Abstracts

An Alternative Parametric Form for Estimating Lorenz Curves
Ibrahim Abdalla and Mohamed Y. Hassan
(United Arab Emirates University, United Arab Emirates)

In this paper a new parametric Lorenz curve is proposed and then fitted to grouped income data utilizing Abu-Dhabi Emirate family income survey, 1997. Given the nature of the distribution of income and the distinct characteristics of Abu-Dhabi Emirate, the proposed function performed well better than other well known parametric forms in the literature. Some of these parametric forms are special cases of this function. Using estimated parameters, based on the new model, the Gini coefficient is calculated to measure inequality of income. Obtained results for the new form satisfy Gastwirth criterion.

Simultaneous Estimation of Indirect and Interaction Effects using Structural Equation Models
Joan Manuel Batista-Foguet (Esade, University Ramon Llull, Spain)
Germà Coenders (University of Girona, Spain)
Willem E. Saris and Josep Bisbe (Esade, University Ramon Llull, Spain)

In most management research studies, interaction effects are modeled by means of moderated regression analysis. Structural equation models with non-linear constraints make it possible to estimate interaction effects while correcting for measurement error. From the various specifications, Jöreskog and Yang's (1996, 1998), likely the most parsimonious, has been chosen and further extended. Up to now, only direct effects have been specified, thus wasting much of the capability of the structural equation approach. This paper presents and discusses an extension of Jöreskog and Yang's specification that can handle direct, indirect and interaction effects simultaneously at any place in a model. The approach is illustrated by a study of the effects of an interactive style of use of budgets on both company innovation and performance.

Evaluating Policy Networks: Abilities and Constraints of Social Network Analysis

Gerd Beidernikl and Dietmar Paier
(Center for Education and Economy, Research and Consulting, Graz, Austria)

In 1997 the European Commission called for the submission of projects under an initiative called 'Territorial Employment Pacts'. TEPs are contractual alliances between protagonists from various sectors on local level in order to develop innovative measures for job creation and job protection in their area. Through this initiative the effectiveness and the relevance of measures is intended to be improved by enriching policy with local partnership. Authority is relocated downwards and sideways from central states and therefore a 'bottom-up' change in the former 'top-down' sector of employment policy is initiated. This leads to the emergence of so-called policy networks. They can be characterized as stable, but only on a low level formalized, communication and cooperation networks between governmental and non-governmental actors, involved in a political process.

Policy networks are kind of a characteristic of modern EU policy in general. More and more initiatives are aiming at the establishment of local partnership and at initiating processes of local policy learning. The social outcomes themselves become more and more important objectives of these initiatives. This on return calls for the development of tools suitable for evaluating these social settings and generating knowledge relevant for the local protagonists. But how to evaluate an initiative based on partnership? Classic evaluation approaches rather focused on measuring effectiveness on base of the jobs created (e.g. input-output-analysis) than on the policy networks established. The underlying research project tries to fill this gap by using the instrument of Social Network Analysis in the evaluation of the Styrian TEPs. 476 local protagonists belonging to the six regional pacts in Styria were questioned in autumn 2002. They gave information about their communication and cooperation patterns and additional information about their involvement in the regional policy networks. Based upon data provided by SNA we were able to deliver results that led to a significant change and improvement of the TEPs.

The proposed paper deals therefore with two aspects of SNA. On the one hand we are going to discuss our experiences in evaluating policy networks in labor market policy using the instrument of SNA in the field of the Styrian TEPs. We are going to unfold our research design and its features developed for initiating policy learning among the pact protagonists. Secondly we are going to focus on a methodologic question: Is SNA an appropriate tool for evaluating policy networks in general? We will try to outline the answer by deriving abilities as well as

constraints of the method based upon experiences from the TEP evaluation. Concluding, we would like to discuss other possible applications for SNA (especially the development of universal performance indicator sets and quality criteria for policy networks as well as web-based monitoring tools for transnational cooperation networks, e.g. EQUAL or INTERREG).

Using Propensity Scores to Reduce Self-Selection Bias: Comparing Effectiveness of the University Training on Professional Outcomes. A Case Study on the University of Bergamo.

Silvia Biffignandi (University of Bergamo, Italy)
Monica Pratesi (University of Pisa, Italy)

In this contribution we are interested in measuring the effect of a voluntary training programme (the university courses) on several professional outcomes (e.g. finding work, finding a coherent work, having career perspectives). In the absence of an observable counterfactual (the outcome that would have resulted had the students not graduated), among several non-experimental evaluation techniques, we use the matching methodology: if the selection can be explained purely in terms of observable characteristics, the matching can be based on propensity scores (Rosenbaum and Rubin, 1983). For every student in the treatment group (a specific training course) a matching individual with the same propensity score is found from among the non-treatment group (another training course). The target population is represented by a statistical register maintained by the university administration: for each person enrolled in a university training course there is an administrative record where each event in the student career - from the university enrolment to the degree or to the abandon - is registered. The professional outcomes of three cohorts of graduates (1999, 2000, 2001) are surveyed by a complete mail-survey.

In the contribution we examine the sensitivity of the results to alternative specifications of the method, we refer to the choice of variables for the propensity score model and the choice of matching method (number of digits in the matching, radius matching). We address also the problem of matching when using survey data (Bryson, 2001; Green et al., 2001). For reasons of survey non-response, or oversampling of particular subgroups, survey data can not be representative of the whole population. It is usual to apply weights to restore the profile of the survey data to that of the population on a number of key characteristics. The role of these weights is explored both in the pre-matching operations, estimating the participa-

tion model which generates the propensity scores, and in the post-matching ones, weighting each treatment group member and the corresponding comparator(s).

References

1. Bryson, A. (2001): *The Union Membership Wage Premium: An Analysis Using Propensity Score Matching*. Centre for Economic Performance Working Paper n. 1160, London: School of Economics.
2. Green, H., Connoly, H., Marsh, A., and Bryson A. (2001): *The Longer-term Effects of Voluntary Participation in ONE*. Department of Work and Pensions, Research report number 149.
3. Parsons, L.S. (2001): Reducing bias in a propensity score matched-pair sample using greedy matching techniques. *Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.
4. Rosenbaum, P.R and Rubin, D.B. (1983): The central role of the propensity score in observational studies for causal effect. *Biometrika*, **70**, 41-55.
5. Dehejia, R.H. and Wahba, S. (1999): Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, **448**, 1053-1062.

Customer Satisfaction in Different Groups: Evaluation and Comparison through the Rasch Model

Andrea Bonanomi (Universita Cattolica del Sacro Cuore di Milano, Italy)
Livio Finos (Universita degli Studi di Padova, Italy)
Silvia Salini (Universita degli Studi di Milano, Italy)

Introduction

In the present paper the Rasch Model (Rasch, 1960) is applied to evaluate the Service Quality. In section 2 Rasch Polytomous Model will be briefly presented; moreover the use of Rasch analysis in the customer satisfaction field will be considered. (Barsotti, 2001). In section 3 a Quality Total Index will be created; the index considers the decomposition between subject factor and attribute factor and ponders the scores with the Rasch parameters. In section 4 distinct models in different groups will be estimated; the aim is to obtain a coefficients test that allows to verify for each attribute if the differences of the parameters are statistically significant.

Rasch Analysis in Customer Satisfaction

In 1960 Georg Rasch proposed a statistical model that complied with a fundamental assumptions made in measurements in physical sciences. It allowed for the transformation of the cumulative raw scores (achieved by a subject across items, or by an item across subjects) into linear continuous measures of ability (for subjects) and difficulty (for items) (Tesio, 2003).

The evaluation of Service Quality is generally obtained through the administration of questionnaires composed of items expressed by an ordinal scale (items). The response to each items depends on only two factors: the effective attribute quality and the personal subject satisfaction. Therefore, in the Quality Service application, there are a correspondence with the duality item/person of the Rasch standard model. In particular the just mentioned measure of ability becomes now a measure of personal satisfaction and a measure of difficulty becomes now a measure of objective quality.

Quality Total Index

A Quality Total Index is a univariate numeric measure that summarizes many variables also categorical. It is used to compare and rank different objects. This aggregated index is often realized in a rudimental way, simply by a linear combination of scores. First the ordinal categories are converted in numerical score through Thurstone transformation (Zanella, 2001) and then the Rasch coefficients are used to weight the single elements of linear combination. A generalization of Thurstone transformation, trough the nonparametric combination of dependent ranking (Pesarin and Lago, 2000), is also possible.

Comparison of groups

The comparison on the satisfaction and quality (evaluated by the Quality Total Index) among groups or customers clusters is often of relevant interest. Such comparison is performable with tools as ANOVA models. In the case of errors with not normal distribution, an alternative way is given by permutation tests use.

References

1. Barsotti, L. (2001): *Analisi della soddisfazione del paziente in una struttura sanitaria: un caso di studio*. Universita Cattolica del S. Cuore, Istituto di Statistica, Serie E. P. N. 104.
2. Pesarin, F. and Lago, A.(2000): Non-parametric combination of dependent rankings with applications to the quality assessment of industrial products, *Metron*, LVIII, **1-2**, 39-52.
3. Pesarin, F. (2001): *Multivariate Permutation Tests with Applications in Biostatistics.* Chichester: Wiley.

4. Rasch, G., (1960): *Probabilistic Mdels for Some Intelligence and Attainment Tests.* Copenhagen, Danish Istitute for Educational Research.
5. Tesio, L. (2003): Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal Rehabilitation Med.*, **35**, 105-115.
6. Zanella, A. (2001): *Valutazione e modelli interpretativi di Customer Satisfaction: una presentazione di insieme*. Universita Cattolica del S. Cuore, Istituto di Statistica, Serie E. P. N. 105.

The Mortality Data Collection Techniques in Tunisia

Sofiane Bouhdiba (Tunis)

The collection of mortality data has always been a problem all over the world. In Tunisia, this problem has been particularly sharp, un till the beginning of the new millennium. In fact, for more than 40 years, since the in dependance of the country (1956), little did we know about causes of death, relationship between life expectancy and feeding, and many other mortality indicators. Now things have changed, and the new projects of collecting mortality data were held successfully in Tunisia.

I propose, in my paper, to show the lessons learnt from these long years of unknown , by discussing the old and new mortality data collection techniques (questionnaires, samples,) and also these data quality (real causes of death, age confusions, validity of responses,...).

I would like to try to find an answer to some questions:

- What are the cultural and social problems that appear in a traditional Arabic country like Tunisia?
- What are the new techniques of collecting mortality data?
- Why do we have today a precise idea about the causes of death in the urban areas in Tunisia, while in a recent period more than 50% of deaths were classified under the item 'natural cause of death'?
- What can we do to improve the mortality data collection in the coming years?

My paper is divided into three parts : first, I will make a flash back to present the main problems found in the past during the collection of mortality data. Then, in the second part, I will discuss the new methods used to obtain good data collected, by eliminating or at least reducing errors. In the last part of the paper, I will present some recommendations to improve the collection of mortality data in Tunisia.

Compositional Data Analysis with R

Matevž Bren (University of Maribor, Slovenia)
Vladimir Batagelj (University of Ljubljana, Slovenia)

R (http://www.r-project.org/) is a free language and environment for statistical computing and graphics. R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, ...).
In the paper we present an R library for compositional data analysis that provides support for:

- operations on compositions such as perturbation, power multiplication, subcomposition, distances ...;
- various logratio transformations of compositions to transform compositions into real vectors that are amenable to standard multivariate statistical analysis;
- compositional concepts such as complete subcompositional independence, the relation of compositions to bases, logcontrast models ...; and
- graphical presentation of compositions in ternary diagrams and tetrahedrons.

The current version of the library is available at
http://vlado.fmf.uni-lj.si/pub/mixture/

References

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

A Multivariate Analysis of the Influences and Profiles of Treated Drug Misuse in Ireland

Paul Cahill and Brendan Bunting (University of Ulster, Northern Ireland)

Objective. This study provides an analysis and profiles of illegal drug usage in the Republic of Ireland. Two questions are addressed: a) can individuals be grouped into heterogeneous classes based upon their type of drug consumption, and b) how much do these heterogeneous classes differ in terms of other key variables,

for example age, gender, education, living with drug misuser, and employment status?

Methodology. The data reported in this study is from the National Drug Treatment Reporting System database in the Republic of Ireland. All analyses were carried out in collaboration with the Drug Misuse Research Division (the Irish REITOX / EMCDDA focal point). This database contains information on all 6994 individuals who received treatment for drug problems in the Republic of Ireland during 2000.

Analysis. The analysis was conducted in four steps. First, a single class model was examined in order to establish the respective probability associated with each drug type. Second, a series of unconditional latent class models was examined. This was done to establish the optimal number of latent classes required to describe the data and to establish the relative size of each latent class. From this analysis, the conditional probabilities for each individual, within a given class, was examined for typical profiles. Third, a series of conditional models were then examined in terms of key predictors, e.g., gender, age etc. This analysis was conducted using MPlus. In the final stage of the research, the parameter estimates obtained form the multinomial logistic regression model (that was previously used to express the probability of an individual being in a given latent class, conditional on a series of covariates) were graphically modelled within EXCEL and the respective functions described.

Conclusions. The results from this analysis will be described in terms of a) the profiling of typical serious drug misuse in Ireland, b) the clustering of drug types and, c) the respective importance of key background variables. The various profiles obtained are discussed in terms of health care strategies both in Ireland and among European counterparts.

The Geometry of Venice: Archaeological Intuition and Statistical Assessment

Luigi Calzavara (University of Padova, Italy)
Maurizio Brizzi (University of Bologna, Italy)

Archaeological intuition

Ordering one's environment has many advantages: it lowers the cost of protection and improves quality of life. To this end, landmarks, such as towers and belfries are key for both the observation and the transmission of visual information. In Venice, it is still obvious even today that the belfries were deliberately positioned at the vertex of particular geometrical forms, i.e. belfries were built according to

precise rules. The belfries of Venice were sited according to an ancient triangulation based on the right-angled triangle. The master builder had a practical method for determining the right angle of a building: he would take a ring of rope, divide it into twelve equal parts, with a knot to mark each section, and then pull it out with a divergent force at the third (3), seventh (3+4) and twelfth (3+4+5) knots, so creating a right-angled triangle.

From now on, the topographer too put this system to good use. This time, however, rings of rope or wooden rods were no longer used to build a right-angle but to trace the vertex points of a specific right-angled triangle. Each time, depending on the city's particular building requirements, the most suitable ring had to be devised. The most suitable ring links the three Pythagorean numbers, and hence the theoretical shape, with the most suitable measurement of the cord segment i.e. with the distance between the vertex points. In this way a bi-univocal correspondence was established between specific numerical values of the infinite number of Pythagorean triples and the infinite points on the plane. Measuring instruments are generally bivalent. They are employed to build but also to verify what has already been built. Our work over forty years with Pythagorean triangles has been to determine the different phases of building development in the city of Venice, and also to provide statistical evidence that the triangulation system was indeed the process employed.

Venice was built on marshland. Church belfries had to rest on labour intensive, and hence costly, foundations. For economic reasons therefore, belfries can be considered topographical reference points down through the centuries. Indeed it could be said that marshland induces immobilism. Furthermore, every belfry has a precise position on the Cartesian grid of the official Italian cartographic system (1). For simplicitys sake, I have limited my survey to 50 belfries linked to churches founded by the XI century (2). Any research into the Pythagorean triangles linking Venice's belfries must obviously focus on the tolerance of right angles. Too high a tolerance value would maximise the number of Pythagorean triangles but would preclude statistical significance. As a result, working on a trial and error basis, I have established an optimal tolerance of $+/-0.125^{\circ}$ (1/8 of a degree). Using this value, 61 Pythagorean triangles have been identified which most importantly, link some 48 of the 50 bell-towers in the survey. The result is a geometrical mesh of triangles each positioned as a function of the other. In this way, the bell-towers form a system because the topographical information passes from one to the other without interruption, as on an electrical circuit. On the basis of these findings, some perfectly measurable topographical archaeology data were prepared and the results analysed statistically. The first survey considers 19

belfries connected to churches founded by the IX century. A system is revealed based on four intermeshing Pythagorean triangles, where several vertex points converge. For example, triangles ABC; ABD; BEC; BEF (Figure 1) have a vertex B and a point C in common. Point C corresponds to the belfry of San Bortolomio at the Rialto bridge in the historic centre of the city. The second topographic system studied looked at the inclusion of the new belfry of San Simeon Grande putatively dating from the X century which became the vertex of two new right-angled triangles. In this way, triangle after triangle we have been able to detect a topographical mesh linking all fifty of the belfries studied. The system appears the fruit of human design: a precise road map for the founding of Venice.

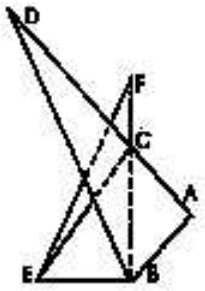


Figure 1: Triangles ABC, ABD, BEC, and BEF.

.

