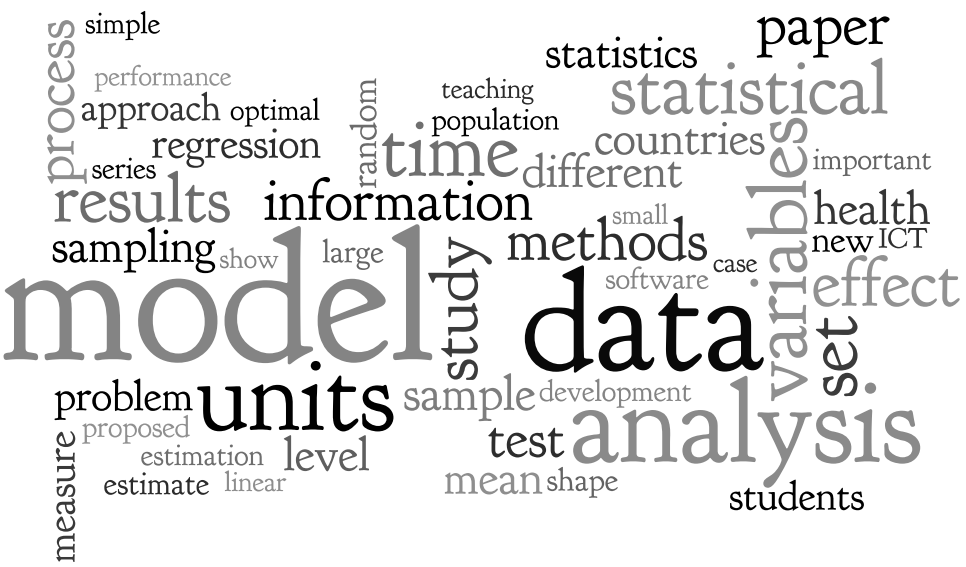
International Conference

APPLIED STATISTICS

2011

PROGRAM and ABSTRACTS



September 25 - 28, 2011
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2011

International Conference

APPLIED STATISTICS 2011

PROGRAM and ABSTRACTS

September 25 – 28, 2011

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2011

Organized by

Statistical Society of Slovenia

Supported by

Slovenian Research Agency (ARSS)

Statistical Office of the Republic of Slovenia

ALARIX

RESULT d.o.o.

VALICON / SPSS Slovenia

ELEARN Web Services Ltd

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(082.034.2)

INTERNATIONAL Conference Applied Statistics (2011 ; Ribno)
Program and abstracts [Elektronki vir] / International
Conference Applied Statistics 2011, September 25–28, 2011, Ribno (Bled), Slovenia
; organized by Statistical Society of Slovenia ; [edited by Lara Lusa
and Janez Stare]. - El. knjiga. - Ljubljana : Statistical Society of Slovenia, 2011

Način dostopa (URL): http://conferences.nib.si/AS2011/AS2011-Abstracts.pdf
ISBN 978-961-92487-7-5

1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
257731328

Scientific Program Committee

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

Organizing Committee

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

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: 200

PROGRAM

		Hall 1	Hall 2
Sunday	10.30 – 11.00	Registration	
	11.00 – 11.10	Opening of the Conference	
	11.10 – 12.00	Invited Lecture	
	12.00 – 12.20	Break	
	12.20 – 13.40	Social Science Methodology	
	13.40 – 15.00	Lunch	
	15.00 – 16.20	Statistical Software	Mathematical Statistics
	16.20 – 16.40	Break	
	16.40 – 18.00	Measurement	Sampling Techniques and Data Collection
	19.00	Reception	
Monday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Statistics in Life	
	11.40 – 12.00	Break	
	12.00 – 13.20	Education	Statistical Applications - Economics
	13.20 – 14.30	Lunch	
	14.30	Excursion	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Biostatistics and Bioinformatics I	
	11.40 – 12.00	Break	
	12.00 – 13.20	Biostatistics and Bioinformatics II	Modeling and Simulation I
	13.20 – 15.00	Lunch	
	15.00 – 16.20	Statistical Applications - Biostatistics	
Wednesday	9.30 – 10.50	Econometrics	
	10.50 – 11.10	Break	
	11.10 – 12.30	Modeling and Simulation II	Network Analysis and Statistical Applications
	12.30 – 12.50	Closing of the conference	
	12.50 – 14.00	Lunch	
	14.00 – 18.00		Workshop

10.30–11.00 **Registration**

11.00–11.10 **Opening of the Conference**

11.10–12.00 **Invited Lecture** (Hall 1) *Chair: Andrej Blejec*

1. **Statistics With a Human Face**
   Adrian Bowman

12.00–12.20 **Break**

12.20–13.40 **Social Science Methodology** (Hall 1) *Chair: Adrian Bowman*

1. **How to Optimally Combine the Fixed and Mobile Sample Frame in a Telephone Survey?**
   Ana Slavec and Vasja Vehovar

2. **A Methodological Framework for a Comprehensive Presentation of Dynamics in Time**
   Katja Prevodnik, Vesna Dolničar, Pavle Sicherl and Vasja Vehovar

3. **Exploratory Structural Equation Model (ESEM) for the Estimation of the Construct Validity. Adaptation of WHOQoL Scale to Iberoamerican Population**
   Joan Guardia-Olmos, Sonia Benítez-Borrego, Alfonso Urzúa-Morales and Maribel Peró-Cebollero

4. **Analyzing Determinants of Health Inequality Among Children in Egypt**
   Nada Abdel Fattah

13.40–15.00 **Lunch**

15.00–16.20 **Statistical Software** (Hall 1) *Chair: Nataša Kejžar*

1. **All of Statistical Software: A Succinct Overview**
   Gaj Vidmar

2. **Integrating R and Excel for Automatic Business Forecasting**
   Giovanni Millo and Fabrizio Ortolani

3. **animatoR: Dynamic Graphics in R**
   Andrej Blejec

4. **Statistical Forecasting of High-way Traffic Jam**
   Igor Grabec and Franc Švegl

15.00–16.20 **Mathematical Statistics** (Hall 2) *Chair: Janez Stare*

1. **Optimal Parameters of EWMA Designs by Integrating Closed Form Formulas and Numerical Integral Equations Methods**
   Saowanit Sukparungsee

2. **Maximum Likelihood Estimation for Ordered Marginal Probabilities of Multivariate Two-Point Distribution**
   Wojciech Gamrot

3. **Explicit Expression of Average Run Length for Exponential CUSUM**
   Yupaporn Areepong

4. **Biased Estimation of Process Capability Indices Using Bootstrap and Jackknife Methods**
   Jeerapa Sappakitkamjorn

16.20–16.40 **Break**

16.40–18.00 **Measurement** (Hall 1) *Chair: Gaj Vidmar*

1. **Transformation of National Income to Gross Domestic Product for the Czech Republic 1970 - 1990**
   Jaroslav Sixta and Jakub Fischer

2. **Multiple Linear Regression Applied to Automatic Target Recognition**
   Gerard Brunet and Abdellah Qannari

3. **The Performance of Forecasting Model for Non Stationary data: Case study: Beer Assumption Model**
   Pathom Glannamtip

4. **Comparison of Pairwise Comparison Methods Under Three Different Variance Levels**
   Krongkaew Wangniwetkul

16.40–18.00 **Sampling Techniques and Data Collection** (Hall 2) *Chair: Vasja Vehovar*

1. **Path Sampling**
   Mena Patummasut and Arthur L. Dryver

2. **MSE Weights of Ratio Estimator in Stratified Random Sampling**
   Vichit Lorchirachoonkul and Jirawan Jitthavech

3. **Statistics Outliers and Remote Sensing Anomalies**
   Jose A. Malpica and Maria C. Alonso

4. **Variable Elimination in Nested DEA Models by the Tukey HSD Procedure**
   Jirawan Jitthavech and Vichit Lorchirachoonkul

19.00 **Reception**

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

1. **The Current Duration (Backward Recurrence Time) Approach to Estimating Time to Pregnancy**
   Niels Keiding

10.00–10.20 **Break**

10.20–11.40 **Statistics in Life** (Hall 1) *Chair: Niels Keiding*

1. **Chasing the Doped Athletes – What can Statistics do About it**
   Maja Pohar Perme

2. **What Medical Researchers Know About Statistics, and is it a Losing Battle to Educate Them?**
   Simon Day and Justyna Stefaniak

3. **Clustering Symbolic Objects Represented With Discrete Distributions**
   Simona Korenjak-Černe, Nataša Kejžar and Vladimir Batagelj

4. **Clustering of Population Pyramids using Wasserstein's Distance**
   Katarina Košmelj and Lynne Billard

11.40–12.00 **Break**

12.00–13.20 **Education** (Hall 1) *Chair: Andrej Blejec*

1. **Exploring the Effects of Information and Communication Technology (ICT) on Educational Achievements**
   Barbara Neža Brečko

2. **Team Teaching With Student – Experimental Study**
   Jerneja Šifrer, Žala Žvab and Matevž Bren

3. **Does Higher Psychometric Score Predict Better Performance in Academic Studies**
   David N. Ben and Tal Shahor

4. **An Application of Teaching for Understanding at the Faculty of Veterinary Science**
   Teresita E. Teran, Omar Cordoba and Augusto Nascimbene

12.00–13.20 **Statistical Applications - Economics** (Hall 2) *Chair: Simona Korenjak Černe*

1. **Maximum Likelihood Estimation of Spatially and Serially Correlated Panels With Random Effects: An Estimation Framework and a Software Implementation**
   Giovanni Millo

2. **Pricing Multi Assets American Options With Monte Carlo Simulation**
   Predrag Popović

3. **Determining Relationship Between R&D and the Market Value: the Case of Turkey**
   Güler Aras, Aslı Aybars, Özlem Kutlu and Nuray Tezcan

4. **Productivity Indicators and their use in Composite Indicators**
   Lenka Hudrlikova and Kristyna Vltavska

13.20–14.30 **Lunch**

14.30 **Excursion**

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

1. **Dynamic Prediction of Survival With Clinical and Genomic Data**
   Hans C. van Houwelingen

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics I** (Hall 1) *Chair: Hans C. van Houwelingen*

1. **Information Growth of Reweighted Cox Model Estimators used in Group Sequential Clinical Trials Under Non-Proportional Hazards**
   Adam P. Boyd

2. **Estimation of Net Survival: Problems and Properties**
   Anamarija Rebolj and Maja Pohar Perme

3. **Fuzzy Bootstraping Aided Extended Mortallity Model of LC-type and its Use to Evaluation of Mortality in Poland**
   Agnieszka Rossa and Andrzej Szymański

4. **Life Tables Model in Life Insurance**
   Abdellah Qannari, Christiana Balan, Serge Sabourin and Gerard Brunet

11.40–12.00 **Break**

12.00–13.20 **Biostatistics and Bioinformatics II** (Hall 1) *Chair: Katarina Košmelj*

1. **Application of a Power Model for Determination of Adventitious Presence of Genetically Modified Organisms in the Case of Maize**
   Katja Rostohar, Andrej Blejec, Vladimir Meglič and Jelka Šuštar-Vozlič

2. **When Vine is not Fine we Fear for Wine**
   Ana Rotter, Petra Nikolič, Kristina Gruden, Andrej Blejec and Marina Dermastia

3. **SMOTE for High-Dimensional Class-Imbalanced Data: A Theoretical and Empirical Analysis**
   Rok Blagus and Lara Lusa

4. **Using Principal Geodesic Analysis on Shape Space**
   Mousa Golalizadeh

12.00–13.20 **Modeling and Simulation I** (Hall 2) *Chair: Aleš Žiberna*

1. **Modelling Functional Relationship Between Longitudinal Data Series**
   Xiaoshu Lu and Esa-Pekka Takala

2. **A Bayesian Analysis of Unemployment Duration Data**
   Mojtaba Ganjali and Taban Baghfalaki

3. **Multi Objective Economic Statistical Design of Control Charts**
   Alireza Faraz, Erwin Saniga and Cédric Heuchenne

13.20–15.00 **Lunch**

15.00–16.20 **Statistical Applications - Biostatistics** (Hall 1) *Chair: Maja Pohar Perme*

1. **Analysis of the Viral Immune Response and Testing the Infection Course Hypothesis**
   Nataša Kejžar, Miša Korva and Tatjana Avšič Županc

2. **The Mammographic Screening Program in Trieste: First Statistical Considerations**
   Fabiola Giudici, Lucio Torelli, Fabrizio Zanconati and Maura Tonutti

3. **How to Measure Patients' Dissatisfaction**
   Mirna Macur

4. **Effect of Repeatedly Self-Assessing the Presence and Severity of Health Symptoms: An Empirical Study**
   Lara Lusa, Daša Stupica and Franc Strle

9.30–10.50 **Econometrics** (Hall 1) *Chair: Giovanni Millo*

1. **Modelling Synergies In Planning Multimedia Activities In Integrated Marketing Communications Perspective**
   Jana Suklan, Vesna Žabkar and Damjan Škulj

2. **Credit Scoring Models and their Quality**
   Jan Kolacek and Martin Rezac

3. **Multidimensional Measures of Happiness and Subjective Wellbeing: Combining Existing Approaches to Develop a New Happiness Model**
   Thanawit Bunsit

4. **Relations between the Czech and German Economies**
   Jakub Fischer and Jana Kramulova

10.50–11.10 **Break**

11.10–12.30 **Modeling and Simulation II** (Hall 1) *Chair: Lara Lusa*

1. **Sample Size in Propensity Score Methods for Estimating Causal Effects**
   Ana Kolar, Vasja Vehovar and Donald B. Rubin

2. **Modeling of Patterns by an Intelligent System**
   Anamarija Borštnik Bračić, Igor Grabec and Edvard Govekar

3. **A Multivariate Markov Modulated Poisson Process Model for Rainfall Intensity**
   Rasiah Thayakaran and Nadarajah I. Ramesh

4. **Is Simple Randomization of Compounds in Training and Test Set as Good as other Methods in quantitative Structure-Activity Experiments?**
   Sorana D. Bolboaca and Lorentz Jäntschi

11.10–12.30 **Network Analysis and Statistical Applications** (Hall 2) *Chair: Matevž Bren*

1. **Blockmodeling of Multilevel Network Data**
   Aleš Žiberna

2. **Networks Generated by Fifa Soccer Games Played between Countries**
   Kristijan Breznik and Vladimir Batagelj

3. **Application of Discriminant Analysis in Investigation of Regional Disparities in Serbia**
   Valentina T. Sokolovska, Katarina J. Čobanović and Emilija Nikolić-Djorić B.

