International Conference

APPLIED STATISTICS

2017

ABSTRACTS and PROGRAM

September 24 - 27, 2017
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2017

International Conference

APPLIED STATISTICS 2017

ABSTRACTS and PROGRAM

2017

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2017

Organized by

Statistical Society of Slovenia

Supported by

D-net

IBMI

NIB

Statistical Office of the Republic of Slovenia

RESULT

The word cloud on the cover was generated using WordArt.com. 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 (2017; Ribno)
Abstracts and program / International Conference Applied Statistics 2017, Ribno (Bled), Slovenia [September 24 - 27, 2017] ; organized by Statistical Society of Slovenia ; [edited by Lara Lusa, Rianne van den Broeke and Andrej Blejec]. - Ljubljana : Statistical Society of Slovenia, 2017

ISBN 9978-961-93547-9-7

1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
291647488

Scientific Program Committee

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

Organizing Committee

Andrej Blejec (Chair)
Lara Lusa
Irena Vipavc Brvar
Bogdan Grmek
Rianne van den Broeke
Jerneja Čuk

Published by: Statistical Society of Slovenia
Litostrojska c. 54
1000 Ljubljana, Slovenia
Edited by: Lara Lusa, Rianne van den Broeke and Andrej Blejec
Printed by: Statistical Office of the Republic of Slovenia, Ljubljana
Produced using: *generbook* R package
Circulation: 120

ABSTRACTS and PROGRAM

PROGRAM

		Hall 1	Hall 2
Sunday	15.00 – 18.00	Workshop	
	18.00 – 19.00	Registration	
	19.00 – 21.00	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	Modeling and Simulation	
	11.40 – 12.00	Break	
	12.00 – 13.40	Modeling and Simulation	Econometrics
	13.40 – 15.00	Lunch	
	15.00 – 16.00	Sampling Techniques	Statistical Applications
	16.00 – 16.20	Break	
	16.20 – 17.40	Other Areas of Statistics	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.20	Network Analysis	
	11.20 – 11.40	Break	
	11.40 – 13.00	Statistical Applications	
	13.00 – 14.30	Lunch	
	14.30	Excursion	
Wednesday	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 – 13.00	Biostatistics and Bioinformatics	Other Areas of Statistics
	13.20 – 13.30	Closing of the Conference	

15.00–18.00 **Workshop**
Hall 2

1. **Survival analysis: how to do it and what to be careful about**
   Maja Pohar Perme and Klemen Pavlič

18.00–19.00 **Registration**

19.00–21.00 **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: Mihael Perman*

1. **Approximate Bayesian computation for inference in models for large-scale network data**
Antonietta Mira, Jukka-Pekka Onnela and Ritabrata Dutta

10.00–10.20 **Break**

10.20–11.40 **Modeling and Simulation**
Hall 1 *Chair: Antonietta Mira*

1. **Bayesian stochastic DEA with time series**
Jhon Vargas, Hugo Mercado and Edwin Causado
2. **Comparison of internal evaluation criteria for categorical data in hierarchical clustering**
Zdenek Sulc, Jana Cibulkova, Jiri Prochazka and Hana Rezankova
3. **A robust stepwise bootstrap regression method**
Tommaso Gennari
4. **Some D-optimal designs: theory and examples**
Małgorzata Graczyk

11.40–12.00 **Break**

12.00–13.40 **Modeling and Simulation**
Hall 1 *Chair: Matevž Bren*

1. **A comparison of filter methods for classification of high-dimensional data sets**
Derya Turfan and Ozgur Yeniay
2. **Semiparametric statistical calibration of the numerical air pollution model outputs**
Marek Brabec, Ondrej Vlcek, Nina Benesova and Pavel Jurus

3. **Quantile estimation and comparing two independent groups by using multiple quantiles**
   Gözde Navruz and A. Fırat Özdemir

4. **Discretization of composite models**
   Yasemin Gençtürk and Ayten Yiğiter

5. **Determining optimal number of specimens per patient required for improved diagnosis**
   Jay Mandrekar

12.00–13.20 **Econometrics**
Hall 2 *Chair: Maurizio Brizzi*

1. **A fast and general cluster-bootstrapping system for panel models**
   Giovanni Millo

2. **The impact of different consumption unit scales on social indicators**
   Petr Musil and Michaela Brázdilová

3. **Applications based on regional input-output tables**
   Jaroslav Sixta and Karel Šafr

4. **Modelling tail dependence for risk measurement: the case of emerging markets**

   Tolga Yamut and Burcu Hudaverdi Ucer

13.40–15.00 **Lunch**

15.00–16.00 **Sampling Techniques**
Hall 1 *Chair: Katarina Košmelj*

1. **Inferences for stress-strength reliability of Burr Type X distributions based on ranked set sampling**
   Fatma Gül Akgül and Birdal Şenoğlu

2. **Power comparisons of several goodness of fit tests under ranked set sampling and simple random sampling**
   Yusuf Can Sevil and Tugba Ozkal Yildiz

3. **Generalized family of estimators with the help of two auxiliary variables for population variance in simple random sampling**
   Hatice Oncel Cekim and Cem Kadilar

15.00–16.00 **Statistical Applications**
Hall 2 *Chair: Nataša Kejžar*

1. **Nowcasting using principal component analysis and linear regression**
   Manca Golmajer

2. **Panel data estimation in regressions for symbolic data: an application to the clustering of cultural entrepreneurial regimes**
   Andrej Srakar and Marilena Vecco

3. **Cluster analysis of European countries according to their happiness levels and trust on their legal system, police force, parliament and politicians**
   Sebnem Er

16.00–16.20 **Break**

16.20–17.40 **Other Areas of Statistics**
Hall 1 *Chair: Gaj Vidmar*

1. **Comparing online change point algorithm with control chart analysis techniques**
   Ayten Yigiter and Canan Hamurkaroğlu

2. **Decision-curve analysis for assessing the net benefit of prediction models**
   Özlem Güllü Kaymaz, Selen Bozkurt and Yasemin Yavuz

3. **Modeling Human Development Index using finite mixtures of distributions**
   Fatma Zehra Doğru

4. **Inferences about Type-O robust correlations with a percentile bootstrap method**

   A.Firat Ozdemir

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

1. **Network models of the diffusion of innovations**
   Thomas W. Valente

10.00–10.20 **Break**

10.20–11.20 **Network Analysis**
Hall 1 *Chair: Thomas W. Valente*

1. **Dynamic multilevel blockmodeling**
   Aleš Žiberna

2. **Building and maintaining co-authorship ties**
   Luka Kronegger, Marjan Cugmas and Anuška Ferligoj

3. **Studying European countries with clustering based on causes of death**
   Aleša Lotrič Dolinar, Jože Sambt and Simona Korenjak-Černe

11.20–11.40 **Break**

11.40–13.00 **Statistical Applications**
Hall 1 *Chair: Maja Pohar Perme*

1. **Application of indirect methods for the estimation of hidden population of problem drug users**
   Katja Rostohar and Ines Kvaternik

2. **Statistical process control in the field of rehabilitation**
   Gaj Vidmar, Helena Burger and Majdič Neža

3. **Patterns in music**
   Tinka Majaron

4. **Medieval burials from Piedmont (North Italy): a statistical analysis of paleodemographic data**
   Alessia Orrù, Rosa Boano, Alessandra Cinti, Sergio De Iasio, Marilena Girotti and Maurizio Brizzi

13.00–14.30 **Lunch**

14.30 **Excursion**

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

1. **Statistical Models for Count Data with Excess Zeros: A Review**
   KyungMann Kim

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics**
Hall 1 *Chair: KyungMann Kim*

1. **Analysing long-term survival outcomes by methods from relative survival**
   Liesbeth C. de Wreede and Johannes Schetelig
2. **The three-zone diagnostic decisions in rare events settings**
   Nataša Kejžar
3. **Confidence intervals for Mann-Whitney test**
   Damjan Manevski and Maja Pohar Perme
4. **On the inexactness of the Clopper-Pearson's confidence interval**
   Anes Valentić and Rok Blagus

11.40–12.00 **Break**

12.00–13.00 **Biostatistics and Bioinformatics**
Hall 1 *Chair: Andrej Blejec*

1. **Goodness of fit test for regression models based on pseudo observations**
   Klemen Pavlič, Torben Martinussen and Per K. Andersen
2. **Relative survival analysis: comparison of net survival estimators on simulated data**
   Nina Ružić Gorenjec and Maja Pohar Perme
3. **Survival analysis on the duration of MeSH terms: preliminary results**
   Andrej Kastrin and Dimitar Hristovski

12.00–13.20 **Other Areas of Statistics**
Hall 2 *Chair: Giovanni Millo*

1. **Statistics and financial mathematics competition for (high school) students in Slovenia – past advances and future challenges**
   Aleš Toman

2. **Statistical disclosure control for Census 2021: feasibility study**
   Junoš Lukan

3. **Electronic data reporting in short-term business statistics in Slovenia**
   Nina Češek Vozel

4. **Validity and reliability test of wellbeing indicators in Thailand**
   Thanawit Bunsit

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

ABSTRACTS

Workshop

Survival analysis: how to do it and what to be careful about

Maja Pohar Perme and Klemen Pavlič

IBMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
maja.pohar@mf.uni-lj.si, klemen.pavlic@mf.uni-lj.si

Topics covered:

- Survival probability, hazard function, Kaplan-Meier estimator
- Fitting the Cox model, checking assumptions
- Immortal bias & time-dependent covariates
- Censoring vs competing risks

Prerequisites: basic knowledge of R (fitting linear regression models, basics plots)
Requirements: laptop with R installed

Invited Lecture

Approximate Bayesian computation for inference in models for large-scale network data

Antonietta Mira[1], *Jukka-Pekka Onnela*[2] *and Ritabrata Dutta*[3]

[1]InterDisciplinary Institute of Data Science Università della Svizzera italiana, Lugano, Switzerland and Università degli Studi dell'Insubria, Como, Italy, Italy

[2]Department of Biostatistics, Harvard University, USA

[3]Università della Svizzera italiana, Lugano, Switzerland

antonietta.mira@usi.ch,,

Many systems of scientific interest can be investigated as networks, where network nodes correspond to the elements of the system and network edges to interactions between the elements. Increasing availability of large-scale data and steady improvements in computational capacity are continuing to fuel the growth of this field. Network models are now commonly used to investigate social, economical and biological complexity at the systemic level. There is a general divide between the two prominent paradigms to the modeling of networks, which are the approach of mechanistic networks models and the approach of statistical network models. Mechanistic network models are knowledge domain driven and assume that the microscopic mechanisms governing network formation and evolution at the level of individual nodes are known, and questions often focus on understanding macroscopic features that emerge from repeated application of these known mechanisms. The statistical approach, in contrast, often starts from observed network structures and attempts to infer some aspects about the underlying data generating process. Mechanistic network models provide insight into how the network is formed and how it evolves at the level of individual nodes, but as mechanistic rules typically lead to complex network structures, it is difficult to assign a probability to any given network realizations that a mechanistic model may generate. Because of this difficulty, there is typically no closed form expression for likelihood for these models and, consequently, both likelihood and posterior based inference for learning from data is not possible. We have developed a principled statistical framework, based on Approximate Bayesian Computation, to bring some of the mechanistic network models into the realm of statistical inference both for parameter estimation, construction of confidence/credible intervals, hypothesis testing and model selection. This approach is feasible because, given a set of parameter values, it is easy to sample network configurations from most mechanistic models. I will introduce the general Approximate

Bayesian Computation framework and demonstrate its application to large-scale mechanistic networks, where it can be used to infer model parameters, and their associated uncertainties, from empirical data. Examples will focus on applications to social and biological networks.

