

International Conference

**APPLIED STATISTICS  
2004**

**PROGRAM and ABSTRACTS**

September 19 – 22, 2004

Ljubljana, Slovenia

Supported by

Ministry of Education, Science and Sport  
Statistical Office of the Republic of Slovenia  
SPSS Division, Slovenia  
Alarix d.o.o.

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

311(063)

INTERNATIONAL Conference Applied Statistics (2004 ; Ljubljana)  
Program and abstracts / International Conference Applied  
Statistics, September 19-22, 2004, Ljubljana, Slovenia ; [edited  
by Janez Stare, Gaj Vidmar, and Gašper Koren]. - Ljubljana :  
Statistical Society of Slovenia, 2004

ISBN 961-90314-3-1

1. Applied Statistics 2. Stare, Janez, 1952-  
215238912

## Organizing Committee

Andrej Blejec (Chair)  
Vladimir Batagelj  
Vesna Dolničar  
Anuška Ferligoj  
Vanja Govednik  
Bogdan Grmek

Gašper Koren  
Andrej Mrvar  
Gregor Petrič  
Irena Vipavc  
Petra Ziherl

## International Program Committee

Janez Stare (Chair), Slovenia  
Tomaž Banovec, Slovenia  
Vladimir Batagelj, Slovenia  
Jacques Billiet, Belgium  
Maurizio Brizzi, Italy  
Brendan Bunting, Northern Ireland  
Anuška Ferligoj, Slovenia  
Herwig Friedl, Austria  
Dario Gregori, Italy  
Katarina Košmelj, Slovenia  
Dagmar Krebs, Germany

Irena Križman, Slovenia  
Mihael Perman, Slovenia  
John O'Quigley, France  
Jože Rován, Slovenia  
Tamas Rudas, Hungary  
Willem E. Saris, The Netherlands  
Albert Satorra, Spain  
Vasja Vehovar, Slovenia  
Gaj Vidmar, Slovenia  
Hans Waeye, Belgium

---

Published by: Statistical Society of Slovenia  
Vožarski pot 12  
1000 Ljubljana  
Slovenia

Edited by: Janez Stare, Gaj Vidmar, and Gašper Koren  
Printed by: BORI Birografika, d.o.o., Ljubljana



**Program**

---

**SUNDAY, September 19, 2004**

---

9.00 Excursion to Dolenjska

---

---

**MONDAY, September 20, 2004**

---

8.00 – 9.15 Registration at City Hotel Turist

9.15 – 9.30 Opening of the Conference (Hall 1)

9.30 – 10.20 **INVITED LECTURE** (Hall 1) *Chair: Mihael Perman*  
**Brian Ripley**, University of Oxford, UK  
**Data mining in large datasets**

10.20 – 11.10 **INVITED LECTURE** (Hall 1) *Chair: Anuška Ferligoj*  
**Patrick Doreian**, University of Pittsburgh, USA  
**A modest agenda of immodest goals for network analysts**

11.10 – 11.30 BREAK

11.30 – 12.30 **Network Analysis I** (Hall 1) *Chair: Patrick Doreian*

1. **Analysis of the buyers' choice networks**  
*Simona Korenjak-Černe, Nataša Kežar, Vladimir Batagelj*
2. **Social networks in Slovenia: a comparative perspective**  
*Valentina Hlebec, Tina Kogovšek*
3. **Islands**  
*Matjaž Zaveršnik, Vladimir Batagelj*

12.30 – 14.30 LUNCH

14.30 – 15.30 **Social Science Methodology** (Hall 1) *Chair: Valentina Hlebec*

1. **Latent change variable models in studying intra-individual variability in coping measures**  
*Vesna Buško, Alija Kulenović*
2. **Estimating the number of foreign-body injuries in childhood with the scale-up method**  
*Silvia Snidero, Bruno Morra, Roberto Corradetti, Dario Gregori*
3. **Alternative approaches to least squares factor analysis**  
*Gregor Sočan*

15.30 – 15.50 BREAK

15.50 – 17.10 **Statistical Theory and Methods** (Hall 1) *Chair: Robin Henderson*

1. **Indices of skewness derived from symmetric quantiles: statistical properties and application to European temperature data**  
*Maurizio Brizzi*
2. **Effect of rounding on expected value and standard deviation**  
*Anton Cedilnik, Katarina Košmelj*
3. **Asymptotic unbiased distribution function estimators on the basis of interval-censored data**  
*Mikhail S. Tikhov, Irina S. Efimenko*
4. **Hierarchical modeling of correlated binary traits with applications to asthma and hay fever twins data**  
*M.M.Shoukri*

15.50 – 17.10 **Statistical Applications I** (Hall 2) *Chair: Tina Kogovšek*

1. **Social integration of elderly in Slovenia**  
*Valentina Hlebec*
2. **Logistic regression analysis of MRI data from breast cancer patients**  
*Masoud Yarmohammadi, Parviz Abdolmaleki*
3. **Absolute, relative and time distance: Contradicting or complementary measures?**  
*Vasja Vehovar, Pavle Sicherl, Andraž Petrovčič, Vesna Dolničar*
4. **Application areas of hidden Markov models**  
*Thomas Benesch*

19.00 WELCOME RECEPTION, *City Hotel Turist*

---

---

**TUESDAY, September 21, 2004**

---

9.10 – 10.00 **INVITED LECTURE** (Hall 1) *Chair: Janez Stare*

**Frank E. Harrell Jr.**, Vanderbilt University, USA

**Statistical principles to live by**

10.00 – 11.20 **Biostatistics** (Hall 1) *Chair: Frank E. Harrell Jr.*

1. **Estimation of optimal adaptive treatments using observational data: a regret-based approach to dynamic anticoagulation**  
*Robin Henderson, Catherine Fullwood, Syd Stewart*
2. **Parsimonious modelling of time-dependent effects in the Cox model**  
*Stephan Lehr, Harald Heinzl, Michael Schemper*
3. **Goodnes of fit in relative survival models**  
*Janez Stare, Robin Henderson, Maja Pohar*
4. **Comparison of analytic models for the costs of the postinfarct patients**  
*Giulia Zigon, Dario Gregori*

11.20 – 11.50 BREAK

11.50 – 12.50 **Modelling and Simulation** (Hall 1) *Chair: Maurizio Brizzi*

1. **Empirical estimation of outliers for different generating mechanisms**  
*D. K. Shangodoyin, Raghunath Arnab*
2. **ML estimation in a random field based forward interest rate model**  
*Jozsef Gall, Gyula Pap*
3. **Bayesian estimation of kinetic parameters: a particle approach**  
*Franz Konecny*

11.50 – 12.50 **Statistical Applications II** (Hall 2) *Chair: Vasja Vehovar*

1. **The expansion of computers and the internet among Slovenian primary-school leavers**  
*Eva Podovšovnik*
2. **A latent class invariant model of first-time clients versus long-term clients receiving drug treatment, controlling for a number of covariates**  
*Paul Cahill, Brendan Bunting*
3. **Intra-assessor inconsistency in setting performance standards**  
*Hans J. Vos*

12.50 – 14.50 LUNCH

14.50 – 16.10 **Network Analysis II** (Hall 1) *Chair: Vladimir Batagelj*

1. **The effects of limiting the number of alters on composition and structure of social support networks**  
*Tina Kogovšek, Valentina Hlebec*
2. **Sub-sampling of alters in collecting the data on ego-centred social networks**  
*Luka Kronegger, Gašper Koren, Valentina Hlebec, Vasja Vehovar*
3. **Analysis of US patents network in time**  
*Nataša Kejžar, Vladimir Batagelj*
4. **What do networks do for companies? Testing the effects of networks on corporate performance**  
*Marko Pahor*

16.10 – 16.30 BREAK

16.30 – 17.50 **Data Analysis I** (Hall 1) *Chair: Matevž Bren*

1. **Steganography detection – noise analysis in image files**  
*Igor Belič, Aleksander Pur*
2. **Quantification of customer satisfaction data: an approach based on the Spline transformation**  
*Michele Gallo*
3. **Different statistical measures provide different perspectives on digital divide**  
*Pavle Sicherl*
4. **The problem of concave sets with clustering**  
*Mateja Nagode, Matej Francetič, Bojan Nastav*

16.30 – 17.50 **Data Collection** (Hall 2) *Chair: Lea Bregar*

1. **Information and communication technologies usage in Slovenian households and enterprises**  
*Andreja Kačič, Eva Belak, Rudi Seljak*
  2. **Contingent versus unconditional incentives in WWW studies**  
*Anja S. Göritz*
  3. **Effect of "grid" questions in Web survey questionnaires**  
*Katja Lozar Manfreda, Gašper Koren, Valentina Hlebec*
  4. **Evaluation of hygienic environmental indices in village schools of Mazandaran state in Iran**  
*Bizhan Shabankhani*
-

---

WEDNESDAY, September 22, 2004

---

9.10 – 10.00 **INVITED LECTURE** (Hall 1) *Chair: Andrej Blejec*

**K. Laurence Weldon**, Simon Fraser University, Canada  
**Less parametric methods in applied statistics**

10.00 – 10.20 BREAK

10.20 – 11.20 **Measurement and Modelling** (Hall 1) *Chair: K. Laurence Weldon*

1. **Assessing the "accuracy" of proxy-reports on attitudes towards immigrants in Germany**  
*Angela Jäger*
2. **Translation of measurement instruments and their reliability: an example of Job-related Affective Well-Being Scale**  
*Nino Rode*
3. **Stability of measures of centrality and prominence: a meta-analysis on fixed choice and free choice data**  
*Barbara Zemljič, Valentina Hlebec*

11.20 – 11.40 BREAK

11.40 – 13.00 **Data Analysis II** (Hall 1) *Chair: Gregor Sočan*

1. **Exploratory analysis of vibration data for damage detection in civil engineering structures**  
*Shola Adeyemi, Michael L. Fugate, Brian J. Williams*
2. **MixeR package for compositional data analysis**  
*Matevž Bren, Vladimir Batagelj*
3. **Table-based visualisation of categorical data using spreadsheets**  
*Gaj Vidmar*
4. **A hierarchical Bayes multilayer perceptron to analyze heterogeneous sales response**  
*Harald Hruschka*

---

11.40 – 13.00 **Data Mining** (Hall 2) *Chair: Dimitar Hristovski*

1. **Web sites and communication: a textmining application in the banking sector**  
*Silvia Biffignandi*
2. **Discovering the most important factors for communities of soil microarthropods using machine learning**  
*Damjan Demšar, Sašo Džeroski, Paul Henning Krogh, Thomas Larsen*
3. **Analysis of example weighting in subgroup discovery by comparison of three algorithms on a real-life data set**  
*Branko Kavšek, Nada Lavrač*
4. **Analysis of excretion of pineal gland hormone melatonin in ex-mercury miners with machine learning methods**  
*Alfred B. Kobal, Bernard Ženko, Darja Kobal, Mladen Krsnik, Sašo Džeroski, Milena Horvat, Joško Osredkar*

13.00 – 15.00 LUNCH

15.00 – 19.00 **WORKSHOP I** (Hall 1)

**Frank E. Harrell Jr.**, Vanderbilt University, USA

**Statistical graphics for exploring data, presenting information, and understanding statistical models**

15.00 – 19.00 **WORKSHOP II** (Hall 2)

**Vladimir Batagelj, Andrej Mrvar**, University of Ljubljana, Slovenia

**Analysing data with PAJEK**

---

## Abstracts

---

### **Exploratory analysis of vibration data for damage detection in civil engineering structures**

*Shola Adeyemi* (Obafemi Awolowo University, Nigeria)

*Michael L. Fugate, Brian J. Williams* (Los Alamos National Laboratory, USA)

Condition monitoring of civil engineering structures is highly important in order to guard against sudden collapse or failure. Therefore, it is important to study the initial (reference) condition of structures to be able to detect any changes with respect to the reference condition at a later time. This initial condition is referred to as the undamaged state and any change in the initial condition is referred to as damage.

Advances in sensor technology and wireless data transmission are making the development of conditions monitoring of civil engineering structures feasible. Time series data can be collected on the conditions of structures at the initial stage and continuously at later times.

In this paper, we present a statistical analysis of bivariate time series data obtained from undamaged and damaged sources to detect changes in the condition of the system. A thorough exploratory data analysis, in both time and frequency domain, is carried out to detect changes that can be noticed in the conditions of the system, if any, and then carry out further statistical tests to detect changes in the conditions using spectral analysis.

---

### **Steganography detection - noise analysis in image files**

*Igor Belič* (University of Ljubljana, Slovenia)

*Aleksander Pur* (Ministry of Interior, Slovenia)

Steganographic methods hide the encrypted message in cover-image carriers so that it cannot be seen while it is transmitted on public communication channels. Steganographic methods can embed a large amount of the secret information in the first  $k$  least significant bits (LSBs) of the pixels of the cover image. To the human eye, the existence of embedded secret information can be imperceptible due to imperfect sensibility of the human visual system. The secret information must not exceed the level of the noise that is always present in any digital image. The emphasis of our work is to statistically analyze the image noise in order to

detect the possibility of embedded information in the so-called cover image. Since the steganographic method hides the information into the image noise, this is the only plausible way to analyze the properties of such noise. The influence of image compression methods on the noise amplitude distribution is also presented.

The simple 2, 3, and 4 LSB insertion method (without further cryptographic methods) has been tested for Slovenian and English text and the noise differences are shown.

The presented image noise analysis can be a coarse test for possible steganography detection, but for more accurate detection it should be combined with more sophisticated methods.

---

### **Application areas of hidden Markov models**

*Thomas Benesch* (Vienna Medical University, Austria)

The paper is organised in two parts. Part one reviews the basic problems concerning Hidden Markov Models. Part two discusses possible application in Medicine (e.g., gene finding, protein secondary structure prediction, and protein homology recognition), to financial data, to speech modelling and in the field of Bioinformatics.