Statistical assessment

In order to evaluate the peculiarity of Venice belfries pattern, we have chosen a group of 19 belfries, located in the historical kernel of Venice, all of them built before 1000 AD., and included in a circle, centered in S.Bortolomio. We have considered all the $\binom{19}{3} = 969$ triangles generated by these belfries; using the topographic coordinates of each belfry, we calculated a quick index for detecting the Pythagorean triangles. If we indicate with a, b, c $(a > b > c)$ the ordered lengths of sides in a triangle, the Pythagorean Ratio $PR = \frac{b^2+c^2}{a^2}$ is equal to one when the main angle is right. We considered as Almost Pythagorean (AP) a triangle for which $0.998 < PR < 1.002$ holds, but we excluded the AP triangles when the smallest angle was less than $9°$. Among the 969 triangles of the whole set, 7 result to be AP, and they are strictly connected, having many vertices in common. The total number of connexions is 14, and every triangle is connected at least with 3 other ones. In order to check if this result may be due to random effects, we simulated a random location of 19 belfries in the same area, considered

the outcoming set of triangles and counted the number of AP triangles and their connexions. After repeating this simulation 3,000 times, we studied the simulated distribution of three statistics:

1. $N(AP)$ =the number of Almost Pythagorean triangles;
2. $N(C)$ = the number of connexions (links) between the triangles;
3. $\min(C)$ = the minimum number of connexions of each AP triangle.

The observed and simulated values are shown in the following table:

	$N(AP)$	$N(C)$	$\min(C)$
Observed (in real Venice)	7	14	3
Simulated mean	3.530	8.914	1.025
Simulated std. dev.	1.879	2.296	0.742
Standard value (observed)	1.847	2.215	2.662
% simulations greater than observed value	2.60%	1.63%	0.17%
% simulations equal to observed value	4.03%	1.94%	2.33%

In particular, the simulated distribution of $N(AP)$ follows almost perfectly a Poisson distribution with parameter $\lambda = 3.53$. If we consider the event 'the number of AP triangles is at least 7, their connexions are at least 14, and each triangle is connected with 3 or more other ones', which involves all the observed results simultaneously, it has a very low simulated frequency, approximately equal to 0.095% (1 out of 1,052). This result seems to suggest that these belfries, located in the ancient kernel of Venice, were not built at random, but rather following a pattern, likely for getting a better way of communication: in that period radio, TV and mobile phones were not invented yet...

References

1. Salzano, E. (Ed): Atlante di Venezia. Comune di Venezia e Marsilio Editori; Venezia,1989-91.
2. Dorigo, W. (1987): Venezia Origini. Electa editrice; Venezia.

Is It Possible to Protect Human Health?

Ida Camminatiello and Biagio Simonetti (University of Naples Federico II, Italy)

The aim of our survey is to study how environmental pollution and institutional capacity (independent variables collected in the X matrix) affect human health (dependent variables collected in the Y matrix).
For this purpose we analyse a data set of 58 variables (53 independent and 5 dependent) measured on 122 world country, the variables are presented in standardized form[1] . Studying the correlation between the variables, it's easy to notice that there a high correlation (sometimes close to 1) between many variables, so we can't apply the Ordinary Least Squares (OLS), but we have to use the PLS regression, to study the relationship between human vulnerability indicators and all the others indicators.
The PLS2 algorithm run by software SIMCA-P 8.0, carries out a model with three significant component that explain the 73% of Y variability an the 40% of X variability. It is interesting to notice that the first two components already explain the 70% of Y variability, so they allow us to have a quite complete view of the relationship between dependent variables and independent variables. In particular, the plot shows that the human vulnerability is greater in those countries where the total fertility rate, project growth rate (present to 2050), air and water pollution are high and water avail-ability per capita, investments are small. Global stewards are helping these countries.
Even though the results obtained by PLS regression are satisfactory, the use of Path modelling methods could describe better the relationship among the variables.

Multilevel Multitrait Multimethod Model

Lluis Coromina and Germà Coenders (University of Girona, Spain)
Tina Kogovšek (University of Ljubljana, Slovenia)

Our goal in this paper is to assess reliability and validity of egocentered network data using multilevel analysis (Muthén, 1989; Hox, 1993) under the multitrait-multimethod approach. The confirmatory factor analysis model for multitrait-multimethod data (Werts & Linn, 1970; Andrews, 1984) is used for our analyses. In this study we reanalyse a part of data of another study (Kogovšek, et al., 2002)

[1] World Economic Forum, Yale Center for Environmental Law and Policy, and CIESIN, 2001 Environmental Sustainability Index, January 2001.

done on a representative sample of the inhabitants of Ljubljana. The traits used in our article are the name interpreters. We consider egocentered network data as hierarchical; therefore a multilevel analysis is required. We use Muthén's partial maximum likelihood approach, called pseudobalanced solution (Muthén, 1989, 1990, 1994) which produces estimations close to maximum likelihood for large ego sample sizes (Hox & Mass, 2001).
Several analyses will be done in order to compare this multilevel analysis to classic methods of analysis such as the ones made in Kogovšek et al. (2002), who analysed the data only at group (ego) level considering averages of all alters within the ego.
We show that some of the results obtained by classic methods are biased and that multilevel analysis provides more detailed information that much enriches the interpretation of reliability and validity of hierarchical data. Within and between-ego reliabilities and validities and other related quality measures are defined, computed and interpreted.

Estimation of a Dynamic Structural Equation Model with Latent Variables
Dario Cziráky (Economics Department, IMO, Zagreb, Croatia)

The general structural equation model with latent variables, originally developed by Jöreskog (1973), has been extensively applied to cross-sectional data for the last three decades, yet time series generalisations are scarce in the literature mainly due to difficulties in estimating dynamic versions of the model.
This paper generalises the Jöreskog's model to a structural autoregressive distributed lag model with latent variables and proposes two general estimation procedures. It is shown how generalised instrumental variable estimation and three-stage least squares methods can be used to estimate dynamic latent variable models and the asymptotic properties of these estimators are described.
A *dynamic structural equation model with latent variables* (DSEM) is formulated as a time series generalisation of the static structural equation model with latent variables[2]

$$\eta_t = \alpha_\eta + \sum_{j=0}^{p} \mathbf{B}_j \eta_{t-j} + \sum_{j=0}^{q} \mathbf{\Gamma}_j \xi_{t-j} + \zeta_t, \tag{1}$$

[2]A static version of this model can be easily estimated by software packages such as LISREL 8.54 (see, e.g., Cziráky, 2003).

where α_η, $\mathbf{B}_0$, and $\mathbf{\Gamma}_0$ are coefficient matrices from the static Jöreskog's model, and $\mathbf{B}_1$, $\mathbf{B}_2$,..., $\mathbf{B}_p$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$,..., $\mathbf{\Gamma}_q$ are the additional $p+q$ matrices that contain coefficients of the lagged endogenous and exogenous latent variables.[3] If we assume time-invariance of the measurement model, the usual specification of the measurement models for $\mathbf{x}_t$ and $\mathbf{y}_t$ applies, thus the structural part of the model from Eq. (1) can be augmented with measurement equations for latent exogenous variables

$$\mathbf{x}_t = \alpha_x + \mathbf{\Lambda}_x \xi_t + \delta_t \tag{2}$$

and for latent endogenous variables

$$\mathbf{y}_t = \alpha_y + \mathbf{\Lambda}_y \eta_t + \varepsilon_t \tag{3}$$

The matrix equations (1)-(3) provide full specification of a general DSEM model directly extending the original Jöreskog's model to time series. It follows that classical (static) structural equation model is a special case of the DSEM model. However, the DSEM model from Eqs. (1)-(3) cannot be directly estimated due to the presence of unobserved latent components. To solve this problem and enable estimation of the model parameters, the latent variable specification is rewritten in terms of the observed variables and latent errors only, following the approach similar to Bollen (1996; 2001; 2002). Bollen developed such specification to enable non-parametric estimation of standard (cross-sectional) structural equation models with an aim of achieving greater robustness to misspecification and non-normality. In this paper it is shown that a similar approach can be used to re-write the DSEM model in the observed form specification (OFS) and to subsequently estimate all model parameters (except latent error terms) by generalised instrumental variables methods. The OFS form of the general DSEM model is given by the structural (matrix) equation

$$\mathbf{y}_{1t} = \alpha_\eta + \sum_{j=0}^{p} \mathbf{B}_j \mathbf{y}_{1t-j} + \sum_{j=0}^{q} \mathbf{\Gamma}_j \mathbf{x}_{1t-j} + \left(\zeta_t + \varepsilon_{1t} - \sum_{j=0}^{p} \mathbf{B}_j \varepsilon_{1t-j} - \sum_{j=0}^{q} \mathbf{\Gamma}_j \delta_{1t-j} \right), \tag{4}$$

and by the measurement models for the latent endogenous variables

[3] Note that Eq. (1) does not require specification of lagged latent variables as separate variables; rather each vector containing all modelled and exogenous latent variables is written for each included lag separately, with separate coefficient matrix. Also note that that Eq. (1) allows different lag lengths for different latent variables (i.e., elements of η and ξ vectors) by appropriate specification of $\mathbf{B}_j$ and $\mathbf{\Gamma}_j$ matrices (e.g., zero elements).

$$\mathbf{y}_{2t} = \alpha_2^{(\mathrm{y})} + \mathbf{\Lambda}_2^{(\mathrm{y})}\mathbf{y}_{1t} + (\varepsilon_{2t} - \mathbf{\Lambda}_2^{(\mathrm{y})}\varepsilon_{1t}), \tag{5}$$

and for the latent exogenous variables

$$\mathbf{x}_{2t} = \alpha_2^{(\mathrm{x})} + \mathbf{\Lambda}_2^{(\mathrm{x})}\mathbf{x}_{1t} + (\delta_{2t} - \mathbf{\Lambda}_2^{(\mathrm{x})}\delta_{1t}). \tag{6}$$

Aside of the specific structure of the latent error terms, the Eq. (4)-(6) present a classical structural equation system with observed variables.
The proposed estimation methods are asymptotically distribution free and thus require no specific assumptions about the joint density of the modelled variables. Additionally, a simple identification procedure is developed which can be used for both static and dynamic latent variable models.

References

1. Bollen, K.A. (1996): An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, **61**, 109-121.
2. Bollen, K.A. (2001): Two-stage least squares and latent variable models: simultaneous estimation and robustness to misspecification. In R. Cudeck, S. Du Toit, and D. Sörbom (Eds.): *Structural Equation Modeling: Present and Future*. Chicago: Scientific Software International, 119-138.
3. Bollen, K.A. (2002): A note on a two-stage least squares estimator for higher-order factor analysis. *Sociological Methods and Research*, **30**, 568-579.
4. Cziráky, D. (2003): LISREL 8.54: A programme for structural equation modelling with latent variables. *Journal of Applied Econometrics*, forthcoming.
5. Jöreskog, K.G. (1973): A general method for estimating a linear structural equation system. In A.S. Goldberger and O.D. Duncan (Eds.): *Structural Equation Models in the Social Sciences*. Chicago: Academic Press, 85-112.

Stochastic Preference Flow and Group Decision

Lavoslav Čaklović (University of Zagreb, Croatia)

Theory: Let us denote the set of alternatives by $S = \{a, b, c, \ldots\}$ and suppose that it is finite. The classical approach to stochastic preference pose the following question:

To each pair of alternatives a, b decision maker assigns probability p_{ab} of choosing a when offered the choice between a and b. We assume $p_{ab} + p_{ba} = 1$, with convention $p_{aa} = \frac{1}{2}$. Which condition should satisfy numbers p_{ab} to generate a value function V on the set of alternatives, i.e. such that

$$p_{ab} \geq \frac{1}{2} \Leftrightarrow V(a) \geq V(b).$$

A binary relation P on the set of alternatives is called *stochastic preference* if

$$aPb \Leftrightarrow p_{ab} \geq \frac{1}{2}.$$

It is well known that if stochastic preference satisfies the following consistency condition

$$\frac{p_{ab}}{p_{ba}} \cdot \frac{p_{bc}}{p_{cb}} = \frac{p_{ac}}{p_{ca}} \tag{1}$$

for all $a, b, c \in S$ then P is necessarily a weak order and function V is given by

$$V(b) := \frac{p_{ba}}{p_{ab}}.$$

In general, we introduce a notion of *stochastic flow*, obtained from stochastic preference by formula

$$\mathcal{F}_\alpha := \log \frac{p_{bc}}{p_{cb}}.$$

It is obvious that:

1. $\mathcal{F}$ is consistent, i.e.
$$\mathcal{F}_\gamma + \mathcal{F}_\alpha = \mathcal{F}_\beta$$
for each $\alpha = (c, b)$, $\beta = (c, a)$ and $\gamma = (a, b)$ such that $\gamma + \alpha = \beta$ if and only if p satisfies (1);

2. the stochastic flow $\mathcal{F} : \mathcal{A} \to \mathbb{R}$ is potential difference, i.e. $BX = \mathcal{F}$ where B is the incidence matrix of the directed graph associated to $\mathcal{F}$, if and only if p satisfies (1).

In general we define *stochastic normal integral* of stochastic preference flow $\mathcal{F}$ as a solution of normal equation

$$B^\tau BX = B^\tau \mathcal{F}, \sum_{i \in S} X_i = 0.$$

Normal integral of a given flow is not a value function in general. If stochastic preference is a weak preference then, the normal integral of unimodular stochastic flow is a value function. More precisely:

Theorem 1 *Let $\mathcal{F}$ be a complete unimodular flow and let us define a relation*

$$a \succcurlyeq b \Leftrightarrow \mathcal{F}_{(b,a)} \geq 0$$

with convention $a \succcurlyeq a,\ \forall a \in S$. If relation $\succcurlyeq$ is transitive then, the normal integral X of $\mathcal{F}$ is a value function, i.e. consistent with $\succcurlyeq$ in the sense

$$a \succcurlyeq b \Leftrightarrow X(b) - X(a) \geq 0.$$

This approach is applied on data collected from a web questionnaire where students were asked to give preference flows, for certain criteria, over the set of their lecturers. It is shown that, in that example, stochastic flow and group flow generate equivalent ranking. At last, we calculated the Condorcet's flow and Savage's value function associated to its unimodular flow. Ranking obtained from Condorcet's flow is not equivalent to ranking obtained from stochastic flow.
Application: Stochastic flow is a simplification of the consensus flow as described in the article L. Čaklović: *Graph Distance in Multicriteria Decision Making Context*. In A. Ferligoj and A. Mrvar (Eds.): Metodološki zvezki, **19**. Collecting data only for stochastic flow from a WEB questionnaire is less time consuming than asking for strength of a preference on a pair of alternatives. It is shown that, in the example from above article, stochastic flow and group flow from generate equivalent ranking.
At last, we calculated the Condorcet's flow and Savage's value function associated to its unimodular flow. Ranking obtained from Condorcet's flow is not equivalent to ranking obtained from stochastic flow.

Estimating Informant Accuracy for Missing Value Imputation in Complete Social Networks

Daniëlle De Lange, Filip Agneessens, and Hans Waege
(Ghent University, Belgium)

Collecting social network data means asking respondents a considerable amount of burdensome questions. Moreover, respondents often perceive these social network questions as being sensitive or threatening. The combination of this burdensome and sensitive nature is often responsible for high unit and item nonresponse.

Both types of non-response are major problems in social network analysis. Especially, when focussing on the measurement of complete social networks both types of nonresponse need to be minimised or properly managed. This paper meets the need for additional research efforts to deal with nonresponse for complete social networks.

A common practice among social network researchers is either to ignore the answers of nonrespondents (leaving the actors out of the analysis) or to replace the missing data by means of particular imputation procedures. This paper continues on the tradition of accuracy research (Bernard et al., 1977, 1980, 1982; Killworth et al., 1976, 1979; and Krackhardt, 1987, 1990). In current accuracy research, the question 'how accurate are respondents in perceiving the network of social relations surrounding them?' is essential. It concerns the respondents own social relations as well as the social relations between the other network members as perceived by the particular respondent. Subsequently, the accuracy of a respondent is measured by comparing the respondent's cognitive report about all social network relations with some measure of the 'actual' social network. This actual network is not measured directly, but is constructed following some specific criteria, either by taking the mean of all answers of all respondents (the Consensus structure) or by calculating the mean of the sender and receiver (the Local Aggregated Structure).

The basic idea of this paper is that when particular types of respondents are highly accurate in their perception of the social network, their reports could be used for missing value imputation. Reports from the most accurate network members are considered as a data source for the imputation of the missing answers. However, a sufficient understanding of informant accuracy and the detection of accurate respondents is a necessary prerequisite for this imputation procedure.

Previous accuracy literature concentrates on estimating the accuracy of respondents in a context of zero nonresponse and dichotomous data (i.e. the absence or presence of a relation). This is a highly specific situation, knowing that nonresponse is inherent to survey research and that working with valued data provides us with a richer source of information. In this paper, we propose a more general approach for the calculation of informant accuracy, taking into account nonresponse and the categorical nature of the data. This new accuracy measure will be used to answer the second, more substantive research question: 'to what extent is it possible to distinguish different types of respondents with varying levels of informant accuracy?'. Because of the use of the new accuracy measure, confirmation of earlier results is not straightforward. Both individual and social network characteristics will prove to be important to consider as determinants of informant accuracy. Subsequently, these results can be used for missing value imputation.

