International Conference

APPLIED STATISTICS

2007

PROGRAM and ABSTRACTS

September 23 – 26, 2007

Ribno (Bled), Slovenia

Supported by

Slovenian Research Agency (ARRS)

Statistical Office of the Republic of Slovenia

Alarix d.o.o.

SPSS Slovenia

Result d.o.o.

Zavod za turizem Ljubljana

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

311(063)(082)

INTERNATIONAL Conference Applied Statistics (2007 ; Ribno)
Program and abstracts / International Conference Applied Statistics 2007, September 23-26, 2007, Ribno (Bled), Slovenia ; [edited by Gaj Vidmar and Janez Stare]. - Ljubljana : Statistical Society of Slovenia, 2007

ISBN 978-961-90314-8-3
1. Applied Statistics 2. Vidmar, Gaj
234880768

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, United Kingdom
Jože Rovan, Slovenia
Tamas Rudas, Hungary
Willem E. Saris, The Netherlands
Albert Satorra, Spain
Vasja Vehovar, Slovenia
Gaj Vidmar, Slovenia
Hans Waege, Belgium

Organizing Committee

Andrej Blejec (Chair)
Bogdan Grmek
Katja Rostohar
Gaj Vidmar
Irena Vipavc Brvar

Published by: Statistical Society of Slovenia
Vožarski pot 12
1000 Ljubljana
Slovenia
Edited by: Gaj Vidmar and Janez Stare
Printed by: Statistical Office of the Republic of Slovenia, Ljubljana

PROGRAM OVERVIEW

Sept. 23 (Sun)	Session	Hall 1	Code	Papers	Hall 2	Code	Papers
	Invited Lecture (A. Agresti)						
	Morning	**Theoretical Statistics I**	T1	5	**Data Collection & Sampling Techniques**	DC	5
	Aftern. 1	**Theoretical Statistics II**	T2	5	**Measurement I**	M1	4
	Aftern. 2	**Spatial Statistics & Time Series Analysis**	ST	4	**Measurement II**	M2	4

Sept. 24 (Mon)	Session	Hall 1	Code	Papers	Hall 2	Code	Papers
	Invited Lecture (I. Olkin)						
	Morning 1	**Biostatistics & Bioinformatics I**	B1	5	**Econometrics & Time Series Analysis I**	E1	4
	Morning 2	**Econometrics & Time Series Analysis II**	E2	5	**Biostatistics & Bioinformatics II**	B2	5

Sept. 25 (Tue)	Session	Hall 1	Code	Papers	Hall 2	Code	Papers
	Invited Lecture (R. Steyer)						
	Morning 1	**Social Science Methodology**	SM	5	**Design of Experiments**	DE	5
	Morning 2	**Network Analysis**	NA	5	**Statistical Applications I**	A1	5
	Aftern. 1	**Data Mining I**	D1	5	**Biostatistics & Bioinformatics III**	B3	5
	Aftern. 2	**Modelling & Simulation I**	L1	5	**Statistical Applications II**	A2	5
	Aftern. 3	**Modelling & Simulation II**	L2	5	**Data Mining II**	D2	5

Sept. 26 (Wed)	Session	Hall 1	Code	Papers	Hall2	Code	Papers
	Morning 1	**Econometrics & Time Series Analysis III**	E3	5	**Theoretical Statistics III**	T3	5
	Morning 2	**Stat. Education & Missing Data in MVA**	EM	5			
	Afternoon	*Workshop (A. Blejec)*					

SUNDAY, September 23, 2007

09.00 – 10.00 Registration

10.00 – 10.10 Opening of the Conference

10.10 – 11.00 **INVITED LECTURE** (Hall 1) *Chair: Janez Stare*

Alan Agresti, Department of Statistics, University of Florida Gainesville, USA
Small-Sample Interval Estimation for Categorical Data

11.00 – 11.20 BREAK

11.20 – 13.00 **Theoretical Statistics I** **T1** (Hall 1) *Chair: Alan Agresti*

1. **Accumulation of Vague Data: Non-Precise Counts** *Bodjanova S*
2. **On Bagging and Estimation in Multivariate Mixtures** *Pakyari R*
3. **Characteristics of a New Bivariate Distribution** *Cedilnik A, Blejec A, Košmelj K*
4. **Properties of Some Statistical Tests to Establish Non-Inferiority for Independent Proportions** *Almendra-Arao F, Sotres-Ramos D*
5. **Reporting Uncertainity by Spline Function Approximation of Log-Likelihood** *Sezer A*

11.20 – 13.00 **Data Collection & Sampling Techniques** **DS** (Hall 2) *Chair: Vasja Vehovar*

1. **Developing a Quality Assurance Model for Iranian Higher Education System: An Exploratory Design** *Mohammadzadeh S, Hedjazi Y, Bazargan A*
2. **Estimation of Non-Response Bias in Round 2 of the European Social Survey: Using Information from Reluctant Respondents** *Beullens K, Billiet J*
3. **Quality Issues in Questionnaire Design** *Petrakos G, Ieromnimon T*
4. **Population in Bosnia and Herzegovina: Survey Versus National Statistical Office Estimates** *Šabanović E*
5. **Some Ratio Estimators in Successive Sampling over Two Occasions** *Ahmed MS, Dorvlo ASS*

13.00 – 15.00 LUNCH

15.00 – 16.40 **Theoretical Statistics II** **T2** (Hall 1) *Chair: Katarina Košmelj*

1. **A Robust Estimation of Spline Function** *Al-Ghamedi AA*
2. **A New Robust Measure of Skewness** *Al-Ghamedi AA*
3. **On Reliability Equivalence Factor** *Pogány TK*
4. **Fitting Mixtures of Poisson Regression Models** *Faria S, Soromenho G*
5. **Robust Test for Testing Hypotheses on Finite Data Sets** *Ratej Pirkovič S, Valentinčič A*

15.00 – 16.00 **Measurement I** **M1** (Hall 2) *Chair: Lluís Coromina*

1. **The Power of the Non-Normality Corrected Chi-Square Statistics in Structural Equation Modeling** *Olsson UH, Foss T*
2. **Measuring Explained Variance in Multiple Correspondence Analysis of Crisp and Fuzzy Coded Data** *Asan Z, Greenacre M*
3. **The Implications of Different Factor Analysis Solutions for the Theoretical Interpretation of the Choice of Educational Path** *Pavlin S, Kogovšek T*
4. **Bootstrapping Congruence Coefficients in Principal Component Solutions** *Sočan G*

16.40 – 17.00 BREAK

17.00 – 18.20 **Spatial Statistics & Time Series Analysis** **TS** (Hall 1) *Chair: Damijana Kastelec*

1. **An Empirical Comparison of Model Selection Criteria for Parametric and Nonparametric Regression** *Şengün Ucal M*
2. **Multivariate Analysis for Space-Time Pollution Estimation** *Palma M, Maggio S*
3. **Space-Time Correlation Analysis: A Comparative Study** *De Iaco S*
4. **Spatial Data Mining and Visualization with GoogleEarth** *Jesenovec D, Lavrač N, Mramor Kosta N*

17.00 – 18.20 **Measurement II** **M2** (Hall 2) *Chair: Gregor Sočan*

1. **Heuristic Evaluation of Website Usability: An Application to the Italian Websites of Municipalities/Communes** *Oehler M, Biffignandi S*
2. **Measuring the Value of Different Categories of Knowledge within a Romanian University** *Lupşa-Tătaru DA*
3. **Measurement Characteristics of the Students' Ratings of the Teachers** *Rode N*
4. **Grades from 1 to 5 or A to E: From Theory to Practice** *Bren M, Zupanc D, Blejec A*

19.00 – RECEPTION

MONDAY, September 24, 2007

09.10 – 10.00 **INVITED LECTURE** (Hall 1) *Chair: Katarina Košmelj*

Ingram Olkin, Department of Statistics, Stanford University, USA
Meta-Analysis; Statistical Methods for Combining the Results of Independent Studies

10.00 – 10.15 BREAK

10.15 – 11.55 **Biostatistics & Bioinformatics I** **B1** (Hall 1) *Chair: Ingram Olkin*

1. **Using Profile Likelihood for Semiparametric Model Selection with Application to Proportional Hazards Mixed-effects Models** *Xu R*
2. **Random Effects Modelling of Patient Pathways** *Adeyemi S, Chaussalet TJ, Xie H, Millard PH*
3. **Testing Dose-Response with Multivariate Ordinal Data** *Solari A, Klingenberg B, Salmaso L, Pesarin F*
4. **A Measure of Prognostic Value of Survival Models** *Stare J, Pohar Perme M*
5. **Checking Hazard Regression Models Assumptions Using Pseudo-Observations** *Pohar Perme M, Andersen PK*

10.15 – 11.55 **Econometrics & Time Series Analysis I** **E1** (Hall 2) *Chair: Jože Rovan*

1. **Comparison of Smooth Transition Stochastic Volatility Models with with Markov switching Stochastic Volatility Models in a Bayesian Approach** *Amiri E*
2. **Time Series Discrimination in State Space Form** *Kalantzis T, Papanastassiou D*
3. **Modelling Moving Feasts Determined by the Islamic Calendar: Application to Macroeconomic Tunisian Time Series** *Grun-Rehomme M, Ben Rejeb A*
4. **How Useful is the LAVE Method** *Lotrič Dolinar A*

11.55 – 12.10 BREAK

12.10 – 13.50 **Econometrics & Time Series Analysis II** **E2** (Hall 1) *Chair: Aleša Lotrič Dolinar*

1. **Comparison of Forecasting Performance of Regime Switching versus Linear Models: Application to Turkish Economy** *Koç S, Özdemir Koç S*
2. **Imperfect Information and Credit Rationing in Financial Markets: Application of Long Range Dependence in Credit Series** *Özdemir Koç S, Koç S*
3. **Nonparametric Evaluation in the Statistical Problems of Finance Theory** *Vavilov SA, Ermolenko KYu*
4. **The Effect of Oil Price Volatility on the Istanbul Stock Exchange** *Demirci E, Er Ş, Ata B*
5. **Model for Evaluating Development of EU Countries** *Ratej Pirkovič S*

12.10 – 13.50 **Biostatistics and Bioinformatics II** **B2** (Hall 2) *Chair: Maja Pohar Perme*

1. **Exploration of Categorical Screening Procedure Data by Multiple Correspondence Analysis** *Rovan J, Urbančič-Rovan V, Slak M*
2. **Education and Second Birth Rates in Denmark 1981-1994** *Gerster M, Keiding N, Knudsen LB, Strandberg-Larsen K*
3. **The Maximum Term for Testing the Homogeneity of Two Multinomial Populations with a Large Number of Categories** *Valente Freitas A, Pinheiro M, Oliveira JL, Moura G, Santos M*
4. **The Impact of Preprocessing on the Differentially Expressed Gene Lists** *Rotter A, Hren M, Baebler Š, Blejec A, Gruden K*
5. **Exact Simultaneous Confidence Regions for Genetic Problems** *Biebler K-E, Jäger B, Wodny M*

14.20 – EXCURSION

TUESDAY, September 25, 2007

09.10 – 10.00 **INVITED LECTURE** (Hall 1) *Chair: Vasja Vehovar*

Rolf Steyer, Institute of Psychology, Friedrich-Schiller-University Jena, Germany
Analysis of Causal Effects in Between-Group and Within-Group Comparisons

10.00 – 10.15 BREAK

10.15 – 11.55 **Social Science Methodology** **SM** (Hall 1) *Chair: Rolf Steyer*

1. **How to Objectively Rate Investment Experts in Absence of Full Disclosure? An Approach Based on a Near Perfect Discrimination Model** *Wessa P*
2. **A Longitudinal Study of Student Performance in English using Repeated Measures Analysis of Variance and Multilevel Modelling** *Camilleri L, Xuereb G*
3. **Assessing Suicidal Intent in an Epidemiological Study Using the Cumulative (Proportional) Odds Model for Ordinal Variables** *Corry C, Bunting B, McCann S*
4. **Pretesting Questionnaires with Expert (Re)appraisal: Comparison of Two Appraisal Schemes** *Hlebec V, Koren G*
5. **Methodological Issues in Analyzing Social Networks in Online Forums** *Vehovar V, Žiberna A, Jakulin A*

10.15 – 11.55 **Design of Experiments** **DE** (Hall 2) *Chair: Bronisław Ceranka*

1. **Variance Balanced Block Designs** *Ceranka B, Graczyk M*
2. **Optimum Chemical Balance Weighing Design under Certain Condition** *Ceranka B, Graczyk M*
3. **An Application of Experimental Design in Inorganic Chemistry** *Bashiri M, Ansar S*
4. **50-50 MANOVA with Rotation Testing: A Framework for Analysing Designed Experiments with Multiple Responses** *Langsrud Ø*
5. **Response Surface Analyses: An Application to the Optimization of Astaxanthin Production by Thraustochytrium CHN-1** *Espina, VD, Carmona, ML, Yamaoka Y, Naganuma T*

11.55 – 12.10 BREAK

12.10 – 13.50 **Network Analysis** **NA** (Hall 1) *Chair: Valentina Hlebec*

1. **Comparison of PhD Students' Performance with Duocentered Network Measures** *Coromina L, Capó AM, Coenders G, Ferligoj A, Matelič U*
2. **Follow-up Qualitative Study for Non-Interpretable Social Network Variables** *Capó Artigues AM, Coenders G, Coromina L*
3. **Valued Two-Mode Blockmodeling for Input-Output Analysis** *Denk M, Weber M*
4. **New Educational Plans for New Professional Profiles: The Social Network Analysis Approach** *Civardi M, Zavarrone E, Zappa P*
5. **Network Analysis of Data from Web of Science** *Batagelj V, Kejžar N*

12.10 – 13.50 **Statistical Applications I** **A1** (Hall 2) *Chair: Gaj Vidmar*

1. **Climate Reconstruction** *Schofield MR, BarkerRJ*
2. **Null Model Analyses of Presence-Absence Data in Ecology: Combining Generalized Linear Models and Monte Carlo Testing for the Detection of Non-Random Patterns** *Navarro JA, Manly BF*
3. **Statistical Analysis of Earthquake Data with Extreme Value Distributions Based on Markov Renewal Process** *Firuzan E, Uzunoğlu Koçer U*
4. **Comparison of Statistical Models Describing Power Semiconductor Cycle Life Data** *Bluder O, Köck H, Glavanovics M*
5. **Application of Mixture Modelling to Personality Disorder Criteria in a General Population Sample** *Devine S, Bunting B, McCann S*

13.55 – 15.20 BREAK

15.20 – 17.00 **Data Mining I** **D1** (Hall 1) *Chair: Nada Lavrač*

1. **Concurrent Programming for Extracting Knowledge with Uncertainty** *Brunet G*
2. **Data Mining Trauma Injury Data with Imputed Values** *Penny KI, Chesney T*
3. **Redundancy Measures for Multivariate Data** *Dhorne T*
4. **Diagnosing Equilibrium Models from Maps Constructed from Logratios of a Data Matrix** *Greenacre M*
5. **Hierarchical Clustering with Relational Constraints of Large Datasets** *Batagelj V, Mrvar A*

15.20 – 17.00 **Biostatistics & Bioinformatics III** **B3** (Hall 2) *Chair: Janez Stare*

1. **Risk Scores for Undiagnosed Diabetic Retinopathy Screening Subjects** *Hosseini SM, Maracy MR, Amini M*
2. **A Probabilistic Fertility Model for First Conception** *Farooqui MZ*
3. **Determining Hidden Markov Models Efficacy in a Gene Finding Problem** *Kazemnejad A, Hajizadeh E, Mirjafari K*
4. **Comparing Generalized Estimating Equation Model with Standard Logistic Regression Model in Determining Back Pain Associated Factors in Iran** *Mohammad K, Saiepour N*
5. **Evaluation of First and Second Order Markov Chains Sensitivity and Specificity and Their Relation with Characteristics of Virus Double-Stranded DNA Genome** *Farzami J, Hajizadeh E*

17.00 – 17.10 BREAK

17.10 – 18.50 **Modelling and Simulation I** **L1** (Hall 1) *Chair: Tibor K. Pogány*

1. **Optimal Design of Bayesian Reliability Test Plans for a Series System Based on Type-II Censoring** *Kumar M*
2. **Stochastic and Monte Carlo Simulation for the Spread of the Hepatitis B Virus** *Alahmed M, Moneim A*
3. **Approximate Maximum Likelihood Estimation for the Scaled Generalized Exponential Distribution Based on Progressive Type-II Censoring** *Asgharzadeh A*
4. **Asymptotic Behaviour of the Friedman Test Statistic** *Jäger B, Biebler K-E*
5. **Comparison of Ordinary and Penalized Power-Divergence Test Statistics for Small and Moderate Samples in Three-Way Contingency Tables via Simulation** *Alin A*

17.10 – 18.50 **Statistical Applications II** **A2** (Hall 2) *Chair: Michael Greenacre*

1. **Do Neural Networks Outperform Classical Logistic Regression and Discriminant Analysis Statistical Classifiers? A Case Study in Selection of Portuguese Air-Force Pilot Candidates** *Maroco J, Bartolo-Ribeiro R*
2. **Canonical Correlation Analysis and Kernel Smoothing and Their Application to Relations and Indices of Organizational Variables** *Bekrizadeh H, Azizi F*
3. **Statistical Analysis of Multiculturalism Research in Vojvodina** *Čobanović K, Sokolovska V, Nikolić-Đorić E*
4. **Edge Correction for Segregation Tests Based on Nearest Neighbor Contingency Tables** *Ceyhan E*
5. **Financial Accounts Visualization** *Komprej I*

18.50 – 19.00 BREAK

19.00 – 20.40 **Modelling and Simulation II** **L2** (Hall 1) *Chair: Matevž Bren*

1. **Nonparametric Inequality Measure Based on Ranks** *Pati D, Bhattacharya A, Sarkar A*
2. **Computer-related Statistical Analysis - Ideas, Otimizations, Problems, Advice** *Tébi G*
3. **A Simulation Study of a Queuing System** *Özdemir Ö*
4. **Estimation of Target Parameters for Small Domains** *Nikič B*
5. **Optimization Problem in Markov Chain Modeling** *Vehovar V, Škulj D, Perman M*

19.00 – 20.40 **Data Mining II** **D2** (Hall 2) *Chair: Thierry Dhorne*

1. **A Hybrid Approach to Data Mining Radiological Medical Records** *Claster W, Shanmuganathan S, Ghotbi N*
2. **Enhancing Performance of Credit Scoring Techniques Based on Logistic Regression and Generalized Additive Models** *Patra S, Shanker K, Kundu D*
3. **Using Behavioral Statistical Scorecards in Portfolio Management and Business Planning** *Klepac G, Kličeк B*
4. **Exploratory Analysis of the ILPNet2 Repository: Research Contents Analysis and Coautorship Network** *Lavrač N, Grčar M, Fortuna B*
5. **Application of Regression Models and Polynomial Equations to Predict Out-Crossing Rate of Maize** *Debeljak M, Ivanovska A, Kocev D, Džeroski S, Rostohar K*

WEDNESDAY, September 26, 2007

09.00 – 10.40 **Econometrics & Time Series Analysis III** **E3** (Hall 1) *Chair: Tina Kogovšek*

1. **Technical Efficiency of Philippine Rice-Producing Regions: A Stochastic Frontier Approach** *Pate NT, Tan-Cruz A*
2. **Liability Dollarization, Exchange Market Pressure and Fear of Floating: Empirical Evidence from Turkey** *Feridun M*
3. **Asymptotics for Periodically Stationary Time Series with Heavy Tails** *Rezakhah S, Ghasemi H*
4. **Nonlinear Time Series Modelling of Lahore's Precipitation** *Khan MS, Iqbal MJ*
5. **Prediction of GDP per Capita of Turkey and Balkan Countries: Comparison of ARIMA Models, Neural Networks and Support Vector Machines** *Tolun Esen S, Er Ş, Kiremitci B*

09.00 – 10.40 **Theoretical Statistics III** **T3** (Hall 2) *Chair: Larry Weldon*

1. **A New Generalized Useful Relative Information Generating Function** *Kumar P, Bhat BA*
2. **Nonparametric Estimation of the Drift Coefficient for a Stationary Process** *Arfi M*
3. **Sample Size in Mupltiple Regression: 20 + 5*k*** *Khamis H, Kepler M*
4. **Parameter Estimation of Bernoulli Mixture Distribution** *Asma S*
5. **Expectation of Maxima of Random Variables: Theory and Application** *Tokarev D, Hamza K, Klebaner FC*

10.40 – 11.00 BREAK

11.00 – 12.40 **Statistical Education & Missing Data in Multivariate Analysis** **EM**(Hall 1) *Chair: Anuška Ferligoj*

1. **Learning Attitudes, Peer Assessment, and Gender in the Context of a Social Constructionist Statistics Course** *Wessa P*
2. **Everyday Benefits of Understanding Variability** *Weldon KL*
3. **M-Shaped Distributions** *Vidmar G*
4. **The Impact of Missing Data Treatment on Results of Ward Hierarchical Clustering** *Žnidaršič A, Garvas T, Planinc S*
5. **Influence of Different Methods for Treatment of Missing Data on Results of Factor Analysis** *Ačimovič J, Kmet A, Kopač P*

12.40 – 12.50 CLOSING OF THE CONFERENCE

12.50 – 15.30 LUNCH

15.30 – 18.00 **WORKSHOP** (Hall 1)

Andrej Blejec, National Institute of Biology, Ljubljana, Slovenia
Introduction to R

INVITED LECTURES

Small-Sample Interval Estimation for Categorical Data

Alan Agresti
(Department of Statistics, University of Florida, USA)

"Exact", small-sample methods for categorical data are exact in terms of using probability distributions that do not depend on unknown parameters. However, they are conservative inferentially, having actual error probabilities for inference that are bounded above by the nominal level. We examine the conservatism for confidence intervals and survey ways of reducing it, illustrating for estimating binomial proportions and related measures such as the odds ratio. Fuzzy inference is an adaptation of randomized inference that achieves the error probability exactly. In practice, many would find this approach unsuitable. However, it motivates inferences based on the mid-P value that are less conservative than standard exact methods yet usually approximate well the desired error probabilities. We also summarize simple ways of adjusting standard large-sample confidence intervals to improve dramatically their small-sample performance.

Analysis of Causal Effects in Between-Group and Within-Group Comparisons

Rolf Steyer
(Department of Methodology and Evaluation Research, Institute of Psychology, Faculty of Social and Behavioural Sciences, Friedrich Schiller University Jena, Germany)

Typically, regression analysis for multistate models has been based on regression models for the transition intensities. These models lead to highly non linear and very complex models for the effects of covariates on state occupation probabilities. We present a technique that models the state occupation or transition probabilities in a multistate model directly. The method is based on the pseudo-values from a jackknife statistic constructed from non-parametric estimators for the probability in question. These pseudo-values are used as outcome variables in a generalized estimating equation to obtain estimates of model parameters. We examine this approach and its properties in detail for two special multistate model probabilities, the cumulative incidence function in competing risks and the current leukemia free survival used in bone marrow transplants. The latter is the probability a patient is alive and in either a first or second post transplant remission. The techniques are illustrated on a data set of leukemia patients given a marrow transplant.

Meta-Analysis; Statistical Methods for Combining the Results of Independent Studies

Ingram Olkin
(Department of Statistics, Stanford University, USA)

Meta-analysis enables researchers to synthesize the results of a number of independent studies designed to determine the effect of an experimental protocol such as an intervention, so that the combines weight of evidence can be considered and applied. Increasingly, meta-analysis is being used in the medical and health sciences to augment traditional methods of narrative research by systematically aggregating and quantifying research literature. A Google Scholar search on meta-analysis plus different fields of research uncovered 229,000 hits in medicine, 84,000 in health policy, and 27,000 in genetics. Similar results occur in the social sciences or in education.