---

### **Web sites and communication: a textmining application in the banking sector**

*Silvia Biffignandi* (University of Bergamo, Italy)

The paper deals with data mining techniques in Web context. First of all, a short discussion on datamining, webmining and textmining approaches is presented, and a classification of the different objectives of each approach is given. The paper then focuses on the objective to highlight the communication strategies via Internet (the Web) of some banking institutions which belong to the main financial groups operating in Italy.

The interest in this analysis is based on the fact that the financial market - as all the markets of goods and services - has undergone significant changes during the 1990s. It was therefore decided to carry out a statistical study on the content of banking institution Web sites, mainly to determine the characteristics of each site, to evaluate how the institutions exploit the web and to resort to the technique of "text mining" in order to identify different behavioural patterns.

The banking institutions have been obliged to adapt to these changes not only by offering new services and products in order to increase and consolidate their

customer base, but also by backing up their strategic decisions by applying statistical analysis to the various aspects of actions taken by their customers and/or by the competitors. In this context, the use of Internet and especially the creation of websites tends to be the best way to rapidly and effectively reach the public, understand what the public wants, and offer the products required.

A better understanding of how the banks use their websites to communicate with their customer base defines an extremely interesting element used to examine both how coherent the communication layout is with the services offered by the bank and with its target market, as well as being used as a reference model for the construction of other web sites.

The paper describes data characteristics, the method which have been applied, and the obtained results. Text mining techniques appear to be a powerful and promising technique in this context.

---

### **MixeR package for compositional data analysis**

*Matevž Bren* (University of Maribor, Slovenia)

*Vladimir Batagelj* (University of Ljubljana, Slovenia)

R (<http://www.r-project.org/>) is 'GNU S' - a language and environment for statistical computing and graphics. R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering etc.). Further extensions can be provided as packages.

In 2003, we started to develop the R package MixeR for compositional data analysis that would provide a support for compositional data analysis: operations on compositions (perturbation and power multiplication, subcomposition with or without residuals, computing Aitchison's, Euclidean, Bhattacharyya distances, compositional Kullback-Leibler divergence etc.) and graphical presentation of compositions with ternary diagrams and tetrahedrons. The current version of the library is available at <http://vlado.fmf.uni-lj.si/pub/mixture/>. In the paper, we will present the procedures for dealing with zeros and missing values in compositional data sets (essential and rounded zeros, replacing of zeros with additive, simple and multiplicative replacement strategy), and we will illustrate the use of MixeR with some generated and real data.

#### **References:**

1. Aitchison, J. (1986): *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York, p. 416.

2. Aitchison, J. (1997): The one-hour course in compositional data analysis or compositional data analysis is simple. Pawlowsky-Glahn, V., Cimne, eds., Proceedings of IAMG'97 The 1997 Annual Conference of the International Association for Mathematical Geology, Barcelona (E), Part I, p. 3-35.
3. Aitchison, J. (1992): On Criteria for Measures of Compositional Difference. *Math. Geology*, vol. 24, No. 4, p. 365-379.
4. Martin-Fernandez, J. A., Barcelo-Vidal, C., Pawlowsky-Glahn, V. (2003): Dealing with zeros and missing values in compositional data sets using non-parametric imputation, *Mathematical Geology*, Vol. 35, No. 3, p. 253-278.
5. Martin-Fernandez, J. A., Barcelo-Vidal, C., Pawlowsky-Glahn, V. (1998): Measures of Difference for Compositional Data and Hierarchical Clustering Methods. Buccianti, A., Nardi, G. and Potenza, R., eds., Proceedings of IAMG'98. The 1998 Annual Conference of the International Association for Mathematical Geology, Napoli (I), Part 2, p. 526-531.
6. Martin-Fernandez, J. A., Barcelo-Vidal, C., Bren, M., Pawlowsky-Glahn, V. (1999): A measure of difference for compositional data based on measures of divergence. Lippard, S.J., Nass, A., Sinding - Larsen, R., eds., Proceedings of the 5th Annual Conference of the International Association for Mathematical Geology, Trondheim, Norway, vol. 1, p. 211-215.

---

### **Latent change variable models in studying intra-individual variability in coping measures**

*Vesna Buško and Alija Kulenović* (University of Zagreb, Croatia)

The study demonstrates the applicability of a particular class of structural equation models to the study of intraindividual variations in coping with low-control stress. These structural equation models, as put forward by Steyer, Eid, and Schwenkmezer (1997; Steyer, Partchev, Shannahan, 2000) are specified so that the true intraindividual change scores between two occasions of measurement correspond to the values of endogenous latent variables. Such a specification further enables examining correlates and predictors of true intraindividual change between the measurements.

According to the transactional stress and coping theory, the processes of coping are context-specific and varying across time and the attributes of concrete stressful situations. Nevertheless, substantial interindividual differences in the intraindividual change scores can still be found in the measures of coping with any kind of situational demands. The true change modeling procedures were hence employed

to test the hypotheses on the antecedents and sources of interindividual differences in changes in the use of specific coping strategies.

The data analyzed here were collected within follow-up study on the processes of coping and adjustment of recruits during military service (Buško, Kulenović, 2003). Along with three broad situation-specific coping measures, personality data derived from NEOFFI questionnaire and the cognitive appraisals data, gathered on two occasions on the sample of 421 military basic trainees, were also used in this study. The tested structural equation models and the parameters obtained point to different patterns of relationships of latent state and latent change coping measures with particular personality and appraisal variables. The findings are discussed within the framework of transactional stress and coping theory and the context of low-control stressful situations.

### **A latent class invariant model of first-time clients versus long-term clients receiving drug treatment, controlling for a number of covariates**

*Paul Cahill and Brendan Bunting* (University of Ulster at Magee Campus, Northern Ireland)

This study examines group differences between long-term drug treatment clients and first time clients in the Republic of Ireland in relation to six established homogenous latent classes of drug consumption behaviours. Drug misusers at first contact with treatment services (n=1978) are more reflective of the true nature of drug use behaviour in society, as opposed to long-term drug treatment clients (n=4900). This group difference is first addressed in terms of measurement invariance and then through the introduction of a range of exogenous variables (employment status, gender and living status). The analysis is conducted and the parameter estimates graphically represented using Mplus 3.01. The key differences between long-term drug clients and those never treated before are discussed in terms of intervention strategies for drug misusers at first point of contact.

### **Effect of rounding on expected value and standard deviation**

*Anton Cedilnik, Katarina Košmelj* (University of Ljubljana, Slovenia)

Rounding is a procedure for reporting real-numerical information as a certain arithmetic sequence, in general as  $(\varepsilon + n\delta)_{n \in \mathbb{Z}}$ . In this notation,  $\delta$ ,  $\delta \in \mathbb{R}^+$ , presents the *rounding level*, and the *shift*  $\varepsilon$  is the first non-negative rounded value ( $0 \leq \varepsilon < \delta$ ). The value  $\varepsilon + n\delta$  is the round-off of the interval  $I_n = [\varepsilon + (n - \eta)\delta, \varepsilon + (n + 1 - \eta)\delta]$ , where  $\eta$ ,  $0 \leq \eta < 1$ , if different from  $\frac{1}{2}$ , presents the

*asymmetry* of rounding.

The *rounding function* assigns the value  $\langle x \rangle$  to  $x$  in the following way:

$$x \rightarrow \langle x \rangle := \varepsilon + \text{int} \left( \frac{x - \varepsilon}{\delta} + \eta \right) \cdot \delta$$

where *int* denotes the well-known *integer function*.

The main objective of our work is to study the effect of rounding on the expected value and standard deviation of an arbitrary random variable  $X$ . We study two differences for the expected value:  $E(\langle X \rangle) - E(X)$  and  $\langle E(\langle X \rangle) \rangle - E(X)$ , and similarly for the standard deviation. We prove that the absolute value of each of these four differences is smaller than  $2\delta$ . The obtained estimates are sharp.

The main message of our work is that the effect of rounding on the expected value and standard deviation can be assessed in terms of the parameters of the rounding function. No uncontrolled effect of rounding can be anticipated.

### **Discovering the most important factors for communities of soil microarthropods using machine learning**

*Damjan Demšar, Sašo Džeroski* (Jožef Stefan Institute, Slovenia)

*Paul Henning Krogh, Thomas Larsen* (National Environmental Research Institute, Denmark)

#### **Abstract.**

In agricultural soil a suite of anthropogenic events shapes the ecosystem processes and populations. The risk of impact from anthropogenic sources on the soil environment is almost exclusively assessed for chemicals, although in agriculture other factors like crop and tillage have large impact too. Thus, the farming system as a whole should be evaluated and ranked according to its environmental benefits and impacts. Our starting point is the availability of a data set describing the agricultural events and the soil biological parameters. Using that dataset and machine learning methods for inducing regression and model trees, we produced empirically based models useful for predicting the soil quality in terms of quantities describing the soil microarthropod community from agricultural measures. However, inducing models for predicting soil quality is not our only goal. What we are also interested is to discover additional knowledge on a higher level and identify the most important factors for population densities of springtails and mites and their biodiversity. We do that by preferring smaller and simpler models to bigger and more complex models, while trying to minimize the performance loss of the models at the same time. Using that approach we identify that microarthropod

communities are most sensitive to crops and tillage.

### **Introduction**

With the final task of designing a decision support system for managing farms, we start at a low level, trying to model effects of different farming practices on community of soil microarthropods. The impact of anthropogenic sources on the soil environment is almost exclusively assessed for chemical factors only, although in agriculture other mechanical factors like tillage, and biological factors like crops, also have large impact. And since farming systems consist of a combination of a certain temporal sequence of interdependent events of different type and duration, it is an imperative to handle a farming system as a whole in order to rank its environmental benefits and impacts. One way to do this ranking is to collect information about the agricultural events and the soil biological parameters reflecting those events, and relate the sequence of agricultural events to the biological parameters. Since the collection of data was in progress, we were able to use the already collected data as an input to machine learning algorithms with the primary task of constructing empirically based models useful for predicting the soil quality in terms of quantities describing the soil microarthropod community from agricultural measures. With the final task of constructing a decision support system in mind, we had additional prerequisites. We wanted to identify the most important factors by preferring small and simple models to bigger and more complex models while limiting the performance costs of small models in terms of decreased accuracy. Identifying the most important factors for the community of soil microarthropods can guide us in further experimentation and data collection where we would pay more attention to identified factors. Additional to the discovery of new knowledge (by identifying the most important factors for the community of soil microarthropods), we wanted to "rediscover" knowledge useful for the decision support system that the experts already know but it is hard to transfer because of the gap between different branches of science and can be difficult to put down in writing. Since we prefer the machine learning tools that produce descriptive models, we can use those as a source of questions that could otherwise not be asked without a lot of background knowledge in the domain of agriculture. The answers from the experts will be used (along with other sources of knowledge) to construct the decision support system.

### **Data**

We combined the two available datasets. The first dataset describes four experimental farming systems (Foulum experimental station, Denmark) in the years from 1989 to 1993, allocated to 15 fields, with pesticide use in a conventional system and in two integrated farming systems and no pesticide use on the other (organic) fields, with 530 microarthropod samples collected. The second dataset

describes several organic farms (Foulum and Flakkebjerg experimental stations plus various farms in Jutland) in the year 2002. 430 samples were collected. To those datasets, we added newly available data from 2003, giving us a total of 1330 samples, of which 1192 were suitable for predicting acari species, 1214 for prediction collembola species and 1138 for predicting biodiversity. The datasets available for the study include the agricultural measures (attributes), for example, ploughing, tillage, fertilizer and pesticide use, crops planted and cattle grazing. The history of crops and grazing for the last 3 years is also available. The datasets also contain environmental variables describing the circumstances of the samples where community data on soil microarthropods have been produced. The dataset also includes over 40 measured species, which are aggregated in 3 aggregate measures: acari (biomass of all mites), collembola (biomass of all springtails) and biodiversity of all species.

### **Experiments and (some) results**

Experiments were performed with the data mining software package Weka 3.2. We used regression methods, as the three class variables are continuous. Model trees were induced with the program M5' : simple trees have simplified equations and are induced with the U option, complex trees are induced by M5' with default parameter settings. We also used regression trees and linear regression. The sizes of both model and regression trees were regulated using post-pruning methods. The nearest neighbour method (IBk) with 1, 5 or 10 neighbours was used as a benchmark for comparing accuracy. Each method was applied to each of the three regression problems. For measuring the predictive performance of the model, we evaluated the correlation coefficient and several error measures using ten-fold cross-validation. We evaluated mean average error, root mean square error, relative average error and root relative square error. Collembola models The questions that follow from the collembola models lead us to knowledge like: Deep tillage has less impact with some crops. Crops that include grass/clover provide protection even if the field is deep tilled because the sods will still be intact and the clover residues add a lot of nitrogen to the soil (which enhances microbial life and thus the food base). In the case that the crops are still there in the current year, it means that there has been no tillage plus clover fertilizes the soil. Lupin also fertilizes the soil. Tillage injures/kills Collembola by physical disruption plus destroys their habitat (pathways in the soil are broken and the soil structure is destroyed).

### **Conclusions**

We tried to model a community of soil microarthropods with machine learning methods from the data describing chemical, biological and mechanical actions on the fields. We then used the produced models to identify the most important

parameters for soil mites, springtails and biodiversity of soil microarthropods. By preferring small and simple models to bigger and complex models. We discovered that the most important factor for community of soil microarthropods are previous crops grown in the observed field, and the different forms of tillage. Furthermore, we used the models as a source of questions for the domain experts. We gained knowledge that will help us in further modeling and building decision support system for the management of farms. We have shown that the machine learning models can be used in multiple ways, from predicting new values to gaining new knowledge about the relation between the attributes and the dependent variable, to extracting knowledge from the domain experts.