Modeling and Simulation

Bayesian stochastic DEA with time series

Jhon Vargas, Hugo Mercado and Edwin Causado

University of Magdalena, Santa Marta, Colombia
jvargass@unimagdalena.edu.co, hmercado@unimagdalena.edu.co, ecausado@unimagdalena.edu.co

Data Envelopment Analysis (DEA) is a non parametric technique for measuring efficiency. Two of the main research lines are stochastic DEA (S-DEA) and DEA in a time horizon. For S-DEA the Bayesian method has arisen the last years and it can deal inference in stochastic efficiency measures. There are no methods in time horizon that explore the correlation structure in the output variables or input variables because variables are not considered as time series, either DEA model known in literature is "Bayesian" and "time horizon" at the same time. We propose to use Bayesian DEA with antedependence structure. This methodology consists in classifying the time series in clusters using maximum likelihood method and optimal Bayes rule. For that, we use the multivariate normal mixture models, specifically the random effect mixture model, and the parameter estimation by using especial algorithms, EM (Expectation Maximization) and AECM (Alternating Expectation Conditional Maximization), then we use the posterior distribution to obtain the efficiency. This new DEA model could be applied in the development of a new DEA methodology in Bayes probability intervals using information in time.

Comparison of internal evaluation criteria for categorical data in hierarchical clustering

Zdenek Sulc , Jana Cibulkova, Jiri Prochazka and Hana Rezankova

University of Economics, Prague, Prague, Czechia
zdenek.sulc@vse.cz, jana.cibulkova@vse.cz, jiri.prochazka@vse.cz, hana.rezankova@vse.cz

Evaluation of the created clusters is an important part of cluster analysis. Internal evaluation indices, which are more suitable for the use in cluster analysis, help to identify the optimal number of clusters or to evaluate compactness of clusters by comparable values. This contribution aims to compare internal indices in hierarchical cluster analysis determined for categorical data, which are not so examined as for their counterparts for quantitative data. Three groups of internal indices for categorical data are compared in this contribution. The first one is based on the within-cluster variability, the second one on the information criteria, and the third one contains indices originally determined for quantitative data, which can be calculated only using a numerical dissimilarity matrix, and thus, they can be used for categorical data. The comparison is conducted on a large number of generated datasets with various properties, such the number of variables or the numbers of categories. The clusters created by several similarity measures for categorical data are compared and evaluated mainly in terms of the optimal cluster solution determination.

A robust stepwise bootstrap regression method

Tommaso Gennari

Hall & Partners, London, United Kingdom
tommaso.gennari@themodellers.com

In market research, regression methods are often used, on survey data, to identify potential drivers of consumer behaviours and attitudes. Several methods are currently used for running optimal regression models while controlling for aspects like collinearity. One of the most used methods at the moment is Elastic Net regularisation. We propose here a regression method based on stepwise bootstrap estimates of regression coefficients. We have implemented this method both with linear and logit regression. We have recently applied it to time series analysis with a very small sample size, obtaining credible results. Earlier tests of this method show similar results compared to Elastic Net regularisation. The advantages of this method are that it controls for collinearity, it produces a coefficient/ importance score for each independent variable, and it can be used with relatively small samples. In preparation of the conference we are planning to conduct additional tests of this method. We will present this method and its validation so far, and we would like to discuss its advantages and potential pitfalls.

Some D-optimal designs: theory and examples

Małgorzata Graczyk

Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland

magra@up.poznan.pl

Let us consider the class $\mathbf{\Phi}_{n\times p}(-1,\ 0,\ 1)$ of the $n \times p$ design matrices having entries $-1,\ 0,\ 1$. The problem considered is the determination of unknown measurements of objects $w_1,\ w_2, \ldots, w_p$, when random observations $y_1,\ y_2, \ldots, y_n$ undergo the model $\mathbf{y} = \mathbf{X}\mathbf{w}$, where $\mathbf{w} = (w_1,\ w_2, \ldots, w_p)^{'}$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{'}$, $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(-1,\ 0,\ 1)$, $\mathbf{e} = (e_1,\ e_2, \ldots, e_n)^{'}$ is a random vector of errors and $E(\mathbf{e}) = \mathbf{0}_n$, $Var(\mathbf{e}) = \sigma^2\mathbf{I}_n$. The form of the variance matrix of errors means that the errors of measurements are uncorrelated and they have the same variances. Among many problems regarding to weighing designs, optimality criteria are discussed. In the present paper, we consider the optimal designs in that the generalized variance of parameter estimates is minimized. However, theoretical work on providing knowledge to guide the selection of optimal designs is not scarce, notwithstanding for any combination of the number of objects and the number of measurements we are not able to determine regular D-optimal design. In literature, some solutions of this problem are presented and some construction methods of D-optimal designs are given. Here, we study classes of design matrices not considered yet and for that reason, we give new construction method of D-optimal design. The idea is to take the regular D-optimal design and add one or more measurements in order to obtain D-optimal design in the class $\mathbf{\Phi}_{n\times p}(-1,\ 0,\ 1)$, i.e. in the class in that regular D-optimal chemical balance weighing design does not exist. For presented cases, some examples are given.

Modeling and Simulation

A comparison of filter methods for classification of high-dimensional data sets

Derya Turfan and Ozgur Yeniay

Hacettepe University, Ankara, Turkey
deryaturfan@gmail.com, yeniay@hacettepe.edu.tr

Classification of high-dimensional data set is a challenging process for statistical learning and data mining algorithms. In order to be efficient while applying classification methods to high-dimensional data sets, feature selection is an essential pre-processing step of learning process. Feature selection aims to identify the most relevant attributes and to remove the redundant and irrelevant attributes. In this study, focus is on the problem of gene selection aiming to contribute to the early diagnosis of cancer disease. With the development of bioinformatics, tumor classification from gene expression data becomes an important and useful technology for cancer diagnosis, for the discovery of the underlying genetic factors that cause diseases and for the development of cure for these diseases. Since a gene expression data set often contains thousands of genes and a small number of samples, gene selection from gene expression data becomes a key step for cancer classification. Thus, this presentation considers the problem of establishing effective feature selection and classification scheme for this type of data sets. To achieve this goal, the study makes a comparison between filter based feature selection methods which are based on different metrics and uses two different classification algorithms to evaluate the performance of the algorithms. During the process, publicly available gene expression data sets will be used and the most relevant genes reponsible for the diseases will be identified.

Semiparametric statistical calibration of the numerical air pollution model outputs

Marek Brabec[1], *Ondrej Vlcek*[2], *Nina Benesova*[2] *and Pavel Jurus*[1]

[1]Institute of Computer Science, The Czech Academy of Sciences, Praha, Czech Republic
[2]Czech Hydrometeorological Institute, Praha, Czech Republic
mbrabec@cs.cas.cz, vlcek@chmi.cz, nina.benesova@chmi.cz, jurus@cs.cas.cz

In this paper, we will consider problem of calibration for relatively complicated numerical air pollutant models (which include numerical weather, chemistry and transport modeling) using a formalized and structured semiparametric statistical approach based on GAM model framework. To this end, we will use a structured model based on several penalized spline components with carefully selected dynamic elements. We will show how such a model formulation is useful not only for the prediction performance improvements, but also for improved understanding of where the original numerical model outputs are fundamentally biased and need improvement either in their functional formulation or in various concomitant information about pollution's time and spatial profiles. After presenting general form of the calibration model, we will illustrate how it can be useful for numerical model improvements. We will also show how the calibrated model has been used for obtaining various pollution characteristics (both in space-time disaggregated and in aggregated forms) in the city of Praha, Czech Republic.
The work has been partly supported by TA CR project TA04020797.

Quantile estimation and comparing two independent groups by using multiple quantiles

Gözde Navruz and A. Fırat Özdemir

Dokuz Eylül University, İzmir, Turkey
gnavruz@gmail.com, firat.ozdemir@deu.edu.tr

Although the most common approach for comparing two independent groups is to use a measure of location such as mean, determination of the differences in the tails of the groups is often of interest. For this purpose, choosing the appropriate quantile estimator becomes a crucial issue. In this study, the default quantile estimator of R, Harrell-Davis estimator, Sfakianakis-Verginis estimators and a recently improved quantile estimator are used in conjunction with a percentile bootstrap based method with the intent of investigating actual type I errors when comparing two groups. Besides, the relative efficiencies of the corresponding estimators are also compared. The performances of the estimators are investigated under both theoretical distributions and real data sets.

Discretization of composite models

Yasemin Gençtürk and Ayten Yiğiter

Hacettepe University, Ankara, Turkey
yasemins@hacettepe.edu.tr, yigiter@hacettepe.edu.tr

As insurance claim data is highly positively skewed and has heavy tail, classical models such as Weibull, Lognormal, Pareto etc. underestimate the tail probabilities. Composite models using one standard distribution up to a threshold and other standard distribution thereafter provide a better fit when data has heavy tail. Although claim amount is continuous in nature, it is generally discrete by observation due to rounding up to nearest integer and may not constitute all points in continuum. In this case, discretization of this continuous variable is required to model the data. There are many methods used to generate the discretization of the underlying continuous variable. In this study, the discretization of some of the composite models is generated and the results of obtained from discretization and original composite models are compared.

Determining optimal number of specimens per patient required for improved diagnosis

Jay Mandrekar

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

As a statistical collaborator at a medical center, you often encounter interesting projects that use both novel clinical approaches and offer an opportunity to use innovative statistical methods. The focus of this presentation is to provide an overview of a project from clinical microbiology. Current study builds on the prior research work that used Bayesian latent class models for estimation of sensitivity and specificity due to imperfect gold standard. We will first outline the key aspects of that study. And then discuss details from recent study, where the goal was to determine number of specimens required per patient for an accurate diagnosis of a Prosthetic Joint Infection (PJI). Findings of this study have direct impact on cost savings, timely and accurate diagnosis of infection.

Econometrics

A fast and general cluster-bootstrapping system for panel models

Giovanni Millo

Group Insurance Research, Assicurazioni Generali, Trieste, Italia
giovanni.millo@generali.com

Block-bootstrapping is popular in panel data Econometrics. It consists in a bootstrapping procedure where resampling with replacement occurs not by individual unit, but by group (cluster) so as to preserve the within-group correlation structure in the bootstrap DGP. Most panel estimators admit a data transformation such that the relevant estimator is equivalent to the OLS estimator on suitably transformed data: notable examples are fixed and random effects, first differences, the GGLS versions of the former, and lastly the common correlated effects pooled estimator of Pesaran. As such, clustered covariance estimators for said models are simply defined by applying the well known sandwich equation to the transformed data. In turn, for any estimator admitting clustered standard errors, a natural block-bootstrapping scheme is defined. Bootstrap standard errors are all the more attractive for estimators whose small-sample properties are less known, or more questionable: e.g., for the different variations of the Generalized General Least Squares estimator of Parks. Moreover, the t-bootstrap variant is pivotal, and therefore provides asymptotic refinement. A cluster resampling procedure might be based on the original data, the original estimator being applied on each resampled dataset. It is numerically equivalent, but computationally much more efficient, to perform the data transformation (which, as observed, is performed cluster-wise) only once, and then apply simple OLS to each resampling from transformed data. Pairs-bootstrap and t-bootstrap for this general category of models are implemented in the 'plm' package for panel data econometrics in R, under the form of 'vcov' methods compliant with the standards of non-bootstrapped covariances; as such, they can be seamlessly employed in testing, e.g., linear restrictions. From the point of view of resources, this packaging allows to perform the computationally intensive bootstrap only once and then keep the resulting covariance object together with the original model, for use with each individual test.

