Studies in Computational Intelligence 448

Araceli N. Proto
Massimo Squillante
Janusz Kacprzyk (Eds.)

Advanced Dynamic Modeling of Economic and Social Systems

Springer

Studies in Computational Intelligence 448

For further volumes:
http://www.springer.com/series/7092

Araceli N. Proto, Massimo Squillante,
and Janusz Kacprzyk (Eds.)

Advanced Dynamic Modeling of Economic and Social Systems

Springer

Editors
Prof. Araceli N. Proto
Laboratorio de Sistemas Complejos
Facultad de Ingenieria
Universidad de Buenos Aires
Buenos Aires
Argentina

Prof. Massimo Squillante
Faculty of Economics and Business Sciences
University of Sannio
Benevento
Italy

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

and

Department of Electrical and Computer Engineering
Cracow University of Technology
Cracow
Poland

ISSN 1860-949X
ISBN 978-3-642-32902-9
DOI 10.1007/978-3-642-32903-6
Springer Heidelberg New York Dordrecht London

e-ISSN 1860-9503
e-ISBN 978-3-642-32903-6

Library of Congress Control Number: 2012945327

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Introduction

In difficult times the world is facing nowadays there is an urgent need for more meaningful and more profound analyses of systems and problems that may be relevant. Of particular interest are here clearly all kinds of analyses of social and economic systems and problems because their concern is what primarily matters for all of us. A natural tendency in this context is to use formal, mathematical approaches. These approaches may concern various aspects of the economic and social systems and problems, and it is obvious that dynamics is omnipresent therein. The development of dynamic, time dependent models is therefore of a primary concern. In this context many diverse mathematical techniques can be employed that can be useful for the analysis of time series, prediction and forecasting, inference procedures, analysis of nonlinear dynamic systems, etc. Clearly, they all can help improve decision making which is a "meta-problem" of an universal importance.

In this volume recent advances in the use of modern quantitative models for the analysis of various problems related to the issues and challenges mentioned above, which can be described as the dynamics of social and economic systems, are presented. The majority of authors have employed tools and techniques of broadly perceived computational intelligence, notably fuzzy logic, evolutionary computation, neural networks and some non-standard probabilistic and statistical analyses. Due to a high complexity of the systems and problems considered, in many situations it is necessary to consider at the same time analytic, topological and statistical aspects and apply appropriate procedures and algorithms.

This volume is a direct result of vivid discussions held during the Fifth International Workshop on Dynamics of Social and Economical Systems (DYSES) was held at Benevento, Italy September 20-25, 2010, as well as a couple of post-workshop meetings and consultations. The meeting was organized by the Asociacion Dyses (Argentina), the Department of Analysis of Social and Economic Systems (Dases) of the University of Sannio (Italy), the Laboratorio de Sistemas Complejos, Facultad de Ingenieria-UBA (Argentina), the GECSI, Grupo de Estudio de la Complejidad Social en la Sociedad de la Informacion, the Facultad de Ciencias Juridicas y Sociales, Universidad Nacional de La Plata (Argentina), the Department of Mathematics and Building Sciences for Architecture, and the Territorial Town

Planning Laboratory, the University of Naples "Federico II", and the Istituto Italiano per gli Studi Filosofici. As one can easily see, though the topics of interest of the particular institutions are different, they concern problems of a high social and economic relevance and hence their analysis should be performed in as much as possible wide and profound a way. It should be noted that we have mentioned above the last DYSES Workshop, it has been just a consecutive one in a long series of DYSES Workshops, more specifically those held in 1998 in La Plata City, capital of the Buenos Aires state, at the Faculty of Engineering, and in 2001 at the Faculty of Juridical and Social Sciences, both at National University of La Plata. The 2005 DYSES was held at the Universidad CAECE (the Mar del Plata branch), and included as one of the Argentinean meetings of the World International Physics Year. The 2009 DYSES Workshop took place also in Argentina, at the coastal city of Pinamar. The consecutive DYSES Workshops have attracted an increased attention and have become true international events.

The primary aim of these meetings was to bring together scientists working on the development of time-dependent models which could help analyze, evaluate and forecast social and economic situations. Mathematical techniques have been widely included, including those for time series forecasting, inference procedures, nonlinear dynamic systems, decision making, decision analysis, uncertainty analyses, etc. As a result of a wider and wider interest of young researchers in the topics covered by the DYSES Workshops, it was decided at the 2010 DYSES Workshop to award best papers by yong reserachers, The members of the international jury composed, Serge Galam, Janusz Kacprzyk and Araceli Proto, awarded the young researchers Silvia Angilella, Bice Cavallo and Giuseppe Gaeta for their outstanding contributions.

Now, we will briefly summarize the consecutive papers included in the volume. We will shortly outline the essence of the problem considered and the tolls employed, as well as the results obtained.

Myriam P. Sassano and Araceli N. Proto, in **Housing prices wavelet entropy to understand macroeconomic states** have proposed wavelet entropy as an indicator of a macroeconomic situation using the past experience in order to extract information about the economic complexity of a country. The series of Argentinean historical construction indexes from 1970 to 2011 are taken as example. For that series, the wavelet entropy has been evaluated for the whole period each 12 and 24 and 32 months. High wavelet entropy variations can be detected in the periods 1974-78, 1985-89 and1990-92 which concide with internal or external political or economical instabilities. A comparison with other temporal series, national and international, is performed in order to justify why the wavelet entropy index provides relevant information about macroeconomic states.

In **Dynamical Models for Representing and Building Consensus in Committees,** Antonio Maturo, Massimo Squillante and Aldo G.S. Ventre are concerned with the problem of how to attain an equilibrium among the features that characterize the alternatives, or the objectives, that constitute the choice which a committee formed with the aim to make a decision of social or economic relevance, is called for to make. In some circumstances, the committee behaves as a unique body whose or-

gans, the experts, share the same opinions and select the same choice. When this occurs, the committee has reached an unanimous consensus. More frequently only a majority of the experts agree about a final choice and circumscribe a precise decision to be made. Also in this case we speak about consensus reached by, or within, the committee. Then authors investigate mechanisms for enhancing, and possibly reaching consensus by means of dynamical models, geometric and game theoretic in nature.

Networks are a topic of a particular interest in the analysis of socio-economic systems. So, the paper **Financial applications of Flow Network theory** by Mario Eboli reviews some recent applications of flow network theory to the modeling of financial systems and of interbank liquidity networks. Three features of the network flows have proven to be particularly useful in this field: i) the modularity of the transmission of flows across a network; ii) the constancy of a flow across all cuts of a network; and iii) the known 'max flow - minimum cut' theorem by Ford and Fulkerson. These properties of the flow networks have been applied to evaluate the exposition to contagion of financial networks and the carrying capacity of interbank liquidity networks.

Luisa Cutillo, Giuseppe De Marco and Chiara Donnini in the paper **Networks of Financial Contagion** discuss some issues related to the fact that banks develop relationships in order to protect themselves against liquidity risk. Despite this benefit, the fragility of financial markets stems from those interconnections. A cornerstone in the microeconomic analysis of contagion in financial systems is the contribution of Allen and Gale (2000). The present work takes up the challenge of generalizing the Allen and Gale contagion analysis to complex networks. Then, the authors provide an alternative procedure to construct a network of financial contagion which enables the analysis of complex networks via simulation. Their study shows that it is possible to find a minimal number of links to guarantee contagion resiliency, and that there holds the Allen and Gale conjecture that the system becomes more fragile as the number of links decreases

In **Rumour Propagation on Social Networks as a Function of Diversity** Bernard Brooks investigates the effect on the rumour propagation of the minority's distribution in a social network in the case of a social network comprised of two interconnected groups: a majority group and a minority group, The rumour is derogatory towards the minority group and its transmission is simulated using the GBN Dialogue model of rumor propagation on realistic social networks. The integration of the minority group into the entire social network is measured by the Minority Integration Metric (MIM). Monte Carlo simulations show that rumour penetration into the minority subgroup is an increasing linear function of the MIM and thus the minority members of the social network will have a higher level of belief in the minority derogatory rumour if the minority is more integrated into the social network. A threshold MIM value can be estimated at which the rumour fails to penetrate the minority subgroup.

Silvana Stefani and Anna Torriero in the paper **Formal and Iniormal networks in organizations** illustrate how to detect informal networks in a company through the study of a restructuring process of a company. A survey was conducted by ques-

tionnaires. The authors have made use of eigencentrality applied to the resulting graphs. Based on the concept of hubs and authorities an informal relationship within the company with a formal structure has been compared with comments on the roles and their significance.

In **Assessing consumer credit applications by a genetic programming approach** by Salvatore Rampone, Franco Frattolillo, Federica Landolfi, the purpose of the analysis is to develop tools and techniques to predict, on a collection of real loan data, whether a credit request has to be approved or rejected. Credit scoring is the assessment of risk associated with lending to an organization or an individual. Genetic programming, an evolutionary computation technique that enables computers to solve problems without being explicitly programmed, is employed and the authors propose a genetic programming approach for risk assessment. The problem is to use existing data to develop rules for placing new observations into one of the sets of discrete groups. The automation of such decision-making processes can lead to savings in time and money by relieving the work load on an "expert" who would otherwise consider each new case individually. The proposed model provides a good performance in terms of accuracy and error rate.

The paper by Janusz Kacprzyk, Sławomir Zadrożny and Tadeusz Baczko, **Towards a human consistent analysis of innovativeness via linguistic data summaries and their protoforms**, is motivated by the fact that in the present, highly competitive world one of key issues that determine the standing and prosperity of a country is innovativeness of companies, branches, industries, etc. In their paper, they are concerned with innovativeness at the national level in the perspective of, for instance, (Archibugi, Howells and Michie, 1999), many papers in (Llerena and Matt, 2004), (Malerba and Brusoni 2007) or (Malerba and Cantner, 2007) but taking into account some specifics of Poland. The work then is based on the use of some public indicators of innovativeness developed specifically for this purpose, along the lines of the Frascati Manual (OECD, 2003) and the Oslo Manual (OECD, 2005), and to international statistical standards according to which these indicators can be easily adjusted to specific requirements of different countries and regions. By using publicly available indicators a comparison of innovativeness can be easier and more objective, with a lower risk assessment cost. This can help rank companies, branches, etc. with respect to their innovativeness.

In **An investigation of computational complexity of the method of symbolic images** Giulia Rotundo starts by observing that a widespread presence of maps in discrete dynamical models implies the necessity to use efficient algorithms for their investigation. The method of symbolic images is more efficient than the exhaustive numerical simulation of trajectories because it transforms a map into a graph through a discretization of the state space, so it opens the way to the usage of graph algorithms and it provides a unified framework for the detection of system features. In this framework, a modification of the algorithm described by Osipenko et al. is proposed and its efficiency is analyzed. Some problems with the convergence of the method appear when the dynamical system is described by a no-Lipschitzian nonlinear map in the plane. It is shown the application of the method on an evolu-

tionary model of a boundedly rational consumer characterized by the presence of a denominator that can vanish.

In **Spacial database quality and the potential uncertainty sources** Šárka Hošková-Mayerová, Václav Talhofer and Alois Hofmann deal with one of the most significant features of geo-information systems, that is, the development of geospatial analyses that are based on fundamental geospatial data which model the landscape in the certain territory of interest. The analyses themselves are often described by a mathematical apparatus which uses tools and techniques of a wide range of branches of mathematics, especially vector analysis, differential geometry, statistics, probability, fuzzy logic, etc. The classical mathematical description of analysis is clear and precisely defined. Complex geospatial analyses, however, work above geospatial data that do not have to be homogeneous from the point of view of quality. With respect to the capacity and technological possibilities of the data supplier, the input data can have different level of geometric and thematic accuracy, their thematic attributes can remain unfulfilled or the data can be obsolete to a certain extent. The imprecision and uncertainty then influence the result of the complete geospatial analysis. The aim of the paper is to find a relation between the quality of input data and the reliability of the result of geospatial analysis. The authors' approach is based on mathematical models, notably the models of vagueness and uncertainty, as well as the models for quality evaluation.

In **Mathematical model used in decision making process with respect to the reliability of geodatabase**, Šárka Hošková-Mayerová, Václav Talhofer, Alois Hofmann and Petr Kubíček show how it is possible - thanks to the use of sophisticated analytical tools for the evaluation of data quality - to better understand geospatial data. Another aim is to assess the impact of data quality on the results of space analyses that are made out of them and that are a basis for such decision-making processes in which it is necessary to take into account an impact of the geographic environment. Organizations that are involved in the development of geospatial databases usually define the content of these databases (i.e. listing of geographical objects and their features) and the quality of data being saved (e.g. geometric, topological and thematic accuracy, level of standardization etc.). As the area of the land that is described with the use of geospatial data is usually significantly larger than the capacity and technological possibilities of the organization responsible, it is not possible to keep the defined content and its quality in the entire secured area on the same level. When performing the geospatial analysis it is therefore necessary to take into account the immediate quality level of data in the particular area and to have the technologies for determining the reliability of the result of the particular analysis.

Statistical methods can serve the purpose of describing and forecasting the behaviour of systems in various frameworks. For instance, in the paper **Complex networks topology: the statistical self-similarity characteristics of the Average Overlapping Index** Francisco O. Redelico and Araceli N. Proto discuss some statistical properties of the Average Overlapping Index (AOI) using the simulated complex networks. The AOI can be interpreted as a measure of local clustering properties of a node indicating the node robustness against an external perturbation. The AOI has been considered in many different disciplines such as computer sci-

ence, macroeconomics, nonlinear dynamics and opinion formation. It reflects the networks topology in the way that according to the complex network generation mechanism some AOI values became forbidden. For that reason the corresponding AOI set for each network has multi-fractal properties. This multi-fractal properties are capable to grasp the generation mechanism of the respective network. The support of the multi-fractal is also a fractal set.

Biagio Simonetti and Antonio Lucadamo in **Taxicab non symmetricalcCorrespondence analysis for the evaluation of the passenger satisfaction** apply the technique mentioned above to some analyses related to public transportation. The Taxicab Non Symmetrical Correspondence Analysis (TNSCA) is a technique which is more robust than the ordinary Non Symmetrical Correspondence Analysis (NSCA). TNSCA is a variant of the classical Correspondence Analysis (CA) for analyzing the two-way contingency table with a structure of dependence between two variables. In order to overcome the influence due to the presence of an outlier, TNSCA gives uniform weights to all points based on the taxicab singular value decomposition. The visual map constructed by TNSCA offers a clearer perspective than that obtained by correspondence analysis and it may be very useful in evaluating the satisfaction of public transportation passengers.

Enrico Ciavolino and Mariangela Nitti, in **Simulation study for PLS path modelling with high-order construct: A jobsatisfaction model evidence,** present a study on the high-order latent variables for the partial least squares path modelling (PLS-PM). A Monte Carlo simulation approach is proposed for comparing the performance of the two best known methods for modelling higher-order constructs, namely the repeated indicators and the two-step approaches. The simulation results, far from covering all the potential uses of the two approaches, could provide some useful suggestions to those researchers who intend to introduce a further level of abstraction in modelling the phenomenon of interest. An illustrative case study on the job satisfaction is reported in order to show how theoretical and statistical instances have to be taken into consideration when modelling higher-order constructs.

In **Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data** Michele Gallo states that, for the exploratory analysis of three-way data, the Parafac/Candecomp model (PC) is one of the most useful models to study three-way arrays when the data are approximately trilinear. It is a three way generalization of the PCA (Principal Component Analysis). The PC model is a common name for the low-rank decomposition of three-way arrays. In this approach, the three-dimensional data are decomposed into a series of factors, each related to one of the three physical ways. When the data are particular ratios, as in the case of the compositional data, this model should consider the special problems that compositional data imply. The principal aim of the paper is to describe how an analysis of compositional data by the PC is possible and how the results obtained should be interpreted.

In **Analyzing AHP-matrices by robust partial least squares regression**, Biagio Simonetti, Gabriella Marcarelli and Viviana Ventre consider the problem of how to compute the values for the weights of a comparison matrix in the AHP framework in order to eliminate the influence of outliers and propose an approach based on the

Robust Partial Least Squares (R-PLS) regression. A simulation study to compare the results with those obtained by using other methods for computing the weights is included.

We hope that the papers included in the volume will be of interest and use for many people interested and involved in the analysis of all kinds of relevant economic and social problems. Moreover, we hope that people interested in the computational intelligence tools and techniques will also find useful information on how those tools can be employed to analyze and solve the problems considered.

We wish to thank all the contributors for their great works, and we also wish to warmly thank Dr. Luca Cirillo for his very valuable involvement in collecting and managing the contributions included in this book.

Araceli N. Proto
Massimo Squillante
Janusz Kacprzyk

Contents

Modeling Economic and Social Systems, and Their Dynamics

Advances in Data Analysis for Economic and Social Systems

Housing Prices Wavelet Entropy to Understand Macroeconomic States

Myriam P. Sassano and Araceli N. Proto

Laboratorio de Sistemas ComplejosFacultad de Ingenieria,
Universidad de Buenos Aires, Argentina

Abstract. Wavelet entropy is proposed as an indicator of a macroeconomic situation using the past experience, in order to extract information about the economic complexity of a country. The Argentinean historical construction indexes series from 1970 to 2011 are taken as example. The wavelet entropy was evaluated for the whole period every 12 and 24 and 32 months, for that series. High wavelet entropy variations can be detected in the periods 1974-78, 1985-89,1990-92, in coincidence with internal or external political or economical instabilities. Comparison with other temporal series, national and international is made in order to justified why wavelet entropy index provide relevant information about macroeconomic states.

Keywords: Wavelet Entropy, Economic Series,Time Series Information Extraction.

1 Introduction

All countries are composed by, oversimplifying, individuals, companies and governments. Governments basic objective is the welfare of the individuals, which compose the society, as well as companies which have a fundamental role as jobs producers. The best situation for a any governemt is when enterprises and indivuals get their maximum profit. On the other side, the extensive use of information technologies has introduced changes all over the society, making that individuals, companies, governments interact at great speed between themselves as well as internally (horizontally and vertically). So, presently, countries/societies function as a complex system , where each decision affects the whole system. This is one of the reasons why the decision making processes become more difficult for all the statements [1]. In this spirit, to project future sceneries, using models for decision making and/or social options, or evaluating the socio-economical data available, would be a paliative even not the universal panacea.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 1–9.
springerlink.com

The main goal of this contribution is to introduce the wavelet entropy as a mesure of the "health" of a given maket, using measured temporal signals "emited" for it. We here use the Shannon [2] information concept to interpret wavelet entropy results. In informational terms, high wavelet entropy means less information, and information lacks affects the degree of certainty of the decisions taken into a given market. When several markets present a high wavelet entropy, the general economics health of the country is in danger as economics decisions are taken in a more uncertain context in all of them. As markets are connected at real time and functioned in a nonlinear (complex) way, the existence of decisions under great uncertainty extend rapidly to other markets. Even when the analysis of time series to make "predictions" based on past situations, is not new, here we added the wavelet entropy as a tool, to reinforce the decision making reliability of the involved actors .Other important advantage when using wavelet entropy (WE) is that series evaluted as sequences of numbers, indexes, percentages, can be treated withiut requiring previuos treatments, as we shall see.

In order to achieve our purpose we select: 1) the Argentinean historical contruction price index (CPI) series from 1970 to 2011[3] for the following reasons: a) they cover 40 years, b) the construction market have no special regulation, further than taxes, so, it could be considered as a "pure" market as no political influences have been introduced, or at least they are not evident; c) to have its own house/appartment is a clear characteristic of the Argentinean middle class.2) The General Activity Index (GAI, Argentina), 3) the Square Meter Cost (SMC) and 4) the Global Food Price Index ($GFPI$)

2 Brief Review about Wavelets

A low pass filter greatly reduces the high frequency components, which often represent noise in the signal, being this is the guiding idea in many problems of signal analysis and synthesis. The main problem is to reconstruct the signal, removing noise, but not the relevant information. In general, the problem is solved using two filters, a high pass (HPF) as well as low pass filter (LPF). This process generates a "filter bank", which sends the signal trough two or more filters, normally structured in tree. Through the Discrete Wavelet Transform (DWT) and the Inverse Discrete Wavelet Transform ($IDWT$) it is possible to decompose and reconstruct the signal. In fact, the word "wavelet " is properly associated with a "multiresolution into different scales". The simplest change of scale comes from down sampling a signal keeping only its even –numbered components. This sampling operation is denoted by the symbol $(\downarrow 2)$, and it is possible to use different scales (fine, medium, and coarse). The overall filter bank Fig. 1 is the successive application of the $(\downarrow 2)$operation, L times. In theory, L can be even infinite, but as in reality

signals have finite length, necessarily the process can done up to the number of samples. So, the basic wavelet idea is to repeat the filter bank: the low pass output becomes the input to the second filter bank.

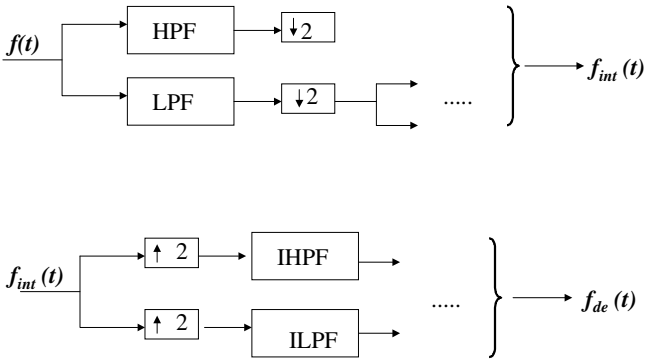


Fig. 1 The WDT filter bank

After the complete down-up filtering the f_{de} signal is obtained. Typical applications of the wavelet transform go to four or five decomposition levels. This completes the iteration of the analysis bank in down sampling. The filtering followed by down sampling is inverted by up sampling $(\uparrow 2)$, followed by filtering. In order of giving a brief idea of the method we are here using, we begin with the fundamental dilation equation or relation between scales also called refinement equation:

$$\Psi(t) = 2\sum_{k=0}^{N} h(k).\Psi(2t - k) \tag{1}$$

$\Psi(t)$ is called the scale function. If $\Psi(t)$ is a solution, so is any multiple of $\Psi(t)$ is a solution too. As usual the normalization condition is $\int \Psi(t)\,dt = 1$. The integral extends over all time, but it is possible to show that the solution $\Psi(t)$ is zero outside the interval $0 \leq t \leq N$. The scaling function has compact support, and the localization of the function is one of its most useful properties. The two key operations of classical wavelet theory are: Dilation of the time scale by 2 or 2^j, which means that the graph of $\Psi(t)$ is compressed by 2, and translation of the time scale by 1 or k, which means that the graph of $\Psi(t)$ is shifted by k. The combination of those operations, from to $2^j t$ to $2^j t - k$, will later produce a family of wavelets $\omega\left(2^j t - k\right)$ from a single wavelet . The scaling function is smoother and live in $\mathbb{L}^2(R)$. If the original input is a box function, we get $\Psi^{i+1}(t)$ from $\Psi^i(t)$, and the filter re-scale results:

$$\Psi^{i+1}\ (t) = 2 \sum_{k=0}^{N} h\,(k)\,.\Psi^{i}\,(2\,t\ -\ k) \qquad 1 \tag{2}$$

where $h\,(k)$ are the decomposition coefficients. This operation is called the cascade algorithm, because the low pass filter is "cascaded" to different time scales. If this iteration converges, $\Psi^{i}\,(t)$ approaches to $\Psi\,(t)$ in the limit $t \longrightarrow \infty$ and then $\Psi\,(t)$ satisfies Eq.(1) or refinement equation. $d_{jk}\ =< f(t), \omega_{\,jk} >$,taken into account the orthogonality properties of the basis into the corresponding spaces.

Applying the method described above, using the Daubechies four wavelet [4] , a succession of wavelets coefficients is obtained for the different decomposition levels taken into account. This procedure generates what is called the multiresolution analysis, and the series could be recomposed applying the inverse transformation. Series can be then written as:

$$\begin{aligned} S\,(t) &= 2 \sum_{j-N_j}^{-1} \sum_{k} C_j(k)\,.\Psi_{jk}\ (2\,t\ -\ k) \\ &= 2 \sum_{j-N_j}^{-1} r_j\ (t) \end{aligned} \tag{3}$$

where the $C_j(k)$ can be interpreted as the residual error among the successive approximations of the signal in the $j,,j \div 1$scales; $r_j\ (t)$ is the residue of the signal at the j scale. The coefficients and the energy in each decomposition level will be the energy (information) of the signal [5, 6] obtained as:.

$$E_{\,j} =\| r_j \|^2 = \sum_{k} |\ C_j(k)\ |^2 \tag{4}$$

where $C_j(k)\ = \langle\ S,\ \Psi_{jk}\ \rangle$ and the total Energy is:

$$E_{tot} =\| S \|^2 = \sum_{j\ <\ 0} \sum_{k} |\ C_j(k)\ |^2\ = \sum_{j\ <\ 0} E_{\,j} \tag{5}$$

Finally the normalized wavelet energy is define as $p_j = \frac{Ej}{E_{\,tot}}$. This p_j will be considered as the wavelet energy distribution function for all $j = -1, \ldots -N_j$ values [5, 6], and by analogy with the Shannon entropy, it is possible to write a wavelet entropy as:

$$S_{wt}\ = -\sum_{j\ <\ 0} p_j\ \ln\ (\ p_j\ ) \tag{6}$$

As in [6] we use the S_{wt} as a measure of *information or energy contained on the significative patterns* in the analyzed series.. Eq. (6) provides the fundamental expression our data analysis.

3 Data Analysis

We start with the CPI. In Fig 2 the original series is shown.

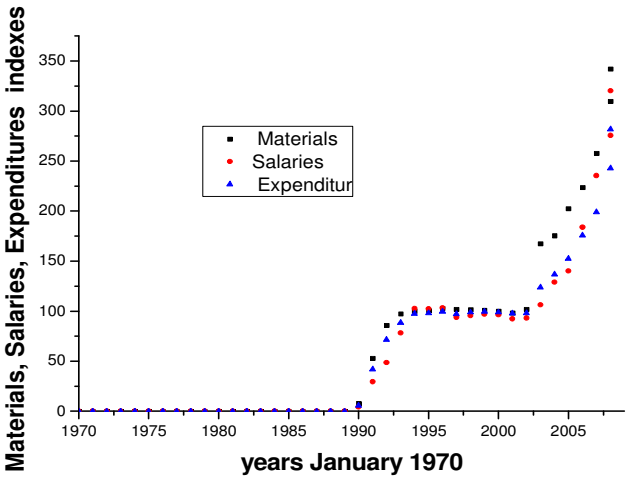


Fig. 2 CPI series period 1970-2011

Fig 2 shows the S_{wt} values for 12, 24, and 36 months windows, and in Fig.4 diferent temporal periods are shown for the 12-months window.

