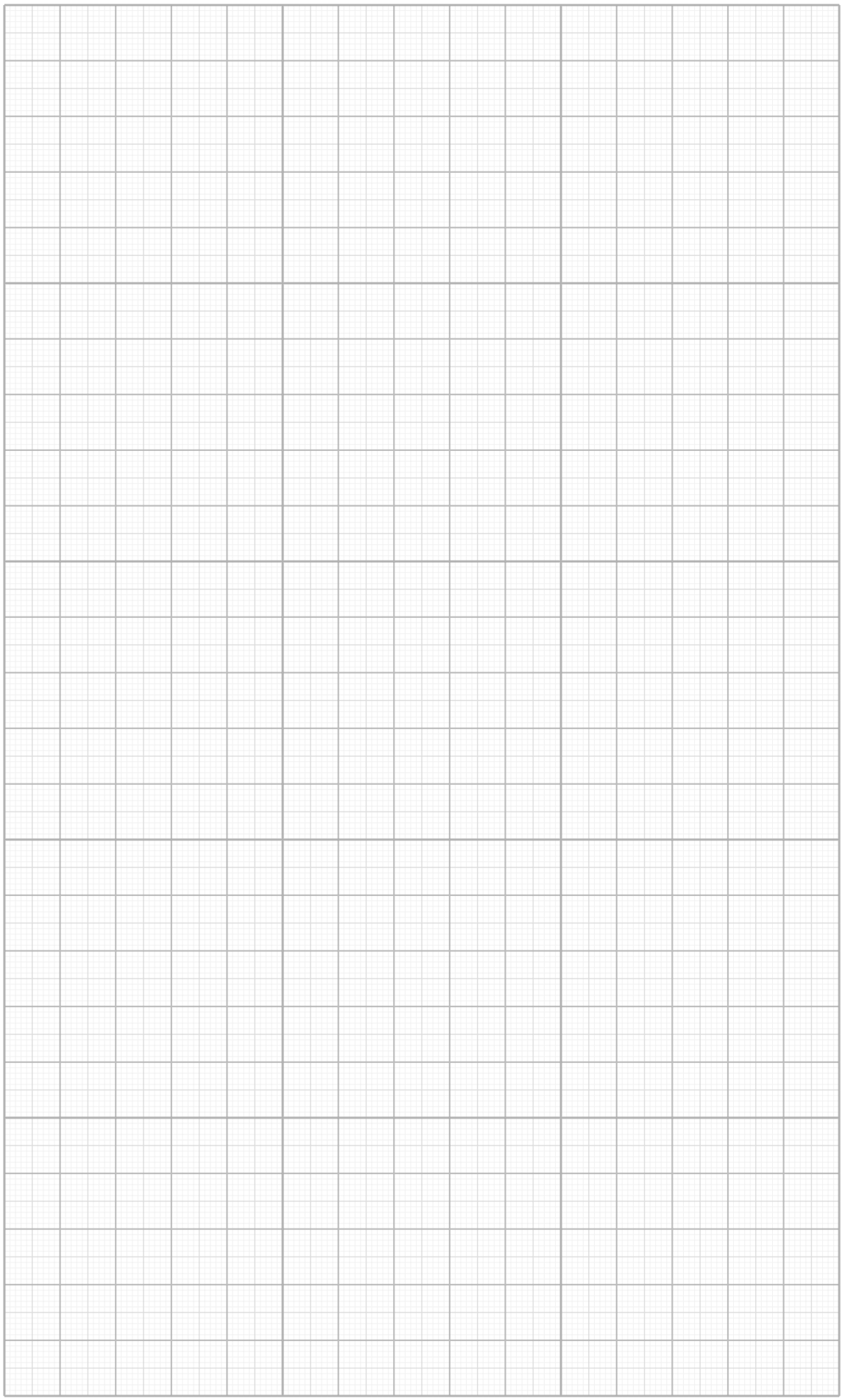