4. **The Impact of Computer Price Indices on Total Factor Productivity Measurement**
   Borut Kodrič and Lea Bregar

12.30–12.50 **Closing of the conference** (Hall 1)

12.50–14.00 **Lunch**

14.00–18.00 **Workshop** (Hall 2)

1. **Statistics of Compositional Data**
   Gerald van den Boogaart

ABSTRACTS

Invited Lecture

Statistics With a Human Face

Adrian Bowman

School of Mathematics & Statistics, University of Glasgow, Glasgow, United Kingdom;
a.bowman@stats.gla.ac.uk

Stereo-photogrammetry provides high-resolution data defining the shape of three-dimensional objects. One example of its application is in the study of facial shape, and indeed of other parts of human anatomy. One particular study aims to describe the facial shape and growth of healthy children and to contrast this with the shape and growth of children who have been born with a cleft lip and/or palate and who have subsequently undergone surgical repair. Information can be extracted in a variety of forms. Methods of analysing landmark shape data are well developed but landmarks alone clearly do not adequately represent the very much richer information present in each digitised face. Facial curves with clear anatomical meaning can be identified. In order to exploit the full extent of the information present in the images, standardised meshes, whose nodes correspond across individuals, can also be fitted. Some of the issues involved in identifying and analysing data of these types will be discussed and illustrated in a variety of surgical and other settings. Statistical issues include how to analyse data objects which express shape, how to measure asymmetry and how to conduct longitudinal modelling.

Social Science Methodology

How to Optimally Combine the Fixed and Mobile Sample Frame in a Telephone Survey?

Ana Slavec[1] *and Vasja Vehovar*[2]

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
[1]ana.slavec@fdv.uni-lj.si
[2]vasja.vehovar@fdv.uni-lj.si

Dual frame sampling is increasingly used in telephone surveys to remove non-coverage bias due to the growing share of households without fixed phone but having at least one mobile phone. A typical dual frame survey has 60-80% of fixed units which is not necessarily optimal – in this paper we aim to find the exact optimal value (mixture parameter) according to survey error and costs. For this purpose, the target population is stratified according to phone use domains. We compare the stratification to 3 domains (mobile-only, overlap, fixed-only) and to 5 domains where the big overlap segment is dissected into two additional strata. By optimizing the product of costs and mean squared error across strata we compute the analytical solution – the roots of a 4th order polynomial. The approach is applied to a 2008 Flash Eurobarometer survey in 8 euro area countries. The optimal mixture parameter is estimated for different variables, countries and cost ratios. The result lies in a relatively flat optimal area from 30-70% and is almost invariant to variables, while the effect of strata weights and especially the cost ratio are quite strong which is tested also with linear regression. In general, countries where the share of the mobile-only segment is not very high), it is optimal to have more fixed units and vice versa: where the mobile-only segment is larger a predominately mobile sample is optimal. Changes in cost ratio, however, are what most importantly determine the optimal allocation: when mobile interviews are much more expensive than fixed, a predominantly fixed sample is optimal in all countries. When we consider the heterogenity of the overlap (5-strata) less mobile units are suggested than in the the case of the ordinary 3-strata segmentation.

A Methodological Framework for a Comprehensive Presentation of Dynamics in Time

Katja Prevodnik[1], *Vesna Dolničar*[2], *Pavle Sicherl*[3] *and Vasja Vehovar*[4]

[1]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
katja.prevodnik@fdv.uni-lj.si
[2]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
vesna.dolnicar@fdv.uni-lj.si
[3]SICENTER, Ljubljana, Slovenia; pavle.sicherl@gmail.com
[4]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
vasja.vehovar@fdv.uni-lj.si

The change in a society should be observed and described with much care, i.e. methodologically correct and with an expert interpretation, taking into account also the guidelines of related ethical codes. Various statistics are nowadays used as benchmarks to identify change in time and to assess the performance of different units (e.g. countries, regions, sociodemographic groups etc.).
The specific challenge addressed here is how to present results of comparative studies by taking into account a static and dynamic aspects simultaneously. Usually, the empirical analyses of compared phenomena focus only on one single measure (i.e. absolute, relative or time differences). However, absolute difference, relative difference and S-time distance provide complementary information on dynamics and sometimes they result even in contradictory conclusions. The question therefore is whether it is possible to present these data in an objective and standardized manner to provide a coherent comparison and interpretation, which would help in avoiding biased or even misleading results and interpretations.
This presentation will provide an overview of the problem and illustrations with the cases when contradictions and inconsistencies among the three measures exist, as well as attempts to shrink/transform the three dimensions into one or two (i.e. composite indicators). General guidelines for future development will be discussed along with some experimental calculations based on the phenomena in the field of information society.

Exploratory Structural Equation Model (ESEM) for the Estimation of the Construct Validity. Adaptation of WHOQoL Scale to Iberoamerican Population

Joan Guardia-Olmos[1] , *Sonia Benítez-Borrego*[2], *Alfonso Urzúa-Morales*[3] *and Maribel Peró-Cebollero*[4]

[1]Depto. de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain; jguardia@ub.edu
[2]Depto. de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain; sbenitez@ub.edu
[3]Escuela de Psicología, Universidad Católica del Norte, Antofagasta, Chile; alurzua@ucn.cl
[4]Depto. de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain; mpero@ub.edu

The emergence of the ESEM is one of the main contributions in recent times in relation to the estimation of factor structures and, therefore, to study the construct validity of measurement scales generated for evaluation of complex traits. As is known, the psychological empirical approach has based much of their budget on the study of factorial solutions for assessing the construct validity and the appearance of confirmatory strategies has complemented the traditional psychometrical studies based on the classical test theory. However, several issues are still offering doubts about the use of these statistical techniques in view of the metric properties of psychometric data. For example, is arguably the use of continuous functions for modeling the items in a Likert scale based on systems or is arguable that the summation of ordinal items discards a good estimate of the trait measured. All this, which is widely known, is still an interesting challenge for the multivariate techniques. ESEM has established a different approach and in this paper we aim to provide the outcome of their application for the psychometric study of the construct validity of the WHOQoL scale adapted to Latin American population in various Spanish-speaking countries.

Analyzing Determinants of Health Inequality Among Children in Egypt

Nada Abdel Fattah

Cairo University, Cairo, Egypt; n_a_abdallah@yahoo.com

Improving health of the poor and reducing health inequalities have become main goals of many national as well as international organizations such as the World Health Organization (WHO) and the World Bank. Children's health is a key health indicator and it appears that there is inequality in the health of children between and within countries.
In Egypt, there is a great health gap between children, and this gap is caused by many direct and indirect factors, so by using path analysis, it can be shown how a series of variables are interrelated. It is therefore often desirable to develop a system of equations, i.e. a model, which specifies all the causal linkages between variables. Also to measure the socioeconomic inequality among children, Relative Index of Inequality (RII), which is defined by Kunst and Mackenbach is used; and to measure economic, social, and demographic statuses, received health services and children's health which are hypothesized to cause children' health inequality, indices were developed from EDHS data using factor scores following factor analysis from a number of measured variables.
In Egypt, although there is decline in children's mortality, there still exists inequality in children's health. Children with higher socioeconomic position in society have greater life chances and opportunities for better life.
The fact that children in Egypt in different circumstances experience avoidable differences in health and wellbeing is unfair. Creating a fairer society is essential, by overcoming the problem of allocating scarce health care resources caused by the existence of tradeoffs between redistribution of health resources and investing in improving the health of the whole population. Also by providing public information campaigns to lower socioeconomic groups to get people adopt healthy life styles. Also by eliminating the real causes behind the existence of children's health inequality caused by low socio-economic and demographic statuses.

Statistical Software

All of Statistical Software: A Succinct Overview

Gaj Vidmar

University Rehabiltation Institute, Republic of Slovenia; Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia;
gaj.vidmar@ir-rs.si

Providing an exhaustive overview, let alone a thorough review, of all statistical software is too large a task for the scope of a book, let alone an article. Furthermore, this refers only to the present-day situation, while describing the interesting and complex history would require a comparably huge amount of words and pictures. Hence, this paper attempts to provide a succinct overview, settling in advance for an imperfect yet useful product. The approach taken is that of a cataloguer – shallow in the eyes of a mathematically minded scientist, but with utility proven through millennia of librarianship and validity grounded in the author's professional credentials. The paper stems from the author's extensive collection of links (ibmi.mf.uni-lj.si/ibmi-english/biostat-center, Selected Links). There are other web resources of similar scope, most notably StatPages.org, Wikipedia's entry on comparison of statistical packages, and an updated online report on popularity of data analysis software. In addition, several statistical journals regularly feature software reviews, most notably The American Statistician. The author of this paper presently counts nearly 300 pieces of obtainable and functioning statistical software. He has seen in operation, briefly tried or seriously applied a good part of all that software, either at work or as a personal hobby. The software ranges from the most general packages to the most specific applications; from statistics in the narrowest sense of an academic discipline to data mining, business intelligence and information visualization that include statistics at their core or at least at their fringes; from stand-alone executables to web applets and spreadsheet add-ins; and over a wide variety of applied fields from geospatial analytics to statistical genetics and bioinformatics. To explain why such variety exists is nearly as gargantuan a task as to describe the variety. Though important, the issue of (direct and indirect) cost (including cost of publicly available packages and of add-ins for proprietary systems) is just one factor. Further insight can be provided from the viewpoints of psychology, social psychology, sociology and microeconomics, whereby the relevant notions include the motives for citations, brand name prestige, subculture and legislation/regulations, as well as political market restrictions in the past that created protracted traditions. A snapshot of statistical software can be obtained through the contents of its dedicated journal – the JSS. Judging by its contents, hardly anything exists outside R. But that is a highly biased picture of software for statisticians rather than of all of statistical software! Analogies like racing cars vs. standard cars clearly illustrate such distinction. So, while the world of software for statisticians may be rightfully and rapidly shrinking, the world of statistical software has been, still is and will, for the reasons outlined above, continue to be much more varied and variable. Therefore, the only tenable conclusion seems to be a paraphrase of last year's paper of Millo: even if there is only one prefe-R-able, R-ecommended and most p-R-omising statistical softwa-R-e, it has hundreds of alternatives (whereby how viable, commendable and/or widespread these alternatives are, is a different issue).

Integrating R and Excel for Automatic Business Forecasting

Giovanni Millo[1] *and Fabrizio Ortolani*[2]

R&D Department, Assicurazioni Generali S.p.A., Trieste, Italy
[1]Giovanni_Millo@generali.com
[2]Fabrizio_Ortolani@generali.com

We present a simple exercise in bridging the gap between statistics and everyday business practice, based on two powerful tools already available in the R system: the forecast package (Hyndman, R. J. (2011). forecast: Forecasting functions for time series. R package version 2.13) for automatic time series forecasting and the RExcel add-in for MS Excel (Baier, T. and E. Neuwirth (2007). Excel :: Com :: R. Computational Statistics 22(1), 91–101.) allowing to embed R functionality into spreadsheets and to interact with their built-in macro language. The application we developed makes forecasting practice accessible to those who are not familiar with statistical programs and, possibly, do not even have a sound statistical background. Many processes inside the firm involve forecasting. Some build on models and relationships between balance sheet items, but sometimes an a-theoretical extrapolation of past tendencies is needed. As few firms can afford to have trained statisticians dedicated to supply-chain forecasting and the like, budgeting and other activities are often based on simple, heuristic extrapolation of past data. It is commonplace, especially in small enterprises, to "pick last year/month's value", either in terms of stocks or of increments, as the best estimate for the coming period. Fully automatic forecasting of time series, based on model fitting and model comparing algorhythms selecting the 'best' model for the data at hand, provides a statistically well founded solution to the forecasting problem and can be of great use to the firm in obtaining accurate predictions for variables like sales, commodities' input needs and the like, where forecast errors cost money. Such fully automatic procedures are implemented in a variety of commercial software. We show how an open-source solution is also very easy to set up.

animatoR: Dynamic Graphics in R

Andrej Blejec

National Institute of Biology, Ljubljana, Slovenia; Andrej.Blejec@nib.si

Graphics, especially dynamic graphics, is an impressive tool in various demonstrations. In statistics teaching, there are many situations where graphics with animation of certain elements of the picture communicate the concepts in obvious way.

Since R graphic devices are in a sense static, several approaches towards dynamic graphics are used. On many occasions, one would like to move certain graphical element, for example one point, on otherwise static picture. One way is to hide the point by re-plotting it in exclusive OR (XOR) mode and plotting the point in a new position. This method can be fast since one is plotting only the elements that are changing on the otherwise static background which can be very complex. R graphic devices are not suitable for such technique. Another way, which is close to this technique, is hiding the dynamic elements by re-plotting them in the background color. This works only for pictures with solid single color background and proves to be unsatisfactory. Another technique is to simply plot a series of complete pictures, each one with relocated picture elements. If the pictures are not very complex, R is fast enough (if not too fast) for producing a flicker free dynamic impression. This is the most popular technique, which can provide satisfactory results.