In Fig. 2 the relation between the CPI and the GAI is compared for the time period for which both data are available. As it could be seen there is not an evident realtionship between both indexes, concerning information or energy series ammount, meaning that, as we said in the introduction that the construction market is a sort of "pure" market, regulated by buyers and sellers. However, the S_{wt} values also shown that it decays strongly around 2002 (2001 external debt default), as the internal situation was extremely severe.

Anyway the construction market recovers its media value around 0.68 (see Fig.5) more quickly, than S_{wt} for IGA values.

In Fig.2 we show the SMP (middle), its entropy (upper plot) and the 12 months entropy in the corresponding period. It could be seen that the 80' hyperinflation processes (also '91 hyperinflation) were more important in terms of information or energy contents, than the impact of the Argentinian external debt default.

In Fig 7 we compare the 12 months window entropy with the Global Food Price ($GFPI$). We have taken a 5 months window for the $GFPI$ data in order to obtain more entropy points.

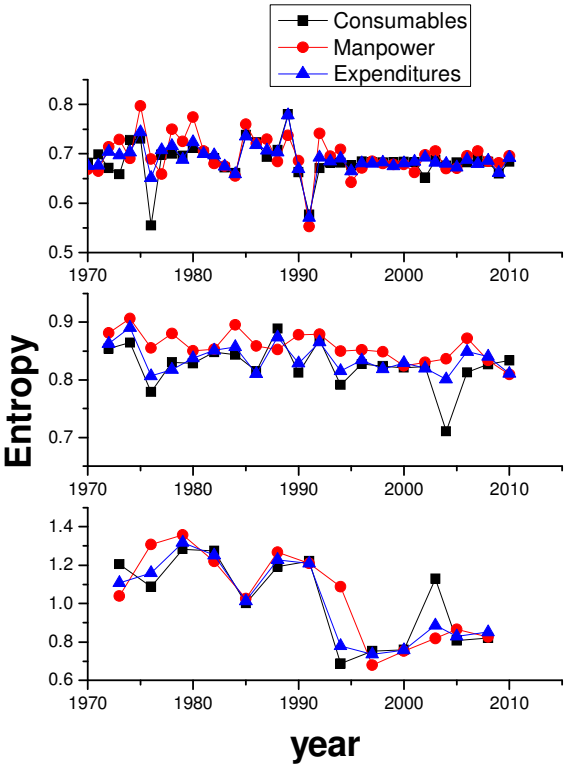


Fig. 3 S_{wt} for time windows of 12, 24 and 36 months for the CPI series (1993 = 100)

4 Conclusions

Wavelet Entropy criteria was applied to analyze: 1) the Argentinean historical contruction price index (CPI) series from 1970 to 2011 [3] 2) The General Activity Index (GAI, Argentina), 3) the Square Meter Price (SMP) and 4) the Global Food Price Index ($GFPI$). From the S_{wt} plots (Fig. 3) we can see that during the last around 15 years the S_{wt} values remain basically constant in mean value. Taking into account only this plot, it would be possible to say that the construction market will not present important turbulences in the future However,.high wavelet entropy variations can be detected in the periods 1974-78, 1985-89,1990-92, in coincidence with internal or external political or economical instabilities. The wavelet entropy was evaluated for the whole period every 12 and 24 and 36 months,time windows Also it is easily seen that the 12 months S_{wt} reflects with more fidelity the market instabilities. Further, the construction market seems to be not very related to the General Activity Index, as it can be observed in Fig.5. Looking Fig.6 the

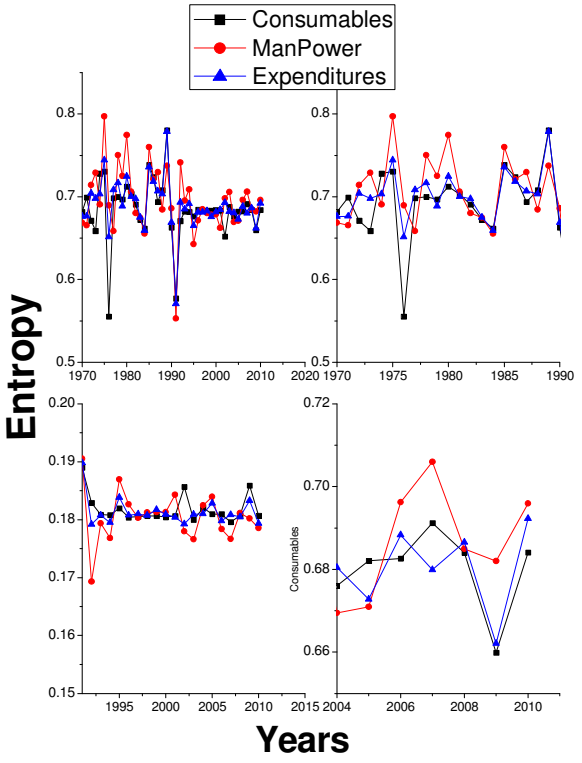


Fig. 4 S_{wt} 12months window, plotted for differen ttime periods

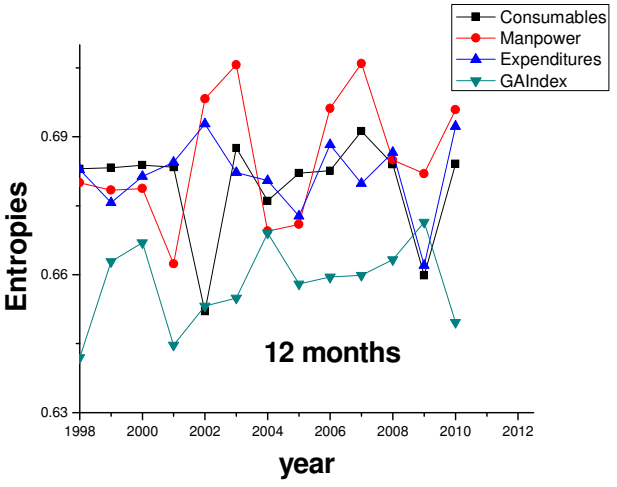


Fig. 5 S_{wt} comparison between 12 months CPI and IGA series

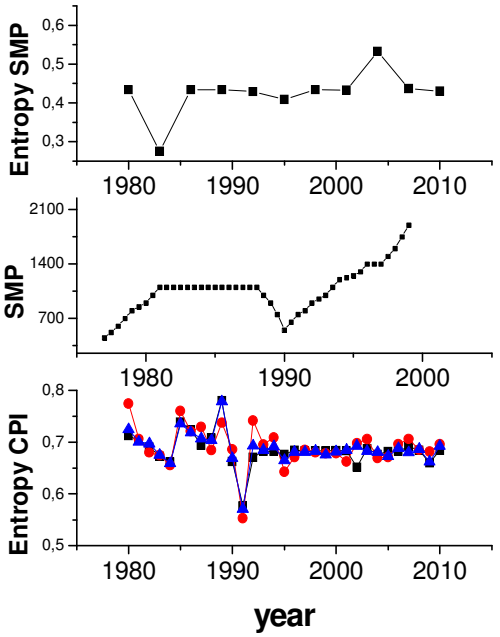


Fig. 6 S_{wt} of SMP (upper), CPI (lower) both for 12 months window and the original SMP series (middle)

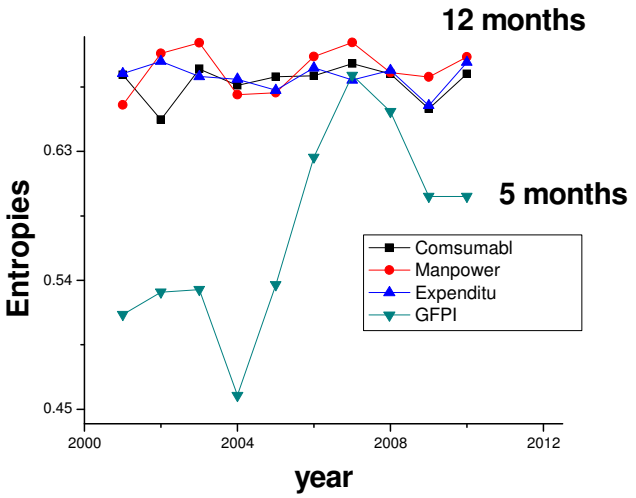


Fig. 7 S_{wt} comparison between 12 months window CPI and $GFPI$ (5 months window). The original GFPI data have been taken from on 1995 = 100 basis.

S_{wt} shows the important square meter value increment is not enough yet to produce alterations in the market. However, the local opinions of economists, construction developers and bankers are not homogeneous. This fact recognizes several resons: a) credit for buying a property are scarse and expensive for the "middle class" and the rentability of investing in a property for rent decays in the last months; most of the new buildings under construction are "high premium" with square meter prices around 6-7000 u$a. Some future constructions are projected for prices even higher. On the other side, the international price of food generates what are called the "sojadollars", and from this side the increasing construction prices can be explained as following the rising of food prices. However, in Fig. 7 the ammount of informational entropy required to buy food changes dramatically after the global 2008 crisis. Considering that S_{wt} nmeasures the ammount of information or energy contained in a temporal series, it is seen that after 2008 the ammout of information/energy to get food (a daily action) is more or less as "expensive" as to buy an apartment, which is done few times during the individuals life. Another ingredient to take into account is that the square meter in Buenos Aires is still very cheap in comparison with i.e. Shangai for an international investor.

Among the advantages of the method the more important is that diferent series coming from different indexes, percentages, or numbers can be compared straightfoward. However, the selection of time windows to apply the wavelet transform is a matter still under our analysis.

Summarizing: up to March 2011 we cannot adhere to the theory that the construction market is in a "bubble" state, although the severity of the global crisis reflected in Fig. 7 seems to be non sustainable. So, if some problem appears in this market, in the near future, it would be expected more from the external than the internal economic conditions.

References

1. Porter, M.A.: Rep. Prog. Phys. 64, 1165–1189 (2001)
2. Caiafa, C.C., Sassano, M.P., Proto, A.N.: Physica A 356, 172–177 (2005)
3. Ballentine, L.E.: Phys. Rev. E 63, 056204 (2001)
4. Daubechies, I.: Ten lectures on Wavelets, p. 3. SIAM, Philadelphia (1992)
5. Mallat, S.: A wavelet Tour of Signal Process sing, 2nd edn. Academic Press (1998)
6. Kowalski, A.M., Martin, M.T., Rosso, O., Plastino, A.: A.N. Proto Phys. Lett. A 311, 180-191 (2003)

Dynamical Models for Representing and Building Consensus in Committees

Antonio Maturo[1], Massimo Squillante[2], and Aldo G.S. Ventre[3]

[1] Department of Social Sciences, University of Chieti-Pescara, Chieti, via dei Vestini, 31, 66100 Chieti, Italy
e-mail: amaturo@unich.it
[2] Faculty of Economics and Business Sciences Sciences, University of Sannio, Via delle Puglie, 82, 82100 Benevento, Italy
e-mail: squillan@unisannio.it
[3] Department of Architecture and Benecon Research Center, Second University of Napoli, via S. Lorenzo, 1, 81031, Aversa, Italy
e-mail: aldoventre@yahoo.it

Abstract. When a committee of experts is formed with the aim to make a decision of social or economic relevance, the various competencies act in order to produce an equilibrium among the features that characterize the alternatives or the objectives, that constitute the choice that the committee is called to make. It is worth to remark that, in some circumstances, the committee behave as a unique body, whose organs, the experts, share the same opinions and select the same choice. When this occurs, the committee has reached unanimous consensus.

More frequently only a majority of the experts agree about a final choice and circumscribe a precise decision to make. Also in this case we speak of consensus reached by, or inside, the committee.

The mechanisms for enhancing, and possibly, reaching consensus are here studied by means of the definition of dynamical models, geometric and game theoretical in nature.

Keywords: Multiperson Decision Making, Consensus, Cooperative Games
2000 MSC: 90B50, 91B10, 91A12.

1 Introduction

Collective decisions are usually given to the responsibility of suitable committees of experts. In particular when the involved acts of choice are related with environmental, social, or economical issues.

Static procedures and dynamic procedures are defined (see e.g. [1], [3], [4]), [5]) in order to organize knowledge, synthesize individual judgments in collective, enhance and evaluate consensus in a group. In fact, an important

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 11–19.
springerlink.com

step in a consensus procedure is the evaluation of the degree of consensus, i. e. the number of individuals decision makers that form a majority: the greater the majority, the higher the degree of consensus.

The maximum possible degree of consensus is, of course, unanimity, what is rare, especially if group decisions that have to be made are related with domains which embody complex issues such as environment, social conflicts, financial crises.

When facing subjects of such a complexity the inputs, that promote and breathe life into the debating group, arise from both the necessity of cooperating and the need of competition and fighting.

The individuals, or groups, in the committee move and search for stability; there are seen as points in a space, that change positions and mutual distances. In a metric space distances between opinions are measured and reveal to what extent a decision maker changes his mind.

In this perspective we deal with a game theoretical model for dynamical consensus searching. Indeed, we mean a consensus procedure as shaping a winning coalition in a cooperative game, where:

1. the decision makers are the players,
2. utility transfer is allowed,
3. only suitable coalitions are admissible.

2 Requirements for a Definition of Consensus

Particular characteristics are needed for a better understanding, or a definition, of the concept of consensus. Let us sketch some of these requirements;

1. Sometimes consensus may be reached immediately, just with the presentation or formalization of the problem, either by means of the unanimous and immediate agreement of the decision makers, or by means of a suitable procedure that rules the search for consensus, e. g., a static average procedure [1], 1987], that works like a black box. In both cases it is a matter of *immediate*, or *one shot*, *consensus*.
2. In a *dynamic procedure*, consensus is related, or determined, by suitable behaviors, such as compromise or agreement. Indeed, consensus should develop like the formation of the opinions or convictions during the debates, or discussions, among persons. Therefore a dynamic procedure soliciting consensus results in a trade-off between agreement and compromise, related with individual decision makers or groups inside the committee. Such a behavior gives rise to movements toward consensus. Any movement is leaded by desire and rationality to get a goal. In other words, to simplify, decision makers are seen as bodies or points that move with their ideas and willingness.
3. *Perfect information* is needed for reaching consensus. Any member in the group knows every act and the behavior of any decision maker.

An additional figure plays a role in our model, that is the *supervisor*, or *chairman*. He/she coordinates the group decision process; the action of the supervisor is *technical*.

3 Evaluation of Resources and a Static Model

A group charged of the duty of reach a sufficiently shared decision, i. e. a decision endowed of a suitable consensus, must first know the elements that are the objects of the judgements. In other words the group must *evaluate the resources*.

Then the group proceeds to the *evaluation of the evaluations*, what leads to enhancing consensus.

In order to evaluate the resources and activate consensus enhancing processes, the decision makers turn to Group Decision Support Systems.

In particular, the achievement of consensus is an objective for *Cooperative Work*. In general, decision makers use *Decision Aids*.

An amount s of quantifiable resources, such as raw materials, energy, money for industrial and scientific projects, grants, must be allocated over m projects. To this purpose a committee of n experts is formed. The expert i recommends to allocate the *amount* x_{ij} over the project j. Then we have the constraints

$$x_{i1} + x_{i2} + ... + x_{im} = s, i = 1, 2, ..., n. \tag{1}$$

Set $\mathbf{x_j} = (x_{1j}, x_{2j}, ..., x_{nj})$, and $\mathbf{s} = (s, s, ..., s)$. Then the system of equations (1) assumes the form:

$$\mathbf{x_1} + \mathbf{x_2} + ... + \mathbf{x_m} = \mathbf{s}. \tag{2}$$

The recommendations of each expert must be aggregated, in a unique consensus allocation [1] by a chairman or external authority: again the shares sum up to s.

The aggregated allocation satisfies suitable reasonable requirements, such as:

- the dependence of the project j only from the recommended allocations by the experts for project j;
- if all the experts recommend to reject the project j, then consensus about the allocation of resources to project j is 0.

Then we assume, for every project j, there is an *aggregation function* $f_j = f_j(y_1, y_2, ..., y_n)$, with values in the set R^+ of nonnegative real numbers, where $y_i = x_{ij} \in [0, s]$ is the allocation proposed by the expert i.

It seems reasonable to assume the following requirements on the functions f_j:

1. $f_j(\mathbf{0}) = 0$;
2. f_j is increasing with respect to every variable;
3. $\sum_{j=1}^{m} \mathbf{x_j} = \mathbf{s} \Rightarrow \sum_{j=1}^{m} f_j(\mathbf{x_j}) = s$.

An important result is the following Aczel's theorem:

Theorem [1] The general solution $(f_1(\mathbf{y}), f_2(\mathbf{y}), ..., f_m(\mathbf{y}))$ of the system of conditions 1., 2., 3., is given, for $m > 2$, by

$$f_1 = f_2 = ... = f_m = \sum_{i=1}^{n} \alpha_i y_i, \tag{3}$$

where:

$$\forall i, \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1. \tag{4}$$

4 Geometrical Representation and Dynamical Model

A dynamical model for building consensus has an operational semantics close to the meetings and discussions in the real life (see [3], [6], [7], [8], [9], [13], and [16]).

Let us sketch a description of a behavioural model of a group of experts that must reach a satisfying internal agreement, or consensus, in order to provide for an aid to the decision to be made, for instance, in a political framework.

Let us suppose that any expert has the opportunity to express his evolving opinion. A coordinator (*chairman*, *supervisor*) solicits the experts to get a satisfying level of consensus, by iterating a loop:

1. at a certain moment the coordinator summarizes the discussion,
2. he makes the point of the current state of consensus,
3. confirms characteristic opinions,
4. stresses conflicting points of view,
5. diplomatically reminds the necessity to maintain united the group.

As a result, the members that are at the opposition, fearing to be emarginated by some majority, react asking for clarifying, or modifying, their opinions in order to weaken the tension between their viewpoints and what defines the current status of consensus. This dynamics runs until the degree of consensus does not seem sufficient to the coordinator, or alternatively, the consensus does not seem, at present, achievable and discussion is postponed.

The growth of consensus is due also to the increase of communication, database, knowledge, technology.

In some models (see, e. g., [3]) the Analytic Hierarchy Process (AHP) [17] is assumed to be the system used by all the experts for ranking alternatives. Of course different evaluation schemes can be considered (see, e. g., [2]).

Let us describe a model of collective decision making and the related consensus achievement procedure (see also, e. g. [3], [9], [10] and [11]). Let D denote the set of the decision makers of a committee, A the set of alternatives, and C the set of the accepted criteria. Let any decision maker $d_i \in D$ be able to assess the relevance of each criterion. Precisely, every d_i assigns a function

$$h_i : C \to [0, 1] \tag{5}$$

such that

$$\sum_{c \in C} h_i(c) = 1. \tag{6}$$

Remark that h_i just denotes the *evaluation* or *weight function* that the decision maker d_i assigns to every criterion $c \in C$.

Furthermore let us consider the function

$$g_i : A \times C \to [0, 1] \tag{7}$$

such that $g_i(a, c)$ is the value of the alternative a with respect to the criterion c, in the perspective of d_i.

The values $h_i(c)$ and $g_i(a, c)$ can be determined by suitable procedures, such as dealt with, e. g., in [2] or [17]. Let n, p, m, denote the numbers of the elements of D, C and A, respectively. The values $(h_i(c))_{c \in C}$ denote the evaluation of the p-tuple of the criteria by the decision maker d_i and the values

$$(g_i(c, a))_{c \in C, a \in A} \tag{8}$$

denote the matrix $p \times m$ whose elements are the evaluations, made by d_i, of the alternatives with respect to each criterion in C.

Function $f_i : A \to [0, 1]$, defined by the scalar product

$$(f_i(a))_{a \in A} = ((h_i(c))_{c \in C}) \cdot (g_i(c, a))_{c \in C, a \in A}) \tag{9}$$

is the evaluation, made by d_i , of the set of alternatives $a \in A$.

Dynamics of consensus enhancing process is managed by an external supervisor that has at his disposal a metric μ, e. g. an Euclidean metric, that acts between couples of decision makers d_i and d_jj, i. e., between individual rankings of alternatives, defined by

$$\mu_{ij} = \mu(d_i, d_j) = \sqrt{\frac{\sum_{a \in A} (f_i(a) - f_j(a))^2}{|A|}}. \tag{10}$$

If the functions h_i, g_i range in $[0, 1]$, then $0 \leq d(i, j) \leq 1$. Hence the decision maker set D is represented by a set of points of the unit cube in a Euclidean space E^m.

The supervisor observes, at any step of the decision making process, the position of each member in the committee and informs the more peripheral expert about the opportunity to revise his judgement.

If we set

$$\mu^* = \max\{\mu(i,j)|i,j \in D\},$$

then a measure of the degree of consensus γ can be defined as the complement to one of the maximum distance between two positions of the experts:

$$\gamma = 1 - \mu^* = 1 - \max\{\mu_{ij}|d_i, d_j \in D\}. \tag{11}$$

5 A Game Theoretic Point of View

We now relate consensus with the construction of a winning coalition in a cooperative game where players are decision makers and utility transfer is allowed. A different game theoretic point of view in dealing with consensus is developed in [12].

We assume given an integer k such that $1 < n/2 < k \leq n$. Let us define a *majority at level* k as a set of decision makers having at least k elements.

Moreover we assume that the members of the committee are points of a metric space (S, μ). In particular, following the notation given in Sec. 6, every decision maker d_i is a point $(f_i(a))_{a\in A}$ of the space R^m with the metric μ given by (10).

Let now δ be a positive real number. We say that q members in the committee agree at level δ if they belong to a ball of diameter in the metric space (S, μ). Reaching consensus can be interpreted as a cooperative game with side payments in which the admissible coalitions are the ones contained in at least a ball of diameter δ of the space (S, μ).

The important concept of admissible coalition was considered in [14]. Admissibility was related with constraints that were in nature ethical, social, etc. Our point of view is a geometrical interpretation of admissibility concept in order to describe possible modifications of coalitions; what is studied, with different methods, also in [14].

Let K be the set of the admissible coalitions. The set K is not empty because every singleton is contained in at least a ball of diameter δ. We can introduce the following classification of the elements of K. In order to get consensus, a coalition H, whose elements are in number of $|H|$, belonging to K, is said to be:

- *winning*, if $|H| \geq k$;
- *losing*, if the coalition $H^c = D - H \in K$, contains a winning coalition;
- *quasi-losing*, if $|H^c| \geq k$, but H^c does not contain any winning coalition;
- *blocking*, if $|H| < k$ and $|H^c| < k$.

It is worth observing that, while in simple games [18] the whole group of players is a winning coalition, in our framework a coalition with at least

k members is winning if and only if it is included in a ball of diameter δ. Winning and losing coalitions were studied in [18], where all coalitions are considered admissible; whereas we assume as admissible only the coalitions satisfying suitable geometric constraints.

Unlike the classical game theory in which we look for minimal winning coalitions [18], finding the consensus means to look for maximal winning coalitions.

We introduce the following further

Definition. Every maximal winning coalition of K is said to be a *solution* of the consensus reaching problem.

One of the following cases occurs:

1. $D \in K$;
2. $D \notin K$, but the consensus problem has a unique solution;
3. $D \notin K$ and the consensus problem has at least two solutions;
4. $D \notin K$ and the consensus problem has no solutions.

In the case (1.) the consensus is reached and the global score of every alternative is obtained by considering a mean of the scores assigned by the decision makers in D.

In the case (2.), if H is the unique maximal solution, the chairman either assumes H as the set D^* of decision makers that give rise to the group decision, or tries to enlarge the set D^*, by persuading some members of $H^c = D - H$ to change their evaluations.

Let us use the procedure introduced in [3] and [9], we call the *Bastian procedure.*

If an element of H^c moves in a maximal winning coalition B, it is possible that B does not contain H. If it happens we fall in the case (3.) and the coalition H may be broken.

In the case (3.) we can use the Bastian procedure as a dynamical procedure to enlarge maximal winning coalitions. The aim is to obtain a unique final maximal winning coalition D^*.

A situation can happen where the players asked by the chairman to change their evaluations can make strategic choices in order to break some coalitions, and cut off some other players to participate to the final aggregation of evaluations.

An alternative to the Bastian procedure is to decide that the maximal winning coalition to enlarge is the one with the maximum number of players. If there are more winning coalitions with these properties, the coalition to be enlarged is the one, if it is unique, that is included in a ball of minimum diameter.

We can also consider fuzzy coalitions [15] and their fuzzy width.

Also in the case (4.) we can use the Bastian procedure, but it is not sure that there is a step in which we have at least a winning coalition, and so it is possible that there is not a solution of the consensus reaching problem.

Some difficulties for the role of the chairman can arise by the blocking coalitions. These coalitions may prevent to obtain solutions in the case (4.) and may give rise to serious problems to the power of the chairman by using the mentioned procedures. Then the existence of blocking coalitions may induce some corrections to our procedures, such as an activation of a form of bargaining or an evaluation of the power of these coalitions.

6 Conclusion

Let us further interpret our model in terms of a metaphor. The way to construct consensus about a social decision usually depends on the particular subject. Psychological and individual propensities and needs are routed in the behaviours of any decision maker in the committee. Each member in the group embark on his way, or programme, but soon he/she has to take into account also all the others' ways that become more or less apparent in time.

Topology provides mainly for the syntax, that explains the formal rules of changing opinions; while game theory plays the role of a semantics when gives intrinsic meanings and motivations to the members of a coalition to modify their thought or feeling.