Modelling the Fractions of Informed and Uninformed Traders in Financial Markets

Giovanni De Luca (University of Naples Parthenope, Italy)

The availability of financial tick-by-tick (or ultra-high frequency) data has allowed a great number of studies aimed at analysing the financial market microstructure corroborating or not the different theories. Tick-by-tick data are irregularly spaced time series and in order to treat them specific econometric techniques have to be used. The Autoregressive Conditional Duration (ACD) type models were introduced by Engle and Russell (1998) and represent a standard way to model durations. Denoted t_i the time of the i-th market event, x_i is the i-th duration, that is $x_i = t_i - t_{i-1}$. In its general formulation, ACD models can be written as follows:

$$\begin{aligned} x_i &= \Psi_i \epsilon_i \\ \Psi_i &= f(x_{i-1}, \ldots, x_{i-q}, \Psi_{i-1}, \ldots, \Psi_{i-p}) \end{aligned}$$

with $\epsilon_i \sim$ iid and $E(\epsilon_i) = 1$.
Denoted $\mathcal{I}_{i-1}$ the information up to time t_{i-1},

$$E(x_i|\mathcal{I}_{i-1}) = \Psi_i,$$

so that Ψ_i is the expected duration conditionally on the information up to the time $i-1$ and is (conditionally) deterministic.
The financial market microstructure theories stress the presence of traders who possess a different degree of information. The notion of market efficiency implies a perfect equality in the knowledge possessed by traders about the future pattern of financial asset. This theoretical notion is often far from true. Actually, the traders have different degrees of information, according to their position or something else. A very simple assumption is the categorization of traders into two classes: informed and uninformed. This assumption is certainly a narrow view, but it allows one to simplify the statistical modelling.[4] A mixture of two distributions seems to be effective in modelling intradaily durations (De Luca, 2003). In this case the mixing proportions can be interpreted as the probabilities of observing a transaction carried out by one of the two types of agents. As pointed out by Ghysels (2000), a more interesting formulation should allow for time-varying probabilities of an informed or uniformed trade to arrive. Define p_{Ii} the fraction of informed traders at time t_i and λ_{Ii} their arrival rate. The arrival rate of informed

[4] A more subtle strategy is to *personalize* the degree of information. See De Luca and Zuccolotto (2002).

traders could be assumed to be temporally dependent, because of the tendency of informed traders to split up their trades, in order to avoid to lose their anonymity. On the other hand, the arrival rate of uninformed traders, λ_{Ui} can be assumed constant, $\lambda_{Ui} = \lambda_U$.
We tried to study the behavior of the time-varying fractions of informed and uninformed traders after selecting some variables. The variables candidate to explain the varying proportions of informed and uninformed traders were the trading intensity (TI) and the average volume per trade (AV).
The variable TI_i is given by the ratio of the number of trades recorded during the price duration x_i and the duration itself. So, the trading intensity increases when the number of trades is high and/or when the duration is short. For duration x_i, the variable AV_i is defined as the ratio of the traded volume over the number of transactions.
A higher than normal trading intensity may be indicative of the release of news, which involves the arrival of informed traders. The same can be said for a higher than normal average volume. So, the idea is to relate p_i to $TI_i/\bar{TI}$ and $AV_i/\bar{AV}$ where $\bar{TI}$ and $\bar{AV}$ represent, respectively, the average trading intensity and the average volume per trade.
The results show a substantial significativity of the trading intensity. On the contrary, the average volume per trade turns out to be not significant. The reason can be found out in the strategy of splitting trades by informed traders.

References

1. De Luca, G. (2003): Estimating the instantaneous volatility of the price process. Paper presented at the Meeting *Linearity and non linearity in economic and financial time series*, Brixen.
2. De Luca, G. and Zuccolotto, P. (2003): Finite and infinite mixture of financial durations, *Technical Report*, University of Brescia.
3. Engle, R.F. and Russell, J.E. (1998): Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, **66**.
4. Ghysels, E. (2000): Some econometric recipes for high-frequency data cooking. *Journa; of Business & Economic Statistics*, **18**.

Methodological Problems with Digital Divide Measuremets

Vesna Dolničar and Vasja Vehovar (University of Ljubljana, Slovenia)

Paper first discusses some specifics regarding terminology and concept of the notion of 'digital gap', which can be further structured into first, second, dual and

third digital divide. The notion of the digital divide is shown to be much more complex than just a distinction between 'haves' and 'have-nots'.

From methodological aspects, the digital divide is typically observed as a comparison of simple percentages related to the Internet penetration among different countries or socio-demographic segments (e.g. young, male, educated, rich, urban). However, such simplified approach has three major drawbacks: it can hide interactions among the variables, it oversimplifies the issues to one variable only and it neglects the component of time. More proper approach to digital divide measurement may thus give radically different conclusions compared to the simple bivariate analysis.

The paper elaborates these problems with an example from Slovenian data. More factors are studied simultaneously with respect to the use of the Internet. The results indicate that the interactions between variables seriously complicate the insights into relevant factors, e.g. the size of the settlement affects the Internet usage only when observed in the bivariate analysis, while its effect disappeared within multivariate context where more variables were studied simultaneously.

Next, a composed measure the Digital Divide Index (DIDIX) in SIBIS (Statistical Indicators Benchmarking Information Society) project - is discussed. This measure provides the integral digital divide measure for benchmarking EU countries and the US. The DIDIX index is constructed as a double average of the percentage difference for the usage of three technologies (computer usage, Internet usage, home Internet access) across four segments (gender, age, education and income). Such measure solves many problems of the simple measurements, however, some serious methodological problems appear here, too.

Finally, the dimension of time is briefly discussed. It is shown that the percentage differences (simple or compound) cannot properly reflect the time changes of the digital divide.

A Quick Procedure for Model Selection in the Case of Bivariate Normal Mixture

Alessandra Durio and Ennio Davide Isaia (University of Turin, Italy)

In many modeling situations, we often have a choiche. We can add to model complexity to describe some unusual observations or groups of observations or we can remove those observations and fit the rest with a different, perhaps simpler, model.

Purpose of this work is to illustrate a procedure in finding the number of the com-

ponents of a mixture of normal bivariate densities, exploiting in this way the properties of robustness of the estimates based on the Minimum Integrated Squared Error (or Minimum L2 distance), e.g. Scott (2001), Durio and Isaia (2003a), Basu et al. (1998). Every step of the procedure consists in the comparison between the estimates of the parameters of the mixture according to the Maximum Likelihood and Minimum L_2 distance criterions. If their difference is significant, then we change the model adding one more component to the mixture.
Theory is outlined and some examples and applications are presented. We shall also suggest a quick rule which allows us to assign data points to each component of the mixture, e.g. Durio and Isaia (2003b).
Software specifically designed to support the entire procedure has been implemented in R computing environment and it will be used to perform all the analysis proposed in the paper.

References

1. Basu, A., Harris, I.R., Hjort, N., and Jones, M. (1998): Robust and efficient estimation by minimizing a density power divergence, *Biometrika*, **85**, 549559.
2. Durio, A. and Isaia, E.D. (2003a): Parametric multivariate density estimation using l2 distance: a simulation study on robustness, S.I.S. Intermediate Scientific Meeting.
3. Durio, A. and Isaia, E.D. (2003b): A parametric regression model by minimum l2 criterion: a study on hydrocarbon pollution of electrical transformers. In A. Ferligoj and A. Mrvar (Eds): *Developments in Applied Statistics*, Metodološki zvezki, **19**, 69-83.
4. Scott, D.W. (2001): Parametric statistical modeling by minimum integrated square error, *Technometrics*, **43**, 274-285.

Stated and Experienced Availability as Clues for Call-Scheduling in Households Longitudinal Cati Surveys

Luigi Fabbris and Maria Cristiana Martini (University of Padua, Italy)

In a telephone survey a large amount of time and money is spent in attempts to contact the sample units and obtain valid and complete interviews. Computer-assisted interviewing allows the implementation of call scheduling procedures aimed at maximising the number of successful attempts. Successfulness of a call attempt consists in:

1. contacting the potential respondent;
2. obtaining a complete interview.

In order to arrange an efficient call scheduling procedure, we need estimates for the probabilities $P(y_{ij})$ of contacting and obtaining a complete interview from the sample unit i, at the time slot j, on the basis of the available information X_{ij}:

$$\hat{P}(y_{ij}) = f(X_{ij})$$

Then, at each time slot, the call scheduling procedure will select first the sample units with the highest probability of success, in absolute terms or compared to the average probability of any other sample unit i over all possible time slots.
Compared to a standard survey, longitudinal surveys provide further information which can be used to predict the presence at home of the potential respondents and their will to collaborate.
In particular, besides the general information that is available for every kind of survey, such as the days of the week and the time of the day usually preferable to obtain an interview, we can use information about:

- household characteristics collected during the previous interview, which allow us to forecast the presence at home of each family, and to tailor the call scheduling to that particular kind of household;
- the call history recorded during the previous survey wave, which outlines the dynamics of presence at home of the household members, and their differential availability across time and days;
- specific questions asked during the previous interview about, on the one side, optimal days and time for further contacts, and, on the other side, time slots in which the contacted person is unavailable to respond.

In this contribution, we analyse data from a rotated panel survey on risk factors of social uneasiness at family level, conducted by the Inter-institutional Permanent Observatory on Padua Families, constituted by the Padua Local Health Service in co-operation with the Statistics Department of the University of Padua. In the first wave, a random sample of 1200 households in the municipality of Padua has been interviewed between October 15th 2001 and February 15th 2002, while in the second wave the rotated sample included 600 of the previously interviewed families, who have been contacted again between April 15th 2002 and July 15th 2002.
Alternative estimation models for the probability of obtaining a valid interview are fitted on the basis of the three different levels of information collected during the

first wave of the survey: household characteristics, call history and respondents stated availability. These models are then assessed and compared by means of the call histories recorded in the second wave.

Item Non-Response Reminder in a Web Survey
Marek Fuchs (Cath. University Eichstaett-Ingolstadt, Germany)

Web Surveys are rapidly becoming state of the art in self administered surveys. Interactive Web Surveysone question at a time is presented on a screen and server contact is made after each screenapply different approaches to deal with item non-response: Some surveys require all items to be answered before the respondent is allowed to move on. Others allow respondents to let proceed thru the questionnaire with some items being unanswered.
Both strategies have advantages:

- Some scholars argue that respondents should be given the opportunity to go thru the questionnaire without being required to answer each and every question. It is assumed that requiring respondent to answer all items increases the respondent burden and thus leads to a higher proportion of abandoned interviews.

- Others point out, that allowing respondents to refuse single items keeps reluctant respondents interested in the survey. It is assumed that they get more involved with the questionnaire once they have scanned the first few pages of the survey and found it easy and interesting.

In this paper we will report results from a field experiment conducted in January of 2003 in Germany. 3000 respondents were randomly assigned to three conditions: Group 1 was required to answer all items in order to move on thru the questionnaire. Group 2 was allowed to move thru the questionnaire even when an item was unanswered. Group3 was reminded of the unanswered item but allowed to move on.
Results indicate that requiring respondents to answer all items right at the beginning of an surveys leads to less item non-response during this phase of the survey. However, in the long run it produces also more abandoned interviews. By contrast, allowing respondents to skip over the first few items involves respondents more pronounced. As a result this group shows a smaller proportion of abandoned interview.

A Comparison of Hierarchical Clustering for Interval and Binary Data

Petra Golob, Nataša Kejžar, and Aleš Žiberna (University of Ljubljana, Slovenia)

The results of a study that is going to be presented in this presentation try to answer the following question: "Using the ordinal scales - when (if ever) does the hierarchical clustering algorithm for binary data outperform the hierarchical clustering algorithm for interval data?". The question is strongly connected to social sciences when using questionnaires with ordinal scales. The results could help us differentiate among the answers to the questions of questionnaires that can be considered as the scale variables and those better considered as binary variables. The initial question was narrowed to a very specific case where only one dataset design was used.

The dataset consisted of 3 groups, each with its own multivariate distribution (3 variables) with the known means and covariances. The groups were of unequal sizes and did slightly overlap. From this design several datasets were simulated. Each variable was cut and recoded in a way to achieve the ordinal scale. Different cutting schemes were used (the intervals were of equal sizes, did increase or decrease from the lowest value or were decreasing from the mean to both extremes). Interval data were created. When using the algorithm for classifying of binary data, dummy variables for each category were created. On these new variables the hierarchical clustering algorithm for interval and binary data were used. Ward's algorithm with Squared Euclidean distance was used when data were considered interval and Ward's algorithm with Jaccard distance when they were considered binary.

The quality of the results was assessed by comparing the gained 3 partitions with the three original groups and those gained from clustering the original (uncut) data. The comparison was made using Rand statistics, original and corrected for chance.

Preliminary results indicate that the clustering algorithm for interval data should usually be the preferred choice, but the difference diminishes when cutting into a lower number of intervals.

Allocation Issues in a Survey of Vehicle Speeds

Annica Isaksson (Linköping University, Sweden)

In a Swedish vehicle speed survey, data are collected for a stratified three-stage sample of road sites. Our concern here is whether the current allocation of the

total sample over sampling stages is the most efficient, or if there is room for improvements. The parameter of main interest is the average speed, R, on the roads. In order to evaluate the present sample allocation, we estimate the components, arising from each sampling stage, of the total variance of the estimator of R. The sampling design is such, that in all stages but the first one, only one sampling unit per stratum is selected. This makes the variance contributions from the first and second sampling stage inseparable. We circumvent this problem by utilizing a 'fictitious' sampling design and some experimental data. In this way, the demanded variance component estimates are calculated for a domain of study.

Our results indicate that for an unchanged total sample size, the precision of the estimator of R would improve if the sample sizes in stage three were increased, and the sample size in stage one decreased correspondingly. Thus, re-allocation of the sample seems to be worth while.

An Analysis of Multivariate Repeated Measures Experiment Design Evaluating the Influence of Gibberellic Acid on Cherry Fruit Quality

Damijana Kastelec and Valentina Usenik-Blazinšek
(University of Ljubljana, Slovenia)

Two between-unit factors and one within-unit factor repeated measures experiment design was used to examine the influence of gibberellic acid on cherry fruit quality. The first one of the two between-unit factors had three levels represented by three varieties of cherries ('Van', 'Sunburst', 'Elise') and the second one had two levels: spraying with gibberellic acid and spraying with pure water as a control. The 3 dependent variables, which describe ripeness of cherries (brightness, intensity of red-green colour and intensity of yellow-green colour) were measured with a chromometer over 6 time points at three or four day intervals during maturation. The experiment was done in 3 replications. One experimental unit was a cherry branch on which 20 cherries were sampled randomly and marked. Measurements were done for each of twenty cherries at each time point. After picking the cherries, measurements of fruit firmness were made on 3 different sites of each marked cherry. The multivariate repeated measures ANOVA was used for the analysis of two sets of data. When we examined the ripeness of fruit as a dependent variable the time of measurements was taken as within-unit factor and in the case when we analysed the fruit firmness as dependent variable, the site of measurements was taken as within-unit factor. The results showed that for all three dependent variables describing fruit ripeness there is no significant interac-

tion between the variety of cherries and the treatment with gibberellic acid, but there exists significant main effect of varieties and also significant main effect of treatment with gibberellic acid. The main effect of time and the interactions between time and varieties and time and treatment were significant and the analysis of polynomial contrasts were used to explain the differences between mean values of dependent variables describing fruit ripeness through the observed time interval. In the case of fruit firmness the assumption of sphericity of covariance matrix can be accepted. Therefore the univariate repeated measures ANOVA can be used instead of multivariate repeated measures ANOVA. The comparison of both results is presented.

Using Subgroup Discovery for Analyzing the Reputation of Selected Slovenian Companies

Branko Kavšek, Nada Lavrač, and Bojan Cestnik
(Jožef Stefan Institute, Ljubljana, Slovenia)

Subgroup discovery, a relatively new technique in machine learning, is aimed at searching for interesting subgroups in data. As 'interesting' is a very subjective term, subgroup discovery is particularly suitable for solving problems involving a human expert. We developed a new algorithm for subgroup discovery - APRIORI-SD - that is aimed at helping the human expert to describe his needs and thus defining what is really interesting for him. We adapted the well known association rule learning algorithm APRIORI by adding descriptive parameters which can help the human expert in giving his definition of interestingness. We then applied both APRIORI and our APRIORI-SD algorithms (in collaboration with the human expert) on the problem of analyzing the reputation of selected slovenian companies from the data of a telephonic survey. We compared the results of the two algorithms first by comparing their descriptive performance measures (number of subgroups, coverage of subgroups, overlap of subgroups, ...) and second by showing the resulting subgroups to the human expert. In both cases our algorithm outperformed APRIORI showing the suitability of the subgroup discovery approach in solving tasks involving the evaluation of a human expert.