---

### **A modest agenda of immodest goals for network analysts**

*Patrick Doreian* (University of Pittsburgh, USA)

Over the years, network analysts have accomplished many things with creative and inventive solutions to many technical problems. The field has progressed in an impressive fashion, so much so that some have suggested it is a scientific revolution for the social sciences. This seems a debatable claim. Until we can solve problems that other scholars cannot solve, the claim will be hollow. A suggested list of such problems is this:

1. test statistically the fit of blockmodels in large networks;
2. model the evolution of such fundamental network structures;
3. incorporate network autocorrelation to a wide variety of statistical methods in a dynamic context to predict actor behavior;
4. incorporate massive shocks to network systems analytically; and
5. build multiple-equation methods for the study of whole networks.

At a minimum, solving these problems would create the necessary conditions for claiming a revolutionary status for social network analysis - if we dare to make predictions ahead of phenomena to create the conditions for a genuine test of our methods. Adopting this mind-set might be the hardest task of all!

---

**ML estimation in a random field based forward interest rate model**

*József Gáll, Gyula Pap* (University of Debrecen, Hungary)

*Martien van Zuijlen* (University of Nijmegen, The Netherlands)

In the talk we shall consider forward interest rate curve models in a discrete time setting. Unlike in the classical Heath-Jarrow-Morton models, in our case the forward rate curves corresponding to different time-to-maturity are not driven by the same process, but a random field is supposed to drive the curves. The models at issue were proposed in Gáll, Pap, and Zuijlen [1], [2]. We note that motivation for studying such discrete time models is based on similar continuous time random field models studied, e.g., in Kennedy [5], Goldstein [4] and Santa Clara and Sornette [6].

We consider Gaussian random field cases, in particular, a Gaussian AR field. After showing the necessary no-arbitrage (drift) conditions that we derived for such models and the role of the market price of risk parameters, we turn to the maximum likelihood estimation of the parameters. Despite the lack of explicit solutions, in many cases we can show the (joint) asymptotic normality of the estimators. We also show our results on the strange behaviour of the different parameters regarding consistency. Finally, we discuss the numerical difficulties occurring in the application of the results.

**References:**

1. Gáll, J., G. Pap and M.C.A. van Zuijlen (2002): Forward interest rate curves in discrete time settings driven by random fields, Technical Report No. 0213, University of Nijmegen, The Netherlands.
  2. Gáll, J., Pap, G. and Zuijlen, M.C.A. v. (2003): Limiting connection between discrete and continuous time forward interest rate curve models, *Acta Applicandae Math.*, 78, pp. 137-144.
  3. Gáll, J., Pap, G. and Zuijlen, M.C.A. v. (2004): The maximum likelihood estimator of the volatility of forward rates driven by geometric spatial AR sheet, *Journal of Applied Math.*, accepted.
  4. Goldstein, R. S. (2000): The term structure of interest rates as a random field, *The Review of Financial Studies*, 13, No. 2, 365–384.
  5. Kennedy, D. P. (1994): The Term Structure of Interest Rates as a Gaussian Random Field, *Mathematical Finance* 4, 247–258.
  6. P. Santa-Clara and Didier Sornette (2001): The Dynamics of the Forward Interest Rate Curve with Stochastic String Shocks, *Rev. Financial Studies*, 14, pp. 149–185.
-

## **Quantification of customer satisfaction data: an approach based on the Spline transformation**

*Michele Gallo* (University of Naples "L'Orientale", Italy)

In the literature, several conceptual models are proposed to obtain a correct evaluation of customer satisfaction (CS). Servqual, Servperf, Two-way, Normed Quality and Qualitometro are only some of them (Franceschini, 2001). In this work, we have not dealt with problems associated with the choice of the measurement framework. Only the Servqual model is considered, but most of the results can be generalized to the other models.

The Servqual model has a structure based on a set of attributes and dimensions, where each attribute is evaluated by an item and sets of items give the evaluation of the dimensions. All the items have the same ordinal seven-point rating-scale. An importance or weight is attached to each dimension (or attribute) that is principally used to weight the gap between performance perception and service quality expectation. The weights can be analyzed independently from the gap in order to obtain information on the nature and causes of the interrelationships between the quality dimensions (or attributes). The nature of data should be considered before we carry out a multidimensional analysis. In particular, perception/expectation evaluation has an ordinal scale. This scale establishes an explicit ranking, but not all arithmetic transformations are meaningful because the distances between points on an ordinal scale are not meaningful. The importance data has a constrained ratio scale. For such data, intervals between values and ratios of values are meaningful, but the constraint of the unit-sum of the composition scale causes problems (Aitchison, 1986).

The principal purpose of this paper is to point out the necessity to transform the raw data before we carry out multidimensional statistics analysis, whereby the transformation should respect the original nature of the data. In particular, after a brief review of some techniques for optimal scaling of ordinal data where the optimal scaling is defined in terms of the correlation matrix of quantified variables (De Leeuw and van Rijkevorsel, 1980), we propose a new way to quantify the perceived/expected data based on the Alternative Least Squares (ALS) algorithm minimizing the loss function and preserving the different subjective scale of each customer. Finally, a comparison with Rating Scale Analysis (Wright and Masters, 1982), Thurston Quantification (Zanella, 1999) and Princals (Gifi, 1982) is presented.

### **References**

1. Aitchison J. (1986): The statistical analysis of compositional data, Chapman and Hall.

2. D'Ambra L., Amenta P., Gallo M. (2002): Riflessioni sulla Valutazione dei Servizi di Day Surgery nel contesto dell'Analisi Multidimensionale dei dati. A cura di
3. B.V.Frosini, U. Magagnoli, G. Boari. Vita Pensieri ISBN 88-343-0945-6.
4. De Leeuw J., van Rijckevorsel J. (1980): Homals en Princals. In E. Diday et al. (eds), *Data Analysis and Informatics*.
5. Franceschini F. (2001): Dai prodotti ai servizi. Le nuove frontiere per la misura della qualit. UTET Libreria.
6. Gallo M. (2003): Partial Least Squares for Compositional Data: an approach based on the splines. *Italian Journal of Applied Statistics*, vol. 3, 2003.
7. Gallo M. (2003): An alternative approach based on the B-Spline Transformation for the quantification of Customer Satisfaction Data. Submitted to *Caribbean Journal of Mathematical and Computing Sciences*.
8. Gifi A. (1982): *Princals user's guide*. Leiden: Department of Data Theory.
9. Wright B. D., Masters G. N. (1982): *Rating Scale Analysis*, Chicago: MESA Press.
10. Zanella A. (1999): A Stochastic model for the analysis of customer satisfaction: some theoretical aspects. *Statistica*, LIX.

---

### **Contingent versus unconditional incentives in WWW-studies**

*Anja S. Göritz* (University of Erlangen-Nürnberg, Germany)

Four experiments examined whether participation in Web based studies was influenced by such a simple measure as framing the reception of an incentive as being contingent on the completeness of the submitted questionnaire.

Three experiments were carried out in a university-based online access panel and one in a market research online panel. Three times the incentive was a prize draw and once it was a personal gift. Two conditions were contrasted in the experiments: one group received an invitation e-mail mentioning that all participants are eligible for the incentive, whereas the other group was told that only those participants who completely fill-out the questionnaire would receive the incentive.

Dependent measures to capture willingness to participate were response rate, dropout rate, and composition of the sample. Dependent measures to capture data quality were number of omitted closed-ended items, length of answers to open-ended

questions, and stereotypical answering of matrix-questions. In all experiments, consistent patterns were found suggesting that conditional incentives decrease response and dropout and influence data quality.

However, in neither study these tendencies reached a conventional level of statistical significance. Therefore, individual results were meta-analytically integrated to find out whether the four single experiments were merely underpowered to detect a small effect.

---

### **Statistical principles to live by**

*Frank E. Harrell Jr.* (Vanderbilt University, USA)

This talk deals with principles derived from over 30 years of applying statistics to biomedical research, collaborating with clinical and basic biological researchers and epidemiologists. The principles relate to statistical efficiency, bias, validity, robustness, interpretation of statistical results, multivariable predictive modeling, statistical computing, and graphical presentation of information. Topics to be discussed include respecting continuous variables, avoiding non-descriptive statistics, problems associated with filtering out negative results, overfitting, shrinkage, adjusting P-values for multiple comparisons without adjusting point estimates for same, and the false promise of multi-stage estimation and testing procedures, related to the use of bogus conditional techniques for computing what is advertised as unconditional variances or type I errors.

---

### **Estimation of optimal adaptive treatments using observational data: a regret-based approach to dynamic anticoagulation**

*Robin Henderson* (Lancaster University, UK)

*Catherine Fullwood* (Lancaster University, UK)

*Syd Stewart* (4S Systems, UK)

We describe an analysis of observational longitudinal anticoagulation data, aimed at determining an optimal reactive dose-changing strategy. We use the regret parameterisation approach recently advocated by Murphy (2003) with additional diagnostic assessment techniques. We believe this to be the first genuinely practical application of Murphy's ideas.

Anticoagulants are administered to patients with high risk of internal blood clotting, such as people with history or high risk of strokes, arterial or venous thrombosis, with certain liver diseases or following heart valve replacements. Blood

thickness is measured through the International Normalised Ratio (INR), a standardised measure of prothrombin time, which is the time it takes plasma to clot. If INR is too high then there is risk of internal bleeding whereas if it is too low there is risk of thrombosis. The aim of anticoagulation is to maintain INR within a given target interval.

At clinic visits by patients on long-term anticoagulation the INR is determined and a decision taken whether or not to change the dosage, with the aim being to find a stable level for that patient. Dose levels may be selected by the clinician or automatically by computer software (Poller et al, 1998). Since INR is highly variable between patients and strongly dependent on diet and other lifestyle characteristics, optimal dosage is both dynamic and patient-specific.

We have data on 303 patients orally treated with warfarin and for this analysis we consider the first 14 clinic visits after induction. The first 4 of these are used as reference data, leaving 10 visits per patient for our main analysis. At visit  $j$  we have  $S_j$ , a standardised measure of the distance of INR from target range, and  $A_j$ , the action taken at the end of the visit. In practice  $A_j$  has three components: a decision whether or not to change dose, the new dose if change is selected, and the timing of the next visit. Visit times are not explicitly considered in this analysis so that  $A_j$  is considered as a scalar quantitative variable measuring the dose change. This has a mixed distribution with a discrete component  $A_j = 0$  if there is no change and is otherwise in principle continuous, though in practice limited by the available tablet strengths.

This set-up so far is essentially that described by Murphy, in her seminal Royal Statistical Society discussion paper (Murphy, 2003). To complete it we need a final response  $Y$  which is to be optimised. After discussion, we decided to use the mean *percentage time in range* as the basis of the target. At each visit linear interpolation between the current and previous INR values were used to estimate the proportion of time between visits during which the INR was in the target range, expressed as a percentage  $P_j$  say. We then took

$$Y = \left(\frac{1}{n}\right) \sum_{j=1}^n P_j$$

so that the optimal value is 100.

The Murphy procedure will be described. It is based on choosing a parametric form for the regret  $\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j)$  at time point  $j$ . Here  $\bar{S}_j = (S_1, S_2, \dots, S_j)$  and  $\bar{A}_{j-1} = (A_1, A_2, \dots, A_{j-1})$  represent the observed history up to visit  $j$  and  $a_j$  is the to-be-decided action at the end of the visit. The regret  $\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j)$  is considered as a function of  $a_j$  and is defined to be the difference between the expected  $Y$  given the best possible action and that if  $a_j$  is selected, in each case assuming that future actions are optimal. A modified form of regret function as-

suming that future actions follow the observable distribution will be described in the paper.

Estimation using two parametric regret functions will be described and their results compared. One assumes regret increases quadratically as action moves away from the optimal, whereas the other imposes a bound to reflect the upper bound of 100 in  $Y$ . In both cases we find that there is a strong penalty for any change in dose if  $S_j = 0$  and INR is in range, whatever the previous history  $\bar{S}_j$  and  $\bar{A}_{j-1}$ . Otherwise the optimal change is proportional to  $S_j$  and inference for the proportionality constant will be discussed. The consequence of following our estimated optimal regime can be assessed, leading to an increase from an observed value in the region of 50% mean time in range to an estimated value just over 60%. This gain is worthwhile yet modest and reflects the high level of between and within patient variability seen in data of this type.

**References:**

1. Murphy, S. (2003): Optimal dynamic treatment regimes (with discussion). *J Royal Statistical Society Series B*, 65, 331-366.
2. Poller, L., Shiach, C.R., MacCallum, P.K. Johansen, A.M., Munster, A.M., Magalhaes, A. and Jespersen, J. (1998): Multicentre randomised study of computerised anticoagulant dosage. *Lancet*, 352, 1505-1509.

---

**Social integration of elderly in Slovenia**

*Valentina Hlebec* (University of Ljubljana, Slovenia)

Owing to growing number of elderly it is necessary to study their social integration and potential sources of social support. Based on representative sample of elderly in Slovenia (2002), six distinct types of social support networks were found, using hierarchical and k-means cluster analysis. These six types of social support networks can be classified into three qualitative types. Local family dependent social support networks represent 65% of all networks, locally integrated social support networks represent 15% of all networks and wider community focused support networks represent 21% of the sample. Gender differences are studied and discussed with regard to capacity of social networks to provide social support and characteristics of obtained groups. Special attention is paid to isolated individuals and individuals with very small networks.