References

[1] Aczél, J.: A short course on functional equations. D. Reidel Publishing Co., Dordrecht (1987)
[2] Bana e Costa, C.A., Vasnik, J.C.: The MACBET approach: basic ideas, software and an application. In: Meskens, N., Roubens, M. (eds.) Advances in Decision Analysis, pp. 131–157. Kluwer Academic Publishers, Dordrecht (1999)
[3] Carlsson, C., Ehrenberg, D., Eklund, P., Fedrizzi, M., Gustafsson, P., Lindholm, P., Merkurieva, G., Riissanen, T., Ventre, A.G.S.: Consensus in distributed soft environments. European Journal of Operational Research 61, 165–185 (1992)
[4] Maturo, A., Ventre, A.G.S.: Models for Consensus in Multiperson Decision Making. In: NAFIPS 2008 Conference Proceedings. IEEE Press, New York (2008)
[5] Maturo, A., Ventre, A.G.S.: Aggregation and consensus in multiobjective and multiperson decision making. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 17(4), 491–499 (2009)
[6] Dalkey, N.C.: Delphi. Rand Corporation, New York (1967)
[7] Dalkey, N.C., Helmer, O.: An Experimental Application of the Delphi Method to the Use of Experts. Management Science 9(3), 458–467 (1963)
[8] Delbecq, A.L., Van de Van, A.H., Gustafson, D.H.: Group Techniques for Program Planning: a Guide to Nominal Group and Delphi Process. Scott Foresman, Glenview (1975)

[9] Ehrenberg, D., Eklund, P., Fedrizzi, M., Ventre, A.G.S.: Consensus in distributed soft environments. Reports in Computer Science and Mathematics, Ser. A, vol. 88. Åbo Akademi (1989)
[10] Eklund, P., Rusinowska, A., De Swart, H.: Consensus reaching in committees. European Journal of Operational Research 178, 185–193 (2007)
[11] Herrera-Viedma, E., Alonso, S., Chiclana, F., Herrera, F.: A Consensus Model for Group Decision Making with Incomplete Fuzzy Preference Relations. IEEE Transactions on Fuzzy Systems 15(5), 863–877 (2007)
[12] Kim, K.H., Roush, F.W.: Introduction to Mathematical Consensus Theory. Marcel Dekker, New York (1980)
[13] Linstone, H.A., Turoff, M.: The Delphi Method: Techniques and Applications. Addison-Wesley, Boston (1975)
[14] Luce, R.D., Raiffa, H.: Games and Decisions. John Wiley, New York (1957)
[15] Mares, M.: Fuzzy Cooperative Games. Springer, New York (2001)
[16] Merton, R.K.: The Focussed Interview and Focus Group: Continuities and Discontinuities. Public Opinion Quarterly, VI 4, 550–566 (1987)
[17] Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
[18] Shapley, L.S.: Simple games. An outline of the theory. Behavioral Sciences 7, 59–66 (1962)

Financial Applications of Flow Network Theory

Mario Eboli*

Dipartimento di Studi Aziendali, Università "G. d'Annunzio", Pescara, Italy

Abstract. This paper reviews some recent applications of flow network theory to the modelling of financial systems and of interbank liquidity networks. Three features of network flows have proven to be particularly useful in this field: i) the modularity of the transmission of flows across a network; ii) the constancy of a flow across all cuts of a nertwork; iii) the known 'max flow – minimum cut' theorem by Ford and Fulkerson. These properties of flow networks have been applied to evaluate the exposition to contagion of financial networks and the carrying capacity of interbank liquidity networks.

Keywords: flow networks, financial networks, contagion, systemic risk, liquidity risk.

1 Introduction

This paper reviews some applications of flow network theory to the modelling of financial networks. A flow network is a weighted and directed graph endowed with source nodes (i.e., nodes with no incoming links) and sink nodes (i.e., nodes with no outgoing links). A flow in a flow network is a value assignment to the links of the network such that: a) the value assigned to a link does not exceed its weight, i.e., the flow assigned to a link does not exceed its capacity (*capacity constraint*); b) for each node in the network that is neither a source or a sink node, the total inflow of a node must equal its total

* This paper was written for the conference 'Dynamics of Social and Economic Systems 2010' and is geared towards an audience of mathematicians and computer scientists interested in the modelling of social and economic networks. I wish to thank the participants to the DYSES conference for their useful comments. As usual, I remain the sole responsible for any possible shortcoming present in this paper. I can be reached at: `m.eboli@unich.it`.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 21–29.
springerlink.com

outflow (*flow conservation property*).[1] Flow networks are a natural modelling tool for the representation of financial networks. The latter arise whenever, among a set of agents, somebody's assets are somebody else's liabilities, i.e., when such agents are connected among themselves by financial obligations of some nature. We have applied flow networks modelling to two kinds of financial networks: financial systems and interbank liquidity networks. In what follows, we describe these models and provide an informal discussion of the properties of flow networks that proved to be useful in the analysis of direct financial contagion and of interbank liquidity transfers.

2 Financial Systems and Direct Contagion

Financial systems intermediate the supply of funds, provided by the non-leveraged sector (savers), and the demand of funds expressed by final users (such as governments, manufacturing companies, households' mortgages, etc.). For our purposes, we represente the liabilities of final users as external assets of the system and model them as source nodes: they represent the uses of funds that generate the value that flows across the network, from final users to final claimants. At the other end of the network, as sink nodes, we have the providers of funds, the final claimants. In between there are the financial intermediaries, the nodes in the network that are neither sink nor source nodes, which intermediate between suppliers and users of funds by providing liquidity and maturity transformations. In Eboli (2011) we represent a financial system as *financial flow network* and analyse the mechanics of direct financial contagion (also known as *domino effect* or *systemic risk*), i.e., the transmission of losses from borrowers to lenders due to an external shock that causes the initial default of some agents.

2.1 Financial Flow Networks

A financial system is composed of a set of operators $\Omega = \{\omega_i | i = 1, ..., n\}$, such as banks and other intermediaries, which are directly or indirectly connected to one another by financial obligations. Let $d_{ij} \in \mathbb{R}^+$ be the value of the liability, if any, issued by agent i and held by agent j. The balance sheet of a member of Ω is $a_i + c_i = e_i + d_i + h_i$, where: i) $a_i \in \mathbb{R}^+$ is the value of the *external assets* owned by the i-th operator. Such assets are liabilities – shares, bonds and bank loans – of final users of funds, which are not financial operators; ii) $c_i = \sum_j d_{ji}$ is the sum of the *internal assets* – which are liabilities of other agents in Ω – held by agent i; iii) $d_i = \sum_j d_{ij}$ is the *internal debt* of agent i, i.e., the sum of the liabilities issued by agent i and held by other agents in Ω; iv) h_i is the *external debt* of agent i, i.e., the amount of debt claims against i held in the form of bonds and deposits by final claimants,

[1] See Ahuya et al. (1993) for a comprehensive treatment of flow networks.

such as households, who do not belong to Ω; and finally v) e_i is the value of the equity of the i-th agent, which is set residually by the budget identity $e_i = a_i + c_i - d_i - h_i$.

Eboli (2011) models such a financial system as a *Financial Flow Network* (henceforth FFN) and the process of direct balance-sheet contagion as a flow of losses that crosses such a network. Formally, a FFN is a *multisource flow network* $N = \{\Omega, L, A, T, H, \Gamma\}$, i.e., a directed, weighted and connected graph, with some sources and two sinks, where:

1. $\Omega = \{\omega_i\}$ is a set of n nodes that represent the above defined financial intermediaries.
2. $A = \{a^k\}$, is a set of m *source nodes*, i.e., nodes with no incoming links, that represent the external assets held by the members of Ω.
3. H is a *sink* node representing the households who hold debt claims, in the form of deposits and bonds, against the agents in Ω.
4. T is a *sink*, i.e., a terminal node with no outgoing links. This node represents the shareholders who own the equity of the agents in Ω.
5. $L = (L^{\Omega} \cup L^{A} \cup L^{T} \cup L^{H})$ is a set of directed links, where i) $L^{\Omega} \subseteq \Omega^2$ is a set of directed links $\{l_{ij}\}$ representing the liabilities d_{ij}, where l_{ij} starts from node ω_i and ends in node ω_j, and $l_{ij} \in L^{\Omega}$ only if $d_{ij} > 0$; ii) $L^{A} = \left\{l_i^k\right\}$ is a set of directed links, with start nodes in A and end nodes in Ω, that connect the external assets a^k to their owners, where $l_i^k \in L^{A}$ only if $a_i^k > 0$; iii) $L^{T} = \left\{l_T^i\right\}$ is a set of n directed links, with start nodes in Ω and end node T; and iv) $L^{H} = \left\{l_H^i\right\}$ is a set of n directed links, with start nodes in Ω and end node H.
6. $\Gamma : L \to \mathbb{R}^{+}$ is a map, called *capacity function*, that associates i) to each l_{ij} the value of the corresponding liability d_{ij}, ii) to each l_i^k the value of the corresponding asset a_i^k, iii) to each l_T^i the equity, e_i, of its start node ω_i, and iv) to each l_H^i the external debt, h_i, of its start node ω_i.

A flow of losses crosses a financial network N when it is perturbated by a *negative exogenous shock*, i.e. by a loss of value of some assets in A. The direct financial contagion induced by a shock in a network N is a diffusion process governed by the laws of limited liability, debt priority and pro-rata reimbursements of creditors. These rules are embedded in the model by the following *absorption* and *loss-given-default* functions.

As a shock occurs, some nodes in Ω suffer a loss which is first offset by the equity of the nodes and borne by their shareholders. This is represented by an *absorption function* that measures the share of net worth lost by a node in Ω

$$\beta_i(\lambda_i) = \min \left(\frac{\lambda_i}{e_i}, 1\right) \tag{1}$$

where λ_i is the total loss borne by the i-th node, received from source nodes and/or from other nodes in Ω. If a node ω_i receives a positive flow of losses,

it sends to the sink T an amount of its own equity equal to $\beta_i e_i$, $\beta \in [0,1]$. If the losses suffered by i-th intermediary are larger than its net worth, $\lambda_i > e_i$, then this node is insolvent and sends the loss which is not absorbed by its equity, $\lambda_i - e_i$, to its creditors, i.e. its children nodes. For each node in Ω, let

$$b_i(\lambda_i) = \left\{ \begin{array}{c} 0, \text{if } \lambda_i < e_i; \\ \dfrac{\lambda_i - e_i}{d_i + h_i}, \text{ if } \lambda_i \geq e_i. \end{array} \right\} \tag{2}$$

be the *loss-given-default function*. If the i-th operator is solvent, b_i is null, while $b_i \in (0,1]$ if the operator defaults. In the latter case, the assets of the insolvent node are liquidated and its creditors get a *pro rata* refund. Households receive a loss equal to $b_i h_i$ (if $h_i > 0$), that ends into the sink H, while a node ω_j which is a creditor of node ω_i receives from the latter a loss equal to $b_i d_{ij}$. The loss borne by a financial intermediary in Ω is the sum of the losses, if any, received from its external and internal exposures, i.e., form the source nodes and from its *parent nodes* $P(\omega_i) = \{\omega_j | \exists l_{ji} \in L\}$:

$$\lambda_i = \sum_{k \in A} b_k a_i^k + \sum_{j \in P(\omega_i)} b_j d_{ji},$$

where b_k is the portion of the value of the k-th asset which is lost because of the initial shock. As a shock perturbates the network, the absorption and loss-given-default functions assign a positive real value to each link in N, given the value taken on by the initial shock vector $[b^k]$. Formally, such a *propagation function* in N is a map $f : L \to \mathcal{R}^+$ such that: $f(l_i^k) = b^k a_i^k$, $f(l_{ij}) = b_i d_{ij}$, $f(l_H^i) = b_i h_i$, $f(l_T^i) = \beta_i e_i$.

In Eboli (2011) we study the properties of such propagation function and use the above model to investigate the vulnerability of differently shaped financial networks. Two properties of flow networks have been particularly useful in this analysis: the modularity of flow propagations and the constancy of a flow across all cuts of a network:

2.1.1 Modularity

The propagation of a flow of losses across a financial network is computed node by node, in a iterated (and recursive, if the network entails strongly connected components) fashion. This modularity of the propagation of losses enabled us to address a known problem of indeterminacy that arises in case of cycles - or, more generally, of strongly connected components (henceforth SCC)[2] – of defaulting agents, establishing necessary and sufficient conditions

[2] A strongly connected component of a directed graph is a subset of nodes such that, for each pair of nodes i and j in the subset, there exist a directed path going from i to j and vice versa. A simple example of a strongly connected component is a directed cyle of nodes.

for the uniqueness of the loss propagation function f. In a SCC of defaulting agents, the loss-given-default functions of the agents are simultaneously determined and this may render undetermined the losses (or, equivalently, the payments) which are passed around within the SCC.[3] Exploiting the modular computation of the above defined propagation f, we analysed the behaviour of SCC's in isolation, separating them form the rest of the network. In so doing, we looked at the matrix of the flows of losses that circulate among the members of a SCC and pinned down the necessary and sufficient conditions for the non-singularity of such a matrix. We found that indeterminacy arises only in case of 'closed' SCC's of defaulting against, i.e. SCC's such that no member of the component has any debt towards agents who do not belong to the SCC.

The modularity of the propagations of network flows is also applied in Eboli (2010), a paper that applies the FFN model to compare the impact on systemic risk of two different accounting rules: the *marking-to-market* rule and the one based on historical cost. The marking-to-market accounting rule (known also as *fair value* rule) has come under scrutiny, after the start of the sub-prime financial cisis, for its alleged role in exacerbating the magnitude and diffusion of financial distress. This rule requires that the marketable assets held by a company, that are not classified as 'held-to-maturity', have to be accounted for at market value (as opposed to historical cost). Using the above described FFN model, we show that the flow of losses that crosses a node, in a network perturbated by a shock, is weakly larger under the mark-to-market rule than with the historical cost one. The modularity of the propagation function f enabled us to show that, for any given exogenous shock, the value taken on by the loss-given-default function of *each* node in N with the mark-to-market rule is weakly larger than the value taken on under the historical cost rule. Therefore, we show that the flow of losses is *linkwise* larger with the former rule and, as a consequence, the set of defaults induced by a shock with marking-to-market is weakly larger than the set of defaults induced by the same shock under historical cost accounting.

2.1.2 Flow Constancy Across Cuts

The constancy of a flow across all cuts of a network is a known property of network flows: for a flow defined in a flow network, the value of the *net forward flow* that crosses a *cut* is the same for all the cuts of the network.

[3] Consider, for example, a simple SCC composed by two defaulting nodes (which is the smallest possible directed cycle), each indebted with the other for one euro and none of them with any residual liquidation value. In this case, any mutual reinbursement comprosed bewteen zero and one would be a legitimate clearing payment flow. This problem was first pointed out by Eisenberg and Noe (2001), who established sufficient conditions for the uniqueness of a clearing payment vector in a financial network.

A *cut* in a flow network is a partition of a network in two subnetworks such that all source nodes are in one part and all sink nodes are in the other part. In the present case, a cut in a FFN N is a partition $\left\{U, \overline{U}\right\}$ of $\{A \cup \Omega \cup T \cup H\}$, where U and $\overline{U}$ are two non-empty sets such that $A \subseteq U$ and $(T, H) \in \overline{U}$. For a cut $\left\{U, \overline{U}\right\}$ of a FFN N, let $L^{+}(U)$ be the set of forward links going from U into $\overline{U}$, and et $L^{-}(U)$ be the set of backward links, going in the opposite direction. Let $f\left[L^{+}(U)\right] = \sum_{L^{+}(U)} f(l)$ be the flow that crosses $\left\{U, \overline{U}\right\}$ in the forward direction, i.e., from nodes in U into nodes in $\overline{U}$, and let $f\left[L^{-}(U)\right] = \sum_{L^{-}(U)} f(l)$ be the flow that crosses the cut in the backward direction. The *net flow* that crosses $\left\{U, \overline{U}\right\}$ is then equal to $\overrightarrow{f}\left\{U, \overline{U}\right\} = f\left[L^{+}(U)\right] - f\left[L^{-}(U)\right]$. Applying the property of flow constancy to a propagation f in a network N, we obtain that the net forward flow across every cut $\left\{U, \overline{U}\right\}$ of N is equal to the flow out of the source nodes, i.e. equal to the value of the exogenous shock: $\overrightarrow{f}\left\{U, \overline{U}\right\} = \overrightarrow{f}\left\{A, (\Omega, H, T)\right\} = \sum_{k \in A} b^{k} a^{k}$, for all $\left\{U, \overline{U}\right\}$ of N.

Using this property, we have characterized the *first* and the *final thresholds* of contagion for different network structures. The first threshold of contagion is the value of the smallest shock capable of inducing default contagion, and the final threshold is the value of the smallest shock capable of inducing the default of all agents in the network. These thresholds provide a measure of the exposition of a financial network to systemic risk. To characterise these thresholds, we applied the above property in a sort of backward induction. We start from the flow of losses going into the sinks T and H, which is the flow $\overrightarrow{f}\left\{\Omega, (H, T)\right\}$, and set it equal to the minimum flow capable of inducing contagion. Then, moving backward from the sinks to the source nodes in A, we compute the value of the shock – i.e. the contagion threshold – that induces such a flow across the cut $\{\Omega, (H, T)\}$ and characterise it in terms of the values of some balance sheet headings of the nodes in Ω.

3 Liquidity Flows in Interbank Liquidity Networks

Banks are exposed to liquidity risk, i.e., the risk of facing a liquidity shortage due to the fluctuations of customers' deposits. As a form of coinsurance, banks share such a risk by holding gross liquid positions: each bank deposits a sum in other banks and receives deposits from other banks. Such a cross-holding of deposits (and/or short term loans) forms what are known as interbank liquidity networks. In Eboli and Castiglionesi (2011) we model an interbank liquidity network as a flow network, with the aim of comparing the efficacy, of different network structures, in transferring liquidity among banks.

3.1 Interbank Liquidity Flow Networks

An *interbank network* is a connected, directed and weighted graph $N = (\Omega, \Lambda)$, where a node ω_i in Ω, $i = 1, 2, ..., n$, represents a banks and the links in $\Lambda \subseteq \Omega^2$ represent the interbank liquid exposures (namely deposits) that connect the members of Ω among themselves. The short term liabilities of a bank ω_i in Ω comprise customers (households) deposits, h_i, and interbank deposits, d_i. For simplicity, we assume that a bank in Ω has no long-term liability but its own equity v_i. On the asset side, a bank holds long-term assets, a_i, which are liabilities of agents that do not belong to Ω, and short-term assets, c_i, which are deposits put by ω_i in other banks of the network. The budget identity of a bank is: $a_i + c_i = h_i + d_i + v_i$. The links in Λ represent the interbank short-term obligations, their direction goes from the debtor node to the creditor node. The weight of a link $l_{ij} \in \Lambda$, that goes from node ω_i to node ω_j, is equal to the amount of money c_{ji} that bank ω_j has deposited in bank ω_i.

To transform an interbank network N into a flow network, we need to add to the network a *liquidity shock*. Such a shock consists of a re-allocations of customer deposits across banks, represented by an ordered vector of scalars $\boldsymbol{\delta} = [\delta_1, \delta_2, ..., \delta_n]$, where δ_i measures the change in customer deposits held by ω_i and $\sum_{\Omega} \delta_i = 0$. We attach a *source node* s_i and a link l_{si} to each node ω_i in Ω that experiences an increase of customer deposits. Correspondingly, to each node ω_i in Ω that faces a decrease of such deposits, we attach a *sink node* t_i and a link l_{it}. A liquidity shock that hits an interbank network N, is then defined as a four-tuple $\Delta = \{S, T, \Lambda^+, \Lambda^-\}$ composed by: the set of source nodes $S = \{s_i | \forall i \in \Omega \text{ s.t. } \delta_i > 0\}$; the set of sink nodes $T = \{t_i | \forall i \in \Omega \text{ s.t. } \delta_i < 0\}$; and the sets of links $\Lambda^+ = \{l_{si}\}$ and $\Lambda^- = \{l_{it}\}$ that connect sources and sinks to the corresponding banks. Adding a liquidity shock Δ to an interbank liquidity network N we obtain an *interbank liquidity flow network* L, which is an n-tuple $L = \{N, \Delta, \Gamma\} = \{\Omega, \Lambda, S, T, \Lambda^+, \Lambda^-, \Gamma\}$, where Γ is a *capacity function* that associates to the links in Λ a capacity equal to the value of the corresponding interbank deposits, and to the links in Λ^+ and Λ^- a capacity equal to the value of the corresponding variations of customers' deposits.

As a liquidity shock occurs, the banks facing a liquidity deficit withdraw their deposits from other banks. Such a behaviour generates an *interbank liquidity flow* in L – i.e. a value assignment to the links in Λ, Λ^+ and Λ^-– in as much as: i) no link carries a flow larger than its own capacity (*capacity constraint*); ii) the *divergence* of a node, i.e., the difference between its inflow and its outflow, is null for all nodes in Ω (*flow conservation property*). Full coverage of liquidity risk is achieved if there is an interbank liquidity flow capable of entirely re-allocating liquidity across banks. The existence and the size of such a flow crucially depend on two features of an interbank network L: its shape and the size of the interbank deposits c_{ij}.

3.1.1 The Minimum Cut-Maximum Flow Theorem

On one hand, the above network of interbank deposits serves the purpose of re-allocating liquidity from banks that have a liquidity surplus to banks that face liquidity deficits. On the other hand, the same network becomes a channel of contagion in case of default of one or more banks in the network. In choosing the amount of interbank deposits, banks face a trade-off: the larger the deposits, the larger the possible liquidity transfers (hence the larger the insurance against liquidity risk) and the larger the exposure to counterparty risk, i.e., the risk of contagion. It is then relevant to identify the network shape that allows the largest liquidity transfer with the smallest interbank exposures. We address this issue applying a fundamental result of network flow theory, the minimum cut – maximum flow theorem, commonly used to assess the maximum carrying capacity of a flow network.

This theorem, put forward by Ford and Fulkerson (1956), states that the maximum flow that can cross a network is equal to the capacity of the cut which, among all cuts of the network, has the smallest capacity. In other words, the capacity of the cut of smallest capacity in a network sets the upper bound of the value of the flows that can cross such a network. Applying this theorem, we evaluate and compare the performance of differently shaped interbank liquidity networks in providing full coverage of liquidity risk. We demonstrate that the star-shaped networks, also known as 'money centre'—i. e., the ones where there is a central node connected to all peripheral nodes and the latter are connected only with the central node—are the most effective in transferring liquidity. Star-shaped networks perform better than complete networks (where every bank is connected to all other banks in the network) which, in turn, perform better than incomplete interbank networks (where each bank is connected only to part of the other banks in the network).

4 Conclusions

Flow network theory has been applied to a vast number of fields, such as telecommunication, electrical and hydraulic engineering, transportation, computer networking, industrial and military logistics, etc. The models reviewed in this paper are the first applications of this theory to economics and finance. As argued above, the modelling of financial systems and of interbank liquidity networks as flow networks has created the grounds for the achievement of analytical results in the fields of direct financial contagion and interbank liquidity transfers. We believe that such a graph-thoretic approach to the modelling of financial networks can and will provide further useful analytical tools for the analysis of systemic risk, banking regulation and financial fragility. This line of research is on our agenda.

References

[1] Ahuja, K.R., Magnanti, T.L., Orlin, J.B.: Network Flows: theory, algorithms and applications. Prentice Hall, New Jersey (1993)
[2] Eboli, M.: The mechanics of direct contagion in financial systems: a flow network approach. Mimeo Previous Versions Presented at the Bank of England and at the National Bank of Denmark (2011)
[3] Eboli, M.: Direct contagion in financial networks with mark-to-market and historical cost accounting rules. International Journal of Economics and Finance 2(5) (2010)
[4] Eboli, M., Castiglionesi, F.: Liquidity transfers in interbank liquidity networks. Mimeo (2011)
[5] Eisenberg, L., Noe, T.H.: Systemic risk in nancial systems. Management Science 47(2), 236–249 (2001)
[6] Ford jr., L.R., Fulkerson, D.R.: Maximal flow through a network. Canadian Journal of Mathematics 8, 339–404 (1956)

Networks of Financial Contagion

Luisa Cutillo[1,2], Giuseppe De Marco[1,3], and Chiara Donnini[1]

[1] Dipartimento di Statistica e Matematica per la Ricerca Economica, Università di Napoli Parthenope. Via Medina 40, Napoli 80133, Italy
[2] Bioinformatics Core, TIGEM, Napoli, Italy
[3] CSEF, Università di Napoli Federico II, Italy
{luisa.cutillo,giuseppe.demarco,chiara.donnini}@uniparthenope.it

Abstract. Banks develop relationships in order to protect themselves against liquidity risk. Despite this benefit, fragility of financial markets stems from these interconnections. A cornerstone in the microeconomic analysis of contagion in financial systems is the contribution of Allen and Gale (2000). Our work takes up the challenge of generalizing their contagion analysis to complex networks. We provide an effective procedure to construct a network of financial contagion which enables the analysis of complex networks via simulations. Our study shows that it is possible to find a minimal number of links to guarantee contagion resiliency, and that the Allen and Gale conjecture holds: the system becomes more fragile as the number of links decreases.