To get an impression of smooth movement, the changes in successive pictures should be small and one needs to get many intermediate point or line positions. Here we provide technique and a set of functions that complement base graphics function for production of dynamic graphics. The basic idea is to define the starting and finishing coordinates of moving picture elements (points, lines, segments, etc.). Then we plot a series of pictures for successive intermediate positions, which are calculated using homotopy between start and end values. If start position is x_0 and end position is x_1 than positions between them can be determined as

$$x_t = x_0(1-t) + x_1 t, \qquad t \in [0,1]$$

for different values of homotopy parameter t. Selection of suitable sequence for homotopy parameter t provides an impression of smooth movement along trajectories from starting to finishing positions. Functions in a package `animatoR` are using homotopy for production of smooth dynamic graphics, with the motive of presentations and use in statistics teaching. In the presentation, several examples that demonstrate the use of dynamic graphics in statistics teaching will be shown.

Statistical Forecasting of High-way Traffic Jam

Igor Grabec[1] *and Franc Švegl*[2]

Amanova doo, Technology Park 18, Ljubljana, Slovenia
[1]igor.grabec@amanova.si
[2]franc.svegl@amanova.si

Traffic flow on high-ways is subject to inherent dynamic instability that leads to evolution of congestions and jams. The instability can also be excited by changes of roads infrastructure such as bottlenecks that are installed due to maintenance works. In order to provide for an optimal installation of a bottleneck and to offer information for traffic participants about the corresponding disturbance in advance, the road maintenance operators have to forecast the length of traffic jam that would probably develop at certain time interval in front of the bottleneck. The article presents a non-parametric statistical model that has been recently developed for this purpose. As the input to the model forecast data of traffic flow field in Slovenia are utilized. A new fundamental diagram of traffic flow is formulated by which the flow through the bottleneck and the evolution of the traffic jam at a selected place can be predicted. Performance of the corresponding computer program is demonstrated by forecasting evolution of rush hour traffic jam in front of a bottleneck installed at the point of maximal traffic activity on a high-way close to Ljubljana.

Mathematical Statistics

Optimal Parameters of EWMA Designs by Integrating Closed Form Formulas and Numerical Integral Equations Methods

Saowanit Sukparungsee

Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand; swns@kmutnb.ac.th

Typically, the performance of control chart is frequently measured by Average Run Length (ARL) and Average Delay Time (ADT) when processes are in-control state and out-of-control state, respectively. Using of Statistical Process Control (SPC) charts play a vital role for detecting small changes, i.e. Exponentially Weighted Moving Average (EWMA), particularly, in disorder and surveillance problems. The objective of this paper is to enhance the numerical algorithm for finding optimal parameters of two-sided Gaussian EWMA procedure by integrating the closed form formulas based on martingale technique and numerical integral equations with Gauss-Legendre rule. However, the proposed numerical algorithm is developed from Sukparungsee and Areepong (2009) is robust to optimal parameter of two-sided Gaussian EWMA chart. The numerical results are compared with the results obtained from Monte Carlo simulation which they perform in good agreement in accuracy terms but the latter is very time consuming.

Maximum Likelihood Estimation for Ordered Marginal Probabilities of Multivariate Two-Point Distribution

Wojciech Gamrot

University of Economics in Katowice, Katowice, Poland;
wojciech.gamrot@ue.katowice.pl

A multivariate Bernoulli distribution that incorporates possible dependencies between individual variables is considered. Auxiliary information taking the form of simple ordering constraints on marginal probabilities is assumed to be accessible. This information is utilized to improve estimates of marginal probabilities by maximizing the likelihood function for corresponding multinomial distribution with respect to suitably chosen constraints on multinomial parameters. Asymptotic properties of the proposed estimator are studied analytically. Some results of a simulation study are also presented.

Explicit Expression of Average Run Length for Exponential CUSUM

Yupaporn Areepong

King Mongkut's University of Technology, North Bangkok, Bangkok, Thailand;
yupaporna@kmutnb.ac.th

The Cumulative SUM chart (CUSUM) is widely used in a great variety of practical applications such as finance and economics, medicine, engineering, psychology, signal processing, and in other areas. A common characteristic used for comparing the performance of control charts is Average Run Length (ARL) - the expected number of observations taken from an in-control process until the control chart falsely signals out-of-control. An ARL will be regarded as acceptable if it is large enough to keep the level of false alarms at an acceptable level. A second common characteristic used for comparing performance is traditionally called Average Delay (AD) time - the expected number of observations taken from an out-of-control process until the control chart signals that the process is out-of-control. Ideally, the AD time should be small as possible. The ARL is usually computed via Markov chain, Monte Carlo simulations or numerical integral equations approaches. In not many cases the solution for the ARL can be found in closed form. In this paper we use the integral equation method to derive analytical solutions for the ARL when CUSUM is employed. We derive the ARL for CUSUM chart assuming that the random observations are iid exponentially distributed. Checking the accuracy of results, we found an excellent agreement between numerical solutions and the closed form expressions.

Biased Estimation of Process Capability Indices Using Bootstrap and Jackknife Methods

Jeerapa Sappakitkamjorn

Department of Applied Statistics, King Mongkut's University of Technology, North Bangkok, Bangkok, Thailand; jsj@kmutnb.ac.th

In this paper, two resampling techniques known as the bootstrap and the jackknife methods are used to estimate the bias of estimators of process capability indices (PCIs). It has been proven that several widely used estimators of indices produce biased estimates. Due to the fact that a large bias indicates a poor performance of an estimator, it is of interest in this study to estimate the magnitude of the bias incurred particularly when the estimates are used to assess the performance of non-normal processes. It is known that the process capability indices are designed to evaluate the process performance under the assumption that the underlying process is normally distributed. In practice, however, this assumption is often violated. The indices are frequently estimated using non-normal or skewed process data. To investigate the effect of non-normally distributed processes and sample sizes on the bias of estimators, a simulation study was carried out using six distributions, a normal and five non-normal—Beta distribution, some bell-shaped (e.g., Student's t and Weibull) and right-skewed (e.g., Gamma and Chi-square) distributions at various sample sizes. Results from the simulation study show that when the underlying processes are non-normal and sample sizes are small, the bias of estimators is significant. Nevertheless the bias is substantially reduced when large sample sizes are used.

Measurement

Transformation of National Income to Gross Domestic Product for the Czech Republic 1970 - 1990

Jaroslav Sixta[1] *and Jakub Fischer*[2]

University of Economics, Prague, Czech Republic
[1]sixta@vse.cz
[2]fischerj@vse.cz

The paper deals with the development of gross domestic product (GDP) of the Czech Republic between the years 1970 and 1990. Official statistical series of gross domestic product (GDP) for the Czech Republic start in 1990; longer series of GDP will not be officially published. Our three-year project is aimed at the transformation of national income prepared according to methodology of Material Product System (MPS) to gross domestic product based on the System of National Accounts (SNA). When estimating GDP we face the several problems relating to data (lots of data were lost), disintegration of former Czechoslovakia and differences in methodology. Our estimates are based on existing figures in MPS methodology, input-output tables and existing comparisons of Czechoslovakia with Western countries (prepared during the socialist period).

Multiple Linear Regression Applied to Automatic Target Recognition

Gerard Brunet[1] *and Abdellah Qannari*[2]

University of Poitiers, Niort, France
[1]gerard.brunet@univ-poitiers.fr
[2]abdellah.qannari@univ-poitiers.fr

The purpose of this project is to build a fast and reliable software for recognition of objects in aerial images. To perform analysis of real images, semi-automatic method is used, in a supervised way, with an human viewer and analysis of a flow of images. The goal is to perform it in an efficient and rapid way. After feature extraction and edge detection a set of parameters is extracted and a set of potential targets is constituted to solve the correspondence problem. Preliminary steps to determine the best conditions of image analysis use multiple linear regression, with SAS (Statistical Analysis System), SAS/IML package, and use of macro-functions. To implement a practical system, concurrent programming uses threads with Java language. The software was applied to 400*400 pixels images in sequence. Real images have been tested to evaluate the performance concerning translation, rotation, zoom, and random noise addition.

The Performance of Forecasting Model for Non Stationary data: Case study: Beer Assumption Model

Pathom Glannamtip

Department of Applied Statistics, Faculty of Applied Science King Mongkut's University of Technology North Bangkok, Bangkok, Thailand; pathom@kmutnb.ac.th

This paper has aimed to compare the forecasting performance for Non Stationary data. The volume of Beer assumption in Thailand was use for model test. The monthly data from January 2000 to November 2009 were corrected by Thai national bank are samples of this research. I was focused to compare the performance of forecasting model with 4 methodologies are Autoregressive (AR), Seasonal Integrated Autoregressive (SAR), Autoregressive Conditional Heteroscedasticity (ARCH) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH). Beer assumption data was tested by Unitroot Stationary Data test. The result has shown that beer assumption model is non-stationary. Dot plot, linear plot, ACF, PACF and Mean Square Error (MSE) were used as comparison tools of the performance's forecasting model by R programming. The results has found that, SAR (2) is the best performance forecasting model (MSE= 526.6885).

Comparison of Pairwise Comparison Methods Under Three Different Variance Levels

Krongkaew Wangniwetkul

Department of Applied Statistics, Faculty of Applied Science, King Mongkut' s University of Technology North Bangkok, Bangkok, Thailand; kws@kmutnb.ac.th

In order to expanding the knowledge of the controlling the normal populations which have variance differ at the three levels; small, medium and large, by the value of noncentrality parameter criteria by Game, Winkler and Robert (1972). At each level of different variance, this study also compares the probability of Type I error and the power of test of the four pairwise comparison methods: Tamhane's T2, Dunnett's T3, Games-Howell and Dunnett's C, by varying the sample sizes : 10, 15, 20, 25, 30, 35, 40, 45 and 50, and the difference of the mean of the controlling normal population: 0% 10%, 30% and 50%. Monte Carlo methods were used again to generate responses based on the sample size and also the difference of the mean 1,000 times. Hypothesis testing in each case was conducted at 0.01 significant level. Every response was confirmed difference of samples variance by Levene's test. In each of the cases of the three different variance levels, it was found that every test was under the capability controlling type I error by using Binomial test. To be concerned about the probability of type I error estimations, small samples had slightly higher trend than the large and medium variance levels at their highest. Games-Howell's test had get quite a bit of fluctuation and Dunnett's C's test had get the least. In the case of probability of the power of the test, they were varied by sample sizes. Whole cases, Games-Howell's test had the highest power of test and Dunnett's C's test had get the lowest in each case. In each case the mean difference and the difference variance level were correlated with the power of the test. Whole sample size cases of the 10% mean difference case, the power of the test of the four methods had been less than 0.9 in the small different variance level and was less than 0.4 in the medium and the large. Very large sample size should be taken for testing in small mean difference and large difference variance levels.

Sampling Techniques and Data Collection

Path Sampling

Mena Patummasut[1] *and Arthur L. Dryver*[2]

[1]School of Applied Statistics, National Institute Development Administration, Bangkok, Thailand; mena_patummasut@yahoo.com
[2]School of Business Administration, National Institute Development Administration, Bangkok, Thailand; dryver@gmail.com

Recently, many sample survey methods have been applied to natural populations in a purpose to estimate total numbers. The population study area is divided up into spatial (squared) units of generally the same size, and the numbers of interest are counted on a selection of the units. Many sampling designs can be used, for example, simple random sampling and cluster sampling. In simple random sampling, the sample consists of n units randomly selected from the N units in the spatial population. Each unit has equal chance of selection. In cluster sampling, a primary unit, which is a sampling unit, consists of a cluster of secondary units, usually in close proximity to each other. In the spatial setting, primary units include spatial arrangements as square collections of adjacent units. A simple random sample of m primary units is taken from M primary units in the population. Simple random sample and cluster sample may coverage all over the region since each sampling unit has equal chance of selection. Unfortunately, traveling from place to place to sample every unit selected can be costly as the distance traveled can be quite large. Therefore, path sampling is a new sampling design proposed in this paper to overcome this disadvantage. Path sampling is a sampling design in which p distinct paths are selected by simple random sampling from the P paths in the population, and the sample consists of all units in the selected paths. A path is defined as the course of sampling from the starting unit to the finishing unit. Path sampling utilizes all the observations over the units traveled. Thus, when the main cost of sampling a unit is the distance traveled, path sampling is a very cost effective design. An estimator of the population total and its variance are derived.

MSE Weights of Ratio Estimator in Stratified Random Sampling

Vichit Lorchirachoonkul[1] *and Jirawan Jitthavech*[2]

School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand
[1]vichit@as.nida.ac.th
[2]jirawan@as.nida.ac.th

The weights of a combined ratio estimator in stratified random sampling is developed by minimizing the MSE of the estimator of the interested variable y without using the stratum weight based on the number of units in the stratum subjected to the sum of the weights equal to 1. The relative efficiency of the new estimator is shown empirically higher than the classical combined ratio estimator and various existing estimators in stratified random sampling by using the simulation data.

Statistics Outliers and Remote Sensing Anomalies

Jose A. Malpica[1] *and Maria C. Alonso*[2]

Mathematics Department, Alcala University, Madrid, Spain
[1] josea.malpica@uah.es
[2] mconcepcion.alonso@uah.es

Objects or materials whose signatures are spectrally distinct from their background are known as anomalies in remote sensing. Anomalies are important features of special interest to image analysts in their daily routines. Methodologies and algorithms are needed to identify these atypical features, thereby allowing image analysts to decide whether to retain them as interesting information or whether to classify them as noise, and remove them. We consider the detection of outliers in the feature space as equivalent to detecting anomalies in hyperspectral images. This work focuses on the very first step in the analysis of an image, the point at which one assumes no prior knowledge about the statistical characteristics of the pixels in the image and where little or nothing is known about the size and shape of the objects to be detected. Therefore, the only available option is to look for a point (or group of points) that deviates so much from other points as to arouse suspicion that it was generated by a different mechanism. This project does that by looking for one-dimensional projection (projection pursuit) optimizing some measurement of interest (index). This work analyzes and compares indexes skewness and kurtosis with the popular RX index. The optimization for the one-dimensional projection is performed with a genetic algorithm. The proposed algorithms are tested in synthetic images and in a hyperspectral imagery. It is showed how these algorithms are superior to RX.