---

---

**Social networks in Slovenia: A comparative perspective**

*Valentina Hlebec and Tina Kogovšek* (University of Ljubljana, Slovenia)

Informal social networks are the most important source of social support, which is an essential foundation of the quality of every-day life. Distributions of various types of social networks have to be studied in comparative perspective to evaluate the effects of change of political, social and economic systems on social network composition and structure.

Data from two studies, done before (1987) and after the transition (2002) on representative samples of adult residents of Slovenia, are compared. In the paper, the ability of informal social networks to provide adequate source of social support is discussed as substantive part of this research. The effects of characteristics of measurement instruments on obtained estimates of network composition are presented and evaluated. Advantages and disadvantages of the relationship approach to measuring personal networks are discussed with regard to complete evaluation of network membership.

---

**A hierarchical bayes multilayer perceptron to analyze heterogeneous sales response**

*Harald Hruschka* (University of Regensburg, Germany)

This paper deals with the question whether for a typical marketing data set functional flexibility produced by a multilayer perceptron is advantageous if heterogeneity is considered as well. The multilayer perceptron is compared to a strict parametric model which also has heterogeneous (i.e. store-specific) coefficients. Both models are specified in a hierarchical Bayesian framework.

There are a few papers estimating multilayer perceptrons by Bayesian approaches. One of these approaches, the evidence framework of MacKay (1992), is essentially a Laplace approximation to the posterior distribution. Alternatively, homogeneous multilayer perceptrons have been estimated by Markov Chain Monte Carlo (MCMC) techniques. But none of these contributions develop or apply a method to estimate multilayer perceptrons with heterogeneous coefficients. In this paper an appropriate estimation technique is introduced which is also capable to satisfy constraints postulated by economic theory. This MCMC technique is based on a method developed by Train (2003). After convergence to stationarity it generates random samples of parameters from the joint posterior density. Statistics of sampled values (e.g. means, percentiles) converge to their population values.

Model performance is evaluated by posterior model probabilities in accordance with the dominant approach in Bayesian statistics. Posterior model probabilities penalize models for complexity, i.e. all else being equal the more complex model receives a lower value. In accordance with a proposal made by Raftery (1996) the multilayer perceptron is judged to perform better if its posterior model probability is greater than 0.75. The empirical study refers to a data base consisting of weekly observations of sales and prices for nine leading brands of a packaged consumer good category (frozen orange juice). Data were acquired in 81 stores. Between 61 and 88 weeks per store lead to a total of at least 4,941 observations per brand. The multilayer perceptron is compared to a strict parametric multiplicative model which is the most popular model for this type of data. Posterior model probabilities of the multilayer perceptrons clearly indicate that using a flexible model is beneficial even if heterogeneity is dealt with. Moreover, price effects implied by the multilayer perceptron differ from those for the conventional multiplicative model for eight out of nine brands, especially at high prices of competitors.

---

**Assessing the "accuracy" of proxy-reports on attitudes towards immigrants in Germany. What does the analysis of the congruence of proxy- and self-reports on attitudes towards immigrants in Germany tell us?**

*Angela Jäger* (University of Mannheim, Germany)

The paper analyses the degree to which subjects are able to report "accurately" the attitudes of their reference group or close discussion partners towards immigrants in Germany. It focuses on factors which affect the probability that ego's proxy-reports are in congruence with alter's self-reports on these attitudes. Additionally, attention will be given to the restriction of the analysis of congruence with respect to identifying indicators of exchange of attitudinal information between ego and alter, and to assessing the extent to which egos' project their own attitudes onto their alter.

To which extent proxy-reports represent useful information on ego's personal reference setting is of fundamental importance for the use of (egocentric) network data. Especially in analyzing the influence of attitudes and opinions of important discussion partners on the attitudes of actors, it is necessary to distinguish between the perceived attributes and attitudes of alteri reported by ego and the actual characteristics of the alteri. Studies using a social network approach have found evidence of a strong association between respondents own attitude reports and their reports on the perceived attitudes of relevant others. Since most researchers have not tested whether this perception of ego is congruent with the actual attitudes of the alteri, it remains unclear to what degree the observed correlations are substan-

tiated in social influence processes or simply result from egos projection of their own attitudes onto their alteri. Apart from this essential question of the quality of proxy-reports, there are few empirical and theoretical insights. Furthermore, past studies are seriously limited in that they consider only determinants on dyadic level, e.g., information transmission between ego and alter or the intensity of the relationship. In doing so, the embeddedness of the dyad in (egocentric) networks has been neglected, as well as the fact that the accuracy depends simultaneously on two different answering processes.

One way to assess the accuracy of ego's proxy-reports is the analysis of the congruence between ego's proxy-reports and the corresponding self-reports of the alteri. Using this measure the amount as well as the causes of the congruence between ego's proxy-reports on attitudes of alteri and the actual self-reports of the alteri can be studied, and determinants of the accuracy can be identified. Another way to assess the accuracy of ego's proxy-reports is the analysis of the extent to which the proxy-reports can be explained by ego's own attitudes or by the actual attitudes of the alteri. Using this approach, a specific problem of proxy-reports on attitudes can be addressed. Proxy-reporters could rely on their knowledge about alter's specific attitude in question, as well as utilize other information as attitude-related general dispositions of alter. In both cases, proxy-reports don't simply result from egos' projection of their own attitudes, but reflect alter's attitudes.

This paper addresses the following questions: Is the accuracy of proxy-reports on attitudes towards immigrants affected by characteristics of the social relationship between ego and alter? How relevant are the perceived or actual features of the whole network? Which role does the direct information exchange between ego and alter play? To which extent do the proxy-reports reflect the attitude of the alteri? In order to answer these questions, we use data collected in 2002 ( $N = 1.693$ ) about egos' attitudes towards immigrants, including information about egos' perception of alter's attitudes and alter's self-reports in this respect. The proxy-reports are measured by items on the preferences about immigration of specific groups. Additionally, the self-reports on attitudes towards immigrants include items about the agreement with discriminating statements about "foreigners". To assess the accuracy of the proxy-reports, both approaches are used.

According to our results, the accuracy of proxy-reports is conditional on respondents' accuracy motivation, as well as on ego's information availability indicated by characteristics of the dyads and the whole network in which these dyads are embedded. Egos in dense and actual homogenous networks regarding the attitude were significantly more able to report alter's attitudes correctly. The frequency of contact and the emotional closeness are found to affect the congruence of the proxy-reports to a minor degree and in a non-additive way.

The discussion about immigrants in Germany as an indicator of the direct information exchange between ego and alter doesn't enhance the congruence. Here the

restriction of the analysis of the congruence of proxy-reports on alteri is apparent. The analyses to which extent the proxy-reports can be explained by ego's own attitudes or by the actual attitudes of the alteri show that in the case of discussions about immigrants, the proxy-reports reflect to a greater extent the general disposition of the alteri than their specific attitude in question. By contrast, the proxy-reports of respondents in dense networks reflect more the specific attitude of the alteri. The results indicate that proxy-reporting could be a reasonable means of gathering informative data on attitudinal dispositions of alteri, but not on specific attitudes.

---

### **Information and communication technologies usage in Slovenian households and enterprises**

*Andreja Kačič, Eva Belak, Rudi Seljak* (Statistical Office of the Republic of Slovenia)

In the spring of 2004, the Statistical Office of the Republic of Slovenia conducted two surveys for the first time: information and communication technologies (ICT) usage in households and ICT usage in enterprises. The research in this field is not new in Slovenia, since numerous surveys have been conducted in the framework of the RIS project. The novelty with this year's surveys is that the methodology (questionnaires, target population) of both surveys is in line with the Eurostat guidelines.

The goal of the enterprise survey was to find out whether and how enterprises are equipped with computers, whether and for which purposes they use the internet etc. The goal of the household survey was to measure the access of households and individuals to different ICTs. With the survey, we can identify factors influencing household's (individual's) access to a particular technology and it also enables us to analyse the purpose of using a particular technology.

The methodology and some results of both surveys will be presented and commented. Results will be compared with the results of other countries and also with the results of previous related surveys.

---

---

**Analysis of example weighting in subgroup discovery by comparison of three algorithms on a real-life data set**

*Branko Kavšek, Nada Lavrač* (Jožef Stefan Institute, Slovenia)

This paper investigates the implications of example weighting in subgroup discovery by comparing three state-of-the-art subgroup discovery algorithms - APRIORI-SD, CN2-SD, and SubgroupMiner - on a real-life data set. While both APRIORI-SD and CN2-SD use example weighting in the process of subgroup discovery, SubgroupMiner does not. APRIORI-SD uses example weighting in the post-processing step of selecting the 'best' rules, while CN2-SD uses example weighting during rule induction.

The results of the application of the three subgroup discovery algorithms on a real-life data set - the UK Traffic challenge data set - are presented in the form of ROC curves. They show that APRIORI-SD slightly outperforms CN2-SD; both APRIORI-SD and CN2-SD are good in finding small and highly accurate subgroups (describing minority classes), while SubgroupMiner found larger and less accurate subgroups (describing the majority class). By using ROC analysis, we show that these results are not surprising and can be attributed to example weighting, which "pushes" the search in the space of potential subgroups towards discovering small and accurate subgroups.

---

**Analysis of US patents network in time**

*Nataša Kejžar and Vladimir Batagelj* (University of Ljubljana, Slovenia)

The network of US patents from 1963 to 1999 (Hall, Jaffe, Tratjenberg 2001, USPTO) is an example of a very large citation network (3774768 vertices and 16522438 arcs).

We selected a group of companies appearing in the main themes in the network. The themes were determined using islands algorithm (Zaveršnik, Batagelj, 2004) for the Search Path Count (SPC) weights (Hummon, Doreian 1989, Batagelj 2003).

We split the complete network into time slices according to time intervals of assigned patents. On these slices we observe the features of the patents assigned to the selected companies, such as autocitations, the SPC weights, number of patents etc.

By studying the development patterns of the network in time, we are trying to determine how the research and development in the companies has evolved over the last thirty years and what are the trends in main themes in the network.

**References:**

1. Batagelj V.: Efficient Algorithms for Citation Network Analysis. Submitted, 2003. <http://www.arxiv.org/abs/cs.DL/0309023>
2. Hall, B.H., Jaffe, A.B. and Tratjenberg M. (2001): The NBER U.S. Patent Citations Data File. NBER Working Paper 8498. <http://www.nber.org/patents/>
3. Hummon N.P., Doreian P. (1989): Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks*, 11, 39-63.
4. (USPTO) The United States Patent and Trademark Office. <http://patft.uspto.gov/netathtml/srchnum.htm>
5. Zaveršnik M., Batagelj V.: Islands - identifying themes in large networks. In preparation.

---

**Analysis of excretion of pineal gland hormone melatonin in ex-mercury miners with machine learning methods**

*Alfred B. Kobal* (Idrija Mercury Mine, Slovenia)

*Bernard Ženko* (Jožef Stefan Institute, Slovenia)

*Darja Kobal* (University Medical Centre, Slovenia)

*Mladen Krsnik* (University Medical Centre, Slovenia)

*Sašo Džeroski* (Jožef Stefan Institute, Slovenia)

*Milena Horvat* (Jožef Stefan Institute, Slovenia)

*Josko Osredkar* (University Medical Centre, Slovenia)

**Introduction**

The toxic effects of mercury are well known and in the case of occupational exposure to elemental mercury vapour ( $Hg^{\circ}$ ) the most frequent symptoms and signs include erythrim, increased irritability, depression, insomnia, psychotic disturbances, tremour and renal impairment. High Hg and Se retention and co-accumulation have been found in the brain, endocrine glands and also pineal gland, kidney and other body tissues in ex-miners from the Idrija Mercury Mine even several years after exposure. This could be associated with possible health problems or adverse biological effects. The long term effects of mercury toxicity are not well studied. Recently the pineal hormone melatonin was found to have a potent-free radical scavenging activity and the protective effect of melatonin towards peroxidative damage was found in some in vivo and in vitro studies. High Hg accumulation has been found in the pineal gland in retired miners which could

modify the synthesis of melatonin. There are no data available in the scientific literature on the possible effects of Hg on melatonin excretion. The purpose of this study was to investigate the impact of occupational mercury exposure on melatonin excretion in mercury miners.

### **Data set and methodology**

Initially, 120 males were examined in the study. After the selection procedure, the study population comprised 53 ex-mercury miners previously exposed to Hg and 53 workers in the control group. The study group of miners comprised 33 active miners not exposed to Hg in the preceding 8 to 60 months, and 20 retired miners who had not been exposed to Hg before the present observations for a period from 32 to 336 months. The miners were employed in the Idrija Mercury Mine. The control workers were taken from "mercury-free" works. They performed jobs in the forests as choppers and transport workers. The medical examination of all subjects included a medical history and lifestyle habits (age, body mass index, smoking, alcohol consumption, dental amalgam score). The examination also included venous blood and urine sampling for determination of blood total (BT-Hg) and methyl mercury (Me-Hg), urine mercury (U-Hg), selected hematological data, selected blood and urinary data of kidney urinary tract disorders, serum gamma glutamyltransferase (GGT), aminotransferases (ALT, AST), bilirubin, blood glucose, c-reactive protein and concentrations of melatonin in blood in urine. Environmental and biological data on the group of miners studied were collected from 1959 onwards from workload records, daily reports on Hg measurements in the workplace, personal medical records and biological monitoring data. On this basis, several parameters of the duration and level of exposure were calculated for each miner, such as years of work in the mercury mine, cycles of exposure (intervals of work at exposure to  $\text{Hg}^\circ$ ), average time-weighted (ATW) air  $\text{Hg}^\circ$  concentration expressed in  $\text{mgHg}^\circ/\text{m}^3$  air.