The impact of different consumption unit scales on social indicators

Petr Musil and Michaela Brázdilová

University of Economics in Prague, Prague, Czech Republic
petr.musil@vse.cz, michaela.brazdilova@vse.cz

At-risk-of-poverty rate is one of the most important indicators used in social statistics. The indicator is highly dependent on the definition of equalized income. Undoubtedly, the composition of household such as number of members, age structure should be considered as households of two or more members realize economies of scales. The fact is taken into account in so called consumption unit scales that are applied in the estimation of income indicators such as average income per consumption unit. OECD scale and OECD modified scale belong among mostly used scales in Europe. However, these scales do not take into account specific economies of scale in each country. The aim of this contribution is to assess the impact of the selection of different consumption unit scales on the indicator 'at-risk-of-poverty rate' in the Czech Republic. We have estimated two scales for the Czech Republic. The first approach is based on expenditures of households as the relationship between expenditures and number of members (of which children) is estimated. The utility function of financial satisfaction is applied with the second approach that uses data from EU_SILC. At-risk-of poverty-rate is estimated using above mentioned approaches. The results are compared with officially published indicators and differences are analysed. Besides the overall indicator the impact on various social groups is examined.

Applications based on regional input-output tables

Jaroslav Sixta and Karel Šafr

University of Economics in Prague, Prague, Czech Republic
sixta@vse.cz, karel.safr@vse.cz

Regional input-output tables are rarely compiled by official statistical agencies and therefore they become the point of interest of academic researchers. The amount of information hidden in regional input-output tables easily satisfies many economists dealing with sophisticated models. For the Czech Republic, regional input-output tables were compiled for 2011 and 2013 in line with national accounts standards ESA 1995 and ESA 2010 by the Department of Economic Statistics of the University of Economics in Prague. The sets of these tables comprise tables for the use of regional output and tables for imports. The dimension is 82x82 products at basic prices. These tables are compiled for 14 regions of the Czech Republic (NUTS 3 level) and they can be arranged into the form of inter-regional model using different methods for allocation of regional trade. Inter-regional model represents a powerful tool for detailed economic analysis on the regional level. The model ensures that the effects are spreading from the first affected region to all other regions since they are linked by the flows of products. This can be used for studies about the sensitivity of employment and regional gross domestic product on external shocks such as government investments or tax incentives. Our contribution briefly illustrates the differences of investment impacts realised in different regions of the Czech Republic.

Modelling tail dependence for risk measurement: the case of emerging markets

Tolga Yamut and Burcu Hudaverdi Ucer

Dokuz Eylul University, Izmir, Turkey
yamuttolga12@yahoo.com, burcu.hudaverdi@deu.edu.tr

In order to clarify the risks of an investment, financial structure of the targeted market requires sensitive modeling that acknowledges the volatile nature. Value-at-Risk (VaR) models help the investors to forecast the consequences of the investment for a particular market. In this study, we explicitly analyze the relation between changes in oil prices and stock market returns for emerging markets. Our goal is to investigate how Brent oil and emerging markets are affected by each other and compute risk measures for these two assets. We propose to forecast the Value-at-Risk of the portfolios on the basis of bivariate copulas using nonparametric estimates of the coefficient of tail dependence which is estimated to fit the data according to its extreme observations. Changes in tail dependence by time are visualized by comparing parametric and non-parametric estimations through a simulation study. Then, VaR for the scenario of equally weighted assets are computed by simulating cumulative returns. Furthermore, validity of VaR models are tested by applying backtesting.

Sampling Techniques

Inferences for stress-strength reliability of Burr Type X distributions based on ranked set sampling

Fatma Gül Akgül[1] *and Birdal Şenoğlu*[2]

[1]Artvin Çoruh University, Artvin, Turkey
[2]Ankara University, Ankara, Turkey
ftm.gul.fuz@artvin.edu.tr, senoglu@science.ankara.edu.tr

In this study, we consider the point and the interval estimation of the stress-strength reliability $R = P(X < Y)$ when the stress X and the strength Y are both independent Burr Type X random variables based on ranked set sampling (RSS). In the context of point estimation, we obtain the maximum likelihood estimator of R by using iterative methods. We also use modified maximum likelihood (MML) methodology as an alternative to ML methodology to obtain the estimator of R. Different than ML estimator of R, MML estimator of R is obtained in closed form. In view of interval estimation, we construct the asymptotic confidence interval (ACI) of R based on the asymptotic distribution of the ML estimator of R. Also, the bootstrap confidence intervals (BCIs) of R are constructed based on two different resampling methods. The performances of the proposed estimators (both point and interval) are compared with their simple random sampling (SRS) counterparts.

Power comparisons of several goodness of fit tests under ranked set sampling and simple random sampling

Yusuf Can Sevil and Tugba Ozkal Yildiz

Dokuz Eylul University, Izmir, Turkey
yusuf.sevil@ogr.deu.edu.tr, tugba.ozkal@deu.edu.tr

Ranked set sampling (RSS) has been a popular sampling scheme in recent years. When sampling observations are expensive or time consuming to measure, RSS is preferable to simple random sampling (SRS). Many authors are used to the sampling method in different fields and suggested several versions of RSS until today. In these studies, many estimators and statistical tests are proposed for RSS. In this work, we discuss the performances of goodness of fit tests (GOF) based on emprical distribution function, quadratic emprical distribution function and kernel function under ranked set sampling and simple random sampling in finite population. Critical values are obtained under standard normal distribution for each GOF tests by using RSS and SRS. To obtain the powers of these GOF tests, many alternative distributions including symmetric and skewed are generated in R software. The powers are computed using RSS and SRS. The three different class of GOF tests are compared for the different alternative distributions. In RSS, the power results are obtained for different set and cycle sizes. Finally, for each class, the best GOF tests will be investigated.

Generalized family of estimators with the help of two auxiliary variables for population variance in simple random sampling

Hatice Oncel Cekim and Cem Kadilar

Hacettepe University, Ankara, Turkey
oncelhatice@hacettepe.edu.tr, kadilar@hacettepe.edu.tr

The variance estimator has been proposed by many authors for many years. It is highly preferred to use auxiliary variable knowledge to obtain more effective estimators. Moreover, the proposed estimators have been studied using the information of two auxiliary variables in recent years. In this study, we propose a family of estimators with the help of two auxiliary variables for the estimation of the population variance. We obtain mean square error (MSE) of the proposed family of estimator under two different conditions. The proposed estimators are compared with the mentioned estimators in theory and numerical illustration. It is seen in practice and theory that the performance of the proposed estimators is better under the conditions obtained.

Statistical Applications

Nowcasting using principal component analysis and linear regression

Manca Golmajer

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
Manca.Golmajer@gov.si

In official statistics, many users wish that good estimates of some indicators would be available more quickly. Therefore, statisticians attempt to produce flash estimates that have a considerably shorter publication lag and low revision error. Various nowcasting methods can be used. The Statistical Office of the Republic of Slovenia tested a nowcasting model that consists of two stages. In the first stage, time series of microdata are prepared; principal component analysis is used and the first few principal components are selected to present microdata using less time series. In the second stage, the selected principal components are used as predictors in linear regression, where the dependent variable is the time series of interest which we wish to nowcast. Other predictors can be used in this linear regression as well. In our examples, we used different time series of interest (e.g. GDP at constant prices), microdata (e.g. turnover for enterprises in industry (from our sample survey)), and other predictors (e.g. economic sentiment indicator, seasonal component, traffic sensor data). We used different conditions for selecting principal components (e.g. take enough principal components to explain 80 percent of the variability of microdata). We also tested using only data that are available less than a specific number of days (e.g. 45 days) after the end of the period we wished to nowcast. Different models were compared using several statistics (e.g. mean absolute error of nowcasts, mean squared error of nowcasts). The obtained results show that using the PCA method in combination with various types of data could be a useful contribution to our effort to nowcast important indicators.

Panel data estimation in regressions for symbolic data: an application to the clustering of cultural entrepreneurial regimes

Andrej Srakar[1] *and Marilena Vecco*[2]

[1]Institute for Economic Research (IER), Ljubljana and Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

[2]Erasmus University Rotterdam, Rotterdam, The Netherlands

andrej_srakar@t-2.net, vecco@eshcc.eur.nl

"Entrepreneurial regimes" is a topic, receiving quite a lot of research attention in the recent years, but the topic has been seldom applied to the field of cultural entrepreneurship. Moreover, the existing studies on entrepreneurial regimes mainly use common methods from multivariate analysis and some type of institutional related analysis. In our analysis we study the cultural entrepreneurial regimes applying a symbolic data analysis approach and using Amadeus data for the period 2006-2015 and for 28 EU countries. On the basis of these data we extract a set of histogram variables in particular related to the characteristics of the firms in culture (socio-economic characteristics and financial variables and ratios). These variables are employed in the symbolic clustering analysis, following the most recent trends, to derive a set of temporally robust clusters, labelled as cultural entrepreneurial regimes. Next, we try to derive mathematical formulas for panel data estimators for regression analysis with symbolic data and explore their statistical behaviour. Finally, the behaviour of clusters (cultural entrepreneurial regimes) is analyzed in regressions for symbolic data, including panel data estimation. The main research questions of the article are: 1) To what extent do the clusters follow the commonly found classifications of entrepreneurial regimes? 2) Are there any significant changes in the positions of individual countries observed in the studied time period? 3) What are the explanations for these changes? 4) Does the inclusion of formal panel data modelling improve results from more common regression specifications for symbolic data in the studied empirical example? The analysis is to our best knowledge the first symbolic data analysis in cultural entrepreneurship and economics. Moreover, it derives formulas for panel data estimators in the case of regressions for symbolic data, contributing to the development of symbolic data analysis research field.

Cluster analysis of European countries according to their happiness levels and trust on their legal system, police force, parliament and politicians

Sebnem Er

University of Cape Town, Cape Town, South Africa
sebnem.er@uct.ac.za

The European Social Survey (ESS) is an academically-driven multi-country survey, which has been administered in over 30 countries to date. Its three aims are, firstly - to monitor and interpret changing public attitudes and values within Europe and to investigate how they interact with Europe's changing institutions, secondly - to advance and consolidate improved methods of cross-national survey measurement in Europe and beyond, and thirdly - to develop a series of European social indicators, including attitudinal indicators. The survey includes data from European Union countries and some non-European countries such as Turkey and Russia. The main aim of this research is to cluster the countries based on their happiness levels and trust levels on their legal system, police force, parliament and politicians. The analysis reveals that Denmark has the highest score for all the variables in consideration whereas Poland, Czech Republic have lower trust levels. Countries such as Germany and England where the population is quite diverse, there is no clear distinction between the levels of trust on legal system, police force, parliament and politicians.