1 Introduction

In modern financial systems, the mutual exposures that banks and other financial institutions adopt towards each other connect the banking system in a network. Banks develop relationships in order to protect themselves against liquidity risk. Despite this benefit, fragility of financial markets stems from these interconnections. Financial crisis demonstrated that, these interdependencies generate amplified responses to shocks in financial systems. They have also made it difficult to assess the potential for contagion arising from the behavior of financial institutions under exogenous shocks. Indeed, the initial default of a financial institutions, can propagate through the entire network. This propagation mechanism is referred to as *financial contagion*. In this paper we study the phenomenon of financial contagion in the banking system with a particular attention on how the structure of links between banks relates to the contagion spread. Our primary focus is on how losses can potentially spread via both direct or indirect exposures to banks following an initial default. Indeed, the exogenous shock on some financial institutions causes losses to financial entities directly linked to it and may cause losses to the indirectly linked ones. Thus we consider the potential for exogenous shocks to trigger further rounds of default to financial entities.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 31–48.
springerlink.com

A cornerstone in the microeconomic analysis of contagion in financial systems is the contribution of Allen and Gale (2000). Using a network structure involving four banks, they model financial contagion as an equilibrium phenomenon and demonstrate that the spread of contagion crucially depends on the nature of the interbank network. They show that the possibility of contagion depends on the number of direct connections between banks and moreover robustness to contagion reveals to be increasing in the number of direct connections. More precisely, Allen and Gale consider a market with different banks each of which is characterized by the liquidity demand of its depositors. In the aggregate, there is no uncertainty so that the (centralized) first best of optimal risk sharing can be obtained as in Allen and Gale (1998). Such a first best is achieved on the basis of the standard assumptions about technology and preferences introduced in the seminal paper by Diamond and Dybvig (1983). However, the structure of the liquidity demand of each bank implies that the first best contracts does not allow to insure banks in every state of the world. This suggests the constitution of an interbank network. In fact, when there is no aggregate uncertainty, the interbank deposits market allow banks to decentralize (achieve) the first-best allocation of risk sharing. However the network is financially fragile since the default of one bank can spread by contagion throughout the economy. In particular, when the network is complete, the impact of a shock is readily attenuated. By contrast, when the network is incomplete, the system is more fragile. This result is obtained by imposing strong assumptions on the distributions of the liquidity demands across banks. Our work takes up the challenge of generalizing the highly stylized Allen and Gale (2000) contagion analysis to complex networks, which could mimic more likely real world ones. Other authors, such as Furfine (2003), Upper (2007) and Eboli (2010) and references therein, produced interesting contribution in the literature of financial contagion in more realistic bank networks, proposing or suggesting numerical simulations. In the present work, we give a contribution to this literature by providing an effective procedure to construct a network of financial contagion. In such a network bank i is linked to bank j if the initial (exogenous) default of j propagates to i causing its failure. We assume that the dynamics of contagion is substantially the one considered in Allen and Gale (2000). This implies that contagion depends only on the interbank market and thus we only deal with systemic risk. We look at the contagion resiliency property of an interbank market network: the default of any bank is not expected to spread to any other bank in the network. In some cases we are able to find a rule of the thumb that suggests the minimal number of links that guarantee resiliency. Otherwise we are able to perform a simulation analysis to give an insight of the contagion spread in more general interbank networks. In conclusion our study shows that it is possible to find a minimal number of links to guarantee resiliency, and that the Allen and Gale conjecture holds: the system becomes more fragile as the number of links decreases.

2 Networks of Banks

Planner's Optimal Risk Sharing

Banks and Consumers
We consider an economy with three time periods, $t = 0, 1, 2$, m equiprobable states of nature, n different banks, a continuum of risk adverse consumers and a single consumption good that serves as the numeraire. We denote by $N = \{1, \ldots, n\}$ the set of banks, that we will index with i, and we denote by $S = \{s_1, \ldots, s_m\}$ the set of states of nature, that we will index with s.

Each consumer is endowed with one unit of the homogeneous consumption good at the date $t = 0$ and nothing at dates $t = 1$ and $t = 2$. Moreover he may be an early consumer, he needs to consume at date $t = 1$, or a late consumer, he needs to consume at date $t = 2$. At time $t = 0$ each consumer is uncertain about his liquidity preferences. At time $t = 1$ the real state of nature is reveled, it becomes known to everyone (that is to every consumer and every bank) and each consumer learns whether he is early consumer or late consumer. A consumer's type is not observable, so late consumers can always imitate early consumers. At time $t = 0$ each consumer deposits his endowment in a bank i. Since ex ante the consumers' type is not known, in each bank the number of early and late consumers depositing their endowments fluctuates randomly. However we assume that each bank knows, for every state of nature, the fraction of its own early consumers and the fraction of early consumers of each other bank. Then, trivially, it knows the average fraction of early consumers of the economy. The uncertainty about the consumers' type creates a demand for liquidity. Following Allen and Gale, we assume that there is no aggregate uncertainty and so the aggregate demand for liquidity is constant.

In order to analyze this framework we associate to each bank i a random variable, X_i, describing the fraction of early consumers depositing at time $t = 0$ their endowments in i. Denoting by $p_i(s)$ the fraction of early consumers and with $1 - p_i(s)$ the fraction of early consumers of bank i at the state s, then

$$X_i(s) = p_i(s) \quad \forall s \in S.$$

The requirement of absence of aggregate uncertainty is summarized in the following condition

$$\frac{1}{n} \sum_{i \in N} X_i(s) = q \quad \forall s \in S. \tag{1}$$

The assumption, clearly implies that each X_i is negatively correlated with $\sum_{j \neq i} X_j$ and q is the average fraction of early consumers in each state s.

At date $t = 2$ the fraction of late consumers in a generic bank i will be $1 - p_i(s)$, where the value of $p_i(s)$ is known at time $t = 1$.

Assets and Deposits
At date $t = 0$ consumers deposit their endowment in the banks, which invest them on behalf of the depositors. In exchange, they are promised a fixed amount of

consumption at each subsequent date, depending on when they choose to withdraw. In particular, the deposit contract specifies that if they withdraw at date $t = 1$, they receive $C' > 1$, and if they withdraw at date $t = 2$, they receive $C'' > C'$.

The banks can invest in two assets, a liquid asset and an illiquid asset. Since the first one pays a return of one unit after one period, we call it short asset. While, since the second one requires more time to mature, we refer to it as the long asset. The long asset is not completely illiquid, indeed, each unit of the long asset can be prematurely liquidated at time $t = 1$ and it pays a return $0 < r < 1$. If it is withdrawn at the final date it pays a return $R > 1$.

The long asset has a higher return if held to maturity, but liquidating it in the middle period is costly, so it is not very useful for providing consumption to early consumers.

The role of banks is to make investments on behalf of consumers and to insure them against liquidity shocks. The requirement that only banks invest in the long asset gives advantages to banks and to consumers. Indeed, the banks can hold a portfolio consisting of both types of assets, which will typically be preferred to a portfolio consisting of the short asset alone, and by pooling the assets of a large number of consumers, the bank can offer insurance to consumers against their uncertain liquidity demands, giving the early consumers some of the benefits of the high-yielding long asset without subjecting them to the high costs of liquidating the long asset prematurely at the second date.

Planner Problem and First-Best

The banking sector is perfectly competitive, so banks offer risk-sharing contracts that maximize depositors ex ante expected utility, subject to a zero-profit constraint.

We assume that consumers have the usual Diamond-Dybvig preferences. That is with probability p they are early consumers and value consumption only at date $t = 1$ and with probability $1 - p$ they are late consumers and value consumption only at date $t = 2$. Then the preferences of the individual consumer are given by

$$U(C', C'') = \begin{cases} u(C') & \text{with probability } p \\ u(C'') & \text{with probability } 1 - p. \end{cases} \tag{2}$$

where C' denotes consumption at $t = 1$ and C' at $t = 2$ and the period utility functions $u(\cdot)$ are assumed to be twice continuously differentiable, increasing, and strictly concave.

We characterize optimal risk sharing as the solution to a planning problem. That is the planner is assumed to make all the investment and consumption decisions to maximize the unweighted sum of consumers expected utility. By (1), assuming $\frac{1}{n}\sum_{s\in S} p_i(s) = q$ for every i we get that all consumers have equal treatment and optimal consumption allocation will be independent of the state. In this way we have a problem similar to the optimization problem (20) in Allen and Gale (1998). So, under the condition $x \leq Ry$, the optimal portfolio is

$$(x, y) = \left(qC', \frac{1-q}{R} C'' \right). \tag{3}$$

In our work we will retain (3) as first best.

Interbank Market

Even if a bank chooses the first best allocation at time $t = 0$ the realization of a state s at time $t = 1$ causes either a liquidity surplus, a liquidity shortage or the bank can perfectly meet withdrawals from early consumers. Moreover there is no aggregate uncertainty and then there are opportunities for risk-sharing. In Allen and Gale it is argued that in the particular case in which banks are partitioned in two regions of identical banks, an interbank market can perfectly decentralize the planner's solution as long as it allows banks that have a high fraction of early withdrawals to raise the liquidity they need from the banks that have a low fraction of early withdrawals. Therefore they can hedge completely individual risk by exchanging interbank deposits at date $t = 0$.

Following Allen and Gale, we assume each bank receives the same return as the consumers for the amounts transferred as deposits: C', if they withdraw after one period, and C'' if they withdraw after two periods. Thus, banks' portfolios consist of three assets: the short asset, the long asset and the interbank deposits. We remind that, while short asset has to be liquidated at time $t = 1$, long asset and interbank deposits can be liquidated in either of the last 2 periods. However, since liquidating the long asset in the middle period is costly, even to withdraw interbank deposits in the middle time is costly. It is assumed that the costliest in terms of early liquidation is the long asset, followed by interbank deposits. This implies the following ordering of returns:

$$1 < \frac{C''}{C'} < \frac{R}{r}.$$

We assume that each bank may exchange deposits with each other banks, so, denoting by a_{ij} the pro capita deposit of bank i in the bank j, we clearly assume that $a_{ij} \geq 0$, for each $i \neq j$, while $a_{ii} = 0$, for each i. We emphasize that deposit contracts between banks, unlike Allen and Gale, are not necessarily bilateral, that is, we do not require $a_{ij} = a_{ji}$.

Definition 1. We define *network of deposits*[1] the $N \times N$ matrix $(a_{ij})_{i,j}$.

3 Insurance against Liquidity Shocks

We impose, for each bank i, the classical budget constraint at time $t = 0$

$$\sum_{j \in N} a_{ji} + 1 \geq x + y + \sum_{j \in N} a_{ij} \iff \sum_{j \in N} a_{ij} - \sum_{j \in N} a_{ji} \leq 1 - x - y \tag{4}$$

[1] In a mathematical framework, a network is referred to as a graph. In this light each bank is a vertex or node and the deposits a_{ij} are links between the nodes of the network that could be called connections, links or edges. From the simplest perspective, a network is just a collection of links.

which, obviously, implies that $1-x-y \geq 0$. For the sake of simplicity, we consider only networks such that $\sum_{j\in N} a_{ij} = \sum_{j\in N} a_{ji}$, for every bank i. Moreover, as previously observed, at time $t=1$, when the real state of nature s is revealed, each bank i incurs either a liquidity surplus ($p_i(s)-q<0$), a liquidity shortage ($p_i(s)-q>0$) or the bank can perfectly meet withdrawals from early consumers ($p_i(s)-q=0$). In order to honor the contracts concluded with consumers and to redeem deposits to the other banks, a bank i that incurs a liquidity shortage needs to withdraw its deposits. On the other hand, a bank j that incurs a liquidity surplus is able to repay C' to a fraction $p_j(s)$ of its own depositors but it may need to withdraw its deposits in order to redeem the deposits eventually claimed by the other banks. Hence, we introduce a demand function depending on banks and on state of nature $d : S\times N\times N \to [0,1]$, which specifies, for every (s,j,i), the portion of a_{ji} that bank j withdraws from i at state s and at time $t=1$. Given a demand function d, we can define when a network of deposits $(a_{ij})_{i,j}$ is *insured against liquidity shocks* under d.

Definition 2. The network $(a_{ij})_{i,j}$ is *insured against liquidity shocks* under a demand function d if

$$p_i(s) + \sum_{j\neq i} a_{ji} d(s,j,i) = \sum_{j\neq i} a_{ij} d(s,i,j) + q \qquad \forall s \in S. \tag{5}$$

Equation (5) means that the budget constraint of bank i is satisfied at time $t=1$ in every state of the world. Moreover, (5) obviously implies that

$$1 - p_i(s) + \sum_{j\neq i} a_{ji}(1-d(s,j,i)) = \sum_{j\neq i} a_{ij}(1-d(s,i,j)) + 1 - q \qquad \forall s \in S. \tag{6}$$

meaning that the budget constraint is satisfied also at time $t=2$.[2]

4 Financial Contagion

Aim of this section is to study the diffusion of financial contagion within the interbank market. To this purpose, we just consider the effects of the default of a bank on the others without considering the reasons leading to the original default. Moreover, the dynamics of contagion is substantially the one considered in Allen and Gale (2000). This latter assumption implies that contagion depends only on the interbank market and not on distribution of the liquidity demands and therefore it does not depend on liquidity shocks at each state in S[3].

[2] We can observe that if, for a bank i, the inequality $p_i(s) + \sum_{j\neq i} a_{ji} d(s,j,i) < \sum_{j\neq i} a_{ij} d(s,i,j) + q$ holds at time $t=1$, then the reverse inequality holds at time $t=2$ which implies that bank i fails at time $t=2$. Moreover, if the surplus at time $t=1$ is such that the bank is not able to guarantee to late consumers at least C' at time $t=2$ then, late consumers would be better off withdrawing at time $t=1$, causing the default of bank i already at time $t=1$.

[3] In order to focus to the systemic risk of financial contagion, the analysis is usually restricted to interbank markets insured against liquidity shocks.

4.1 Contagion Threshold

Losses Given Default

Suppose that an exogenous shock leads to the default of bank j and there are no other liquidity shocks [4]. Then, the other banks will be trying to withdraw their claims on the bank j and, at the same time the bank j liquidates its deposits in the other banks. All depositors must be treated equally; that is, every depositor gets the same liquidation value c_j from the bank j for each unit invested at the first date, whether the depositor is a consumer or another bank. Since default can occur to many banks simultaneously, then the values $(c_i)_{i\in N}$ must be determined simultaneously from a system of equations, each of them representing the budget constraint of each bank j:

$$\left(1+\sum_{i\in N} a_{ij}\right) c_j = x + ry + \sum_{i\in N} a_{ji} c_i, \tag{7}$$

where $c_i = C'$ if bank i has not failed while $c_i < C'$ otherwise. The left hand side of (7) gives the value of the deposits in the bank j, while the right hand side gives the value of the assets and the deposits of bank j in the other banks. Now we give an upper bound for each liquidation value c_j, denoted with $\overline{c}_j$, in which the liquidation values of the other banks are supposed to be equal to the first-best values at date 1, that is:

$$\left(1+\sum_{i\in N} a_{ij}\right) \overline{c}_j = x + ry + C' \sum_{i\in N} a_{ji} \implies \overline{c}_j = \frac{x + ry + C' \sum_{i\in N} a_{ji}}{1+\sum_{i\in N} a_{ij}}. \tag{8}$$

Therefore we define *losses of bank i given the default of bank j* to be

$$\lambda_{ij} = a_{ij}\left(C' - \overline{c}_j\right). \tag{9}$$

Note that this formula gives a lower bound for the real losses incurred by bank i given bank j's default.

Buffer Level

Following Allen and Gale, we construct a threshold level for the propagation (in expectation) of the default to a bank i. This threshold is based on the condition that bank i is expected to fail as long as late consumers would be better off withdrawing at date 1. This condition is satisfied when bank i is not able to repay (to late consumers and at date $t = 2$) at least the return C' for each unit invested at $t = 0$. Denote with $y - \overline{y}$ the fraction of the long asset liquidated at date 1. In order to repay C' to late consumers at date $t = 2$, bank i with a fraction q of early consumers must keep (at $t = 1$) at least a fraction $\overline{y}$ of the long asset such that

[4] For instance, Allen and Gale (2000) consider an additional zero probability state $\bar{s}$ such that $p_j(\bar{s}) = q + \varepsilon$ and $p_i(\bar{s}) = q \ \forall i \in N \setminus \{j\}$.

$$(1-q)C' = \overline{y}R \implies \overline{y} = \frac{(1-q)C'}{R}.$$

The value, at time $t = 1$, of the maximum amount of long asset that can be liquidated without causing a run on the bank i can be therefore interpreted as a contagion threshold:

$$b(q) = r\left[y - \frac{(1-q)C'}{R}\right]. \tag{10}$$

More precisely, suppose that F_i is the subset of failed banks in $N \setminus \{i\}$ then the bank i is expected to fail if

$$\sum_{h \in F_i} \lambda_{ih} > b(q). \tag{11}$$

4.2 The Contagion Network

Given the network of deposits with adjacency matrix $(a_{ij})_{i,j}$ and building upon the sufficient contagion condition (11) which guarantees the contagion of a bank i given the default of the others, we can construct networks of financial contagion with adjacency matrix $(f_{ij})_{i,j}$ where $f_{ij} = 1$ means that bank i is expected to fail given that bank j has failed and $f_{ij} = 0$ otherwise. Obviously such networks should satisfy the following

$$\begin{array}{l} f_{ii} = 1 \quad \forall i \in N \\ f_{ij} = \chi\left(\sum_{h \neq i} \lambda_{ih} f_{hj} - b(q)\right) \ \forall i, j \text{ with } \quad i \neq j \end{array} \tag{12}$$

where

$$\chi(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}.$$

The previous conditions provide a system of nonlinear equalities which may have many solutions. However we give a constructive method to find a particular network satisfying (12). This network, that we call *contagion network*, will be denoted again with $(f_{ij})_{i,j}$. The idea underlying the definition of this network is to start from the failure of a bank (say j) and then to construct the j-th column of the adjacency matrix by looking at the final propagation of the default under the contagion condition (11). In particular, the element f_{ij} will be equal either to one, if the default has reached bank i, or to 0 otherwise.

In order to provide the precise definition of the contagion matrix, we give the following:

Definition 3. Given two different banks i and j, we define *contagion path* from j to i a set of banks $CP_{ji} = \{j_0 = j, j_1, ..., j_t = i\}$ s.t. $\chi(\lambda_{j_k j_{k-1}} - b(q)) = 1 \ \ \forall k = 1, ..., t$. We denote by SP_{ij} the shortest path.

We observe that the contagion path from j to i exists if and only if the financial contagion spreads all over the path finally causing the failure of bank i.

1) Define $A_1^j = \{j\} \cup \left[\bigcup_{i \in N \setminus \{j\}} SP_{ji} \right]$

2) Given the set A_{r-1}^j, let

$$B_r^j = \left\{ i \in N \setminus A_{r-1}^j \mid \chi \left(\sum_{h \in A_{r-1}^j} \lambda_{ih} - b(q) \right) = 1 \right\}$$

and

$$A_r^j = A_{r-1}^j \cup B_r^j$$

3) Since the set of banks is finite, there exists $w \leq n-2$ such that $A_w^j = A_r^j$ for all $r > w$. Hence the iterative construction stops after at most $n-1$ steps.

Now we can give the following

Definition 4. The contagion network is the directed network given by the adjacency matrix $(f_{ij})_{i,j}$ satisfying

$$f_{ij} = \begin{cases} 1 \text{ if } i \in A_n^j \\ 0 \text{ otherwise.} \end{cases} \tag{13}$$

By construction the contagion network satisfies the condition (12); however, it is not the unique solution of such system in general. We implemented in *Matlab* our iterative method described so far. The source code is available on demand. For sake of simplicity, we can summarize the overall algorithm as follows:

ALGORITHM

- Step 0: Define F_0 as the Identity matrix $(N \times N)$.
  We suppose that at the beginning one bank per time can fail.
- Step 1: Define F_1 as the direct contagion matrix of F_0:

$$F_1(i,j) = \begin{cases} 1 \text{ if } i \text{ is directly infected by the default of } j \\ 0 \text{ if otherwise} \end{cases}$$

- Repeat
  - Step r: Define F_r as the new contagion matrix[5] given F_{r-1}:

$$F_r(i,j) = \begin{cases} 1 \text{ if } \sum_h \lambda_{ih} F_{r-1}(h,j) > b(q) \\ 0 \text{ if otherwise} \end{cases}$$

- Until(F_r==F_{r-1}).

[5] In practice, given that bank j fails first, $F_r(i,j) = 1$ if and only if $F_{r-1}(i,j) = 1$ or the banks failed at step $r-1$ cause the default of bank i.

4.3 Contagion

When the contagion network $F = (f_{ij})_{i,j}$ corresponds to the identity matrix then the default of any bank is not expected to spread to any other bank in the network. In the following we will refer to this property as the *contagion resiliency* of the interbank market:

Definition 5. A network of interbank deposits $A = (a_{ij})_{i,j}$ is *resilient* from contagion if $f_{ij} = 1$ if and only if $i = j$.

It can be easily checked that if $\lambda_{ij} \leq b(q)$, for every pair of banks (i, j) with $i \neq j$, then the interbank market is resilient from contagion. Indeed, the previous condition implies that there are no contagion path between any pair of players so that $f_{ij} = 0$ whenever $i \neq j$.

If $\lambda_{ij} > b(q)$ for every $i \neq j$, then we obviously get $f_{ij} = 1$ for every pair of banks (i, j). Finally, if $\lambda_{ij} > b(q)$ for some pair of banks (i, j) with $i \neq j$, then we do not have, in general, one of the two extreme situations (resiliency or full contagion) since it may be $f_{ij} = 1$ even if $\lambda_{ij} \leq b(q)$.

Building upon the previous resiliency condition, that is, $\lambda_{ij} \leq b(q)$ for every $i \neq j$, we give two conditions to obtain resilient interbank deposits network $(a_{ij})_{i,j}$. To this purpose we state the following assumptions on $(a_{ij})_{i,j}$.

Assumption 1. *For each bank i,*

$$\sum_{h \in N} a_{ih} = \sum_{h \in N} a_{hi} = M$$

where M is a fixed positive number.

Assumption 2. *If*

$$N_i^{in} = \{h \in N \,\text{s.t.}\, a_{hi} \neq 0\}$$

and

$$N_i^{out} = \{h \in N \,\text{s.t.}\, a_{ih} \neq 0\}$$

then

$$|N_i^{in}| = |N_i^{out}| = K \qquad \forall i \in N.$$

Assumption 3. *For every bank i and for every* $j \in N_i^{out}$

$$a_{ij} = \frac{M}{|N_i^{out}|}.$$

Proposition 1. *If* $(a_{ij})_{i,j}$ *satisfies Assumptions 1,2,3, and*

$$|N_i^{out}| \geq K^* = \frac{RC' - rC''}{r(C'' - C')} - \frac{1}{L}$$

then it is resilient.

Proof. Let us assume that the exogenous shock strikes a bank j. Let $K = |N_j^{out}|$, $a_{ij} = a_{ji}$ and $L = a_{ij}$ for each $a_{ij} \neq 0$. By Assumption 2, $K = |N_i^{in}|$, so

$$\overline{c}_j = \frac{x + ry + C'M}{1 + M} = \frac{x + ry + C'KL}{1 + KL}.$$

Since $x = qC'$ and $Ry = (1-q)C''$, then, for every $i \in N_j^{in}$, it follows that

$$\lambda_{ij} = \frac{(1-q)L(RC' - rC'')}{R(1+KL)} = \lambda$$

and

$$b(q) = \frac{r}{R}(1-q)(C'' - C').$$

So $\lambda_{ij} \leq b(q)$ if and only if

$$K \geq K^* = \frac{RC' - rC''}{r(C'' - C')} - \frac{1}{L}.$$

Under Assumptions 1,2,3 each link in the interbank deposits network is $a_{ij} = 0$ or $a_{ij} = c$ where $c = \frac{M}{K}$. It can be checked that, in this case, losses given default are $\lambda_{ij} = 0$ or $\lambda_{ij} = \lambda$ for every pair of banks (i, j). Given the default of a bank j, if $\lambda > b(q)$ (i.e. $K < K^*$) then the default reaches every bank $i \in N_j^{in}$. Therefore, the default of a bank j spreads all over the set of banks which are indirectly connected with j in the network $(a_{ij})_{i,j}$; where a bank i is indirectly connected with the bank j if there exists a subset $\{i_0 = i, i_1, \ldots, i_k = j\} \subseteq N$ such that $a_{i_t i_{t+1}} \neq 0$ for all $t \in \{0, \ldots, k-1\}$. In particular, if the interbank deposits network is complete and $\lambda > b(q)$ then we get full contagion, that is $f_{ij} = 1$ for every pair (i, j).

Finally, we emphasize that the value K^* depends on the amount M and on the market parameters, but not on number of banks n.

Table 1 Minimum number K of interbank deposits to reach contagion resiliency of the system

Case	K
0.01≤M≤ 0.10	1
0.11≤M≤ 0.22	2
0.23≤M≤ 0.37	3
0.38≤M≤ 0.50	4

Table 1 shows the minimum number of interbank deposits required to obtain contagion resiliency of networks satisfying Assumptions 1,2,3, when the market parameters are $R = 1.1$, $r = 0.5$, $q = 0.5$ $x = y = 0.5$.

Proposition 2. *If $(a_{ij})_{i,j}$ satisfies Assumption 1 and for every i and j*

$$\frac{a_{ij}}{1+M} \leq r\frac{C'' - C'}{C'R - rC''} \tag{14}$$

then it is resilient.

Proof. Let $(a_{ij})_{i,j}$ be a network satisfying Assumption 1 and (14). Let us assume that the exogenous shock strikes a bank j. Let us consider $i \in N_j^{in}$. Since $x = qC'$, $Ry = (1-q)C''$, and, by Assumption 1, $\overline{c}_j = \frac{x+ry+C'M}{1+M}$, then

$$\lambda_{ij} = \frac{a_{ij}}{R(1+M)}(1-q)(C'R - rC'').$$

Hence $\lambda_{ij} \leq b(q)$ if and only if

$$\frac{a_{ij}}{1+M}(C'R - rC'') \leq r(C' - C').$$

The conditions $1 < \frac{C''}{C'} < \frac{R}{r}$ imply $C'R - rC'' > 0$ that, with condition (14), gives $\lambda_{ij} \leq b(q)$. In this way none in N_j^{in} has losses given default greater than the threshold and then the contagion does not spread.

5 Simulation

Using a four banks market, Allen and Gale (2000) show that financial contagion is influenced by the *degree of connectedness* of interbank network structure. They show, by means of examples, that the spread of contagion can range from a best to a worst scenario as far as the topology of the network ranges from complete to incomplete. In summary they get that:

- When the network is complete (i.e. $a_{ij} \neq 0 \ \forall i \neq j$) and satisfies Assumptions 1,2,3, then it is resilient.
- When the network is a directed cycle, then the system is fragile since the entire system is affected by full contagion.

Our study takes up the challenge of generalizing the Allen and Gale (2000) contagion analysis to more complex networks and we show that the conjecture that interbank markets are more fragile as the number of links decreases (given that the network is connected) is consistent with the results for many and quite general deposits network structures.