Variable Elimination in Nested DEA Models by the Tukey HSD Procedure

Jirawan Jitthavech[1] *and Vichit Lorchirachoonkul*[2]

School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand
[1]jirawan@as.nida.ac.th
[2]vichit@as.nida.ac.th

A novel statistical procedure for backward elimination of input and output variables in nested DEA models is developed by multiple comparisons of DMU efficiency scores in nested DEA models with DMU efficiency scores in the full DEA model, referred as the reference model, which consists of all predetermined input and output variables. The Tukey HSD procedure is shown analytically to be more effective in detecting the change in DMU efficiency scores among nested DEA models by testing the differences of DMU efficiency scores than by testing the DMU efficiency scores directly. The strategy of backward elimination is to eliminate either one input or output variable at a time by the Tukey HSD procedure testing the mean of differences of efficiency scores of the same DMU in the reference model and in the reduced model. The simulation results show that the proposed procedure outperforms the method suggested in the literature.

Invited Lecture

The Current Duration (Backward Recurrence Time) Approach to Estimating Time to Pregnancy

Niels Keiding

Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark;
nike@sund.ku.dk

Time to pregnancy is the duration from a couple starts trying to become pregnant until they succeed and is considered one of the most direct methods to measure natural fecundity in humans. Statistical tools for designing and analysing time to pregnancy studies belong to the general area of survival analysis, but several features require special attention. Prospective designs are difficult to carry out and retrospective (pregnancy-based) designs, being widely used in this area, do not allow efficiently including couples remaining childless, cf. Scheike and Keiding (2006), Scheike et al.(2008).
A third possible design starts from a cross-sectional sample of couples currently trying to become pregnant, using *current duration* (backward recurrence time) as basis for the estimation of time to pregnancy/end of pregnancy attempt (Keiding et al., 2002). Regression analysis is then most conveniently carried out in the accelerated failure time model (Yamaguchi, 2003, Keiding et al., 2011). This paper surveys some practical and technical-statistical issues in implementing this approach in practice in a large telephone-based survey, the *Observatoire de la fertilité en France (OBSEFF)* (Slama et al., 2006).
It is not trivial to recover the current duration of a pregnancy attempt from a telephone interview with the woman. I focus on two problems:

- It is hard to distinguish between couples who do not use prevention because they want to become pregnant and couples who know or believe that they are unlikely to succeed anyway. This issue is particularly important because of the length-biased nature of the sampling of current durations.

- There are several causes of stopping attempts that compete with the desired endpoint, the pregnancy. First, there may be several reasons for the couple to give up trying; and secondly, the couple may seek medical fertility treatment. The statistical treatment of these competing risks has required further development and interpretation of the current duration approach.

The statistical inference is quite sensitive to observations near zero; indeed Woodroofe and Sun (1993) pointed out that the nonparametric maximum likelihood estimator is inconsistent at zero. Switching to parametric models, we are faced with balancing between stability or flexibility. The former represented by a simple mixed exponential model such as the Pareto distribution, the latter by the class of generalized gamma distributions studied by Farewell and Prentice (1977) and applied to backward recurrence times by Yamaguchi (2003) in a sociological context. A simulation study illustrates these issues.
The study design of OBSEFF includes two follow-up interviews with the women. This allows alternative estimation of time to pregnancy using the *prevalent cohort* approach: follow-up of couples already trying until success or right censoring at the next interview. This would provide practical validation of the results obtained by the current duration approach.
The results reported in the talk were developed in my long-standing collaboration with Rémy Slama (Inserm-Grenoble) and his group and Oluf Hansen (Copenhagen).

References

Statistics in Life

Chasing the Doped Athletes – What can Statistics do About it

Maja Pohar Perme

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

Doping is without doubt the most severe problem in the world of sports and the fight against it is the main agenda of all international sport's organizations. Athletes are under constant supervision and blood and urine controls are regular.
But while a decision on guilt is rather easy after detecting a foreign substance in a sample, judgement is much more complicated when it comes to doping methods that only change blood or urine values. In such cases, the proofs can only be indirect and statistics is used to help in deciding.
In this talk, we shall review the statistical methods used in the Athlete's Biological Passport and discuss what can and cannot be proven with them. The examples of profiles and interpretation shall be taken from recent trials at the Court of arbitration for sports.

What Medical Researchers Know About Statistics, and is it a Losing Battle to Educate Them?

Simon Day[1] *and Justyna Stefaniak*[2]

[1]Roche Products Ltd, Welwyn Garden City, United Kingdom; simon.day@Roche.com
[2]Data Management and Statistical Analysis, Krakow, Poland; jusstefa@gmail.com

We have surveyed 80 physicians, based in University Hospitals in Krakow, Poland, who work in clinical trials to determine some of the statistical methods of analysis commonly used in trials that they consider they do, or do not, understand. Examples include p-values, confidence intervals, parametric and non-parametric tests, ANOVA, survival analysis, odds ratios and risk ratios, propensity score, ROC and regression analysis. We present the results of this and compare to experiences found in other similar surveys. The challenges of teaching even some of the most fundamental ideas in statistics are, we believe, enormous, and under-rated by many statisticians. We give examples, particularly concerning "adjusted" estimates resulting from models (ANCOVA, logistic regression, Cox regression), and of equivalence and non-inferiority. Good graphics, which can sometimes take much effort to produce, undoubtedly help but still do not adequately address all the issues. We believe, because we are optimistic, that progress is being made. However, much of this progress is limited to individual statisticians finding good ways to explain particular methods and having successful teaching/consulting interactions with selected physicians. Whilst this may contribute to an increase in the mean level of teaching skills amongst statisticians, and it may contribute to an increase in the mean level of understanding amongst physicians, we fear that the advancement of statistical techniques – many of them used quite routinely – is moving ahead faster than we are educating. This is resulting in a widening gap between what is understood and what needs to be understood.

Clustering Symbolic Objects Represented With Discrete Distributions

Simona Korenjak-Černe[1], *Nataša Kejžar*[2] *and Vladimir Batagelj*[3]

[1]Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia;
simona.cerne@ef.uni-lj.si
[2]Institute for Biostatistics and Medical Informatics, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia; natasa.kejzar@mf.uni-lj.si
[3]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia;
vladimir.batagelj@fmf.uni-lj.si

Many real data, especially in social sciences, are composed of categorical values. Their components can be plotted as histograms (numeric data) or bar charts (nominal data). Their representation with discrete distributions produces a special kind of symbolic objects. Such a representation preserves more information than the representation with only mean values and enables us to consider the variables of different types in the clustering process at the same time. Another advantage of such a symbolic data representation is that it enables to combine related data sets into one (e.g. ego-centered networks: ego-SO = ego-properties + SO of ego's alters).
Classical clustering methods were adapted for discrete distributions and implemented in the R package clamix (Kejžar and Batagelj). The adapted leaders method and the adapted Ward's method are compatible – they can be viewed as two approaches for solving the same clustering optimization problem. We applied both methods on several data sets and some of the results obtained on one of them will be presented.

Clustering of Population Pyramids using Wasserstein's Distance

Katarina Košmelj[1] *and Lynne Billard*[2]

[1]University of Ljubljana, Ljubljana, Slovenia; katarina.kosmelj@bf.uni-lj.si
[2]University of Georgia, Athens, USA; lynne@stat.uga.edu

In many real situations, data are collected/presented as histograms. Such an example are population pyramids, which present the age distribution of a population by gender for a particular country. The objective of this paper is to partition countries into homogenous groups according to the similarity of the shape of the population pyramids in each particular year and to observe the time-trend. We use a modified Wasserstein distance for this purpose. A case study on East European countries in the period 1995-2015 is presented. The results reflect that the countries are becoming more and more similar and follow a pattern of aging populations. For the majority of countries, this process started long before 1990, for Kosovo, Albania and Macedonia it started after 1990.

Education

Exploring the Effects of Information and Communication Technology (ICT) on Educational Achievements

Barbara Neža Brečko

Centre for Social Informatics, Faculty of Social Sciences, Ljubljana, Slovenia;
barbara.brecko@fdv.uni-lj.si

Exploring the effects of information and communication technology (ICT) on the educational achievements of pupils, represent a unique challenge, because learning outcomes are influenced by several interrelated factors and processes. The use of ICT for teaching and learning is often associated with factors such as method of teaching and learning, student characteristics, characteristics of school leadership, assessment features, etc. Therefore in analyzing the effects of ICT in pupils' achievement it is crucial to consider ICT as a factor in conjunction with other factors and with understanding of the relationship of ICT and other factors. For instance in measuring the impact of computer use on student achievement a causal relationship can be expressed as the basic linear regression model; however, the pupils' achievement in addition to computer use is also influenced by other factors not included in the model. Also it should be realized that the relationships between variables are not always linear, in relationship between variables may be a loop - for example, the use of ICT impacts greater motivation for learning, which improves achievement, and the return affect of higher achievement could be increased motivation to use ICT.
In order to better explain the relationship between the dependent (pupils' achievement) and independent (ICT indicators, sociodemographic variables, student characteristics...) variables, we will use linear structural modeling, which combines regression and factor analysis. With models we will show causality and intertwining of relationships between factors which affect both - the achievement as well as the use of ICT. Data used for measuring the effects of ICT on the achievement, must allow regression analysis. For the analyses international educational surveys measuring pupils' achievement (where Slovenia also participated) - TIMSS and PISA will be used.

Team Teaching With Student – Experimental Study

Jerneja Šifrer[1], *Zala Žvab*[2] *and Matevž Bren*[3]

Faculty of Criminal Justice and Security, University of Maribor, Maribor, Slovenia
[1]jerneja.sifrer@fvv.uni-mb.si
[2]zala.zvab@gmail.com
[3]matevz.bren@fvv.uni-mb.si

The team or collaborative teaching idea is not new; there were several pilot studies reports and articles published in the nineties. 'Teaching can be a lonely and burdensome experience, but team teaching can transform it from a source of stress to a source of innovation and success' was the motto of these studies that reports professor-professor or professor-student collaborative teaching. In our contribution an experiment study performed with the undergraduate students class of Statistics and professor-student collaborative teaching at the Faculty of Criminal Justice and Security will be reported: comparison of this and previous year students outcomes, the results of quantitative analysis of students questionnaire on theirs experience on team teaching, and qualitative analysis on several benefits to students, student teacher and professors. Hypothesis tested are that team teaching contribute to the better students outcomes, more collaboration and every day students' work and more positive attitude of students.

Does Higher Psychometric Score Predict Better Performance in Academic Studies

David N. Ben[1] *and Tal Shahor*[2]

Department of Economics, Max Stern Academic College Of Emek Yezreel, Migdal Hemek, Israel
[1]missimb@yvc.ac.il
[2]tals@yvc.ac.il

In this paper we estimated the affect of psychometric score and high school grade on bachelor's degree score in several social science departments at Yezreel Valley College. In some regressions the coefficients of psychometric score and high school grade were not significantly different from zero. In cases where the regression coefficients were significant, the effect of high school grade was significantly higher than the effect of the psychometric score. To examine the effect of raising admission requirements of psychometric score, we defined several dummy variables that represent different levels of psychometric score. Regression equation includes: dummy variables, the interaction of these variables with the psychometric score, the interaction of the dummy variables with high school grade and other explanatory variables. The results show that in most cases, raising psychometric score does not make a significant change in final grade of academic degree.

An Application of Teaching for Understanding at the Faculty of Veterinary Science

Teresita E. Teran[1], *Omar Cordoba*[2] *and Augusto Nascimbene*[3]

Faculty of Veterinary, University of Rosario, Rosario, Argentina
[1]teresitateran@hotmail.com
[2]odcordoba@hotmail.com
[3]aguna2003@hotmail.com

Positioned in "Teaching for understanding" developed by Gardner (1993), we carried out the following experience last year in the Faculty of Veterinary Science of Casilda, in the chair of Biostatistics that is taught in the 2nd year of this career. The aim of this work was to obtain a better understanding of the topic Linear Regression by the students.
After having reviewed the previous topics, we proposed comprehensive goals, based on the book we used in the chair : "Statistics Inference with applications. Specific didactic units for Veterinary Science". (Teran,et al, 2010). We developed the concept of Regression, association and causality, we raised the model of Linear Regression, verified the model and evaluated the regression equation; we carried out the test of hipothesis for β and interpreted the analysis of the linear regression. We inferred on the averages of and for each x value, and built up the Confidence Intervals for the Regression straight line. We analysed the residuals and made predictions of a value of and for x along with their Confidence Intervals. After we had developed the main ideas, we put forward problematic situations that the students solved in group with the support of the computer as a tool and which allowed to value the whys of having learnt this topic since all the problem-solving activities that they dealt with were about the field of Veterinary Medicine.
The evaluation was perfomed through non-participant observation, following a protocol that had already been evaluatetd through the α Cronbach and with the recording of the dialogues of the groups taking into account the flexible Knowledge – based understanding arising from the Teaching for Understanding .
The qualitative results have been highly satisfactory .
We believe that the theory of Teaching for Understanding developed by Gardner at Harvard University is an incentive in our own practice to obtain a better understanding, which becomes a significant learning of students in a career where Biostatistics plays an instrumental role .