Other Areas of Statistics

Comparing online change point algorithm with control chart analysis techniques

Ayten Yigiter[1] *and Canan Hamurkaroğlu*[2]

[1]Hacettepe University, Ankara, Turkey
[2]Karabük University, Karabük, Turkey
yigiter@hacettepe.edu.tr, cananhamurkaroglu@karabuk.edu.tr

Change point problems arise in many fields such as seismology, genetic, meteorology, economy etc. To solve these problems, online and offline change point detection methods are available in the literature. The online change point detection algorithm, proposed by Fearnhead and Liu (2007), is basically based on a Bayesian product partition and standard recursive filtering. Other methods based on control charts are used to detect a change point in the process. In this study we consider multiple change points problems in mean of the normally distributed sequence. We conduct a simulation study to compare the online change point algorithm with CUSUM, EWMA, Shewhart charts by using root mean square error(RMSE) and average run length (ARL).

Decision-curve analysis for assessing the net benefit of prediction models

Özlem Güllü Kaymaz[1], *Selen Bozkurt*[2] *and Yasemin Yavuz*[1]

[1]Ankara University, Ankara, Turkey
[2]Akdeniz University, Antalya, Turkey
ozlem.gullu@gmail.com, selenb@gmail.com,
yasemin.genc@medicine.ankara.edu.tr

Decision-curve analysis (DCA) is a simple method to quantify the clinical usefulness of a prediction model (or an extension to a model). In order to assess a biomarker as a determinant to improve patient care, it has to improve decision making in the clinical setting. Unlike the other methods, DCA, which is recently proposed as a novel method for evaluating predictive models, incorporates consequences of clinical decisions, such as an increased number of unnecessary or missed treatments. DCA calculates the net benefit of a model by the difference between the number of true-positive and false-positive results, weighted by the odds of the selected threshold probability of risk. It also visualizes the clinical consequences of a treatment strategy. The aim of the present study is to use DCA to determine whether adding a biomarker to a previously developed baseline model of risk stratification improves clinical decision making for the early detection of disease. DCA represents the clinical net benefit of prediction models: one sums the benefits (true positives) and subtracts the harms (false positives) at each threshold probability of the outcome. As the threshold probabilities can vary from patient to patient, the net benefit is calculated across a range of probabilities. A model can be compared with competing models or default strategies of treating all or treating none by using DCA. As an application of this study, DCA of the prediction model will be applied on a medical data set.

Modeling Human Development Index using finite mixtures of distributions

Fatma Zehra Doğru

Department of Econometrics, Faculty of Economics and Administrative Sciences, Giresun University, Giresun, Türkiye

fatma.dogru@giresun.edu.tr

The Human Development Index (HDI) measures the development level of a country which was laid out by the United Nations Development Programme (UNDP). Due to the fact that the values of HDI for different countries become different with respect to the development level of a country, the distribution of HDI may have one more mode, thick tail or skewness. Thus, this type of data sets can be modeled by using the mixtures of distributions to deal with modality, heavy-tailedness and/or skewness. In this study, we propose to model the data set from the HDI report 2015 for 188 countries with finite mixtures of distributions. We give the basic scheme of the maximum likelihood (ML) estimation procedure for the finite mixture model based on the Expectation-Maximization (EM) algorithm. To get the best model for HDI data set, we first determine the appropriate cluster number using model-based clustering. Then, to find the best model for HDI data set, we use the finite mixture models obtained from some symmetric and/or heavy-tailed and skew and/or heavy-tailed distributions.

Inferences about Type-O robust correlations with a percentile bootstrap method

A.Firat Ozdemir

Dokuz Eylül University, Department of Statistics, İzmir, Turkey
firat.ozdemir@deu.edu.tr

A measure of the linear association between two random variables X and Y is a fundamental component of statistical methods. It is clear that the most frequently applied choice in applied work is Pearson's correlation which is very weak in terms of robustness. Wilcox (2012) classified robust analogs of Pearson's correlation into two types: those that protect against outliers among the marginal distributions without taking into account the overall structure of the data (type M), and those that take into account the overall structure of the data when looking for outliers (type O). Before calculating a Type-O robust correlation a multivariate outlier detection rule is applied and there are very limited results for making inference with Type-O robust correlations in the literature. In this study, the performance of a hypothesis testing procedure based on a percentile bootstrap method was investigated for different Type-O robust correlations in terms of actual significance level and power.

Invited Lecture

Network models of the diffusion of innovations

Thomas W. Valente

Institute for Prevention Research, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, USA

tvalente@usc.edu

Diffusion of innovations theory posits that new ideas and practices spread within and between communities via interpersonal influences. Network models of the diffusion of innovations have been developed in order to measure these interpersonal influences and debates exist regarding appropriate ways to estimate influence effects. In this talk I describe the many ways interpersonal influence has been measured and modeled. In addition to exposure/contagion effects, other interpersonal influences include structural equivalence, indirect ties, thresholds, Simmelian ties, homophilous ties and so on. The interaction between these individual level influences and macro network properties are discussed. The talk concludes with observations on the intersection between theory and interpersonal influence which leads to implications for using social network data for program implementation and intervention. Strategies and algorithms used to accelerate the diffusion of innovations and results from current studies are presented.

Network Analysis

Dynamic multilevel blockmodeling

Aleš Žiberna

University of Ljubljana, Ljubljana, Slovenia
Ales.Ziberna@fdv.uni-lj.si

In this work, blockmodeling of multilevel network data gathered in several time periods (here named dynamic multilevel blockmodeling) will be presented. Multilevel network data consist of networks that are measured on at least two levels (e.g. between organizations and between people) and information on ties between these levels (e.g. information on which people are members of which organizations). Dynamic networks are here meant networks which are measured in several (at least two) time points. I have argued before that both of these two types can be treated as special cases of linked networks and that the same procedure can be used for blockmodeling them. Here I present for the first time blockmodeling of a network that is both multilevel and dynamic. This means that we have in the presented example network measured at two levels (and ties between them), where all measurements are taken in two time points. The k-means like approach followed by generalized blockmodeling approach both adapted to lined networks is applied to this network. As usual with the linked/multilevel blockmodeling, special attention must be put to weighting parts of this linked network.

Building and maintaining co-authorship ties

Luka Kronegger, Marjan Cugmas and Anuška Ferligoj

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

From everyday personal experiences we know, that establishing relationship to someone is different process from maintaining relationship for a longer period. If we narrow our general experience to scientific community, do we recognise a certain collaborative connection to a fellow researcher as a one night stand?
Research collaboration is one of the fundamental principles of modern science. The collective nature of the research work in modern science represents the driving mechanism of contemporary scientific advancement. The study of scientific collaboration, of co-authorship networks, enables us to get new insights into the dynamics and structure of mechanisms; on one hand, for establishing, and on the other, for maintaining scientific collaborations. In this presentation we will introduce the study of factors that lead to the establishment or maintenance of ties in co-authorship networks. The presented analysis is additional research based on results of Ferligoj et al. (2016).
Analysed co-authorship networks were generated based on the Slovenian national bibliographic database (COBISS) and analysed with available stochastic methods for modelling of network dynamics (RSiena)

Studying European countries with clustering based on causes of death

Aleša Lotrič Dolinar, Jože Sambt and Simona Korenjak-Černe

Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia
alesa.lotric.dolinar@ef.uni-lj.si, joze.sambt@ef.uni-lj.si, simona.cerne@ef.uni-lj.si

People in countries with different sex-age specific mortality suffer from different health problems. Consequently, the countries with different mortality level and/or structure face different demographic issues as well as health costs that both have important implications for labor markets, social security, health care systems and related development strategies. For implementing proper health policy and improving health cost management it is important to know which countries are similar in a sense of sex-age-cause specific mortality and what are the differences between the groups of similar countries. We study the mortality in most European countries using classical and symbolic data analysis clustering methods, based on different types of input data: death level only, relative structure of deaths by causes of death, and combination of the two. As causes of death are strongly related to sex and age, the input data are provided separately for each sex-age group. The clustering results are compared based on how they capture both the level of mortality and the relative distribution of deaths by causes of death. We compare the clustering results also with the clusters based on life expectancy at birth that is considered the main demographic indicator of country development.

Statistical Applications

Application of indirect methods for the estimation of hidden population of problem drug users

Katja Rostohar[1] *and Ines Kvaternik*[2]

[1]National Institute of Public Health, Ljubljana, Slovenia
[2]National Institute of Public Health, Koper, Slovenia
Katja.Rostohar@nijz.si, Ines.Kvaternik@nijz.si

The assessment of problem drug users is problematic in many aspects. First, because drug use is a sensitive topic and drug users do not want to expose themselves and disclose their use, while they officially violate law and are socially excluded as such. Second, they are often homeless so it is difficult to find and assess them. Therefore the use of direct methods in the estimation of the number of drug users is ineffective or impossible to apply as we cannot obtain relevant data on their drug use. In this case, we applied for Slovenia two indirect methods which were recommended by the EMCDDA. The first applied method was Capture-Recapture, where the datasets of drug related deaths from mortality register and the data on drug use from medical treatment centers were used. The estimation of problem drug use was obtained for the whole Slovenia. An estimate shows that from 2009 to 2012 there were about 16100 problem drug users and about 12800 poly-drug users, where the estimates change through years. The estimates seem to be overestimated due to dependency of datasets and the confidence interval of estimation is quite wide, mostly due to a low number of drug related deaths. In the second case, the Single list method was applied using Poisson distribution for the estimation of hidden population. The data on the frequency of visits in low threshold program in the coastal area were used for selected time periods. Review of the data showed that there were about 195 different drug users that visited low-threshold programs in the coastal area in one year. The estimates on two months data collection showed an estimate of about 200 drug users and on 4 months or more about 350 drug users. The estimates are different when using different time periods of observation.

Statistical process control in the field of rehabilitation

Gaj Vidmar[1], *Helena Burger*[2] *and Majdič Neža*[2]

[1]University Rehabilitation Institute; University of Ljubljana, Faculty of Medicine, Institute for Biostatistics and Medical Informatics; University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Ljubljana; Ljubljana; Koper, Slovenia

[2]University Rehabilitation Institute, Ljubljana, Slovenia

gaj.vidmar@ir-rs.si, helena.burger@ir-rs.si, neza.majdic@ir-rs.si

Statistical process control (SPC) and statistical quality control (SQC) have gained popularity and appreciation in health care during the last 15 years. However, they have found virtually no application in rehabilitation. Apart from our work, there are only two published scientific articles on this topic (one from 2005 and one from 2017). They present SPC at individual-patient and patient-subgroup level (i.e., down-top), whereas our approach is at hospital and ward level (i.e., top-down). We have been introducing SPC/SQC at our Institute over the recent years for practical and research purposes. Here we present three completed projects: auditing of efficiency and effectiveness of inpatient rehabilitation (as assessed using the Functional Independence Measure at admission and discharge); evaluation of a fall-prevention programme for stroke-rehabilitation inpatients; and monitoring of falls across all our hospital wards. Exploratory graphics, regression models (linear, robust and Poisson) and control charts (p-charts, u-charts and XmR-charts) were applied. In the first project, we found no clear trend in the proportion of inpatients admitted directly from other hospitals or in the proportion of first admissions to primary rehabilitation over the last 10 years, and a clearly increasing trend in rehabilitation efficiency and effectiveness. In the second project, the analyses confirmed the reduction in the falls rate due to the programme. In the third project, the required IT infrastructure was set up and the control charts have been introduced into internal periodic reporting. Hence, SPC is feasible, can yield important insight and provides valuable decision-support in rehabilitation practice.