In this talk we provide a historical perspective of meta-analysis, and discuss some of the nonparametric and parametric methods. With time permitting, we will discuss regression and multivariate models, including the treatment of missing data.

CONTRIBUTED PAPERS

Influence of Different Methods for Treatment of Missing Data on Results of Factor Analysis

EM

Jure Ačimovič, Andreja Kmet, Primož Kopač

jure.acimovic@mf.uni-lj.si; andrejakmet@yahoo.com; primoz.kopac@adacta.si
Postgraduate Study Programme in Statistics, University of Ljubljana, Ljubljana, Slovenia

Influence of different methods for treatment of missing data on results of factor analysis was studied. Initially, multivariate normal data with clear factor structure was generated. Afterwards, the data were deleted using MCAR method with repeatedly increasing ratio of missing data (1%, 2% 3%, 5%, 7%, 10%, 12%, 15%, 20%, 30%). On each of the new datasets, the following methods for treatment of missing data were used: listwise deletion, pairwise deletion, mean imputation, K-nearest-neighbour (for $K = 1$ and $K = 7$), EM algorithm and multiple imputation.

Measures of similarity were used to assess the difference between original factor loadings and factor loadings that were obtained from factor analysis where one of the methods for treated missing data was used. Four measures of similarity were applied: graphical representation of factor loadings, root mean square (RMS), coefficient of congruence (CC) and measure based on correlation matrix between variables (M_S).

The following conclusions were reached:

1. K-nearest-neighbour method for $K = 7$ works well only for the cases with maximum 12% of missing data;
2. According to RMS and M_S, the most efficient methods are pairwise deletion and K-nearest-neighbour for $K = 7$ (for maximum 12% of missingdata);
3. According to CC, the most efficient method is K-nearest-neighbour for $K = 7$ (for maximum 12% of missing data);
4. According to all quantitative measures, the least efficient method is listwise deletion.

Random Effects Modelling of Patient Pathways B1

Shola Adeyemi, Thierry J. Chaussalet, Haifeng Xie, Peter H. Millard

s.adeyemi3@wmin.ac.uk; chausst@wmin.ac.uk; h.xie@wmin.ac.uk, phmillard@tiscali.co.uk
Health and Social Care Modelling Group, University of Westminster, London, UK

In healthcare systems, the improvement of patients depends either on activities (or interventions) in each state visited in the care process or the recovery of patients progresses in stages (states) until discharge. The states visited represent patient journeys (pathways) that a particular clinical process may entail. Patients flow through the pathways attaining different health states as they journey through the system. Movement or flow between states is thus governed by subject specific random effects (or severity of disease). In this paper, we introduce a random effects modelling framework for patients flow through a health care system based on patients specific pathways. The model presented is based on the assumption that the random effects (or severity) govern the flow of patients between the states of the system. The model explains patients' journey through the system until discharge, based on patients' specific random effects, which can be used to predict subject specific discharge probabilities from each state. The discharge process induces some missingness which needs to be modelled since dropout cannot be ignored. A model is needed for the missing data as well as the measurement model and the dropout needs to be fitted jointly.

The model can be used to explain the system operations, predict patients specific discharge probabilities and to detect unusual system/patient behaviours. We present an application to a UK Department of Health dataset.

Some Ratio Estimators in Successive Sampling over Two Occasions DS

M. S. Ahmed, Atsu S. S. Dorvlo

msahmed@squ.edu.om; atsu@squ.edu.om
Department of Mathematics and Statistics, Sultan Qaboos University, Muscat, Sultanate of Oman

Key words: *successive sampling, chain-type ratio estimator, bias, mean square error*

In many situations, information on auxiliary variables may be readily available on different occasions (e.g., 1st, 2nd and so on) for all the units of a study variable in sample surveys. In successive sampling, it is advantageous to utilize the entire information collected in the previous occasions (Jessen, 1942; Patterson, 1950; Rao & Graham, 1964; Gupta, 1979; Das, 1982; Chaturvedi and Tripathi, 1983).

Sen (1971) developed estimators for the population mean on the current occasion using information on two auxiliary variables available on previous occasion. Further, Sen (1972, 1973) extended his work for the *p* auxiliary variables. Singh et. al. (1991) and Singh & Singh (2001) used the auxiliary information on current occasion for estimating the current population mean in two occasions successive sampling. Singh (2003) extended the work of Singh & Singh (2001) for *h*-occasions successive sampling. Feng & Zou (1997) and Biradar & Singh (2001) used the auxiliary information on both the occasions for estimating the current mean in successive sampling. In this paper, we propose a general chain type ratio estimator by using multivariate auxiliary information in successive sampling for multi-occasions. Most of recent estimators are the particular case of our general estimator. The properties of proposed general estimator have been studied for different sampling schemes.

Stochastic and Monte Carlo Simulation for the Spread of the Hepatitis B Virus

L1

M. Alahmed, A. Moneim

alahmed@ksu.edu.sa
King Saud University, Riyadh, Saudi Arabia

Modeling and simulation of human biology is one of the most recent and interesting techniques. Infectious and cell diseases are very serious major public health problems. These diseases are still an open field of study. Some them have not yet been understood while others need more effort to be completely solved. Constructing mathematical and simulation models for hepatitis B virus (HBV) as an example of such a disease is the main aim of our paper. Understanding, predicting and possibly controlling the nature and dynamic of these diseases are also important tasks. The HBV is considered to be among the most dangerous infectious diseases in the world. It is one of the common causes of cirrhosis and hepatocellular carcinoma (HCC). It has been shown that there is a threshold level of the basic reproduction rate R0 below which the disease dies out from the population, and above that threshold value the disease fires up. It is a parameter in both stochastic and Mont Carlo simulations of the disease. In the a stochastic model for the spread of the HBV, we use the method of stochastic partial differential equations to drive our stochastic model and then try to solve it. Monte Carlo simulations have also been conducted for this disease using random infection rates.

A Robust Estimation of Spline Function

T2

Ateq Ahmed Al-Ghamedi

drateq@yahoo.com
Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Key words: *spline function; heavy-tailed distributions; robust estimation*

The goal of the paper is to provide a robust estimation of spline function. In particular, we used t-distribution $\varepsilon_i \sim t(\nu)$. We showed that there is a unique solution of the problem which can be obtained through iterative procedures.

A New Robust Measures of Skewness

T2

Ateq Ahmed Al-Ghamedi

drateq@yahoo.com
Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Key words: *skewness; trimmed mean; heavy-tailed distributions; robust estimation*

The classical skewness coefficient is based on the first three moments of the data set, and it is affected by one or more outliers. In this paper we introduced new robust measures of skewness which depend on the trimmed mean. Monte Carlo simulations results show that the new measures give more accurate estimates of skewness of asymmetric univariate continuous distributions.

Comparison of Ordinary and Penalized Power-Divergence Test Statistics for Small and Moderate Samples in Three-Way Contingency Tables via Simulation

L1

Aylin Alin

aylin.alin@deu.edu.tr
Department of Statistics, College of Arts and Sciences, Dokuz Eylul University, Izmir, Turkey

Cressie & Pardo (2000) defined the family of ordinary power-divergence test statistics to test nested sequence of log-linear models. This family is determined by the values of λ and includes the well known likelihood ratio test statistic. In this paper, the family of penalized power-divergence test statistics is defined to

cope with some problems that the family of ordinary power-divergence test statistics cause when there is/are empty cell/cells in the contingency table. Their size and power properties to test a nested sequence of log-linear models is compared with the ordinary power-divergence test statistics for various penalization and λ values for small and moderate samples. Three-way contingency tables distributed according to a multinomial distribution are considered. The comparison is based on both asymptotic and designed simulation study results. Even though the likelihood ratio test statistic is the mostly used test statistic to test nested sequence of log-linear models, the results reveal that under the considered situations, the penalized power-divergence test statistics with penalization value of one and negative values of λ are better than the ordinary power-divergence test statistics for small samples, whereas for moderate samples, the ordinary power-divergence test statistics with negative values of λ perform better.

References

1. Cressie, N., Pardo, L. (2000). Minimum ϕ-divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, *10*, 867-884.

Properties of Some Statistical Tests to Establish Non-Inferiority for Independent Proportions

T1

Félix Almendra-Arao[1], David Sotres-Ramos[2]

1 falmendra@ipn.mx
UPIITA del Instituto Politécnico Nacional, México
2 davida.sotres@kendle.com
Colegio de Postgraduados, México

Key words: *non-inferiority, Barnard convexity condition, proportions, independent samples*

Non-inferiority tests are used very often in clinical trials to demonstrate that a new therapy (with minimum side effects or low cost) is not substantially less efficacy than the standard therapy.

For exact tests of non-inferiority, Röhmel and Mansmann (1999) proved that if the rejection region fulfills the Barnard convexity condition, then the level of significance can be computed as the maximum in a part of the boundary of the null space instead of being computed as the supremum in the whole null space.

This is particularly important due to the great amount of time required to compute levels of significance for non-inferiority tests.

However, there are some non-inferiority tests for which rejection region does not satisfie the Barnard convexity condition, but it satisfies a less restrictive condition. In this work, we derive some results for these more general rejection regions. These results extend those by Röhmel and Mansmann; the theorem of Röhmel and Mansmann is generalized in two directions: firstly the result is extended to general statistical tests (including exact and asymptotic tests); secondly, the Barnard convexity condition is relaxed to a less restrictive condition.

These results include hypotheses of non-inferiority for parameters such as difference, ratio, and odds ratio. For example, for the Blackwelder test, taking the difference of proportions as the parameter of interest, and 0.1 as the non-inferiority limit, the time of computing is reduced to approximately 1% of the original time.

References

1. Röhmel, J., Mansmann, U. (1999). Unconditional nonasymptotic one-sided tests for independent binomial proportions when the interest lies in showing noninferiority and/or superiority. *Biometrical Journal*, *2*, 149-170.

Comparison of Smooth Transition Stochastic Volatility Models with with Markov switching Stochastic Volatility Models in a Bayesian Approach

E1

Esmail Amiri

e amiri@ikiu.ac.ir
Department of Statistics, Imam Khomeini International Universiy, Ghazvin, Iran

The results of time series studies show that a sequence of returns on some financial assets often exhibit time dependent variances and excess kurtosis in the marginal distributions. Two kinds of models have been suggested by researchers to predict the returns in this situation: observation-driven and parameter-driven models. In parameter-driven models, it is assumed that the time dependent variances are random variables generated by an underlying stochastic process. These models are called stochastic volatility models(SV). In a Bayesian framework, we assume the time dependent variance of a Stochastic Volatility model(SV), follow a non-linear autoregressive model known as smooth transition

autoregressive(STAR) model and a Markov switching autoregressive model (MSAR), and estimate the parameters of these SV models using Markov chain Monte Carlo (MCMC) methods. We use deviance information criterion (DIC) for model comparison. A financial dataset is analyzed with the proposed models.

Nonparametric Estimation of the Drift Coefficient for a Stationary Process

T3

Mounir Arfi

m-arfi@hotmail.com
Department of Mathematics, College of Sciences, University of Bahrain, Kindom of Bahrain

Key words: *kernel, ergodic process, drift coefficient*

In this paper we investigate the kernel estimation of the drift coefficient under the ergodic hypothesis which is a condition implied by all the mixing hypotheses but less restrictive. We consider a diffusion solution of the stochastic differential equation

$$dX_t = \mu(X_t)\, dt + \sigma(X_t)\, dW_t, \quad t \in \Re^+,$$

where $(W; t>0)$ is a standard Brownian motion; μ and σ are two Lipschitz and unknown functions.

A kernel estimate of the drift coefficient μ is established and the almost sure convergence is obtained over a sequence of compact sets which increases to the real line $\Re$ when n approaches infinity. The observed process is supposed to be stationary and ergodic.

Measuring Explained Variance in Multiple Correspondence Analysis of Crisp and Fuzzy Coded Data

M1

Zerrin Asan[1], Michael Greenacre[2]

1 zasan@anadolu.edu.tr
Department of Statistics, Anadolu University, Eskişehir, Turkey
2 michael@upf.es
Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

It is well-known that in multiple correspondence analysis (MCA) of a multivariate

categorical data set, the eigenvalues severely underestimate the parts of explained variance, and proposals have been made to adjust these (Greenacre, 1988). The idea is to adjust the results of a MCA to explaining only the off-diagonal tables in the Burt matrix (the matrix of all two-way cross-tabulations of the variables). In this case a very good approximation of the adjusted solution is given by a simple calculation using the full set of eigenvalues from the MCA (Greenacre, 2007, chapter 19).

When continuous data are coded to categorical variables, two types of coding are possible: crisp coding, leading to a categorical data matrix which can be analyzed by MCA, or fuzzy coding, where each observation is coded by a set of probabilities of being in the set of defined categories. In the CA of the fuzzy coded data, exactly the same problem arises as in MCA, namely the eigenvalues underestimate the explained variance. In this case, a similar adjustment can be performed, by calculating rescaling factors for each axis to best approximate the off-diagonal matrices in the "fuzzy" Burt matrix. This calculation has to be performed using all the results of MCA – the coordinates as well as the eigenvalues – since there does not appear to be any shortcut to determining the adjustments as in "crisp" MCA.

References

1. Greenacre, M.J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, *75*, 457-467.
2. Greenacre, M.J. (2007). *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC Press.

Approximate Maximum Likelihood Estimation for the Scaled Generalized Exponential Distribution Based on Progressive Type-II Censoring

L1

Akbar Asgharzadeh

a.asgharzadeh@umz.ac.ir
Department of Statistics, Faculty of Basic Science, University of Mazandaran, Babolsar, Iran

For the scaled generalized exponential distribution, the maximum likelihood method does not provide an explicit estimator for the scale parameter based on a progressively Type-II censored sample. This paper provides a simple method of deriving an explicit estimator by approximating the likelihood function. We examine numerically the bias and variance of this estimator and show that this

estimator is as efficient as the maximum likelihood estimator (MLE). We present a numerical example to illustrate the methods of inference discussed here.

Parameter Estimation of Bernoulli Mixture Distribution T3

Senay Asma

senayyolacan@anadolu.edu.tr
Department of Statistics, Faculty of Science, Anadolu University, Eskisehir, Turkey

Key words: *latent class analysis, Bernoulli mixture, identifiability, maximum likelihood estimators*

The mixture of Bernoulli distribution is known to be non-identifiable. However, the estimation of this class of mixtures produces meaningful results in practice. So, this kind of mixtures are said to be practically identifiable. From this point of view, the aim of this study is the parameter estimation of the Bernoulli mixture distribution. Then, the practicality of the estimators is shown by using latent class analysis.

An Application of Experimental Design in Inorganic Chemistry DE

Mahdi Bashiri[1], Sheida Ansar[2]

1 bashiri@shahed.ac.ir
Industrial Engineering Department, Shahed University, Tehran, Iran
2 Tehran Azad University, Tehran, Iran

Statistical designs of experiments refer to the process of planning the experiment so that appropriate data that can be analyzed by statistical methods will be collected, resulting in objective and valid conclusions.

In this paper, a factorial experiment from inorganic chemistry is presented. The results show that synthesis of diazaphospholes and diazaphosphorinans is significantly affected by solvent type, whereby the effect interacts with amin structure and intermediate type.

Network Analysis of Data from Web of Science NA

Vladimir Batagelj[1], Nataša Kejžar[2]

1 vladimir.batagelj@fmf.uni-lj.si
Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
2 natasa.kejzar@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

Web of Science is a database that provides information about current and past articles published in approximately 8,700 of the most prestigious, high impact research journals in the world. One can download a file with full information about searched records (articles) and analyze that further.

Our general search query consisted of term "social network*" and it gave us 6257 hits (April 2007). A new program WoS2Pajek was developed in order to convert these data into Pajek network files. Due to the known problems of non-unique authors' names and citation descriptions, abbreviated record/citation descriptions were used. The network of articles, network of articles × authors, and partitions (according to year published and description) were analyzed in order to obtain the most important people and works that have been involved in the social networks field for the last few decades. A special emphasis was given to determining the boundary of the network and to the exploration of the network of citations among authors.

Hierarchical Clustering with Relational Constraints of Large Datasets D1

Vladimir Batagelj[1], Andrej Mrvar[2]

1 vladimir.batagelj@fmf.uni-lj.si
Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
2 andrej.mrvar@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

In the paper, an adaptation of the hierarchical clustering with relational constraints approach (Ferligoj & Batagelj, 1982, 1983) to large data sets is presented.

To obtain an efficient algorithm for large networks, we:

- compute the dissimilarities between units (vertices of network) only for endpoints of existing links (of constraining relation);
- define the dissimilarities between clusters based only on the dissimilarities of the corresponding links and derive the update relations.

We also show that for selected dissimilarities between clusters, the Bruynooghe (1977) reducibility property holds. This allows us to speed-up the hierarchical clustering procedure by using the RNN (reciprocal nearest neighbors) approach.

The developed algorithms are implemented in *Pajek* – a program for analysis and visualization of large networks (http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm).

References

1. Bruynooghe, M. (1977). Méthodes nouvelles en classification automatique des données taxinomiques nombreuses. *Statistique et Analyse des Données*, *3*, 24-42.
2. Ferligoj, A., Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, *47* (4), 413-426.
3. Ferligoj, A., Batagelj, V. (1983). Some types of clustering with relational constraints. *Psychometrika*, *48* (4), 541-552.

Canonical Correlation Analysis and Kernel Smoothing and Their Application to Relations and Indices of Organizational Variables

A2

Hakim Bekrizadeh[1], Fazlolah Azizi[2]

1 h_bekri@yahoo.com
Department of Statistics, Payam Noor University, Ilam, Iran
2 Department of Statistics, Shahid Chamran University, Ahwaz, Iran

Key words: *canonical correlation analysis, kernel smoothing, Gaussian kernel function*

Canonical correlation analysis is a technique applied for extracting common properties of a pair of multivariate datasets to find their linear transformation in a way that the correlation is maximized. Kernel canonical correlation analysis can be used when the relation among the data pairs is known to be nonlinear. Using a desirable kernel function (Gaussian), this method finds the proper nonlinear transformation and avoids the unwanted local optima.

In this papers, we use the kernel method and canonical correlation analysis to determine basic relations and main indices among two sets of organizational variables.

Estimation of Non-Response Bias in Round 2 of the European Social Survey: Using Information from Reluctant Respondents

DS

Koen Beullens, Jaak Billiet

koen.beullens@soc.kuleuven.be; jaak.billiet@soc.kuleuven.be
Department of Sociology, Faculty of Social Sciences, University of Leuven, Leuven, Belgium

In five countries participating in the European Social Survey, round 2 has been examined with regard to the differences between initially cooperative and reluctant respondents, whereby it is assumed that reluctant respondents are indicative or informative of final refusers. In two of the five countries, the Netherlands and Germany, a considerable amount of refusal data (n = 500) allows one to explore more fruitful distinctions between hard and soft refusals, so that reluctance toward surveys can be studied under various assumptions and definitions. This paper illustrates the importance of contact forms and the necessity to further standardize these means of survey research, as contact data keep track of every step of the contact process between sample unit and interviewer(s). Although the results are far from decisive, they indicate that reluctant respondents differ from cooperative ones in a myriad of ways. Not only background variables are sensible to reluctance, also target variables (attitudes, reported behaviour) differentiate between cooperativeness and recalcitrance. Multivariate approaches suggest that these biases can not be undone by introducing weighting variables based on known population distributions such as gender, age, residence or level of education.

Exact Simultaneous Confidence Regions for Genetic Problems

B2

Karl-Ernst Biebler, Bernd Jäger, Michael Wodny

biebler@biometrie.uni-greifswald.de; bjaeger@biometrie.uni-greifswald.de
Institut für Biometrie und Medizinische Informatik, Ernst-Moritz-Arndt-Universität, Greifswald, Germany

The problem of estimation of allele probabilities from samples appears in genetic studies. For more than two alleles, the problem is no longer trivial. Simultaneous confidence estimation is possible if the associated point estimator is unbiased and effective. We calculate exact simultaneous confidence intervals with the help of Dirichlet integral and compare them with asymptotic results.

Comparison of Statistical Models Describing Power Semiconductor Cycle Life Data

A1

Olivia Bluder[1,3], Helmut Köck[1,2], Michael Glavanovics[1]

1 olivia.bluder@k-ai.at
KAI – Kompetenzzentrum Automobil- und Industrie-Elektronik GmbH, Villach, Austria
2 Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
3 Carinthia University of Applied Sciences, Klagenfurt, Austria

Introduction

Lifetime testing of semiconductor devices often needs months until enough devices have failed, or until the required cycles to failure are reached. This means that the effort for getting valid statistical data is high. The aim is to receive a good and efficient model from a series of measurements for the prediction of small quantiles, so that only a necessary minimum of devices needs to be tested, and testing time can be saved.

Data analysis has been based on the log-normal distribution as appropriate life time distribution for our measurements. This paper shows other possible models, such as the Weibull-model, the Generalized linear model (GLM) and the nonparametric Cox-Proportional-Hazard model, and compares the results for a given set of data.

Data

Statistical lifetime test data are provided by a temperature cycle stress test system for integrated power semiconductor devices (Glavanovics et al., 2007). For this study, one type of power switches was tested under different ambient conditions with different electrical and thermical stress conditions. The test equipment measures, monitors and records the state of every device under test (DUT) during test (Glavanovics et al., 2007). To analyze stress-related degradation mechanisms of smart power devices, satisfactory flexibility in the definition of test conditions can be achieved with 9 measurement parameters that affect the Cycles to Failure (CTF) of the DUTs: ambient temperature, pulse shape, clamping voltage, peak current, pulse length, repetition rate (frequency), clamping energy, temperature of the case and the rise of temperature during the load pulse. The devices may fail in a short circuit condition (switch continuously on) or in an open load mode (switch continuously high-ohmic). Non-failed devices are called survivors. There are two conditions for ending the testing period: (a) all DUTs have failed or (b) the maximum duration of the test has been reached. This leads to two kinds of data, uncensored or censored.

Methods

The aim of this work is to find a more explanatory model for fitting data and for predicting CTFs, for example a model that includes the measurement parameters. For well behaved survival data, that represent a single distribution without outliers, both the Log-normal and Weibull distribution are suitable. The decision for either one of them was based on graphical (e.g. histogram, kernel estimator) and analytical (Kolmogorov-Smirnov test) methods (Hartung, 2002).

Another approach for the prediction was to generate a GLM

$$\log_{10}(CTF) = \vec{x}_i * (\vec{\beta}_i)^t \quad \text{for } i = 1, \ldots, n$$

with $\vec{x}$ being the vector of the measurement parameters and $\vec{\beta}$ the vector of the model parameters. The elements of $\vec{\beta}$ were calculated under the assumption of the log-normal distribution with the least-squares method.

For the decision which one of the proposed models fits the data best, the coefficient of determination (R^2), the Bayes Information Criteria (BIC) and Mallow's Cp were used (Toutenberg, 2003). To compare the goodness of the GLM with a nonparametric model we used the Cox Proportional Hazards models. This type of model is appropriate for survival data with and without censoring (Kalbfleisch & Prentice, 1980; Lee, 1980).

Results and Conclusion

For well-behaved data both distributions, the log-normal and the Weibull, are appropriate, but the errors when using the log-normal distribution are smaller. Both models are not suitable for data that leads to a double distribution.