Collecting Data on Ego-Centered Social Networks on the Web

Gašper Koren, Katja Lozar Manfreda, Vasja Vehovar, and Valentina Hlebec
(University of Ljubljana, Slovenia)

Owing to complex structure of questionnaires on ego-centered social networks interviewer assisted data collection modes are usually used. Web surveys, on the other hand, are offering variety of efficient solutions with sophisticated questionnaire interfaces and time-and-cost effectiveness. As shown in our experiments, single name generators for evaluating ego-centered networks can be effectively applied also within Web surveys, but visual elements of the measurement instrument have to be designed very carefully.
In this study several experiments with varying instructions for name generators and space provided for listing alters were applied within Web questionnaire. Visual characteristics of measurement instrument and effects of experimental instructions on number of listed alters and composition of the network as well as dropout rate, one of the main concerns of self-administered data collection modes, were studied.
For collecting the data on alters two different techniques for assessing name interpreters were applied: alter-vise (complete information on name interpreters is obtained for each alter separately and the variable context is changing) and variable-vise (complete information on name interpreters is obtained for each variable and the alter context is changing).

Probability Distribution for x-axis Intercept in Linear Regression

Katarina Košmelj, Anton Cedilnik, and Andrej Blejec
(University of Ljubljana, Slovenia)

A particular nonlinear relationship describing a chemical dependence can be transformed into a standard linear regression of the form

$$Y = \alpha + \beta X + \epsilon$$

In this setting, x-axis intercept has a specific meaning (concentration) and is of major importance for the experimenters. They are interested in the confidence interval for this intercept.
To solve this problem, the experimenters use the upper and lower hyperbola presenting the endpoints for confidence intervals for predicting $E(Y|x)$. They consider the interval defined by the intercepts of the two hyperbolas with x-axis as

the confidence interval for the x-axis intercept. This approach is commonly used, however it is wrong. The obtained interval is not the confidence interval for the x-axis intercept but is the 'fiducial interval' for X at $Y = 0$.
We are interested in the probability distribution of the random variable

$$X_0 = -A/B$$

which describes the intercept; A and B denote the estimators for α and β. X_0 is expressed as the ratio of two normally distributed and dependent random variables, their distributions and dependence are known from the regression theory.
It is well known that the ratio of two independent unit normal distributions follows the standard Cauchy distribution which is symmetric around 0. Cauchy distribution does not posses finite moments of order greater or equal 1, and so it does not posses a finite expected value or standard deviation.
We derived the distribution for the ratio of two random variables following a general bivariate normal distribution. The obtained distribution is asymmetric, however of Cauchy type. Applying this result to linear regression we obtained the analytical form for the probability distribution for x-axis intercept. The analytical results were compared with the results obtained by extensive Monte-Carlo simulations. The confidence interval for the intercept was obtained on the basis of the analytical density function as well as on the basis of the simulated empirical distribution functions.

Direction of Answering Categories: Does It Make a Difference in Response Behavior?
Dagmar Krebs (University of Giessen, Germany)
Jürgen H.P. Hoffmeyer-Zlotnik (ZUMA, Mannheim, Germany)

Should a scale offering answering categories to respondents start on the right side with the positive pole or with the negative pole? Does it make a difference in response behavior whether the positive or negative pole comes first?
Based on a split-ballot design, the paper approaches this question and tries to give a solution. Respondents were asked to answer different attitude questions by using an 8-point scale offering increasing (split 1) or decreasing (split 2) intensity of how much an item applies to the feeling of the respondent. Only the extreme points of the 8-point scale were labeled. For one group of items the answering categories were labeled as does not at all apply and applies completely (split 1) versus applies completely and does not at all apply (split 2). For a second group of

items answering categories were labeled as not at all important and very important (split 1) versus very important and not at all important (split 2).

The most general hypothesis is that response behavior differs depending on the direction of the answering categories. This very general hypothesis postulates that respondents answer in the direction of reading from left to right and tend to mark (check) answering categories that are more to the left. Thus, if the direction of answering categories is from negative to positive (split 1), the proportion of negative answers is expected to be higher than under the condition where the direction of answering categories is from positive to negative (split 2), where the proportion of positive answers is expected to be higher.

The results show that, depending on item content, there are exceptions to this postulated general response behavior. If items are of a rather unspecific content or describe rather general states of mind then distributions (as well as means) do not differ depending on split versions 1 or 2. Additionally, if items contain statements referring to a state of mind that can be described as external control then again no difference between the two split versions can be observed.

If, however, items are of specific content and additionally describe a state of mind that can be described as internal control the direction of answering categories produces a difference in response behavior resulting in different distributions (as well as different means) between the two spilt versions. Although the occurrence of differences seems to support the general hypothesis, this is true only at first glance. The observed differences between the two split versions reveal that responses are more positive if the direction is from positive to negative (split 2). If the direction of answering categories is, however, from negative to positive (split 1), responses are not more negative (with higher proportions of responses on the negative end of the scale) but they are less positive. Although there are differences in response behavior, these do not systematically reflect the reading direction from left to right.

Additionally, differences in response behavior between the two split versions occur systematically in combination with the labels of the answering categories (applies completely / does not at all apply on the one hand side and very important / not at all important on the other hand side). If the item content is judged on the dimension of importance, the observed differences again show the bias into the positive direction as described above. If items are judged on the apply-not apply dimension, differences in response behavior are nearly not observable.

Consequences of these observations together with a conclusion will be given in the conference presentation.

Component Analysis of Multivariate Spatial and Temporal Data

Wojtek J. Krzanowski (University of Exeter, UK)

Multivariate measurements are often made in either a spatial or a temporal context. Examples of the former occur in environmental reconnaissance programs that are carried out in many countries, where measurements are taken on a large set of different chemical elements at various locations throughout the study area. An example of the latter occurs in industrial process control, where a range of output measures is monitored on a regular basis, such as daily or weekly, across a period of time.

Principal component analysis (PCA) is often used on such data sets to identify important combinations of the original variables, either as a focus for more detailed study or to explain any significant outcomes shown up by omnibus tests. However, while PCA and related projection techniques from the standard multivariate repertoire are optimal under independence of observations and random sampling, they are not explicitly designed to address or to exploit the strong auto-correlation and cross-correlation structures that are often present in multivariate spatial or temporal data. Consequently their use may produce misleading results.

In this talk, several alternative projection techniques that are tailored to multivariate spatial and temporal data will be introduced and described. Like PCA these methods linearly transform the original variables into uncorrelated components, but instead of maximising variance these components are designed to have particular spatial or temporal properties. They can thus be viewed as optimal components to use in a spatial or temporal setting.

The technical derivation of the methods will be presented, their general performances and properties will be demonstrated via simulation results, and their specific applications will be illustrated on several real data sets. The advantages of the new methods over the more traditional ones such as PCA will be highlighted in these examples.

Effects of Sampling Mode on Factor Loadings

Bojan Leskošek and Gašper Koren (University of Ljubljana, Slovenia)
Rudi Seljak (Statistical Office of Republic of Slovenia, Ljubljana, Slovenia)

The aim of study is to investigate the effects of different modes of sampling on the results of factor analysis. Two-stage cluster sampling is compared to simple

random sampling. Using maximum-likelihood factor analysis with 11 variables with real-world correlation matrix the bias, variability and distribution shape of loadings on the first factor were compared.

On How to Establish Preference (Joint) Scales by Means of Multiple Unidimensional Unfolding

Herbert Matschinger and Matthias C. Angermeyer
(University of Leipzig, Germany)

The exploration of public preferences concerning the financial efforts in health care is of vital importance particularly for health politics. For which kind of a disease expenses are either accepted or disapproved by the general public provides the basis for concerted efforts for allocation and the distribution of means. The focus of this study is to explore whether or not such preferences or a preference order really does exist and how it could be established.

In 2001, a representative survey was conducted among the adult population of Germany ($n = 5025$). The respondents were presented with a list of nine different diseases, including medical conditions like cancer, cardiovascular disease, diabetes, rheumatism and AIDS as well as mental disorders such as Alzheimer's disease, schizophrenia, depression and alcoholism. Then, they were asked: 'Please select the three diseases for which, according to your opinion, available resources should on no account be shortened'. By restricting the number of options, we intended to force the respondents to really make a choice, otherwise the respondents would have been allowed to choose either all or none of the items. In a second question they were asked to choose those three out of the same set for which expenses preferably should be cut down, since we must assume that the preference of certain stimuli not necessarily indicates exactly the opposite of rejecting the same item. Comparing the results for both questions sheds light on the possibility to elicit and measure latent preferences.

In the first step we examined whether an underlying continuum can be identified representing both the various diseases and the preferences of the respondents for both questions. This leads to the problem of picking three out of nine dichotomous items equally available to all respondents which can be solved by means of multiple unidimensional unfolding (Coombs, 1964; Post, 1992; van Schuur, 1993; van Schuur and Kiers, 1994). However, in order to be able to apply this ordinal scaling technique, the following two assumptions must be fulfilled: (1) All subjects agree about the location of the stimuli (diseases) along the latent preference

dimension, although they may differ with respect to their individual preferences. (2) All subjects choose only those diseases that are close to each other on the continuum. These assumptions clearly mark the main differences between the better known cumulative scales (e.g., Guttman scale) and so called unfolding scales. The probability of a disease to be selected by a respondent decreases as the distance between the respondent's 'ideal point' on the scale increases; the relation between the latent preference dimension and the probability of a particular disease to be selected is single peaked and not monotonic as usually assumed in Guttman scales. The analysis was carried out using the programme MUDFOLD 4.0 (van Schuur and Post, 1998). Results clearly show that a sound preference order a joint scale for both the diseases and the respondents - can only be established if a real preference is asked for. Surprisingly enough a plausible, general preference order of diseases for the question: ...' available resources should on no account be shortened' does exist. Of the nine conditions, six could be arranged along a preference scale. One pole of the scale was made up by diabetes, the other one by alcoholism. Between these two poles, there were cardiovascular disease and cancer on one side and schizophrenia on the other side. AIDS constituted the centre of the scale. The diseases were not ordered on the scale according to how frequently they were chosen, but according to the respondents' preferences with respect to the latent dimension. The 2nd question (cut down expenses) does not provide an unfolding scale which can be employed for further analysis.

References

1. Coombs, C.H. (1964): *A Theory of Data*. New York, London, Sidney: Wiley.
2. Post, W.J. (1992): *Nonparametric Unfolding Models*. Leiden: DSWO Press.
3. van Schuur, W.H. (1993): Nonparametrical unidimensional unfolding for multicategory data. *Political Analysis*, **4**, 41-74.
4. van Schuur, W.H. and Kiers, A.L.H. (1994): Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement*, **18**, 97-110.
5. van Schuur, W.H. and Post, W.J. (1998): MUDFOLD – A program for Multiple Unidimensional Unfolding Version 4.0. Groningen: iec ProGAMMA.

Latent Class Models for Studying Measurement-Related Mode Effects in Mixed-Mode Surveys

Allan L. McCutcheon (Gallup Research Center, USA)

Mixed-mode survey data collection plays an increasing role in modern survey research. Telephone surveys and self-completion–both traditional paper and pencil, as well as the more recent internet-based–questionnaires often complement face-to-face surveys, in an effort to contain costs (e.g., Biemer and Lyberg, 2002). Also, mixed-mode data collection strategies have been used to improve response rates (e.g., Shettle and Mooney, 1999). For example, those who fail to respond to mail-out or internet questionnaires, may be called by telephone to improve completion rates.

The mixing of survey data collection modes, however, include a number of problematic issues. Recent research evidence clearly indicates that there exist substantial differences in response rates between modes of data collection (e.g., Dillman et al., 2001; Dillman 2000). Thus, the differential response rates across modes of data collection may give rise to differential response distributions in the measures of interest. Yet, even controlling for these differential response rates, among respondents who do participate in surveys, it is clear that differential modes of data collection lead to differential response patterns. De Leeuw and van der Zouwen (1988), for example, report that telephone surveys are less subject to social desirability bias than are face-to-face surveys. Similarly, Tourangeau and Smith (1998) report a substantial degree of mode-induced measurement bias in responses to sensitive questions.

Saris and Hagenaars (1997) propose using the latent class model (LCM) to examine mode-induced measurement error. This paper extends their use of LCMs to examine a number of plausible models for testing specific hypotheses regarding the nature of mode effects. The paper demonstrates the use of LCMs for examining mode effects for nominal and ordinal level measures. Data from a variety of public-release data sources, as well as recently collected mode effect experimental data, are used to illustrate the utility of this approach.

References

1. Biemer, P.P. and Lyberg, L.E. (2002): *Introduction to Survey Quality*. New York: Wiley.
2. De Leeuw, E. and van der Zouwen, J. (1988): Data quality in telephone and face-to-face surveys: A comparative analysis. In R. Groves et al. (Eds): *Telephone Survey Methodology*. New York: Wiley.

3. Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., and Berck, J. (2001): Response rates and measurement differences in mixed-mode surveys: Using mail, telephone, interactive voice response, and the internet. Paper presented at the American Association for Public Opinion Research meeting.
4. Dillman, D.A. (2000): *Mail and Internet Surveys*. New York: Wiley.
5. Saris, W.E. and Hagenaars, J.A. (1997): Mode effects in the standard Eurobarometer questions. In W.E. Saris and M. Kaase (Eds.): *Eurobarometer: Measurement Instruments for Opinions in Europe*. ZUMA Special Band 2. ZUMA: Mannheim, Germany.
6. Shettle, C. and Mooney, G. (1999): Evaluation of using monetary incentives in a government survey. *Journal of Official Statistics.*
7. Tourangeau, R. and Smith, T.W. (1998): Collecting sensitive information with different modes of data collection. In M.P. Couper et al. (Eds): *Computer Assisted Survey Information Collection*. New York: Willey.

Analysis of Genealogies with Pajek

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

Genealogies can be represented as graphs in several different ways: as *Ore graphs*, as *p-graphs* or *bipartite p-graphs*. p-graphs are more suitable for analyses and visualization.
Some approaches to analysis of large genealogies implemented in program Pajek are presented. A special emphasis is given to the relinking patterns in genealogies. These approaches will be illustrated with analyses of some typical large genealogies.

References

1. Dremelj, P., Mrvar, A., and Batagelj, V. (2002): Analiza rodoslova dubrovačkog vlasteoskog kruga pomoču programa Pajek. *Anali Zavoda povij. znan. Hrvat. akad. znan. umjet. Dubr.* **40**, 105-126.
2. White, D.R., Batagelj, V., and Mrvar, A. (1999): Analyzing Large Kinship and Marriage Networks with Pgraph and Pajek. *Social Science Computer Review – SSCORE*, **17**, 245-274.
3. White, D.R. and Jorion, P. (1996): Kinship Networks and Discrete Structure Theory: Applications and Implications. *Social Networks*, **18**, 267-314.

How the Measure Process can Influence the Evaluation of Public Services: A Case Study

Laura Pagani (University of Udine, Italy)
M. Chiara Zanarotti (Università Cattolica del S.C. of Milan, Italy)

The comparative assessment of public sector activities (typically education, health, social services), from the social and political perspectives, is an increasingly important problem. If focus is on evaluation of 'effectiveness', one approach is to evaluate the 'appreciation' i.e. the perceived quality of the user of a public services (Customer Satisfaction). Usually primary information are elicited via questionnaires, where the possible responses to single items are dichotomous or polytomous: this kind of data reflects the aim at measuring quality or another latent variable. Several disturbing factors affect the quality of the measure process: the individual perceptions, the relevance that every individual assigns to different aspects of the service, the differences in expected quality, the measurement scale adopted for data collection. The paper focuses on the latter mentioned trouble, with particular reference to role of arbitrary response categories. In general the responses are coded assigning numerical scores (*interval scale*) to categories (*ordinal scale*) but the coding reflects only ordinality, not intensity. Although the use of these arbitrary scores undoubtedly simplifies the analysis, inference can be unduly influenced by the initial scoring system. The aim of this paper is thus to compare the results obtained changing measurement scales differing in the number and semantic differential of the categories, using different models for Customer Satisfaction. The models considered are: multilevel models (for continuous and ordered response variables) and Rasch models. In particular, we want to test the robustness, that is the stability of the results under variations of the scale. A data set in evaluating teachers' effectiveness in University courses is used for this purpose.

References

1. Bertoli-Barsotti, L. and Franzoni, S. (2001): Analisi della soddisfazione del paziente in una struttura sanitaria: un caso di studio, Università Cattolica del S.C., Serie E.P. 104, Milano, Italy.
2. Bond, T.G. and Fox, C.M. (2001): *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, London: Lawrence Erlbaum Associates, Publishers.
3. Fischer, G.H. and Molenaar, I.W. (1995): *Rasch Models. Foundations, Recent Developments, and Applications*, New York: Springer-Verlag.

4. Goldstein, H. and Spiegelhalter, D.J. (1996): League tables and their limitations: Statistical issues in comparison of institutional performance (with discussion). *Journal of the Royal Statistical Society*, **159**, 385-443.
5. Snijders, T. and Bosker, R. (1999): *Multilevel Analysis. An Introduction to Basic and Advanced multilevel Modelling*. London: Sage Publications.

Reasons for Corporate Cross-Holdings

Marko Pahor (University of Ljubljana, Slovenia)

Research has found evidence that companies don't act independently from each other, but that they tie with other companies forming groups (e.g., Granovetter, 1995). Studies in the past examined different types of ties between firm, the bulk of the research was however concentrated on studying interlocking directorates and alliance networks. This research however concentrates on ownership relations, and in particular direct ownership, that are arguably a much stronger type of relations than other types of relations usually studied between companies (see, e.g., Kogut and Walker, 2001).