The data were analysed with machine learning methods in order to gain insight in the background of the melatonin excretion process. We have used the algorithm for induction of model trees, which can be interpreted by a domain expert. A model tree is similar to a regression tree, except that it has linear equations instead of a single class label in the leaves. The models presented in this study were built with the M5' model tree learning algorithm as implemented in the Weka data mining environment. The default parameters of M5' were used. We have built two models, one for the concentration of melatonin in blood and another for melatonin (melatonin sulfate) in urine, but because of space limitations we only present the latter model in this abstract.

### **Results and discussion**

The following model was built for the concentration of melatonin in urine. The

tree has only one node, i.e., one linear equation:

$$\begin{aligned}
 \text{melatonin\_in\_urine} = & \\
 -1.1228 * \text{age} + & [25/64/44.6/45/ - 50.0] \\
 -0.8144 * \text{no\_of\_cigarettes\_per\_day} + & [0/40/9.4/0/ - 7.66] \\
 0.5893 * \text{no\_of\_years\_smoking} + & [0/40/9.3/0/5.48] \\
 -1.8774 * \text{body\_mass\_index} + & [16.8/38.5/26.8/26.0/ - 50.3] \\
 0.5118 * \text{sistolic\_blood\_press} + & [100/170/129.4/128/66.2] \\
 1.3376 * \text{se\_in\_urine} + & [7.2/41.3/18.4/17.5/24.6] \\
 8.8882 * \text{cd\_in\_blood} + & [0.23/9.51/1.58/11.1/14.0] \\
 -0.6878 * \text{pb\_in\_total\_blood} + & [14.7/91.3/48.3/49.6/ - 33.2] \\
 -29.2923 * \text{mda} + & [0.001/1.26/0.19/0.081/ - 5.57] \\
 -0.0026 * \text{u - hg\_sum\_max} + & [0/11365/1831/0/ - 4.76] \\
 79.6388 &
 \end{aligned}$$

The numbers in the brackets at each attribute are as follows: min value, max value, average value, median value and relative importance factor, where the latter is calculated as a product of the coefficient and the average value. Correlation of the model predictions with the actual values (calculated on training data) is 0.60.

Age and body mass index are the most important factors that decrease the concentration of melatonin in urine, which is also known from previous studies. Factors that promote the formation of free radicals (Pb, Cd, Hg, and tobacco smoking) and affect the peroxidative degradation of the lipids and phospholipids of cellular membrane additionally decrease the concentration of melatonin in urine. In accordance with this is the negative influence of a lipid peroxidation product (MDA - malondialdehyde) on the melatonin concentration. This supports the assumption that melatonin also has antioxidative effect. In the process of antioxidation activity, melatonin decomposes, and we assume this is the reason for decreased excretion of melatonin in mercury miners.

### **The effects of limiting the number of alters on composition and structure of social support networks**

*Tina Kogovšek and Valentina Hlebec* (University of Ljubljana, Slovenia)

Social networks can be operationalized as potential and actual source of social support. Comparing cross-sectional data sets from various points in time allows us to compare overall characteristics of social support networks and their changes over time in the population. However, the measurement instruments should be identical or at least very similar.

Burt's name generator (1984) for assessing discussion networks of Slovenians

was applied in the 1987 and 2002 surveys. Data were collected on representative samples of adult residents of Slovenia. Owing to the complexity of the personal network questionnaire, the number of alters was limited to 5 in the 1987 survey, whereas there was no upper limit in the 2002 survey. In the paper, both surveys are compared with regard to the network composition and structure. Special attention is given to the effects of limiting the number of alters on statistical estimates of network composition and structure.

### Bayesian estimation of kinetic parameters: a particle approach

*F. Konecny* (BOKU - University of Natural Resources and Applied Life Sciences, Austria)

Monod's microbial growth model (*Monod*, 1941) has long been found to have wide application in the fields of biotechnology and environmental engineering. It defines a relation between nutrient utilization and the growth of microorganisms:

$$\frac{dx}{dt} = \mu_m \frac{s}{K_s + s} x - K_d x, \quad \frac{ds}{dt} = -\frac{\mu_m}{Y} \frac{s}{K_s + s} x,$$

where

- $x$  = concentration of microorganisms,
- $s$  = concentration of growth limiting substrate,
- $\mu_m$  = maximal specific growth rate,
- $K_s$  = substrate saturation constant,
- $Y$  = yield coefficient, and
- $K_d$  = decay rate.

A typical batch experiment starts with an inoculum ( $x(0) = x_0$ ) and an initial substrate concentration ( $s(0) = s_0$ ) and consists of observations of biomass and/or substrate concentrations. The final biomass is reached, when all substrate is exhausted. The yield is a real measurable quantity given by

$$Y = \frac{\text{specific substrate uptake rate}}{\text{specific biomass growth rate}}.$$

In biotechnical applications the parameters  $\mu_m$  and  $K_s$  are often considered to have biological significance and are treated as characteristics of the process. Practical experience in determining the growth parameters of different growth processes, however, led to a critical analysis of the practical identifiability of the Monod parameters. The following difficulties were encountered in connection with parameter estimation (*Holmberg*, 1982):

1. Unique sets of parameters could rarely be obtained, as many sets seemed to explain the measurements equally well.
2. Parameter estimated from data obtained during apparently similar conditions showed great variations.
3. Nonlinear regression routines showed poor convergence properties.

This paper examines the problem of parameter uncertainty and its effect on the prediction of microbial in a Bayesian framework. We view the estimation problem as an optimal filtering problem based on noise-contaminated observations of the substrate concentration at discrete observation times

$$y_k = s(t_k; \theta) + w_k,$$

where  $\theta = (\mu_m, K_s)$  and  $(w_k)$  is white noise. We adopt a particle filter (*Doucet et al.*, 2001) to compute the the optimal filter and (approximative) posterior distributions of the parameters to be estimated. This particle method is compared to the classical approach in application to simulated data sets and to Monod's original data for the growth of *Escherichia coli* in a batch experiment with lactose as the limiting substrate to identify similarities and differences between both methods.

#### References:

1. Doucet, A., de Freitas, J.F.G. and Gordon, N.J. (eds.) (2001): Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag.
2. Holmberg, A. (1982): On the practical Identifiability of Microbial Growth Models Incorporating Michaelis-Menten Type Nonlinearities. *Math.Biosciences* 62, 23-43.
3. Monod, J. (1941): Recherches sur croissance des cultures bacteriennes. These de docteur es sciences naturelles, Paris.

#### Analysis of the buyers' choice networks

*Simona Korenjak-Černe, Nataša Kejžar, Vladimir Batagelj* (University of Ljubljana)

Buyer's choice implies some kind of relations among products. Products and their relations produce buyers' choice network. Analyses of such networks can offer interesting information for marketing. Our purpose is to analyze the networks obtained from Amazon Internet bookstore and CD-store. All the analyses were done with the program Pajek.

### **Sub-sampling of alters in collecting the data on ego-centred social networks**

*Luka Kronegger, Gašper Koren, Valentina Hlebec and Vasja Vehovar*  
(University of Ljubljana, Slovenia)

The importance of the data on ego-centred social networks is lately highly increasing within social research. Collecting the data in this field is, however, not trivial. Usually, owing to their complex structure, it represents a difficult and demanding task for the respondent and therefore increases the survey costs. In the past, interviewer assisted data collection modes were preferred over the self-administered modes. However, recent research within this field shows that self-administered data collection modes can be sufficiently used also for collecting the data on complex issues, such as ego-centered social networks. On the one hand, self-administered modes make easier for the researchers to collect the data on more delicate issues, but on the other hand, they shift all the burden of the survey process to the respondent. The problem which arises is that within self-administered survey modes, the respondent can easily quit the survey process at any stage, which questions the reliability and validity of the collected data.

The survey process can be divided into two stages. Within the first stage, the respondent (i.e. ego) provides the list of names (i.e. alters) from his/her social network. The network is defined with the social tie which connects the ego with his/her alters. In the second stage, the ego reposts more detailed information on each alter from his/her network. So the ego is the main source, which provides all necessary information. This second stage is very demanding, while the ego should compile all the information about the alters and report them through the questionnaire. With the increasing number of listed alters (i.e. larger networks), the number of questions posed to the ego increases. That prolongs the questionnaire and, by repeating questions for each of the alters, increases the burden put on the respondent.

Reducing the number of questions within the survey process can be achieved through sub-sampling of the alters listed within the first stage of the survey process. Instead of collecting the details on each listed alter, the ego can be asked detailed questions for only a smaller number of alters, who are sub-sampled from the reported network.

This method can provide sufficient information about the characteristics of the network, while it reduces the effort of the respondent, as well as the time needed for collecting the data and the survey costs. In this paper we focus on the pay-offs between the corresponding loss of information on one side and the benefits arising from reduced respondent burden and survey costs on the other side.

As it is shown, this approach can be sufficiently applied in different self-administered survey modes as well as interviewer-administered survey modes. Usage of computer-

assisted data collection is preferable due to the automation of the alter sampling within the sampling procedure, which is conducted in real time.

---

### **Parsimonious modelling of time-dependent effects in the Cox model**

*Stephan Lehr, Harald Heinzl and Michael Schemper* (Medical University of Vienna, Austria)

The influence of prognostic factors on survival is often evaluated by the proportional hazards model (Cox, 1972). The model in its original form assumes the effect of a covariate on survival to be constant over the whole follow-up period. This assumption may not hold in some clinical studies, for instance, if the importance of a treatment diminishes in the course of time. Various methods have been proposed to extend the original model to allow for time-dependent effects of the prognostic factors. These methods can roughly be classified into (i) inclusion of interactions with parametric functions of time, (ii) partitioning the time axis and piecewise estimation and (iii) penalized likelihood approaches.

The empirical support provided by clinical data sets of typical sample sizes and censoring proportions might be far too weak for confirming more complex time-dependencies than monotonic ones and the 'price' to be paid for using the flexible approaches could be a substantial loss in power to confirm any time-dependent effect or even any effect of the prognostic factor.

In this presentation, we quantify the amount of over-fit and its impact on testing power if time-dependent effects are analysed by means of varyingly flexible approaches. Based on a simulation study depicting scenarios likely in practical situations, guidelines are established that guard against the unduly increase in the variability of the estimates. The power of the different methods is assessed by specifying certain time-varying effects in a background population. The performance of the various approaches is compared by taking into account the degrees of freedom used for estimation, and the issue whether information criteria can provide suitable guidance on the most adequate modelling of time-dependent effects, is addressed.

We can conclude that for a fixed number of parameters to model time-dependence of prognostic factors, differences in over-fit between the methods considered tend to be small. However, an increase of such parameters has a large effect on over-fit. Therefore, unless there is no scientific evidence for complex time-dependencies, the modelling approach should be based on the principle of parsimony and the proportional hazards assumption should be checked only against simple alternatives. Furthermore, a hierarchical approach due to Ambler and Royston (2001) and the Bayesian information criterion are especially considered, since they both provide

parsimonious guidance and are powerful enough to detect substantial deviations from a constant effect.

**References:**

1. Cox, D.R. (1972): Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187-220.
2. Ambler G. and Royston P. (2001): Fractional polynomial model selection procedures: investigation of type I error rate. *Journal of Statistical Computation and Simulation* 69, 89-108.

---

**Effect of "grid" questions in Web survey questionnaires**

*Katja Lozar Manfreda, Gašper Koren, Valentina Hlebec*  
(University of Ljubljana, Slovenia)

Survey methodologists have already performed quite some testing regarding Web survey questionnaire design in order to efficiently collect survey data via the Web. However, in comparison to the amount of research dealing with questionnaires in other survey modes, these studies represent only the first steps towards systematic research regarding possible effects of Web questionnaires on non-response and measurement survey errors.

Our study contributes to the body of knowledge on designing principles of Web questionnaires. We focus our attention to the problem of "grid" questions, where items are presented in a matrix form, i.e., several items with the same answer categories are positioned together in a table. The rationale behind using this form is the fact that the questionnaire appears shorter than when each item is presented separately as an individual question. However, "grid" questions may be unsuitable if there is a tendency to answer questions uniformly. The cognitive effort in the case of "grid" questions is smaller than it would be if respondents always had to choose an answer from separate items. This problem may be even larger in Web questionnaires due to two reasons. First, respondents tend to read and perform tasks on the Web very quickly as if they were "scanning" the content. In addition, the problem may be larger owing to the use of mouse and cursor. The effort involved in positioning the cursor to a new position in each row is larger than that of continuing in the same or neighboring column for each row. This effect may occur more often the less interested respondents are in the survey, the more tired or inattentive (bored) they already are, and the more difficult it is for them to find the right answer.

We are thus testing how does the presentation of a Likert-scale set of questions in a grid table or separately influence the collected data. The experiment addressing

this issue was implemented in a Web survey among the first grade students at the Faculty of Social Sciences, Ljubljana, Slovenia, in spring 2002. Three hundred fifty respondents were randomly administered to different experimental conditions using split-ballot experimental design. We will show that some of the basic principles for designing self-administered questionnaires from paper-and-pencil self-administered modes hold true also for Web surveys. However, specifics can also occur. We must regard the Web as a special medium with special design options, visual features and required respondent's actions. These all require a special treatment of the Web questionnaire.

---

### **The problem of concave sets with hierarchical clustering**

*Mateja Nagode, Matej Francetič, Bojan Nastav* (University of Ljubljana, Slovenia)

Clustering methods are among the most widely used methods in multivariate analysis. Roughly speaking, there are two main groups of clustering methods: hierarchical and non-hierarchical. While the latter are mainly used in applied statistics, our paper focuses on hierarchical methods due to the nature of the problem. It is assumed that nonhierarchical methods (the most widely used being the K-means method) are an inadequate and inappropriate approach with concave set of points and that the centroids of the groups may lead to false conclusions. Hence, hierarchical methods, such as the nearest neighbour, the farthest neighbour, the Wards, between-groups linkage, within-groups linkage, centroid and median clustering methods, are examined.