For the best GLM, 11 variables were used (containing also combinations of variables). Between all possible GLMs, the decision for this model was based on comparisons of the R^2, the residual standard errors, the BIC and the Cp values.

The assumed Cox Proportional Hazards model uses the same variables as the GLM, but does not have the same potential of explanation like the GLM The advantage is that no underlying distribution needs to be evaluated. Another advantage of this method is that for the estimation of the model parameters, the baseline hazard needs not be calculated. But the quality of the prediction turned up to be useless, because the model predicted a negative CTF. Our explanation is

that for getting a good prediction out of a nonparametric model, the amount of data needs to be larger.

With the data collected until now, the GLM leads to the best predictions, because it considers the underlying distribution and the involved parameters. A good alternative for the GLM is the log-normal model, because the potential of explanation is 0.9738, but with the disadvantage that for every type of power switch and for every kind of variable a new model needs to be generated. The Cox model is not a good alternative because the amount of collected data is not sufficient. For increasing the number of measured points, more tests are running and planned for the next few months. As a next step, nonparametric models will be investigated for the explanation of not-so-well-behaved data, which can not be described by assuming a single GLM.

References

1. Glavanovics, M., Köck, H., Eder, H., Košel, V., Smorodin, T. (2007). A new cycle test system emulating inductive switching waveforms. In 12th European Conference on Power Electronics and Applications (in press).
2. Hartung, J. (2002). *Statistik*. München, Wien: Oldenburger.
3. Kalbfleisch, J.D., Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
4. Lee, E.T. (1980). *Statistical Methods for Survival Data Analysis*. Belmont, CA: Wadsworth.
5. Toutenberg, H. (2003). *Lineare Modelle: Theorie und Anwendung*. Heidelberg: Physica..

Accumulation of Vague Data: Non-Precise Counts **T1**

Slavka Bodjanova

kfsb000@tamuk.edu
Department of Mathematics, Texas A&M University, Kingsville, Texas, USA

Vague data incorporating non-statistical uncertainty are often analyzed by methods of fuzzy sets. We will consider a sample of vague observations described by fuzzy numbers defined on a closed interval of real numbers. A crisp or fuzzy partition of this interval will discretize the sample into a collection of crisp or fuzzy granules. Because a vague observation may belong to different granules with different degree of membership, cardinality of each granule is, in general, non-precise. Work on accumulation of vague data has already been initiated by

several researchers with the goal to construct a "fuzzy histogram". In this paper we propose an axiomatic definition of non-precise scalar counts and an axiomatic definition of non-precise fuzzy counts. Non-precise scalar counts are values of a special type of a non-negative real function. Non-precise fuzzy counts are convex fuzzy sets defined on a set of all natural numbers including zero. We suggest a general approach to the construction of non-precise counts. Lower and upper frequency functions are discussed and evaluated on a small sample of vague data.

Bibliography

1. Bodjanova, S. (1999). Fuzziness and roughness of nonprecise quantities. *Austrian Journal of Statistics*, *28* (4), 173-194.
2. Bodjanova, S. (2000). A generalized histogram. *Fuzzy Sets and Systems*, *116*, 155-166.
3. Viertl, R. (1996). *Statistical Methods for Non-Precise Data*. Boca Raton: CRC Press.
4. Viertl, R., Trutsching, W. (2006). Fuzzy histograms and fuzzy probability distributions. In B. Bouchon-Meunier & R.R. Yager (Eds.), *Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2006), Paris, France, 2-7 July 2006*. Berlin: Springer, 957-964.
5. Wygralak, M. (2003). *Cardinality of Fuzzy Sets*. Berlin: Springer.
6. Zadeh, L.A. (1965). Fuzzy sets. *Information Control*, *8*, 338-353.

Grades from 1 to 5 or A to E: From Theory to Practice **M2**

Matevž Bren[1], Darko Zupanc[2], Andrej Blejec[3]

1 matevz.bren@fov.uni-mb.si
Faculty of Organizational Sciences, University of Maribor, Kranj, Slovenia
2 darko.zupanc@guest.arnes.si
National Examination Center, Ljubljana, Slovenia
3 andrej.blejec@nib.si
Department of Entomology, National Institute of Biology, Ljubljana, Slovenia

In some countries, grades in schools range from A to E, or common categories are officially used (Excellent, Very Good, Good, Sufficient), meaning that grades are ordinal measurements with the mode and the median as appropriate measures of central tendency.

In other countries, grades at schools are awarded using numbers. Common grades range from 1 (insufficient) to 5 (excellent), or from 1 to 10, or even 20, somehow

pretending to be measured on an interval scale, thus allowing addition and substraction and therefore also implying arithmetic mean as an appropriate measure of central tendency. Indeed, using grade point average (GPA) and standard deviation is common practice when analysing gradings in those countries.

We discuss (1) appropriate presentation of grading results in the students grading of their teachers, (2) methods for comparing students achievements in one subject, and (3) overall achievement presentation as applied in the ALA (Assessment of/for Learning Analytic) Tool for gathering and analysing grades in upper secondary education in Slovenia.

Moreover, we hope to make a step further in appropriate presentation and analysis of grading.

Concurrent Programming for Extracting Knowledge with Uncertainty | D1

Gerard Brunet

gerard.brunet@univ-poitiers.fr
IUT-STID, Centre de Noron, Niort, France

Key words: *images, object shape, concurrent programming, uncertainty*

The aim of the paper is to use concurrent programming for extracting the shape of an object in an image. The object is not drawn precisely – there is only a probability of presence of pixels, which is higher in the place where the object is located. Concurrent programming uses Threads with Java language and Semaphores for synchronisation of processes. A parallel determination is performed for the whole image, divided into squares, then merging is performed from the information on each square with the description of the square with uncertainty.

A Longitudinal Study of Student Performance in English using Repeated Measures Analysis of Variance and Multilevel Modelling

SM

Liberato Camilleri, Georgiana Xuereb

liberato.camilleri@um.edu.mt; gxue006@um.edu.mt
Department of Statistics and Operations Research, University of Malta, Malta

Longitudinal data arise when multiple observations are made on each subject over time. Repeated measurements on the same subject are more likely to be correlated than measurements on different subjects. For this reason, models that are fitted to longitudinal data involve the estimation of covariance parameters to capture this correlation and make valid statistical inference. There are, basically, two approaches to modelling such data. One approach is a repeated measures analysis of variance that allows explicitly the selection of a plausible variance-covariance structure. The second approach is multilevel modelling which provides a powerful framework for exploring how average relationships vary across the hierarchical structure of the study design. These statistical procedures are employed to analyze the marks obtained in English by a number of students during the last three years in primary schools. The random sample comprises 325 male and female students, who attend either a state or a private school.

In a Repeated Measures setting the within-subjects factor is defined by grouping the responses (English marks) measured for each student. This within-subjects factor has three levels – one level for each repetition. The student gender and the type of school attended by the student are the between-subjects factors. In practice, the covariance structure for the measurement of each student is unknown and has to be estimated from the data. There are several forms for this variance-covariance matrix, which include the Unstructured, Compound Symmetry, Autoregressive, Toeplitz and Banded structures. The selection of this covariance matrix is based on three information criteria – AIC (Akaike's Information Criteria), AICC (AIC Corrected) and BIC (Bayesian Information Criteria). In this study we conduct several hypothesis tests to examine the within-subjects and between-subjects effects, equality of covariance matrices and sphericity. We also fit a parsimonious repeated measures model to these normal responses.

Multilevel modelling is an alternative approach for analyzing longitudinal data. These models are hierarchical structures that accommodate nested observations within several levels of classification. In this study, the hierarchical model has two levels reflecting the contributions of the time level (level 1) and the student level (level 2). We fit a random coefficient model to examine the effect of time on student performance in English and simultaneously investigate the amount of

between-subject variance in the effects of student gender and the type of school attended by the student across level-2 units. This model allows the student-specific coefficients describing individual trajectories to vary randomly. We employ a number of diagnostics, using graphical procedures, to assess the assumption of constant variance for the residuals. We check the normality assumption of the random effects and residuals and evaluate the agreement between the actual observed responses and the conditional predicted values..

Bibliography

1. Hand, D., Crowder, M. (1996). *Practical Longitudinal Data Analysis*. London: Chapman & Hall.
2. Rabe-Hesketh, S., Pickles, A., Skrondal, A. (2001). GLLAMM: A General Class of Multilevel Models and Stata Program. *Multilevel Modelling Newsletter*, *13*, 17-23.
3. Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling*. Boca Raton, FL: Chapman & Hall / CRC.
4. West, B.T., Welch, K.B., Galecki, A.T. (2007). *Linear Mixed Models – A Practical Guide using Statistical Software*. Boca Raton, FL: Chapman & Hall / CRC.

Follow-up Qualitative Study for Non-Interpretable Social Network Variables

NA

Aina Maria Capó Artigues[1], Germà Coenders[1], Lluís Coromina[2]

1 aina.capo@udg.edu; germa.coenders@udg.es
Faculty of Economics, University of Girona, Girona, Spain
2 lluis.coromina1@esade.edu
ESADE Business School, University Ramon Llull, Barcelona, Spain

We try to explain some unexpected results from a previous quantitative research. The aim of the quantitative research was to predict the PhD students' academic performance at the University of Girona. Explanatory variables were characteristics of the PhD students' research group understood as a social network, background and attitudinal characteristics of the PhD students and some characteristics of the supervisors. Academic performance was measured by the weighted number of publications.

The unexpected results were the lack of predictive power of social network variables on PhD students' academic performance. If network variables fail to

predict performance it is because the four possible profiles of PhD students are in more or less equal proportions: good research group high performer, good group low performer, bad group high performer, and bad group low performer.

These unexpected results are analyzed by a qualitative research in order to generate hypothesis about lack of significant effects.

The qualitative design is to identify a few students in each of the four profiles and learn which other unknown variables make the difference between good and bad performers given a particular group type, by means of in-depth personal interviews. We expect that the qualitative study can uncover the reasons why the quality of the group fails to translate into the quality of the student's work (e.g., a lot of collaboration contacts within the research groups may imply a big workload of the students, which diverts them from their main research and PhD related tasks).

Characteristics of a New Bivariate Distribution **T1**

Anton Cedilnik[1], Andrej Blejec[1,2], Katarina Košmelj[1]

1 anton.cedilnik@bf.uni-lj.si; katarina.kosmelj@bf.uni-lj.si
University of Ljubljana, Ljubljana, Slovenia
2 andrej.blejec@nib.si
National Institute of Biology, Ljubljana, Slovenia

We constructed the density for a new bivariate distribution for $\mathbf{Z}=[X,\ Y]^{\mathrm{T}}$ for $x, y > 0$:

$$z(x,y) = C \cdot \left(\frac{x}{a}\right)^{mr-1} \left(\frac{y}{b}\right)^{ns-1} \exp\left(-\left(\left(\frac{x}{a}\right)^{r} + \left(\frac{y}{b}\right)^{s}\right)^{\frac{1}{p}}\right).$$

It has 7 parameters: r and s are power parameters, m and n are moment parameters, a and b are scaling parameters, and p is the linking parameter. The distribution is very general and has some nice mathematical properties, i.e., moments of all kinds are easily calculated. In addition, we propose an iterative procedure for parameter estimation.

We define two random variables U and V, where U is the sum of $(X/a)^r$ and $(Y/b)^s$, and V their ratio. We derive densities and moments for U and V, and show that U and V are independent.

Described characteristics are illustrated on a specific dataset. A sample of 500 randomly generated pairs (x, y) was obtained from the distribution with a known set of parameters; for this purpose, the acceptance-rejection method was used. Parameters were estimated by the iterative estimation procedure and used to transform the (x, y) space into (u, v) space. Data analysis techniques were used to demonstrate the interesting features of the (X, Y) and (U, V) correspondence.

These results are applicable in particular when studying the ratio of two positive continuous variables (e.g., body mass index in medicine, dimension ratio in forestry, etc.).

Variance Balanced Block Designs **DE**

Bronisław Ceranka, Małgorzata Graczyk

bronicer@au.poznan.pl; magra@au.poznan.pl
Department of Mathematical and Statistical Methods, Agricultural University of Poznán, Poznán, Poland

Key words: *balanced incomplete block design, repeated blocks, variance balanced block design*

We present some types of block designs that are use full in practice as well as in the general theory of block design. The designs discussed are designs with repeated blocks with the equireplications and equiblock sizes. They are available in the literature. From the practical point of view, sometimes it may be not possible to construct the design with equisize blocks accomodating the equireplication of each treatment in all the blocks. Here we consider a class of block designs called variance balanced block designs, which can be made available in unequal block sizes and for varying replications, and we present new construction method of these designs. Some new construction methods of variance balanced block designs with repeated blocks are also given. For construction methods, we use the incidence matrices of the balanced incomplete block designs and the specialised product of two matrices introduced by Pal and Dutta.

Optimum Chemical Balance Weighing Design under Certain Condition | DE

Bronisław Ceranka, Małgorzata Graczyk

bronicer@au.poznan.pl; magra@au.poznan.pl
Department of Mathematical and Statistical Methods, Agricultural University of Poznán, Poznán, Poland

Key words: *chemical balance weighing design, ternary balanced block design*

We consider the standard linear model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad \mathrm{E}(\mathbf{e}) = \mathbf{0}_n, \quad \mathrm{E}(\mathbf{ee'}) = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{y} = [y_1, y_2, ..., y_p]'$ denotes the recorded observations in n operations, and $\mathbf{X}$ is the matrix of order $n \times p$ that is called the weighing design matrix for determining unkown measurements of p objects in n measurement operations. The elements of $\mathbf{X}$ are x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, and a typical element x_{ij} is -1 if the jth object is placed on the left pan during the ith weighing operation, $+1$ if the jth object is placed on the right pan during the ith weighing operation and 0 if the jth object is not utilized in either pan during the ith weighing operation. Hence, $\mathbf{w} = (w_1, w_1, \ldots, w_p)'$ is the vector of true unknown weights (of parameters). The vector $\mathbf{e}$ is the so-called vector of error components satisfying the usual homoscedasticy condition.

The problem is how to find the design matrix $\mathbf{X}$ in such a way that the variance factors are minimized. We are interested in optimal estimation in the sense of Hotelling. Additionally, we assume that not all objects are included in each measurement operation and the errors are equally correlated. We present new construction methods of the optimal chemical balance weighing designs based on the incidence matrices of the ternary balanced block designs.

Edge Correction for Segregation Tests Based on Nearest Neighbor Contingency Tables

A2

Elvan Ceyhan

elceyhan@ku.edu.tr
Department of Mathematics, College of Arts and Sciences, Koç University, Sariyer, Istanbul, Turkey

The spatial patterns of segregation and association between two or more classes (or species) have important implications. These patterns can be studied using a nearest neighbor contingency table (NNCT). Pielou's test of segregation is equivalent to the usual (Pearson's) test of independence, but is liberal under complete spatial randomness (CSR) or random labeling (RL). Dixon's test of segregation based on NNCTs has the desired significance level α. Previously, we proposed three new segregation tests based on NNCTs using the appropriate sampling distribution of the cell counts and suggested a simple correction to adjust Pielou's test for data with rectangular support (Ceyhan, 2007), and demonstrated that one of our newly proposed segregation tests outperformed all the others. The null patterns for all these tests are either complete spatial randomness (CSR) or random labeling (RL). The former model assumes that the study region is unbounded for the analyzed pattern, which is not the case in practice. In this article, the edge or boundary effects on various exact tests for NNCTs under the CSR pattern are discussed. It is demonstrated that (inner and outer) buffer zone edge corrections can severely affect the results of these tests; and the toroidal edge correction has a moderate influence on the results of the tests. It is not recommended to use buffer zone corrections for the tests, but one can use toroidal edge correction, provided that no clusters exist around the edges. For illustrative purposes, Pielou's Douglas-fir/ponderosa pine data, swamp tree data, and an artificial data set are used as examples.

New Educational Plans for New Professional Profiles: The Social Network Analysis Approach

NA

Marisa Civardi, Emma Zavarrone, Paola Zappa

emma.zavarrone@unimib.it
Department of Economics, University of Milano – Bicocca, Milano, Italy

This work aims at developing an evaluation methodology of compatibility between demand and supply of labor by elaborating the "ideal composition of specific knowledge" of a degree course. The ideal course is best able to meet labor

market requirements in terms of knowledge, flexibility and problem-solving abilities of given professional profiles in the chosen areas of economic activities (as identified based on NACE classification).

To this purpose, a dataset of professions referring to specifically selected areas has been constructed from the O*NET database (Occupational Information Network, developed and run by the US Department of Labor, Employment and Training Administration) and several data have been collected (or derived through properly defined rules) for each of them: relevance of a given profession within the activity (i.e., pertinence with the activity considered), specific knowledge required, composition, weight and macro areas of study (subjects taught at the university, as defined by the ministry in charge of the university system and scientific research).

Then, social network theory (Scott, 1991; Snidjers, 2005; Wasserman & Faust, 1994) was applied. As it considers each economic activity as a set of professions (the nodes) linked to each other by a symmetric relation of similarity of the knowledge set needed (the edges), this methodology provides an original way to investigate and describe the labor market. By adopting the network theory, some descriptive measures cane be obtained (centrality and degree of centralization, density, clusters or cohesive subgroups, bridges –Borgatti, 2005), and it is possible to identify which kinds of professional profiles are more important within each economic activity, which share more knowledge and competencies with others (i.e., people covering these positions have chances to be employed in many more jobs) and what is the level of use of specific academic knowledge in each profile. Including these outcomes in elaborating the "ideal composition of specific knowledge" of a Bachelor degree as described above can provide a more realistic and pragmatic view of the labor market.

Although this study is directed towards the educational system, mainly the universities, most findings are useful for recruiters and applicants as well, providing them with indications on which candidates best match specific profiles and on feasible career paths.

References

1. Borgatti, S.P. (2005). Centrality and network flow. *Social Networks*, *27* (1), 55-71.
2. Scott, J. (1991). *Social Network Analysis*. New York, London: Sage.
3. Snidjers, T.A. (2005). Models for longitudinal network data. In P.J. Carrington & S. Wasserman (eds.), *Models and Methods in Social Network Analysis*, New York, Cambridge University Press.

4. Wasserman, S., Faust, K. (1994). *Social Networks Analysis: Methods and Applications*. New York: Cambridge University Press.

A Hybrid Approach to Data Mining Radiological Medical Records D2

William Claster, Subana Shanmuganathan, Nader Ghotbi

wclaster@apu.ac.jp; s5subana@apu.ac.jp; nader@apu.ac.jp
Ritsumeikan Asia Pacific Univeristy, Beppu, Japan

In this paper we extend our previous work, in an effort to extract meaningful information from medical records using data mining techniques. The medical data derive from patients' radiology department records where CT scanning was used as part of a diagnostic exploration. The records are from the digital records of about 900 pediatric patients who were CT scanned through a one year period in 2004 at the Nagasaki University Medical Hospital in Japan. This approach led to a model based on SOM clusters and statistical analysis which allow for the prediction of when a particular medical screening procedure may be unnecessary. The procedure involves CT scans of patients. This is important because radiation at levels ordinarily used for CT scanning may pose significant health risks especially to children. The medical records are employed to compare clustering and classification methods on text data. SOM and K-means are compared to develop alternative cluster perspectives. Then we use a novel approach developing cluster profiles by using association rules for describing individual clusters. We then compare SOM based hierarchical trees to C5.0 trees to measure the accuracy of these two approaches to decision rules. Finally we compare classification methods on the medical records including neural networks, vector space classification using hyperplanes, k-nearest neighbors, and logistic regression using k-fold validation to measure accuracy.

Statistical Analysis of Multiculturalism Research in Vojvodina

A2

Katarina Čobanović[1], Valentina Sokolovska[2], Emilija Nikolić-Đorić[1]

1 katcob@polj.ns.ac.yu; emily@polj.ns.ac.yu
Department of Agricultural Economics and Rural Sociology, Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia
2 valentina@unsff.ns.ac.yu
Department of Sociology, Faculty of Arts, University of Novi Sad, Novi Sad, Serbia

The paper is based on the analysis of a survey conducted in 2006. A simple random sample was used, consisting of about 1200 persons distributed over seven counties of Vojvodina.

The paper deals with the part of the study concerning the problem of socio-cultural and economic aspects of multiculturalism in Vojvodina. The data consist of a multivariate matrix of categorical and numerical variables.

Multigroup analysis is based on methods for classification of variables and some nonparametric methods. Multicultural aspects are analized using cluster analysis and contingency table analysis. Contingency tables in two and three dimension are the basis for estimation of the indicators of association and correlation between variables. Multigroup analysis supposes specific problems of comparisons over different variables representing different cultural, national, ethnic and other levels of the investigation. Hence, some comparisons of socio-economic, cultural, ethnic, national and other aspects are made on the county level and for the Province of Vojvodina.

Comparison of PhD Students' Performance with Duocentered Network Measures

NA

Lluís Coromina[1], Aina Maria Capó[2], Germà Coenders[3], Anuška Ferligoj[4], Uroš Matelič[5]

1 lluis.coromina1@esade.edu
ESADE Business School, University Ramon Llull, Barcelona, Spain
2 aina.capo@udg.edu
Faculty of Economics, University of Girona, Girona, Spain
3 germa.coenders@udg.es
Faculty of Economics, University of Girona, Girona, Spain
4 anuska.ferligoj@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
5 uros.matelic@valicon.net
VALICON, Ljubljana, Slovenia

The main goal of this article is to predict PhD students' academic performance in the universities of Girona (Spain) and Ljubljana (Slovenia) and compare this performance across countries. Explanatory variables are characteristics of PhD student's research group understood as a social network, background and attitudinal characteristics of the PhD students and some characteristics of the supervisors. Academic performance, which is the dependent variable, was measured by the weighted number of publications.

Comparable versions of the questionnaire were designed in each language (Catalan and Slovenian) taking into account the differences in university systems, which involved two independent translations, a pre-test of the translated questionnaires and further discussions and modifications. When direct comparison was not possible, we created comparable indicators.

The measures for network variables are substantially different from those found in the literature on social network analysis. We used measures computed from duocentered networks structure. The main characteristics of this type of network is that it is based a pair of central actors, and their contacts, but the ties among their contacts are unobserved. This network provides more information than the egocentered network, because even if there are two central actors (such as PhD student and supervisor), the egocentered networks would consider one of them as a conventional alter. Several measures adapted from complete networks and tailor-made measures can be created for a duocentered network structure.

We used a regression model in which the country was introduced as a dummy coded variable, including all possible interaction effects. The optimal transformations of the main and interaction variables are discussed.

The paper is based on a previous comparative study between Slovenia and Girona based on egocentered network measures (Capó et al., in press). As in that paper, attitudinal, background and network variables prove to be all predictors of academic performance, though some differences between countries emerge. The predictive power of the network variables in the present study is higher due to the use of duocentered network measures.

Assessing Suicidal Intent in an Epidemiological Study Using the Cumulative (Proportional) Odds Model for Ordinal Variables

SM

Colette Corry, Brendan Bunting, Siobhan McCann

corry-c2@ulster.ac.uk; bp.bunting@ulster.ac.uk; sm.mccann@ulster.ac.uk
School of Psychology, University of Ulster, Magee campus, Londonderry, Northern Ireland, UK

The cumulative odds model conceptualizes how data may be sequentially compartmentalized into dichotomous groups, while taking into account the ordering of responses. It considers the impact of a set of independent variables across these possible consecutive cumulative splits to the data, which may be utilized to provide a single parsimonious prediction model.