Patterns in music

Tinka Majaron

Konservatorij za glasbo in balet Ljubljana, Ljubljana, Slovenija
tinka.majaron@gmail.com

Even a small child can recognize patterns in music. Having that in mind, is it reasonable to look for patterns using statistics and computers? Statistically speaking, almost nothing changes, if we modify a single note in a song, but in musical terms the difference is enormous. Musical composition is not just a series of decisions on notes, so at first glance it would seem that music and statistics have nothing in common. The truth is quite opposite, only by using statistics (and computers) we can see some surprising patterns. Around the world there are many researches in the field of statistical analysis of music – computers can identify the type of music, write a song in a certain style and so on. In my research, I was looking for the simplest patterns in music that would tell something to the musicologist who would like to analyze one specific composition or composer. This lecture would be the presentation of my journey towards these patterns with the following itinerary: • the basics of musical theory, • converting musical notes into numbers, • analysing a melody using the R programming language, • looking at the most repeated patterns with a length from 4 to 31 notes (from motive to theme), • music (at the last stop we must hear the results).

Medieval burials from Piedmont (North Italy): a statistical analysis of paleo-demographic data

Alessia Orrù[1], *Rosa Boano*[1], *Alessandra Cinti*[1], *Sergio De Iasio*[2], *Marilena Girotti*[1] *and* *Maurizio Brizzi*[3]

[1]Department of Life Sciences and Systems Biology, University of Turin, Torino, Italy
[2]Department of Chemical, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy
[3]Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Bologna, Italy
aorru@unito.it, rosa.boano@unito.it, alessandra.cinti@gmail.com, sergio.deiasio@unipr.it, marilena.girotti@unito.it, maurizio.brizzi@unibo.it

Single burials dated to the period between the 8th and 13th century were discovered under the floor of the Cathedral of San Lorenzo (Alba, CN) during an archaeological excavations undertaken by the Soprintendenza Archeologica (Archaeologic Authority) during 2007–2011. The number of individuals recovered is about 350, both adults and non adults, males and females. Skeletal remains are well placed into chronological phases of the cemetery, based on stratigraphic evidences and archaeological data. The statistical analysis of paleo-demographic data aims to verify the differences in the demographic composition and in the biological characteristics of the examined sample.

Invited Lecture

Statistical Models for Count Data with Excess Zeros: A Review

KyungMann Kim

University of Wisconsin-Madison, USA
kyungmann.kim@wisc.edu

Count data are routinely analyzed using Poisson (P) distributions. Due to population heterogeneity, however, they often exhibit over-dispersion known as the extra-Poisson variation. This extra-Poisson variation can be handled in one of two ways, maximum quasi-likelihood method or a latent variable model leading to negative binomial (NB) distribution with a gamma mixing distribution for the Poisson mean. Still there are situations where these models perform poorly because of excess zeros in the count. There are two similar, but conceptually different approaches to handling excess zeros. In what is commonly known as zero-inflated (ZI) models, we may view the data as being generated from a mixture model with a point mass at zero representing "excess" zeros and a standard non-degenerate distribution including "true" zeros. This mixture model allows for mixture of two different populations, one non-susceptible for events (resulting in excess zeros) and the other susceptible (including true zeros). In contrast, the so-called hurdle (H) models may be conceptualized as having zeros only from a non-susceptible population and can be modeled using two processes, one generating zeros ("choice") and the other generating only the positive counts ("intensity") from a truncated count distribution. In this presentation, I will show examples of count data with excess zeros from the literature in various disciplines and applications and review recently developed marginal mean models for count data with excess zeros for illustration.

Biostatistics and Bioinformatics

Analysing long-term survival outcomes by methods from relative survival

Liesbeth C. de Wreede[1] *and Johannes Schetelig*[2]

[1]Dept of Medical Statistics and Bioinformatics, LUMC, Leiden, the Netherlands
[2]Universitätsklinikum, Technische Universität, Dresden, Germany
l.c.de_wreede@lumc.nl, johannes.schetelig@uniklinikum-dresden.de

When long-term survival outcomes of a group of patients are analysed, information about the mortality of the general population may be considered to study excess mortality of the patient population and the risk factors associated with it. This is especially relevant for an older patient population. We studied two large multi-national cohorts of patients who had received an allogeneic haematopoietic stem cell transplantation and whose data had been collected by the European Society for Blood and Marrow Transplantation, one of CLL (chronic lymphocytic leukaemia) and one of MDS (myelodysplastic syndrome) patients. Allogeneic transplantation is the only curative treatment for these patients, yet it is also associated with high mortality, due to the underlying disease, previous treatment or the transplantation itself. Each patient was matched to an artificial control from the general population with the same sex, age and nationality in the year of transplantation. Population tables were obtained from the Human Mortality Database. In both cohorts, we separated background mortality, mortality after relapse of the underlying disease and excess non-relapse mortality through cumulative incidence and crude mortality curves. Cox proportional hazards models for excess mortality were fitted to investigate the impact of risk factors. We also studied changing impact of risk factors over time on death after relapse and non-relapse mortality by means of landmark analyses and models for the relative contribution of causes of death over time. A large part of the analyses was performed by means of the 'relsurv' package in R of Maja Pohar Perme. The presentation will show how methods from relative survival and competing risks may help to understand the different components of mortality better.

The three-zone diagnostic decisions in rare events settings

Nataša Kejžar

University of Ljubljana, Faculty of medicine, IBMI, Ljubljana, Slovenia
natasa.kejzar@mf.uni-lj.si

Assessment of diagnostic and prognostic biomarkers allows one to determine whether biomarker could be used as a surrogate for the diagnosis (prognosis). Usually the candidates for surrogate tests are less expensive, less invasive and easier to perform. It is therefore of crucial importance to quantify the candidates with regard to the gold standard. This is done with computing diagnostic accuracy (plotting ROC curves) etc.
In this talk we concentrate on candidate tests with numeric values and discuss the problems in detecting the optimal cut point for differentiating between cases and controls. The notion of "gray zone" is used with ROC curves and likelihood ratios (LR) that is useful to compute for the candidate test. It forms three regions: the likely cases, controls and the still undecided.
Sensitivity, specificity and LR are all invariant with regard to the prevalence of the cases in the population and they take part in the computation of the gray zone. Usually one of the classes is more prevalent. We explore the behaviour of optimal cut points and gray zone in this setting.

Confidence intervals for Mann-Whitney test

Damjan Manevski[1] *and Maja Pohar Perme*[2]

[1]Student of Applied statistics, University of Ljubljana, Ljubljana, Slovenia
[2]IBMI, Medical faculty, University of Ljubljana, Ljubljana, Slovenia
man.damjan@gmail.com, maja.pohar@mf.uni-lj.si

The Mann-Whitney test is a commonly used non-parametric alternative of the t-test. Despite its frequent use, it is only rarely accompanied with confidence intervals, furthermore, the confidence intervals are often reported either for the difference in the medians or, more generally, for a shift of the two distribution locations. Neither of these two measures directly coincides with the test statistic, so no duality between the confidence intervals and the hypothesis testing can be expected. In this paper, we focus on the measure of distributions' overlap, which is in direct one-to-one relationship with the Mann-Whitney test statistic. It is also equal to the area under the ROC curve, which greatly improves both its interpretability and usability. We present the ideas behind the estimation of the variance and construction of the confidence intervals. We find examples in which the existing estimators perform poorly, explain the reasons for this behaviour and propose alternative approaches to confidence intervals construction that ensure better coverage.

On the inexactness of the Clopper-Pearson's confidence interval

Anes Valentić[1] *and Rok Blagus*[2]

[1]Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Univerza na Primorskem, Koper, Slovenia

[2]Medical faculty, University of Ljubljana, Ljubljana, Slovenia

anesvalentic@hotmail.com, rok.blagus@mf.uni-lj.si

One of the most commonly used methods for calculating interval estimates of a binomial parameter is Clopper-Pearson's confidence interval, also known as the exact binomial confidence interval. As most exact methods, the Clopper-Pearson's interval gained its name for being exact in terms of using probability distribution that does not depend on any unknown parameters. However it is inexact in the sense that its actual coverage probability is not equal to the nominal coverage probability.

In this talk we will present a formal proof of the inexactness of Clopper-Pearson's interval and show how the actual coverage probability can be calculated without doing simulations for any sample size n, binomial parameter p and confidence level $1 - \alpha$. Not only will in the proof be shown that the interval is inexact, it will be shown that in fact Clopper-Pearson's interval is conservative.

We compare the coverage probability of Clopper-Pearson's interval with some other commonly used interval estimators of the binomial parameter, namely the standard interval which is based on inverting the Wald large sample normal test and Wilson's confidence interval which is based on inverting the approximately normal test that uses the null standard error, rather than the estimated standard error.

The results based on extensive simulations show that Clopper-Pearson's test is giving too large actual coverage probabilities, which agrees with our theoretical results. On the other hand the Standard interval has too small actual coverage probabilities which behave unpredictable as a function of p or n having α fixed. Wilson's interval has the actual coverage probability closest to the nominal coverage probability among this three confidence interval estimators.

Biostatistics and Bioinformatics

Goodness of fit test for regression models based on pseudo observations

Klemen Pavlič[1], *Torben Martinussen*[2] *and Per K. Andersen*[2]

[1]Institute for Biostatistics and Meidcal Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

[2]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

klemen.pavlic@mf.uni-lj.si, tma@sund.ku.dk, pka@biostat.ku.dk

Regression models for survival time are often formulated on the hazard scale even though a mean value parameter is of primary interest. The main disadvantage of such models is that the relation between a given set of covariates and the hazard function does not directly translate to the relation between the mean value parameter of interest and a given set of covariates. To overcome this difficulties, models based on pseudo-observations were proposed. Such models have some underlying assumptions that should be checked before valid inference could be made and results could be interpreted. We present a technique for checking some of the underlying assumptions for a general set of mean value parameter models. This approach is based on the cumulative sum of pseudo-residuals. The distribution of this process can be approximated by some Gaussian process and this allows us to plot it together with some realizations of this Gaussian process and visually assess the misspecification. To obtain a more objective measure, this process can be supplemented with a supremum test. In the sequel, we focus on models for survival probability, restricted mean lifetime, cumulative incidence function for a given cause and years lost due to a specific cause and show how their respective processes can be constructed.

Relative survival analysis: comparison of net survival estimators on simulated data

Nina Ružić Gorenjec and Maja Pohar Perme

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

nina.ruzic.gorenjec@mf.uni-lj.si, maja.pohar@mf.uni-lj.si

In the field of relative survival, we are analyzing the survival of patients with a certain disease in the presence of other possible causes of death. The most common measure for comparison of the burden of a disease between different populations, with different hazards for death from other causes, is net survival which is the survival in the world where the only possible cause of dying is the studied disease. There are many estimators used for net survival in the literature, but recently introduced Pohar Perme estimator is the only consistent one. Nevertheless, based on simulation studies, some authors claim that otherwise biased (age-standardised) Ederer II estimator is still preferable based on mean squared error.