I describe and test the reasons, that lead to the emergence of cross-ownership ties between non-financial companies. Whereas the reasons for establishing controlling ownership ties, i.e. mergers and acquisitions, are well covered in the existing literature, much less can be found about the reasons for minority cross-holdings. I argue that under certain conditions minority and controlling holdings can be substitutes for each other. Thus, I propose and test the reasons for both types of holdings at the same time. In this I lean on the work of Watts (2003) who argues that ties emerge because actors of a network are physically socially close in one or another dimension. I apply the same argument for the cross holdings network between companies and propose that ownership ties between companies emerge in settings, in which these companies are embedded. Overlap in settings will foster the emergence of clusters, while strategic behavior of some companies in selecting the settings and ties within those settings will lead to the creation of cluster-spanning ties.

I test the proposition on a sample of 413 non-financial joint-stock companies in Slovenia. I use five observations of the complete directed network between companies, six months apart, starting with January 2000. I apply the methodology for the statistical evaluation of social network dynamics (e.g., Snijders, 2001) and test, what are the effects that foster the emergence of ownership ties between companies. Empirical findings confirm the theoretical expectations. Settings, that

tie companies together (industry, geographical closeness, common owner) greatly increase the probability of an ownership tie being created. Also, tendency to reciprocate the relations, formation of transitive triplets and the aversion towards indirect relations promotes clustering and leads to the creation of a small-world network.

A Behaviour Study of K-Means Method When the Starting Conditions are Changed: A Simulation Study

Anja Pajtler, Rok Podgornik, and Marijan Zafred
(University of Ljubljana, Slovenia)

The report consists of two parts. The first part is focused on theoretical foundations of clustering, especially on method of large data sets clustering (k-means method). In the experimental part of the work the data sets with known structure in three dimensional space were simulated. Each set consists of three or two groups where centroids were kept identical, while the dispersions ascended for three different structures. On simulated data sets we then analyzed the effectiveness and stability of k-means method, which was measured by several criteria. The results were verbally explained and also graphically presented.

Harmonic Markov Switching Autoregressive Models for Bayesian Analysis of Air Pollution

Roberta Paroli (Università Cattolica, Milano, Italy)
Spezia Luigi (Athens University of Economics and Business, Greece)

Markov switching autoregressive models (MSARMs) are efficient tools to analyse non linear and non Gaussian time series. A special MSARM with a harmonic component in the Bayesian framework is here proposed to analyse periodic time series. We present a complete Gibbs sampling algorithm for model choice (the selection of the autoregressive order and of the cardinality of the hidden Markov chain state-space), for constraint identification (the research of the identifiability constraints which respect the geometry and the shape of the posterior distribution) and for the estimation of the unknown parameters and the latent data. These three consecutive steps are developed tackling the problem of the hidden states labeling by means of random permutation sampling and constrained permutation sampling. We illustrate our methodology with two examples about the dynamics

of air pollutants.

Graphical Gaussian Models with Latent Variables: Interpretations, Estimation Algorithms and Identifiability Problems

Fulvia Pennoni and Giovanni Marchetti (University of Florence, Italy)

Linear triangular Gaussian systems

We consider a set of nodes N, begin completely ordered as $N = (1, ..., d_V)$ having node i corresponding to a random variable X_i so that each variable in the ordering is potentially a response variable for the proceeding one and explanatory for the following. The joint density f_V can be factorized into d_V univariate conditional densities:

$$f_V = \prod_i f_{i|i+1,....,d_V} = \prod_i f_{i|par_i}$$

If we consider a mean centered Gaussian column vector variable X of dimension d_v, the system of equations is

$$AX = \epsilon$$

where A is $d \times d$ upper triangular matrix with ones along the diagonal and $a_{ij} = -\beta_{ij.par(i)\setminus j}$ are partial regression coefficients; $Cov(\epsilon) = \Delta$ is a diagonal matrix with elements of partial variances $\delta_{ii} = \sigma_{ii.par(i)} = (\sigma^{rr})^{-1}$ along the diagonal. The ith row specifies X_i via a linear least squares regression on $X_{i+1}, ..., X_{d_v}$ with the residual uncorrelated with these latter variables. The covariance matrix of X and its inverse are

$$\Sigma = B\Delta B' \qquad \Sigma^{-1} = A'\Delta^{-1}A$$

where $B = A^{-1}$ is upper triangular, (A, Δ^{-1}) is a triangular decomposition of the concentration matrix. When a unique full ordering is provided from substantive knowledge the generating process determines uniquely a $p \times p$ indicator matrix $E = [e_{ij}]$ representing the set of edges with $e_{ij} = 1$ if in the Directed Acyclic Graph (DAG) there is an arrow between $i \longleftarrow j$, zero otherwise. Thus E has the same structure of zeros as A.

It can be seen that this type of DAG is also an elegant and informative graphical representation for causal relations.

Given a random sample of size n $X_1^1, ..., X_n^n$ from X The log-likelihood of the model is

$$l(\Sigma; X) = \frac{n}{2}[\ln|\Sigma^{-1}| - \mathrm{tr}(\Sigma^{-1}S)] = -\frac{n}{2} \quad [\sum_r ln\delta_{ii} + tr(A'\Delta^{-1}AS) \quad]$$

where $S = (\sum_i X_i X_i')/n = [\hat{\sigma_{ij}}]$

Unobserved variables
Supposing that we observe only a subset $Y = (Y_1, ..., Y_p)$ of the variables, thus the system can be seen as $X = (Y, Z)$ where Y denotes the observed components of X and Z denotes the unobserved components.
The EM algorithm is a method for solving the likelihood equations when the model is identified. Criteria for global identifiability of DAG models with latent variable are given in Stanghellini and Wermuth 2002.
We describe the computations required to implement the *EM algorithm* to such models: in the *E-step* we must compute $Q(\Sigma|\Sigma_r)$ the conditional expected value of the complete data log-likelihood given data Y and a guessed initial value of a complete data covariance matrix $\hat{\Sigma}_r$. We call this quantity

$$Q(\Sigma|\Sigma_r) = E(l(\Sigma, |Y_1, ..., Y_p, \hat{\Sigma}_r))$$

We note that in (Kiiveri, 1987) this quantity

$$C(S_{yy}|\Sigma_r) = \begin{pmatrix} S_{yy} & S_{yy}B' \\ . & BS_{yy}B' + (\Sigma^{zz})^{-1} \end{pmatrix} = E(S|Y, \Sigma)$$

Where $B = -(\Sigma^{zz})^{-1}\Sigma^{zy} = \Sigma_{zy}\Sigma_{yy}^{-1}$
It follows that

$$Q(\Sigma|\Sigma_r) = \frac{n}{2}[\ln|\Sigma^{-1}| - \mathrm{tr}[(\Sigma^{-1})C(S_{yy}|\Sigma_r)]]$$

Therefore in the *M-step* we maximize maximize $Q(\Sigma|\Sigma_r)$ as a function of Σ to produce Σ_{r+1}. This maximization is carried out by fitting using least squares the linear recursive regressions generating over the DAG to the sample covariance matrix $C(S_{yy}|\Sigma_r)$. The deviance is $n[\mathrm{tr}(S_{yy}\Sigma_{r+1}^{-1}) - \ln|S_{yy}\Sigma_{r+1}^{-1}| - m]$ and degree of freedom $df = \frac{m(m+1)}{2} - m - k$ where m is the number observed variables and k is the number of edges.

We provide an analytical solution to evaluate the asymptotic variance-covariance matrix of estimates. Following Oakes (1999) equalities hold:

$$\frac{\partial^2 l(\phi_K;\Sigma)}{\partial \phi_K^2} = \left[\frac{\partial Q(\phi'_K|\phi_K)}{\partial \phi_K^2} + \frac{\partial Q(\phi'_K|\phi_K)}{\partial \phi'_K \partial \phi_K}\right]_{\phi'_K=\phi_K}$$

which are valid for all elements of the parameter vector ϕ. The first term corresponds to the complete information and the second term can be interpreted as a correction for the 'missing information' (Louis, 1982). Most of the quantities involved evaluated at the maximum likelihood estimates are obtained as a by-product from the algorithm above. The asymptotic variance-covariance matrix is then derived as the inverse of the observed information.
We develop an R code to estimate the previous models.

References

1. Kiiveri, H.T. (1987): An incomplete data approach to the analysis of covariance structure. *Psychometrika*, **52**, 539-554.
2. Louis, T.A. (1982): Finding the observed information matrix when using the EM-algorithm. *Journal of the Royal Statistical Society*, **44**, 1-38.
3. Oakes, D. (1999): Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society*, **44**, 1-38.
4. Stanghellini, E. and Wermuth, N. (2002): On the identification of of the directed acyclic graph models with one hidden variables. *Submitted*. http://psystat.sowi.uni-mainz.de/wermuth/.
5. Wermuth, N. and Cox, D.R. (2003): Joint response graphs and separation induced by triangular systems. *Revision submitted JRSS B*. http://psystat.sowi.uni-mainz.de/wermuth/.

Hypertextual Characteristics of the Slovenian World Wide Web

Gregor Petrič (University of Ljubljana, Slovenia)

The substantial concern of the presented research is the question to what extent does the contemporary World Wide Web as an information retrieval system reflect certain attributes of ideal hypertextual systems. The topic is relevant, since notions of hypertext or hypertextual systems are in the literature accompanied with strong implications not only for the speed and efficacy of access to information, but also for fostering democratisation, augmenting creativity and cooperativeness of human beings. After the brief presentation of the theoretical problem the paper concentrates on the methodology of analysing it definition of relevant dimensions of hypertext in the World Wide Web, their operationalisation and empirical verification.

Hypertext has become with regard to a relatively simple idea, originating in the work of Vannevar Bush (1945), a very complex term, which is discussed in many scientific fields, from information retrieval and management, especially in the framework of internet search systems (Bieber et al., 1997; Chakrabarti, 1999), to poststructuralism and its focus on the relation between author and reader in contemporary literature (Landow, 1997; Ryan, 1999). In its essence, hypertext however consistently refers to the mode of data organisation in information systems, where logically separated individual pieces of information are interconnected in an associative or connectivist manner (Bardini, 2000), while implications of such an organisation stimulate research in various scientific fields.

Taking for granted the relevance of social implications of hypertextual systems, the research focuses on the existence of hypertextual features in the contemporary World Wide Web, since the Web was conceived in the beginning of 90's explicitly on the basis of preceding ideas or realisations of hypertextual systems (Berners-Lee, 1996). Several relevant dimensions of hypertextual systems are explicated and defined interconnectedness, decentrality and nonlinearity whose operationalisation and verification is presented.

Empirical verification deserves special attention, since it includes a procedure of generating a whole network of 26.954 web sites on the basis of approximately 1.8 million of web pages in the Slovenian World Wide Web, identified by search system Najdi.si, owned by company Noviforum. After the definition of units and relations, relevant methods and their results are presented in order to assess the hypertextuality of the Slovenian World Wide Web. Social network analysis offers a suitable analytical instrument, while the analysis was performed with program Pajek (Batagelj & Mrvar, 2002), specialised for analysis of large and sparse networks. Relevant characteristics, such as centrality measures, centralisation indexes, diameter, k-cores and components are reported and informative graphical presentations included.

It is shown that a relatively great proportion of web sites do not follow the expectation of the designers of the World Wide Web technology for it to be a globally interconnected "Docuverse", however a large minority of web sites are in aggregate reflecting the attributes of ideal hypertext systems. The results can be informative for the global World Wide Web since the essential characteristics of the Slovenian World Wide Web have similar distributions to the ones assessed in other researches (Broder et al., 1999; Kleinberg, 1998) on significantly larger, although not sufficiently adequate for complete network analysis, proportions of the World Wide Web.

References

1. Bardini, T (2000): *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing*. Stanford: Stanford University Press.
2. Batagelj, V. and Mrvar, A. (2002): Pajek.
http://vlado.fmf.uni-lj.si/pub/networks/pajek
3. Berners-Lee, T. (1996): *The World Wide Web: Past, Present and Future*. Proceedings of the Fifth International World Wide Web Conference, Computer Networks and ISDN systems, 7-11.
4. Bieber, M., Vittali, F., Ashman, H., and Oinas-Kukkonen, H. (1997): Fourth generation hypermedia: Some missing links for the World Wide Web. *International Journal of Human Computer Studies*, **47**, 3165.
http://www.hbuk.co.uk/ap/ijhcs/webusability/
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (1999). *Graph Structure in the Web*. Proceedings of the 9th WWW conference.
http://www.almaden.ibm.com/cs/k53/www9/final/
6. Bush, V. (1945): As we may think. *Atlantic Monthly*, **176**, 101108.
http://www.isg.sfu.ca/~duchier/misc/vbush
7. Chkarabarti (1999): *Surfing the Web Backwards*. Proceedings of the 8th WWW conference.
http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html
8. Kleinberg (1998): Authoratitve sources in hyperlinked enivorment. Proceedigns of 9th ACM-SIAM SODA.
9. Landow, G.P. (1997): *Hypertext 2.0: The Convergence of Contemporary Critical Theroy and Technology*. London, Baltimore: The Johns Hopkins University Press.
10. Ryan, M.-L. (1999): Cyberspace, virtuality, and the text. In M.-L. Ryan (Ed.): *Cyberspace Textuality: Computer Technology and Literary Theory*, 78110. Indiana: Indiana University Press.

Autologistic Regression Model for Poverty Mapping and Analysis

Alessandra Petrucci, Nicola Salvati, and Chiara Seghieri
(University of Florence, Italy)

The mapping of poverty in developing countries has become an increasingly important tool in the search for ways to improve living standards in an economically

and environmentally sustainable manner. The methods used to generate poverty maps naturally come under more intense scrutiny as their policy implications become more apparent. Those most commonly used until now have involved the application of econometric models to generate local indicators of poverty.

Although the classical econometric methods provide information on the geographic distribution of poverty, they do not take into account the spatial dependence of the data and generally they do not consider any environmental information. Therefore, methods which use spatial analysis tools are required to explore such spatial dimensions of poverty and its linkages with the environmental conditions.

This study applies a spatial analysis to determine those variables that affect household poverty and to estimate the number of poor people in the target areas. This type of analysis is based on the assumption that measured geographic variables often exhibit properties of spatial dependency (the tendency of the same variables measured in locations in close proximity to be related) and spatial heterogeneity (non-stationarity of most geographic processes, meaning that global parameters do not well reflect processes occurring at a particular location). While traditional statistical techniques have treated these two last features as nuisances, spatial statistics considers them explicitly.

The data are arranged from three sources in a GIS database in order to manage the spatial dimension. The study applied a five-step spatial analysis to determine the variables affecting household poverty and to estimate the number of poor people in the target areas. Spatial autocorrelation exists whenever there is a pattern in the values recorded at locations in a map. Global and local spatial statistical techniques to detect association and autocorrelation share two aspects: (i) the assumption of a spatially random distribution of data; and (ii) the spatial pattern, structure and form of dependence are typically derived from the data. A pseudo-likelihood estimation procedure of an autologistic model is used, avoiding the difficulties of the full maximum likelihood approach. A case study is carried out to test the methodology.

The method applied for poverty mapping at the level of small geographical areas provided a more accurate specification of poverty by incorporating a spatial component into the classical regression model.

The results of the fitted spatial model demonstrate the statistical significance of environmental variables and suggest that environmental indicators could be an important component of poverty reduction strategies. Moreover it is important to underline the significant effect of the spatial correlation variable, introduced in the autologistic model, that denotes the presence of clusters in the spatial distribution of poverty and the influence betweenneighbour households on the probability of

being poor.
Overall, the results of the study demonstrate the usefulness of the spatial analysis method in poverty targeting.

Role of ICT Knowledge in the First Educational Transition: Measurement and Modelling Aspects

Eva Podovšovnik (University of Primorska, Koper, Slovenia)
Anton Kramberger (University of Ljubljana, Slovenia)

The main idea of the paper is to empirically investigate how natural ability, attained skills and competences, and contextual factors of pupils affect the achieved level of their ICT knowledge, their wishfull educational choices, and the final outcomes of the first educational transition (from primary to secondary education). A special survey research[5] among the Slovenian elementary school leavers[6] started in May and June, 2003; hopefully, it will be finished in September 2003.[7]
We organized the paper presentation in two parts. Firstly, we present an extensive measurement effort aimed at robustly estimating the achieved level of ICT knowledge among the pupils. Secondly, we discuss a set of models, by which the relative impacts (effects) of different factors, being effective before and during the process of educational transition, are statistically assessed.
Details are the following. In the first part of the paper, after a brief terminological discussion, we explain basic methodological issues: measurement steps (research design, development of survey questionnaires, sampling procedures, field work issues), quality of data gathered (missing data treatment, sample sources of variations in error terms), along with the tests of external/internal validity and reliability of the key numeric variable (ICT knowledge as a synthetic index construct). External validity of the ICT index's measurement is assessed by matching the variability in ICT indices with the variability in the individual concept maps on likely ICT usage (qualitative perceptions of the ICT), developed by the pupils. In the second part of the paper we model the following three processes: gaining ICT knowledge, formulating educational choices, and exercising first educational transition. Dependent variables for this processes are three, respectively: ICT

[5]The research is part of the doctoral thesis which Eva Podovšovnik is currently elaborating on under the supervision of prof. Kramberger, at the Faculty of Social Sciences in Ljubljana.