A total of 9282 English speaking adults aged 18 years or older, living in the non-institutionalised civilian household population of the coterminous United States completed the National Co-morbidity Survey-Replication (2001-2003). The diagnostic instrument was the World Health Organisation (WHO) Composite International Diagnostic Interview (CIDI), including demographic characteristics, history of substance use/abuse, psychiatric history and treatment, risk behaviours and physical/mental health status.

In this study, the cumulative (proportional) odds model was applied to the suicidality section of the CIDI. Results support the underpinning theory that an inherent ordering exists between levels of the behaviour, from ideation (thoughts of taking one's own life) towards enactment.

Application of Regression Models and Polynomial Equations to Predict Out-Crossing Rate of Maize

D2

Marko Debeljak[1], Aneta Ivanovska[1], Dragi Kocev[1], Sašo Džeroski[2], Katja Rostohar[2]

1 marko.debeljak@ijs.si; aneta.ivanovska@ijs.si; dragi.kocev@ijs.si; saso.dzeroski@ijs.si
Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
2 katja.rostohar@kis.si
Crop and Seeds Science Department, Agricultural Institute of Slovenia, Ljubljana, Slovenia

Key words: *gene flow, regression trees, predictive clustering trees, polynomial equations, WEKA, CLUS, CIPER*

Introduction

Pollen dispersal can represent a significant proportion of the gene flow in flowering plants and has long been of interest in agriculture as a potential source of admixture of one crop variety with the pollen of another. This became more important with the advent of genetically modified (GM) crops and related regulations, where the potential of transgenic pollen to cross pollinate with non-transgenic or even wild relatives, and thereby spread the modified genes, needs to be estimated (Commission 2003; European Parliament and the Council, 2003a,b).

To estimate the impact factors on outcrossing frequency between two varieties of maize – donor variety with yellow kernels (simulating the GM variety) and the recipient variety with white kernels (non-GM variety) – a field trial was designed in the year 2006. The site of 120 by 120m was allocated in central part of Slovenia. A central square field (20 by 20m) was planted with yellow kernels variety surrounded with white kernel variety. In total, 1470 samples were collected. A yellow coloured grain in a white coloured variety was considered as an outcross event. Every sampling location was determined with spatial coordinates for further spatial modelling of pollen distribution. During the growing period, the meteorological parameters were monitored and data describing properties of boundary layer (temperature, humidity, air pressure, wind direction and wind velocity) were measured. Phenological parameters were monitored as well. Each sampling point (1470) was described with the following set of attributes: angle from the centre of the donor field, distance from the centre and from the nearest edge of the donor field, visual angle of the donor field, the percentage of appropriate wind (the percentage of flowering time when the wind was blowing over the donor field to the sample plot), and the length of the wind ventilation route (the cumulative lengths of wind paths multiplied by wind strength over the donor field during flowering).

Methods

Regression trees are a representation for piece-wise constant or piece-wise linear functions. Like classical regression equations, they predict the average value of a dependent variable from the values of a set of independent variables (called attributes). Leaf nodes give a linear equation (model trees) or a constant (regression trees) that applies to all instances that reach the leaf. Regression trees partition the space of examples into axis-parallel rectangles and fit a model for each of these partitions.

In our analyses, we used three different approaches (tools): WEKA, CLUS and CIPER. The WEKA workbench, which is a collection of data mining algorithms and data preprocessing tools, was used for building model trees and regression trees. CLUS is a system for constructing decision trees for prediction and clustering tasks which supports different types of constraints, such as minimal size and maximal accuracy. CIPER (Constrained Induction of Polynomial Equations for Regression) is a system that uses a beam search algorithm that heuristically searches through the space of possible polynomial equations that best fit the data. It is comparable in performance to other commonly used methods for regression, such as model trees.

Results and Discussion

Analysing the data, we discovered a few outliers, i.e., samples which show high percent of outcrossing despite their relatively large distance from the donor field. The reason that these outliers exist is believed to be a data error or an unusual event that happened while the field experiments were carried out. Therefore, we decided to remove them from the dataset.

We built model trees and regression trees (WEKA), predictive clustering trees (CLUS) and generated polynomial equations (CIPER) on the data with and without the outliers. In each of the analysis on the data with removed outliers, we obtained a high correlation coefficient of around 0.80 and a RMSE and RRMSE of around 2.25 and 0.60, respectively. These results were far better than the results obtained from the data that included the outliers, which shows that they have a big effect on the model building process.

The results showed that the distance is the crucial parameter that determines the outcrossing and therefore the future work includes analysis of a similar scenario where a denser net of samples is situated around the donor field. The reason for

doing this is to get an insight of what is happening with the outcrossing in the nearest distance of the donor.

Space-Time Correlation Analysis: A Comparative Study ST

Sandra De Iaco

s.deiaco@economia.unile.it
Faculty of Economics, University of Salento, Lecce, Italy

Key words: *space-time random field, space-time covariance, characteristic behaviour, product-sum model*

In literature, many space-time covariance models are available and the main features, such as the behaviour near the origin and the asymptotic behaviour, have been studied for several classes by De Iaco.

Estimating and modeling the correlation of a space-time process is a relevant issue, since only if the correlation model is appropriate for the variable under study, one can rely on further kriging results.

Since covariance models have different features, a comparative study among some of them is useful to underline the importance of choosing a suitable model by taking into account the characteristic behaviours of the models. In this paper, some applications are performed on observed and simulated data sets, in order to prove the efficiency and flexibility in fitting the product-sum model compared with other classes of covariance functions. The relevance of choosing a suitable model by taking into account the characteristic behaviours of the models is proved by using a space-time data set of daily hazardous polluntant averages, and the flexibility of the product-sum model is also highlighted through simulated data sets. The applications are divided in two parts: the first part is aimed at fitting different space-time models to an empirical variogram surface and evaluating which one worksbetter; the second part is intended to prove the flexibility of two significant classes of covariances, such as product-sum models and Gneiting models by using simulated space-time data.

The Effect of Oil Price Volatility on the Istanbul Stock Exchange E2

Ebru Demirci[1], Şebnem Er[2], Burak Ata[3]

1 edemirci@istanbul.edu.tr
School of Transportation and Logistics, Istanbul University, Istanbul, Turkey
2 sebnemer@istanbul.edu.tr
Department of Quantitative Methods, Faculty of Business Administration, Istanbul University, Istanbul, Turkey
3 ataburak@istanbul.edu.tr
Department of Finance, Faculty of Business Administration, Istanbul University, Istanbul, Turkey

Researchers have been examining the role and impact of oil prices on financial markets and stock prices of corporations at the Istanbul Stock Exchange (ISE). Oil price volatility constitutes a systematic influence over many economic indicators especially in many countries such as Turkey that are highly depended on oil price volatility. In this paper, our core aim is to examine how asset prices of corporations in different industries of ISE are influenced by the oil price volatility. We expected to find that oil price changes have a negative or positive significant effect on asset prices of corporations operating in different industries. Regression and panel data analysis methods are applied on monthly data from the period 2002-2006 in order to analyze the effects.

Valued Two-Mode Blockmodeling for Input-Output Analysis NA

Michaela Denk, Michael Weber

michaela.denk@ec3.at; michael.weber@ec3.at
EC3 – E-Commerce Competence Center, Vienna, Austria

Blockmodeling aims at gaining insight into the complex interweavements of relational data (networks or graphs) by simultaneously clustering network nodes (actors) and partitioning network edges (relations). Current developments further the analysis of two-mode instead of one-mode data, i.e., the analysis of relations between disjoint sets of nodes or nodes with differing roles. Thereby, the nodes (or their roles) are clustered and blocks are generated that conform to a specific type of equivalence with respect to their relational structure. Originally, blockmodeling was developed for applications in sociometry and psychometry, primarily making use of binary data. However, a wide variety of applications can be found for valued relations, especially in economic contexts. Recently proposed valued approaches indicate the potentials of blockmodeling for econometrics,

especially input-output analysis. Even though the methods are in their infancy, it has become apparent that blockmodeling provides an interesting way to generate information to support the coordination of relations between economic units, which might eventually benefit application fields of input-output analytic methods, such as supply chain management. This contribution shows the application of two-mode impact blockmodeling, a method introduced at the GfKl 2007, that enables direct two-mode analysis of valued relational data, to input-output analytic questions. Based on the relative values (weights) of the relations, economic units are clustered and their relations partitioned, supporting both exploratory and confirmatory structural analysis.

Application of Mixture Modelling to Personality Disorder Criteria in a General Population Sample

A1

Sharon Devine, Brendan Bunting, Siobhan McCann

devine-s@ulster.ac.uk; bp.bunting@ulster.ac.uk; sm.mccann@ulster.ac.uk
School of Psychology, University of Ulster, Magee campus, Londonderry, Northern Ireland, UK

Diagnosis of personality disorders is increasingly recognised as problematic, including high co-morbidity rates within the personality disorder category and with other clinical mental disorders. This has led to a trend in research towards dimensional models for diagnosis as alternative to the psychiatric categorical models. Advances in statistical methodology and software allow for combination of dimensional and categorical models.

The structure and classification of personality disorders is evaluated and compared with (a) the summed index approach used in DSM, (b) item response theory ranking of severity, and (c) mixture model (factor analysis and latent class analysis).

Information from over 8,000 participants from the British Psychiatric Morbidity Survey, 2000 (Singleton et al., 2001) was obtained from the Data Archives, Essex, England. Presence and absence of personality disorder criteria were assessed using the Structured Clinical Interview for DSM Axis II disorders (SCID-II; First et al., 1992) self report screening questionnaire.

Inconsistencies between different results are examined and discussed in the context of assessment practice.

Redundancy Measures for Multivariate Data

D1

Thierry Dhorne

thierry.dhorne@univ-ubs.fr
SABRES Laboratory, University of South Brittany, Vannes, Brittany, France

There are a lot of statistical problems connected with the redundancy among multivariate data. One of those is the limiting case which corresponds to collinearity and which is at the origin of many statistical developments, for example in the field of model selection or shrinkage estimators. On the other hand, redundancy may be interesting (or at least exploited) in data compression or variables selection.

There is therefore a need for measures of redundancy consistent with the former objectives. Some new proposals have recently been made in order to fill this gap (Kovács, Petres & Tóth, 2005; Shimansky, 2000).

In order to appreciate the redundancy among multivariate data, we propose an axiomatic approach based on some natural requirements. This approach turns to an algebraical formulation and can then be expressed in a linear or more precisely metrical way.

We prove the existence of some measures consistent with the former axiomatic and show that it cannot be reduced to any function of the data covariance spectrum. We propose some approximations only depending on the covariance spectrum and examine the link with the above mentioned literature (Kovács, Petres & Tóth, 2005; Shimansky, 2000).

Finally some comparisons are made on two data sets, the first one with real data and the other one with simulated data in order to appreciate the link between measures and actual redundancy.

References

1. Kovács, P., Petres, T., Tóth, L. (2005). A new measure of multicollinearity in linear regression models. *International Statistical Review*, *73* (3), 405-412.
2. Shimansky, Y. (2000). Continuous measure of significant linear dimensionality of a waveform set. *Computational Statistics & Data Analysis*, *35* (1), 1-10.

Response Surface Analyses: An Application to the Optimization of Astaxanthin Production by Thraustochytrium CHN-1

DE

Virgilio D. Espina[1], Marvelisa L. Carmona[2], Yukiho Yamaoka[3], Takeshi Naganuma[3]

1 heliumx2003@yahoo.com
Department of Mathematics, College of Science and Mathematics, MSU-Iligan Institute of Technology, Iligan City, Philippines
2 maruveris@hotmail.com
National Institute of Advanced Industrial Science and Technology, Kure, Japan
School of Biosphere Sciences, Hiroshima University, Kagamiyama, Higashihiroshima, Japan
3 yamaoka-yu@aist.go.jp; takn@hiroshima-u.ac.jp
School of Biosphere Sciences, Hiroshima University, Kagamiyama, Higashihiroshima, Japan

A class of statistically designed experiments, Response Surface Methodology is introduced and compared to the conventional One-Factor-at-a-Time (OFAT) method in search for the optimum production of astaxanthin by the marine protist Thraustochytrium CHN-1 under varied conditions (15-23 ^{o}C, 7.5-10% glucose, and pH 5-9). Response surface analyses showed that the optimum response was achieved at 17.9 ^{o}C, pH 5.24 and 8.65% Glucose concentration with an estimated maximum response of 0.9199 (mg/L).

Fitting Mixtures of Poisson Regression Models

T2

Susana Faria[1], Gilda Soromenho[2]

1 sfaria@mct.uminho.pt
Department of Mathematics for Science and Technology, Research Centre Officina Mathematica, University of Minho, Guimarães, Portugal
2 soromenhop@sapo.pt
Faculty of Psychology and Sciences of Education, Research LEAP, University of Lisbon, Portugal

Key words: *Poisson mixture regression models, maximum likelihood estimation, EM algorithm , simulation*

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena (McLachlan & Peel, 2000; Frühwirth-Schnatter, 2006). In such finite mixture models, it is assumed that a sample of observations arises from a specified number of the underlying populations of unknown proportions.

Within the family of mixture of distributions models, the class of mixture of regression models has also been studied fairly extensively. In this work, we study Poisson mixture regression models, which are commonly used to analyse heterogeneous count data (Wang et al., 1996; Wedel et al. 1993).

When fitting Poisson mixture regression models, the observed counts are supposed to come from two or more latent subpopulations, and parameter estimation is typically achieved via the EM algorithm to ensure convergence (McLachlan & Peel, 2000). In this study, we develop a procedure for fitting these models using a classification EM algorithm (Celeux & Govaert, 1992) and compare it to the EM approach. The comparison of the two procedures is done through a simulation study of the performance (computational effort and goodness of fit) of these approaches on simulated data sets in a target number of iterations. Simulations show that the choice of the approach depends on the size of the sample. The comparison of the two procedures is also illustrated by analyzing a real data set.

References

1. Celeux, G., Govaert, G. (1992). A classification EM algorithm and two stochastic versions. *Computational Statistics & Data Analysis*, *14*, 315-332.
2. Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Heidelberg: Springer.
3. McLachlan, G.J., Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.
4. Wang, P.M., Puterman, M.L., Cockburn, I., Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, *52*, 381-400.
5. Wedel, M., Desarbo, W.S., Bult, J.R., Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, *8*, 397-411.

A Probabilistic Fertility Model for First Conception

B3

M. Z. Farooqui

nuzza1959@yahoo.com
Department of Statistics, Maharshi Dayananad College, Parel, Mumbai, India

Key words: *fertility, fecunability, amenorrhea, probabilistic models*

An important factor in the determination of the fertility behavior of women is the interval between marriage and the first conception that leads to a live birth. At the

time of marriage, a woman is considered susceptible to conception and the time elapsed before a conception is a random variable determined by fecundability, which is defined as the monthly chance of conception. The analysis of waiting time of first conception shows couple's fertility at early stages of married life. This variable is widely used to study fertility characteristic of a woman, since it is independent of effect of amenorrhea period and since in general, women do not like to use contraceptives to postpone the first birth. There is little chance of recall lapse in reporting the time of first birth from the date of marriage.

As the first conception since marriage can be considered a random phenomenon, probabilistic models can be developed by treating time as discrete or as continuous variable. Gini (1924) derived the geometric distribution for the waiting time of first conception, treating the time as a discrete random variable. He defined the term "fecundability" as the monthly chance of conception for woman living in the married, fecund and exposed state.

Treating the time elapsed from the marriage or from the beginning of the reproductive process to first conception as continuous makes mathematical treatment more convenient. Singh (1964), Henry (1953) and Vincent (1961) developed models treating the waiting time of the first conception as continuous. The negative exponential distribution plays the role of geometric distribution for studying the waiting time of conceive after marriage. Thus, if T denotes the time of first conception then its density function, $f(t)$, is given by

$$f(t) = \delta\, e^{\delta t}, \qquad t > 0, \delta > 0,$$

where δ is instantaneous fecundability. A number of authors made modifications of the above simple distribution to study realistic situations.

In the present model of waiting time of first conception, the time elapsed is defined over the range $(0, \infty)$. But in practical problems, the upper limit may be considered as finite since a woman can conceive up to an age limit. So, there is a need for introducing a new continuous model with finite range. Keeping this is view, an attempt has been made to characterize an existing model derived by Mukeherjee & Islam (1983), defined over a finite range for the purpose of life testing analysis yet suitable for real-life situations.

A Finite Range Continuous Model

The continuous model introduced by Mukherjee & Islam (1983) is considered for the purpose of studying waiting time:

$$F(t,\theta,p) = (p / \theta^p)\, t^{p-1}, \qquad p\ , \theta > 0,\ t \geq 0.$$

The above model is monotonically decreasing and highly skewed to the right. The graph is J-shaped, thereby showing the unimodal feature. The same model has been modified in the sense of range of parameter p for the specific use in the study as

$$f(t,\theta,p) = (p / \theta^p)\, t^{p-1}, \qquad \theta > 0,\ 0 \leq t \leq \theta,\ 0 \leq p \leq 1,$$

where p is instantaneous fecundability and θ is considered as the age limit beyond which a married woman cannot conceive. We define p over the range [0, 1]. The distribution function of the above model is

$$F(t) = \mathrm{P}\ (T \leq \mathrm{t}) = \int_0^t f(t)dt\ = \int_0^t \left(\frac{p}{\Theta^p}\right) t^{p-1} dt\ = \left(\frac{p}{\Theta^p}\right) \int_0^t t^{p-1} dt\ = \frac{p}{\Theta^p}\left[\frac{t^p}{p}\right]_0^t = [t / \theta]^p.$$

The survival function at time t is given by

$$S(t) = P(\mathrm{T} > t) = 1 - P(\mathrm{T} \leq \mathrm{t}) = 1 - F(t) = 1 - [t / \theta]^p.$$

The conception rate function at time t is then given by

$$w(t) = \frac{P(t = t)}{P(t \geq t)} = \frac{(p / \Theta^p) t^{p-1}}{1 - (t / \Theta)^p} = \frac{p t^{p-1}}{\Theta^p - t^p}.$$

Characterization of the Model

The main characteristics of the proposed model are as follows:

- Mean:

$$\mathrm{E}(t) = \int_0^{\Theta} t f(t)dt\ = \left(\frac{p}{\Theta^p}\right) \int_0^{\Theta} t^p dt\ = \frac{p}{\Theta^p}\left[\frac{t^{p+1}}{p+1}\right]_0^{\Theta} = \frac{p}{p+1}\Theta.$$

- Variance:

In order to calculate Var(t), we first calculate E(t^2) as

$$\mathrm{E}(t^2) = \int_0^{\Theta} t^2 f(t)dt\ = \int_0^{\Theta} t^2 \left(\frac{p}{\Theta^p}\right) t^{p-1} dt\ = \ldots = \frac{p}{p+2}\Theta^2.$$

Putting that and the mean into the expression $\mathrm{Var}(t) = \mathrm{E}(t^2) - [\mathrm{E}(t)]^2$, we get

$$\mathrm{Var}(t) = \frac{p}{p+2}\Theta^2 - \left(\frac{p}{p+1}\Theta\right)^2 = \frac{p}{(p+1)^2(p+2)}\Theta^2 .$$

- Maximum-likelihood estimate:

The likelihood function for the model is given by

$$L(t) = p^n \Theta^{-np} \prod_{i=1}^{n} t_i^{p-1} .$$

Hence, $\log L(t) = n \log p - np \log \theta + (p-1) \sum_{i=1..n} \log t_i$. Differentiating the above equation partially with respect to p and equating it to zero, we get

$$\frac{\partial \log L(t)}{\partial p} = \frac{n}{p} - n \log \Theta + \sum_{i=1..n} \log t_i .$$

The MLE of p is then obtained as

$$p = \frac{n}{n \log \Theta - \sum_{i=1..n} \log t_i} .$$

However, differentiating the log-likelihood partially with respect to θ and equating it to zero to obtain the MLE of θ, we get

$$\frac{\partial \log L(t)}{\partial \Theta} = \frac{np}{\Theta} = 0.$$

Hence, for obtaining the MLE of θ, the traditional method is not applicable. The MLE is obtained through order statistic technique. Since the upper limit of the model is θ, it is reasonable to take $t_{(n)}$ i.e. maximum t_i as the MLE for the parameter θ:

$$\theta = t_{(n)} = \max (t_1, t_2, \ldots, t_n).$$

- Median

By the definition of median, we have

$$\int_0^{Me}\left(\frac{p}{\Theta^p}\right)t^{p-1}dt=\frac{1}{2}=\frac{p}{\Theta^p}\left[\frac{t^p}{p}\right]_0^{Me}=\frac{p}{\Theta^p}\frac{Me^p}{p}\Rightarrow Me^p=\frac{\Theta^p}{2}\Rightarrow Me=\frac{\Theta}{2^{1/p}}.$$

- The r^{th} moment about the origin:

$$\mu_r^t=\int_0^{\Theta}t^r\left(\frac{p}{\Theta^p}\right)t^{p-1}dt=\left(\frac{p}{\Theta^p}\right)\int_0^{\Theta}t^{p+r-1}dt=\frac{p}{p+r}\Theta^r.$$

References

1. Gini, C. (1924). Premiers recherches sur la fécondabilité de la femme. *Proceedings of International Mathematic Congress*, *2*, 889-992.
2. Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley.
3. Mukherjee, S.P., Islam, A., (1983). A finite range distribution of failure times. *Naval Research Logistics Quarterly*, *30*, 487-491.
4. Singh, S.N. (1964). On the Time of First Birth. *Sankhya*, *26* (B), 95-102.
5. Vincent, P. (1961). Recherches sur la fécondité biologique. Etude d'un groupe de familles nombreuses. *Population (French Edition)*, *16* (1), 105-112.

Evaluation of First and Second Order Markov Chains Sensitivity and Specificity and Their Relation with Characteristics of Virus Double-Stranded DNA Genome

B3

Jalal Farzami, Ebrahim Hajizadeh

jfarzami@gmail.com; hajitm@yahoo.com
Departmen of Biostatistics, Tarbiat Modares University, Tehran, Iran

Growing amount of information on biological sequences has made application of statistical approaches necessary for modeling and estimation of their functions. In this paper, sensitivity and specificity of the first and second order Markov chains for prediction of genes and their relation with genome characteristics are evaluated using complete double-stranded DNA genome of viruses. In order to compare the first and second order Markov chains results, we developed four algorithms for each Markov model with some differences in prediction of Markov chain parameters. We compared eight algorithms regarding sensitivity and specificity by repeated measure analysis of variance with 3 factors (Markov model, type of selection and estimation of transition probabilities). Results

revealed that the second order Markov chain had significantly better sensitivity and specificity than the first order Markov chain (p<0.001).

For evaluation of effects of genome characteristics on the criteria of quality of algorithms, we added some covariates in the repeated measures model. Adding the covariates – the number of annotated genes per length of genome as well as the A&T and C&G contents of genomes – showed an insignificant difference between the sensitivities of the two Markov models. It was also established that gene base-pairs per genome length and A&T contents of genome as model covariates result in significant differences between the specificities of the Markov models.

Liability Dollarization, Exchange Market Pressure and Fear of Floating: Empirical Evidence from Turkey

E3

Mete Feridun

m.feridun@lboro.ac.uk
Department of Economics, Loughborough University, Loughborough, United Kingdom

The objective of the paper is to examine the relationship between liability dollarization and the exchange market pressure in Turkey within an autoregressive distributed lag (ARDL) and Granger causality framework using monthly data from 1991:12 to 2006:08. The findings suggest that there exists a long-term equilibrium relationship between exchange market pressure and liability dollarization, where liability dollarization Granger causes exchange market pressure both in short-run and long-run, with no evidence of reverse causality. This suggests that the predominance of foreign currency liabilities in the banks' balance sheets in Turkey induces a selling pressure in the exchange market as well as a fear of floating.