The paper's primal goal is to assess the clustering methods' validity and successfulness when using concave sets, and to establish in which types of data structures can the methods reveal and correctly assign group membership. Sets of points differing in shape and inter-point distance were used in the analysis. Since each point was defined once by two variables and once by three variables, our simulations were run in a two- and three-dimensional space.

Knowing the actual state (cluster membership) is essential in comparing clustering methods. Thus, the analysis was based on generating sets of points. Three parameters defined the simulations: the skeleton (defining the shape of a set), the standard deviation of points around the skeleton, and the number of points generated. This generated various shapes of sets with different inter-cluster distances (from clearly separate groups to almost completely overlapping). Certain limitations were imposed, since the used parameters can lead to a vast number of generated sets. Applying different hierarchical clustering methods to these generated sets was the basis for assessing clustering accuracy and hence the validity of

different hierarchical clustering methods for concave sets.

The paper consists of two parts. First, we review the published work in this field and introduce our research methodology. In the second part, we report how we generated the data sets and present the results of assessing how well different clustering methods perform on the generated data. We believe that our conclusions are important and interesting since real-life data seldom follow the simple convex-shaped structure.

---

### **What do networks do for companies? Testing the effects of networks on corporate performance**

*Marko Pahor* (University of Ljubljana, Slovenia)

Companies don't act independently from one another, but instead they form ties with other companies. The ties aren't formed randomly - they are carefully chosen in order to maximize company's social capital. This social capital as a productive means has the possibility of increasing company's performance.

Social capital is produced in networks in two ways - as network closure and as network possibilities. Although the first one is usually associated with dense networks, which would normally preclude the existence of brokering possibilities (or structural holes), they can coexist in a network that has the characteristics of a "small world" network. In a small world network, locally dense clusters in a generally sparse network exist.

How the social capital will be produced for a company depends on the characteristics of company's egonet, which is in turn dependent on the general strategy of the company. For a company that chooses cost-leadership strategy, a mechanistic egonet is more appropriate, and the social capital will be produced in the form of network closure. On the other hand, companies that choose differentiation as their generic strategy will find advantage in maintaining an organic network and will gain social capital from brokering possibilities.

In the paper, the theory is modelled in the framework of the joint evolution of networks and behavior. Company's properties that affect and also reveal the chosen strategy also affect the selection of ties to other companies, giving the company more network closure or more brokering possibilities. Network configuration, in turn, affects company's performance as measured by the added value of the company. The model is tested on the data for Slovenian joint-stock companies in the period 1998-2002.

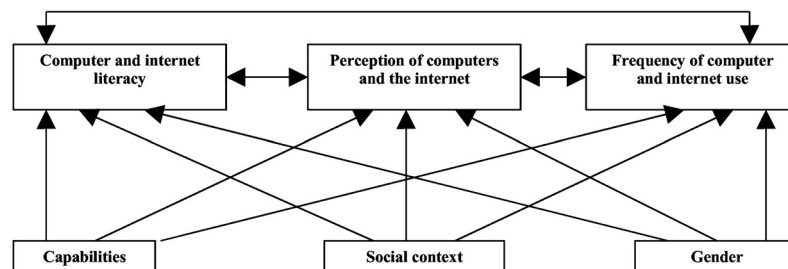
---

## The expansion of computers and the internet among Slovenian primary-school leavers

*Eva Podovšovnik* (University of Primorska, Slovenia)

In this paper I will be mainly interested in the process of the expansion of computers and the internet among Slovenian youngsters, especially the primary-school leavers. I will present the results of the survey I have conducted in the year 2003 among Slovenian 14-16 year old students.

The focus of the paper will be on identifying the most important factors that affect the expansion of computers and technology. Almost all the theories that study the expansion of the new technologies in society agree that the process follows the s-curve distribution over time. Different social theories focus on different factors that affect the shape of this process. For example, economists emphasise the role of possession of technologies, sociologists focus on the role of sociodemographic factors, while social psychologists notice the importance of cognitive processes. The research is based on the findings of the social cognitive career theory, which was developed by Lent, Hackett and Brown. On the basis of the social cognitive career theory, I formulated the following research model:



On the basis of the research model, I developed the hypothesis that the expansion of computers and the internet is a multidimensional concept formed by perception, literacy and frequency of use of computers and the internet. Each of the three variables is affected by the other two variables, individual's capabilities, gender and social context (especially family and school).

While presenting the results of the survey, I will first present the operationalisation of the dependent variables: the perception (cognition) of the computers and the internet in relation to future career, the computer and internet literacy and the frequency of use of computers and the internet. The independent variables in the research model are the capabilities, gender, father's and mother's education,

home and school computer equipment, parent's and teacher's computer use and the location (home or school) of computer and internet learning.

I have tested my hypotheses with regression analysis. There were three dependent variables, so three regression models were fitted. Within each model, the impact of each of the variables that shape the process of expansion of computers and the internet was tested.

The findings of the research show that the three dependent variables are correlated. The independent variable that affects all of them is the location of computer and internet learning.

#### References:

1. Audretsch, David B.; Bozeman, Barry; Combs, Kathryn L.; Feldman, Maryann; Link, Albert N.; Siegel, Donald S.; Stephan, Paula; Tasse, Gregory; Wessner, Charles (2002): The Economics of Science and Technology. *Journal of Technology Transfer*, 27, p. 155-203.
2. Bandura, Albert (1986): *Social foundation of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
3. Bandura, Albert (2001): *Social cognitive theory: An agentic perspective*. *Annual Review of Psychology*, 52, p. 1-26.
4. Compeau, Deborah; Higgins, Christopher A.; Huff, Sid (1999): *Social cognitive theory and individual reactions to computing technology: A longitudinal study*. *MIS Quarterly*, 23, 2, p. 145-158.
5. DiMaggio, Paul E.; Hargittai, W. Russel; Neuman, J. P.; Robinson (2001): *Social Implications of Internet*. *Annual review of Sociology*, 27, p. 307-336.
6. Freeman, Christopher; Soete, Luc (1997): *The Economics of Industrial Innovation*. Third Edition. Pinter, London and Washington.
7. Harrison, Allison W.; Rainer, R. Kelly Jr.; Hochwarter, Wayne A.; Thompson, Kenneth R. (1997): *Testing the self-efficacy-performance linkage of social-cognitive theory*. *The Journal of Social Psychology*, 137, 1, p. 79-87.
8. Lee, Alice Y.L. (1999): *Infomedia Literacy*. *Information, Communication Society*, 2, 2, p. 134-155.
9. Lent, Robert W.; Brown, Steven D.; Hackett, Gail (1994): *Toward a unifying social cognitive theory of career choice and academic interest, choice, and performance*. *Journal of Vocational Behaviour*, 45, p. 79-121.
10. Lent, Robert W.; Brown, Steven D. (1996): *Social cognitive approach to career development: An overview*. *The Career Development Quarterly*, 44, 4, 310-321.

11. Lent, Robert W.; Hackett, Gail; Brown, Steven D. (1999): A social cognitive view of school-to-work transition. *The Career Development Quarterly*, 47, 4, p. 297-311.
12. Mukoyama, Toshihiko (2003): Rosenberg's Learning by Using and Technology Diffusion.  
[http://Artsandscience.concordia.ca/faculty/mukoyama/pdf\\_word\\_excel/research/learningbyusing.pdf](http://Artsandscience.concordia.ca/faculty/mukoyama/pdf_word_excel/research/learningbyusing.pdf)  
(downloaded: 16. 01. 2004).
13. Salomon, Gavriel (1990): Cognitive Effects with and of Computer Technology. *Communication Research*, 17, 1, p. 26-44.
14. Smith, Sheila, M. (2002a): Using the social cognitive model to explain vocational interest in information technology. *Information Technology, Learning, and Performance Journal*, 20, 1, p. 1-9.
15. Smith, Sheila, M. (2002b): The role of social cognitive career theory in information technology based academic performance. *Information Technology, Learning, and Performance Journal*, 20, 2, p. 1.

---

### **Data mining in large datasets**

*Brian Ripley* (University of Oxford, UK)

Data mining used to mean over-analysing a data set, but that is now called "data dredging", and data mining is a newly popular term for finding structure in large-scale databases.

The talk will give an introduction to the area and illustrate it by two case studies from Magnetic Resonance Imaging of human brains, one on classifying tissue and one on determining where in the brain the subject is thinking.

---

### **Translation of measurement instruments and their reliability: an example of Job-related Affective Well-Being Scale**

*Nino Rode* (University of Ljubljana, Slovenia)

The translation of measurement instruments, like all cross-cultural research, presents many problems for researchers. Because of cultural and linguistic differences the questions or items in the translated instruments can have quite different meaning,

thus threatening the validity and reliability of measurement. These problems are (or should be) addressed by the translation/back-translation procedure.

To illustrate the problems, the case of Job-related Affective Well-Being Scale is presented, which was translated from English to Slovene and applied directly in research without back-translation or other form of testing the translation.

Cronbach's alpha is used to compare reliability of the results obtained with original (English) version reported by authors of the scale and with the translated (Slovene) version. Some items are singled out as problematic by item diagnostics. Especially the item "excited" translated as "vznemirjen" is discussed, since it is possible that back-translation procedure could fail to detect the problem. At the end, some tentative solutions are suggested.

---

### **Evaluation of hygienic environmental indices in village schools of Mazandaran state in Iran**

*Bizhan Shabankhani* (Faculty of Health, Iran)

The effect of schools on education and human development is crucial and schools are the center of concern in many societies. In this descriptive study, 102 village schools of Mazandaran state were sampled by an environmental hygienist during 2 months. Data collection was done by means of questionnaire, measurement and observation.

Sixty-three percent of the schools were primary schools, 30% were of intermediate level and 7% were high schools. One third of the schools were girl schools, 30% were boy schools and 37% were mixed-sex schools with split shifts for boys and girls. Average area of school was 3650m<sup>2</sup>, whereby 18% of the schools fell below the standard area level per student. In only 15% of the schools, drinking water and toilets were separate. In only 42% of the classrooms, light radiation angle was appropriate. Average number of classrooms per school was 6, with an average area of 27.8m<sup>2</sup> and an average of 18 students per class, which results in 1.5m<sup>2</sup> of space per student. Average distance between school and road was 3265m, whereby 48% of the schools were below the standard distance. Variances in all the observed data were large.

---

**Empirical estimation of outliers for different generating mechanisms**

*K. Shangodoyin and R. Arnab* (University of KwaZulu-Natal, South Africa)

The analytical construction of outlier estimates for different generating mechanisms (GMs) has shown that the magnitude of outlier depends on the respective weight of their generating mechanism. In this paper, we have generated time occurrence of outliers by simulation and compared precision and variation of their derived estimates for different GMs of order 1. We have empirically established that for both innovational (IO) and additive (AO) outlier models, if the first observation is a significant outlier, its estimated magnitude ( $\widehat{D}_T$ ) is the same as the simulated value at  $t = T - 1$ . For innovational model, the moving average (MA) generating mechanisms gave a smaller variation and lower precision compared with autoregressive moving average (ARMA) and autoregressive (AR) generating mechanisms. But the AR and the ARMA generating mechanism gave a high precision and low variation for AO model. A comparison between AO and IO models for the three simulated series indicated that the former gave a high precision for all the GMs.

---

**Different statistical measures provide different perspectives on digital divide**

*Pavle Sicerl* (University of Ljubljana, Slovenia)

A brief explanation of time distance methodology as a new view of time series data is provided. Existing static measures are left unchanged, complemented by proximity in time. The novel statistical measure S-distance measures the distance (proximity) in time between the points in time when the two series compared reach a specified level of the indicator  $X$ . It is a generic concept like static difference or growth rate.

In the empirical part its application to the gap between North America and Europe in Internet users per capita will serve as a vivid example of how different statistical measures lead to different conclusions, even about the direction of change in digital divide. A further application provides a time distance analysis of the personal computers per capita indicator for 27 countries over the period 1990-2001. Time distance analysis also brings new insights to survey results. One of such examples will be the analysis of digital divide for selected disadvantaged categories for the EU 15 and selected countries, derived from projects of the 5<sup>th</sup> Framework Programme.

---

---

**Estimating the number of foreign-body injuries in childhood with the scale-up method**

*Silvia Snidero* (University of Torino, Italy)

*Bruno Morra* (Hospital "San Giovanni Battista", Torino, Italy)

*Roberto Corradetti* (University of Torino, Italy)

*Dario Gregori* (University of Torino, Italy)

The foreign-body injury in the upper aerodigestive tract is a rare but not negligible event. This study is aimed at estimating the size of the population of people who had a foreign-body injury, using the scale-up method. This is a novel approach to estimate the size of hidden or hard-to-count subpopulations.

This estimator is based on the concept of social networks. Respondents are interviewed about the number of people known in several subpopulations (of known size) and a subpopulation  $E$  (the size of which is to be estimated). Assuming that the proportion of subjects belonging to  $E$  over the number  $c$  of people in the social network of a person is the same that in the overall population, we get the scale-up estimate of the size of the target subpopulation  $E$ .

All the otorhinolaryngologists of the Piemonte region in Italy were interviewed about the number of people they know in several known subpopulations, and the number of people they remember were hospitalized in their hospital for choking injuries in the years 1999-2001 (the target subpopulation).

This estimate was then compared with the hospital records of the Piemonte region about all injuries with ICD9 codes 931 to 934 which occurred in children aged 0-14 in the years 1999-2001.