Statistical Applications - Economics

Maximum Likelihood Estimation of Spatially and Serially Correlated Panels With Random Effects: An Estimation Framework and a Software Implementation

Giovanni Millo

R&D Department, Assicurazioni Generali S.p.A., Trieste, Italy;
giovanni_millo@generali.com

I describe maximum likelihood estimation of panel models incorporating: random effects and spatial dependence in the error terms; and/or a spatially lagged dependent variable; and possibly also a serial dependence structure in the remainder of the error term. I derive an operational version of Anselin's general estimation framework, discuss the computational challenges of estimation and describe an open source implementation in the R system for statistical computing. Applications of spatial panel models in the literature are usually restricted to the standard spatially autoregressive (SAR) and spatial error (SEM) models, possibly with random or fixed individual effects. While a combination as well as extensions of these models to richer correlation structures have been considered by methodologists, starting with Anselin who described a general estimation procedure, practical feasibility has been hampered by computational difficulties. This implementation allows estimating a complete taxonomy of panel models with a combination of spatial lags, spatial errors, serially correlated errors and random individual effects, distinguishing between two different specifications for the random effects (spatially correlated or not). Likelihood ratio and Wald tests for significance of the spatial lag and of any error covariance component are also available. I validate the estimation routines by means of Montecarlo simulations and finally illustrate the package functionalities by applying them to some well-known datasets from the literature.

Pricing Multi Assets American Options With Monte Carlo Simulation

Predrag Popović

University Of Niš, Niš, Serbia; predrag.popovic@gaf.ni.ac.rs

Option pricing is one very important framework in the world of finance. One of the methods widely used for option pricing is the Monte Carlo method. In this paper we describe how the method can be used to price multi assets American option. This kind of options are influenced by multiple sources of uncertainty so there is no explicit solution for pricing them, or if there is then it can't be solved analytically. The method is based on Least-Square Linear Regression model that approximate conditional expectation of the payoff to the option holder. Also, we give an insight in using different basis functions in the regression equation in order to increase robustness of Monte Carlo approximation.

Determining Relationship Between R&D and the Market Value: the Case of Turkey

Güler Aras[1], *Aslı Aybars*[2], *Özlem Kutlu*[3] *and Nuray Tezcan*[4]

[1]Yıldız Technical University, Istanbul, Turkey; aras@yildiz.edu.tr
[2]Marmara University, Istanbul, Turkey; aybars.asli@yahoo.com
[3]Yıldız Technical University, Istanbul, Turkey; okutlu@yildiz.edu.tr
[4]Haliç University, Istanbul, Turkey; nuraytezcan@hotmail.com

In recent years, firms' research and development expenditures have been used as important indicators in various areas, one of which is considered to be the determination of firms' market value. Depending on this issue, numerous studies have been conducted in the literature with empirical results showing a positive and statistically significant relationship between R&D expenditures and market value. Therefore, the existence of this relationship, especially in emerging markets, has become an important factor in terms of encouraging investors to make valuable investment decisions. Furthermore, firms are able to produce high value added products and increase their productivity with investments in this rather important field. In line with these developments, the level of countries' growth and development improves.

This study aims to determine whether R&D expenditures have a positive effect on the market value of firms' in Turkey by utilizing multivariate techniques together with a two stage methodology. First of all, exploratory factor analysis is employed in order to reduce numerous financial ratios. After determining the factors, multiple regression analysis is used to further reveal the relationship. The data set used in the analysis consists of 75 firms quoted on the Istanbul Stock Exchange (ISE). The data pertaining to the firms' R&D expenditures is obtained from the database of ISE.

Productivity Indicators and their use in Composite Indicators

Lenka Hudrlikova[1] *and Kristyna Vltavska*[2]

[1]University of Economics, Prague, Prague, Czech Republic; lenka.hudrlikova@vse.cz
[2]University of Economics, Prague, Czech Republic; kristyna.vltavska@vse.cz

The use of composite indicators is recently developing significantly. The composite indicators measure multidimensional concepts which cannot be captured by single indicator. Building composite indicators contains several steps – from choosing quality theoretical framework to visualization of the results. Sub-indicators should meet certain requirements based on statistical methods. The productivity measurement as a very popular part of economic research is one of the most suitable measurement for compiling the composite indicator. The aim of the paper is to go through indicators of productivity, such as trends in total factor productivity, trends in apparent work productivity, productivity per hour worked, changes in unit labour costs, costs / revenue ratio in the banking sector etc. and compile the composite indicator of productivity measurement for EU members.

Invited Lecture

Dynamic Prediction of Survival With Clinical and Genomic Data

Hans C. van Houwelingen

Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; J.C.van_Houwelingen@lumc.nl

An important clinical application of biostatistics is the development of statistical models for the prognosis of a patient at the moment of diagnosis. In cancer the usual way of giving a prognosis is by means of the x-year survival probability, with $x = 1, 5$ or 10, for example. Traditionally, the prognosis is based on clinical information at the start of the treatment, like age, gender, size of the tumor, tumor stage etc. In the last decade new types of genomic information have become available like micro-array gene expression and proteomic mass spectrometry data. The problem with this new type of data is its abundance. Micro-arrays can measure the expression of tens of thousands of genes, for example.
The talk will address three issues:

1. How to obtain valid prognostic model based on high-dimensional genomic data.
2. How to assess the added value of the genomic information.
3. How to obtain robust dynamic predictions (predictions available later on in the follow-up).

Biostatistics and Bioinformatics I

Information Growth of Reweighted Cox Model Estimators used in Group Sequential Clinical Trials Under Non-Proportional Hazards

Adam P. Boyd

Novartis Pharma AG, Basel, Switzerland; adam.boyd@novartis.com

Randomized controlled clinical trials are often conducted using group sequential designs, which provide guidelines for early stopping if the scientific aim of the study is answered prior to the planned maximal study duration. For trials with time-to-event outcomes, the Cox proportional hazards model is commonly used to estimate the hazard ratio. When the proportional hazards assumption does not hold, estimates at interim analyses can differ from those expected at final analyses due to the change in the hazard ratio over time and the observed censoring pattern, complicating the estimation and inference problem at the interim analysis timepoints. To address this problem, a reweighting of the Cox model estimator can be used to standardize the estimate to more closely reflect what may be observed at the planned final analysis timepoint. In this work, we show how the information of the reweighted estimator grows over time, allowing for improved planning and implementation of interim analyses on the information time scale. The issue is illustrated with an oncology clinical trial example .

Estimation of Net Survival: Problems and Properties

Anamarija Rebolj[1] *and Maja Pohar Perme*[2]

Institute for Biostatistics and Medical Informatics, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia
[1]anamarija.rebolj@mf.uni-lj.si
[2]maja.pohar@mf.uni-lj.si

Estimation of relative survival has become one of the the most basic steps when reporting cancer survival statistics. Its aim is to evaluate the burden of a disease using the observed survival data and the population mortality tables. Most often, the measure of interest is *net survival*, the probability of surviving cancer in a hypothetical situation when cancer is the only cause of death. A new estimator of net survival, which is unbiased under the usual assumption of non-informative censoring, was proposed. We shall describe its idea and properties.
In practice, the age structure of the diagnosed cohort may change in calendar time, thus causing the censoring due to the end of study to become age-dependent. We shall illustrate the problem on simulated data and propose weighting that ensures an unbiased estimator despite informative censoring.

Fuzzy Bootstraping Aided Extended Mortallity Model of LC-type and its Use to Evaluation of Mortality in Poland

Agnieszka Rossa[1] *and Andrzej Szymański*[2]

University of Lodz, Lodz, Poland
[1]agrossa@uni.lodz.pl
[2]anszyman@math.uni.lodz.pl

In the original Lee-Carter stochastic mortality model (LC) the Singular Value Decomposition is usually used for parameters estimation. What is more, the original LC approach assumes the homoscedascity of random errors, what means that the model errors have the same variance over all ages and time periods, what is not always a real and correct assumption.
Koissi and Shapiro (2006) have formulated the fuzzy version of the LC model (FLC), where the model coefficients are assumed to be fuzzy numbers. The advantage of such a fuzzy approach is that the errors are viewed as fuzziness of the model structure; hence the homoscedasticity is not an issue. However, the fuzzy version of the LC model needs some fixed membership functions to be employed.
Koissi and Shapiro assumed the symmetric triangular membership function (STMF) for representation of model coefficients. Every STMF is characterized by the so-called central value and spread. For each age group the log-central death rate in FLC is treated as a linear function of time and is estimated by ordinary least-square method. The spread is obtained by using the minimum fuzziness criterion.
However, the empirical scatterplots show that the regression should be rather piecewise linear instead of linear. Thus, we have formulated the Extended Fuzzy LC model (EFLC), which will be a subject of our presentation. We have used the bootstrap simulation method to determine the membership functions for all the linear pieces of the EFLC model solution.

Life Tables Model in Life Insurance

Abdellah Qannari[1], *Christiana Balan*[2], *Serge Sabourin*[3] *and Gerard Brunet*[4]

[1]University of Poitiers, Niort, France; abdellah.qannari@univ-poitiers.fr
[2]Universitatea Alexandru Ioan Cusa, Iasi, Romania; christiana.balan@uaic.ro
[3]University of Poitiers, Niort, France; serge.sabourin@univ-poitiers.fr
[4]University of Poitiers, Niort, France; gerard.brunet@univ-poitiers.fr

The new EU regulation on insurance (Solvency II) will allow insurance companies to prove their abilities to quantify the risks inherent in the business for a better allocation of their capital.
The implementation of Solvency II requires insurers to provide "mathematical provisions", the level of which is set on the basis of the expected value of future claims. In this context, insurers are interested in predicting future claims costs, which involves predicting the death rate.
In the field of life insurance, the method used for predicting death rate is life tables which describes the law of accident incidence and assesses the average cost for a better provisioning and pricing.
In this paper we propose to model life tables using data on death rates by age (20 years to 84 years) between 1961 and 2006. The purpose of this study is to predict future life tables for a given time horizon.
To model life tables we express mortality according to age and date, using two established models: Lee Carter's model, and the logarithmic regression model. We compare the estimation errors of mortality rates between the two models using the squared error and the Chi square distance, by age and year, and predict future mortality rates using a trend modelling tool.

Biostatistics and Bioinformatics II

Application of a Power Model for Determination of Adventitious Presence of Genetically Modified Organisms in the Case of Maize

Katja Rostohar[1], *Andrej Blejec*[2], *Vladimir Meglič*[3] *and Jelka Šuštar-Vozlič*[4]

[1]Agricultural Institute of Slovenia, Ljubljana, Slovenia; katja.rostohar@kis.si
[2]National Institute of Biology, Ljubljana, Slovenia; andrej.blejec@nib.si
[3]Agricultural Institute of Slovenia, Ljubljana, Slovenia; vladimir.meglic@kis.si
[4]Agricultural Institute of Slovenia, Ljubljana, Slovenia; jelka.sustar-vozlic@kis.si

The question 'Can different types of production systems, such as genetically modified (GM), conventional or organic, co-exist?' is still being considered and the problem is examined case-by-case for each crop species. Gene flow has been studied for certain field positions by taking samples at different distances from the GM pollen source (donor) and defined as outcrossing rate (OCR). Since the OCR decreases with increased distance from the donor field, curve fitting methods have been used to estimate parameters of the power model: $OCR(x) = K \cdot x^a$ (x is the distance from the pollen source). To estimate parameters K and a at least two points of measurements are needed, which are obtained by sampling in the field. Power functions have been applied to estimate the overall mean of OCRs in the field and the distance from the pollen source, where the mean OCRs drops below the threshold level for labelling of the adventitious presence of GM organisms. The simulations have shown that the overall mean is dependent on the field length. The OCRs are the highest near the pollen source. The overall means were below the threshold level of 0.9% in the fields longer than 50m. To reach the overall mean OCR below the threshold of 0.1% in the fields shorter than 500m the buffer zone is needed.

When Vine is not Fine we Fear for Wine

Ana Rotter[1], *Petra Nikolič*[2], *Kristina Gruden*[3], *Andrej Blejec*[4] *and Marina Dermastia*[5]

National Institute of Biology, Ljubljana, Slovenia
[1]ana.rotter@nib.si
[2]petra.nikolic@nib.si
[3]kristina.gruden@nib.si
[4]andrej.blejec@nib.si
[5]marina.dermastia@nib.si

The agronomical and economical importance of wine is obvious. In recent years, small plant pathogenic bacteria called phytoplasmas, causing grapevine yellows, have been identified in the majority of grapevine growing countries. They pose a danger to the wine-producing community due to the potentially devastating effects of grapevine's infection with phytoplasmas. When the pathogen, as in this case, does not immediately destroy the targeted organism, several scenarios arise. A plant in a vineyard may remain infected; it may recover or even get infected again. The immediate question is then whether there are differences between all possible scenarios of the plant-pathogen infection and whether it is possible to differentiate between them? What statistical tools are there to be used? Are the results gained relevant from a biological perspective? We will present our approach to this problem and try to answer the questions.

SMOTE for High-Dimensional Class-Imbalanced Data: A Theoretical and Empirical Analysis

Rok Blagus[1] *and Lara Lusa*[2]

Institute for Biostatistics and Medical Informatics, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia
[1]rok.blagus@mf.uni-lj.si
[2]lara.lusa@mf.uni-lj.si

Classification using class-imbalanced data is biased in favor of the majority class. For high-dimensional data, where the number of variables greatly exceeds the number of samples, the bias is even larger. The class-imbalance problem can be attenuated by training the classifiers on class-balanced training sets, which can be obtained using down-sizing or oversampling techniques. Down-sizing uses only a subset of the data, while oversampling artificially increases the sample size; however, simple oversampling generally is not an effective choice. Synthetic Minority Oversampling TEchnique (SMOTE) is a very popular oversampling method that seems to perform better than simple oversampling but its behavior on high-dimensional data has not been thoroughly investigated.
In this paper we investigate the properties of SMOTE from the theoretical and from the empirical point of view, using simulated and real high-dimensional data. We show that SMOTE reduces the variability of the minority class and introduces correlation between samples; a specificity in the high-dimensional setting is that SMOTE reduces the Euclidean distance of test samples from the minority class. SMOTE does not modify the classification rules of most classifiers; the results obtained by variable selection methods should be interpreted with care if SMOTE was performed prior to selecting the variables. The simulation studies and real data analysis show that SMOTE is likely to be helpful for k-NN classifiers that use Euclidean distance as a similarity measure, provided that some type of variable selection is performed. The other six types of classifiers that we consider (discriminant analysis, CART, random forests, penalized logistic regression, support vector machines and PAM) do not benefit from SMOTE, which performs worse than simple downsizing, and only slightly improves the classification obtained using the original class-imbalanced training sets.