Statistical Analysis of Earthquake Data with Extreme Value Distributions Based on Markov Renewal Process

A1

Esin Firuzan, Umay Uzunoğlu Koçer

esin.firuzan@deu.edu.tr; umay.uzunoglu@deu.edu.tr
Department of Statistics, Faculty of Arts and Sciences, Dokuz Eylul University, Izmir, Turkey

The paper presents the comparative discussion between two advance distributions and Markov renewal process in providing accurate and reliable earthquake estimates for two cities, which have the highest probability of earthquake occurrence in the Aegean Region in Turkey. Earthquakes of magnitudes between 4.0 and 7.9 had occurred often both in Izmir (29:09E-38:25N) and in Mugla (38:22E-37:12N) between January 1900 and December 2006. In this study, earthquake occurrences are assumed to be a Markov renewal process since the sequence of earthquakes is a Markov chain and the waiting time distributions depend only on the types of the last and the next earthquakes. The state-space is supposed to be finite and the related Markov chain to be stationary. Weibull distributions are proposed for the waiting times between transitions and their parameters are estimated jointly with the transition probabilities through maximum likelihood (MLE) and least squares estimation (LSE) methods.

Fitting a Weibull Markov renewal process, the exceedance probabilities of waiting times between transitions are obtained. In addition, recurrence times are derived from conditional probabilities. Historical earthquake data is also analyzed by means of Gumbel distribution. There are two reasons for choosing Gumbel distribution: it is an extreme value distribution like Weibull (Extreme Value Distribution Family Type III) and it also has annual maximum magnitude random variable. Hence, recurrence periods are obtained for both magnitudes and waiting times between transitions. Based on Izmir and Mugla earthquake data, using the Gumbel distribution, it is estimated that compared to Izmir, Mugla has a greater probability of earthquake occurrence of magnitude $M \geq 6.0$ in fifteen years. However, using the Weibull distribution, at the present time when one year has elapsed after the earthquake occurrence of $4.0 \leq M \leq 4.9$, the probability of earthquake of $6.0 \leq M \leq 7.9$ to occur in two years is 0.632 for Mugla, while this probability is 0.742 for Izmir. The results of the paper are supported by the corresponding Markov renewal process.

Education and Second Birth Rates in Denmark 1981-1994 B2

Mette Gerster, Niels Keiding, Lisbeth B. Knudsen, Katrine Strandberg-Larsen

meha@biostat.ku.dk
Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

A high educational attainment is shown to have a positive effect on second birth rates for Danish one-child mothers during the period 1981-1994. We examine whether a time-squeeze is a possible explanation: due to the longer enrolment in the educational system, highly educated women have less time at their disposal in order to get the desired number of children. Also, we examine to what extent the partner's education can explain some of the positive effect. We find no evidence that the positive effect of education is due to either time-squeeze or partner effect.

Diagnosing Equilibrium Models from Maps Constructed from Logratios of a Data Matrix D1

Michael Greenacre

michael@upf.es
Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

A competitor to correspondence analysis for analysing tables of positive data is weighted logratio analysis, also known in the biomedical literature as the spectral map (Lewi, 1976). The two approaches have a lot in common: first, they are both calculated using the singular-value decomposition of a weighted double-centred matrix; second, when variance in the data is low, the two methods give very similar results (Greenacre and Lewi, 2005); and third, it can be shown that CA of the data submitted to a power transformation converges to logratio analysis as the power tends to zero.

Logratio analysis has an edge over correspondence analysis, however, when it comes to diagnosing models in the data or in subsets of the data. Points lying on straight lines in the solution space identify subsets of the data which obey certain equilibrium models, as we shall demonstrate in the case of an example from population genetics and from linguistics. In correspondence analysis these equilibrium models appear as curves (also known as the horseshoe, or arch, effect) and these are effectively straightened out by the logratio transformation. The disadvantage of logratio analysis is that it can not handle data zeros, which correspondence analysis copes with easily – this is the reason why

correspondence analysis is so popular in ecology and archeology, for example, where data tables are often sparse with lots of zeros

References

1. Greenacre, M.J., Lewi, P.J. (2005). Distributional equivalence and subcompositional coherence in the analysis of contingency tables, ratio scale measurements and compositional data. *Department of Economics and Business, Universitat Pompeu Fabra, Economics Working Papers* 908. http://www.econ.upf.edu/docs/papers/downloads/908.pdf
2. Lewi, P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneimittel Forschung / Drug Research*, *26*, 1295–1300.

Modelling Moving Feasts Determined by the Islamic Calendar: Application to Macroeconomic Tunisian Time Series

E1

Michel Grun-Rehomme, Amani Ben Rejeb

grun@u-paris2.fr; amani.ben-rejeb@u-paris2.fr
ERMES, CNRS-UMR7181, Université Panthéon-Assas Paris II, Paris, France

National and religious events usually influence the economic activity. Production and consumption show many fluctuations around these feasts. We notice for example, that the retail sales increase in the western countries before Christmas, stop in the vacation-period and fall down just after the holiday.

Islamic events, subject of our paper, are determined by the Hegirian calendar, which is based on the lunar cycles. Some Islamic countries use officially the Gregorian calendar. In such countries the Islamic feasts are moving over time. The lunar calendar is shorter than the Gregorian calendar, which is based on the cycles of the Earth revolution. Consequently, every year the dates of religious events change in the official calendar. Tunisia is a good example of this phenomenon. Twelve relevant time series are analysed in this paper. Our work consists of modelling the whole holidays' effects in order to have a better seasonal decomposition. There are five religious events in Tunisia and the other Muslim countries; The holy month of Ramadan, the feast of the end of Ramadan, the feast of sacrifice, the birthday of the prophet Mohammed and finally the Islamic New Year, that is the first day of the lunar New Year. Our purpose is to measure the impact of these moving feasts.

Removing the effect of the Islamic feasts from time series is very important for both forecasting and comparison purposes. An erroneous seasonal adjustment distorts the forecasting results and influences the decision policy related to costs, employment, production, consumption, import, export etc. It also complicates the comparison between data from countries which do not have the same feasts and events.

For the seasonal adjustment procedure we use the X-12-ARIMA developed by the U.S Bureau of Census. We adopt an approach initially used by Bell and Hillmer (1983) to analyse Easter effect. This method consists in introducing the moving holidays' regressors in the initial regARIMA model. Since the effect is not the same, we consider three regressors to measure the effect before, during and after the feast. The value of the regressor in a given month is the proportion of this interval that falls in the month. The feast is supposed to have the same effect for each day of the interval over which the regressor is nonzero. The length and the number of the intervals are determined by a model-selection procedure. Two methods can be used: the AICC criterion and the of out-of-sample forecast performance. Graphics and various properties of the seasonal adjustment can be used for diagnostic checking.

The empirical results confirm this approach for all the macroeconomic time series considered in the paper, except for the exports which are not affected by the religious feasts, and the broad money due to its composition. Time series for which the holiday regressors were accepted show an improved seasonal decomposition. Hence, the effect of moving holidays can be controlled by adding the appropriate regressors.

Pretesting Questionnaires with Expert (Re)appraisal: Comparison of Two Appraisal Schemes

SM

Valentina Hlebec, Gašper Koren

valentina.hlebec@guest.arnes.si, gasper.koren@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

Expert (re)appraisal is one of several cognitive methods used for questionnaire pretesting. Usually a panel of experts (at least three experts) evaluates the quality of a draft questionnaire from several perspectives. Several different appraisal schemes are used to detect various potential problems with survey questions, such as *Designing Questionnaires Checklist* (Statistics Sweden, 1992), *Questionnaire Appraisal Coding System* (Lessler and Forsysth, 1996), *An eclectic classification*

of measurement error risks to assess questionnaires (Akkerborn and Dehue, 1997) or *Operationalization of science concepts by intuition* (Saris and Galhoffer, 2001). Each of these appraisal systems has its own focus and logic behind it and they tend to detect different potential problems in questionnaire design.

We will present application of two appraisal schemes in several expert panels – *Designing Questionnaires Checklist* and *Questionnaire Appraisal Coding System* – to evaluate two different questionnaires. The first one was selected for its practical orientation and the second one for its conceptual background rooted in the Tourangeau and Rasinski model of question-and-answer process. Expert appraisals of these panels are systematically compared in meta-analysis to show the usability of these appraisal schemas and show the strength of expert evaluation in the process of questionnaire testing.

Risk Scores for Undiagnosed Diabetic Retinopathy Screening Subjects | B3

Sayed Mohsen Hosseini[1], Mohammad Reza Maracy[2], Massoud Amini[3]

1 hosseini@hlth.mui.ac.ir
Department of Biostatistics and Epidemiology, Faculty of Health, Isfahan University of Medical Sciences, Isfahan, Iran
2 maracy@med.mui.ac.ir
Department of Community Medicine, Medical Faculty, Isfahan University of Medical Sciences, Isfahan, Iran
3 m_amini@med.mui.ac.ir
Isfahan Endocrinology and Metabolism Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

Key words: *IEMRC, ROC analysis, diabetic retinopathy, risk score*

The objective of the study was to develop and test the validity of a diabetes retinopathy risk score in the diabetic population. We wanted to evaluate the usefullness of the Simplified Iranian Diabetic Risk Score for identifying undiagnosed diabetic retinopathy subjects.

A total of 12645 patients diagnosed with type 2 diabetes were recruited to Isfahan Endocrinology and Metabolism Research Center (IEMRC) in the studied period. Among those, 3735 patients with complete data were included for the purposes of the present study. The following risk factors were analyzed: sex, age, duration of diabetes, BMI, levels of HbA1C, FBS, cholesterol, triglyceride and the presence or absence of high blood pressure. Multiple logistic regression analysis was used to detect diabetic retinopathy as the dependent variable and a score for each

significant risk factor was based on the regression coefficient. ROC curves were constructed to identify to optimum cut-off value of the predicted probability (>60%) of diabetic patients for determining diabetic retinopathy.

In the sample, 36% patients were male and 64% were female; 54% of patients were diagnosed as having retinopathy. According to the model, being female, having lower BMI, older age, longer duration of diabetes and an increased Hba1c was associated with increased risk of diabetic retinopathy. The AUC for the ROC was 0.699 (95% CI 0.680-0.718). For the IEMRC clinical patients, the regression score threshold of 55 had the optimum sensitivity (62%) and specificity (63%) for determining diabetic retinopathy.

The Simplified Iranian Diabetic Retinopathy Risk Score has been design to be a screening tool for identifying high risk subject in the population and for increasing awareness of the modifiable risk factors and healthy lifestyle. We believe that the public health benefit of the risk score is considerable. It is a cost-effective and practical way to identify individuals at high risk for diabetic retinopathy in the diabetic population.

Asymptotic Behaviour of the Friedman Test Statistic **L1**

Bernd Jäger, Karl-Ernst Biebler

biebler@biometrie.uni-greifswald.de; bjaeger@biometrie.uni-greifswald.de
Institut für Biometrie und Medizinische Informatik, Ernst-Moritz-Arndt-Universität, Greifswald, Germany

The Friedman test is a multiple rank test and is used for correlated samples. The distribution of the test statistic is obtained from combinatorial considerations. For larger sample sizes, calculation problems arise. One therefore prefers the asymptotic chi-square distribution in practical application. We carried out computer simulations and compare exact distribution, asymptotic distribution and empirical distribution of the Friedman test statistic.

Spatial Data Mining and Visualization with GoogleEarth

ST

Domen Jesenovec[1], Nada Lavrač[2], Neža Mramor Kosta[3]

1 domen.jesenovec@gmail.com
2 nada.lavrac@ijs.si
Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
3 neza.mramor-kosta@fri.uni-lj.si
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Spatial data mining is a new area of research concerned with identifying interesting spatial patterns from data stored in spatial databases and geographic information systems (GIS). This area involves specialized statistical and data mining techniques, adapted for use of GIS and spatial relationships between objects. The main aim of these techniques is to identify spatial patterns through the identification of spatial objects as potential pattern generators, to identify information relevant for explaining the spatial pattern, and to present it in a way that is intuitive to the analyst and supports further analysis. The main contribution of this paper is a methodology that enables effective spatial pattern explanation and visualization using GoogleEarth. The main advantages of the proposed methodology are that the software can be used free of charge, that it includes detailed satellite images of the whole earth surface, and that it provides the possibility of various visualization techniques using HTML descriptions. The usefulness of the proposed methodology is demonstrated in the analysis of a relational database of Slovene traffic accidents, including accident and casualty data for ten consecutive years. The spatial patterns, discovered by time series clustering have been visualized using a simple regional map of Slovenia, while the new GoogleEarth visualization method was used for visualizing patters discovered by subgroup mining. Our method enables also the explanations through a customized user interface, explaining traffic accident groupings formed of spatially similar (geographically close) traffic accident subgroups.

Time Series Discrimination in State Space Form E1

Thomas Kalantzis, Demetrios Papanastassiou

tkalant@uom.gr; papanast@uom.gr
University of Macedonia, Thessaloniki, Greece

Key words: *discrimination, time series, parametric models, time domain*

In this paper, we discuss the discrimination of time series data generated by a parametric model casted in a general linear Gaussian state space formulation. A series is assigned into one of two possible groups according to a likelihood based criterion. We derive an asymptotic approximation to the misclassification probabilities based on a time domain approach. The state space formulation entails as a special case the well known ARMA family. We compare our findings with numerical results from relevant works on specific simple ARMA models.

Determining Hidden Markov Models Efficacy in a Gene Finding Problem B3

Anoushirvan Kazemnejad, Ebrahim Hajizadeh, K. Mirjafari

aklili@yahoo.com; hajitm@yahoo.com; kmirjafari@gmail.com
Tarbiat Modares University, Tehran, Iran

In all fields of scientific researches one can see how statistical models are important and useful for modeling various phenomena. One class of these models are Hidden Markov Models (HMM), which are extensively used to analyze several problems, such as speech recognition and finding genes which are implicated in causing cancer. Nowadays most of the methods for prediction of genes are based on these models.

In this paper, we show how weighting affects Generalized HMM efficacy in gene finding problems in eukaryotes. For this aim, we chose a simple structure of HMM and entered Open Reading Frames (ORF) information with a weight matrix into the model, whereby the appropriate values of weights were estimated based on sensitivity and specificity in samples of human DNA sequence.

Results showed that entering weights into the model has significant effect on efficacy and is comparable to the existing gene finding methods. It even seems that the presented method is superior because it has fewer parameters and is simpler to analyze.

Sample Size in Multiple Regression: 20 + 5*k* T3

Harry Khamis[1], Mike Kepler[2]

1 harry.khamis@wright.edu
Statistical Consulting Center, Wright State University, Dayton, Ohio, USA
2 helpdesk@wright.edu
Computing and Telecommunications Services, Wright State University, Dayton, Ohio, USA

There are many instances when a statistical power analysis is not reliable due to lack of information or because it doesn't suit the research goals. In these cases, other performance characteristics of the regression should be considered when trying to determine an appropriate sample size. Many informal rules of thumb have been provided as recommendations for the minimum sample size to be used in a multiple regression. The goal of these simple sample size formulas is to avoid overparameterization and insure generalizability rather than to accommodate a given statistical power. They are useful when conducting pilot studies, or working with new variables and/or new populations, or when reliability and predictive discrimination is of primary interest. However, these recommendations are largely unsubstantiated, somewhat arbitrary, and vary quite widely from one another. In this paper, reliability is used as the criterion for developing a formula for minimum sample size in the multiple regression model with continuous predictors. The ultimate formula, $n = 20 + 5k$, where k = number of predictors, is simple and optimal with respect to a principle based on the rate of change of the reliability criterion relative to n.

Nonlinear Time Series Modelling of Lahore's Precipitation E3

Muhammad Saleem Khan, Muhammad Jawed Iqbal

jiqbalku@yahoo.com
Institute of Space and Planetary Astrophysics, University of Karachi, Karachi, Pakistan

The highly nonlinear relationships governing climate phenomena suggest an application of the nonlinear time series analysis. We examine the presence of nonlinearity in the climate system using surrogate data. This paper aims to investigate the existence of chaos in the time series of Lahore precipitation. The predictability of daily rainfall is the most difficult task because of the nonlinear behavior of climate system. This paper employs radial basis function and neural network for constructing the nonlinear model of summer monsoon precipitation of

Lahore. We also compare both algorithms with respect to prediction accuracy and storage requirements.

Using Behavioral Statistical Scorecards in Portfolio Management and Business Planning

D2

Goran Klepac[1], Božidar Kliček[2]

1 goran.klepac@rba.hr; goran@goranklepac.com
Raiffeisen Bank, Zagreb, Croatia
2 bozidar.klicek@foi.hr
Faculty of Organization and Informatics Varaždin, University of Zagreb, Varaždin, Croatia

Key words: *behavioral statistical scorecards, data mining, Bayesian networks*

Behavioral statistical scorecards have increased their popularity with BASELII standards in financial institutions. This article recognizes new potential and fruitful application area of behavioral statistical scorecards for portfolio management and holistic business planning from the perspective of risk measurement. This methodology could not be applied only in financial institutions but also in different types of business like insurance or trade. Risk measurement through behavioral statistical scorecards can be the base for business planning with the perspective of client/customer future behavior considering delays in paying invoices, delays in paying loans (default), fraud, churn and similar risky behavior. Behavioral statistical scorecards method can clearly explain influences on target variable and perform client/buyer profiling. Campaign planning could be focused on potential clients with low risk profile from the perspective of fraud, delays in paying or/and churn, depending on company policy (conservative/liberal). All those parameters could be measured periodically with trend monitoring of risk parameters on whole portfolio or in portfolio segments. Behavioral statistical scorecards method can be useful for the construction of an early warning system that is sensitive to structural changes in portfolio. Another potential application is "what-if" or sensitivity analysis with Bayesian networks using calculated risk measures (scores from statistical scorecard) as an input. This article describes several innovative applications performing problem analysis first, then explaining the underlying theoretical model, and then describing the practical examples of application. Theoretical and implementation issues of these applications are discussed, and future research directions are outlined.

Comparison of Forecasting Performance of Regime Switching versus Linear Models: Application to Turkish Economy

E2

Selcuk Koç, Selin Özdemir Koç

selcukkoc@marmara.edu.tr; sozdemir@marmara.edu.tr
Marmara University, Department of Econometrics, Istanbul, Turkey

Key words: *linearity, TAR, Markov switching, rate of net import to GNP, forecasting performance*

Most of the time series exhibit break(s) in their behavior associated with structural changes in government policy or financial crises, which makes it hard to model these series. When time series have changes in their structure, the linearity assumption can not be used. Linearity is an important assumption in traditional econometrics, but in practice most of the series do not meet it. Such series are called nonlinear series. Hence, it is important to test the linearity assumption. If a series is nonlinear and has changes in the structure, it can be modelled with TAR or Markov Switching Model. Both belong to regime switching models. There are two main groups of them: the first one comprises TAR and its variations (STAR, ESTAR, LSTAR etc.), while the second is based on the Markov switching model.

The nonlinear models which we use for our analysis are:

a) The Threshold Autoregressive Model (TAR), which was first proposed by Tong (1978), is popular due to the fact that it is relatively simple to specify, estimate and interpret, at least in comparison with many other nonlinear time series models. This model allows several distinct regimes, and allows that transition between these regimes can be rapid.

b) The other popular nonlinear time series model, which was first proposed by Teräsvirta (1994), is called Smooth Transition Autoregressive Model (STAR). The assumption that the economy can be only in two states is its basis. This model allows two distinct regimes which represent two different phases of the data, but transition between these regimes can be smooth. The two-regime TAR model is a special case of the STAR model. It has two sub-approaches called LSTAR and ESTAR. The LSTAR model states that the contraction and expansion phases of an economy may have different dynamics, and a transition from one regime to other can be smooth. The ESTAR model implies that the contraction and expansion have rather similar dynamic structures, whereas the middle ground can have different dynamics.

c) The Markov switching models, which were first proposed by Hamilton (1989), are another group of regime switching models. They are a generalisation of the simple dummy variables approach. Movements of the state variable between regimes are governed by a Markov process. If a variable follows a Markov process, all we need to forecast the probability that it will be in a given regime during the next period is the current period's probability and a transition probability matrix.

In this paper, we question whether there is any nonlinearity in rate of net import to GNP of Turkey. We use nonlinearity tests such as McLeod-Li, BDS, and Tsay and Kaplan test. After detecting the nonlinearity, we attempt to construct the best fitting nonlinear model among the competing TAR, STAR and Markov switching models. We make the static and dynamic forecasts for each nonlinear model and linear model, and compare residual variances and root mean squared errors. The model with lower residual variance and root mean square error is considered superior. We first compare forecasting performance between different nonlinear models, and then between linear and the selected best nonlinear model.

Imperfect Information and Credit Rationing in Financial Markets: Application of Long Range Dependence in Credit Series

E2

Selin Özdemir Koç, Selcuk Koç

sozdemir@marmara.edu.tr; selcukkoc@marmara.edu.tr
Marmara University, Department of Econometrics, Istanbul, Turkey

Key words: *credit rationing, imperfect information, long range dependence, ARFIMA, R/S statistic, GPH test, Robinson's semi-parametric test*

Credit rationing is financial behavior to reduce credit supply because of its adverse selection effect and moral hazard effect. Both effects derive directly from imperfect information in credit markets. Credit information is present in loan markets after banks have evaluated loan applications. However, banks do not have enough information to identify different borrowers as having different probabilities of repaying their loans. The bank would like to be able to identify borrowers who are more likely to repay. Interest rate can reflect borrower's repaying behavior. Those who are willing to pay high interest rates may present higher risk due to their probability of repaying the loan being low. Therefore, banks would not lend money to an individual who offers to pay more than ordinary interest rate level. In the bank's judgment, the expected return of a loan at an interest rate above ordinary interest rate level is actually lower than the expected return of the loans the bank is presently making. To sum up, it may not be profitable to raise the interest rate when a bank has excess demand for credit. In this situation, banks deny loans to borrowers who are not well known from past experience. In addition to adverse selection effect, moral hazard effect also leads to credit rationing. Naturally, banks are not able to directly control all the actions of the borrowers. Because of imperfect information about the borrowers, banks may encounter default risks in some credit contracts.

Abstaining from credit risk problem is the main aim for banks. They gather the credit information set from credit markets through bank's past experience. Because of imperfect information explained above, this information set is the basic determinant for credit policy. Evaluating loan applications mostly depends on customer's repayment habit in the past. If the customer has had problems repaying a loan in the past, the new loan application may be rejected by the credit committee. The customers' credit limits are also determined by their history. Hence, various kinds of credit time series may exhibit the long memory (long range dependance) property.

The potential presence of stochastic long memory in economic and financial time series has been an important subject of both theoretical and empirical research.

The long-memory, or long-term dependence, property describes the high-order correlation structure of a series. If a series exhibits long memory, there is persistent temporal dependence between observations widely separated in time. Fractionally integrated processes can give rise to long memory.

The main aim of our paper is to investigate Turkish credit markets in the light of the efficient market hypothesis. Long memory features provide the evidence of invalidity of this hypothesis for credit markets. The presence of long memory in credit series also implies that past credit decision can help making future credit decisions. Thus, borrower's repaying behavior in the past is more important than latest financial information about the borrower. Long range dependence is investigated using *R*/*S* statistic, GPH parametric, Robinson's semi-parametric method and the frequency-domain method of ARFIMA(*p*,*d*,*q*) models. The fractional differencing parameter is estimated using the spectral analysis method.