[6]In Slovenia the elementary school goes up to the eighth or ninth level. This means that the research was conducted among children aged 13-15 years.

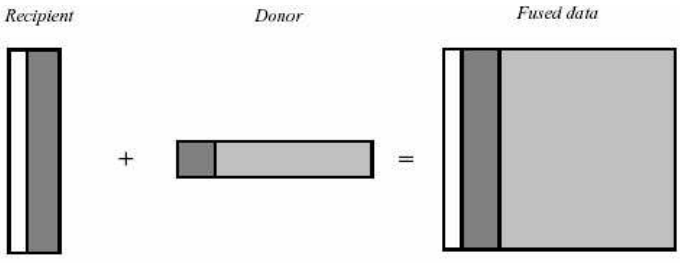
[7]This part of the research is still not finished yet, because pupils will know the results of their transition in September 2003.

knowledge (numeric), educational choices (categorical), and real educational destinations (categorical). Among background factors, we introduce a set of possible explaining variables: more personal characteristics (general ability, competences, overall skills of pupils) and more contextual variables (family, class size & structure, teachers ICT involvement, school ICT equipment,[8] school regional location). These factors we understand as most plausible 'causes' of the variations in pupils ICT knowledge and their educational decisions. Statistical models, applied as to fairly describe the above mentioned processes, are: education production function (process: gaining ICT knowledge), Logan's two-sided logit model or similar approach (process: educational choices due to educational opportunity structure), and Mare's logit/probit models (process of social selection in school transition, by omitting temporal/trend dimension). Models are (re)designed in such a way as to gradually and meaningfully delineate the relative factor effects, by separating factors of merit-selection on the one side from the factors of social selection on the other. We finish the paper with preliminary empirical results of the still ongoing research.

On Data Integration and Public Statistics: The Statistical Matching
Gabriella Schoier and Adriana Monte (University of Trieste, Italy)

The main idea of statistical matching is to give joint information on variables available from different sources. It is a procedure for data integration when these derived from different sets or when there is the lack of units identifiers (for privacy constraints) (see e.g., Räessler and Fleischer, 2002; D'Orazio, Di Zio, and Scanu, 2002).
Among the different advantages of these procedures we may mention the possibility of obtainining timely results and cutting costs, moreover there is a reduction of the response burden of the statistical units. Regarding drawbacks there is the necessity of specify a correct model on which build up the statistical matching procedure otherwise the overall quality of the final results may be poor.
The statistical matching concepts may be described by the figure. There are a certain amount of variables occurring in both the database (the recipient) and the survey (the donor) called common variables. For each record in the database the procedure try to predict the most probable answers for the survey, the variable to be predicted are called fusion variables. The resulting set is called synthetic set.

[8] While measuring the use of the ICT the focus is on the intensity and the type of the use of the ICT. The use is measured only as the use of the ICT of teachers.



Our data regards the information needs (open learning) of the small enterprises (with at max. 49 employees) of . The sample is formed by 205 enterprises. The recipient set (variables X and Y) is formed by all the 205 records. The donor set (variables X and Z) is formed by a random sample of 134 records drawn by the recipient set. The objective is to integrate the recipient set with the information (variables Z) contained in the donor set.
The common variables X are: sex, age and studies of the entrepreneur, year of foundation of the enterprise, percentage of clients who are final consumers. The variables Y are: number of personal computers, numbers of Internet connections; presence of the enterprise on the web; use of Internet. The variables Z are: human resources, participation to distance courses, future interest in distance courses.
A traditional way to handle this situation is to look at it as an imputation problem. In particular, in this paper, the multiple imputation approach has been considered (see e.g., Rodgers, 1984; Rubin, 1986). The results of this approach have been compared with a new approach based on neural networks (see e.g., Van der Putten, 2000). The final result is a synthetic set of 205 records containing the three vectors of variables X, Y and Z. It is possible to compare the distributions of the synthetic set and of the donor set for the Z variables. The best results rergard the variables: participation to distance courses and future interest in distance courses.
The results are encouraging as regards the introduction of Data Mining techniques to deal with these problems.

References

1. D'Orazio, M., Di Zio, M., and Scanu, M.(2002): Statistical matching and official statistics. *Rivista di Statistica Ufficiale 1*, Istat, Roma.
2. Räessler, S. and Fleischer, K. (2002): *Aspects Concerning Data Fusion Techniques*. http://www.uni-leipzig.de/wifa/emp/d/public.html
3. Rodgers, W.L. (1984): An evaluation of statistical matching, *Journal of Business and Economic Statistics*, **2**, 91-102.
4. Rubin, D.B. (1986): Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic*

Statistics, **6**, 87-94.

5. Van der Putten, P. (2000): Data fusion: A way to provide more data to mine in. *Proceedings 12th Belgian-Dutch Artificial Intelligence Conference BNAIC2000*, De Efteling, Kaatsheuvel, The Netherlands, http://www.liacs-nl/ putten/library/fusiebnaic00.htlm

---

The Bernard Estimator: A Simulation Study in Case of Overlapping Sub-Populations

Silvia Snidero (University of Trieste, Italy)
Alessandra Durio (University of Torino, Italy)
Giulia Zigon (University of Firenze, Italy)
Dario Gregori (University of Torino, Italy)

The Bernard estimator is a simple and appealing approach to the estimation of the size of hidden and unreacheable sub-populations. Each person is interviewed to detail the number of people known in several sub-populations (of size known) and in the target subpopulation E (which size is to be estimated). Assuming that the proportion of subjects belonging to E over the number c of people in the social network of the person is the same that in the overall population, we can get a reasonable estimate of the target population. In case the size of the social network c is not known, it has to be estimated in a two step procedure.
Although appealing becasue of the potential efficiency w.r.t. traditional sampling schemes, it remains unclear how to choose the sub-populations, particularly in case of a (partial) overlap of them.
In our study we perform an extensive simulation to determine the bias and the efficency of various Bernard-type estimators in several conditions of overlap.

What Makes the Estimates of Factor Loadings More or Less Accurate?

Gregor Sočan (University of Groningen, The Netherlands)

In the psychometric literature, surprisingly few information can be found about factors influencing the accuracy of factor loadings estimates in samples: most often just some simple rules of thumb with regard to sample size are offered. In our talk we shall report some results on a comparison of three factor analytic methods in various conditions of sample size, average communality and factor overdetermination. The three methods included Maximum Likelihood Factor Analysis,

Minres Factor Analysis (a variant of the least squares factoring) and Minimum Rank Factor Analysis. The latter is a recently developed method of common factor analysis with two unique features. First, it produces communality estimates yielding a proper (positive semidefinite) reduced correlation matrix. Second, it makes possible to estimate the proportion of common variance left unexplained after extraction of r common factors; this measure is a useful descriptive diagnostic for determining the optimal number of extracted factors. Our simulation study used simulated sample correlation matrices, based on contrived population loading matrices. The accuracy of the estimates of loadings was assessed by the congruence coefficient and by the mean absolute deviation. All three investigated factors (and the interaction of sample size and average communality) notably influenced the accuracy. The three methods were not equally accurate: the estimates by Minimum Rank Factor Analysis had the highest average congruence, and the Minres estimates had the smallest mean absolute deviation. The differences among the methods were larger in the less favorable conditions. Further, the presence of a Heywood case notably reduced the accuracy of all methods, too. Heywood cases seem to deteriorate the performance of Maximum Likelihood Factor Analysis more than the performance of Minres and Minimum Rank Factor Analysis.

An Individual Measure of Relative Survival

Janez Stare (University of Ljubljana, Slovenia)
Robin Henderson (Lancaster University, United Kingdom)
Maja Pohar (University of Ljubljana, Slovenia)

The ratio between the observed and the expected survival is often used in studies of survival, cancer being the most common example. This ratio is called the relative survival and it is a measure of the net effect of the disease on survival. While the expected survival is, in principle, a simple concept, there are some problems in calculating it in a way that it controls for heterogeneity in the observed follow-up times and making it independent of the observed mortality at the same time. Of the methods proposed the best approach is that of Hakulinen. Still, all these methods focus on group experience, and say nothing about individual values. For example, they do not answer a very natural question "How long, relative to the general population, has a certain person lived?"

We introduce an individual measure of relative survival that provides answers to such and similar questions. But, more importantly, it enables simple and effective

modelling of excess risk via regression models. In this presentation we focus on properties of the measure, and the comparison of our regression approach to those of Hakulinen and Tenkanen (1987), and Andersen et al (1985). The main advantage of our approach is that it accounts for the population risk without assuming anything about its relation to the excess risk. Further, in regression models, variables other than those entering population tables can be used. In fact, the method enables usage of all known regression approaches without restrictions, using only standard statistical software.

The approach is illustrated using data on survival of patients after acute myocardial infarction.

Measuring Information Accessibility and Predicting Response Effects: The Validity of Response Certainty and Differently Transformed Response Latencies

Volker Stocké (University of Mannheim, Germany)

Research has shown that survey respondents answers are often influenced by seemingly irrelevant features of response options. For instance, subjects reports about frequency and duration of everyday behavior has been found to be strongly affected by the range of the response scales. Instead to ask the focal information, for example the number of products purchased in a reference period, directly as an open-ended questions, respondents often have to select between quantitative alternatives, each representing a range of the response continuum. Depending on how many options are used to represent the higher and lower part of the continuum, the reported quantity under investigation has been found to differ substantially.

One explanation for these as well as for other response effects is an insufficient cognitive availability of the requested information. Under this condition, subjects use the relative number of alternatives representing the higher and lower part of the frequency continuum to infer an answer as appropriate as possible. Research has supported the role of imperfect information as a determinant for this sort scale effects. Accordingly, stronger effects are found for survey topics that can be assumed to be less available in memory. It is, however, unclear whether individual differences in information availability with regard to the same topic can predict the respondents susceptibility to the frequency range of the response scales. This is an important question for survey research, because it might allow to identify respondents who are more or less prone to response effects.

In order to realize this aim, a valid measure for the cognitive availability of factual

information is needed. For this purpose, meta-cognitive and operational indicators have been proposed. Whereas the first type relies on the respondents self-reports, the second group of measures is based on objective and therefore directly observable features of the response generation process. The subjective response certainty is probably the most frequently used meta-cognitive indicator and the time needed to answer the questions the most prominent operational measure. In general, operational measures are assumed to be the more valid indicators, because they do not require the respondents to judge their answers from an unfamiliar meta-perspective. However, it has not yet been analyzed systematically which indicator is in fact more valid and thus the better predictor for respondents susceptibility to response scale effects.

Although response latencies are a measure that is innovative and available at no additional cost, evidence about their predictive power is rather mixed. One reason for this might be that the way how response latencies are transformed prior to data analysis varies broadly. In order to reduce the characteristic skewness of latency data the distributions are treated with logarithmic, square-root or reciprocal transformations. Whether these transformations affect the validity of response latencies has so far not been systematically studied. The same holds true for the question whether and how latency data should be standardized, using the individual respondents average response time and its variation across different survey topics. On the one hand these individual differences are assumed to be an error component and should therefore be eliminated. On the other hand they are regarded to represent individual differences in the global information accessibility between the subjects. We are not aware of any systematic studies which analyze the relative merit of these arguments.

The present paper has three closely related aims. First, it is analyzed whether individual differences in the information availability predict the degree to which respondents answers are affected by the range of frequency scales. The second goal of the study is to address the question whether meta-cognitive or operational indicators are better predictors in this respect. The third an most important goal is to provide evidence about whether the transformation and standardization of response latencies affects their predictive power.

Data from a local survey based on a random probability sample is used to answer these questions. In this survey, respondents reported the length of their daily television consumption with high or low frequency scales. This is done with a split-ballot design. As in many other studies, subjects reported significantly longer viewing-times under the condition of the high frequency scale. As a first result, both indicators for information availability, response certainty and response la-

tency, predict how strong subjects answers are influenced by the characteristics of the response scales. However, using multivariate regression analysis, response latencies are found to be the stronger predictor. The comparison of altogether six differently transformed versions of response latencies reveals that one treatment in particular increases their predictive power substantially. Here, individual respondents mean and standard deviation of response times, observed when answering the other question in the interview, are used to compute z-scores for the target response latencies. With this kind of standardization, the raw response latencies ability to predict how strong respondents are affected by the response effect under consideration is nearly doubled. Other types of transformations do have either no or a much weaker effect on the response latencies predictive power.

Dynamic Process Simulation as Tool to Connect Theory and Data
Frans N. Stokman (University of Groningen, The Netherlands)

My focus is on simulation as a tool in macro-micro-macro analyses of social systems. Most simulation models make too many ad hoc assumptions and are not or only very loosely connected with empirical data. In contrast, I argue for simple models, well grounded on theory, adding complexity stepwise. Following the paradigms of structural individualism, the emphasis lies on multi-agent modeling, taking the incentive structures of social actors as starting points. The model(s) should focus on the basic processes through which simultaneous and interdependent individual actions are transformed into collective outcomes. By feeding the models with parameters that specify the incentive structures of social actors and their social constraints (and are obtained in efficient data collection procedures), powerful models can be developed that have high predictive value and provide valuable instruments for social intervention.
I will illustrate the approach with models of collective decision making and social network evolution. Particularly the collective decision making models have been validated in extensive scientific and applied research, providing important and valuable extra insights even to 'experienced negotiators'.

Comparative Analysis of AHP and DEX Decision Making Methods

Robert Špendl (OIKOS Inc., Domžale, Slovenia)
Marko Bohanec (Jožef Stefan Inst., Ljubljana, Slovenia)
Vladislav Rajkovič (University of Maribor, Slovenia)

Two multiattribute decision making methods, AHP and DEX, are compared in the paper with respect to decision knowledge management. Both methods rely on hierarchical decomposition of criteria. AHP (Analytic Hierarchy Process) is a quantitative method based on a matrix of relative importance of criteria and a matrix of relative comparison of options. DEX (Decision Expert) is a qualitative method whose models use variables with descriptive values, and utility functions that are expressed by decision rules. The knowledge management is discussed in the frame of organizational effectiveness, comprehensibility and explanation of results. The comparison is carried out on a real-life example of environmental projects evaluation.

In multiattribute decision making, preferential modeling is based on knowledge about (1) attributes and their structure, (2) alternatives, and (3) utilities. The decision model has to express all decision-relevant features of alternatives according to the decision maker's objectives. Knowledge about alternatives represents the decision maker's view on alternatives; the decision maker should know them sufficiently well to describe or measure them along attributes. Knowledge about utilities can be treated in two steps. The first step is to establish a dependence between attribute's real values (money, size, age, etc., measured on alternatives) and their utilities as a reflection of decision maker's preferential opinion. The second step is the aggregation of these partial utilities into the final overall utility value.

A number of methods have been developed to support the decision knowledge management. One of the well known approaches is hierarchical decomposition in which the decision problem is decomposed into smaller and less complex subproblems. The result of decomposition is a hierarchical decision model. Such decomposition contributes to a better fit between decision models and human understanding. By this, the decision makers' cognitive processes are supported by decision knowledge which can be easily understood, and can be easily updated and actively used by all participants in the decision making process.

A distinguishing characteristic of AHP is a method to obtain the elements of decision models that is based on a pairwise comparison of criteria and options. For the purpose of this paper, we used AHP as implemented in the program Expert Choice. The second approach, DEX, uses qualitative decision models. Each variable in the model can take only symbolic values, which are usually expressed

with words such as good or unacceptable. The aggregation of partial evaluations into the overall evaluation is carried out by decision rules. This methodology is supported by an expert system shell DEX.

The goal of this paper is to analyze the strengths and weaknesses of both approaches and to propose a complementary usage of them for effective decision knowledge management. The comparison is explained on a real-life example of environmental projects evaluation. Both methodologies and the corresponding programs are presented in parallel following the major steps in decision model development: (1) hierarchical decomposition of criteria, (2) aggregation, (3) assessment of options and (4) analysis of results.

Comparison of the methods on a case of project evaluation showed that DEX is more comprehensible method. AHP enabled better resolution of similar options, but the results were not clearly comprehensible. The results of DEX described exactly what was expected from the decision model: whether the proposed project is promising enough to be performed or not. For each project, DEX also highlighted its strengths and weaknesses, which is particularly important for successful project management. The most important advantage of AHP is the ability to distinguish between similar options, for which creating comparison matrices is easier than for complex problems.

An interesting further work could be implementing a combination of DEX and AHP, where DEX would classify an option into a certain class, and AHP would be used for detailed evaluation within a certain class.

An Exploratory Analysis of Image and Perception of Risk of a Tourist Destination

Nirundon Tapachai (Kasetsart University, Bangkok, Thailand)

This study aims to 1) identify the underlying dimensions of the destination images of Trang, one of well-known tourist destination in Thailand 2) investigate the perceived risk of tourists towards Trang as a tourist destination 3) determine differences in the image and perceived risk of Trang held by first-time and repeat tourists, domestic and international tourists and among travelers with different demographic and psychological profiles and 4) to explore the relationship between destination image and perceived risk level.