Using Principal Geodesic Analysis on Shape Space

Mousa Golalizadeh

Tarbiat Modares University, Tehran, Iran; golalizadeh@modares.ac.ir

The scope of the linear statistics consists of analyzing data only on Euclidean space. However, there are many examples, such as DNA molecular topological structure, in which the initial or transformed data lie on non-Euclidean space (Mardia et al. 2003). In this case, so-called the manifold valued statistics should be used (Sommer et al. 2010). An example of manifold valued statistics is the shape statistics which mainly deals with geometrical objects when the rotation, scale and translation effects are not of any interest (Dryden and Mardia, 1998). To provide a measure of shape variability, the Principal Component Analysis (PCA) is usually performed on a tangent space, which is a linearized Euclidean space. This is because the PCA cannot be directly implemented on non-Euclidean shape space. A recently developed tool, Principal Geodesic Analysis (PGA), is a feasible measure to explain variability for nonlinear statistics or, generally, manifolds valued statistics (Fletcher et al. 2004). In my talk, using a real data set representing DNA molecular structure the performance of this new tool is compared with usual PCA on the tangent space to the shape space. Particularly, it will be shown that although the PCA explains the relative variability around the mean, the PGA outperforms better in this case because of taking the entire variability of geometrical objects into account. Some comments on other aspects of the statistical shape analysis of DNA molecular structure will also be given.

Modeling and Simulation I

Modelling Functional Relationship Between Longitudinal Data Series

Xiaoshu Lu[1] *and Esa-Pekka Takala*[2]

Finnish Institute of Occupational Health, Helsinki, Finland
[1]xiaoshu@cc.hut.fi
[2]eas-pekka.takala@ttl.fi

In this paper a new method is presented to infer functional relationship between simultaneous longitudinal data series based on singular value decomposition through extracting temporal patterns, performing functional regression for relating temporal patterns, and modelling functional relationship. The application and the utility of the model are illustrated by the measurement data. The performance of the model is evaluated by comparing the model predictions using actual measurements. The results obtained provide new insights into the structural-functional relationship of the studied longitudinal data.

A Bayesian Analysis of Unemployment Duration Data

Mojtaba Ganjali[1] *and Taban Baghfalaki*[2]

Shahid Beheshti University, Tehran, Iran
[1]m-ganjali@sbu.ac.ir
[2]t.baghfalaki@yahoo.com

In this paper a parametric Bayesian approach is used for analysing unemployment duration data with right and interval censored values. The effects of some important factors on the duration time of unemployment are investigated by a regression modeling approach. Some sensitivity analysis on the choice of prior parameters is performed. Also, posterior predictive distribution is used for goodness of fit of the model. The Bayesian model is applied on duration of unemployment of people in Iran.

Multi Objective Economic Statistical Design of Control Charts

Alireza Faraz[1], *Erwin Saniga*[2] *and Cédric Heuchenne*[3]

[1]Quantitative Methods and Operation Research Department, HEC Management School, University of Liege, Liege, Belgium; and Industrial Engineering Department, Masjed Soleiman Branch, Islamic Azad University, Masjed Soleiman, Iran; alireza.faraz@gmail.com
[2]Department of Business Administration, University of Delaware, Newark, Delaware, USA; sanigae@lerner.udel.edu
[3]Quantitative Methods and Operation Research Department, HEC Management School, University of Liege, Liege, Belgium; C.Heuchenne@ulg.ac.be

Control charts are the primary tools of statistical process control. These charts may be designed by using a simple rule suggested by Shewhart, by a statistical criterion, an economic criterion or a joint economic-statistical criterion. Each method has its strengths and weaknesses. One weakness of the methods of design listed above is their lack of flexibility and adaptability, a primary objective of practical mathematical models. In this paper, we explore multi objective models as an alternative for the methods listed above. These provide a set of optimal solutions rather than a single optimal solution and thus allow the user to tailor their solution to the temporal imperative of a specific industrial situation. We present a solution to a well known industrial problem and compare optimal multi objective designs to economic designs, statistical designs, economic statistical designs and heuristic designs.

Statistical Applications - Biostatistics

Analysis of the Viral Immune Response and Testing the Infection Course Hypothesis

Nataša Kejžar[1], *Miša Korva*[2] *and Tatjana Avšič Županc*[3]

[1]University of Ljubljana, Faculty of Medicine, IBMI, Ljubljana, Slovenia;
natasa.kejzar@mf.uni-lj.si
[2]IMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia;
misa.korva@mf.uni-lj.si
[3]IMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia;

Classical primary response graphs to a viral infection yield diagrams that are well known to every microbiologist. If a virus shows a different immune response behavior it should be subjected to more specific statistical analyses.
In the work we analyze an example of such a virus. We investigate the associations between viral replication (viral load) and specific antibodies measured several times during the patients' infection. Longitudinal data analysis, i.e. mixed effect models are applied.
The second part of the analysis consists of testing the novel expert's hypothesis about the infection course. How does the immune response differ if the infection is fatal with regard to its milder courses? Bayesian evaluation of informative hypothesis is going to be used to show which hypothesis is more likely regarding the data.

The Mammographic Screening Program in Trieste: First Statistical Considerations

Fabiola Giudici[1], *Lucio Torelli*[2], *Fabrizio Zanconati*[3] *and Maura Tonutti*[4]

[1]Department of Medical, Surgical and Health Sciences, University of Trieste , Trieste, Italy;
fabiette84@yahoo.it
[2]Department of Mathematics and Informatics, University of Trieste, Trieste, Italy;
torelli@units.it
[3]Department of Medical, Surgical and Health Sciences, University of Trieste , Trieste, Italy;
fabrizio.zanconati@aots.sanita.fvg.it
[4]Clinical Operative Unit (UCO), Radiology-University Hospital, Trieste, Italy;
mtonutti@sirm.org

The province of Trieste is one of the Italian areas with the highest incidence of breast cancer. Mammographic screening program has started for women aged 50 to 69 since January 2006. This program relies on the collaboration of professional figures who deal with breast disease in our area. Our team of biostatisticians deal in particular with the quality of the data and its subsequent analysis. Women with non-negative mammography are invited to a second phase of cytological study. Each nodule subjected to cytological examination, is classified using the five diagnostic categories (C1-C5) proposed by the European guidelines. Thanks to the cyto-histologic correlation of each lesion, it is possible to calculate sensitivity, specificity, negative and false positive rate, and compare those values with relative standards proposed by the screening's guidelines. The quantitative and qualitative impact of mammography screening, is demonstrated by comparing breast cancers in the two years period before the activation of the program (2004-2005), with those found in the same population during the second round (2008-2009). The screening program has identified a greater number of breast cancers within the same range and in particular it has detected very small lesions. It was also possible to assess risk factors for this pathology.

How to Measure Patients' Dissatisfaction

Mirna Macur

School of Advanced Social Studies, Nova Gorica, Slovenia; mirna.macur@fuds.si

Quantitative research instrument is often used to measure users' satisfaction. Such research instruments (either satisfaction survey or other more or less standardised questionnaires) are used also in health care where patients are being asked about satisfaction with health care services. Does this research instrument work well in such settings? Do patients talk about their dissatisfaction with health care service? If they do under what circumstances? Paper discusses strengths and weaknesses of quantitative research instrument in health care and suggests alternatives. Health care is specific in terms of importance to patients (health being one of the most important values) on the other hand significant level of trust is often required between physicians and patients. We also believe that patients' dissatisfaction is not shown immediately and is hard to measure with validity. Data show that dissatisfied patients will most likely change personal physician, than explain what is wrong. Importance of complaints in health care is discussed, because they reveal information that is hard to obtain with questionnaires. Being so rare complaints represent important information; they should be taken seriously and lead to improvements. But do they?

Effect of Repeatedly Self-Assessing the Presence and Severity of Health Symptoms: An Empirical Study

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

[1]Institute for Biostatistics and Medical Informatics, Medical Faculty, University of Ljubljana, Ljubljana, Slovenia; lara.lusa@mf.uni-lj.si
[2]Department of Infectious Diseases, University Medical Center Ljubljana, Ljubljana, Slovenia; cerar.dasa@gmail.com
[3]Department of Infectious Diseases, University Medical Center Ljubljana, Ljubljana, Slovenia; franc.strle@kclj.si

In this talk we focus on the effect of repeatedly self-assessing the presence and severity of health symptoms. The 14 investigated symptoms were: fatigue, malaise, arthralgias, headache, myalgias, parasthesias, dizziness, nausea, insomnia, sleepiness, forgetfulness, concentration difficulties, irritability and pain in the spine. The instrument used was a written 14-symptom questionnaire, asking whether the subjects had had any of the 14 symptoms in the previous week. The severity of each individual symptom was graded by the subject on an 8-cm visual analog scale (8 = most severe). The subjects included in the study were two cohorts of patients with typical solitary erythema migrans enrolled in two clinical studies, who filled out the questionnaire at enrollment, and 6 and 12 months thereafter. Each study included also a cohort of controls, for whom the same questionnaire and time points were used.
We estimated the effect of repeating the questionnaire using a multivariable logistic regression model, the self-assessed presence of symptoms was the dependent variable and the analysis was controlled for seasonality, age and gender of the subjects. To account for multiple measurements in each patient, the analysis was also adjusted for a subject variable as a random effect.
Our results show that the frequency and severity of symptoms dramatically decrease during follow-up, for both patients and controls. While this might be explained as an improvement of general health in patients due to treatment, it is a puzzling result for controls. We discuss the methodological issues related to this unexpected result; moreover, we also identify which symptoms are associated with seasonality, age and gender, and show the patterns of association between symptoms.

Econometrics

Modelling Synergies In Planning Multimedia Activities In Integrated Marketing Communications Perspective

Jana Suklan[1], *Vesna Žabkar*[2] *and Damjan Škulj*[3]

[1]Interdisciplinary Postgraduate Study in Statistics, University of Ljubljana, Ljubljana, Slovenia;
jana.suklan@gmail.com
[2]Faculty of Economics, University of Ljubljana,, Ljubljana, Slovenia;
vesna.zabkar@ef.uni-lj.si
[3]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
damjan.skulj@fdv.uni-lj.si

When planning communications costs it appears that marketers have only the minimum information about the relative effectiveness of individual advertising channel. As a result, companies have little information on whether their input is intended to promote sales.
The concept of integrated marketing communications provides for the consolidation of advertising channels, leading to the emergence of synergy. Effects of individual channels and each other are not independent. Synergy means interaction between the various communication channels, when acting simultaneously. The term refers to combined sales impact when exceeds the sum of the independent effects (Naik and Raman, Understanding the Impact of Synergy in Multimedia Communications, 2003, Vol. XL, 375-388). The objective of integrated marketing communication is to increase the final effect, such as sales and brand recognition by applying various communication channels.
Much has been done on this field trying to model synergy in traditional media, using a common advertising model. Early studies on the field have been conducted in terms of static models (Gatignon and Hanssens 1987). Recent models have been redefined by adding time components and other modifications (Gopalakrishna and Chatterjee 1992, Naik and Raman 2003 and 2006). Our case study provides for collecting the data in the specified way (tracking sales for each channel separately) that can lead to an added value in modeling synergy between channels. To this end we introduced our dynamic model $S_t = \alpha + \beta_1 u_t + \beta_2 v_t + \lambda S_{t-1} + \kappa_1 u_t v_t + u_t S^1_{t-1} + v_t S^2_{t-1} + \nu_t$, where κ represents synergy because the combined sales impact of (u, v) exceeds the sum of the independent effects $(\beta_1 u_t + \beta_2 v_t)$ when $\kappa > 0$.
Having this information we studied the impact of synergy on media budget, media mix, and advertising carryover effect. Our aim is to test the model empirically against previously introduced models.

Credit Scoring Models and their Quality

Jan Kolacek[1] *and Martin Rezac*[2]

Masaryk University, Brno, Czech Republic
[1]kolacek@math.muni.cz
[2]mrezac@math.muni.cz

Lenders, such as banks and credit card companies, use credit scoring to evaluate the potential risk posed by lending money to consumers. Once a scoring model is available, it is natural to measure its quality. To evaluate the effectiveness of credit scoring models, it is possible to use quantitative indexes such as Gini index, K-S statistics, Lift, Information statistics etc. The presentation deals with mentioned quality indexes, their properties and relationships. Finally, a new approach to measure power of scoring models is discussed and a new quality index is proposed. Also an application of mentioned theory on real data set is demonstrated.

Multidimensional Measures of Happiness and Subjective Wellbeing: Combining Existing Approaches to Develop a New Happiness Model

Thanawit Bunsit

University of Bath, Bath, United Kingdom; tb238@bath.ac.uk

With decades of research in welfare economics, psychology and neuroscience, new approaches have been found to provide a better understanding of how to measure the self-reported happiness and subjective wellbeing of individuals. This paper attempts to compare different types of happiness and subjective wellbeing measures and determines which ones are the most powerful in order to evaluate these abstract ideas.

Data used in this paper was collated from fieldwork in rural Thailand. Different techniques for measuring happiness included a single item question to provide ordinal scale or binary measures as a response. In addition, indicators were measured as an interval or ratio scale. These different measures of happiness and subjective wellbeing were compared for validity and reliability.