Financial Accounts Visualization **A2**

Irena Komprej

irena.komprej@bsi.si
Bank of Slovenia, Ljubljana, Slovenia

In the paper, we present the idea of visualizing Financial Accounts data as a network. Financial Accounts data, as a constitutive part of National Accounts, consist of a set of macroeconomic accounts based on internationally agreed concepts, definitions, classifications and accounting rules. Being the most comprehensive macroeconomic standard, they provide a framework within which economic data can be not only compiled, but also presented in a format that is designed for purposes of economic analysis, decision-taking and policy-making. Being so comprehensive, though, they are not easy to visualize with conventional visualizing techniques.

Financial Accounts are specific in the sense they provide information on financial instruments exchanged between institutional sectors. This makes them ideal to be observed as a network of institutional sectors (vertices) being related with financial instruments (multiple relations). We describe the data model of Financial Accounts and propose a suitable transformation to network. The results of the transformation are data in a format appropriate for the Pajek software for network analyses and vizualizations. In Pajek, we perform interesting vizualizations, some of which are presented in the paper.

The proposed technique can be useful for the compilers as well as for the analysts of Financial Accounts data. The compilers can appreciate added value to their product and use it for quality checks, whereas the analysts can easily observe interesting economic phenomena without cumbersome browsing through large amounts of data.

Optimal Design of Bayesian Reliability Test Plans for a Series System Based on Type-II Censoring

L1

Mahesh Kumar

mahesh@caledonian.edu.om
Department of Mathematics and Science, Caledonian College of Engineering (affiliated to Glasgow Caledonian University, Scotland, UK), Sultanate of Oman

Key words: *exponential lifetime, quasi density, iid prior, squared error loss, type-II censoring, non-linear integer programming, optimization, type-I error, type-II error*

Consider a series system with n independent components, where the i-th component has exponential lifetime with parameter θ_i. Assume θ_i's have *iid* prior with quasi density given by $g(\theta_i) = 1 / \theta^d_i$, $0 < \theta_i < \infty$, $d \geq 0$, i = 1,2,…,n. In this paper, using the data obtained from Type-II censoring, we investigate the optimal Bayesian estimator for the system reliability, based on squared error loss. Further, we design the reliability test plan to test the reliability of a series system, satisfying the usual probability requirements, i.e., Type-I and Type-II error constraints. The non-linear integer programming optimization problem is solved for some special cases. We compare this new plan with existing classical test plans. Furthermore, our plan has minimum testing cost and provides 66% reduction in testing cost.

A New Generalized Useful Relative Information Generating Function T3

Parmil Kumar[1], Bilal Ahmad Bhatt[2]

1 parmil@yahoo.com
Department of Statistics, University of Jammu, Jammu, India
2 bhat_bilal@rediffmail.com
Division of Agricultural Economy and Statistics, Faculty of Agriculture, Sher-e-Kashmir University of Agricultural Sciences and Technology, Jammu, India

Key words: *information generating function, entropy, information measures, useful information*

We define relative information generating function with utilities. We also discuss its particular and limiting cases. It is interesting to note that differentiation of this relative information generating function at t=0 produces various well-known measures of information. The relative information generating function for uniform and exponential probability distributions is also studied.

50-50 MANOVA with Rotation Testing: A Framework for Analysing Designed Experiments with Multiple Responses DE

oyvind.langsrud@matforsk.no
MATFORSK, Norwegian Food Research Institute, Ås, Norway

This talk presents an overview of a unified framework for analysing designed experiments with univariate or multivariate responses.

The single response special case is ordinary general linear modelling. Type II sums of squares are used to handle unbalanced designs. Furthermore, the Type II philosophy is extended to continuous design variables. This means that the method is invariant to scale changes. Centring of design variables is not needed. The Type II approach ensures that common pitfalls are avoided.

Multivariate testing of all response variables is performed by the 50-50 MANOVA method, which is a modified variant of classical MANOVA made to handle several highly correlated responses. Classical MANOVA performs poorly in such cases and it collapses when the number of responses exceeds the number of observations. The 50-50 MANOVA method is suggested as a general method that will handle all types of data. Principal component analysis is an integrated part of the algorithm.

A univariate *F*-test *p*-value of each response variable can be reported when several responses are present. However, with a large number of response variables, these results are questionable since we will expect a lot of type I errors. Therefore the p-values need to be adjusted.

By using rotation testing it is possible to adjust the single response *p*-values according to the familywise error rate criterion in an exact and non-conservative (unlike Bonferroni) way. It is also possible to adjust p-values according to a false discovery rate criterion. Our method is based on rotation testing and allows any kind of dependence among the responses. Note that rotation testing is closely related to permutation testing. One difference is that rotation testing relies on the multinormal assumption. All the classical tests (*t*-test, *F*-test, Hotelling T^2-test, ...) can be viewed as special cases of rotation testing.

The methodology is illustrated by a microarray example. Software is available at www.matforsk.no/ola.

Exploratory Analysis of the ILPNet2 Repository: Research Contents Analysis and Coautorship Network

D2

Nada Lavrač, Miha Grčar, Blaž Fortuna

nada.lavrac@ijs.si; miha.grcar@ijs.si; blaz.fortuna@ijs.si
Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia

The ILPnet2 database is publicly available on the Web and contains information about publications in the area of Relational Data Mining and Inductive Logic Programming, published in the period 1970-2003. This paper presents the results of co-authorship analysis obtained using the Pajek network analysis program. Next, the paper presents the analysis of ILPnet2 publications with OntoGen, a system for data-driven semi-automatic ontology construction. We prepared two different datasets based on the same domain. In the first dataset the instances represent the authors from the ILPnet2 database and in the second the instances represent the publications found in the ILPnet2 database. In the first case, a document is named after the corresponding author and contains his/her entire bibliography, and in the second case a document bears the name of the corresponding publication and contains the title and the abstract of the publication.

The constructed ontologies are presented and analysed. Finally, we show the recent advances in ILPNet2 ontology construction through the use of automatic term extraction, and the visualization of the ILPNet2 co-authorship network on the Document Atlas, using a 2D visualization of the main ILPNet2 research topics.

How Useful is the LAVE Method

E1

Aleša Lotrič Dolinar

alesa.lotric.dolinar@ef.uni-lj.si
Institute of Statistics, Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

Key words: *multivariate LAVE, GARCH, ARIMA, volatility, emerging markets, stock exchange index*

We show that the findings of Mercurio and Spokoiny (2004) concerning their new LAVE (Local Adaptive Volatility Estimate) approach for volatility estimation are not necessarily true for another type of volatile time series, as far as comparison to the usual GARCH(1,1) process is concerned. However, we propose another use of LAVE for the purpose of level, not volatility, modeling – and in our application it turns out to be successful.

Measuring the Value of Different Categories of Knowledge within a Romanian University

M2

Dana Adriana Lupşa-Tătaru

lupsad@unitbv.ro
Faculty of Economic Sciences, Transilvania University of Braşov, Braşov, Romania

Considering the role of knowledge in knowledge-based society, the importance of universities within this context and the goal of EU to become the most developed knowledge-based economy in the world, determining the perceived value of different categories of knowledge by the teaching staff is an important matter for the management. It helps identifying which knowledge is critical and crucial for the development and the competitive advantage of a university.

The paper aims to presents the results of the application of an original model of knowledge evaluation within a Romanian university, considering the opinions of experts compared with the results regarding the perceived value of different types

of knowledge obtained via questionnaire at the same university. The model used for knowledge evaluation, originally based on subtle sets theory, is enriched and made more objective by using factor analysis and cluster analysis.

The study revealed that the results of theoretical knowledge value model application and the perceived value of knowledge coincide, highlighting the major role of one specific category of knowledge, namely knowledge about / generated through the application of quality management within the studied university, with the perspective of extrapolating the results to all Romanian universities.

Do Neural Networks Outperform Classical Logistic Regression and Discriminant Analysis Statistical Classifiers? A Case Study in Selection of Portuguese Air-Force Pilot Candidates

A2

João P. Maroco [1], Rui Bártolo-Ribeiro[2,3]

1 joao.maroco@ispa.pt
Department of Statistics, Instituto Superior de Psicologia Aplicada, Lisboa, Portugal
2 Psychology Centre of the Portuguese Air Force, Lisboa, Portugal
3 Department of Organisational Psychology, Instituto Superior de Psicologia Aplicada, Lisboa, Portugal

We evaluated the classification accuracy of discriminant analysis, logistic regression and four neural network topologies (multi-layer perceptron, radial basis network, probabilistic neural network, and linear neural network) in classifiying approved vs. failed air force's pilot candidates using several psychometric pre-admission tests as predictors. A stepwise (for logistic regression and discriminant analysis) and sensitivity (for neural networks) selection procedure retained as significant predictors the tests that evaluated the abilities to read and interpret a series of airplane instruments, eye-hand-foot coordination and vigilance alert. Performance in an early flight train was also retained as significant predictor of final scoring.

Regarding the accuracy of predictions, logistic regression showed the highest accuracy (77%) with high sensitivity (92%) but low specificity (31%). Discriminant analysis had high sensitivity (77%) and high specificity (64%). However, it had the second lowest accuracy (74%). The best performing neural network was the multi-layer perceptron (5:5-6-1:1), which showed high sensitivity (85%), second highest specificity (47%) and high accuracy (76%). Radial basis networks and probabilistic neural network both failed to predict correctly the failed candidates (0% specificity). The parametric classical statistical classifiers – logistic regression and linear discriminant analysis – were not outperformed by three of the four neural networks evaluated, which included topologies devised specifically for classification tasks.

Comparing Generalized Estimating Equation Model with Standard Logistic Regression Model in Determining Back Pain Associated Factors in Iran

B3

Kazem Mohammad, Nargess Saiepour

mohamadk@tums.ac.ir; saiepour@razi.tums.ac.ir
Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

Our study is based on the information obtained from the second National Health Survey (NHS) conducted in the year 2000 in Iran. Cluster sampling was used, with each cluster covering 8 households. The relation between back pain as the dependent variable and factors such as residential area, gender, skeletal deformation, education, marital status, body mass index (BMI), smoking habits and mental health as independent variables was examined. Using generalized estimation equation (GEE) and standard logistic regression models, the results indicated that in both models, all mentioned factors were associated with back pain. The estimated odds ratios (ORs) by the two models were approximately similar, but overall, logistic regression resulted in slightly smaller standard errors of the estimated ORs.

Developing a Quality Assurance Model for Iranian Higher Education System: An Exploratory Design

DS

Saeed Mohammadzadeh[1], Yousef Hedjazi[2], Abbas Bazargan[3]

1 saeed_rhm@yahoo.com
Department of Agricultural Extension and Education, University of Tehran, Tehran, Iran
2 yhejazi@ut.ac.ir
Department of Agricultural Extension and Education, University of Tehran, Tehran, Iran
3 abbas.bazargan@pedagogy.ir
Faculty of Psychology and Education, University of Tehran, Tehran, Iran

The purpose of this study was to develop a Quality Assurance (QA) model for Iranian higher education system as perceived by faculty members. To accomplish that purpose, the study proceeded in two parts. In part A, we studied the current situation of Iranian higher education QA model through departmental self-assessment, and in part B we investigated QA model's components.

Part A unfolded in two-phases Exploratory Design as a kind of mixed methods research. This design is based on the premise that an exploration is needed for several reasons like new area of inquiry, phenomena that have not been studied previously, unknown variables or lacking of guiding framework or theory. It is also appropriate when we want to generalize results to different groups. Since QA through departmental self-assessment is a new phenomenon in Iranian higher education system that has not been studied previously, so the factors related to its implementation are unknown, we started with a qualitative phase to explore the current self-assessment situation in the departments and then build to a second, quantitative phase.

Phase one was accomplished by asking faculty members to describe their experience of the departmental self-assessment. To ensure that the concepts would be appropriate for better understanding of the self-assessment process in the sense of being candidate for departmental self-assessment or not, six executives of departmental self-assessment, five members of the departmental self-assessment committee and seven faculty members participated in the interview as key informants from self-assessment candidates' departments, and seven heads of departments and eight faculty members participated from departments that were not self-assessment candidates. According to the interview protocol, we designed the questions in such a way as to help the interviewees to think about factors related to implementation of self-assessment and reasons of its acceptance or rejection by departments. The participants took part in the interview individually at their respective departments, except for three people who were interviewed over the telephone. Interviews continued until the point of theoretical saturation

was reached, which indicated diminishing original insights. These interviews generated two types of qualitative data: interviewer field notes and transcripts of the interviews. Using Glaser and Strauss' method of constant comparison, and Miles and Haberman's suggestions for coding qualitative data, we identified and categorized all points that the participants described in the interviews as pertaininng to their experience with self-assessment. After coding and labeling the interviewees' responses, we developed a general category scheme of the participant responses. Then, we began to identify themes by sorting the initial scheme into concrete categories and subcategories. We categorized the responses according to several themes.

Building from findings of the qualitative phase, we developed items to represent the themes identified in part A in a close-end questionnaire. Part B of this research unfolded in a quantitative exploratory design. In this part of the study, we first reviewed the related literature. The literature showed that a "general model" of QA does not universally apply, but that most components of it do apply in most countries. Based on international experiences, we identified eight QA model's components. Every component includes several options. Based on these options, we developed items to represent the components in part B of the close-end questionnaire.

The population included faculty members of public universities, which are the target of the departmental self-assessment, from which 219 faculty members were sampled by using multistage random sampling procedure, and the questionnaire was administrated to them. This survey generated quantitative data. Data analysis for phase two of Part A was performed using exploratory factor analysis and descriptive statistics, and for part B by ranking. For efficiently ranking the options of every component, we used both the average and the coefficient of variation (CV) at the same time. This procedure of ranking could better indicate the participants' preferences. Through the results of analyzing the current situation of QA through departmental self-assessment and participants' preferences on the options of every component, we completed the components of a QA model for Iranian higher education system.

Null Model Analyses of Presence-Absence Data in Ecology: Combining Generalized Linear Models and Monte Carlo Testing for the Detection of Non-Random Patterns

A1

Jorge A. Navarro[1], Brian F. Manly[2]

1 nalberto@uady.mx
Departamento de Ecologia, Universidad Autónoma de Yucatán, Merida, Mexico
2 bmanly@compuserve.com
Western Ecosystems Technology, Inc., Cheyenne, Wyoming, USA

Null model analyses of presence/absence data in ecology (e.g., occurrences of species on particular locations) can be characterized into two broad categories: those where the simulation protocols keep row (species) and column (location) totals fixed in the null matrices, and those where row and/or column totals are allowed to vary. In contrast to the research devoted to the first type of null models, relatively little research has been done to study the properties of the latter.

In this work, we describe a strategy for null model construction by means of generalized linear models for presence-absence data. Assumptions for the generalized linear models (GLMs) are that (1) occurrences are independent of each other; (2) species and island effects are the only explanatory variables for each observation in the matrix; and (3) the relationship between the occurrence of each species-location combination and the species and location effects is non-linear. For model definition, observable presence-absence data are related with unobservable hypothetical distributions of the number of elements (called the "quasiabundance") of each species-location combination; these distributions are interpreted as different scenarios of species occurrences from where the best fitting model is selected among a range of competitor null models. The method produces fitted cell probabilities, which are subsequently used for the detection of non-random patterns in the observed matrices, via parametric bootstrap. As a consequence, the simulation protocol allows both row and column totals to vary from one simulation to the other.

Monte Carlo tests applied to suitable metrics for the observed and simulated matrices are then used to evaluate the adequacy of species and location effects for the prediction of each species-location combination. Properties of the observed data matrices (e.g., sparseness and degeneracy) and constraints in the simulation protocols are also evaluated. Finally, using as statistic the estimated proportion of allocated presences in each cell of randomly generated matrices, it is shown that the set of null matrices in the GLM approach can be different from the set of null matrices obtained with three algorithms keeping row and column totals fixed. It is confirmed also that there may be differences in the null universe of matrices produced by different versions of this latter simulation protocol.

Estimation of Target Parameters for Small Domain

L2

Boro Nikič

boro.nikic@gov.si
Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia

Key words: *small area estimation, GREG estimator, EBLUP estimator*

There is constantly growing demand for the estimated parameters for small domains in the case of social and business sampling surveys. For estimating parameters of total in small domains, design-based estimators such Horvitz-Thompson (HT) and Generalized Regression estimator (GREG) have been used so far. Due to the fact that design-based estimators depend on inclusion probabilities and sufficiently large sample size, they are not appropriate for areas with small sampling size. In that case, it is more appropriate to use the model-depended estimators which rely on superpopulation model. One of the best known model-depended estimators is the Empirical Best Linear Unbiased Predictor (EBLUP).

In our simulation study, we estimated domain total of simple random sampling (SRS) design by using HT, GREG and EBLUP estimators with regard to reducing sampling size. Results of simulations shows us that design-based estimators performs better then the EBLUP estimator if the sample size is large, while the EBLUP estimator outperforms design based estimators with decreasing sample size. In the paper, we describe results of applied methods, problems that we faced and our plans for the future work in the area of small area estimation in order to apply small area models in practice.

Bibliography

1. Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.
2. Rao, J.N.K. (2003). *Small Area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Wiley.
3. Longford, N.T. (2005). *Missing Data and Small Area Estimation*. New York: Springer.
4. Pfeffermann, D. (2002). Small area estimation: new developments and directions. *International Statistical Review*, *70* (1), 125-143.

Heuristic Evaluation of Website Usability: An Application to the Italian Websites of Municipalities/Communes

M2

Matthias Oehler[1], Silvia Biffignandi[2]

1 matthias.oehler@gmail.com
Université Lumière Lyon I, Lyon, France
2 silvia.biffignandi@unibg.it
DMSIA, Università degli studi di Bergamo, Bergamo, Italy

Key words: *websites, usability, heuristic evaluation, commune, data mining, text mining*

Today the web is becoming the most common way to broadcast information. However, in the face of this mass of knowledge, it is *more and more difficult* for the user to extract the really interesting and relevant information. This is why there is a real need to concretely understand what are the structures and directives that webmasters and web designers have to respect. They are needed to create a website of "quality" in such a way that allows an efficient spread of information to be achieved.

Accordingly, a website has to respect some *standards and guidelines*, which are legitimately named "usability". According to Jarrett (2006), the usability task denoted the capacity of a tool (i.e., in our study a website) to be useful, usable and used (Powerful, Simple and Sexy, respectively, according to Nielsen, 2000). In other words, *usability* is a way to evaluate the ease of use of a particular tool by people in order to achieve a precise goal without error in a short lapse of time. But it is important to know that a system can respect all the criteria of usability, yet be useless. It is the *adequacy between the activity and the tool* which will make it possible to say that a tool is useful. Hence, the ideal configuration for a usable tool is to allow convergence between the user's expectations and those the owner's goals.

There are many possibilities to evaluate the usability, but the best method for quick, cheap, and easy evaluation of a user interface design is the so-called heuristics evaluation. "The goal of *heuristic evaluation* is to find the usability problems in the design (structure and content) so that they can be attended to as part of an iterative design process" (Nielsen, 2000). Nevertheless, all existing usability evaluation tools do not allow taking on board the activity of the application to evaluate.

Thus, in this paper, we present how to *set up and evaluate* some heuristics than can measure the usability of municipality/commune websites according to their economic profile. Our hypotheses was that the information search of website users, as well as the objectives for creating a website, depend on the economic profile defined by socio-economic, geographic and technological characteristics of the municipalities/communities (i.e., their profile being touristic, industrial, agricultural etc.). The main problem being that municipality/commune websites generally contain bulky and heterogeneous information; we propose a process that can be useful to evaluate their usability. Therefore, we developed software for data extraction to evaluate (based on the static HTML source code) the structure and the contents of the sampled websites. In addition, some variables were selected for carrying out the classificaion of municipalities/communes. It was then possible to classify the results according to various criteria or define some atypical results that *could be useful for municipality/commune councils to help them orient and develop their websites*.

References

1. Jarrett, C. (2006). *Caroline's Corner: Useful, Usable and Used – Your New Look Council Website*. http://www.usabilitynews.com/news/article3296.asp
2. Nielsen, J. (2000). *Designing Web Usability*. Indianapolis: New Riders Publishing.

The Power of the Non-Normality Corrected Chi-Square Statistics in Structural Equation Modeling

M1

Ulf Henning Olsson, Tron Foss

ulf.h.olsson@bi.no; tron.foss@bi.no
Department of Economics, Norwegian School of Management BI, Sandvika, Norway

Key words: *scaled chi-square statistic, asymptotic covariance matrix, kurtosis*

There are several different chi-square statistics offered in typically SEM software to deal with non-normal data. In this paper we examine the power of two such chi-square statistics, namely the SB statistic (Satorra & Bentler, 1988) and the ADF statistic (Browne, 1984). The SB statistic corrects the normal theory chi-square with a scale factor which is estimated from the sample and involves the estimated asymptotic covariance matrix (ACM). The scale factor is estimated so that the SB statistic has an asymptotically correct mean. The ADF statistic under the assumption of correct model has an asymptotic chi-square distribution.

Following the notation of Jöreskog, Sörbom, Du Toit, & Du Toit (2003), these chi-square statistics are denoted by c_3 and c_4, respectively, while the normal theory chi-square statistic is denoted by c_2.

In this study we demonstrate how the power of c_3 and c_4 varies with increasing kurtosis in a homogeneous and non-homogenous way. Since c_3 and c_4 depend on the ACM and the ACM depends on kurtosis, c_3 and c_4 are affected by kurtosis (Olsson, Foss & Troye, 2003).

We use two different cases: the model of parametric drift, where the models hold in the population (when sample size goes to infinity), and a case where the models do not hold in the population. In the first case, we study the non-centrality parameter as it can be read as the power of the test. In the second case, there is no real non-centrality parameter, so we base the discussion on the population value of the fit function, i.e., the minimum value of the fit function when the model is fitted to the population covariance matrix $\mathbf{\Sigma}_0$.

References

1. Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.

2. Satorra, A., Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308-313.
3. Jöreskog, K.G., Sörbom, D., Du Toit, S., Du Toit, M. (2003). *LISREL 8: New Statistical Features*. Third printing with revisions. Lincolnwood, IL: Scientific Software International.
4. Olsson, U.H., Foss, T., Troye, S.V. (2003). Does the ADF fit function decrease when the kurtosis increases? *British Journal of Mathematical and Statistical Psychology*, *56*, 289-303.

A Simulation Study of a Queuing System **L2**

Özer Özdemir

ozerozdemir@anadolu.edu.tr
Department of Statistics, Faculty of Science, Anadolu University, Eskişehir, Turkey

Key words: *simulation, queuing system, C programming, customer satisfaction, quality of service*

A simulation is an imitation of some real thing, state of affairs, or process. The act of simulating something generally entails representing certain key characteristics or behaviors of a selected physical or abstract system. A queue is not only a physical queue of people, it can also be a task list, a buffer of finished goods waiting for transportation or any place where entities are waiting for something to happen for any reason. An example of a queue is also the accumulation that occurs when customers come for a service from a system when the system is servicing another customer. We wrote a computer simulation (in the C language) in order to simulate queue systems. The purpose of the simulation is increasing customer's satisfaction by decreasing service cost and increasing quality of service with minimum waiting time of customers. By defining alternative solutions, the optimal system is obtained according to its economical stability between the interests of the customers and the management.