According to the Proposition 1, we can directly observe that:

Proposition 3. *Under the the Assumptions 1, 2 and 3, a complete interbank network of size $n \geq 4$ is resilient $\iff n - 1 \geq K^*$.*

This proposition provide us with a first detailed generalization of Allen and Gale (2000) for a complete interbank network of arbitrary size $n \geq 4$. If we neglect

Assumptions 2 or 3, we could not obtain any general results of contagion resiliency. This consideration suggested us to perform a simulation analysis to give an insight of the contagion spread in more general interbank networks.

In the following we depict three different network topologies. In each of them we look for the minimum number of interbank deposits that guarantee contagion resiliency. We start from undirected network structures and we end up to a more general situation where the interbank deposits are randomly distributed among a fixed number of banks. For our simulation study we consider the market parameters $R = 1.1$, $r = 0.5$, $q = 0.5$ and $x = y = 0.5$. Moreover, we assume that any of the bank network topology proposed in the following satisfies Assumption 1, we allow the amount invested by each bank to be $M \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and we consider networks of size $n \in [4, \ldots, 100] \subset \mathbb{N}$. To visualize our entire collection of examples, the proposed interbank networks and the contagion networks identified by means of our algorithm were displayed using the Cytoscape 2003 software environment.

5.1 *Undirected Complete Topology*

Our first attempt of generalization consists in complete and undirected network structures in which deposits are not necessarily constant across players. In this general setting, Assumption 3 does not hold any more and hence we can't retain Proposition 3. Our simulation analysis shows that, in most of the cases, this scenario lends to an identity contagion matrix that expresses contagion resiliency. These results are summarized in Table 2.

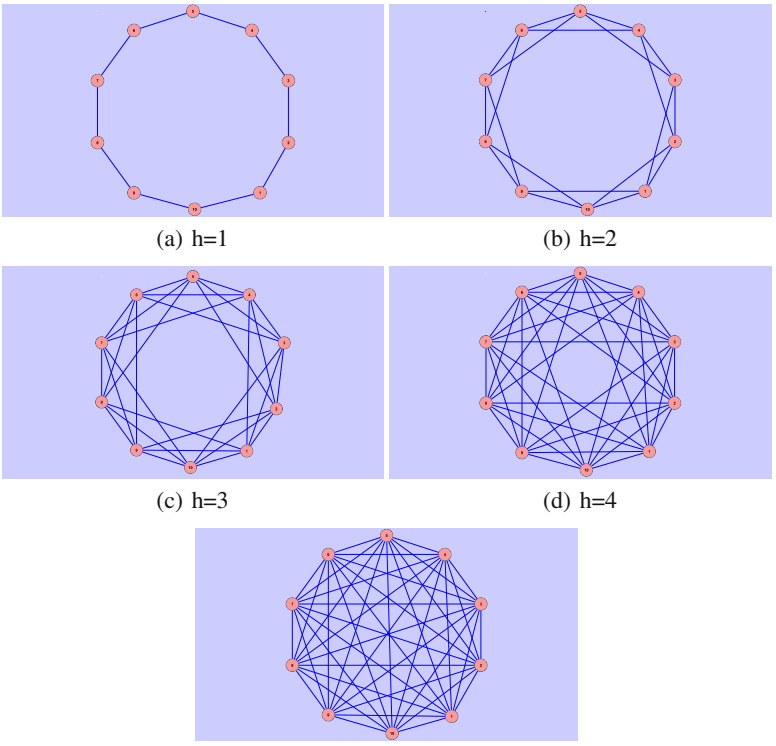
Table 2 Simulation results table for a general undirected complete topology

M	n range	Resiliency
0.1	[4, 100]	Yes
0.2	[4, 100]	Yes
0.3	[4,8]	No
	[9,100]	Yes
0.4	[4, 7]	No
	[8, 100]	Yes
0.5	[4, 8]	No
	[9, 100]	Yes

5.2 *Connected K-Regular Topology*

Aim of this subsection is to investigate by means of simulations, the contagion diffusion in connected but not necessarily complete networks. We recall that in a connected graph two nodes are always indirectly linked in the sense that there is a path from any node to any other node in the graph. The total number of

(not necessarily connected) labeled n nodes graphs for $n = 1,2,3,4,5,6,...$ is given by $1,2,8,64,1024,32768,...$ while the number of connected labeled graphs on n nodes is given by the logarithmic transform of the preceding sequence: $1,1,4,38,728,26704,...$ (see also Sloan and Plouffe (1995)). Given the great variety of connected networks, for the sake of simplicity we focus on a subset of connected k-regular networks. A graph where each vertex has the same number of neighbors (i.e. every vertex has the same degree) is defined to be regular. A regular graph with vertices of degree k is called a k-regular graph. Then, to keep the analysis as simpler as possible, we impose an additional condition on the set of neighbors of each node. Assumed to label the nodes circular wise, we allow each node to link only to the h preceding nodes and to the h following nodes in the circle. This topology reveals to be a $k = 2h$ regular graph and we will refer to it as $2h$ regular *ring* graph. The rationale of this choice relies on the fact that for $h = 1$ this is the topology of a ring network.



(a) h=1 (b) h=2

(c) h=3 (d) h=4

(e) h=5, complete case

Fig. 1 Ring $2h$ regular graph of 10 nodes with $h \in \{1,\ldots,5\}$

Figure 1 depicts a 2h regular ring for $n = 10$ nodes. As you can see in this case h can range in the set $\{1,2,\ldots 5\}$. We can observe two extreme situations: for $h = 1$ the graph turns to be a ring, while for $h = 5$ it becomes complete. It is easy to show that, when the number of nodes is an arbitrary integer $n \geq 4$, a 2h regular ring graph is a ring again for $h = 1$ and it is the complete network for $h = [\frac{n}{2}]^+$ where $[\frac{n}{2}]^+$ is the integer part of the number $\frac{n}{2}$[6].

Given the topology defined so far, we explored both the undirected and the directed cases. It is easy to see that a $2h$ regular ring network satisfies Assumption 2, in fact it is $|N_i^{in}| = |N_i^{out}| = h, \forall i \in \{1,\ldots,n\}$. In the undirected case, we impose the additional requirement that the amount each bank exchanges with its neighbors is the same all over the network, according to Assumption 3. Thus we refer improperly to such network structures as *constant*[7] k regular networks. As previously observed, this topology satisfies Assumptions 1, 2 and 3, and hence the Proposition 1 and Table 1 hold. As a consequence, we can state that when $0.01 \leq M \leq 0.22$, contagion resiliency is ensured by constant $2h$ regular ring networks for every $h \geq 1$ and for every $n \geq 4$. At the same time, when $0.23 \leq M \leq 0.37$, contagion resiliency is ensured by constant $2h$ regular ring networks for every $h \geq 2$ and for every $n \geq 4$. Finally, when $0.38 \leq M \leq 0.50$, contagion resiliency is ensured by constant $2h$ regular ring networks for every $h \geq 2$ and for every $n \geq 5$.

In the case of directed k regular ring network structures, we impose only Assumption 1. Hence we do not have general results and we can only proceed by simulation to inspect the contagion spread in this case. Results tell us that when $M \in \{0.1,0.2\}$, a 2 regular ring networks is resilient for every $n \geq 4$. Similarly when $M \in \{0.3,0.4,0.5\}$ there exist a k regular ring graph that ensures the contagion immunization with $k \in [4,14]$ and $n \geq 6$. More explicitly Table 3 shows the minimal h such that the simulated $2h$ regular ring networks are resilient for $M \in \{0.3,0.4,0.5\}$. Note that we used the symbol $\emptyset$ to refer to the eventuality that such a value was not found. Finally observe that in case $n \geq 10$, it is always possible to obtain contagion resiliency if each bank is directly linked to *at list* k banks where k can be calculated exactly for a constant $2h$ regular ring topology, while it could be inferred case by case by simulation analysis for a generic $2h$ regular ring topology. As example consider the case of a 10 banks network. Suppose we seek for a star k regular graph that enables contagion resiliency. Our simulations reveals that the minimal configuration is a 8 ring regular graph, corresponding to $h = 4$. Figure 2 shows the contagion networks derived for $h \in \{1,2,3,4\}$. It is evident that the contagion spread ranges from a full contagion ($h = 1$) to a diagonal contagion ($h = 4$) that is resiliency. As further example suppose we seek for a star k regular that enables contagion resiliency for a 53 banks network. In this case the minimal configuration is a 10 ring regular graph, corresponding to $h = 5$. Figure 3 shows the contagion networks derived for

[6] Note that in general each bank can have at most $n-1$ direct links. Thus, when n is odd, $h = \frac{n-1}{2}$ and each bank is exactly linked to the h previous and to the h followings banks. On the other hand, when n is even, $h = \frac{n}{2}$ and the sets of h following and h preceding banks have a bank in common.

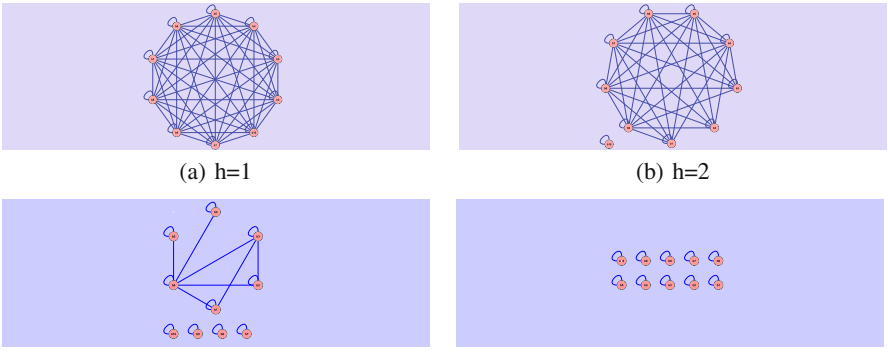
[7] The use of the term *constant* is improper because the adjacency matrix is not constant as $a_{ij} = 0$ if and only if bank i is not directly linked to bank j and $a_{ij} = \frac{M}{|N_i^{out}|}$ otherwise.

$h \in \{1,2,3,4,5\}$. Also in this case the contagion spread ranges from a full contagion ($h = 1$) to a diagonal contagion ($h = 5$). Note that the number of contagion links reduces dramatically as h grows.

Table 3 minimal h such that the simulated $2h$ regular ring graph is immunized from certain contagion

	M=0.3	**M=0.4**	**M=0.5**
n	h	h	h
4	$\emptyset$	$\emptyset$	$\emptyset$
5	$\emptyset$	$\emptyset$	$\emptyset$
6	3	$\emptyset$	$\emptyset$
7	3	3	$\emptyset$
8	4	$\emptyset$	$\emptyset$
9	3	4	$\emptyset$
10	3	3	5
11	4	4	5
15	4	4	5
20	2	4	6
40	4	5	5
70	2	4	5
90	5	5	6
100	4	5	6

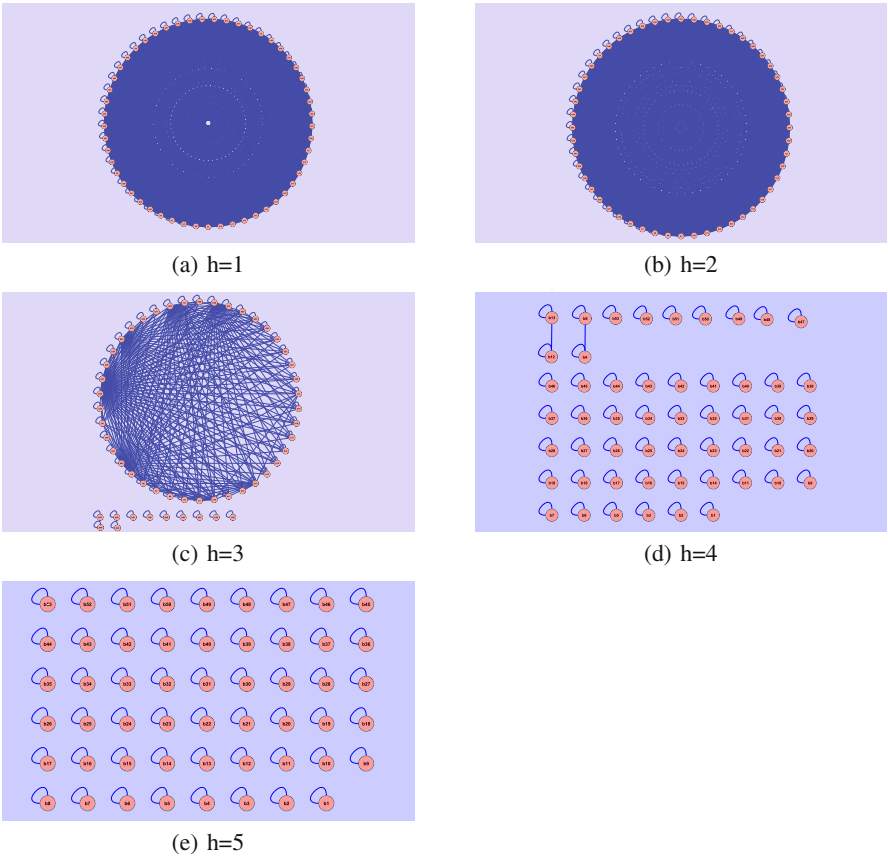
These final examples provide a description of the spread of contagion as the number of links varies, in the framework of $2h$ regular ring structures. They clearly show that the Allen and Gale conjecture holds: the system becomes more fragile as the number of links decreases.



(a) h=1 (b) h=2

(c) h=3 (d) h=4

Fig. 2 Contagion Networks derived for a $2h$ regular ring graph of 10 nodes with $h \in \{1,\ldots,4\}$ and $M = 0.4$



(a) h=1 (b) h=2

(c) h=3 (d) h=4

(e) h=5

Fig. 3 Contagion Networks derived for a $2h$ regular ring graph of 53 nodes with $h \in \{1, \ldots, 5\}$ and $M = 0.4$

Remark 1 (Directed k-Regular Topology). We recall that k-regular network structures can be obtained by permutations of $2h$ regular ring networks. Our simulation results in this case are similar to the $2h$ regular ring networks case and are available on demand.

References

1. Allen, F., Gale, D.: Optimal Financial Crisis. Journal of Finance 53, 1245–1284 (1998)
2. Allen, F., Gale, D.: Financial Contagion. Journal of Political Economy 108, 1–33 (2000)
3. Dyamond, D.W., Dybvig, P.: Bank Runs, Deposit Insurance and Liquidity. Journal of Political Economy 91, 401–419 (1983)
4. Eboli, M.: An Algorithm of Propagation in Weighted Directed Graphs with Applications to Economics and Finance. International Journal of Intelligent Systems 25, 237–252 (2010)

5. Furfine, C.H.: Interbank Exposures: Quantifying the Risk of Contagion. Journal of Money, Credit and Banking 35, 111–138 (2003)
6. Shannon, P., et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003)
7. Sloane, N.J.A., Plouffe, S.: The Encyclopedia of Integer Sequences. Academic Press, San Diego (1995)
8. Upper, C.: Using counterfactual simulations to asses the danger of contagion in interbank markets, BIS Working Paper No. 234 (2007)

Rumour Propagation on Social Networks as a Function of Diversity

Bernard Brooks

School of Mathematical Sciences
Rochester Institute of Technology

Abstract. In the case of a rumour propagating across a social network comprised of two interconnected groups, a majority group and a minority group, the effect on the rumour propagation of the minority's distribution in the social network is investigated. The rumour is derogatory towards the minority group and its transmission is simulated using the GBN-Dialogue model of rumour propagation on realistic social networks. The integration of the minority group into the entire social network is measured by the Minority Integration Metric (MIM). Monte Carlo simulations show that rumour penetration into the minority subgroup is an increasing linear function of the MIM and thus minority members of the social network will have a higher level of belief in the minority-derogatory rumour if the minority is more integrated into the social network. A threshold MIM value can be estimated below which the rumour fails to penetrate the minority subgroup.

Keywords: rumor, minority, GBN-Dialogue, out-group negative, clique.

1 Introduction

People tend to divide themselves into two groups, us and them, along many different arbitrary lines of distinction. The social networks considered here are comprised of two different types of people, a majority and a minority with both groups being interconnected within themselves and with one another. There are rumours that circulate, are believed and repeated in one sub-population that are rarely encountered by members outside that sub-population. After the 9-11 terrorist attacks in New York City a rumour circulated in the Arab community in Toronto that no Jews reported for work on that day. This untrue rumour is believed strongly in the Toronto Muslim community but is almost unknown in Toronto outside of that community [1]. This large difference in the levels of belief in the rumour found in the two groups, the Muslims and non-Muslims, is surprising because the two groups are not separated from each other. Relationships exist between the two groups and every day Muslim--non-Muslim dialogues occur about many rumours; the Toronto Maple Leafs trading a veteran player or the possibility of a transit strike, and yet somehow some rumours fail to make the leap from one group to another.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 49–60.
springerlink.com

Transmitting rumours is an information gathering exercise [2]. Rumours are unverified information and we share them in dialogues in order to try and make sense of uncertain information and to reinforce relationships by using the rumour discussion as a bonding mechanism [3]. Thus the type of relationship that links the two interlocutors greatly affects the probability that the rumour will be a part of a given discussion. The 9-11 rumour mentioned above is unlikely to be discussed between a Muslim salesman and his Jewish customer because it would weaken the existing relationship but that same salesman might ask a Muslim taxi driver his opinion of the rumour in order to facilitate bonding and to learn more about the veracity of the rumour. So intuition would be that in a very disperse and integrated Muslim community the rumour would spread more slowly and belief levels would generally be lower (because of rebuttals) as compared to what might occur in a very tight knit insular Muslim community. The social structure of the subgroup would affect the extent of the rumour's propagation. Regarding this intuition; there is much to be quantified.

To understand how the social structure affects rumour propagation Monte Carlo experiments were conducted on realistic artificial social networks. In order to conduct these experiments realistic social networks must be generated using an algorithm that produces social networks that are similar to real social networks. In a given experiment, once the social network is randomly generated its diversity needs to be quantified. The independent variable in these numeric experiments will be the integration of the minority sub-group into the social network. The measure of the minority integration is given by the Minority Integration Metric (MIM) outlined below. The rumour propagation is simulated using the GBN-Dialogue model [4] which was created by an interdisciplinary team of psychologists and mathematicians and is rooted in the social-psychological literature. The dependent variable of the Monte Carlo experiments presented here must quantify the extent to which the belief in the rumour has spread in the minority community. Because the rumour considered here reflects negatively on the minority group the level of belief in the rumour amongst the members of the minority community is expected to be initially less than a neutral ambivalence. The dependent variable that measures the extent of the rumour spread is the time from the inception of the rumour until the average level of belief in the minority members reaches a neutral ambivalence. The time when the average belief level in the minority members reaches the neutral ambivalence level is called the time to neutral belief, TNB.

Thus the hypothesis is that the TNB is an decreasing function of the MIM in a realistic social network, $\frac{\partial TNB}{\partial MIM} \leq 0$. This of course seems intuitive and our experiments will confirm this obvious intuition. Less obvious is the sign of $\frac{\partial^2 TNB}{\partial MIM^2}$ which the simulations will determine.

2 Generating the Artificial Social Networks

A social network is an undirected graph. It consists of a set of people and a set of pair-wise connections between the people. The template for the artificial social

networks used in the Monte Carlo experiments is the Facebook network gathered in 2007 at the Rochester Institute of Technology [5]. The RIT Facebook social network that was collected in 2007 had 4160 people connected by 46936 edges. The RIT students are connected by an undirected edge if they are Facebook friends which each other. The average degree of an individual in the RIT Facebook network is 22, that is, the expected number of people in an individual's neighbourhood is 22 RIT friends. The real social network has properties that the artificial networks must also exhibit in order for the simulations to produce realistic results. The real social network is scale-free meaning the degree distribution follows a power law with most people being connected to a relatively small number of friends and a few people being connected to a very large number of friends. The RIT Facebook network also displays a small-world property [6] because the average path length between people was only 3.37 and yet the clustering coefficient was a relatively high 0.53. Average path length is the average shortest path a rumour could take between any two individuals in the network (degrees of separation). The clustering coefficient is the proportion of an individual's friends that know each other [6]. The clustering coefficient for a network is the average of the individual clustering coefficients.

Thus each Monte Carlo experiment requires the random generation of a scale-free small-world network with a mean degree of 22, an average path length of approximately 3.4 and an approximate expected clustering coefficient of 0.53. Because the typical method of generating scale free networks, the Barabási-Albert (BA) method [7], produces networks with clustering coefficients that are much too small to be realistic approximations for social networks of humans the Mutual Neighbor Preferential Attachment (MNPA) [8] was used. This algorithm for randomly generating scale-free small-world networks can produce the desired simulated social networks with metrics approximating the real RIT Facebook network. Simulated social networks of 400 people were generated using MNPA. The MNPA method is similar to the standard BA method of generating a network except the number of edges linking a new node is drawn from a geometric distribution and the probability that a new node is linked to existing members that are neighbours is weighted. This weighting allows the generation of scale-free networks with higher target cluster coefficients matching the target network.

Once the MNPA algorithm randomly generated a network of 400 people 50 of those people were selected to be minority members. In a real social network whether or not a person is a minority member is dependent on the content of the rumour. The fraction of 1/8 was chosen as the proportion of minority members in the simulated social network as a generalization. One can find many different examples in which the minority is approximately 1/8: 12.4% of the population of the United States of America is African American, 14% of graduates in US mechanical engineering programs are women, 13.2% of Americans are nonreligious or secular. That said, the minority fraction of 1/8 was chosen as a generalization in which the minority members would be numerous enough to have the potential to produce a significant clique but certainly not so numerous as to be a threat to the majority's rumour belief level.

3 Measuring Diversity

The social networks in question are comprised of two groups, a minority group and a majority group. In order to quantify the diversity of the social network as a whole a numeric measure of how well the minority is integrated into the social network is needed.

Consider two different example networks each comprised of 400 people including 50 of the minority group. Each of the two connected social networks has a 12.5% minority but the distribution of that minority in the two groups is different. In the first network the minority consists of a connected complete subgroup with each minority member connected to 1 non-minority member and all the other 49 minority members while in the second network the every minority member is connected to 25 non-minority members and 25 minority members . The degree of each minority member in both cases is 50 and the minority percentage of both networks is 12.5%, but in the second network the minority is clearly more integrated into the entire social network. Metrics such as minority percentage or degree are not adequate to describe social network diversity.

The Minority Integration Metric (MIM) is defined as the fraction of the total edges with at least one minority member that join a minority member to a non-minority member, that is, the ratio between heterogeneous edges (minority--non-minority) to total number of edges (minority--non-minority plus minority--minority).

$$MIM = \frac{total\ number\ of\ edges\ that\ connect\ minority\ to\ majority}{total\ number\ of\ edges\ that\ connect\ minority\ to\ minority\ or\ minority\ to\ majority} \quad (1)$$

In the two example networks above the MIM would be 50/1275 = 0.04 for the less integrated network and 1250/1875 = 0.67 for the more integrated network. Thus the MIM provides a quantitative measure of how integrated is the minority. If the MIM = 0 then the minority is not connected to the rest of the social network; if the MIM = 1 then every minority member is only connected to people outside their minority group. The two example networks are presented to illustrate the MIM but are very dissimilar to the target RIT Facebook network.

Notice in the two example networks described above if 100 more majority members are added to each of the networks by connecting them to only existing majority members the minority percentage drops to 10% but the MIM's remain unchanged. The MIM can be calculated by considering only the edges that emanate from minority members in the social network. The MIM can be estimated by sampling a subset of the minority members. For example consider a large social network where a sample of 100 people are polled and 20 are minority. If the 20 minority members have on average 2 heterogeneous connections and 4 homogeneous connections then the expected MIM is 0.5.

An Erdos-Renyi network consist of n people where a fraction, p, of the total $n(n-1)/2$ possible edges is present. Such a network is constructed by considering each of the possible $n(n-1)/2$ edges in turn and with probability p including that edge. The fraction of the n people that are minority is m and so there are nm minority people and $n(1-m)$ of the majority. The expected MIM for this Erdos-Renyi random graph is m.

The MIM is the independent variable in the Monte Carlo simulations. It will be shown that as the MIM increases so does the penetration of the minority negative rumour into the minority population. The MIM values for network used in the simulations presented here range from 0.8 to 0.98. Lower MIM values are not possible in networks matching the targeted mean degree, mean path length and clustering coefficient.

4 GBN-Dialogue Model

The propagation of the rumour over the social network is simulated using the GBN-Dialogue model [4]. The GBN-Dialogue model is an interdisciplinary research-based mathematical model of rumour propagation on a social network. The GBN-Dialogue model includes a rumour probability transmission function derived from the current socio-psychological literature which is dependent on the personal, contextual, and relational factors. Aspects associated with each participant, the rumour itself, and with the interaction between the participants have been incorporated into the model. Each member of the social network has characteristics that are modeled by parameters such as group membership (majority or minority) that are constant over the time span of the rumour. The two variables associated with each person in the network are the level of belief in the rumour, B_i, and the time of initial exposure, t_{oi}. The belief level of each person changes over time as the rumour propagates across the social network. The time at which a person in the social network is initially exposed to the rumour, t_{oi}, is their reference time which determines their perception of the age of the rumour. People are less likely to pass on rumours they perceive as old news [3]. As the difference between the current time and t_{oi} increases, the probability of a person transmitting the rumour decreases.

The belief level in the rumour is modeled as a continuous variable ranging from a possible $B_i = 0$ meaning the person absolutely does not believe the rumour through $B_i = 0.5$ for a neutral ambivalence to $B_i = 1$ meaning the person believes the rumour to be absolutely true. This belief level changes when an interlocutor discusses the rumour with one of their neighbours.

The GBN-Dialogue model is a two step algorithm. During each iteration of the algorithm one edge in the social network is chosen uniformly at random and the probability of the rumour being transmitted between the two interlocutors connected by that edge is calculated.

$$P = G_{ij}\left(w\left[B_i + B_j - B_i B_j \right] + (1-w)\, N_{ij}\left(1 - \frac{1}{2} H\left(t - t_{0,i} - h \right) - \frac{1}{2} H\left(t - t_{0,j} - h \right) \right) \right) \quad (2)$$

The GBN-Dialogue model does not assume that the probability of a rumour spreading from one person to another is uniform in any given encounter. The transmission probability (Eqn 2) is a function of the belief levels of the two dialogue participants, their perceptions of how new is the rumour and their group statuses. The probability of transmission is an increasing function of the belief levels of the two interlocutors, people tend to spread rumours they strongly belief

to be true. The probability of transmission is a decreasing function of time, people tend not to spread rumours they perceive as old news. The probability of transmission is also a function of the group statuses of the two interlocutors; the probability of transmission between two members of the majority is greater than the probability between two minority members which is greater than the transmission probability between a member of the majority and the minority. The rumour is derogatory towards the minority so it is least likely to be discussed in the heterogeneous relationships as it would negatively affect the existing bond. The transmission probability function of the GBN-Dialogue model has three components; one that is a constant over the time span of a rumour (G), one that can capture aspects that increase rumour transmission (B), one that can capture aspects that decrease rumour transmission such as the loss of Novelty (N). These three factors and their parameters accurately and realistically model the propagation of a rumour.