We are comparing Ederer II and Pohar Perme estimators on simulated data. Our simulation design follows a recent work in the field where so-called realistic scenarios for simulations were used. First, we show the patterns of the hazard functions used in this simulations. Then we compare Ederer II and Pohar Perme estimators on data simulated from three different patterns of hazard functions. We comment on the problems in net survival estimation, and shed light on the advantages and disadvantages of the compared estimators.

Survival analysis on the duration of MeSH terms: preliminary results

Andrej Kastrin and Dimitar Hristovski

Institute of biostatistics and medical informatics, Faculty of medicine, University of Ljubljana, Ljubljana, Slovenia

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

MEDLINE is the main and largest literature database in the biomedical domain. MEDLINE records have been manually annotated using the MeSH vocabulary, a controlled vocabulary thesaurus consisting of biomedical terms at various level of complexity. We explore how the temporal characteristics of the MeSH terms can be used to provide insight into the historical evolution of the scientific thought. We propose an approach based on the Kaplan-Meier estimator to study how dead MeSH terms affects evolving MEDLINE distribution. The survival theory assumes that death happens just once for each subject in the analysis. In order to define terminal event we have made the assumption that a MeSH term is considered to be inactive if it has appeared less than n times per year. If this happens then the MeSH term is considered inactive for this year. If for the next year the MeSH term in inactive again, then it is considered abandoned and dead. This study demonstrates how survival analysis can be applied to text mining field. To the best of our knowledge this is the first such analysis performed on documents from biomedical domain. The analysis we presented has two main drawbacks. The first one is to expand the analysis to whole MEDLINE domain. The second issue is the existence of sporadic biomedical concepts, i.e., MeSH terms stopping and then starting after some time. However, inclusion of such MeSH terms is a very challenging problem.

Other Areas of Statistics

Statistics and financial mathematics competition for (high school) students in Slovenia – past advances and future challenges

Aleš Toman

University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia
ales.toman@ef.uni-lj.si

Mathematics competitions for primary and high school students have a long history in Slovenia with the first secondary school competition organized in 1950 and the first primary school competition organized in 1965. Ever since, the competitions have broadened to include students at different education levels, specialized into more specific disciplines, and became part of an international competitions scheme.

The series of competitions in business mathematics started in Slovenia in 2003 and competitions in statistics started in 2012. Both competitions were and still are organized by The Society of Mathematicians, Physicists and Astronomers of Slovenia (DMFAS) and only the students enrolled into one of the four specific business related secondary school programs are allowed to participate. In 2014 DMFAS organized the first statistics and financial mathematics competition intended for all secondary school students. The competition syllabus extends the basic high school mathematics syllabus and the vast majority of students participating at the competition so far were high school students. The competition is organized on a school and national level.

In this presentation we will outline the aims of the competition, summarize the structure of the problems and analyze students' performance with respect to topics covered in individual problems. We will see that the performance is good at the problems related to the core high school mathematics syllabus and quite poor at the problems where students were asked to study additional materials – quite independently of the complexity and specificity of the questions asked. Modifying and improving the materials offered to students and their mentors have not changed the situation. Additional challenges for the future of the competition will also be presented. Stagnating number of schools and participants will be on top of the list. Comments and suggestions from the audience will be highly appreciated.

Statistical disclosure control for Census 2021: feasibility study

Junoš Lukan

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
junos.lukan@gov.si

European Commission regulations prescribe the technical specifications of the topics and breakdowns of the data to be published for the Census 2021. Some are planned to be published as tabular data or hypercubes, defined in Annex 1 of Regulation C(2017)2433, with topics established in Annex 1 of Implementing regulation C(2017)1728. Others are going to be presented on a grid map with a cell 1 km^2 of size as proposed in the Annex I of Task Force's ESTAT/F2/CENS/2016/06 and 07. Regardless of the dissemination practice, some published data will contain geographical variables. This requires specific methods of statistical disclosure control. The UK's Office of National Statistics (ONS) developed two methods intended for this purpose, which are meant to be used concurrently. First, record swapping is applied on microdata. Records are marked as high risk according to a set of specified variables and the corresponding households that are determined to be similar are then swapped between different geographical areas, such as territorial (NUTS) and local administrative units (LAU). Next, a perturbation method is applied to the cell in a table or on a grid. Perturbation is random, but chosen according to cell keys, which ensure that the same cell in different tables is perturbed identically. Within the scope of the Harmonization Protection of Census Data in the ESS, these two methods were applied to Slovenian Census 2011 data on chosen hypercubes and on 1 km^2 grid in order to test their feasibility. The protected values resulting from swapping and perturbation were compared to the original values, using different measures of distance. The test results were analysed with respect to the parameters of the methods and their spatial distribution.

Electronic data reporting in short-term business statistics in Slovenia

Nina Češek Vozel

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
nina.cesek-vozel@gov.si

Due to the efficient use of resources there is a need to modernize and rationalize statistical processes. Modernization of data collection is one of the key elements. This paper illustrates the practical implementation and outcome of introducing online surveys in short-term business statistics in Slovenia. Short-term business statistics are of great importance because they enable early detection of changes in economic development. Timeliness and quality of data are significant. The Slovenian statistical office met the challenge of conducting online surveys in 2013. Initial deficiency became the challenge and over the years the technology improved. It became more user-friendly although we still face certain challenges especially when introducing a new online survey because each has its speciality. A lot of the initial effort to replace paper questionnaires with electronic ones reflected in the outcome such as providing the quick response rate when reminding enterprises, fewer errors due to additional data control, but on the other hand different types of errors, standardized statistical processes across similar surveys and in the end lower costs for the Office and enterprises. Though the majority of enterprises send their data electronically, a small percentage still use paper questionnaires and reducing their number still remains our goal for the future. In order to provide good representativeness of the published data, we need to consider all enterprises in the sample due to the fact that the Slovenian economy is relatively small. Since the Slovenian government has been promoting electronic communication with enterprises for quite some time, for the future we expect more enterprises to be ready for electronic data reporting.