On Bagging and Estimation in Multivariate Mixtures **T1**

Reza Pakyari

r-pakyari@araku.ac.ir
Department of Mathematics, Arak University, Arak, Iran

Key words: *multivariate mixture model, maximum likelihood estimation, EM algorithm, bagging, subagging*

Two bagging approaches, say ½ *n*-out-of-*n* without replacement (subagging) and *n*-out-of-*n* with replacement (bagging) have been applied in the problem of estimation of the parameters in a multivariate mixture model. It has been observed by Monte Carlo simulations and a real data example that both bagging methods have improved the standard deviation of the maximum likelihood estimator of the mixing proportion, whilst the absolute bias increased slightly. In estimating the component distributions, bagging could increase the root mean integrated squared error when estimating the most probable component.

Multivariate Analysis for Space-Time Pollution Estimation **ST**

Monica Palma, Sabrina Maggio

m.palma@economia.unile.it; s.maggio@economia.unile.it
Faculty of Economics, University of Salento, Lecce, Italy

Key words: *multivariate space-time random field, space-time linear coregionalization model, space-time prediction*

The space-time linear coregionalization model, based on the generalized product-sum variogram model, is considered for modeling and prediction purposes in a spatial-temporal context. As known, a very difficult task in multivariate prediction is the modeling of variogram matrix. The space-time linear coregionalization model, which is a straightforward extension of the spatial linear coregionalization model to the spatialtemporal case, represents a simple way to solve the mentioned problem.

The case study presented in the paper highlights the flexibility of the space-time linear coregionalization model based on the generalized product-sum variogram in the modeling and interpolation techniques for a multivariate space-time random field.

Technical Efficiency of Philippine Rice-Producing Regions: A Stochastic Frontier Approach

E3

Niño T. Pate[1], Agustina Tan-Cruz[2]

[1] ntpate30@yahoo.co.uk
Palau Community College, Administration Division, Koror State, Republic of Palau

[2] ttcruz@usep.edu.ph
University of Southeastern Philippines, School of Applied Economics, Davao City, Philippines

This study attempts to measure the technical efficiency of irrigated and rainfed rice production using stochastic frontier approach with a balanced panel data of fifteen regions in the Philippines from period 1991-2002. The study measures the adequacy of different estimation methods to find out a more accurate model to represent irrigated and rainfed rice production in the Philippines and to provide a thematic map of the spatial distribution of technical efficiency among rice-producing regions using Geographic Information System (GIS). The frontier function involves inputs such as area of production, fertilizer applied in kilogram of nutrients (N:P2O5:K2O), cost of labor, seeds/planting materials in pesos, crop protection products, other miscellaneous inputs, year of observation, dummy for severe drought due to El Niño phenomenon and tropical cyclone passage. The stochastic frontier production functions and technical efficiency models were jointly estimated by the maximum-likelihood method and least squares method.

All the estimates have expected signs. The results showed that half-normal distribution with time-varying technical efficiency is an adequate representation of irrigated rice-producing regions in the Philippines. The preferred model for the rainfed rice production in the Philippines is the stochastic frontier production function with time-invariant technical efficiency, having half-normal distribution. About 76% of the technical efficiencies for irrigated regions clustered around 0.901 to 1.000. For the rainfed regions, the figures are 51% clustering around 0.901 to 0.950. The mean technical efficiency of irrigated rice-producing regions reveals that the Caraga, Cagayan Valley and Northern Mindanao regions are considered to be the most efficient, while Ilocos region, Central Luzon and Southern Mindanao are considered to be the most efficient in rainfed rice production.

The applied stochastic frontier approach did not involve farmer-specific variables, which could have provided an insight for policy framework.

Nonparametric Inequality Measure Based on Ranks

L2

Debdeep Pati, Anirban Bhattacharya, Abhishek Sarkar

debdeeppati@gmail.com; anirban86@gmail.com; abhishek.sarkar@gmail.com
Indian Statistical Institute, Kolkata, India

The aim of the project is to study the presence of inequality among the states of India with respect to several attributes. Given the ranks of 19 states of India on 8 different attributes during a period of time summarized in a "rank matrix" $A_{19\times 8}$ whose each column is a permutation of 1 to 19, our problem involves proposing a statistic that would be an indicator of inequality in the country for that period of time. Next, on the basis of data on several time instants, we would like to conclude if there is any trend in inequality amongst the states over the past decade.

Firstly, we tried out several common measures in this regard, e.g., the Friedman's statistic, a statistic involving the Cayley's distance, etc., but they reflected the inequality only in a crude way in the sense that they are likely to inflate or deflate mainly due to the inflation or deflation of the inter-column distances of A. So we have modified the statistic in such a way that it must represent how the matrix as a whole deviates from the perfect inequality situation i.e., matrices having identical columns. To this end, we proposed a statistic which represents the minimum of the minimum number of transpositions required to transform our given matrix into a perfect inequality matrix. Clearly, we could interpret smaller values of the statistic as more inequality in the sense that it is closer to perfect inequality situation.

Computing the aforesaid statistic for large dimensional matrices through complete enumeration is beyond the scope of a modern computer. Also no closed form expression of our proposed statistic for an arbitrary rank matrix is available. So we took resort to several combinatorial optimization techniques, though for two fairly restrictive situations we were able to give an exact solution to the proposed statistic. Again, the objective function (i.e. the minimum number of transpositions required to transform the given rank matrix into a perfect inequality matrix) that we are going to minimise is likely to have several local minima and hence standard optimisation techniques like gradient based random search or the Metropolis Hastings algorithm being stuck at a local minima are bound to give inaccurate results. We therefore used the Simulated Annealing algorithm, which is similar to hill-climbing or gradient search with a few modifications. We chose that algorithm as it works well with functions having lots of local minima. We got

accurate results using simulated annealing. We also obtained the bootstrap distribution of our statistic and as apprehended, we were able to detect high inequality amongst the states of India in all the four rounds of survey conducted by the National Sample Survey Organization.

Enhancing Performance of Credit Scoring Techniques Based on Logistic Regression and Generalized Additive Models

D2

Sabyasachi Patra[1], Kripa Shanker[1], Debasis Kundu[2]

1 sspatra@iitk.ac.in; ks@iitk.ac.in
Department of Industrial and Management Engineering, Indian Institute of Technology, Kanpur, India
2 kundu@iitk.ac.in
Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur, India

Credit scoring has gained attention of the managers and researchers as the credit industries have been experiencing enormous growth and severe competition during the past few years. The goal of credit scoring models is to categorize the applicants as either accepted (good) or rejected (bad) debtors prior to granting credit, on the basis of their financial and demographic conditions. Therefore it is necessary to evaluate the performance of the classifier, as high risks associated with inappropriate credit decisions may result in huge amount of loss to the creditors.

A well-established literature exists examining the relative performances of parametric and nonparametric techniques. However, to determine model and technique superiority, little organized research attention has been given to the evaluation methods. Considering the popularity in current business applications, in this paper we restrict our attention to logistic regression and Generalized Additive Models (GAM) and investigate how the performance of procedures is linked to amount of information, total number of covariates, number of influential covariates, correlation between covariates, and extent of nonlinearity. We also examine the classification accuracy of above techniques through three methods of evaluation – error rates, ROC (Receiver Operating Characteristics) curves and Root Mean Square Error (RMSE).

To improve generalization capability and model interpretability, stepwise logistic models based on Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are chosen. There are several procedures for fitting generalized additive models, i.e., multivariate logistic regression models where the influence of each single covariates is assumed to have unknown, potentially non-linear

shape. We considered two regression splines for implementing GAM: Thin Plate Regression Spline (TPRS) and Cubic Regression Spline (CRS). The terms were fitted by penalized maximum likelihood estimation and the actual degrees of freedom had been chosen by Generalized Cross-Validation (GCV). Two real-life credit scoring data sets (available at the STATLOG Database of UCI Repository) are used to compare logistic regression and GAM approaches.

The most popular metrics to assess classifiers' performance are error rate, precision and recall. We have shown how choice of cut-off points affects the performance. In general, when the class distribution is highly imbalanced for a two-class problem, classification accuracy can be misleading. A suboptimal choice of cut-off may increase predictive power, but they just overfit the current data. We have proposed a range of cut-off points for the two-class classification problem to overcome the addressed problem. On the other hand, ROC curve is more robust in the sense that it is insensitive to the change in the class distribution. The area under the ROC curve (AUC) aggregates the models behavior for all possible decision thresholds.

As our results revealed, logistic regression and GAM are quite competitive with each other. At optimal cut-off point logistic regression is superior to other techniques for both datasets. However, after calculating AUC for each model, it has been observed that GAM-CRS outperforms other models by a small margin. Our next consideration is to improve model performance by maximizing AUC and we look forward to building stable classification models based on that.

The Implications of Different Factor Analysis Solutions for the Theoretical Interpretation of the Choice of Educational Path

M1

Samo Pavlin[1], Tina Kogovšek[2]

1 samo.pavlin@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
2 tina.kogovsek@guest.arnes.si
Faculty of Arts, Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

Key words: *factor analysis, career orientation theory, theories of motivation*

The aim of the paper is to compare several Factor Analysis solutions with regard to different possible theoretical explanations of the choice of educational path. The paper begins with a description of the emerging career orientation theory and different theories of motivation. Firstly, the main factors and motives that impact the choice of educational path in the secondary and tertiary level of education are

highlighted and described. On the basis of empirical data, the need for (re)classification of these factors is shown. Comparison of the results of several Factor Analyses opens very different, but theoretically possible interpretations. However, some solutions have very clear implications, while others can blur the main message to a large extent. Finally, it is stressed that utilization of a single method, especially in the social sciences, does not assure reliable and valid data and research findings.

Data was collected in Slovenia from January to July 2006 on a non-random sample of 1512 occupational practitioners classified into 63 occupations. Those individuals were equally distributed among sectors and accordingly among the labour force. Occupations with different levels of occupational professionalisation were included (e.g., cleaners, farmers, drivers, hairdressers, teachers, lawyers, medical doctors etc.).

Data Mining Trauma Injury Data with Imputed Values

D1

Kay I. Penny[1], Thomas Chesney[2]

1 k.penny@napier.ac.uk
Napier University, Edinburgh, UK
2 tomas.cesney@nottingham.ac.uk
Nottingham University Business School, Nottingham, UK

The aim of this study is to investigate the accuracy of modelling patient death following trauma injury in conjunction with missing value imputation.

Methods for analysing trauma injury data with missing values, collected at a UK hospital, are reported. Missing data do not always cause concern when using data mining techniques. However, one measure of injury severity, the Glasgow coma score, which is associated with patient death, is missing for 12% of patients in the dataset. Applying the standard practice of complete-case analysis therefore means that 12% of the dataset would be excluded from the modelling, which may lead to bias in the results.

In order to include these 12% of patients in the analysis, three different data imputation techniques are used to estimate the missing values. The imputed data sets are then analysed by an artificial neural network and logistic regression, and the results are compared in terms of sensitivity, specificity, positive predictive value and negative predictive value.

Quality Issues in Questionnaire Design

DS

George Petrakos, Tonia Ieromnimon

george.petrakos@agilis-sa.gr; tonia.ieromnimon @agilis-sa.gr
Agilis SA – Statistics & Informatics, Athens, Greece

Questionnaire in any form is the basic instrument of the data collection process and also the communication means between the data provider and the data collector. Therefore several quality issues referring to questionnaire like relevance, completeness, clarity, accuracy, etc., can be addressed, defined and measured during the design and implementation of a statistical survey. The evaluation of these issues provides a methodological tool for assessing not only the quality of the questionnaire itself but also certain quality issues of the statistical data produced by the survey. A set of metadata referring to questionnaire characteristics and functionalities is developed in order to summarize and quantify, where possible, these quality issues.

On Reliability Equivalence Factor

T2

Tibor K. Pogány

poganj@pfri.hr
Department of Sciences, Faculty of Maritime Studies, University of Rijeka, Rijeka, Croatia

It is shown how system reliability is improved by three methods: (1) decreasing the scaling parameter of the argument in the probability distribution function of general positive *i.i.d.* random life components; (2) involving new components in hot duplication manner; (3) using the cold duplication method. Numerical results are presented using the Leipnik – Pearce life distribution.

Checking Hazard Regression Models Assumptions Using Pseudo-Observations

B1

Maja Pohar Perme[1], Per Kragh Andersen[2]

1 maja.pohar@mf.uni-lj.si
Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
2 p.k.andersen@biostat.ku.dk
Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Hazard regression models provide a convenient way of specifying how covariates affect survival time distributions. Typical examples include the multiplicative Cox proportional hazards model or the additive hazard model of Lin and Ying. Both models rely on certain assumptions concerning the functional form of the covariate effects and the scale on which the covariate effects are time-constant.The choice of the model depends on the data in hand and for that purpose a graphical evaluation of the data can be very helpful. The plotting of survival data, however, is hampered by its main discerning property, i.e., the presence of censored observations.

While model specific solutions do exist and are commonly used, we present a more general approach that covers all the hazard regression models using the same framework. The pseudo-observations are defined for each individual at all times regardless of censoring and therefore provide residuals that can be plotted without further difficulties. We present methods for simultaneously checking all the assumptions of both the Cox and the additive model, comment on their extensions to the multiple covariate case and compliment them with corresponding goodness-of-fit tests. We illustrate the methods using both simulated and real data sets..

Model for Evaluating Development of EU Countries

E2

Sonja Ratej Pirkovič

sonja.ratej@ef.uni-lj.si
Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

In this paper, I propose an alternative model for evaluating the development of Slovenia within the European Union. The offcial model and ranking is performed by the Institute of Macroeconomic Analysis and Development of the Republic of Slovenia (UMAR) and is part of the yearly national development report.

I present differences between the two models, which are due to different selection of data and different selection of reference point. A sensitivity analysis regarding different metrics and different ponders is presented.

I distinct between data on state and data on changes. In the ranking, I use only data on state. In projections, I use both data on state and data on changes. I show the differences in rankings of EU countries between my model and the official model in the years 2000-2006.

Robust Test for Testing Hypotheses on Finite Data Sets **T2**

Sonja Ratej Pirkovič, Aljoša Valentinčič

sonja.ratej@ef.uni-lj.si; aljosa.valentincic@ef.uni-lj.si
Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

In this paper, we propose a simple test for testing different hypotheses on finite data sets. The main advantage of our test is that no assumptions regarding the underlying distribution function are necessary. Nonetheless, by use of the Chebyshev inequality we are able to define the upper limit of probabilities of test values.

We can test the continuity of a distribution function or the type of distribution. Results of Monte Carlo simulations indicate the robustness of the test in that the hypothesis of continuity for distribution functions with jumps is rejected, whilst for continuous distributions it is not rejected. We also show that the test appropriately does not reject/rejects hypotheses regarding the type of distribution that a set of data follows. The test is particularly reliable for samples of more than 5,000 observations. It is robust and has time requirement of $\mathcal{O}(n)$.

We can use a similar test for comparing two unknown distributions. If it is expected that two data sets will have different expected values and different variations, but both should follow the same type of distribution, then we can compare standardized data sets. Our test is more robust and much less time consuming than bootstrap methods.

We have applied this test in the research of earnings management in Slovenian private firms. Some main results are presented.

Asymptotics for Periodically Stationary Time Series with Heavy Tails E3

Saeid Rezakhah, Hassan Ghasemi

rezakhah@aut.ac.ir; h.ghasemi@aut.ac.ir
Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology, Tehran, Iran

A stochastic process X_t is called periodically stationary if $\mu_t = \mathrm{E}(X_t)$ and $\gamma_t(h) = \mathrm{E}(X_t X_{t+h})$ for $h = 0, \pm1, \pm2, \ldots$ are all periodic function of t with the same period $\nu \geq 1$. If $\nu = 1$ then the process is stationary. The periodic ARMA process $\{\widetilde{X}_t\}$ with period ν, denoted by $\mathrm{PARMA}_\nu(\mathrm{p}, \mathrm{q})$, has representation

$$X_t - \sum_{j=1}^{p} \phi_t(j)\ X_{t-j} = \varepsilon_t - \sum_{j=1}^{q} \theta_t(j)\ \varepsilon_{t-j}$$

where $X_t = \widetilde{X}_t - \mu_t$ and $\{\varepsilon_t\}$ is a sequence of random variables with mean zero and standard deviation σ_t such that $\{\delta_t\} = \{\sigma_t^{-1}\ \varepsilon_t\}$ is i.i.d. The autoregressive parameters $\phi_t(j)$, the moving average parameters $\theta_t(j)$, and the residual standard deviation σ_t are all periodic function of t with the same period $\nu \geq 1$. We also assume that the model admits a causal representation $X_t = \sum_{j=0}^{\infty} \omega_t(j)\varepsilon_{t-j}$ where $\omega_t(0) = 1$ and $\sum_{j=0}^{\infty} |\omega_t(j)| < \infty$ for all t, and satisfies an invertibility condition $\varepsilon_t = \sum_{j=0}^{\infty} \pi_t(j) X_{t-j}$ where $\pi_t(0) = 1$ and $\sum_{j=0}^{\infty} |\pi_t(j)| < \infty$ for all t. In this paper we use the innovation algorithm to obtain parameter estimates for periodically stationary time series models. We also compute the asymptotic distribution for these estimates in the following cases:

- The case that the innovations have finite fourth moment;
- The case that the innovations have infinite fourth moment but finite second moment. In such a case we also assume that $P[|\delta_t| > x]$ varies regularly with index α and $P[\delta_t > x] \,/\, P[|\delta_t| > x] \to p$ for some $p \in [0, 1]$, denoted by δ_t is $\mathrm{RV}(\alpha)$.

These asymptotic results are useful to determine which model parameters are significant. Finally, we develop asymptotics for the Yule-Walker estimates.

Measurement Characteristics of the Students' Ratings of the Teachers | M2

Nino Rode

nino.rode@fsd.uni-lj.si
Faculty of Social Work, University of Ljubljana, Ljubljana, Slovenia

Within the quality assurance of the University of Ljubljana, the pedagogical work of the teachers is routinely rated by the students. The rating is performed yearly for the teachers' performance of the previous year. The ratings are part of the criteria for the teachers' promotion. There are some discussions on how such a measure can reflect on the quality of an individual teacher. One of the percieved problems of the instrument was lack of explicit criteria for the rating. There was also some evidence of the hallo effect in rating of individual caracteristics of the teachers.

To look into these problems, the data on the students' ratings at the School for Social Work were used to test the measutement characteristics of the instrument. The threats of individual differences between the students in rating criteria and the hallo effect were tested. The stability of criteria between the generations of the students was also assessed.

The Impact of Preprocessing on the Differentially Expressed Gene Lists | B2

Ana Rotter[1], Matjaž Hren[1], Špela Baebler[1], Andrej Blejec[2], Kristina Gruden[1]

[1] ana.rotter@nib.si
Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia
[2] andrej.blejec@nib.si
Department of Entomology, National Institute of Biology, Ljubljana, Slovenia

Microarrays provide a unique tool for analysing the expression of thousands of genes simultaneously within a single experiment. From a data analyst’s point of view, microarray technology offers a great challenge simply because of the nature of the microarray data. Instead of having large numbers of sample observations for a few variables, microarray data usually involve thousands of gene variables but only a few samples.

The first phase of microarray data analysis consists of data preprocessing and quality control. In the second phase, statistical analysis is performed in order to find potentially differentially expressed genes. Finally, the differential expression

of genes is biologically interpreted using various visualization tools or clustering methods.

This study focused on the influence of different preprocessing combinations on the outcome of statistical analysis, i.e., presence of genes in differentially expressed gene lists. Two independent experiments were used to investigate the influence of background correction, normalization and duplicate spots correlation calculation on discovery of differentially expressed genes. The intersection of genes resulting as differentially expressed and the biological relevance of results obtained was inspected.

We propose a rough guideline for finding genes which are *de facto* differentially expressed. Those are the genes whose membership in the differentially expressed gene lists is more robust.

Grant Acknowledgement

The work was financed through programme P4-0165 and projects 1000-05-310172 and Z4-9697 of the Slovenian Research Agency, and project 4302-38/2006/4 of the Ministry of Higher Education, Science and Technology of Republic of Slovenia.

Exploration of Categorical Screening Procedure Data by Multiple Correspondence Analysis

B2

Jože Rovan[1], Vilma Urbančič-Rovan[2], Mira Slak[2]

1 joze.rovan@ef.uni-lj.si
Department of Statistics, Faculty of Economics, University of Ljubljana, Slovenia
2 vilma.urbancic@kclj.si; mira.slak@kclj.si
Department of Endocrinology, University Medical Centre, Ljubljana, Slovenia

Key words: *screening procedures, diabetic foot, multiple correspondence analysis*

In November 1996, foot screening procedure was introduced at the out-patient diabetes clinic in Ljubljana. It consisted of medical history, foot examination (skin, toe nails, deformities etc.) and classification into risk groups. In 10 years, 7700 patients were examined. Most of the variables under consideration were categorical.

The relationship between variables was summarized by cross-tabulations and analyzed by multiple correspondence analysis (MCA). Using this methodology,

we came to a set of conclusions mostly described before by diabetic foot specialists, namely (a) for most of the variables under consideration, there was not much difference in the characteristics of the left and the right foot; (b) the patients with impaired arterial blood supply form a special group – possibly due to symptoms they seek help in earlier stages and are therefore not discovered by screening; (c) the category-points representing the groups with an acute foot ulcer, loss of protective sensation, absent foot pulses, foot deformity, abundant callus and history of previous foot ulcer were close together, confirming the influence of the known risk factors on ulcer development; (d) in the same way, the connection between loss of protective sensation and abundant callus formation was shown; (e) abundant callus, hallux valgus and toe nail deformities seem to be more frequent in women than in men – possibly due to fashionable footwear.

To summarise, multiple correspondence analysis was applied to the data on foot pathology in the population attending the out-patient diabetes clinic. The method proved to be a useful statistical tool for analysing the data of the screening procedures.

Bibliography

1. Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
2. Urbancic-Rovan, V. (2005). Causes of diabetic foot lesions. *The Lancet*, *366*, 1675-1676.

Population in Bosnia and Herzegovina: Survey Versus National Statistical Office Estimates

DS

Edin Šabanović

edin.sabanovic@efsa.unsa.ba
Faculty for Economics and Management, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

The last population census in Bosnia and Herzegovina was conducted in 1991. According to its results, 4,38 million inhabitants were living in the country. The 1992-1995 war had a huge negative impact on the population. The lack of recent population data was a significant problem for statistical offices in Bosnia and Herzegovina and their scheduled activities. They based their job on population estimates instead of census data.

In the early post-war situation, the demographic divisions in the Agency for Statistics of Bosnia and Herzegovina and entity offices started to produce population estimates based on the municipality lists created during the war for different purposes, UNHCR estimates, current registry data and fertility rates. The state level estimate was the simple total of entity and District estimates and these data were published on regular basis.

Some years after the war, the statistical offices also introduced surveys in their programmes. The first survey conducted in the post-war period was the "Living Standard Measurement Survey" in 2001, which was followed by three panel surveys "Living in Bosnia and Herzegovina" from 2002 to 2004. The Household Budget Survey was conducted in 2004 and Labour Force Survey in 2006 and 2007. Currently, the statistical offices are conducting the Household Budget Survey 2007.

The above mentioned surveys produced a wide range of data. The population estimates were one of their first outputs showing huge differences in figures in comparison with the NSO`s estimates. The differences were more evident on entity and District levels and were pronounced in both the size and the structure of the population. This subject matter was discussed not only within the statistical offices but also with many of their project partners, donors and international organizations. The discussions led to an independent expert population estimate made a few years ago, but that neither satisfied the data users nor offered a proper solution.

In this paper, we treat this issue, summarize and compare all survey and NSO's estimates in the post-war period and make some proposals for future work.