In Equation 2 w is the weight between the belief motivation, $[B_i + B_j - B_i B_j]$, and the novelty motivation component, N_{ij}. The novelty component contains two Heaviside step functions that become zero when the time elapsed since t_{oi} (or t_{oj}) is greater than the parameter, h. The time that a rumour is still perceived as fresh news is modeled by its hang-time, h. The group status' of the interlocutors is modeled by the parameter G_{ij} which can be one of three values corresponding to the three types of relationships that occur in the social network: majority—majority, majority—minority and minority—minority.

If the rumour is discussed during a given dialogue the belief levels of the two interlocutors are updated. The resulting change in each belief level is dependent on the credibility that each has in the eyes of the other as well as their viewpoints on the rumour's authenticity. Discussing a rumour with a like minded individual (B_i, $B_j < 0.5$ or B_i, $B_j > 0.5$) makes both interlocutors' belief levels more extreme. This results in the two belief levels moving farther away from the neutral value of 0.5 reinforcing prior beliefs. Discussing a rumour with a non-likeminded individual ($B_i < 0.5 < B_j$) results in both belief levels becoming closer to the neutral value of 0.5.

This belief updating combined with the probability transmission function's dependence on the interlocutors' belief levels contribute to the GBN-Dialogue's dynamic feedback features. These feedback features are based on well-established findings in the social influence literature that one's belief is affected by those he interacts with; and, of course, vice-versa. Transmission probability is based upon each dialogue participant's level of belief in the rumour, but these levels of belief are then in turn affected by rumour transmission. The GBN-Dialogue model thus reflects the fundamentally iterative nature of rumour spread.

5 Parameter Set

The type of rumour being simulated is derogatory to the minority group. In a university environment an example of the rumour could be that members of the athletic teams are held to lesser academic standards as compared to the rest of the students. In modeling the group status effect two athletes who have heard the

rumour would discuss this rumour between themselves half as often as two regular students. Discussing an unpleasant possibility in a dialogue between two athletes would still occur more often than the rumour occurring in a dialogue between an athlete and a regular student. Discussion in such a heterogeneous relationship would be the most uncomfortable of the three relationship types. The parameter set of the GBN-Dialogue model simulated here was calibrated to model a general rumour that is derogatory to the minority group. Thus the parameters used to model the effect of group status on transmission probability are set so that, *caeteris paribus*, the probability of transmission between two minority members is half the transmission probability between two majority members and twice the transmission probability between a minority and majority member,($G_{Maj\text{-}Maj} = 1$, $G_{Maj\text{-}Min} = 0.5$, $G_{Min\text{-}Min} = 0.25$) . The parameters noted in this section refer to those in the standard GBN-Dialogue model.

In order to not have either the novelty or the belief factor dominate the other they are given equal weight in the probability transmission function, $w = 0.5$. The effect of novelty in both groups is assumed to be equal and so without loss of generality, $N_{ij} = 1$ and $h = 200$.

People that share the same group status as us are generally viewed with a higher credibility than those from a different group, especially regarding rumours that provide foundation for the differences between the two groups. Thus the intragroup credibility parameters are twice the intergroup credibility parameters, ($c_{Maj\text{-}Maj} = c_{Min\text{-}Min} = 0.5$, $G_{Maj\text{-}Min} = 0.25$).

6 Experimental Procedure

Multiple Monte Carlo simulations of rumour propagation on artificial social networks were run to determine just how the TNB is a function of the MIM. Each data point was obtained by first generating a 400 person social network using the MNPA method and randomly assigning 50 of the people to be minority members. The artificial social network generated is scale-free and small-world. It has network metrics--clustering coefficient, mean path length-- similar to the targeted actual RIT Facebook network. The MIM of the network was calculated and one non-minority member randomly selected as the rumour instigator. The rumour instigator differs from the other non-minority members only by knowing the rumour. The initial belief level of the rumour instigator is equal to the non-minority's prejudice value of 0.75. All the other non-minority members have prejudice level of 0.75 and have not been exposed to the rumour. All of the minority members have prejudice levels of 0.25 and have not been exposed to the rumour.

Each iteration of the simulation begins with an edge being randomly chosen from the set of all edges in the social network and the probability of the rumour being discussed between the two people connected by that edge is calculated using the probability function from the GBN-Dialogue model. If neither of the two people has been exposed to the rumour the probability of discussing the rumour is necessarily zero and a new iteration begins. If at least one of the two interlocutors has been exposed to the rumour the probability function of the GBN-Dialogue

model is used to determine if the rumour is discussed in that dialogue during that iteration. If the rumour is discussed in that dialogue during that iteration the two belief levels are updated using the second step of the GBN-Dialogue model. If one of the two interlocutors had not yet been exposed to the rumour then that person's first exposure time, t_{oi}, is fixed at that iteration number.

Thus the only variables in the simulation are the first exposure time and the belief levels of each person in the social network. The mean belief levels of the majority and minority groups are calculated as an average over all their members who were exposed to the rumour. Similarly the mean belief level of the entire social network is calculated as the average over all the people in the social network exposed to the rumour. These mean belief levels are plotted as a function of time (Figure 1); the top curve is the mean belief level of the majority, the middle curve is the mean belief level of the entire social network and the bottom curve is the mean belief level of the minority group. The iteration number at which the bottom curve, the mean belief level of the minority group, crosses 0.5 is the TNB for that experimental run. This yields the data point from that particular experiment, (MIM, TNB). The simulations were run using Maple.

7 Results

The simulation results confirmed the initial hypothesis. The more integrated the minority was in the social network the quicker their mean belief level reached the neutrally ambivalent value of 0.5. That is, as the MIM is increased the TNB decreases. The plots in Figure 1 show three example simulations with decreasing MIM values. The top trace in each graph is the mean belief level of the exposed majority members, the middle trace is the mean belief level of all the exposed people in the social network and the bottom trace is the mean belief level of the minority members. The time at which the bottom trace reaches 0.5 is the TNB. The values 5800, 6000 and 7500 are the TNB numbers for their associated MIM's.

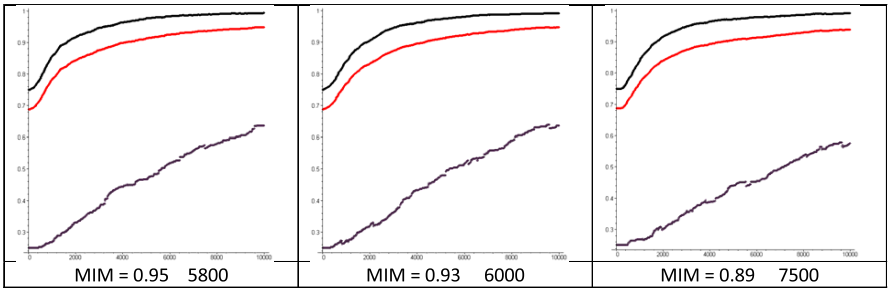


Fig. 1 The three plots show the mean belief levels in three example simulations with time on the horizontal axis. The top trace in each graph is the mean belief level of the exposed majority members, the middle trace is the mean belief level of all the exposed people in the social network and the bottom trace is the mean belief level of the minority members. The time at which the bottom trace reaches 0.5 is the TNB. The resulting three data points are (MIM, TNB) = (0.95, 5800), (0.93, 6000) and (0.89, 7500).

The results not only confirm one's intuition but show that the belief level in the minority can be modeled by a linear relation of time. The mean belief level of the exposed members of the minority group always has an intercept near 0.25 since that is the initial prejudice level of the minority group members. The smaller TNB that is found in networks with a higher MIM is the result of the greater slope of the linear bottom traces in plots such as those in Figure 1. That is, networks with a greater MIM have a greater constant rate of increase in minority mean belief as a function of time and thus reach the mean belief level of 0.5 earlier (smaller TNB).

For the networks simulated to match the target RIT Facebook network the linear relation is $TNB = -29034(MIM) + 33802$ with $R^2 = 0.94$. It must be noted that this equation is only suitable for networks similar to the target network with MIM's in the range of 0.8 to 0.98. That said, in numeric experiments with other types of networks such as random, torus and family torus the results were qualitatively similar in that $\frac{dTNB}{dMIM} \ll 0$. In the simulations of the target network $\frac{\partial^2 TNB}{\partial MIM^2} = 0$ because the relation is linear.

The linear relation gives an expected TNB as a function of the MIM of the network. Because these are Monte Carlo experiments there is a variation in the TNB's for any given MIM. The variation in the TNB times decreased as the integration of the minority increased. This can be seen in Figure 2 which shows the spread in the data points is much less for larger MIM values than for lower MIM values. Figure 2 also illustrates that the expected TNB decreases as MIM increases.

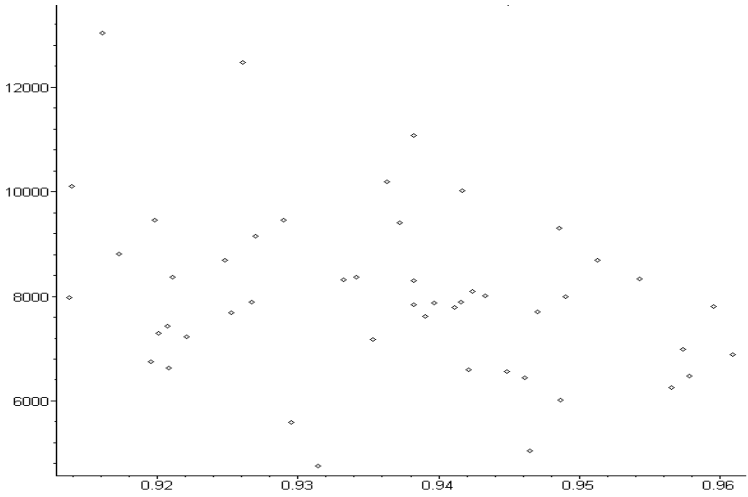


Fig. 2 Each data point represents the result of a simulation. The independent variable on the horizontal axis is the MIM and the dependent variable on the vertical axis is the TNB. This plot shows that as the MIM increases both the TNB and its variance decreases.

8 Threshold Value

In all of the rumour simulations using networks similar to the RIT Facebook target network with 1/8 minority members the mean belief level of the minority population eventually crosses the neutral level of 0.5. That is, given enough time the minority will eventually believe the negative rumour about themselves. Because the simulated networks were generated to be similar to the target network with a 1/8 minority they never obtained a MIM less than 0.8. If a network had a MIM much less than 0.8 it would be conceivable that the minority group would not be convinced to believe the negative rumour about them. The lack of minority integration would protect them from the negative rumour.

In general, given the calibrated set of parameters above, the GBN-Dialogue model has an analytical threshold MIM of 2/3. That is, when a minority is more integrated than MIM = 2/3 the belief in the minority derogatory rumour among the minority members will eventually grow. This results in an integrated minority community with a belief in a negative rumour about themselves. All the rumour simulations involving the networks similar to the RIT Facebook network had MIM's greater than 2/3 and the minority group was eventually convinced at iteration TNB.

Consider a network of n people with m being the proportion of minority members and where the people—minority and majority--have an expected degree of k. There are $\frac{nk}{2}$ total edges in this network. The number of edges that link minority members to other minority members is $mnk\left(\frac{1-MIM}{2-MIM}\right)$ and the number of edges that link a minority member with a majority member is $mnk\left(\frac{MIM}{2-MIM}\right)$. Thus the chance of choosing an edge that links two minority members is $2m\left(\frac{1-MIM}{2-MIM}\right)$ and the chance of choosing an edge that links a minority and majority member is $2m\left(\frac{MIM}{2-MIM}\right)$. The parameters of the GBN-Dialogue model calibrated for a minority negative rumour are such that the rumour is half as likely to be discussed in a dialogue between a minority and a majority member as in a dialogue between two minority members. Therefore the probability of choosing a minority—minority edge and the rumour being discussed is $2m\left(\frac{1-MIM}{2-MIM}\right)$. The probability of choosing a minority—majority edge and the rumour being discussed is $m\left(\frac{MIM}{2-MIM}\right)$.

Initially minority members tend to disbelieve the negative rumour about themselves and consequently when two minority members discuss the rumour they reinforce their prior held disbeliefs. In that case the minority members' belief level would each drop by a value d and the total belief level in the minority drops by $2d$. In conversations between a minority member and a majority member in which the rumour is discussed the minority interlocutor's belief level increases by a value of u since the minority member's new belief level is a credibility weighted average of the two interlocutors' prior belief levels.

The expected change in minority belief per iteration of the GBN-Dialogue model is $2m\left(\frac{1-MIM}{2-MIM}\right)(-2d)+m\left(\frac{MIM}{2-MIM}\right)(u)=m\left(\frac{(u+4d)MIM-4d}{2-MIM}\right)$ which is negative if $<\frac{4d}{u+4d}$. Using the parameter set calibrated to match the minority negative rumour d = 0.0625 and u = 0.125 resulting in the MIM threshold of 2/3. The threshold MIM does not depend on m, the proportion of the population that is minority, or n, the total number of people in the social network. Simulations on other network types such as random and torus with lower MIM values confirm the threshold MIM value.

If one were to use the idea of altering the threshold MIM in order to increase or decrease the minority-negative rumour's penetration into the minority group the two parameters to modify through some sort of marketing effort would be u or d. Notice that $\frac{dMIM}{du}=-\frac{dMIM}{dd}$ and so increasing u or decreasing d would have equal effect at changing the threshold MIM.

9 Conclusions

It has been shown that as a minority becomes more integrated into the social network the time until a minority-negative rumour is believed by that minority decreases linearly in realistic small-world networks. The MIM, a metric for minority integration was presented that betters the degree or minority proportion as a measure of the integration of the minority into the entire social network. A threshold MIM value can be estimated below which the minority-negative rumour fails to infiltrate the minority subgroup. The ratio between the decrease in total minority belief that results when a minority member discusses the rumour with a likeminded minority member and the increase in total minority belief that results when a minority member discusses the rumour with a majority member who believes the negative rumour is $\frac{d}{u}$. If one's goal were to increase the likelihood of a minority-negative rumour penetrating the minority subgroup then one must aim to have at least one majority-minority relationship for every $4\frac{d}{u}$ minority-minority relationships resulting in a MIM greater than the threshold.

References

[1] The Secret History of 9/11: Terrorist Threats Ignored. Canadian Broadcasting Corporation. September 10 (2006)

[2] DiFonzo, N.: The Watercooler Effect: A Psychologist Explores the Extraordinary Power of Rumors. Avery (Penguin), New York (2008), http://www.thewatercoolereffect.com

[3] Bordia, P., DiFonzo, N.: Psychological motivations in rumor spread. In: Fine, G.A., Heath, C., Campion-Vincent, V. (eds.) Rumor Mills: The Social Impact of Rumor and Legend, pp. 87–101. Aldine Press, NY (2005)

[4] Brooks, B.P., DiFonzo, N., Ross, D.S.: The GBN-Dialogue Model of Outgroup-Negative Rumor Transmission: Group Membership, Belief, and Novelty (2011) (in press)
[5] Longo, D., Brooks, B.P.: Modeling the RIT Facebook Social Network. Presented to the Centre for Applied and Computational Mathematics at the Rochester Institute of Technology (2008)
[6] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 409–410 (1998)
[7] Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
[8] Burgers, K.M., Upton, J., Brooks, B.P.: Generating Networks with Target Clustering and Path Length. 2010 NSF-REU in Extremal Graph Theory and Dynamical Systems. School of Mathematical Sciences. Rochester Institute of Technology (2010)

Formal and Informal Networks in Organizations

Silvana Stefani[1] and Anna Torriero[2]

[1] Dipartimento Metodi Quantitativi
Università Milano Bicocca - Italy
silvana.stefani@unimib.it
[2] Dipartimento di Econometria e Matematica
Università Cattolica del S.C. Milano- Italy
anna.torriero@unicatt.it

Abstract. Through the study case of a restructuring process of a company, we illustrate how to detect informal networks in a company. A survey was conducted by questionnaires. We made use of eigencentrality applied to the resulting graphs. Based on the concept of hubs and authorities, we compare the informal relationships within the company with the formal structure and comment the roles and their significance.

Keywords: organizational networks, eigenvector, centrality.

1 Introduction

Organizations are complex structures and often the success of a company depends not only on the ability to produce or sell but also on how efficient the ties areand how smooth the information flows within the organization. The pattern of interactions must be understood, so as to use this methodology to improve information and knowledge ([2]). Moreover, it is also known that within an organization, the activities and the work process coordination often occur through informal relationships rather than on the formal structure ([5]). The informal roles have a great influence in the decision processes, and it has been shown that they can have a strong impact on performance both at the individual level and in the internal organization network ([7]).

The notions and the instruments provided by the Social Network Analysis (SNA) allow the decision maker to model and understand the informal ties. This can be done through specific surveys and assessment of suitable questionnaires. Using the concept of hubs and authorities [8] we can model the flow of information and detect the bottlenecks or the sources of inefficiency in communication. From the analysis of informal networks, the knowledge sharing flows can be enhanced and the formal organizational processes can be joined to the informal, or viceversa and an assessment of the whole organization can follow. The objectives of an informal network analysis within

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 61–77.
springerlink.com

an organization regard the support to: a) integration and efficient cooperation among strategic groups; b) integration and efficient cooperation within strategic groups, c) key roles ([7]). Moreover, power is not only hierarchical and individuals who are central in the informal networks often are even more influential than the established ones. In this paper we will use techniques for SNA, focusing in particular on centrality measures. The use of centrality in organizational networks is widespread and motivated in [9], [1], [2], [5], [11]. We will explain how eigencentrality can be a useful tool for our purposes. Theoretical results on centrality and extensive reviews can be found in [16], [13] for eigencentrality, and [14] for betweenness. Applications of centrality to various fields such as financial markets can be found in [15].

In this paper, we will describe how to detect informal networks through eigencentrality, calculated on asymmetric relationships. This gives rise to the concept of hubs and authorities, i.e. the seekers and the providers of information. Information, as is well known, can be perceived and given at various levels, from the more superficial to a deeper one: information, advices, help, creation of value. Trust is more transversal and covers all levels of information. We will appy those techniques to an Italian consulting company.

The structure of the paper is as follows. Preliminaries with basic definitions from graph theory are in Section 2; Methodology in Section 3 illustrates the tools for detecting hubs and authorities. Conditions are given for the proper use of eigencentrality. The case study is analyzed in Section 4. Section 5 concludes.

2 Preliminaries

First we recall some basic definitions about graph theory.

A graph $G = (V, E)$ is a pair of sets (V, E) , where V is the set of n vertices and E is the set of m pairs of vertices of V; the pair $(i, j) \in E$ $(i \neq j)$ is called an edge of G and i and j are called adjacent $(i \sim j)$; an undirected graph is a graph in which $(j, i) \in E$ whenever $(i, j) \in E$, whereas a directed graph (digraph) is a graph in which each edge (arc) is an ordered pair (i, j) of vertices. Let's denote with $|V|$ and $|E|$ the cardinality of the sets V and E respectively. $G = (V, E)$ is simple if there is one edge between two adjacent vertices. Moreover, a weight w_{ij} is possibly associated to each edge (i, j), in this case we will have a weighted (or valued) graph.

The degree d_i of a vertex i $(i = 1, ..., n)$ is the number of edges incident to it. In a directed graph the indegree of a vertex i is the number of arcs directed from other vertices to i and the oudegree of a vertex i is the number of arcs directed from i to other vertices.

A path is a sequence of distinct adjacent vertices; a $i - j$ path is a path from i to j; a shortest path joining vertices i and j is called a $i - j$ geodesic. A walk is an alternating sequence of vertices and edges. A directed walk is a walk in which each edge e_j is directed from $j - 1$ to j. The distance $d(i, j)$

between two vertices i and j is the length of the $i-j$ geodesic. A graph is connected if for each pair of nodes i and j, $(i, j = 1, 2, ..., n)$, there is a walk from i to j. A digraph is strongly connected if for each pair of vertices there is a directed walk.

A nonnegative n-square matrix A representing the adjacency relationships between vertices of G is associated to the graph (the adjacency matrix); the off -diagonal elements a_{ij} of A are equal to 1 if vertices i and j are adjacent, 0 otherwise. If the graph is simple, the diagonal elements of A are zero; let $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ be the set of the eigenvalues of A, $\rho = \max_i |\lambda_i|$ its spectral radius; if $G = (V, E)$ is undirected, it is well known that A is symmetric and its eigenvalues are real; also, when $G = (V, E)$ is connected, A is irreducible (see [18]). If $G = (V, E)$ is a digraph, its adjacency matrix A is in general asymmetric and G is strongly connected if and only if A is irreducible.

In the case of a weighted graph the adjacency matrix has zero diagonal elements, and nonnegative off-diagonal entries (denoted by W, the weighted adjacency matrix).

For other graph definitions we refer for instance to ([17]).

3 Methodology

Eigenvector centrality and related measures have many advantages over other centrality measures like degree, betweenness and closeness. Indeed they can be used also in weighted or directed graphs and they take into account not only direct, but also indirect links, being proportional to the sum of centralities of the nodes to which each node is connected: each individual status is proportional to the status of the individual to whom he/she is connected.

We recall that in an undirected graph the eigencentrality ([4]) of a node v is defined as the $i-th$ component (the i-th score) of the eigenvector associated to the maximum eigenvalue ρ of the adjacency matrix. This eigenvector is known as the principal, or Perron-Frobenius eigenvector being $v_i = \rho^{-1} \sum_{j=1}^{n} a_{ij} v_j$.

When we deal with asymmetric relations between actors, as in the case of an informal organizational network in which information flows from one node to another in a given direction, the corresponding graph is directed and the adjacency matrix is, in general, asymmetric.

So, in order to provide centrality measures for each node, we consider a generalization of the eigenvalue centrality that, as for standard eigenvector centrality, takes into account the scores of neighboring nodes. This measure allows nodes to have two attributes (status): authority and hubness ([19]). Authority measures *prestige*: actors that many other actors point to are called authorities. If a node has a high number of nodes pointing to it, it has a high authority value, that quantifies its role as a source of information. On the contrary, a hub is an actor asking information to many high authorities and its score measures *acquaintance*. Essentially, a good hub points to many good authorities and a good authorithy is pointed to by many good hubs.

More formally, let $G = (V, E)$ be the directed graph on n nodes modelling the network and $A(G)$ the asymmetric adjacency matrix of G of order n. Two scores (centralities) are associated to each node: x_i, the authority score for node i and y_i the hub score for node i, as follows:

$$x_i = \sum_{j=1}^{n} a_{ji} y_j, \quad i = 1, ..., n; \tag{1}$$

and

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, ..., n; \tag{2}$$

or equivalently

$$\mathbf{x} = A^T \mathbf{y} \text{ and } \mathbf{y} = A\mathbf{x}$$

where vectors $\mathbf{x} = [x_1, ..., x_n]^T$ and $\mathbf{y} = [y_1, ..., y_n]^T$ collect respectively the autorithy and the hub scores on all nodes. Note that if the hub vector $\mathbf{y}$ is known with accuracy the authority scores could be computed from (1), similarly for the hubs scores using (2). Since it is not so, the idea of the HITS algorithm is to start with initial vectors $\mathbf{x}^{(0)}$ and $\mathbf{y}^{(0)}$ and to update hub scores and authority scores, by repeating the process.

Note that hub scores are computed by the current authority values, which in turn were computed from the previous hub scores. It follows that hub and authority values are reciprocally defined in a recursive scheme that is the basis of HITS (Hyperlink-Induced Topic Search) algorithm ([19]). The convergence of HITS algorithm is guaranteed by the normalization of solutions after each iteration and by the condition $\lambda_1(A^T A) > \lambda_2(A^T A)$,being $\lambda_1$and λ_2 the first two eigenvalues of $A^T A$. Denoting by $\mathbf{x}^{(k)}$ and $\mathbf{y}^{(k)}$ authorities and hub scores we get:

$\mathbf{x}^{(1)} = A^T \mathbf{y}^{(0)}$, $\widehat{\mathbf{x}}^{(1)} = \mathbf{x}^{(1)} / \left\|\mathbf{x}^{(1)}\right\|$, $\mathbf{y}^{(1)} = A\widehat{\mathbf{x}}^{(1)}$and at the $k-th$ iteration, $\alpha \mathbf{x}^{(k)} = A^T A \mathbf{x}^{(k-1)}$,$\alpha \mathbf{y}^{(k)} = AA^T \mathbf{y}^{(k-1)}$, being α a normalization coefficient.

Performing power iteration method on AA^T and $A^T A$, $\mathbf{x}^{(k)}$ and $\mathbf{y}^{(k)}$ will converge respectively to the principal eigenvectors $\mathbf{x}^*$ and $\mathbf{y}^*$ of the symmetric semi-positive definite matrices $A^T A$ and AA^T . If we consider the singular value decomposition (SVD) of A ([10]), given by $A = UDV^T$, it is well known that the columns of U are the eigenvectors of $AA^T$and the columns of V are the eigenvectors of $A^T A$, called the left singular vectors and right singular vectors of A, respectively. It follows that $\mathbf{x}^*$ and $\mathbf{y}^*$ correspond respectively to the principal right and left singular vectors of A.

To avoid that zero entries are assigned to important nodes and the non uniqueness of the ranking vector (see following examples) we mention a characterization of graphs on which the HITS algorithm badly-behaves ([20], [8]).