Data were collected from a convenience sample of 500 tourists who visited Trang (400 Thai tourist and 100 international tourists using self-administered questionnaires. The questionnaires were administered between February and April 2003.

The questionnaires consisted of three components: an image measurement component, a perceived risk measurement component and a set of demographic and personal variables. The image measurement component consisted of 22 selected image items derived from the pilot study and the perceived risk component consisted of 16 risk related items derived from relevant literature. The respondents were asked to state their extent of agreement to these image and risk items on a five-point Likert type scale (1 = strongly disagree, 5 = strongly agree). In addition not sure/dont know answer was provided along with each question to help ensure that no response in the scale was applicable.

For pretesting purposes, the questionnaire was administered to 30 tourists visiting Trang (20 for Thai tourists and 10 for international tourists. The tourists were asked to provide feedback on readability where not any change was suggested. The reliability of image and perceived risk set of questions were also tested.

Both univariate and multivariate statistics were applied to analyze the data. First, descriptive statistics was used to determine the frequency distribution of travel behavior and demographic profiles and the mean and standard deviation of image and perceived risk of Trang of the surveyed respondents. Second, an exploratory factor analysis was used to identify the underlying dimensions of the image of Trang. Based on the resulting factor structures, summary-scale scores for subsequent analyses were constructed. Third, analysis of variance (ANOVA) was conducted to determine whether different groups of tourists held different images and perceived risk of Trang. Finally, the Pearsons correlation was used to determine the relationship between image and perceived risk level.

Preliminary analyses show final factor analysis of 22 image attributes compiled seven image factors, the relatively high level of confidence on the destination, and differences in image perceptions according to demographic and behavioral factors. The details of the research results are shown in the full paper.

Discovery of Polynomial Equations for Regression

Ljupčo Todorovski, Sašo Džeroski, and Peter Ljubič
(Jožef Stefan Institute, Ljubljana, Slovenia)

Equation discovery (Langley et al., 1987) aims at developing methods for computational discovery of quantitative laws or models, expressed in the form of equations, in collections of measured numerical data. Equation discovery methods are mainly used for automated modeling of real-world systems from measurements and observations. Since they operate on numerical data, they are strongly related

to the regression methods used in statistics and data mining for inducing predictive models of numerical variables.

The difference between equation discovery methods are mainly in the application focus. While equation discovery methods aim mainly at inducing *comprehensible* and *general* models of the observed system, regression methods focus on the problem of inducing *accurate* predictive models of a designated target variable. Due to the difference in the focus, both methods are evaluated on different kinds of problems and using different evaluation criteria. In this paper, we evaluate the performance of an equation discovery method on fourteen tasks from the UCI Repository of Machine Learning Databases and Domain Theories (Blake and Merz, 1998) and compare it to the performance of standard state-of-the-art regression methods built in the WEKA data mining suite (Frank and Witten, 1999).

Equation discovery method follows the generate-and-test approach. They start the search with the simplest equation structure, i.e., $v_d = const$, where v_d is the dependent (or target) variable. At each search iteration, more complex equations are generated. In case of polynomial equations, the complexity of the polynomial on the right-hand side of the equation is increased in one of two directions. First, we can add an arbitrary linear term, that is a single variable to the current polynomial on the right-hand side of the equation. Second, we can add a variable to one of the current terms in the polynomial and increase its degree. At each search step the values of the constant parameters in the equations are fitted against training data using linear regression methods. The equation discovery method searches for the equation that fits the training data best.

The quality of the obtained equation is usually evaluated using a degree of fit measure that measures the discrepancy between the observed values of v_d and the values predicted using the equation. One such measure is mean squared error (MSE), calculated as $\mathrm{MSE}(v_d = P) = 1/m \cdot \sum_{i=1}^{m}(v_d(i) - \hat{v_d}(i))^2$, where $v_d(i)$ is the value of the target variable v_d for the i-th training example, $\hat{v_d}(i)$ is the value of v_d for the same example, but predicted using equation $v_d = P$, and m is the number of training examples.

To avoid over-fitting, we use a MDL (minimal description length) based heuristic function for evaluating the quality of equations that combines the degree of fit with the complexity of the equation. In the literature, the following combination has been proposed:

$$\mathrm{MDL}(v_d = P) = (d + r) \log m + m \log \mathrm{MSE}(v_d = P),$$

where d is the degree of P, r is the number of terms in P, and m number of training examples. While the second term the MDL heuristic function measures the

degree of fit of a given equation, the first term introduces a penalty for complexity of the equation. Through this penalty the MDL heuristic function introduces preference toward simpler equations.

In performed experiments, we evaluated the performance of the equation discovery method, outlined above, and compared it to the performance of standard regression methods, such as linear regression and regression trees, as implemented in WEKA (Frank and Witten, 1999). The comparison was performed on a collection of fourteen regression datasets from the UCI repository (Blake and Merz, 1998). In all the experiments, regression performance is estimated using 10-fold cross validation. The regression performance is measured in terms of RE defined as $RE = \sum_{i=1}^{m}(v_d(i) - \hat{v_d}(i))^2 / \sum_{i=1}^{m}(v_d(i) - \overline{v}_d)^2$, where $v_d(i)$ and $\hat{v_d}(i)$ are the observed and predicted values of the dependent variable for the i-th training example, m is the number of examples, and $\overline{v}_d$ is the average value of the dependent variable. Note that RE gives a normalized value of the mean squared error, that is independent on the magnitude of the dependent variable v_d. The normalization allows for comparison and aggregation of the performance measure across different datasets.

On average, the polynomial equations, induced using equation discovery, perform better (average RE on the fourteen datasets 0.5465) than linear regression (average RE: 0.5885) and regression trees (average RE: 0.6746). To our surprise, the performance of model trees (average RE: 0.5404) is comparable to the one obtained using polynomial models. We should note that this is achieved using a single equation/model over the entire instance space, rather than a piece-wise model as in model (or regression) trees.

Furthermore, we compared the complexities of different models. The complexity of polynomial equations is assessed in terms of polynomial degree d (average 1.64), and number of terms r (which equals the number of constant parameters in the equations, average 4.93). The complexity of linear regression is assessed in terms of number of constant parameters. Finally, the complexity of tree models is measured in number of decision nodes and number of constant parameters used in the leaf nodes. While regression trees use one constant parameter in each leaf node, model trees use linear regression model in each leaf node. Thus, the number of the constant parameters in the model trees is higher compared to the number of constant parameters in the regression trees. The results show that the average complexity of the polynomial models compares favorably with complexities of the other regression methods. The average model tree consists of 3 decision nodes and includes 14 constant parameters in the leaf nodes, which leads to the average number of constant parameters of 17. The average number of constant parameters

in regression tree models is 13. Both numbers are significantly higher than the average of 5 parameters in the polynomial models.
The directions for further work include integration of clustering methods with equation discovery in order to obtain piece-wise models (one piece for each cluster) based on polynomial equations. This might increase the accuracy of a single equation induced over the entire instance space.

References

1. Blake, C.L. and Merz, C.J. (1998): *UCI Repository of Machine Learning Databases.*
2. Frank, E. and Witten, I.H. (1999): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Mateo: Morgan Kaufmann.
3. Langley, P., Simon, H.A., Bradshaw, G.L., and Żythow, J.M. (1987): *Scientific Discovery*. Cambridge: MIT Press.

Response Trends in a National RDD Survey

Robert D. Tortora (The Gallup Organization - Europe)

This paper examines response trends in a national RDD survey over 23 consecutive quarters of data collection starting in 1997. The target population is adults that have used the Internet in the last 30 days. The constant survey design includes stratification by modeled household income and uses a maximum of five calls for each sampled telephone number. The analysis includes comparisons by sampling stratum and over time for the following key variables: refusal, working number, contact, cooperation, completion, response, busy, answering machine, no answer, deafness/language barrier and 'other' barriers to response.
Findings include that upper income households, defined by membership in the highest income stratum, have poorer response characteristics, such as higher refusal rates, lower contact and cooperation rates and hence lower response rates, than in the lower household income sampling strata. In addition, while refusal rates are decreasing over time a decrease in contact and cooperation rates is causing a reduction in response rates both at the stratum level as well as overall.

Methodological Issues in Mobile Phone Surveys

Vasja Vehovar (University of Ljubljana, Slovenia)
Zenel Batagelj (CATI, Ljubljana, Slovenia)
Mario Callegaro (University of Nebraska, Lincoln, USA)
Marek Fuchs (University of Eichstaett, Germany)
Kuusela Vesa (Statistics Finland)

The methodological issues of mobile phone surveys are becoming increasingly problematic. The share of mobile-only households (with no fixed telephone line) is steadily growing and it has already exceeded 20% in some European countries. It thus seems that fixed telephone surveys will no longer provide representative samples of general population. However, the costs of mobile phone surveys are much higher and the response rates also seem to be lower compared to the fixed telephone surveys.

The paper first overviews the status of mobile phone interview surveys across different countries. Next, the issues of self-administered mobile phone surveys are discussed. Finally, a fresh Slovenian mobile phone research study is presented to illustrate the contact and cooperation problems with mobile phone surveys. The study basically confirmed a whole array of disadvantages of mobile phone surveys. It also revealed considerable national specifics.

A Study of Socio-Economic Determinants of Mortality in Slovenia: Findings and Methodological Issues

Gaj Vidmar and Barbara Artnik (University of Ljubljana, Slovenia)
Jana Javornik (Institute of Macroeconomic Analysis and Development, Ljubljana, Slovenia)

A pioneering joint research has been conducted by the Faculty of Medicine of the University of Ljubljana and the Institute of Macroeconomic Analysis and Development, to study the relationship between selected biological (age, sex), socio-economic (marital status, education, occupation, nationality, income) and geographical factors (region), and the cause of death (classified according to the ICD-10) and age at death. With the help of the Statistical Office of the Republic of Slovenia and the Slovenian Institute of Public Health, a database was compiled for all deceased in Slovenia in 1992, 1995 and 1998 (almost 60.000). The diagnoses of the cause of death were linked to the data on the deceased from the 1991 Census; for individuals who died in 1998, personal income tax data for 1996 were

appended.
The first part of the paper gives an overview of the technical and methodological issues in collecting, cleaning and recoding the data. The second part summarises the findings that have been published as part of the annual Human Development Report for Slovenia, which were obtained mainly by means of bivariate analyses and focus on premature mortality, i.e., deaths at age 25 to 64 (N=14816). Attempts to develop multivariate models of cause of death and life expectancy, as well as to identify clusters of individuals based on socio-demographic and mortality data, are reported in the third part of the paper.
The study is still a work in progress and it has been designed so as to allow periodic updates of the database. The data and the ongoing analyses should aid in understanding the nature, prevalence and consequences of health problems as related to socio-economic inequalities in Slovenia, and thus serve as a basis for setting healthcare policy goals and planning healthcare measures.

Effects of Categorization of Variables on Discriminant Analysis: A Simulation Study

Irena Vipavc, Simona Pustavrh, and Barbara Zemljič
(University of Ljubljana, Slovenia)

The present study is focused on measuring the influence of categorization of interval variables on the error of classification in discriminant analysis. Discriminant analysis is multivariate technique concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. Using Monte Carlo simulation, the research in this study addresses two different issues regarding categorization and classification. Categorization was done by dividing values of each interval variable in three to ten intervals, first, when they were of equal width, and second, when they were of different width. Widths of intervals in the second case form a geometric sequence, which leads to asymmetric distribution of variables. The interpretation is focused on classification of units from classification table into belonging groups and on the effect of categorization on position of centroids. The results indicate, that discriminant analysis is robust enough, when categorizing normally distributed variables in five to ten categories. The same was obvious when variables were asymmetric, however this was true only for certain level of asymmetry.

Random Effects Models for Correlated Censored and Uncensored Data

Ronghui Xu (Harvard University, USA)

In this talk we describe our recent work on random effects models for right-censored data. Vaida and Xu (2000) provided a general framework for handling random effects in proportional hazards regression, in a way similar to the linear, non-linear and generalized linear mixed effects models that allow random effects of arbitrary covariates. This general framework includes the frailty models as a special case. Maximum likelihood estimates of the regression parameters, the variance components and the baseline hazard, and empirical Bayes estimates of the random effects can be obtained via an MCEM algoritm. Variances of the parameter estimates are approximated using Louis' formula.

We show interesting applications of the random effects Cox model to a US Vietnam Era Twin Registry study on alcohol abuse, with the primary goal of identifying genetic contributions to such events. The twin pairs in the registry consist of monozygotic and dizygotic twins. After model fitting and for interpretation purposes, the proportional hazards formulation is converted to a linear transformation model before the results on genetic contributions are reported. The model also allows examination of gene and covariate interactions, as well as the modelling of multivariate outcomes (comorbidities).

Time permitting we will discuss methods for model selection. These include likelihood ratio tests for the variance components, and information criteria under the random effects models.

Search of Latent Dimensions to Analyse the Withdrawal Risks of the University Studies

Emma Zavarrone (Universita di Milano-Bicocca, Italy)

Every student is at risk of withdrawal from university since the first year of his/her academic career. Withdrawal is a complex phenomenon, determined by various causes that may have important influence on university administration. This paper will prove the hypothesis that both causes and effects of withdrawal from university are influenced by the same latent variable, namely Withdrawal inclination (B1). This hypothesis provides Milano-Bicocca University with a useful support for determining the reasons ('causes') that lead students at risk of early university leaving into such behaviour and consequently for containing withdrawal ('effects'). This study uses data on Faculty students at Milano-Bicocca University

enrolled between 1992 and 1995. A MIMIC model has been applied to allow for simultaneous control of both 'causes' indicators of withdrawal inclination and the 'effects' indicators (Blalock, 1964; Bollen and Lennox, 1991; Hox, 2002; Joreskog, Goldberger; 1975).
The course indicators are: marks in the high schools (VT) and the type of the high schools (TSC) while the indicators effects are linked to the job market (age at the first job (ETD) and a proxy variable to estimates the presence of the student in the job market before the University (DING)).
A past paper (Civardi and Zavarrone, 2002) highlights that the variables VT, TSC and DING, measure a latent dimensions, called school background (BF), that determines the exact conclusion of university study.
The common latent dimension allows applying the segmentation on the basis of the spent time before the withdrawal has allowed for the identification of withdrawal inclination modalities and of groups of students at similar risk of withdrawal. The focus of the segmentation is the withdrawal time linked to the schools background (measured with hazard ratio - HR). The results highlight eleven groups with different characteristics.

Short Cycles Connectivity
Matjaž Zaveršnik and Vladimir Batagelj (University of Ljubljana, Slovenia)

Short cycle connectivity is a generalization of ordinary connectivity. Instead by a path (sequence of edges), two vertices have to be connected by a sequence of short cycles, in which two adjacent cycles have at least one common vertex. If all adjacent cycles in the sequence share at least one edge, we talk about edge short cycles connectivity.
It is shown that the short cycles connectivity is an equivalence relation on the set of vertices V, while the edge short cycles connectivity components determine an equivalence relation on the set of edges E. Efficient algorithms for determining equivalence classes are presented.
Short cycles connectivity can be extended to directed graphs (cyclic and transitive connectivity). For further generalization we can also observe connectivity by small cliques or other families of graphs.
Some applications of short cycle connectivity in analysis of large networks will be presented.

Reliability of Measures of Centrality and Prominence Based on Fixed Choice Data Collection Modes

Barbara Zemljič and Valentina Hlebec (University of Ljubljana, Slovenia)

This research evaluates reliability of measures of centrality and prominence of social networks among high school students. Authors show results from fourteen experiments altogether. In the first set of eight experiments four types of social support were measured three times within each class. Four measurement scales (1) binary, (2) categorical, (3) categorical with labels and (4) line production, as well as two measurement techniques for listing alters (free recall and recognition) were applied. In the second design two measurement scales (5) categorical and (6) 11 point scale with limitation of naming three or up to five most important alters were applied. Measurement technique was limited to free recall. Reliability of in- and out-degree, in- and out- closeness, betweenness and flow-betweenness was estimated by Pearson correlation coefficient. Meta analysis of factors affecting the reliability of measures of centrality and prominence was done by Multiple Classification Analysis. Apart from previous findings with regard to global and local measures, in- and out-measures, time between repetitions, we expect that reduction of number of choices would increase the reliability as only the most important actors are asked for.

MCMC Estimation of the p_2 Model

Bonne J.H. Zijlstra and Marijtje A.J. van Duijn
(University of Groningen, The Netherlands)

The p_2 model is a model for the analysis of social network data. The dependent variable is a (binary) network, or directed graph, on a given set of actors, or nodes; there can be explanatory variables on the levels of the actor and the ordered pair of actors. In the p_2 model, differences between actors in attractiveness and productivity (outgoingness) are modeled by random effects. The model includes variance parameters and a covariance for attractiveness and productivity, and parameters for density (log-odds of existence of a tie) and reciprocity; these parameters can be related to explanatory variables.

One way to estimate the parameters in the p_2 model is to use an IGLS estimation procedure that applies a first-order Taylor approximation of the non-linear link function. Such a procedure has been shown to sometimes underestimate variance parameters in non-linear mixed models.