---

**Alternative approaches to least squares factor analysis**

*Gregor Sočan* (University of Ljubljana, Slovenia)

The least squares method (LSM) is probably the most frequently used estimation method for the factor analysis model, the most popular implementations being known as Minres, Principal Axis (or Iterated Principal Factors) and Unweighted Least Squares. Although LSM lacks the elegant asymptotic properties of the Maximum Likelihood Estimation (MLE), recent empirical results suggest that it is empirically superior to MLE. I shall discuss two recent applications of the least squares principle. The first one, Minimum Rank Factor Analysis (MRFA), is based on minimising some function (for instance, sum or sum of squares) of the smallest  $p-r$  eigenvalues of the reduced correlation matrix (where  $p$  is the number of variables and  $r$  is the number of extracted factors). An especially attractive dis-

tinctive feature of this approach is that its solutions do not imply factors with negative variances. The second recent application, Direct Factor Analysis, analyses the data matrix rather than the corresponding correlation matrix. Consequently, this method provides a unique solution for factor scores and thus circumvents the notorious factor indeterminacy problem. In my talk I shall compare the rationales of the three approaches as well as present some empirical results indicating their potential practical usefulness.

---

### **Goodness of fit in regression models for relative survival**

*Janez Stare* (University of Ljubljana, Slovenia)

*Robin Henderson* (University of Lancaster, England)

*Maja Pohar* (University of Ljubljana, Slovenia)

Additive and multiplicative regression models represent the two basic ways of modelling relative survival. The first are often preferred in practice, where experience shows that the basic assumption of the additivity of hazards is likely to be met with cancer registry data. Additive models are by definition not suited for data containing subsets of subjects living longer than the general population. Multiplicative models have no such limitations.

The multiplicative model can be seen as a Cox model with time dependent coefficients and consequently the many existing methods of checking the goodness of fit can be used. In this presentation we therefore focus on checking the goodness of fit for additive models. Two of the methods discussed are based on residuals and are therefore independent of the fitting procedure and furthermore follow the same logic as the analog residual based methods for the multiplicative model. They provide a new graphical method for checking goodness of fit of the additive models and a goodness of fit test based on the Brownian bridge.

---

### **Asymptotic unbiased distribution function estimators on the basis of interval-censored data**

*Mikhail S. Tikhov and Irina S. Efimenko* (Nizhny Novgorod State University, Russia)

Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed (iid) as  $X$  random variables (rv) with unknown distribution function  $F(x)$ ;  $U_1, U_2, \dots, U_n$  are iid rv, independent of  $\{X_i, 1 \leq i \leq n\}$ , with unknown distribution function  $G(x)$ . We observe a sequence of identically distributed pairs  $\mathcal{U}^{(n)} = \{(U_i, W_i), i =$

$1, 2, \dots, n\}$ , where  $W_i = I(U_i < X_i)$  is the indicator of event  $\{U_i < X_i\}$ ,  $(U_i, X_i)$ , have joint distribution density  $g(u, x) = g(u)f(x) > 0$ . The problem of estimating the distribution function  $F(x)$  and the quantile  $x_\lambda$  of the order  $0 < \lambda < 1$  of distribution  $F(x)$  on sample  $\mathcal{U}^{(n)}$  is considered. The asymptotic behavior of the constructed estimators is established.

In works of S.V. Krishtopenko and M.S. Tikhov (1997), and M.S. Tikhov (1999), the methods for estimation of effective doses on the data in the form of binary responses were suggested and analyzed, and the algorithms implementing the considered ways of estimation were developed. The effect of a substance on an organism was estimated, whereby the mathematical model consists of the following. The random dose  $U$  of the substance is entered into an organism. Let  $X$  be the lower bound with which the effect begins: if  $U \leq X$ , the effect is absent; otherwise (i.e. when  $U > X$ ) the effect is present. In other words, consider a binary random variable (rv)  $W$  which is an indicator of event  $(U > X)$ :  $W = I(U > X)$ . As the outcome of the experiment we have a sample of pairs of values  $\mathcal{U}^{(n)} = \{(U_i, W_i), i = 1, 2, \dots, n\}$ . The considered characteristic is regression of  $W$  on  $U$ , i.e. conditional expectation  $\mathbf{E}(W|U = x)$  of rv  $W$  with the fixed value  $U = x$ . If  $U$  and  $X$  are independent, then  $\mathbf{E}(W|U = x) = F(x)$ , where  $F(x)$  is the cumulative distribution function (cdf) of rv  $X$ .

The following class of statistics will be considered (the estimators of Nadaraya-Watson):

$$\hat{F}_n(x) = \frac{S_{2n}(x)}{S_{1n}(x)},$$

where

$$S_{1n}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \quad S_{2n}(x) = \frac{1}{nh} \sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right),$$

Statistics  $\hat{F}_n(x)$  is used as the estimators of the function  $F(x)$  on the sample  $\mathcal{U}^{(n)}$ . Here  $K(x)$  is the kernel,  $h = cn^{-1/5}$  being the width of the window, and  $c$  is some constant.

We assume that the smoothing kernel  $K(\cdot)$  satisfies the following:

**K1** The kernel  $K(\cdot)$  is a bounded symmetric probability density function.

**K2** The kernel  $K(\cdot)$  vanishes outside the interval  $[-1, 1]$  and, for any  $0 < \varepsilon < 1$ ,

$$\inf\{K(u) : u \in [-1 + \varepsilon, 1 - \varepsilon]\} > 0.$$

**K3** The  $L_2$ -norm  $\nu^2 = \|K\|_2^2 = \int K^2(u) du$  and second moment  $\sigma_K^2 = \int u^2 K(u) du$  are finite.

Under the usual regularity conditions, kernel distribution function estimators from iid samples of distributions with twice differentiable densities and distribution

functions satisfy the central limit theorem

$$n^{2/5}(\hat{F}_n(x) - F(x)) \xrightarrow[n \rightarrow \infty]{d} N(a(x), \sigma^2(x))$$

with asymptotic bias and variance

$$a(x) = \frac{f'(x)g(x) + 2g'(x)f(x)}{g(x)} \neq 0, \quad \sigma^2(x) = \frac{F(x)(1 - F(x))}{g(x)},$$

respectively (see, for example, M.S. Tikhov, 2004).

This paper considers simple kernel based distribution function estimators that are both rate-optimal and have zero asymptotic bias. First, a kernel distribution function estimator with zero asymptotic bias is presented. However, the asymptotic variance of this estimator is larger (by some constant) than the asymptotic variance of the usual kernel distribution function estimator. Second, a two-step distribution function estimator is seen to be asymptotically normal with mean zero bias and the same variance as the usual kernel distribution function estimator.

The proposed estimator is obtained in two steps. First, compute pilot kernel estimators

$$\tilde{g}(x) = \frac{1}{n} \sum_{j=1}^n K_{h_0}(x - U_j), \quad \tilde{m}(x) = \frac{1}{n} \sum_{j=1}^n W_j K_{h_0}(x - U_j),$$

where  $K_h(u) = (1/h)K(u/h)$ ,  $m(x) = F(x)g(x)$ . Second, estimate the ratios  $\alpha(x) = g(x)/\tilde{g}(x)$  and  $\beta(x) = m(x)\tilde{m}(x)$  by

$$\hat{\alpha}(x) = \frac{1}{n} \sum_{j=1}^n \frac{K_{h_1}(x - U_j)}{\tilde{g}(U_j)}, \quad \hat{\beta}(x) = \frac{1}{n} \sum_{j=1}^n \frac{W_j K_{h_1}(x - U_j)}{\tilde{g}(U_j)}$$

Multiplying the pilot estimators by  $\hat{\alpha}(x)$ ,  $\hat{\beta}(x)$ , we obtain

$$\hat{g}(x) = \hat{\alpha}(x)\tilde{g}(x) = \frac{1}{n} \sum_{j=1}^n K_{h_1}(x - U_j) \frac{\tilde{g}(x)}{\tilde{g}(U_j)},$$

$$\hat{m}(x) = \hat{\beta}(x)\tilde{m}(x) = \frac{1}{n} \sum_{j=1}^n W_j K_{h_1}(x - U_j) \frac{\tilde{m}(x)}{\tilde{m}(U_j)}.$$

Let's consider the statistics

$$V_{1n}(x) = \frac{1}{n} \sum_{j=1}^n K_{h_1}(x - U_j) \cdot H_j(x), \quad \text{where} \quad H_j(x) = \frac{\sum_{i=1}^n K_{h_1}(x - U_j)}{\sum_{i=1}^n K_{h_1}(U_i - U_j)},$$

$$V_{2n}(x) = \frac{1}{n} \sum_{j=1}^n W_j K_{h_1}(x - U_j) \cdot M_j(x), \quad \text{where} \quad M_j(x) = \frac{\sum_{i=1}^n W_i K_{h_1}(x - U_j)}{\sum_{i=1}^n W_i K_{h_1}(U_i - U_j)},$$

$$\hat{F}(x) = V_{2n}(x)/V_{1n}(x).$$

**Theorem 1** Assume that kernel  $K(\cdot)$  satisfies conditions **(K1)**-**(K3)**. Denote by  $\bar{g}(x) = \mathbf{E}(\tilde{g}(x))$ ,  $\bar{m}(x) = \mathbf{E}(\tilde{m}(x))$  the expected value of the pilot estimators. If the bandwidths  $h_0, h_1$  satisfy

$$h_\ell \rightarrow 0, \quad nh_\ell / \log n \rightarrow \infty \quad \text{for} \quad \ell = 0, 1,$$

then

$$V_{1n}(x) = \frac{1}{n} \sum_{j=1}^n K_{h_1}(x - Y_j) \frac{\bar{g}(x)}{\bar{g}(Y_j)} + O_p \left( \sqrt{\frac{\ln n}{nh_0}} \right),$$

$$V_{2n}(x) = \frac{1}{n} \sum_{j=1}^n W_j K_{h_1}(x - Y_j) \frac{\bar{m}(x)}{\bar{m}(Y_j)} + O_p \left( \sqrt{\frac{\ln n}{nh_0}} \right).$$

**Theorem 2.** Assume that the functions  $g(x) > 0$  and  $f(x)$  are continuous and limited, there exists  $L_0 > 0, M_1 > 0, M_2 > 0$ , so that

$$|g(u_2) - g(u_1)| \leq L_0 |u_2 - u_1|, \quad g(u) \leq M_1, \quad f(x) \leq M_2.$$

If the ratio  $r(x) = g(x)/\bar{g}(x)$  is twice continuously differentiable, then

$$\hat{F}_n(x) = \frac{V_{1n}(x)}{V_{2n}(x)} \xrightarrow[n \rightarrow \infty]{p} F(x).$$

**Theorem 3.** Assume the smoothing kernel  $K(\cdot)$  satisfies conditions **(K1)**-**(K3)**. Let  $h_1 = c \cdot n^{-1/5}$  and  $h_0 = cn^{-\alpha}$  for  $0 < \alpha < 1/5$ . Let's assume the ratio  $r(x) = g(x)/\bar{g}(x)$  is twice continuously differentiable and there exists  $L_3$  such that  $|r'''(x)| \leq L_3$ . Then

$$n^{2/5}(V_{1n}(x) - g(x)) \xrightarrow[n \rightarrow \infty]{d} N(0, \nu^2 g(x)c), \quad n^{2/5}(V_{2n}(x) - F(x)g(x)) \xrightarrow[n \rightarrow \infty]{d} N(0, \nu^2 F(x)g(x)c),$$

$$\frac{n^{2/5}(\hat{F}_n(x) - F(x))}{b(x)} \xrightarrow[n \rightarrow \infty]{d} N(0, 1).$$

## References:

1. Krishtopenko, S.V. and Tikhov, M.S. (1997): Toxicometry of effective dozes, NNSU, Nizhny Novgorod – 156 p.

2. Tikhov, M.S. (1999): Linear functions of induced order statistics and non-parametric estimation of distributions in dose-effect dependence, *Surveys in Appl. and Industr. Math.*, 6, no 1, 244 – 245.
3. Tikhov, M.S. (2004): Statistical Estimation on the Basis of Interval-Censored Data. *J. Math. Sciences* , 119, no 3, 321 – 335.

---

### **Absolute, relative and time distance: Contradicting or complementary measures?**

*Vasja Vehovar, Pavle Sicherl, Andraž Petrovčič, Vesna Dolničar* (University of Ljubljana, Slovenia)

The comparison of socio-demographic variables ( $y_1, y_2$ ) in time (e.g. internet penetration in two comparable countries in two consecutive years) can be expressed in absolute ( $y_1 - y_2$ ), relative ( $y_2/y_1$ ), but also as time distance (or S-distance, after its founder prof. Sicherl). These three measures can be in tune, but they can also be contradicting, i.e. one measure can show an increasing difference, while the other two can show a decrease. The paper presents all 27 possible combinations of increasing, decreasing and stagnating measures of the relations of two phenomena, observed in two time points. It is shown that, in principle, all 27 combinations are possible in reality. The examples of most frequent combinations are presented and discussed. The general guidelines are outlined for the interpretation of these cases. The special case of linear trend function is also discussed, where the number of potential combinations is more restricted.

---

### **Table-based visualisation of categorical data using spreadsheets**

*Gaj Vidmar* (University of Ljubljana, Slovenia)

Various statistical visualisations of categorical data can be implemented with tables, and the ideal software tool for that purpose are spreadsheets. The paper presents a novel type of chart and some of the well-known ones, all of which are implemented in Microsoft<sup>®</sup> Excel by using only data rearrangement, cell formatting, basic formulae, sorting and a minimum of simple VBA programming.