Finally, the happiness and wellbeing indicators were employed in the happiness model and assessed using binary and ordinal probit model, and multiple regression analysis, to find out the determinants of happiness. Using multidimensional aspects from psychology, economics and neuroscience perspective, the proposed indicators are expected not only to give a new way of measuring happiness and subjective wellbeing but also to give an alternative appraisal of a nation's development.

Relations between the Czech and German Economies

Jakub Fischer[1] *and Jana Kramulova*[2]

University of Economics, Prague, Czech Republic
[1]fischerj@vse.cz
[2]jana.kramulova@vse.cz

This presentation is focused on statistical verifications of relations between the Czech and German economies. Many economists and journalists say that there is a strong dependency of development of the Czech economy on German economic trends. At a level of statistical indicators, there is a hypothesis of dependency of GDP series of the Czech Republic on GDP series of Germany. This hypothesis is based also on the fact that the Czech Republic is a small open economy and the main part of goods and services is exported to Germany. In this presentation, long-term and short-term relations between GDP development of Germany on one hand and GDP development and Czech exports to Germany on the other hand are verified. Exports are also broken down by SITC classification, Error Correction Model is used for analysis.

Modeling and Simulation II

Sample Size in Propensity Score Methods for Estimating Causal Effects

Ana Kolar[1], *Vasja Vehovar*[2] *and Donald B. Rubin*[3]

[1]University of Ljubljana, Ljubljana, Slovenia; annakolar@yahoo.com
[2]University of Ljubljana, Ljubljana, Slovenia; vasja.vehovar@fdv.uni-lj.si
[3]Harvard University, Cambridge, United States of America; rubin@stat.harvard.edu

Propensity score methods for estimating causal effects are becoming increasingly important within observational study designs. Having the property of being able to replicate randomized design greatly increases the applicative value of the methods, not only within observational studies, but also within broken randomized experiments, or even partly randomized experiments. Accordingly, the methods are becoming increasingly alternatives to various correlation and or/regression based methods, which are conceptually problematic for estimating causal effects.
Propensity score methods are based on the Rubin Causal Model, and their corresponding development over the past three decades has now enabled a clearer guidance on the process of estimating causal effects. However, this guidance is well established predominantly for the case of large data sets, whereas the question of "how large" data set should be remains in large part unanswered. Particularly in social science and medical research, where we often face relatively small samples, an answer to the question remains a highly important issue. Essential questions here are related to the factors that have an impact on required/desired sample size: (a) criteria (e.g. bias, precision), (b) characteristics of the data (i.e. number of covariates, correlation structure, ratio between control and treated group) and (c) statistical approaches (e.g., matching algorithms). The paper presents results of initial simulations that address the role of these three factors.

Modeling of Patterns by an Intelligent System

Anamarija Borštnik Bračič[1], *Igor Grabec*[2] *and Edvard Govekar*[3]

[1]Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana, Slovenia;
anamarija.bracic@fs.uni-lj.si
[2]Amanova d.o.o., Technology Park, Ljubljana , Slovenia; Igor.grabec@amanova.si
[3]Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana, Slovenia;
edvard.govekar@fs.uni-lj.si

A two dimensional pattern represents a fingerprint of the process, which was used to produce the pattern. Therefore, it is expected, that the information about the production process can be estimated from the pattern. A non-parametric statistical method for modeling chaotic two dimensional patterns and estimation of type of production process and the characteristic parameters is proposed. It is based on joint probability density function of samples taken from known two dimensional patterns representing the database. A new pattern with unknown production process is reproduced by comparing parts of the new pattern with samples taken from the data base. Since the samples in the data base also include the information about the production process, relevant parameters and type of production process can be estimated simultaneously with reproduction of patterns. In the presentation the proposed method is demonstrated on two types of patterns. The first type originates from a milling process with intentionally invoked chatter. The produced surface structures are digitalized by using a white light microscope. Different values of relevant milling parameters are chosen to obtain a rich variety of patterns, which exhibit chaotic behavior. The second type of patterns is obtained by digitalizing artistic paintings, which are produced in different techniques in order to obtain a rich database. The proposed method has been used to reproduce a selected testing pattern, which has not been a member of the database used to build the model, and to estimate its production process parameters. The estimated type and the parameters of the process have shown a very good agreement with the true values, which makes this method a candidate for successful solution of the problem of reverse engineering of surface structures.

A Multivariate Markov Modulated Poisson Process Model for Rainfall Intensity

Rasiah Thayakaran[1] *and Nadarajah I. Ramesh*[2]

School of Computing and Mathematical Sciences, Old Royal Naval College, University of Greenwich, Greenwich, United Kingdom
[1]R.Thayakaran@gre.ac.uk
[2]N.I.Ramesh@gre.ac.uk

Point process models for rainfall are constructed generally based on Poisson cluster processes. Most commonly used point process models in the past are either based on Bartlett-Lewis or Neyman-Scott cluster processes. In this paper, we explore the application of a class of Cox process models, termed Markov Modulated Poisson Process (MMPP), in the field of rainfall modelling. We use this class of models to analyse rainfall data observed in the form of tip time series from rain gauge tipping-buckets. Univariate and multivariate models are employed to analyse data recorded at a single and multiple sites in a catchment area.

We analyse the pattern of rainfall from a network of gauges in Somerset, South-West of England (HYREX data). As the structure of this proposed class of MMPP models allows us to construct the likelihood function of the observed tip time series, we utilize the maximum likelihood methods in our analysis to make inferences about the rainfall intensity at sub-hourly time scales. Univariate version of the model is used to study the rainfall tip-times series at a single location. The multivariate models are then applied to model rainfall time series jointly at four stations in the region. Results of fitting 3-state models are presented. Properties of the cumulative rainfall in discrete time intervals are studied.

Is Simple Randomization of Compounds in Training and Test Set as Good as other Methods in quantitative Structure-Activity Experiments?

Sorana D. Bolboaca[1] *and Lorentz Jäntschi*[2]

[1]Iuliu Hatieganu University of Medicina and Pharmacy Cluj-Napoca, Cluj-Napoca, Romania; sbolboaca@umfcluj.ro
[2]Technical University of Cluj-Napoca, Cluj-Napoca, Romania; lori@academicdirect.org

The present research aimed to assess if the simple random sampling is a proper method for splitting the set of compounds in training and test sets.
Four sets of compounds were included in the analysis: 1) a set of 83 of drug-like compounds with blood-brain barrier permeation; 2) a set of 18 sulfanilamide derivatives with carbonic anhydrase II isozyme inhibitory activity; 3) a set of 34 taxoids with inhibitory activity on cell growth; and 4) a set of 25 triphenylacrylonitriles with affinity on estrogen receptor. A qSAR experiments was carried using the Molecular Descriptors Family on Vertices for computing structural descriptors and multiple linear regressions were identified. Each set of compounds was split in training and test sets using a simple randomization approach. The reliability of randomization was tested using the generalized cluster analysis with K-means algorithm (Statistica 8; Euclidian distance and maximization of the initial distance in regards of cluster center using a cross-validation with 10-folds).
The following number of molecules was included in training:test sets: 55:28 for 1st set; 12:6 for 2nd set; 23:11 for 3rd set; and 19:6 for 4th set. Both the experimental data in training set and test set proved to be normal distributed (Anderson-Darling and Kolmogorov-Smirnov statistics with p-value ¡ 0.05). The proper number of clusters identified using the observed activity and identified descriptors varied from 3 (1st set) to 6 (4th test). With some exceptions (clusters with just one compound), the clusters proved to contain compounds from both training and test set. The descriptors and observed activity proved to have significant contribution in clusterization ($p < 0.001$)
. Simple randomization proved to be a proper method for splitting the set of compounds in training and test sub-sets.

Network Analysis and Statistical Applications

Blockmodeling of Multilevel Network Data

Aleš Žiberna

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

In the talk 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.

Networks Generated by Fifa Soccer Games Played between Countries

Kristijan Breznik[1] *and Vladimir Batagelj*[2]

[1]International School for Social and Business Studies, Celje, Slovenia;
kristijan.breznik@mfdps.si
[2]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia;
vladimir.batagelj@fmf.uni-lj.si

In recent years, home advantage in soccer has been analysed abundantly, but mainly at the club level. In the following paper home advantage in soccer games among national teams was analysed. The data were obtained from the Fifa (world football association) website. It include over 30000 games played among over 200 country from the first official game in 1872 to the end of 2010. Additional data about countries (number of registered and unregistered soccer players, number of soccer clubs and officials etc.) are also available and are going to be presented.
The analysis was performed using descriptive and inferential procedures. A more precise insight at the level of individual country was performed by network analytic methods. In addition, some clustering methods for classification into clusters were applied. Various problems arising during data analysis, e.g. disintegrating countries and the new emerging countries etc., and their solution are also going to be presented. The results are consistent to previous finding at the club level and clearly confirm the existence of home advantage at the country level.
The programs for harvesting and analysing the data and producing networks were written in R. For analysis of networks we used the Pajek program.

Application of Discriminant Analysis in Investigation of Regional Disparities in Serbia

Valentina T. Sokolovska[1], *Katarina J. Čobanović* [2] *and Emilija Nikolić-Djorić B.*[3]

[1]Department of Sociology, Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia; valentinas@neobee.net
[2]Department of Agricultural Economics and Rural Sociology, Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia; katcob@polj.uns.ac.rs
[3]Department of Agricultural Economics and Rural Sociology, Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia; emily@polj.uns.ac.rs

In making strategy of economic development of the country it is very important to find out the differences between regions. The results of previous investigations showed the existence of three main methodological approaches. The first one concerns the classification of analyzed units in groups, in dependence of homogeneous level. The second one emphasizes the ranking of regions according to development level. The third one is the selection of variables which emphasize the differences between existing regions.
The paper analyses the four regions of Serbia which are defined according to NUTS 2 criterion: Vojvodina region, Belgrade region, Šumadija and western Serbia region and South and East Serbia region. Applying the actual methods of discriminant analysis, from a set of variables that characterize the economic and demographic development, it would be selected variables that mostly contribute to regional differences.

The Impact of Computer Price Indices on Total Factor Productivity Measurement

Borut Kodrič[1] *and Lea Bregar*[2]

[1]Faculty of Management, University of Primorska, Koper, Slovenia; borut.kodric@fm-kp.si
[2]Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia;
lea.bregar@ef.uni-lj.si

Multifactor productivity is a central economic concept that measures the efficiency gains in the production process. In theory, it's a more comprehensive measure than labour productivity, but it's also more difficult to estimate. One of the most difficult and contentious tasks of multifactor productivity analysis is the construction of the reliable measure of capital services. Even within the comprehensive and consistent framework of the neoclassical theory of capital, there are numerous conceptual dilemmas and practical problems regarding how to measure aggregate capital services. While there has been a lot of debate regarding the choice of expected rate of return, the choice of expected capital gain, and the treatment of taxes in the user cost estimation, not much attention has been paid to the issues of appropriate price indices of assets which are required in the process of assessment of capital services. The price indices for a given asset type affect the aggregate index of capital services through estimates of the productive stock of the asset type as well as through relative user costs of the asset type in a direct and indirect way. The use of inappropriate asset price indices results in a biased estimate of capital services index and consequently influences the estimate of multifactor productivity index. Compilation of reliable price indices of fixed assets is, however, particularly difficult due to the heterogeneity among capital goods and rapid changes of models as a result of fast technological change. The later is particularly the case with ICT goods, where the conventional methodologies to derive price indices for ICT goods usually understate true price changes. The purpose of this paper is to evaluate the impact on estimates of capital services and multifactor productivity when using alternative ICT price indices in the process of capital services estimation. The impact on capital services and multifactor productivity was assessed for Slovenian manufacturing for the period 1995-2008 by sensitivity analysis.