Climate Reconstruction S1

Matthew R. Schofield, Richard J. Barker

mschofield@maths.otago.ac.nz; rbarker@maths.otago.ac.nz
University of Otago, Dunedin, New Zealand

The study of climatological data is inhibited by the availability of data. Inference about the climate over the past hundreds or thousands of years cannot be based on direct observations, which are only available for the past century or two. To obviate this problem proxies with many more observations, such as isotopes, tree rings and ice cores are used to predict the missing climate observations using calibration/inverse regression methods. In this talk we will investigate the

assumptions and corresponding limitations of various calibration strategies and make suggestions about the use of such methods. If time permits, an example will also be given.

An Empirical Comparison of Model Selection Criteria for Parametric and Nonparametric Regression

ST

Meltem Şengün Ucal

msengun@khas.edu.tr
Department of Economics, Faculty of Economics and Business Administrative Sciences, Kadir Has University, Fatih, Istanbul, Turkey

Key words: *model selection, Kullback-Leibler information, AIC, BIC, bootstrap, cross-validation*

The purpose of this paper is to survey the different model selection criteria and compares them with each other.

"Which variables are important? How to select a model?" kind of questions are very important for modeling. A good model certainly fits well to the data under investigation (in econometrics). The econometrician and statistician would like to select the most appropriate model from data sets, whereby there may be more than one definition of "appropriate". Model selection criteria are one way to decide on the most appropriate model.

The analyzed model selection criteria are based on the information theory and are quite different from the usual methods based on null hypothesis testing. Information theory approaches were popular in the 1970s with the landmark Akaike Information Criterion based on the Kullback-Leibler discrepancy. Later, those approaches were diversified and such criteria as Bayes Information Criterion (BIC), Schwartz Information Criterion (SCI), and Mallow's Cp, as well as resampling methods (bootstrap and cross-validation) were developed.

This paper finds that most leading criteria perform well in selecting the best model, and several criteria also produce accurate probabilities of model superiority. Computationally intensive criteria failed to perform better than criteria which were computationally simpler. The use of several criteria in the application failed to appreciably outperform the use of one parametric or nonparametric model.

Reporting Uncertainity by Spline Function Approximation of Log-Likelihood

T1

Ahmet Sezer

ahsst12@yahoo.com
Department of Statistics, Anadolu University, Eskişehir, Turkey

Key words: *uncertainty, modified likelihood, raindrop plot*

Reporting uncertainty is one of the most important tasks in any statistical paradigm. Likelihood functions from independent studies can be easily combined, and the combined likelihood function serves as a meaningful indication of the support the observed data give to the various parameter values. This fact has led us to suggest using the likelihood function as a summary of post-data uncertainty concerning the parameter.

However, a serious difficulty arises because likelihood functions may not be expressible in a compact, easily-understood mathematical form suitable for communication or publication. To overcome this difficulty, we propose to approximate log-likelihood functions by using piecewise polynomials governed by a minimal number of parameters.

Our goal is to find the function of the parameter(s) that approximates the log-likelihood function with the minimum integrated (square) error over the parameter space. We achieve several things by approximating the log-likelihood; first, we significantly reduce the numerical difficulty associated with finding the maximum likelihood estimator. Second, in order to be able to combine the likelihoods that come from independent studies, it is important that the approximation of the log-likelihood should depend only upon a few parameters so that the results can be communicated compactly.

By the simulation studies we compared natural cubic spline approximation with the conventional modified likelihood methods in terms of coverage probability, confidence interval length of highest density region of the approximated likelihood and the mean squared error of the maximum likelihood estimator. We also showed how to use raindrop plots for the approximated likelihood functions.

Bootstrapping Congruence Coefficients in Principal Component Solutions

M1

Gregor Sočan

gregor.socan@ff.uni-lj.si
Department of Psychology, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

Structural equation modelling (SEM) is currently the prevailing approach to the analysis of stability of covariance structures across time or across samples, respectively. In recent years, however, several researchers expressed doubts concerning the optimality of this approach in the area of personality structure, and proposed a more frequent use of the Procrustes rotation in combination with the congruence coefficient. In the past, a decisive drawback of the Procrustes-based approach was its tendency to capitalise on chance. More recently, however, computationally intensive methods, like baseline permutation tests and especially the bootstrap, provided means to control for the upward bias of congruence coefficients. Chan et al. (1999), for instance, proposed a bootstrap method for testing the stability of factor structures. We present a modification of the Chan et al. method that can be used for testing the component structure rather than the common factor structure. The main feature of our modification is a different definition of the resampling space, specified accordingly to the principal component model. The analysis is illustrated using results from a study investigating the stability of the personality structure in early childhood. We further present results of a simulation study that evaluated the determinants of the accuracy of the method and compared it to the "naïve" bootstrap. Finally, we discuss the practical aspects of our procedure, especially with relation to the results obtained by SEM.

References

1. Chan, W., Leung, K., Chan, D.K.-S., Ho, R.M., Yung, Y.-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods*, *4*, 378-402.

Testing Dose-Response with Multivariate Ordinal Data **B1**

Aldo Solari[1], Bernhard Klingenberg[2], Luigi Salmaso[3], Fortunato Pesarin[1]

1 solari@stat.unipd.it; fortunato.pesarin@unipd.it
Department of Statistics, University of Padova, Padova, Italy
2 bklingen@williams.edu
Department of Mathematics and Statistics, Williams College, Williamstown, Massachusetts, USA
3 salmaso@gest.unipd.it
Department of Management and Engineering, University of Padova, Vicenza, Italy

Key words: *adverse events, closed testing, combining P-values, correlated ordinal observations, multiple endpoints, safety analysis, stochastic order*

Many assessment instruments used in the evaluation of toxicity, safety, pain or disease progression consider multiple ordinal endpoints to capture the presence and severity of dose effects. Contingency tables underlying these correlated responses are often sparse and imbalanced, rendering asymptotic results unreliable or model fitting prohibitively complex without simplifying assumptions. Instead of modeling the dose response directly, we look at stochastic order as an expression or manifestation of a dose effect under much weaker assumptions. Several approaches, ranging from statistics that contrast mean scores to combining dependent P-values for score-free tests are derived for the two and many multivariate sample case. Often, endpoints are grouped together into physiological domains or by the body function they describe. We derive tests based on these subgroups which might supplement or replace the individual endpoint analysis because they are more powerful. The permutation approach is used throughout to obtain global, subgroup and individual significance levels as it naturally incorporates the correlation among endpoints. Multiplicity adjustments for the individual endpoints are obtained via step-down procedures, while subgroup significance levels are adjusted via the full closed testing framework. The proposed methodology is illustrated using a collection of 25 correlated ordinal endpoints to evaluate toxicity of a chemical compound.

A Measure of Prognostic Value of Survival Models B1

Janez Stare, Maja Pohar Perme

janez.stare@mf.uni-lj.si; maja.pohar@mf.uni-lj.si
Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

There is no shortage of proposed measures of prognostic values of survival models in statistical literature. They come under different names (explained variation, correlation, explained randomness, ...), but their goal is common: to define something analogous to the coefficient of determination (R^2) in linear regression. Such measures will typically (surprisingly, not all of them do!) lie between 0 and 1, reaching 1 under a "perfect" model. Some of the measures are only defined for the Cox model, and only a few can meaningfully deal with time-dependent covariates and time-dependent effects.

We present a new measure that has no problem with time-dependency, is easily calculated, has a straightfoward interpretation, and can be used with any survival model. In fact, it is naturally adapted for comparison of different models. We give the population definition, discuss properties of the estimator, and illustrate its usage on some real data sets.

Computer-related Statistical Analysis – Ideas, Otimizations, Problems, Advice L2

Gábor Tébi

varietas1@t-online.hu
Faculty of Science & Faculty of Economics and Business Administration, University of Szeged, Szeged, Hungary

At the Applied Statistics 2006 conference, Kovács and Petres introduced a new indicator of multicollinearity in linear regression models based on eigenvalues of the correlation matrix, called Petres' Red. The distribution of the indicator is yet to be determined and confirmed using custom-built software. The first program was too slow and the number of indicators were limited, thus limiting the number of explanatory variables as well. The second one reduced the running time from days to minutes and eliminated the limitation of the number of indicators. This example proves how important it is to use proper programming language, algorythms and data structures, to check for possible optimizations, to watch out for problems like floating-point number representation, etc., in computational statistics.

Expectation of Maxima of Random Variables: Theory and Application | T3

Daniel Tokarev, Kais Hamza, Firma C. Klebaner

{daniel.tokarev;fima.klebaner;kais.hamza}@sci.monash.edu.au
School of Mathematical Sciences, Monash University, Australia

In a recent paper (with K. Hamza, P. Jagers and A. Sudbury; submitted), tight bounds for expectation of a maximum of a system of k random variables were obtained. This paper focuses on the asymptotics of this expectation as k goes to infinity where all random variables are *iid*. Regular behaviour of tails of their common distribution is assumed and Abelian/Tauberian theorems for the expectation of maximum are derived for different cases of tail behaviour. Immediate applications include asymptotics of mean time to extinction of critical Galton-Watson processes with infinite variance as studied by Slack. Expectations of maximum of k copies of a random variable (which we call meanmax) are analogous to moments in some respects. We showt hat they determine the random variable uniquely and solve some optimisation problems with the aid of meanmax generating function (MMGF).

Prediction of GDP per Capita of Turkey and Balkan Countries: Comparison of ARIMA Models, Neural Networks and Support Vector Machines | E3

Seda Tolun Esen[1], Şebnem Er[1], Barış Kiremitci[2]

1 stolun@istanbul.edu.tr; sebnemer@istanbul.edu.tr
Department of Quantitative Methods, Faculty of Business Administration, Istanbul University, Istanbul, Turkey
2 baris@istanbul.edu.tr
School of Transportation and Logistics, Istanbul University, Istanbul, Turkey

This paper presents comparison of ARIMA models, neural networks (NN) and support vector machines (SVM) in predicting GDP per capita values of Turkey and Balkan countries for the period of 1995-2006. The GDP is equal to the total of the gross expenditure on the final uses of the domestic supply of goods and services valued at price to the purchaser minus the imports of goods and services. The GDP can serve as an indicator of the scale of a country's economy. However, to judge a country's level of economic development, it must be divided by the country's population. GDP per capita is one of the main indicators that express the

average standard of living of individuals in a country. Economic growth is often seen as indicating an increase in the average standard of living, and that is why the prediction of GDP per capita is crucial for economic planners to evolve new policies to support economic growth. GDP per capita can be estimated by various time series models such as decomposition, moving average, exponential smoothing and ARIMA models, and also by artificial intelligence techniques such as NN and SVM. Our essential objective in this paper is to state which method is best suited for the countries considered in the study.

The Maximum Term for Testing the Homogeneity of Two Multinomial Populations with a Large Number of Categories

B2

Adelaide Valente Freitas[1], Miguel Pinheiro[2], José Luís Oliveira[2], Gabriela Moura[3], Manuel Santos[3]

1 adelaide@mat.ua.pt
Department of Mathematics, University of Aveiro, Aveiro, Portugal
2 IEETA, University of Aveiro, Aveiro, Portugal
3 Department of Biology, University of Aveiro, Aveiro, Portugal

Key words: *extreme value distribution, Pearson chi-squared statistic, contingency table, ORFeome*

For comparing two populations A and B, each one described by a unknown multinomial probability distribution with a large number of mutually exclusive categories, we propose a new test statistic (T_m) defined as the maximum term of components of the partitioning of the Pearson chi-squared statistic (X^2_P) formulated by Kimball (1954), and its limiting distribution is derived (Freitas et al., 2006).

Furthermore, we compare the results obtained from the statistics X^2_P and T_m for testing the homogeneity of codon contexts of the complete ORFeome sequences of *Homo sapiens* and *Pan troglodytes*. The statistic T_m has the advantage for identifying the nucleotide A-Adenine as one of the responsible for the rejection of the hypothesis of homogeneity between the distributions of the codon pairs in these two species.

References

1. Freitas, A.V., Pinheiro, M., Oliveira, J.L., Moura, G., Santos, M. (2006). A new limiting distribution for a statistical test for the homogeneity of two multinomial populations. In W. Urfer & M.A. Amaral Turkman (Eds.),

Proceedings of the Workshop in Statistic on Genomics and Proteomics, Monte Estoril, Portugal, October 5-8, 2005. Coimbra: Centro Internacional de Matemática, 113-120.
2. Kimball, A.W. (1954). Short-cut formulae for the exact partition of chi-square in contingency tables. *Biometrika*, *10*, 452-458.

Nonparametric Evaluation in the Statistical Problems of Finance Theory

E2

Sergey A. Vavilov[1], Konstantin Yu. Ermolenko[2]

1 svavilov@som.pu.ru
Department of Finance Theory, School of Management, St. Petersburg State University, St. Petersburg, Russia
2 k.ermolenko@econ.pu.ru
Department of Economic Cybernetics, Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia

Classical theory of pricing under uncertainty implies that the price of high liquid asset x_t is a random process on the time interval [0, T] and follows stochastic differential equation

$$dx_t = c_t x_t\, dt + \sigma_t x_t\, dW_t, \tag{1}$$

where a factor of volatility $\sigma_t = \sigma(t, \omega)$ and $c_t = c\ (t, \omega)$ are, generally speaking, measurable random functions. The pointed out model of pricing corresponds to geometrical Brownian motion. To avoid misunderstanding, the realization of any random process in contrast to the process itself is denoted by the corresponding letter with wave. Then, integrated volatility on the time interval [0, T] is defined as follows:

$$J(t) = \int_0^t \tilde{\sigma}_s^2 ds\ ,\ 0 \le t \le T \tag{2}$$

The problem to calculate integrated volatility consists of quantity (2) evaluation on the basis of the observable realization $\tilde{x}_t$ of the stochastic process x_t. The problem to calculate integrated volatility for different financial assets is the inseparable part of financial engineering topics. Hence, it is not surprising that a substantial amount of literature concerning the problem of integrated volatility evaluation has been published. These publications are based on deriving various sufficient conditions to provide the convergence in probability to $J(t)$ of specially

constructed sums (realized variance) calculated on the basis of statistical data. The application of such approach to solve the problem under consideration imposes many artificial restrictions generated by the specificity of the method applied. To what extent such restrictions correspond to the realities of financial world seems to be unclear.

In the present work, a new approach to tackle the problem of integrated volatility evaluation is proposed. The integral equation to provide the calculation of integrated volatility (2) is derived. The solving of this integral equation turns out to be a standard ill-posed problem of mathematical physics and is solved by making use of the well-known methods of functional analysis. It is worth noting that the proposed approach also imposes a number of artificial constraints which may be attributed to the class of sufficient conditions. Nevertheless, there is the possibility to compare the numerical estimates of the evaluated quantity based on the application of two different arguments. It is reasonable to believe that if the numerical results coincide, the plausibility of the integrated volatility evaluation increases markedly. It should be added that the reduction of the original problem to the ill-posed problem under consideration makes its solution robust towards different kind of errors, for instance, the presence of anomalous data.

Optimization Problem in Markov Chain Modeling **L2**

Vasja Vehovar[1], Damjan Škulj[1], Mihael Perman[2]

1 vasja.vehovar@fdv.uni-lj.si; damjan.skulj@fdv.uni-lj.si
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
2 mihael.perman@fmf.uni-lj.si
Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

The problem that stimulated this research arose from human resources management within Slovenian Army. There, the employees are structured into around 70 status categories, according to the hierarchy (seniority), from soldiers to generals, from civil servants to reserve army staff. One of the key challenges in this research was related to the specifications and modeling of the (future) transition probabilities between statuses. The aim is that within certain number of years the structure (i.e. the number of persons in each of the status categories) will match the desired structure.

To address this problem, various approaches were considered, including simulations. The paper presents methodological challenges and approaches used to solve this optimization problem.

Methodological Issues in Analyzing Social Networks in Online Forums **SM**

Vasja Vehovar, Aleš Žiberna, Aleks Jakulin

vasja.vehovar@fdv.uni-lj.si; ales.ziberna@fdv.uni-lj.si; jakulin@acm.org
Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

Online forums enable people to discuss different topics by posting messages into the existing topics (threads) or by opening new ones. We will focus on the usual type of online forums, where there is no exact information whether a post was a response to some other specific post, or the post was included simply as a general contribution to the discussion.

We can observe online forum networking at three levels:
1. person-person: some participants can be related more often with each other via direct links (i.e., reply, comments) or indirect ones (i.e., participating in the same topics/thread);
2. person-topic: participants can be related indirectly by the common interest, which is demonstrated by participating in similar topics (but not necessarily in the same thread);
3. topic-topic: the topics themselves are also indirectly related via participants, whereby topics can be operationalised by name of the thread, but also with tags or with descriptors.

In this paper, we address the key methodological issues: how to define metrics, how to generate tags, how to deal with the time component, and how to solve the problem of non-registered participants and missing data. Methodological dilemmas are illustrated by using empirical data from real online forums.

M-Shaped Distributions

EM

Gaj Vidmar

gaj.vidmar@ir-rs.si
Institute for Rehabilitation of the Republic of Slovenia, Ljubljana, Slovenia

A non-exhaustive compendium of eleven distributions with probability density function (or, in discrete cases, probability mass function) resembling the capital letter M is presented, including both continuous and discrete random variables, and both single random variables and mixtures of two random variables. The exhibits range from banal to relatively advanced, and from purely academic exercises to a real-life example of reported annual stock-exchange returns.

All the calculations and graphics are combined with interactive input of parameters in a spreadsheet, which also serves as the means of presentation to the audience. This provides yet another demonstration of the usefulness of spreadsheets in mathematics and statistics education, particularly for non-mathematicians, undergraduate students and self-study.

To conclude in a recreational mathematics manner, which also has notable paedagogical value for the same target population, the natural sequel to M-shaped distribution is presented, i.e., the N-shaped distribution.

Everyday Benefits of Understanding Variability

EM

K. Larry Weldon

weldon@sfu.ca
Simon Fraser University, Vancouver, Canada

Statistics professionals usually focus on the methods of formal inference. Informal inference is left to investment analysts, sport commentators, government bureaucrats and others who may rely on their intuition for guidance rather than a formal education in statistics. In this paper, I provide some examples of often-overlooked phenomena that would be useful for the layman. The contexts for these examples are investment, sport, academic research, health, and lotteries. I suggest that statistics professionals should allot some energy to communicating such examples to the general public.

How to Objectively Rate Investment Experts in Absence of Full Disclosure? An Approach Based on a Near Perfect Discrimination Model

SM

Patrick Wessa

patrick.wessa@lessius.eu; patrick@wessa.net
Lessius Business School, Integrated Economics Faculty, Association Catholic University of Leuven, Leuven, Belgium

The result of this investigation is an operational model that can be used to accurately identify real stock market time series. In other words, if we are presented with a collection of blinded time series (real-life time series and simulated Random-Walks) then the proposed model will allow us to discriminate between both categories. In addition, it is shown that the type II error of this model quickly converges to zero as the time series length increases. The most remarkable feature of this model is its simplicity: a (bias-reduced) logistic regression with a single exogenous variable (kurtosis p-value) based on the Quasi Random-Walk model that relates returns of equity and the entire market in times of large market returns. This model can be used as an objective rating benchmark for the models that are used by hedge funds to identify the stocks that should be used in a market neutral arbitrage strategy of long and short positions. In addition, it allows independent auditors to objectively evaluate the added value of statistical and technical analysis techniques that are often used in investment decisions. A rating mechanism that is based on the proposed benchmark provides valuable information about the investment strategy even in absence of full disclosure.

Learning Attitudes, Peer Assessment, and Gender in the Context of a Social Constructionist Statistics Course

EM

Patrick Wessa

patrick.wessa@lessius.eu; patrick@wessa.net
Lessius Business School, Integrated Economics Faculty, Association Catholic University of Leuven, Leuven, Belgium

This paper provides an illustrated, explorative, and empirical analysis, based on a student-centered, constructivist statistics course for a large student population (150+ students). It shows that not learning attitude but gender plays an important role in relation to the performance at the final examination (aimed at measuring analytical competences). It is shown that a variety of external stimuli, such as a learning environment that is based on social constructivism, and the competences

of the educator, perform different roles for female and male students. It is also concluded that Peer Assessment is an excellent learning activity but not necessarily an accurate evaluation tool.

Using Profile Likelihood for Semiparametric Model Selection with Application to Proportional Hazards Mixed-effects Models

B1

Ronghui Xu

rxu@ucsd.edu
Department of Family and Preventive Medicine, Department of Mathematics, University of California, San Diego, USA

We consider selection of nested and non-nested semiparametric models. Using profile likelihood, we can define both a likelihood ratio statistic and an Akaike information for models with nuisance parameters. Asymptotic quadratic expansion of the log profile likelihood allows derivation of the asymptotic null distribution of the likelihood ratio statistic including the boundary cases, as well as unbiased estimation of the Akaike information by an Akaike information criterion.

Our work was motivated by the proportional hazards mixed effects model (PHMM), which incorporates general random effects of arbitrary covariates and includes the frailty model as a special case. The asymptotic properties of its parameter estimate has recently been established, which enables the quadratic expansion of the log profile likelihood. For computation of the (profile) likelihood under PHMM, we apply three algorithms: Laplace approximation, reciprocal importance sampling and bridge sampling. We compare the three algorithms under different data structures, and apply the methods to a multi-center lung cancer clinical trial.

Impact of Missing Data Treatment on Results of Ward Hierarchical Clustering

EM

Anja Žnidaršič[1], Tanja Garvas[2], Saša Planinc[3]

1 anja.znidarsic@siol.net
Postgraduate Study Programme in Statistics, University of Ljubljana, Ljubljana, Slovenia
2 tanja.garvas@gov.si
Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
3 sasa.planinc@turistica.si
Turistica – College of Tourism, University of Primorska, Portorož, Slovenia

Missing data are frequently present, especialy in surveys. There are several ways how to treat such data, which give more or less good results with a particular statistical method. In the paper, some known imputation techniques and two most frequently used treatments – casewise (listwise) and pairwise deletion – are used and compared. We studied the impact of missing data treatment on the results recieved from hierarchical clustering using Ward method under different conditions. These conditions were obtained by generating two kinds of missing data: missing completely at random (MCAR) and missing at random (MAR) on otherwise complete real data. Different treatments on different percentage of missing data were used. The impact of missing data treatment was then estimated by external and internal cluster validation techniques (criteria). Using external technique (Rand index), we compared if the same units of two diferent partitions (without and with missing data by using a particular treatment) are classified into the same group. Using internal techniques, we studied characteristics of a particular partition (separation and homogeneity of groups). The aim of the study was to find out which missing data treatment is the most suitable when analysing data by Ward hierarchical clustering method considering particular percentage of missing data and the type of missing data.

WORKSHOP

Introduction to R

Andrej Blejec
(National Institute of Biology, Ljubljana, Slovenia)

R is a freely available implementation of the S language. It is a statistical environment particularly suitable for development of new methods, characterised by high-quality graphics. In the academic setting, it has become the dominant tool for statistical exploration.

The workshop will introduce the participants to the use of of R, comprising:

- Basic functions for data processing and analysis;
- Graphical displays;
- Writing new functions;
- Connecting R with other programs (Excel, SPSS).

NOTES

AUTHOR INDEX

Supported by

Slovenian Research Agency (ARRS)
http://www.arrs.gov.si/en

Statistical Office of the Republic of Slovenia
http://www.stat.si

Alarix d.o.o.
http://www.alarix.si

SPSS Slovenia
http://www.spss.si

Result d.o.o.
http://www.result.si

Zavod za turizem Ljubljana
http://www.ljubljana-tourism.si