Validity and reliability test of wellbeing indicators in Thailand

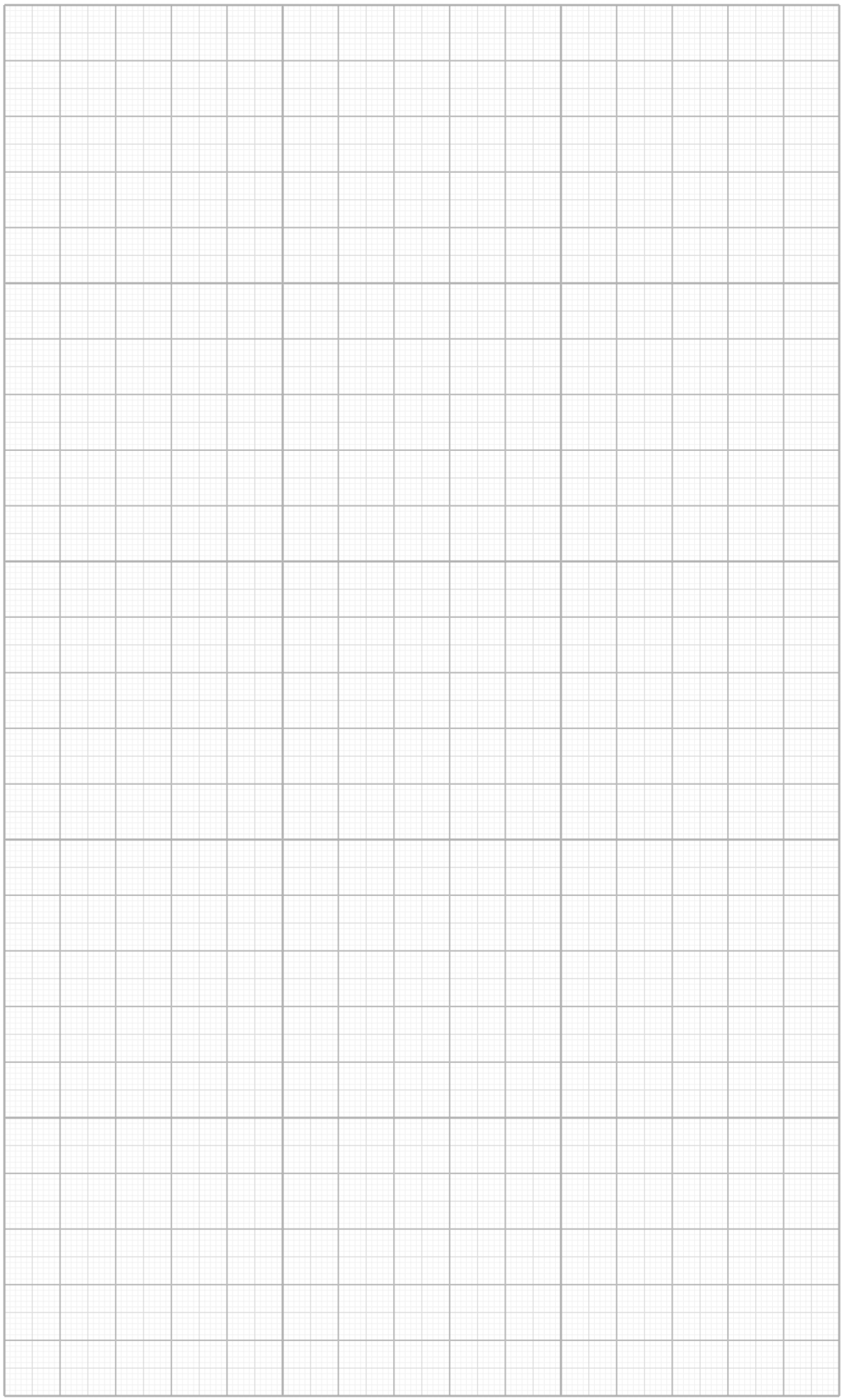
Thanawit Bunsit

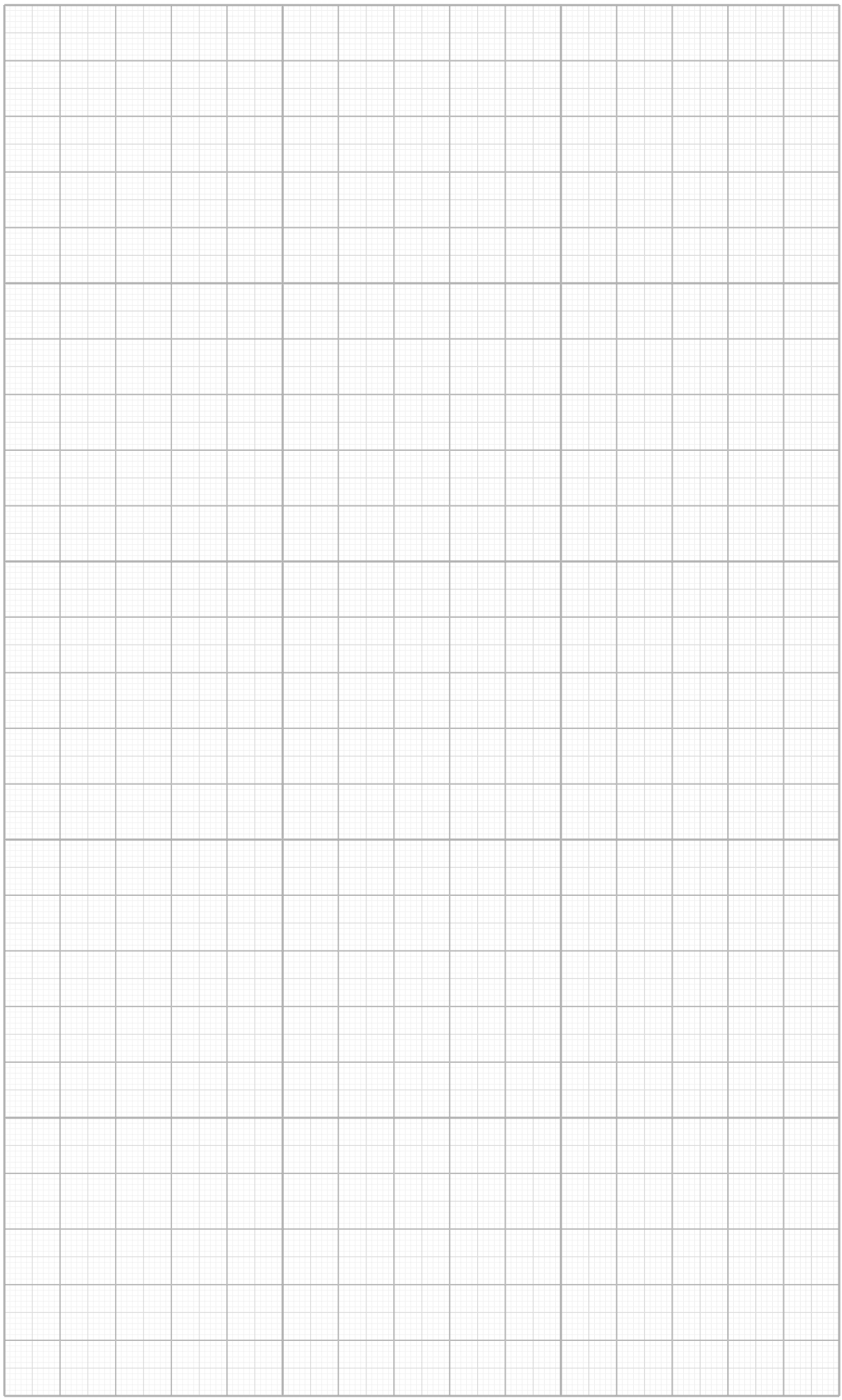
Thaksin University, Songkhla, Thailand
thanawit.b@tsu.ac.th

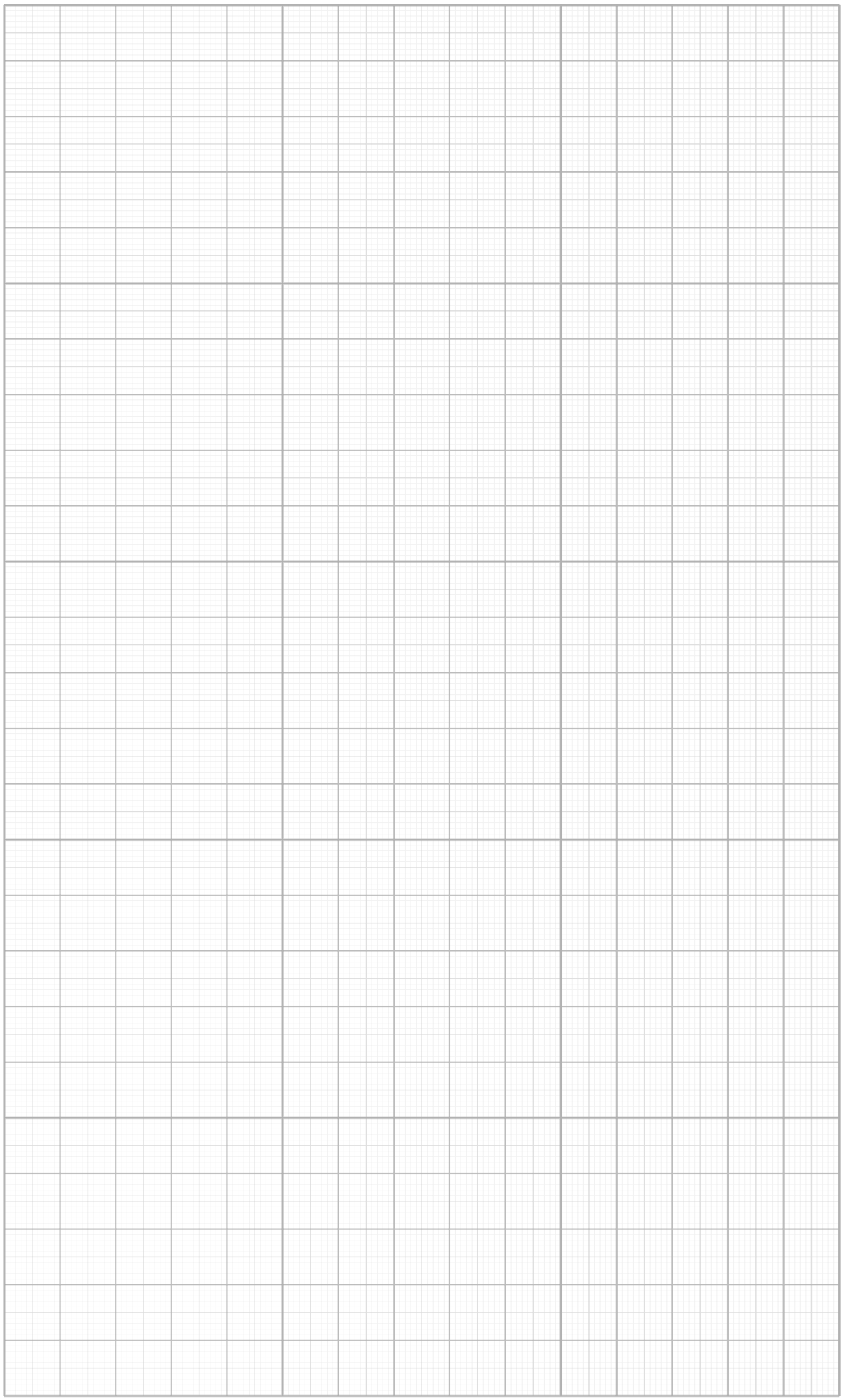
Measuring abstract concepts such as happiness or wellbeing is a challenging task for social scientists. This study aims to examine the validity and reliability of different measures of wellbeing. Wellbeing can be measures using various constructs with different indicators, dimensions and scales. In this study, wellbeing constructs such as positive and negative affects or feelings (PANAS), happiness questions, objective versus subjective indicators were employed in order to capture wellbeing perspective of Thai samples. The dimensionality, validity and reliability were assessed. The results supported the 20-item psychological constructs and 20-item objective wellbeing indicators for measuring wellbeing of people with a high score of Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity. By comparing Cronbach's alpha, Spearman-Brown coefficient and Guttman split-half coefficient showed some constructs exhibited higher reliability in capturing wellbeing characterisation than others. The results provide a useful research tool for measurement in a specific area and a global wellbeing-related scales.