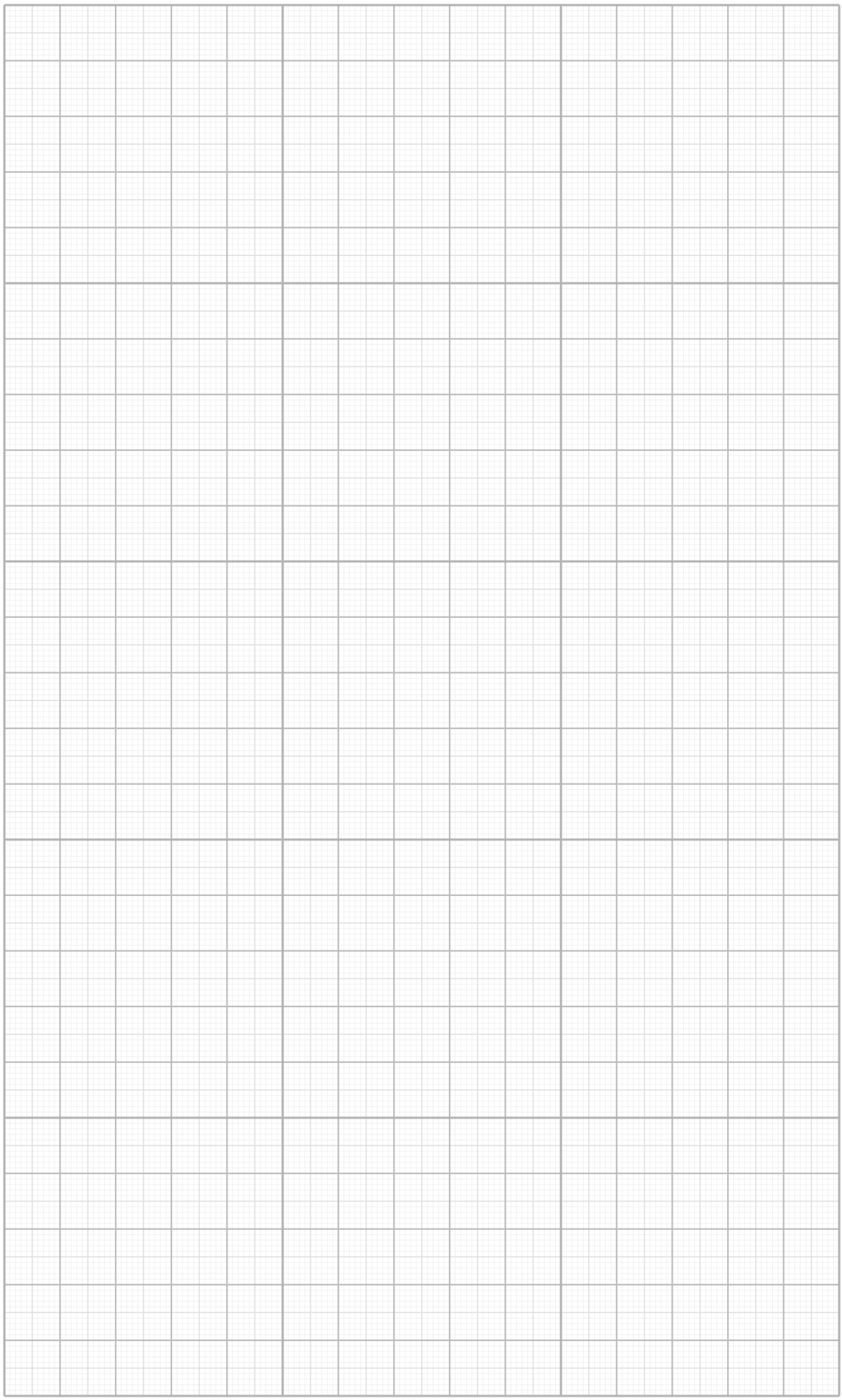












SUPPORTED BY