Workshop

Statistics of Compositional Data

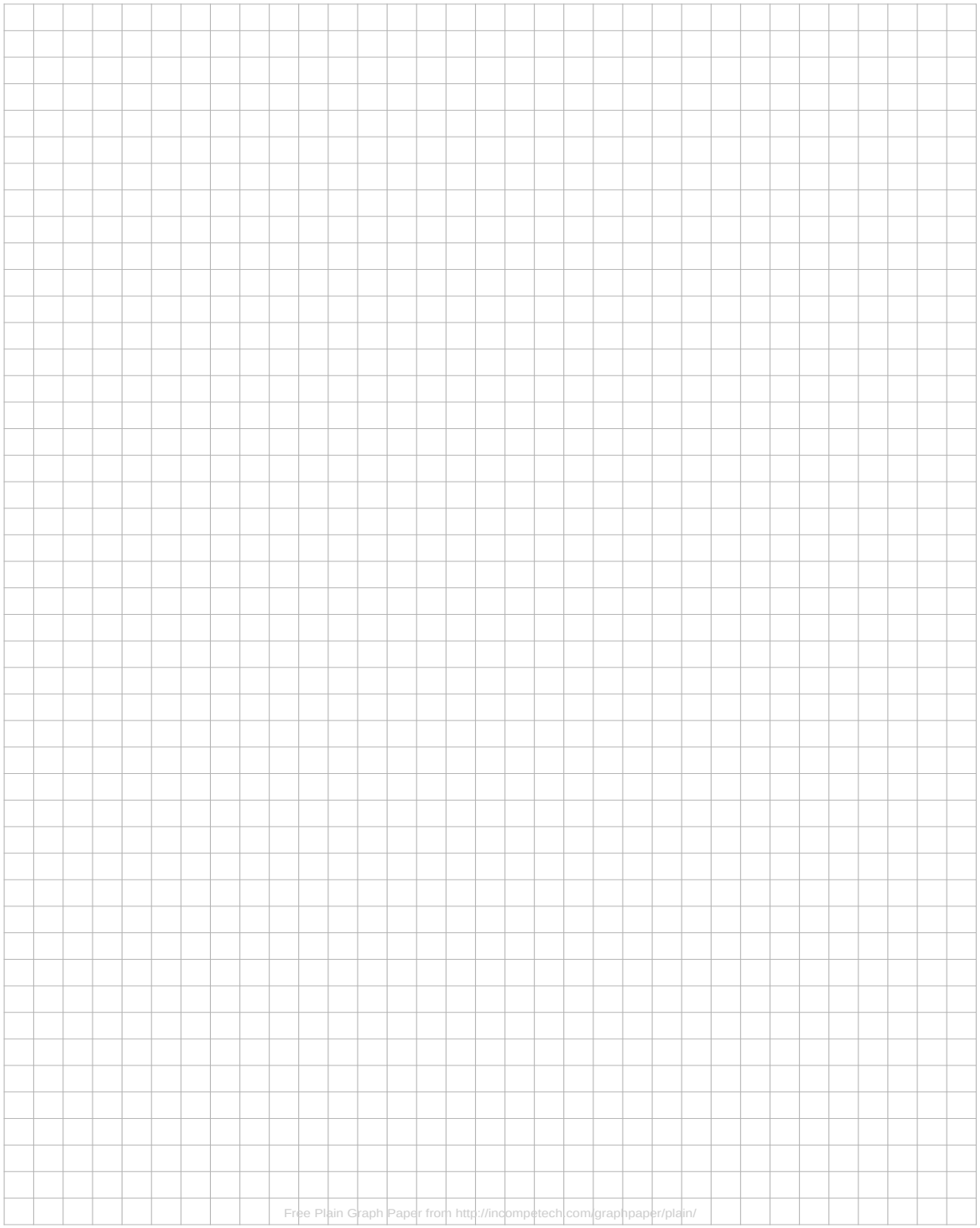
Gerald van den Boogaart

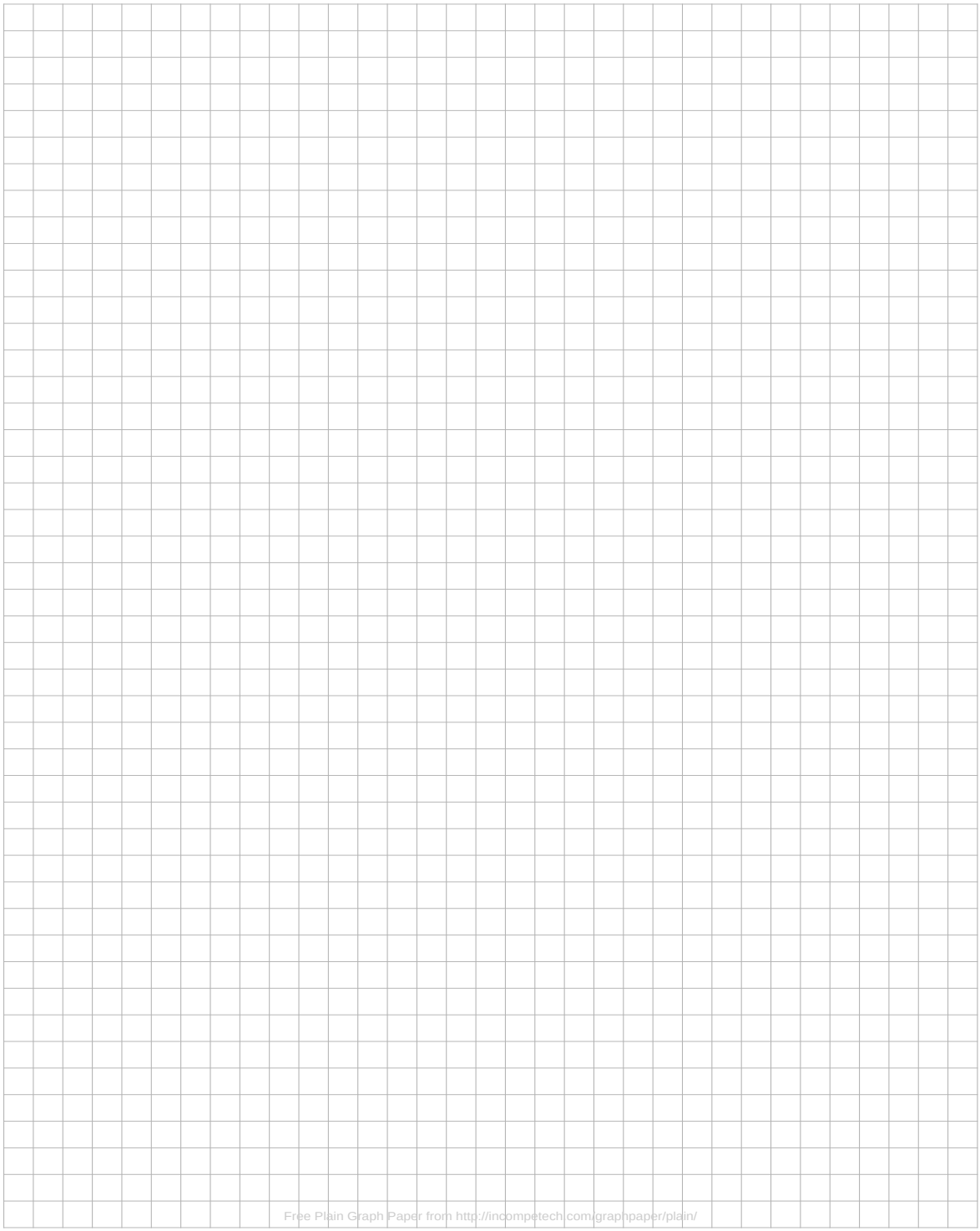
Technical University Freiberg, Freiberg, Germany; boogaart@math.tu-freiberg.de

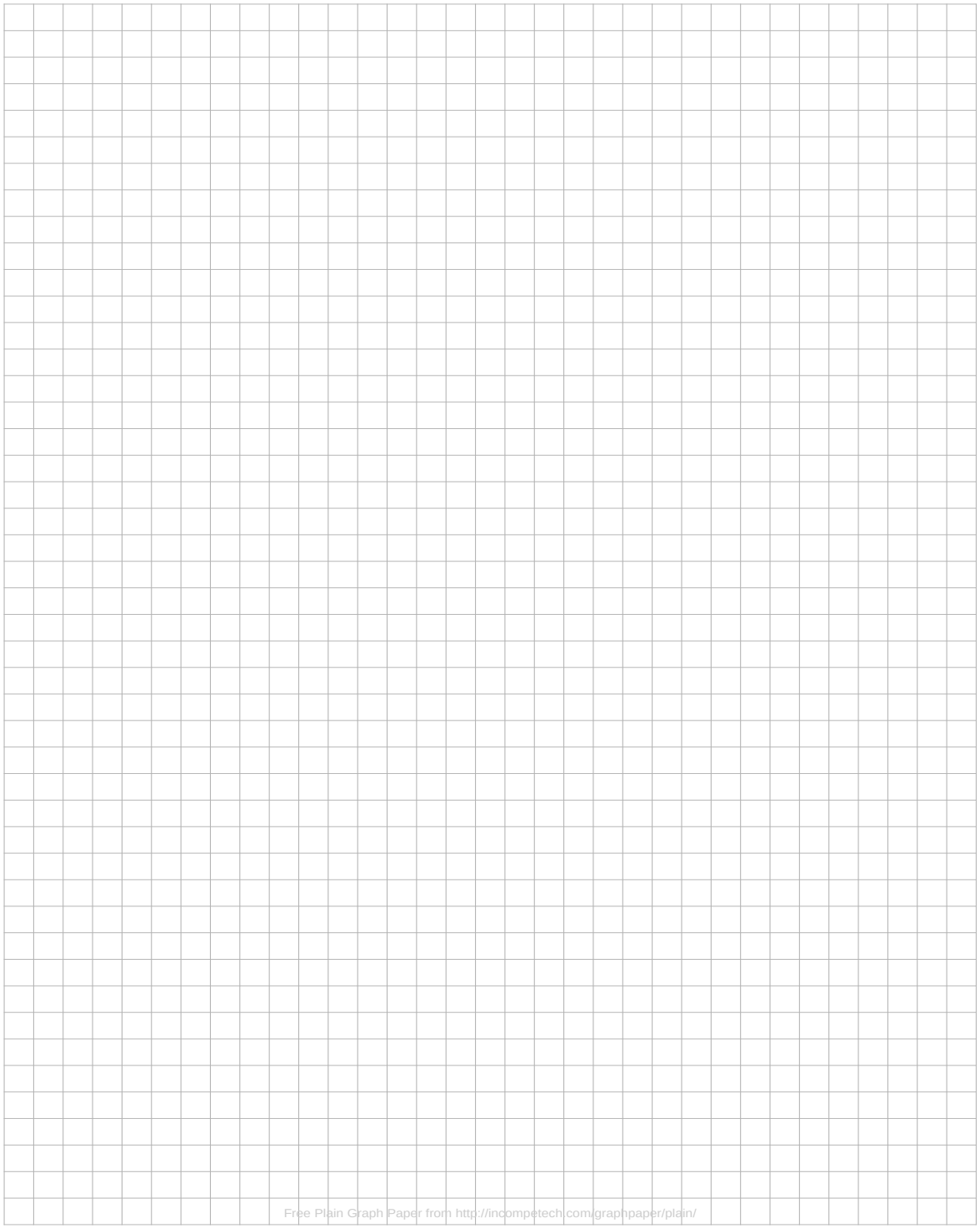
The workshop will teach statistics of compositional data analysis using R and the R package compositions. Compositional data is characterised by the fact that the sum is an artefact of the measurment process (e.g. a constant like100%) and not meaningfull for process to analysed. Analysing compositional data with classical multivariate tools leads to many artefacts. However the so called ”working in coordinates principle” allows to create similar tools for statistical analysis especially adapted for compositional data.
Participants are expected to bring a laptop or computer to the workshop with R and the R package compositions installed.) The help with installing the package can be offered by the instructures prior to the workshop during the conference. Instructures are Gerald van den Boogaart and Matevž Bren, two of the authors of the compositions package.

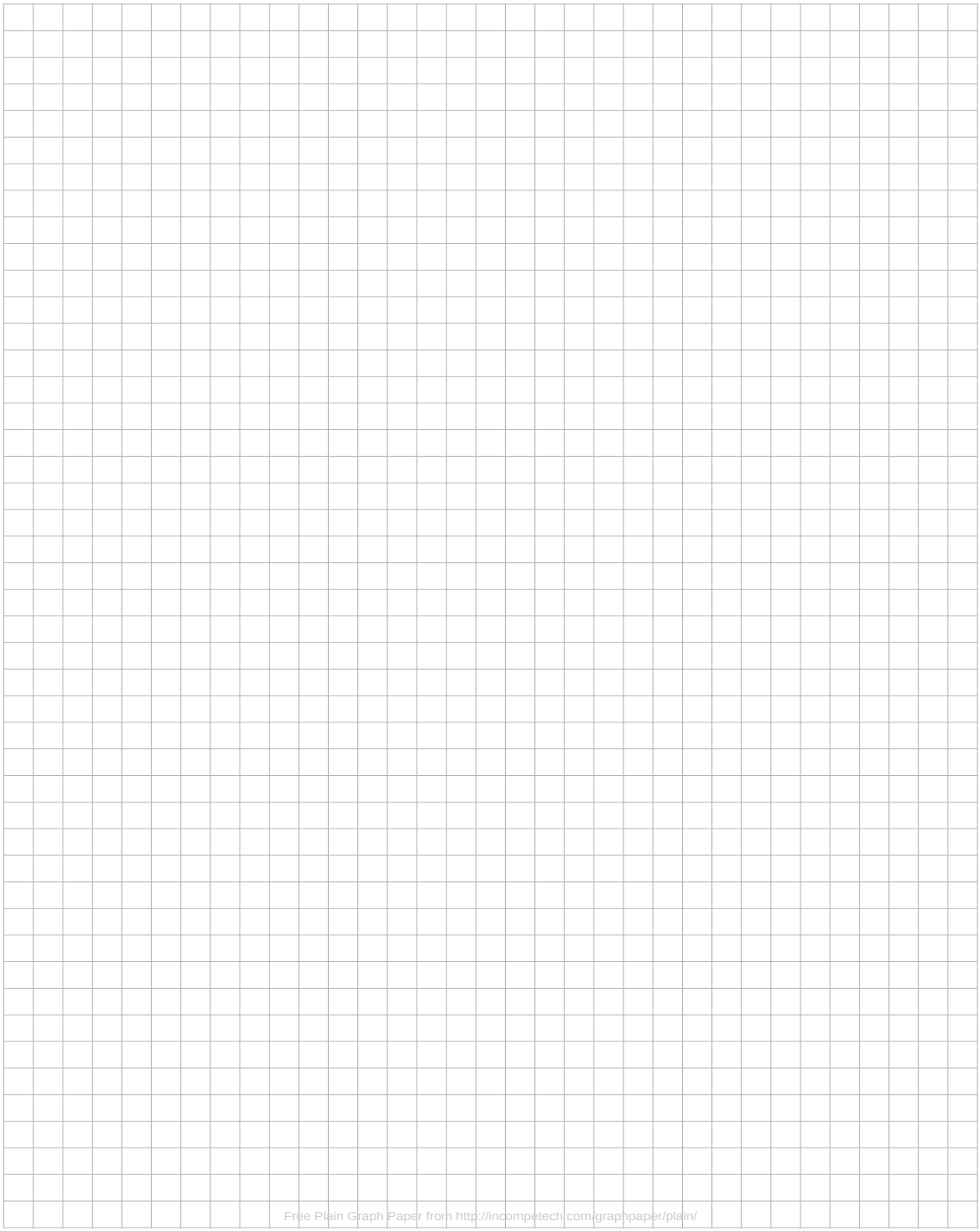
INDEX OF AUTHORS

Index of Authors


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/

SUPPORTED BY

www.arrs.gov.si/en

www.valicon.si

www.alarix.si

www.result.si

www.sweetsurveys.com