A possible solution to this problem of underestimation is to apply simulation-based estimation techniques. We propose here an MCMC algorithm for Bayesian parameter estimation in this model, using Gibbs and Metropolis updating steps. Some preliminary results will be presented to illustrate the differences between the IGLS and the MCMC algorithms.

Centroid Rotation in Canonical Discriminant Analysis
Aleksandar Zorić (Faculty of Philosophy, Belgrade, Serbia)

Traditionally the canonical discriminant analysis is used for identification of taxons gained from some taxonomic technique. This identification is done by investigation of discriminant functions and centroids (means of taxons on these functions). A lot of different techniques were proposed to make interpretation of discriminant function easier, rotating them in some parsimonic position, as it is done in factor analysis. But interpretation of taxons is done by investigation of their centroids, too, and it is facilitated if centroids are as far as possible on all functions. This paper is suggesting a method for solving this problem as a method of rotation of centroid matrix of discriminant functions.
As discriminant problem is done by solving characteristic equation $(A-\rho_p^2 R)x_p = 0$, where the R is matrix of correlations among measured variables defined as $R = Z^t Z^{n-1}$; and A matrix of intergroup correlation defined as $A = Z^t S(S^t S)^{-1} S^t Z$ where S is selector matrix defined as function of belonging of each case to each taxon. If the $X = (x_p), p = 1, \ldots, b$ is a matrix of nontrivial eigenvectors, then the $K = ZX$ is a matrix of discriminant functions, and the $C = (S^t S)^{-1} S^t K$ is a matrix of centroids on those functions. Suggested method could be stated as orthogonal rotation of centroid matrix: $CT = D | f(D) = \mathrm{ext}, \mathrm{T^t T} = \mathrm{T T^t} = \mathrm{I}$ where function f could be some parsimonic function as Kaiser's varimax function which maximizes the column variance of rotated matrix (D).

Assessing Measurement Invariance in Cross-National Research of Business-to-Business Relationships: Confirmatory Factor Analysis Model
Vesna Žabkar and Barbara Žužel (University of Ljubljana, Slovenia)

Assessing of the applicability of frameworks cross-nationally opens the field for generalizations of theories in behavioural sciences. In the article we follow a coherent conceptual framework as well as a sequential testing procedure for assess-

ing measurement invariance in a cross-national research (Steenkamp and Baumgartner, 1998) to be able to compare business-to-business relationships cross-nationally. The Steenkamp and Baumgartner (1998) approach is based on multi sample confirmatory factor analysis, taking into account the conditions under which meaningful comparisons of construct conceptualisations, construct means and relationships between constructs are possible.

The instrument used to measure the theoretical constructs of interest should exhibit adequate cross-national equivalence. Although a variety of techniques have been used to assess various aspects of measurement equivalence, there is general agreement that the multi-group confirmatory factor analysis model (Joereskog, 1971) represents the most powerful approach to testing for cross-national measurement invariance. Although the evidence of measurement equivalence is important for scientific inference, it is often not found in cross-national research in behavioural sciences (Hui and Triandis, 1985).

There are extended lists of variables that researchers in developed market economies have used in examining different business-to-business relationships (Cannon and Perreault, 1999; Jap, 1999; Joshi and Stump, 1999; Morris et al., 1998; Wilson, 1995). From these variables, we test a conceptual model that depicts the linkages between cooperative norms, complementary competencies of the dyad, beliefs in interpersonal trustworthiness and profit performance/ competitive advantages. Our research interests are in examination of structural relationships among these constructs cross-nationally. Complementary competences and interpersonal trustworthiness are focal constructs that relate to cooperative norms and profit performance/competitive advantages constructs in a nomological net.

Measures of all constructs were developed based on literature review, conceptual domains and field interviews. The set of items for each construct was initially examined, in the pretest, with exploratory factor analysis to identify items not belonging to the specified domain.

Data for the research were gathered through a telephone survey. Institute for South-Eastern Europe of the Faculty of Economics Ljubljana (ISEE) funded the research. Medium sized and large companies (more than 200 employees), producers of consumer goods and durables (except raw-material industries) were interviewed in September and October 2001. The two countries included were Serbia and Croatia. From both countries, in more than 200 of the largest companies (204 in Croatia and 216 in Serbia), either the CEO, marketing director or sales director, was interviewed (over 70% response rate). Marketing research specialists and their trained subcontractors in each of the countries gathered the data. Based on the data provided we test measure equivalence within the confirmatory factor

analysis framework. In line with Steenkamp and Baumgartner (1998) we therefore demonstrate greater concern with measurement equivalence in cross-national research of business-to-business relationships.

References

1. Cannon, J.P. and Perreault, Jr., W.D. (1999): Buyer-seller relationships in business markets. *Journal of Marketing Research*, **36**, 439-460.
2. Hui, C.H. and Triandis, H.C. (1985): Measurment in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, **16**, 131-152.
3. Jap, S.D. (1999). Pie-expansion efforts: Collaboration process in buyer-supplier relationships. *Journal of Marketing Research*, **36**, 461-475.
4. Jöreskog K.G. (1971): Simultaneous factor analysis in several populations. *Psychometrika*, **36**, 409-426.
5. Joshi, A.W. and Stump, R.L. (1999): Determinants of commitment and opportunism: Integrating and extending insights from transaction cost analysis and relational exchange theory. *Canadian Journal of Administrative Sciences*, **16**, 334-352.
6. Morris, M.H., Brunyee, J., and Page, M. (1998): Relationship marketing in practice: Myths and realities. *Industrial Marketing Management*, **27**, 359-371.
7. Steenkamp, J.E.M. and Baumgartner, H. (1998): Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, **25**, 78-90.
8. Wilson, D.T. (1995): An integrated model of buyer-seller relationships. *Journal of the Academy of Marketing*, **23**, 335-345.

Using Constraints in Relational Subgroup Discovery

Filip Železný (Czech Technical University, Prague, Czech Republic)
Nada Lavrač and Sašo Džeroski (Jožef Stefan Institute, Ljubljana, Slovenia)

Developments in descriptive induction have recently gained much attention of researchers developing rule learning algorithms. These involve mining of association rules (e.g., the APRIORI association rule learning algorithm), clausal discovery (e.g., the CLAUDIEN system), subgroup discovery (e.g., the MIDOS and EXPLORA subgroup discovery systems) and other approaches to nonclassificatory induction aimed at finding interesting patterns in data.

We consider the task of subgroup discovery: given a population of individuals and a specific property of those individuals that we are interested in, find population subgroups that are statistically most interesting, e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest. While the goal of standard rule learning is to generate models, one for each class, inducing class characteristics in terms of properties occurring in the descriptions of training examples, in contrast, subgroup discovery aims at discovering individual patterns of interest.

Our approach adapts classification rule learning to relational subgroup discovery, which is achieved by (a) propositionalization through first-order feature construction, (b) incorporation of example weights into the covering algorithm, (c) incorporation of example weights into the weighted relative accuracy search heuristic, (d) probabilistic classification based on the class distribution of covered examples by individual rules, and (e) area under the ROC curve rule set evaluation. The main advantage of the proposed approach is that each induced rule with a high weighted relative accuracy represents a chunk of knowledge about the problem, due to the appropriate tradeo between accuracy and coverage, achieved through the use of the weighted relative accuracy heuristic.

The outlined strategy, implemented in the software system RSD (downloadable from http://labe.felk.cvut.cz/~zelezny/rsd), is heavily based on the use of constraints. Constraints play a central role in data mining and constraintbased data mining is now a recognized research topic. The use of constraints enables more ecient induction as well as focussing the search for patterns on patterns likely to be of interest to the end user. While many dierent types of patterns have been considered in data mining, constraints have been mostly considered in mining frequent itemsets and association rules, as well as some related tasks, such as mining frequent episodes, Datalog queries, molecular fragments, etc. Few approaches exist that use constraints for other types of patterns/models, such as size and accuracy constraints in decision trees. A general formulation of data mining involves the specification of a language of patterns and a set of constraints that a pattern has to satisfy with respect to a given database. The constraints a pattern must satisfy can be divided in two parts: language constraints and evaluation constraints. The first only concern the pattern itself, the second concern the validity of the pattern with respect to a database.

As for the language constraints, we apply them to define the language of possible subgroup descriptions. These are applied both in feature generation and rule induction. RSD generates a set of first-order features that are used as new attributes in the transformed (single-relational) form of the original multirelational

data. Here the language constraints are expressed through declarations of feature syntax, specifying namely the types and modes of arguments in the used predicates (relations). In addition, other language constraints can be specified. These are: the maximum length of a feature (number of contained literals), maximum variable depth and maximum number of occurrences of a given predicate symbol. Some language constraints are imposed by the system automatically, eg. the constraint that no produced feature can be decomposable into a conjunction of two independent (roughly: not connected by variable sharing) features. Such features would be redundant, as the subsequently used rule inducer is itself able to construct conjunctions. The language constraint of undecomposability plays a major role: it enables pruning the search for possible features without losing any solutions, thus providing often dramatic speedup of the algorithm. In the subgroup discovery (rule induction) phase, a language constraint employed is the prescription of a maximal number of conditions/features in the description of a subgroup. The evaluation constraints are also used in both phases of RSD. In the feature construction process, it holds that (a) no feature should have the same value for all examples and (b) no two features should have the same values for all examples. For the latter case, only the syntactically shortest feature is chosen to represent the class of semantically equivalent features. In the subgroup discovery phase, several evaluation functions are considered. These include accuracy, weighted relative accuracy (WRAcc), significance, and area under the ROC curve. Accuracy and WRAcc are used in optimization constraints, i.e., RSD looks for rules with high accuracy/WRAcc. In fact, they are used as heuristic functions in RSD. Significance is used in evaluation constraints, i.e., one can prescribe a significance threshold that rules have to satisfy. WRAcc may be used in a similar fashion.

RSD has so far been succesfully tested in several relational-data domains, including the well known benchmarks of East-West trains and King-Rook-King chess endgame as well as in real-world problems concerning telecommunication and mutagenicity of chemicals.

Acknowledgements. We are grateful to Peter Flach for the collaboration in the genesis phase of the RSD algorithm. F.Ž. is supported by the DARPA EELD grant F30602-01-2-0571 and the Czech ministry of education grant MSM 21230013. S.D. acknowledges the support of the cInQ (Consortium on discovering knowledge with Inductive Queries) project, funded by the European Commission.

Estimating Probabilities of Continuous Attributes in Naive Bayesian Classifier

Martin Žnidaršič (Jožef Stefan Institute, Ljubljana, Slovenia)
Janez Demšar and Blaž Zupan (University of Ljubljana, Slovenia)

Naive Bayesian classifier is a very simple, but successful machine learning method to obtain a predictive model from a set of classified data. Fast learning and classifying, robustness in respect to missing values and useful explanation (Kononenko, 1993) of classification are its main advantages. It classifies an unseen example E in a class with the highest probability, given the example. The estimation of this conditional probability is made using the formula:

$$p(C_k|E) = p(C_k) \prod_{i=1}^{a} \frac{p(C_k|A_i = v_i)}{p(C_k)}$$

where the attributes contributions are considered independent. This assumption of independence is the reason the method is called "naive".
There have been many attempts to evaluate and to correct the impact of the independence assumption. Effort has been also made to deal with the second major problem of the naive Bayesian classifier, the estimation of conditional probabilities for continuous valued attributes, that is the focus of our work.
The estimation of probabilities in the formula is straightforward for the nominal attributes, but in the case of continuous attributes it requires special attention, as there is no standard solution to follow. The probability $p(C_k)$ may be obtained from the training data set as the relative frequency of instances classified to class C_k. Similarly, for nominal attributes, to estimate conditional probabilities $p(C_k|A_i = v_i)$, we only need to count how many examples within those that have the value of i-th attribute equal to v_i are classified to the class C_k. For continuous attributes, the simplest and most frequently used approach is categorization (discretization), see, Kohavi and Sahami (1996). As it has some unwanted effects, for instance the abrupt changes in estimates at interval bounds, some methods were proposed that deal with continuous attributes without the drawbacks of categorization. Kononenko (1992) proposed fuzzy discretization, that fuzzyfies crisp probability boundaries at cutoff points. A different approach is to estimate a probability density function of continuous values, to be then used for probability estimation (Hastie et al., 2001). A Gaussian probability density function is often used, but a non-parametric kernel density estimation was also studied and proposed by John and Langley (1995).
Fuzzy discretization was compared to standard crisp discretization and the kernel

density approach was compared to Gaussian density approach, both showing some improvement in classification accuracy. We could not trace any report that would deal with cross-comparison of all of these methods.
We review several methods including fuzzy discretization and kernel densities approaches which use have been reported in the past and propose a new approach based on probability estimation through local regression. An experimental comparison of all presented methods is made and the results are discussed.

References

1. John, G.H. and Langley, P. (1995): Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference On Uncertainty in Artificial Intelligence*. Montreal, Quebec: Morgan Kaufmann, 338–345.
2. Kononenko, I. (1992): Naive bayesian classifier and continuous attributes. *Informatica*, **6/1**, 1-8.
3. Kohavi, R. and Sahami, M. (1996): Error-based and entropy-based discretization of continuous features. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 114–119.
4. Hastie, T., Tibshirani, R., and Friedman, J. H. (2001): *The Elements of Statistical Learning*. Springer Verlag.
5. Kononenko, I. (1993): Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, **7**, 317–337.

Network Analysis of Relationships among Marketing Research Firms and their Clients: The Case of Slovenian Market

Barbara Žužel and Vesna Žabkar (University of Ljubljana, Slovenia)

Over the last few years, there has been increased interest in business networks and their importance for business activities (Araujo & Easton, 1996; Moller & Halinen, 1999; Turnbull et al., 1996). In studying business networks, most of the relationships analysed have been vertical (buyer-seller) relationships. Less attention has been given to the relationships between competitors (Bengtsson & Kock, 1999). The reason is that cooperative relationships between vertical actors are easier to comprehend because they are built on a distribution of activities and resources among actors in a supply chain, while horizontal relationships are more informal and invisible. Additionally, vertical relationships often contain economic exchange, while horizontal relationships are built mainly on information and social exchange (Bengtsson & Kock, 1999).

The empirical focus of this study is the marketing research industry. Measured against the scale of most industrial companies marketing research firms in Slovenia are small (the largest firm has 35 employees). There are eight biggest companies covering around 90% of the market, with four biggest firms having over 10% market share each. There is medium competition present in the market. Marketing research firms compete with each other and also with companies own marketing research departments.

The purpose of this study is to describe and analyse the structure of the field through the study of horizontal (competitors) and vertical (buyer-seller) relationships among marketing research firms and their biggest clients. Since companies are interdependent with each other through inter-organisational relationships (Ritter, 2000), their interconnectedness is best analysed using network approach. In the paper, a research framework will be presented with preliminary results based on depth interviews with managing directors of selected research firms and clients in the industry.

Due to the small number of marketing research firms sampling of units (firms) will not be necessary; a complete population of marketing research companies in Slovenia will be analysed. Analysis will be done separately for different types of research. Realistic approach will be used for determining boundaries of the network. After exploratory research (depth interviews planned for the first half of September 2003) survey will be directed at managing directors of marketing research firms. In the next step, survey will be directed at biggest clients enumerated by marketing research firms. Quantitative methods used in this study are based on Knoke and Kuklinski (1982), Wasserman and Faust (1994) and Scott (2000).

On the basis of obtained results relationships between competitors (horizontal network) and marketing research firms and their clients (vertical network) will be analyzed. The structure of relationships (economic and non-economic exchange) and competition as seen by respondents will be revealed for each type of research (e.g. media research, advertising research, etc.). Economic exchange is expected to dominate in vertical relationships and non-economic exchange in horizontal relationships. Different effects of interconnectedness (Ritter, 2000) are expected to be revealed at the triad level (2 competitors, a client) of the research. It is also assumed that firms handle their relationships with larger competitors in one way, with smaller competitors in another way and with their clients in a third way. Outcomes are expected to add to understanding of business networks consisting of horizontal and vertical relationships.

References

1. Araujo, L. and Easton, G. (1996): Networks in socioeconomic systems:

A critical review. In D. Iacobucci (Ed.): *Networks in Marketing*, 63-107. Thousand Oaks: Sage Publications.

2. Bengtsson, M. and Kock, S. (1999): Cooperation and competition in relationship between competitors in business networks. *Journal of Business and Industrial Marketing*, **14**, 178-193.
3. Knoke, D. and Kuklinski, J.H. (1982): *Network Analysis*. Newbury Park: Sage Publications.
4. Moller, K. and Halinen, A. (1999): Business relationships and networks: Managerial challenge of network era. *Industrial Marketing Management*, **28**, 413-427.
5. Ritter, T. (2000): A framework for analyzing interconnectedness of relationships. *Industrial Marketing Management*, **29**, 317-326.
6. Scott, J. (2000): *Social Network Analysis*. London: Sage Publications.
7. Turnbull, P., Ford, D., and Cunningham, M. (1996): Interaction, relationships and networks in business markets: an evolving perspective. *Journal of Business and Industrial Marketing*, **11**, 44-62.
8. Wasserman, S. and Faust, K. (1994): *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.