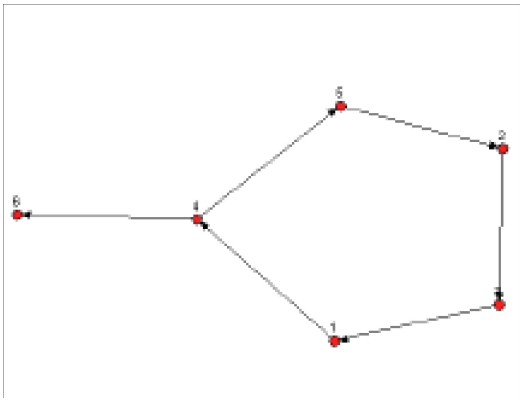


Fig. 1 Example of a bad graph

Starting from G, we construct the undirected HITS authority graph G' taking the nodes of G with positive indegree, where $(i, j) \in E(G')$ if there is a node $k \in G$ such that (k, i), (k, j) are directed edges of G. The following Theorem holds ([8]):

Theorem 1. *The HITS algorithm is badly behaved on a graph G if and only if the HITS authority graph G' is disconnected.*

Example 1. *The graph in Fig.1 is an example of a bad graph. In fact the computation of the first eigenvector of the authority matrix $A^T A$ yields*

$$\mathbf{x} = [0.44721, 0.44721, 0.44721, 0.44721, 0.44721, 0]^T,$$

assigning zero score to the last vertex, even though it has indegree 1. *This result is also confirmed by Theorem 1 , being the resulting graph G' disconnected (all vertices isolated).*

Example 2. *Let us consider the graph H in Figure 2. The spectral radius of the authority matrix is $\rho(A^T A) = 2$ with algebraic multiplicity equal to* 2 *and the corresponding (first) eigenvector $\mathbf{x} = [0\ \alpha\ \beta\ \alpha]^T$. This implies different ranking of authority scores depending on the choice of parameters α and β such that $2\alpha^2 + \beta^2 = 1$, due to the normalization. For example both pairs $\alpha = \frac{1}{2}$, $\beta = \frac{1}{\sqrt{2}}$ and $\alpha = \frac{1}{4}$, $\beta = \sqrt{\frac{7}{8}}$ give authority scores, but the uniqueness of ordering fails and the result is of no practical use. Also in this case the resulting graph H' obtained by Theorem 1 is disconnected so that H turns out to be a bad graph.*

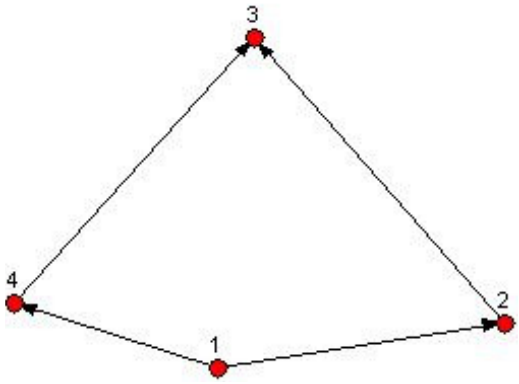


Fig. 2 Example of a bad graph

4 The Case Study

A family-owned consulting company has to deal with power transition. The owner wants to retire and leave control to heirs, part of whom are involved directly in the management.

Since business has always been doing very well, a need was never perceived to structure the organization formally, i.e. assigning people to specific sectors.

The company, of small size[1], is characterized by flexibility and ability to quickly adjust to changes. In particular, people working in the front office area were left free to establish contacts with clients and assist them from the beginning to the end.

Even though people were motivated in pursuing the interests of the company, on the other hand it left too much freedom and gave the opportunity to some brilliant and ambitious consultants (Frank, not belonging to the family) to manage too much information within the organization and even to attempt a sort of take over. Thus, a restructuring process is needed, to allow a smooth transition and to give again the necessary power to family members.

After extensive interviews and talks, we proposed a new formal organization of the company. This new formal chart (Fig.3) fits all Ted requests.

The detailed description of the process of restructuring the organization chart can be found in [11]. We made use of eigenvector centrality for weighted and unweighted bipartite graphs, as well as betweenness and flow betweenness centrality.

However, after restructuring, the committant still felt that a more thorough analysis on the informal relationships within the organization should

[1] The company staff is composed of : Ted (the owner); Alex and Anna (his daughters), Charly (Alex's husband), Frank and Coch (consultants close to the family), Chin, Ary, Dominic, Four, Dellor, Taggy (external consultants), Jac (a new employee), Joan and Laura (secretaries), Sara, Anna (other roles).

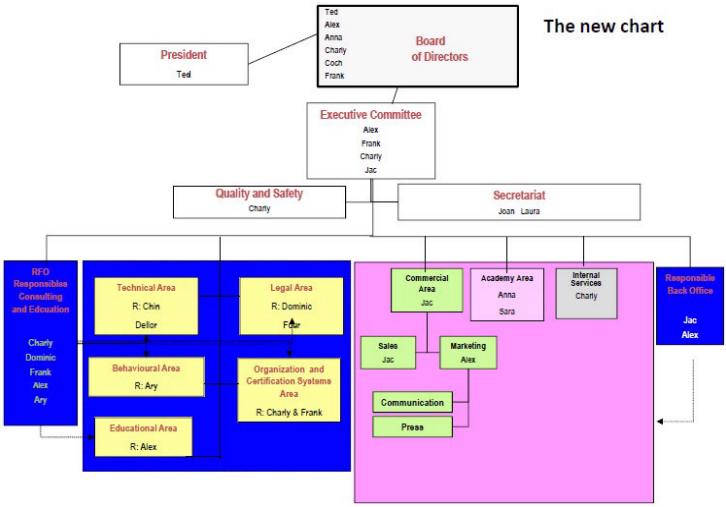


Fig. 3 New formal chart

be performed. The flow of information is probably not efficient and we have to understand what the informal ties are and try to "adapt" the informal to formal, or viceversa, through specific actions. In fact, if the analysis shows that the leading and crucial roles in the organization chart are assigned to people who are collaborative, give advices and understand the need of the others, then the whole organization is efficient. If, on the other hand, we discover that help is required or some sectors do not interact effectively, then an intervention is needed.

In synthesis, the restructuring process can be divided in three phases

1st Phase: Formal restructuring
2nd Phase: Understanding informal ties
3rd Phase: The decision
In this paper we describe the instruments, the methodology and the results related to the second phase.

The third phase is still in progress.

4.1 The Questionnaire

A questionnaire, composed of ten questions, was distributed to all members, including the owner. A typical format is in Fig. 4. The six possibile answers range from 0 to 6: 0=no connection, 1-5= various levels of connections, from the weakest to the strongest level of connection.

Level / People	1 Very seldom	2 Seldom	3 Fairly seldom	4 Fairly frequently	5 Frequently	6 Very frequently
Charly						
Ary						
Chin						
Frank						
Coch						
Delbr						
Dominic						
Joan						
Laura						
Jac						
Four						
Sara						
Anna						
Alex						
Taggy						
Ted						

Fig. 4 The questionnaire

The objective was to discover links at various levels of information, from the most superficial and operative up to the decision level. A couple of questions on trust were included:

1. *Simple information - I ask information from and I give information to*
   How often do you receive information from (Question Q1) or give information to (Question Q6), this person to accomplish work?
   These questions establish the frequency of contacts, valuing the strength of the direct relationship
2. *Problem resolution - I ask advices to solve problems and I give advices to*
   How often do you contact (Question Q2), or are contacted by (Question Q7), this person to get (or give) help and advice to solve a problem at work?
   These questions establish how well people know each other's skills and how available people are to solve problems.
3. *Decision making - I ask for help to take a decision and I give help to*
   How often do you contact (Question Q3), or are contacted by (Question Q8), this person before taking a decision?

These questions can discover potential bottlenecks among managers in an organization but also the strength of ties at the decision level among people.

4. *Creation of value - I spare time thanks to him/her and I let him/her spare time thanks to me*

   How much time was spared last month by contacting (Question Q4), or being contacted by (Question Q9), this person and receiving or giving help and advices?

   These questions reveal the key contributors in an organization, along with a baseline time value that can be used for future metrics and evaluation

5. *Mistrust - I hide information and information is kept hidden from me*

   How often do you get the feeling that this person is not sharing with you useful information (Question Q5)? How often, for various reasons, don't you share useful information with this person (Question Q10)?

   These questions identify the levels of trust and mistrust existing in a company. Information may be hidden not purposely, by mistake or simply by lack of time to meet and talk.

The format and the structure of the questionnaire were taken from [1] and [2], but we added new insights. Take for instance *Problem resolution*, in which questions Q2 and Q7 are coupled. Matching questions on the various levels allowed us to understand if the perception of giving advices was as strong as the perception of receiving them. We can call it a sort of *bilateral awareness*. If A claims to give advices to B but B does not perceive to receive advices from A, the level of communication and understanding between A and B is not good. Thus, from the coupling of questions we can compare for instance hubs of Q2 (measuring the intensity on getting help, from the point of view of the asking person) and authorities of Q7 (measuring the intensity in the perception of getting help, from the point of view of the giving person).

All the resulting adjacency matrices satisfy Theorem 1 so that the analysis of centralities can be properly performed. We focused on the analysis of Questions Q2 and Q7. For computational purposes we set a threshold and grouped the answers to two categories (0=0-2, 1=3-5).

For each question, we computed hubs and authorities. The results for Question Q2 are reported in Table 1 and in Figure 5. Each person is represented by a dot. On the x-axis we read the intensity (normalized eigencentrality) of each person in contacting other people to be helped in problem resolution (Hub), on the y-axis we read the intensity (normalized eigencentrality) of the person being contacted by other people to solve probems (Authority).

The line $x = y$ is the boundary between hubs and authorities. Dots above this line indicate people who are more authorities than hubs, while the opposite happens below that line. Dots close to the line are at the same time givers and takers, and show a good level of communication, mutual help and contacts with other people.

Table 1 Hubs and auth. Q2

People	Hubs Q2	Authorities Q2
Charly	0,32535	0,38952
Ary	0,42749	0,19437
Chin	0,13923	0,36418
Frank	0,36303	0,37755
Coch	0,25577	0,13634
Dellor	0,10576	0,20725
Dominic	0,22215	0,21008
Joan	0,3623	0,171
Laura	0,16191	0,18238
Jac	0,41806	0,28665
Four	0,17588	0,33279
Sara	0,17452	0,11841
Anna	3,032E-57	0,12873
Alex	0,095011	0,2665
Taggy	0,086781	0,12891
Ted	0,15512	0,23662

Charly and Frank are good help givers and takers, while Chin, Ted and Alex are more help givers than takers. Jac asks for help more than he gives, but this is quite reasonable, since he is new in the organization.

The same analysis was performed for Question Q7. Authorities of Q7 are people who perceive being contacted frequently by others. Results are in Figure 6 and in Figure 2.

Finally we checked for bilateral awareness. We compared Hubs of Q2 and Authorities of Q7. Results are in Table 2 and in Table 3.

On the line $x = y$ we find people who ask for help or advice and are perceived by the others doing so. In the region below the line $(x > y)$ we find people who ask for help or advice but are not perceived doing so. Thus may be due to shyness, difficulty to communicate or sporadic relationships. This is the case especially of Ary, who is an external consultant and visits rarely the company's headquarters. On the region above the line $(x < y)$ we find people who are perceived by others asking but who claim non to ask so frequently for help or advice. The general perception does not correspond to what the person claims. This is the case of Dellor, Four and Chin, all external consultants. Also Alex is in a critical position: she claims asking other people for help rarely, but this not exactly what is perceived by others.

A special attention deserves questions on trust and mistrust (Q5 and Q10). Once again, hiding information is not necessarily done purposely, but possibly by lack of time or occasions to meet. In Table 4 and Figure 8 we report Hubs and Authorities of Q5 and the graph Q5 Hubs - Q10 Authorities respectively.

Since Q5 is a sensible question, we include the graphical detail of the answers (se Fig .9).

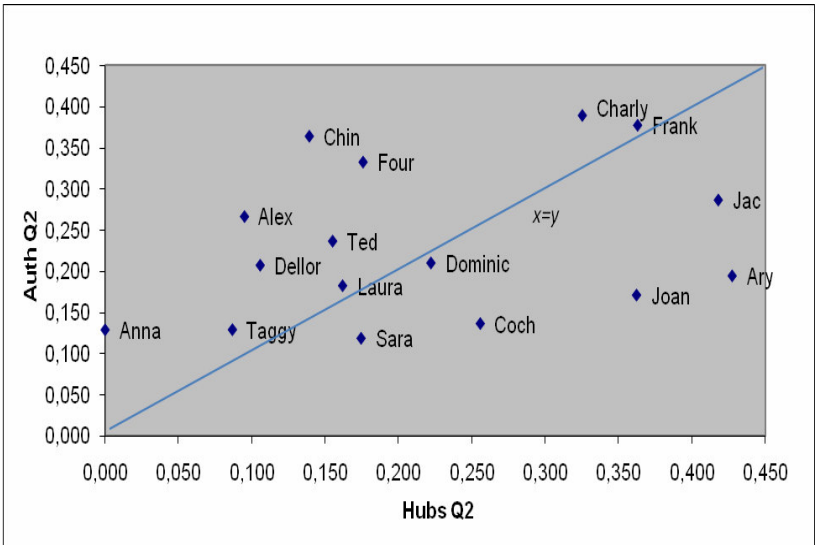


Fig. 5 Problem solving (Q2) hubs-authorities

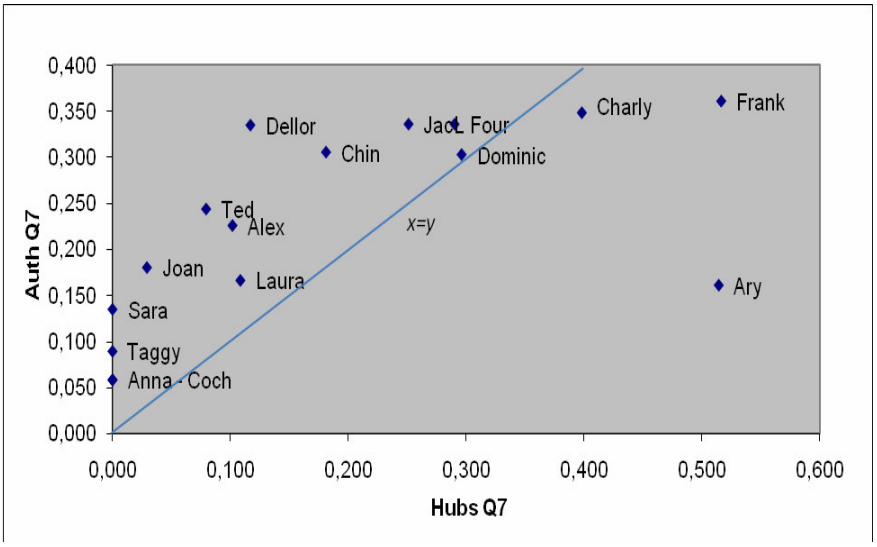


Fig. 6 Problem solving (Q7) hubs-authorities

Table 2 Hubs and auth. Q7

People	Hubs Q7	Authorities Q7
Charly	0,39806	0,34762
Ary	0,51417	0,16083
Chin	0,1811	0,30495
Frank	0,51627	0,36018
Coch	0	0,058139
Dellor	0,117	0,33417
Dominic	0,29608	0,30231
Joan	0,029194	0,17991
Laura	0,10847	0,16597
Jac	0,25109	0,33523
Four	0,2901	0,33522
Sara	0	0,13461
Anna	0	0,058139
Alex	0,10177	0,22543
Taggy	0	0,089099
Ted	0,079422	0,24333

Table 3 Hubs Q2, auth. Q7

People	Hubs Q2	Authorities Q7
Charly	0,32535	0,34762
Ary	0,42749	0,16083
Chin	0,13923	0,30495
Frank	0,36303	0,36018
Coch	0,25577	0,058139
Dellor	0,10576	0,33417
Dominic	0,22215	0,30231
Joan	0,3623	0,17991
Laura	0,16191	0,16597
Jac	0,41806	0,33523
Four	0,17588	0,33522
Sara	0,17452	0,13461
Anna	3,032E-57	0,058139
Alex	0,095011	0,22543
Taggy	0,086781	0,089099
Ted	0,15512	0,24333

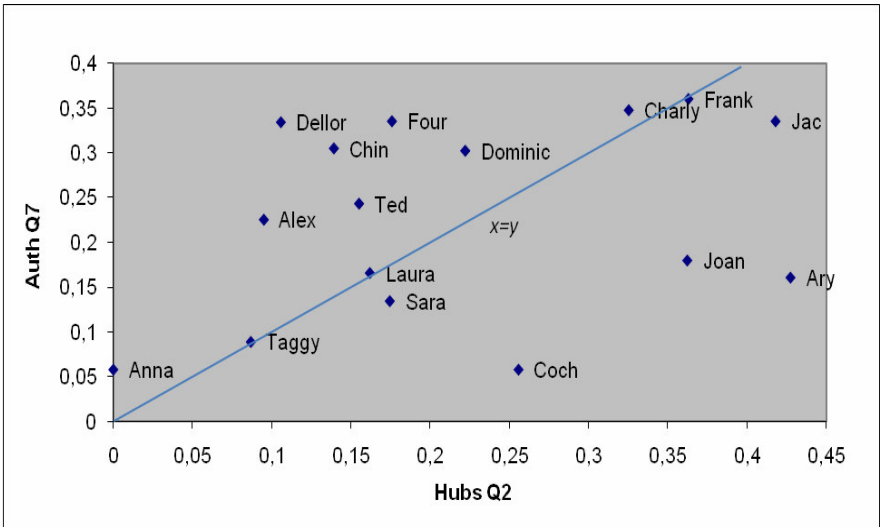


Fig. 7 Bilateral awareness in problem solving (Q2-Q7)

Fortunately, there are many 0 entries, meaning that most people do not believe that others hide information, thus many nodes are isolated. However, it is apparent that Frank is the main collector of mistrust within the company. This is confirmed by Figure 4 in the column of Authorities. Frank has the highest value among authorities, while on the other hand has a hub value 0: he feels that nobody is hiding information from him. On the other hand, high positive values in the hub column are associated to Charly, Dominic, Alex, Ted. It is quite strange that this feeling is shared by all managers of the company, except Frank. Continuing the analysis and checking for bilateral awareness, from Figure 8 we compare hubs of Q5 (people who feel not receiving information by others) with authorities of Q10, that is people with whom others do not share information. Again, the $x{=}y$ line indicates the exact balance between feeling and what is actually happening (i.e. not sharing information). It seems that people do not like sharing information with Frank, but Frank does not feel the same. The same happens to Chin, Ary and Dellor. A solution to this situation, that can be pathological if not treated, is to organize periodical meetings and brainstormings, involving in particular external consultants and all managers.

In conclusion, Charly has a formal leading role in the organization, but is also first among authorities (AQ2, AQ7) and a good recognized problem solver (HQ7) in the informal network. Frank has a leading role in the organization, is second among authorities (AQ2), but he underestimates the value

Table 4 Q5 - Hubs and Authorities

People	Hubs Q5	Authorities Q5
Charly	0,48155	0,28992
Ary	8,40E-41	6,11E-40
Chin	0	0
Frank	0	0,7537
Coch	0	0
Dellor	0	0
Dominic	0,42601	0,46377
Joan	0	0
Laura	0,48867	0
Jac	0	0,15611
Four	0,48155	0
Sara	0	0
Anna	0	0
Alex	0,24077	0
Taggy	0	0,15611
Ted	0,24077	0,28992

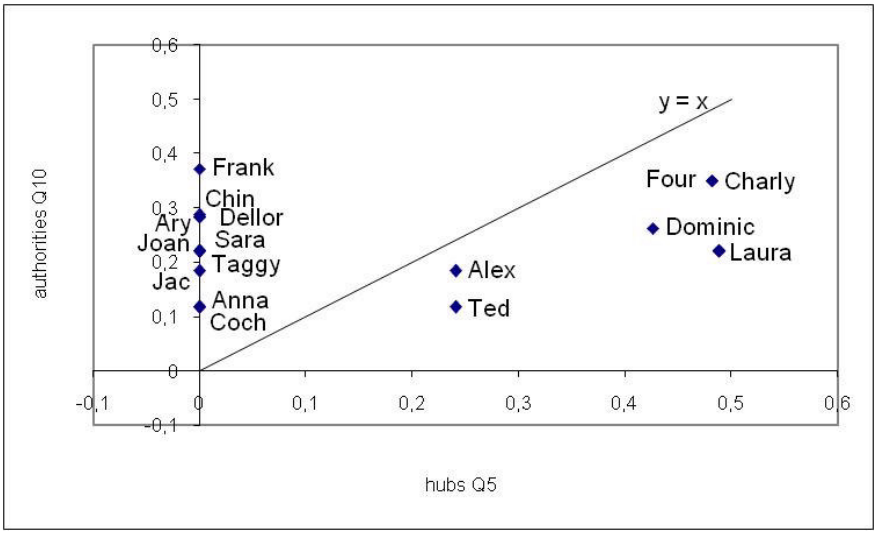


Fig. 8 Hubs Q5 - Authorities Q10

of asking other people for help (HQ2) and does not seem to care much of what other people think (A5, A10). Ary is first among hubs (HQ2, HQ7), he asks a lot for help but people do not perceive his requests properly. Dellor, Four and Chin are in a more critical situation. Jac is second among hubs (HQ2), a new entry in a managing role in the organization, asks for help and people recognize his needs.

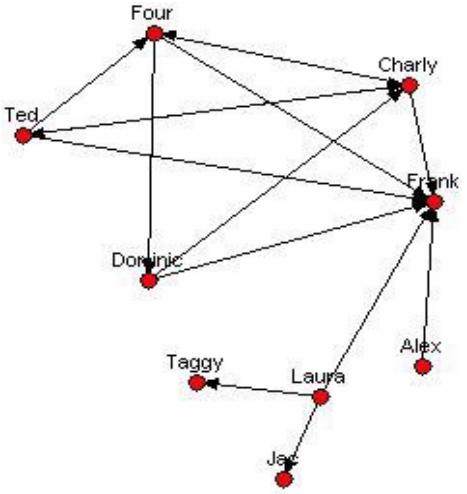


Fig. 9 Answer to Q5

As a final comparison with the official chart, we can conclude that the informal role of Charly and Frank is coherent with their official position; Alex, important as a formal leading role, does not emerge in the informal network; Ted, by his own choice, keeps a low profile.

5 Conclusions

We limited so far the analysis to Questions Q2, Q7 and Q5,Q10. From this partial results, we concluded that the formal structure reflects the informal one quite nicely. However, thorough work has to be done to establish more ties with the external consultants and to reconstruct trust among managers, in particular Frank. Besides the extension to the other coupled questions, our aim is to add a further network, obtained as the intersection of the first four informal networks (simple information, problem resolution, decision making, creation of value) [7]. Formally, this can be obtained through the Hadamard Product of the adjacency matrices ([22]). This extension will contribute to the understanding of the inner ties in the organization.

Acknowledgments. We would like to thank Giovanni Zambruno for his careful reading and comments and Luca D'Ambrosio for providing the numerical support.

References

[1] Anklam, P.: Knowledge Management: the Collaboration Thread. Bulletin of the American Society for Information Science and Technology 28(6) (2002)
[2] Anklam, P.: KM and the Social Network, Knowledge Management Magazine (2003)
[3] Anklam, P., Cross, R., Gulas, V.: Expanding the field of vision. The Learning Organization 12(6) (2005)
[4] Bonacich, P.: Power and centrality: A family of measures. American Journal of Sociology 92, 1170–1182 (1987)
[5] Cross, R.: The hidden power of social networks: understanding how work gets really done in an organization. Harvard Business School Press (2004)
[6] Cross, R., Thomas, R.: How Top Talent Uses Networks and Where Rising Stars Get Trapped. Organizational Dynamics, 165–180 (2008)
[7] De Toni, A.F., Nonino, F.: The key roles in the informal organization: a network perspective analysis. The Learning Organization 10(17), 86–103 (2010)
[8] Farahat, A., Lofaro, T., Miller, J.C., Rae, G., Ward, L.A.: Authority rankings from HITS, PageRank and SALSA: existence, uniqueness and effect of inizialization. SIAM Journal of Scientific Computing 27(4), 1181–1201 (2006)
[9] Friedkin, N.E.: Structural Bases of Interpersonal Influence in Groups: A Longitudinal Case Study. American Sociological Review 58(6), 861–872 (1993)
[10] Golub, G., Van Loan, C.: Matrix computations, 3rd edn. The Johns Hopkins University Press, London (1996)
[11] Grassi, R., Stefani, S., Torriero, A.: Using bipartite graphs to assess power in organizational networks: A case study. In: Rotundo, G. (ed.) Dynamics of Socioeconomic Systems, DYSES Journal, pp. 199–216 (2011) ISSN 1852-379X
[12] Grassi, R., Stefani, S., Torriero, A.: Centrality in Organizational Networks. International Journal of Intelligent Systems 25(3), 253–265 (2010)
[13] Grassi, R., Stefani, S., Torriero, A.: Extremal Properties of Graphs and Eigencentrality in Trees with a given Degree Sequence. The Journal of Mathematical Sociology 34(2), 115–135 (2010)
[14] Grassi, R., Scapellato, R., Stefani, S., Torriero, A.: Betweenness centrality: extremal values and structural properties. In: Naimzada, A.K., Stefani, S., Torriero, A. (eds.) Networks, Topology and Dynamics - Theory and Applications to Economics and Social Systems. Lecture Notes in Economics and Mathematical Systems, vol. 613, pp. 161–176. Springer, Heidelberg (2009) ISBN: 978-3-540-68407-7
[15] D'Errico, M., Grassi, R., Stefani, S., Torriero, A.: Shareholding Networks and Centrality: an Application to the Italian Financial Market. In: Naimzada, A.K., Stefani, S., Torriero, A. (eds.) Networks, Topology and Dynamics - Theory and Applications to Economics and Social Systems, vol. 613, pp. 215–228. Springer, Heidelberg (2009) ISBN: 978-3-540-68407-7
[16] Grassi, R., Stefani, S., Torriero, A.: Some Results on Eigenvector Centrality. The Journal of Mathematical Sociology 31, 237–248 (2007)
[17] Harary, F.: Graph theory. Addison-Wesley, Reading (1969)
[18] Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press (1985)
[19] Kleinberg, J.M.: Authorative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)

[20] Miller, J.C., Rae, G., Schaefer, F., Ward, L.A., Farahat, A., LoFaro, T.: Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 444–445 (2001)
[21] Golub, G., van Loan, C.: Matrix computations, 3rd edn. The Johns Hopkins University Press, London (1996)
[22] Schott, J.R.: Matrix Analysis for Statistics, 2nd edn. John Wiley & Sons, HoboKen (2005)

Assessing Consumer Credit Applications by a Genetic Programming Approach

Salvatore Rampone[1], Franco Frattolillo[2], and Federica Landolfi[2]

[1] Department for the Study of Biological, Geological and Environmental Sciences (DSBGA) University of Sannio, Benevento, Italy
rampone@unisannio.it
[2] Department of Engineering, University of Sannio, Benevento, Italy
{frattolillo,landolfi}@unisannio.it

Abstract. Credit scoring is the assessment of the risk associated with lending to an organization or an individual. Genetic Programming is an evolutionary computational technique that enables computers to solve problems without being explicitly programmed. This paper proposes a genetic programming approach for risk assessment. In particular, the study is set in order to predict, on a collection of real loan data, whether a credit request has to be approved or rejected. The task is to use existing data to develop rules for placing new observations into one of a set of discrete groups. The automation of such decision-making processes can lead to savings in time and money by relieving the load of work on an "expert" who would otherwise consider each new case individually. The proposed model provides good performance in terms of accuracy and error rate.