The introduced chart is suitable for presentation of hundreds of cases and dozens of categorical variables. It has been developed for depicting accuracy of automated assignment of MeSH<sup>®</sup> thesaurus descriptor headings (15 binary attributes) to abstracts (308 cases). First, within each case, the classification probabilities

(obtained with k-nearest neighbour or some other algorithm) for the attributes are ranked. Then, an actually present/absent attribute (assigned/not assigned to a given abstract by a human information specialist in the example) is depicted with a blank/black cell; in the neighbouring cell, high support rank is depicted with a dark green cell and low support with a dark red cell for an actually present attribute, and vice versa (using the same 9-colour scale) for an actually absent attribute. Thus, each case is depicted by one row of a table and each attribute is depicted with two adjacent columns. Next, a classification accuracy measure is calculated for each case (Jaccard-type coefficient based on ranks), and cases in the plot are sorted from better accuracy to worse, thus producing increasing share of red from top to bottom of the chart. Finally, an equivalent measure is calculated for each attribute (simply by summing discrepancies for a column rather than for a row), and the attributes are sorted on its basis from left (more accurately predicted attributes - prevailingly green) to right (less accurately predicted attributes - prevailingly red). Two other examples of the same chart are presented: one from clinical biostatistics (data on morbidity after conization for 800 women) and one from bioinformatics (genetic data for 600 infertile men), whereby the attributes were predicted with logistic regression. Numerous applications of the same charting procedure can easily be envisioned, with binary or polytomous attributes and various classification methods.

In addition, it is demonstrated how sieve diagrams, association plots and observer agreement (Bangdiwala) charts can be constructed with spreadsheets simply by means of cell formatting. Such procedures could be automated with macros, while various excellent implementations of related charts (mosaic plots, biplots for correspondence analysis, ternary diagrams etc.) already exist in publicly available add-ins and workbooks. Furthermore, many built-in chart types can be efficiently adapted for visualisation of categorical data (e.g., bubble plots for logistic regression diagnostics and stacked bar-charts for hanging rootograms), and vast expert knowledge in the field is publicly accessible (c.f. websites by J.Peltier and F.Cinquegrani). Hence, implementation of all the visualisations discussed in the seminal book *Graphical Methods for Categorical Data* (M.Friendly, 2001) in Microsoft<sup>®</sup> Excel is a worthy goal for statistical practice and a valuable potential project for statistics educators.

---

**Intra-assessor inconsistency in setting performance standards**

*Hans J. Vos* (University of Twente, The Netherlands)

A common task for assessors working in the field of HRD is setting performance standards on assessments composed of complex, multiply scored performance-based exercises. Intra-assessor inconsistency arises when assessors specify performance standards which are incompatible with each other and, consequently, imply different standards. The purpose of this paper is to propose a method for analyzing intra-assessor inconsistency by comparing assessor's subjective performance standards using the extended Angoff method (Hambleton, 1995) with those obtained under an IRT model for polytomously ordered items. More specifically, the well-known two-parameter generalized partial credit model (GPCM) for rating scales (Muraki, 1992) will be used as a polytomous IRT model. The proposed method will be demonstrated with an empirical study in which rating scores are obtained by providing assessors with hypothetical profiles on several dimensions of the complex tasks to be performed by employees in an Assessment Center. The technique of systematically generating profile scores is borrowed from Judgmental Policy Capturing (JPC), a technique well-known in the field of personnel selection in industrial psychology and elsewhere. Fitting the profile scores to an underlying multiple regression model, it will also be investigated to what extent the assessors are capable to assign relative weights to the several dimensions of the tasks as part of the extended Angoff method for subjective performance standard setting.

**References:**

1. Hambleton, R.K. (1995): Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.
  2. Muraki, E. (1992): A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
-

**Less parametric methods in applied statistics**

*K. Laurence Weldon* (Simon Fraser University, Canada)

Despite forty years of revolution in the tools available for statistical analysis, the current academic tradition in statistics is remarkably similar to the pre-computer tradition. This tradition is rooted in parametric modeling, least squares, and linear models. The paper argues for a shift in emphasis away from parametric modeling and data reduction to graphical display, from omnibus optimal techniques to those that are more context-specific, and from goals of objectivity to goals of revelation. It is suggested that certain modern topics should be included in a basic statistical education, and that more academic work is needed to achieve this goal.

---

**Logistic regression analysis of MRI data from breast cancer patients**

*Masoud Yarmohammadi* (Payame Noor University, Iran)

*Parviz Abdolmaleki* (Tarbiat Modarres University, Iran)

Logistic regression analysis was used to differentiate malignant disease from benign in a group of patients with proved breast lesions on the basis of morphological data extracted from MR imaging. Our database included 161 patients' records consisting of six qualitative variables. The database was randomly divided into training and validation sample including 110 and 51 records, respectively. The training sample was used to construct the logistic regression model as a classifier, while the validation sample was used to validate the model's performance. Finally, sensitivity, specificity, accuracy and receiver operating characteristic curve (ROC) for this method as well as that of the radiologist were compared. Our results show that the logistic regression model is able to classify correctly 42 out of 51 cases in the validation sample. Comparing the output of this method with that of the radiologist reveals reasonable diagnostic accuracy 82%, remarkable specificity (90%) and relatively high sensitivity (80%).

---

## Islands

*Matjaž Zaveršnik and Vladimir Batagelj* (University of Ljubljana, Slovenia)

In large network analysis we are often interested in important parts of a given network. There are several ways how to determine them. Our algorithm is based on the importance weights of vertices.

Let  $t$  be any real number. If we delete all vertices (and corresponding edges) with the importance weights less than  $t$ , we get subnetwork called vertex-cut at level  $t$ . The number and sizes of its components depend on  $t$ . Often we consider only components of size at least  $k$  and not exceeding  $K$ . The components of size smaller than  $k$  are discarded as noninteresting while the components of size larger than  $K$  are cut again at some higher level.

Vertex-island is connected subnetwork which vertices have greater importance weights than the vertices in its neighborhood. It is easy to see that the components of vertex-cuts are all vertex-islands. We developed an algorithm that identifies all maximal vertex-islands of sizes in the interval  $k \dots K$  in a given network. Each island is identified with its port - its lowest vertex. The main problem are the vertices at the same level - flat regions.

For networks with weighted edges we can similarly define edge-islands. The edge-islands algorithm is based on edge-cuts.

Both algorithms and some applications of islands in analysis of large networks will be presented.

---

## Stability of measures of centrality and prominence: a meta-analysis on fixed choice and free choice data

*Barbara Zemljič* (Radiotelevision Slovenia, Slovenia)

*Valentina Hlebec* (University of Ljubljana, Slovenia)

An important aspect of network structure is network centrality and prominence, which seeks to quantify theoretical ideas about an actor's importance within the network on the individual level by summarizing the structure of relations among actors. A number of measures of centrality and prominence have been devised to capture importance, addressing different aspects of the intuitive notion. This empirical study evaluates the stability of measures of centrality and prominence of social networks among high school students. The authors present and discuss results from two studies carried out in January 1998 and November 2001 among high school students in Slovenia. Four types of social support - instrumental, informational, social companionship and emotional support - were measured

with four measurement scales (binary, categorical and line production in the first study, and categorical and 11-point scale in the second study). Two methods - recognition and free recall - were used for collecting social network data. In the second study, the number of choices was limited to between three and five and was assessed against the free choice method. Measurement of each social network was repeated two or three times. The stability of fourteen measures of centrality and prominence was estimated by the Pearson correlation coefficient. Apart from previous findings with regard to global and local measures, in- and out-measures, time between repetitions and social support, the reduction in the number of choices and reversed questions decreased the stability of measures of centrality and prominence.

---

### **Comparison of analytic models for the costs of postinfarct patients**

*Giulia Zigon* (University of Firenze, Italy)

*Dario Gregori* (University of Torino, Italy)

#### **Introduction and aim**

Studies of the costs of myocardial infarction (MI) and the factors affecting such costs are becoming more and more important for clinicians and policy-makers aiming specifically at the most cost-effective strategy for treatment of postinfarct patient.

Different stratification modalities have been proposed (simple clinical data obtained during the acute phase, the most commonly used exercise testing and, more recently, coronary angiography and stress echocardiography), but it is still unclear what is the better choice between invasive and non-invasive strategy in terms of cost-efficacy. Furthermore, the analysis of medical costs presents several difficulties from the statistical point of view.

The data referring to medical costs are characterized by an asymmetric distribution (because of a minority with high medical costs compared to the rest of the population) and the presence of dependent censoring (because of correlation between cost at censoring and cost-to-event) due to the patient deaths during follow-up.

According to the data characteristics and particularly to the presence of censoring, several works [1] have proposed to use survival models like the Weibull model and the Cox regression model, because these models are based on few and/or more realistic assumptions concerning the distribution of the cost variable. Nevertheless, accrual of costs at different rates leads to dependent (or informative) censoring within subgroups defined by covariate levels, and the proportional hazards (PH) assumption of these models is not in general satisfied [2].

In the light of these considerations, the purpose of this study is a comparison

of analytic models for estimating the effect of clinical factors and management strategies on the costs of postinfarct patients. The innovative application of the Aalen additive regression model [3, 4] to medical costs is stressed.

### Methods

A follow up of 1 year for medical costs was carried out in 10 general hospitals, eight in Italy and two in Turkey. Patients were uncomplicated MIs admitted within 24h of the onset of symptoms; four-hundred eighty-seven patients were enrolled and randomly assigned to three different strategies: 1) (132 patients) Early use of pharmacological stress echo cardiography and immediate subsequent discharge on day 3-5; 2) (130 patients) Maximal symptom limited exercise testing under therapy, discharge on day 7-9; 3) (22 patients) Clinical evaluation and hospital discharge on day 7-9. Cost of hospitalization was estimated referring to mean reimbursement for the diagnosis-related groups (DRG).

The clinical variables considered were age, gender, previous MI, diabetes, ejection fraction (EF), MI antero/lateral, and strategy type.

Five different models were utilized: ordinary least square (OLS) linear regression, binary logistic regression (with median and third quartile as cut-point), Cox PH model, parametric survival model assuming the Weibull distribution, and the Aalen additive regression model.

### Results and Discussion

The considered covariates are not significant except for the MI location (antero/lateral). There is agreement between all the models regarding this variable provided that the third quartile is considered as cut-point for the logistic model. The cost data appear to be approximated well by a Weibull distribution (estimated scale parameter 0.88), but the key assumption of proportional hazards of the Cox and the Weibull model is not met (global chi-squared=22.88,  $p = 0.001$ ), particularly for age and strategy.

To compare the quantitative cost predictions of the models, we computed the predicted costs relative to the mean and median (Table 1). Logistic regression predicts well the proportion of costs greater than 12319 Euro (third quartile) ( $p = 0.26$ ) and 4845 Euro (median) ( $p = 0.51$ ). The OLS linear regression model predicts the mean cost well enough, but it overestimates the median, and the same goes for the Weibull model. The Cox model and the Aalen model perform well, the later particularly regarding the median value.

Table 1: Mean and median of the cost values predicted by the models (in Euro).

	Obs. data	OLS m.	Weibull m.	Cox m.	Aalen m.
Mean	9162.1	9352.7	9938.5	9378	9281
Median	4845	9447.4	9822.7	4967	4906

**References:**

1. A.R. Dudley, F.E. Harrell, R.L. Smith, D.B. Mark, R.M. Califf, D.B. Pryor., D. Glower, J. Lipscomb, and M. Hlatky (1993): Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology*, 46(3):261 - 271.
  2. R.D. Etzioni, E.J. Feuer, S. Sullivan, D. Lin, C. Hu, and S. Ramsey (1999): On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics*, 18:365 - 380.
  3. O.O. Aalen (1989): A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907- 925.
  4. O.O. Aalen (1993): Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine*, 12:1569 - 1588.
-

## Workshops

---

### Workshop 1: **Statistical Graphics for Exploring Data, Presenting Information, and Understanding Statistical Models**

*Frank Harrell* (Vanderbilt University, USA)

Graphical methods are being increasingly used for exploratory data analysis. Some of the many graphical tools that are useful in this setting are scatterplot matrices, nonparametric smoothers, and tree diagrams. Statistical graphics for presenting information have been used much longer, but most of the commonly used graphics used in papers, presentations, and the popular media, such as bar charts and pie charts, are either poor or misleading in communicating information to the reader. This short course begins with a series of graphical horror stories from the scientific and lay press. Then elements of graphical perception and good graph construction, many from the writings of Bill Cleveland, are covered. Practical suggestions for choosing the best chart or graph type, making good and clear graphics, and formatting are covered. Techniques for simultaneous presentation of multiple variables are described.

Complex outcome or risk adjustment models are not easily grasped by non-statisticians. Special graphics such as effect charts and nomograms can assist physicians and other consumers of statistical analysis in understanding statistical models and in using them for obtaining predictions for individual subjects. Examples of model presentation graphics will be given.

At the close of the short course some graphical marvels from the literature (especially from Edward Tufte and Howard Wainer) are presented.

---

**Workshop 2: Analyzing data with PAJEK**

*Vladimir Batagelj and Andrej Mrvar* (University of Ljubljana, Slovenia)

We start with a discussion of connections between data analysis and network analysis - the relational data analysis.

One among the approaches is the network analysis of neighborhood networks derived from the (multivariate) data. We present several possibilities offered by Pajek for their analysis and visualization.

We also demonstrate how we can complement the analyses in Pajek with analyses using statistical programs R and SPSS.

Program Pajek is available at:

`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

---



## Notes

---















Supported by



---

REPUBLIKA SLOVENIJA  
MINISTRSTVO ZA ŠOLSTVO, ZNANOST IN ŠPORT

---

**Ministry of Education, Science and Sport**



---

STATISTIČNI URAD REPUBLIKE SLOVENIJE  
STATISTICAL OFFICE OF THE REPUBLIC OF SLOVENIA

---



**SPSS Division, Slovenia**

<http://spss.cati.si/>



**Alarix d.o.o.**

<http://www.alarix.si/>