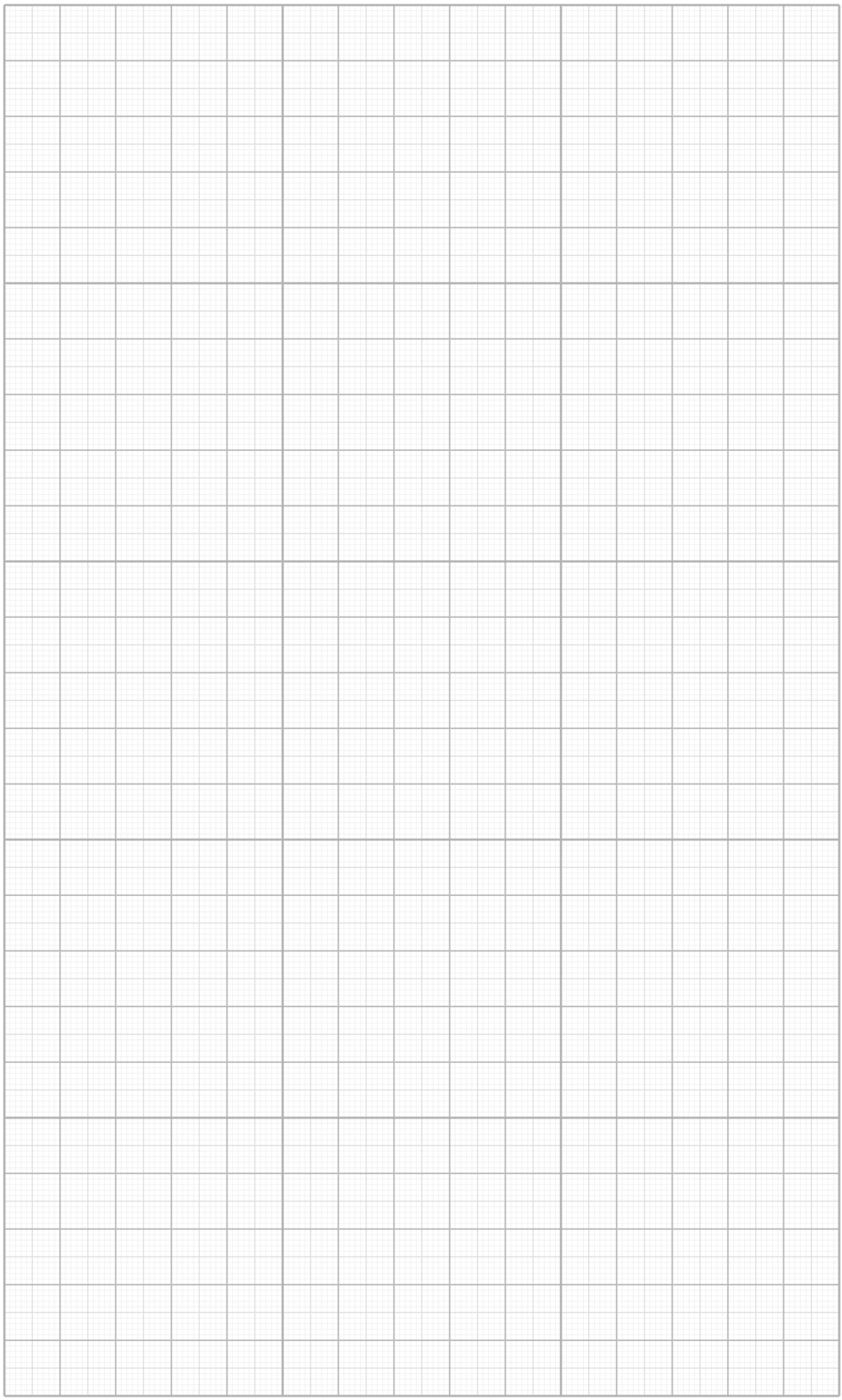
INDEX OF AUTHORS

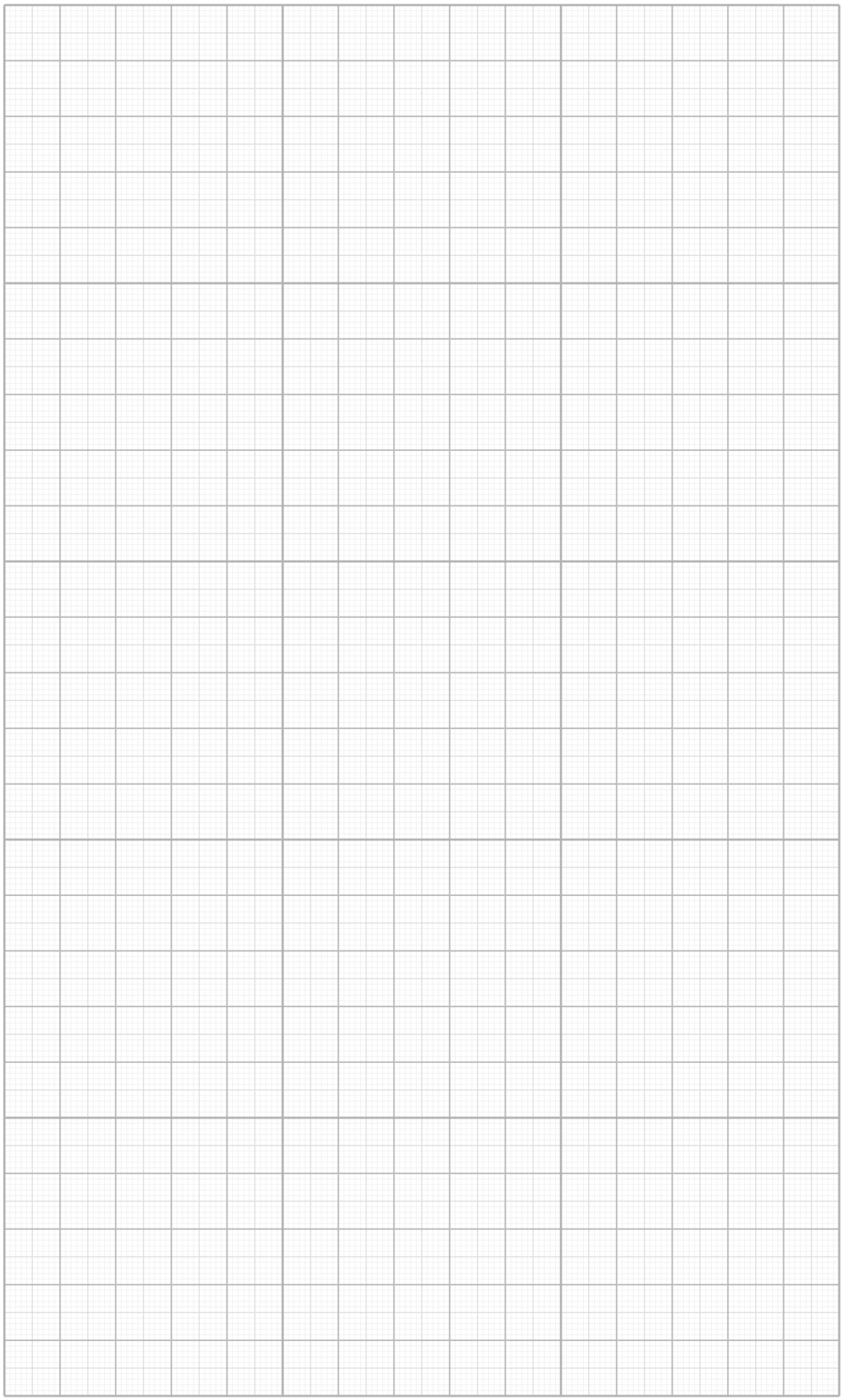
Index of Authors

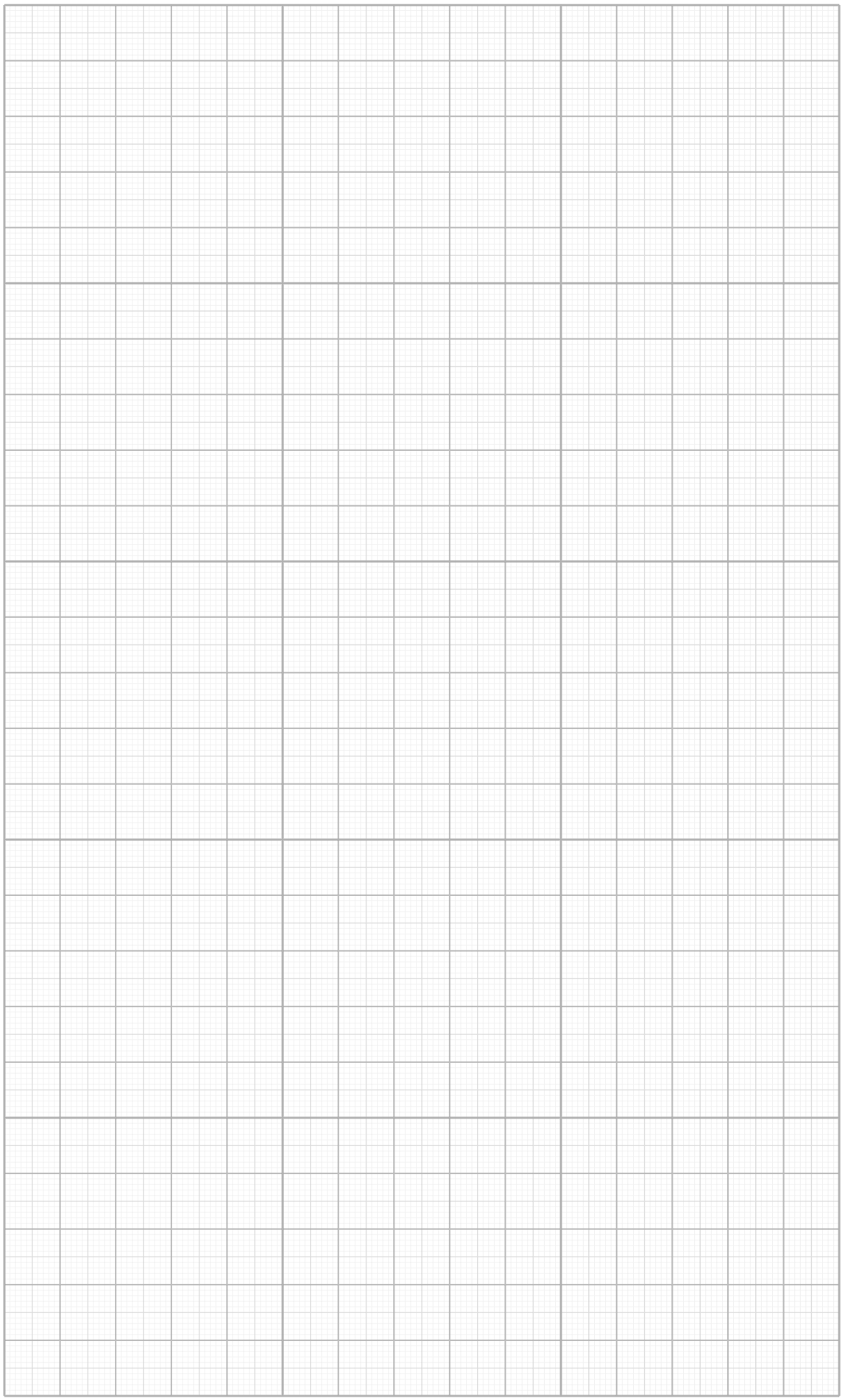


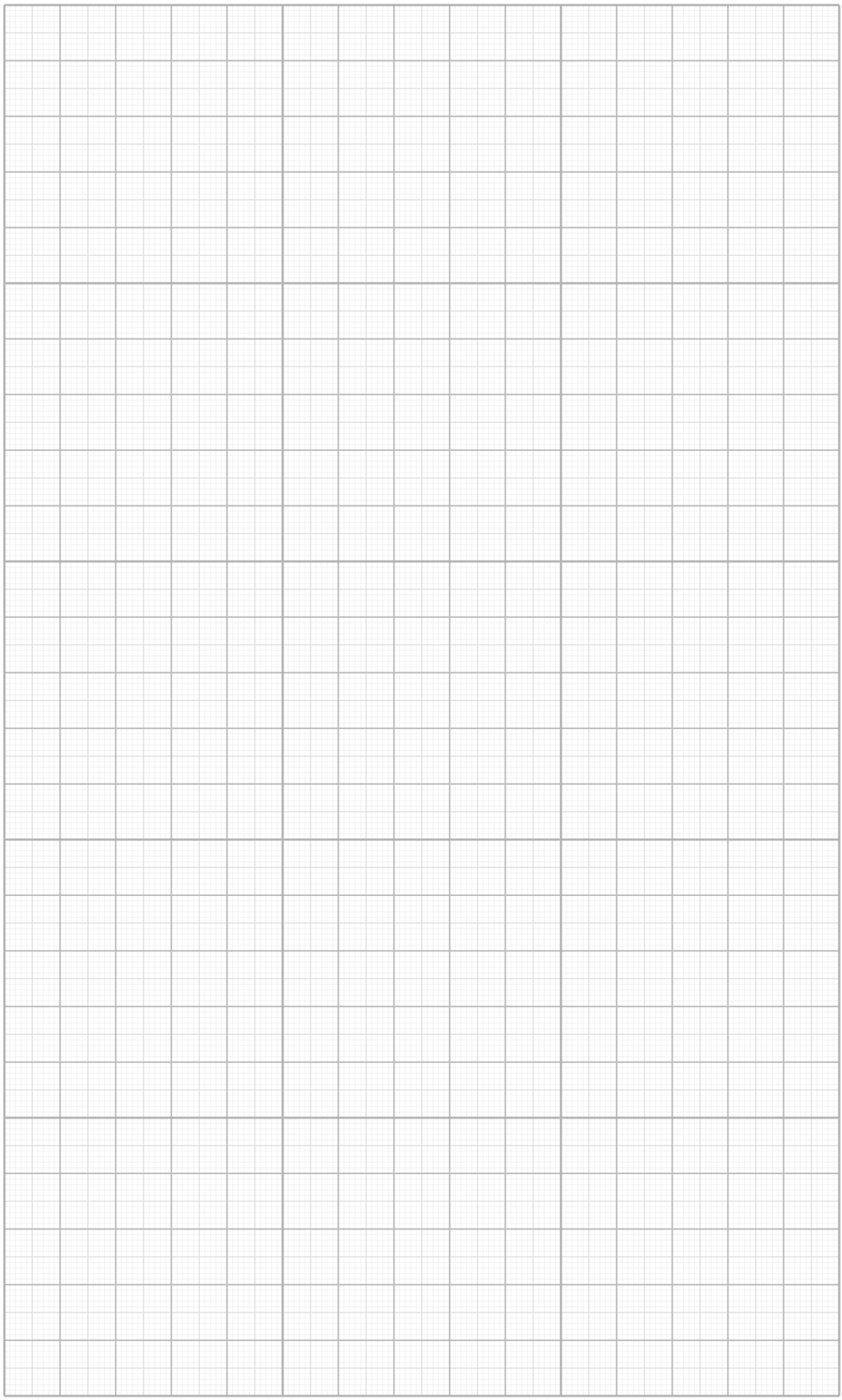












SUPPORTED BY