Keywords: Consumer credit, Genetic Programming, Risk Assessment, Credit Request.

1 Introduction

Credit scoring is the assessment of the risk associated with lending to an organization or an individual. The history of credit scoring begins in 1941 with the publication by Durand (1941) of a study that distinguished between good and bad loans made by 37 firms. Since then the already established techniques of statistical discrimination have been developed and an enormous number of new classificatory algorithms have been researched and tested. Virtually all major banks use credit scoring with specialized consultancies providing credit scoring services and offering powerful software to score applicants, monitor their performance and manage their accounts (Crook et al., 2007).

The main role of such systems is in a binary classification setup represented by an automated system for a lending institution that can assist credit professionals in

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 79–89.
springerlink.com

assessing consumer credit applications, for instance, in deciding whether to accept or reject a credit request. Lenders can resort to automate the credit review process to automatically approve customers that would be good risks, automatically reject customers who would be bad risks, and select a subset of applications for further review by an expert. In fact, the credit approval by a financial institution, such as a bank, is an important decision problem. Credit risk evaluation decisions are key determinants of success for financial institutions in the lending industry due to the heavy losses associated with wrong decisions. Moreover, humans are not good at evaluating loan applications. For instance, the presence of a physical or emotional condition might distort the judgmental capability, so a knowledge discovery tool is needed to assist the decision maker to make decisions regarding credit approval applications. This knowledge can be used as a risk management tool to enhance the economic efficiency and the profitability of credit providers. To this end, consumer credit reporting can be considered an important instrument for lenders in order to evaluate a consumer's credit application and his or her creditworthiness (Lahsasna et al. 2010).

The commonest method of estimating a classifier of applicants into those likely to repay and those unlikely to repay is logistic regression. Many alternative classifiers have been developed in the last thirty years, as linear regression, mathematical programming, and classification trees (Crook et al 2007). A number of other classification techniques such as genetic algorithms (GA), support vector machines (SVM) and nearest neighbours have been piloted but have not become established in the industry (Handzic et al.2003) (Ong et al., 2005).

Genetic Programming (GP) is an evolutionary computational technique that enables computers to solve problems without being explicitly programmed. It works by using genetic algorithms (Goldberg, 1989) to automatically generate computer programs. The most common representation, used in genetic programming, of an individual, a computer program or a function, in the population is in the form of parse trees. Trees can perform classification by returning numeric values and then translating these values into class labels. This Data Classification consists in assigning a class label to a data instance based upon knowledge gained from previously seen class labeled data (Hajira and Baig., 2010). For binary classification problems, the division between negative and non-negative numbers acts as a natural boundary for a division between two classes. This means that genetic programs can easily represent binary class problems (Faraoun and Boukelif, 2007).

In this paper we propose a genetic programming approach for risk assessment. In particular, the study is set in order to predict, on the collection of real loan data, whether a credit request has to be approved or rejected. The task is to use existing data to develop rules for placing new observations into one of a set of discrete groups.

The paper is organized as follows: in the next section we describe the genetic programming approach details; then, in section 3, we show its use in assessing

consumer credit applications; the experimental results are reported in section 4. A toy example of parse tree representation of consumer credit evaluation models is reported in the Appendix.

2 Genetic Programming

Genetic programming was proposed by Koza (1994) to automatically extract intelligible relationships in a system and has been used in many applications such as symbolic regression (Davidson et al., 2003), and classification (De Stefano et al., 2002; Zhang & Bhattacharyya, 2004).

GP evolves computer programs, traditionally represented as tree structures (Cramer, 1985). Every tree node has an operator function and every terminal node has an operand, making mathematical expressions easy to evolve and evaluate. The function set is composed by arithmetic operators ($\{+,-,x,/\}$) or conditional statements ($\{>,<,=\}$) and/or any other kind of function. The terminal set contains all inputs, constants and other zero-argument in the GP tree. An example of this structure is reported in Figure 1 and a simple application to consumer credit is reported in the Appendix.

Many candidate solutions are considered and a measure of the predictive accuracy or fitness is computed for each solution. From this sample of solutions a random sample is selected where the probability of selection depends on the fitness.

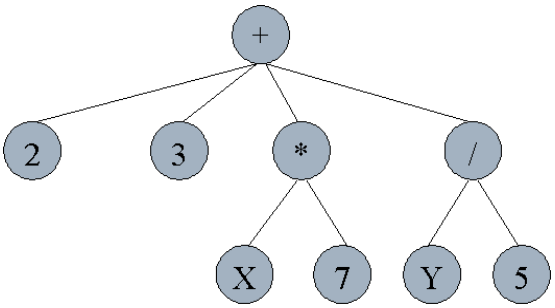


Fig. 1 Tree structure representation of $f(x,y)=2+3+7x+y/5$

Then genetic operators such as crossover and mutation are applied. The crossover operator is used to swap a subtree from two 'parents' to reproduce different 'children'. An example of a crossover in GP is shown in Figure 2.

The mutation operator is used to randomly choose a node in a subtree and replace it with a new created random subtree. An example of a crossover in GP is shown in Figure 3.

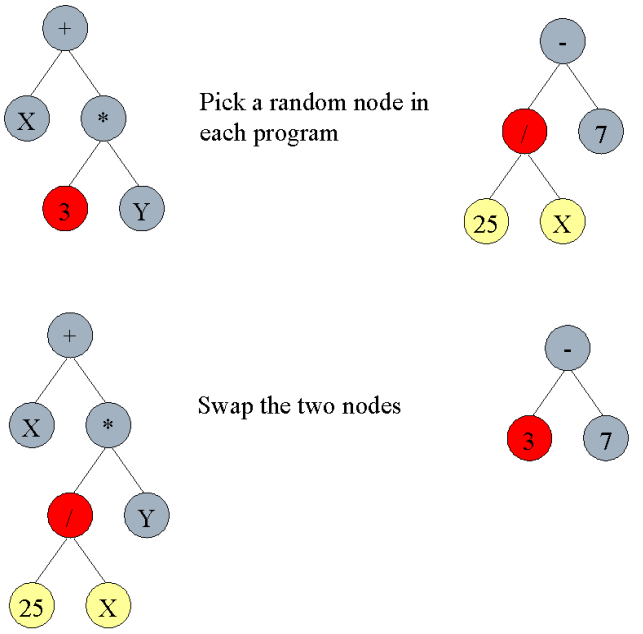


Fig. 2 Example of crossover

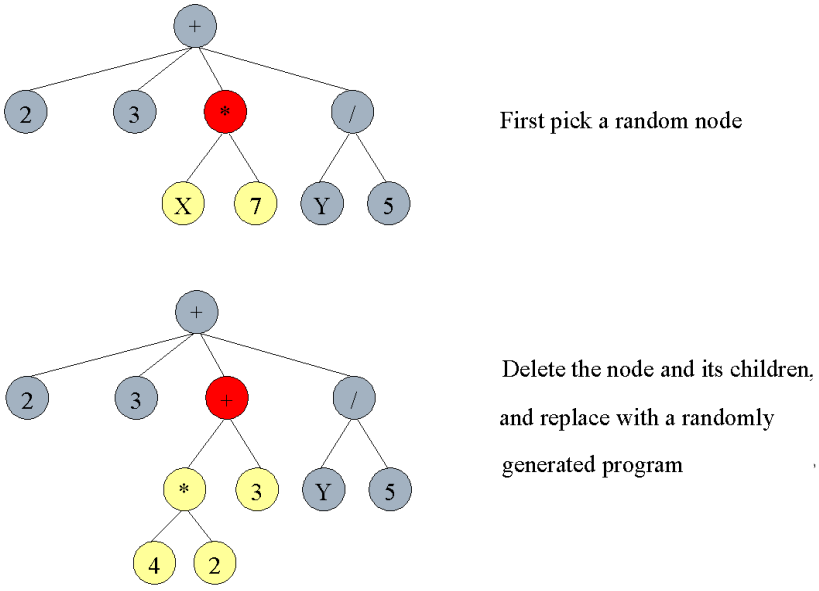


Fig. 3 Example of mutation

Moreover, in order to satisfy the principle of parsimony, the depth of the GP-tree should also be limited. So, essentially, as reported in Figure 4, Genetic programming, uses four steps to solve problems:

- Generate an initial population of random compositions of the functions and terminals of the problem (computer programs).
- Execute each program in the population and assign it a fitness value according to how well it solves the problem.
- Create a new population of computer programs by applying genetic operators (mutation, crossover, etc.) to some selected tree (best fit trees are selected most likely)
- The best computer program that appeared in any generation, the best-so-far solution, is designated as the result of genetic programming.

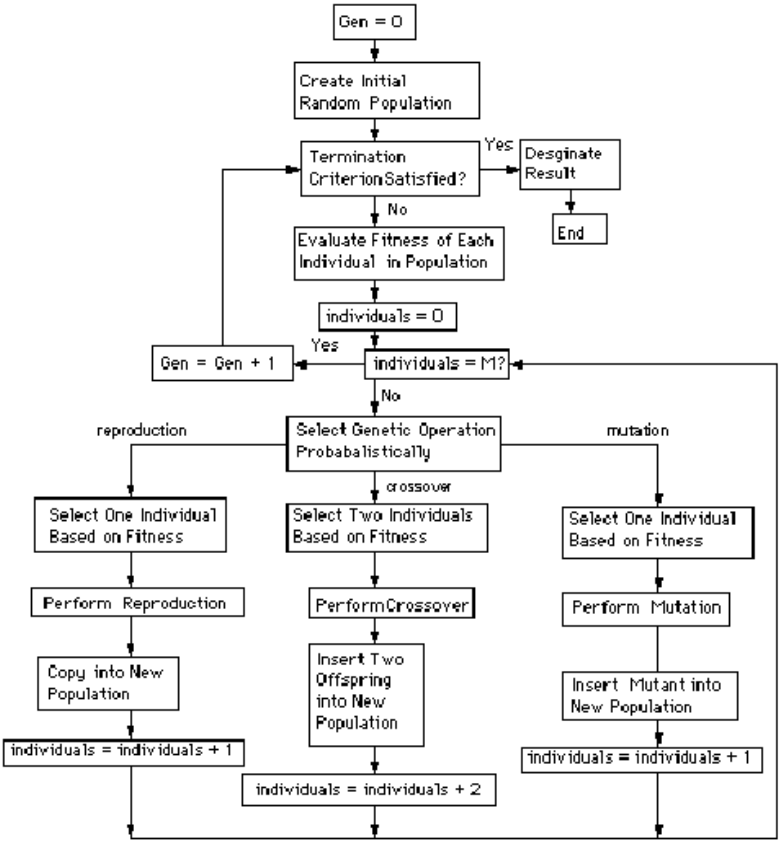


Fig. 4 Genetic programming flowchart

3 Assessing Consumer Credit Applications

3.1 Problem Formalization

We are looking for a function that maps between a set of customer features (inputs) and the customer creditworthiness (output) in order to predict whether a credit request has to be approved or rejected (binary classification).

$$f(x_1,x_2,\ldots,x_m)= y_n \quad (1)$$

where $x_1,x_2,\ldots,x_m$ are the customer features that in the following we call *attributes*, and y_i denotes the type of customer, for example good (1) or bad (0).

We propose to capture this relationship from a set of historical information. So our problem is: given a set of labeled feature vectors (data points), find a mathematical model that explains the data.

3.2 Data Set

The actual data used for this study, called *Credit Approval Data Set*, was taken from the UCI Repository of Machine Learning Databases and Domain Theories (Frank and Asuncion, 2010). This file concerns credit card applications. This data set, in a slightly different form, is also known as *Australian Credit Data Set.*

The selection of the relevant attributes and the related assigned points are based on the experience of credit professionals. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. However, important information include: customer's credit history, the amount of the loan, the things that the applicant wishes to accomplish with the loan, the applicant's repaying ability depending on income, number of he/she have or how long has he/she been working in their job, and, finally, the condition of current economic market.

Table 1 Summary of Credit Approval Data Set characteristics

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer, Real
Associated Tasks	Classification
Number of Instances	690
Number of Attributes	15
Missing Values	Yes
Class Distribution	+: 307 (44.5%) -: 383 (55.5%)

Six of the attributes are continuous, four are binary, and the remaining 5 are nominal with a large number of values.

Of the 690 observations, 307 (44.5%) correspond to accepted applications, and 383 (55.5%) correspond to rejected applications. There are also 37 (5%) cases which have missing values, but they are not considered in this study.

A summary of the characteristics of this data set is reported in Table 1.

3.3 Preparatory Steps

For generating the model by using genetic programming, there are a number of preparatory steps to do.

First we have to determine the representation scheme, i.e. the:

- set of terminals (ex: { $x_1,x_2,\ldots,x_m$})
- set of functions (ex: *{=, +, -, *, /, sin, cos,…}*)

Then we have to determine the fitness measure. This specifies what type of error to measure when comparing and optimizing solutions.

$$err(f) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 \tag{2}$$

For example, it is preferable to minimize "Squared Error" if the data has normally distributed noise, or "Logarithmic Error" if it contains many outliers.

Next step is in determining the parameters:

- Population size,
- Number of generations
- Probability of crossover, mutation, reproduction

Finally we have to determine the criterion for terminating a run (max number of generations or exact solution).

Performing optimal choices for all these parameters is still a quite dark side of the procedure and must be achieved on the basis of personal experience and a trial and error approach.

4 Experimental Results

Given the data set characteristics, we are looking for a formula *f()* that satisfies

$$y = f(x_1,x_2,\ldots x_{15}) \tag{3}$$

We limit the set of functions to the arithmetic operators (*{+,-,x,/}*) plus two trigonometric functions (*{sin, cos }*)

As fitness measure we use the mean square error (MSE).

The exemplars are split in two groups for training and validation. The training set is created by randomly selecting samples from the whole dataset.

A summary of the GP selected parameters is reported in Table 2.

Table 2 Summary of GP selected parameters

PARAMETERS	
Search for a formula f() that satisfies	$y = f(x_1,x_2,\ldots x_{15})$
Building blocks operations (6)	+ , - , *, /, *sin*, *cos*
Variables (16)	$x_1,x_2,\ldots x_{15}$, y
Training samples	656
Validation samples	321

We use a genetic programming software tool called Eureqa, for detecting equations and hidden mathematical relationships in a given data set. Eureqa works in order to reduce the error function, given by the discrepancy between the data and the generated model (Schmidt and Lipson, 2009). Its goal is to identify the simplest mathematical formulas which could describe the underlying mechanisms that produced the data. It also allows to evidence the behaviour of each solution with respect to its size (see Figure 5 and Figure 6).

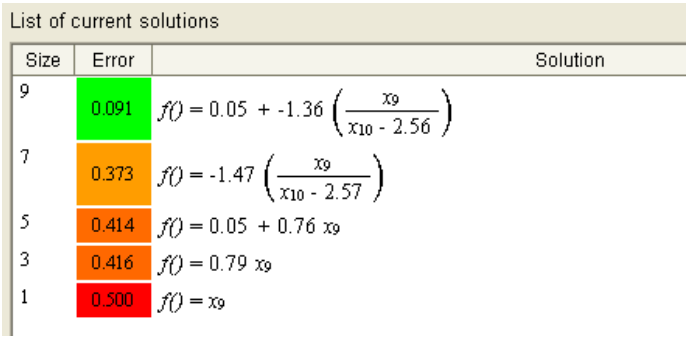


Fig. 5 Representation of a solution set in Eureqa. The first column reports the solution size, the next one the error, and the last one the solution.

The results on the *Credit Approval Data Set* are displayed in the first row of Table 3. The table also reports the results obtained by using a set of alternative models as Decision tree (SEE 5), Multi Criteria Convex Quadratic Programming (MCCQP), Linear Discriminant Analysis (LDA), Fuzzy Type 2 Inference System (FT2IS), and Support Vector Machines (SVMLIGHT and LIBSVM). The best classification performance is offered by our GP approach. Decision trees and a Multi Criteria Convex Quadratic Programming Model are the second choice, as Linear Discriminant Analysis and Fuzzy Type 2 Inference Systems. Support Vector Machines (SVMLIGHT and LIBSVM) appear quite poor in performing this task, and while other works report them most accurate, much more evidence is needed to confirm this.

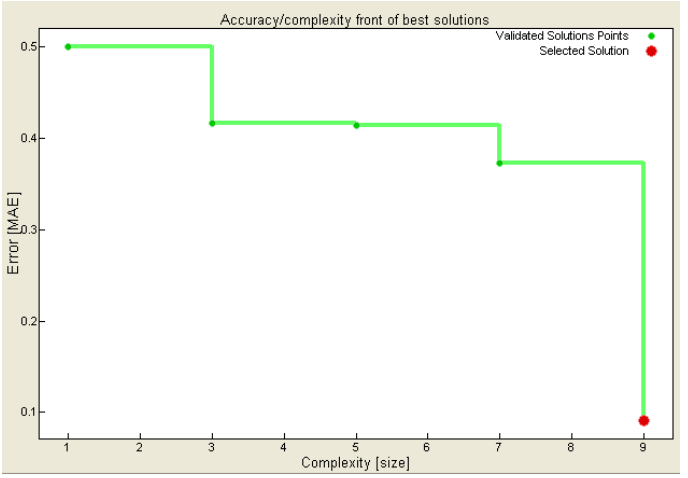


Fig. 6 Error/complexity plot of each solution. The x axis is the formula size, while the y axis is the error.

Table 3 Comparison of credit scoring models on the Credit Approval Data Set

MODEL	OVERALL PERFORMANCE (%)
GP (Genetic Programming)	90.90
SEE 5 (Decision tree)	86.52
MCCQP (Multi Criteria Convex Quadratic Programming Model)	86.38
LDA (Linear Discriminant Analysis)	85.80
FT2IS (Fuzzy Type 2 Inference System)	84.49
SVMLIGHT	44.83
LIBSVM	44.83

5 Conclusions

In recent years, the developments of financial markets have made sophisticated and reliable methods able to model complicated real world applications necessary. In this context, genetic programming has been successfully applied to find approximate solutions for real-world problems which contain various kind of uncertainties.

Namely customer credit is an important concept in the banking industry, which reflects a customer's non-monetary value. Using credit scoring methods, customers can be assigned to different credit levels.

Many classification tools can deal with this task. However, from the point of view of a customer manager, the classification results from the above tools are often too complex and difficult to comprehend. The proposed Genetic programming approach offers an explicit rule representation (a formula) while providing low error rates.

Furthermore, the traditional statistical methods need several a priori assumptions, while GP is a nonparametric tool and suitable for any situations and data sets.

On the basis of the empirical results, we can conclude that GP is more flexible and shows significantly better performance in the credit scoring problem in terms of accuracy and error rate.

Appendix

In this appendix we introduce a very simple example of credit score model representation by a parse tree. We want to distinguish good from bad loan applicants, and we are looking for a model that matches historical data. Let us consider a set of historical data represented in Table 4.

Table 4 A set of historical data

ID	NO OF CHILDREN (NOC)	SALARY (M)	MARITAL STATUS (MS)	OK?
ID-1	2	45000	Married	0
ID-2	0	30000	Single	1
ID-3	1	40000	Divorced	1
…				

A possible model is

IF (NOC = 2) AND (S > 80000) THEN good ELSE bad

In this case and in general the only unknown part is the right formula, hence our search space is the set of formulas. Now a natural representation of a formula is a parse tree, as reported in Figure 7.

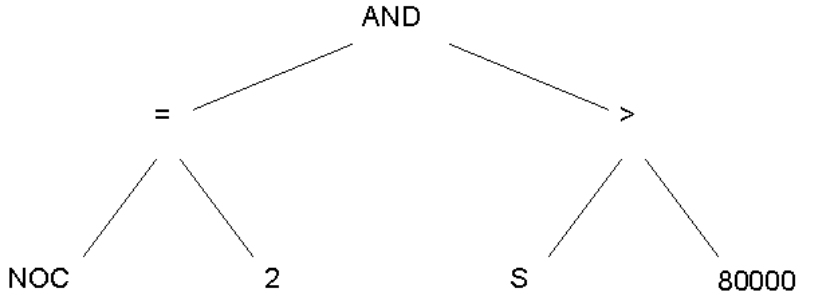


Fig. 7 A parse tree representing the formula f(NOC, M, MS) = (NOC = 2) AND (S > 80000)

References

Cramer, N.L.: A representation for the Adaptive Generation of Simple Sequential Programs. In: Grefenstette, J.J. (ed.) Proceedings of an International Conference on Genetic Algorithms and the Applications, Carnegie Mellon University (1985)

Crook, J.N., Edelman, D.B., Thomas, L.C.: Recent developments in consumer credit risk assessment. European Journal of Operational Research 183, 1447–1465 (2007)

Davidson, J.W., Savic, D.A., Walters, G.A.: Symbolic and numerical regression: Experiments and applications. Information Sciences 150(1/2), 95–117 (2003)

De Stefano, C., Della Cioppa, A., Marcelli, A.: Character preclassification based on genetic programming. Pattern Recognition Letters 23(12), 1439–1448 (2002)

Durand, D.: Risk Elements in Consumer Installment Financing. National Bureau of Economic Research, New York (1941)

Faraoun, K., Boukelif, A.: Genetic Programming Approach for Multicategory Pattern Classification Applied to Network Intrusions Detection. International Journal of Computational Intelligence and Applications (IJCIA) 6(1), 77–99 (2006)

Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2010), http://archive.ics.uci.edu/ml, http://archive.ics.uci.edu/ml/datasets/Credit+Approval

Goldberg, D.E.: Genetic Algorithms in Search, Optimization & Machine Learning. AddisonWesley (1989)

Jabeen, H., Baig, A.R.: Review of Classification Using Genetic Programming. International Journal of Engineering Science and Technology 2(2), 94–103 (2010)

Handzic, M., Tjandrawibawa, F., Yeo, J.: How Neural Networks Can Help Loan Officers to Make Better Informed Application Decisions. Informing Science 6, 97–109 (2003)

Koza, J.R.: Genetic Programming II: Automatic Discovery of Reusable Programs. The MIT Press, MA (1994)

Lahsasna, A., Ainon, R.N., Wah, T.Y.: Credit Scoring Models Using Soft Computing Methods: A Survey. The International Arab Journal of Information Technology 7(2), 115–123 (2010)

Ong, C.S., Huang, J.J., Tzeng, G.H.: Building credit scoring systems using genetic programming. Expert Systems with Applications 29, 41–47 (2005)

Schmidt, M., Lipson, H.: Distilling Free-Form Natural Laws from Experimental Data. Science 324(5923), 81–85 (2009)

Zhang, Y., Bhattacharyya, S.: Genetic programming in classifying large-scale data: an ensemble method. Information Science 163(1/3), 85–101 (2004)

Towards a Human Consistent Analysis of Innovativeness via Linguistic Data Summaries and Their Protoforms

Janusz Kacprzyk[1,2], Sławomir Zadrożny[1], and Tadeusz Baczko[3]

[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
{kacprzyk,zadrozny}@ibspan.waw.pl
[2] Warsaw School of Information Technology
ul. Newelska 6, 01-447 Warsaw, Poland
[3] Institute Institute of Economics, Polish Academy of Sciences
Pl. Defilad 1, 00-901 Warsaw, Poland
tbaczko@inepan.waw.pl

Abstract. We present the application of linguistic data summaries exemplified by, for a personnel database, "most employees are young and well paid" (with some degree of truth) for a human consistent verbalization of data analysis and data mining results in the context of the assessment and evaluation of innovativeness of companies. We present the linguistic summaries in the perspective of Zadeh's protoforms (prototypical forms), and their derivation as an interactive process through a fuzzy querying interface. We show that a relevant class of linguistic summaries that may be of use for our purposes can be obtained by using association rules mining. We show some results of linguistic summaries for innovativeness assessment and evaluation of SME (small to medium) companies in Poland taking into account both quantitative and qualitative attributes.

1 Introduction

In the present, highly competitive world one of key issues that determine the standing and prosperity of a country is innovativeness of companies, branches, industries, etc. All over the world governments put much emphasis on innovations and try to devise a national innovation system, as well as corresponding systems for the monitoring, evaluation and supporting of innovativeness and innovations.

In this paper, which is an extension of our former paper (Baczko, Kacprzyk and Zadrożny, 2010), we are concerned with innovativeness at the national level in the perspective of, for instance, (Archibugi, Howells and Michie, 1999), many papers in (Llerena and Matt, 2004), (Malerba and Brusoni 2007) or (Malerba and Cantner, 2007) but taking into account some specifics of Poland – cf. Baczko (2007a; 2007b; 2008; 2009a). For a similar perspective and solutions in the

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 91–107.
springerlink.com

regional context cf. papers in Baraczyk, Cook and Heidenreich (1996) or Howells (2005), or in a sectorial context cf. Malerba (2004).

Basically, the work is based on the use of some public indicators of innovativeness developed specifically for this purpose, along the lines of the Frascati Manual (OECD, 2003) and the Oslo Manual (OECD, 2005), and to international statistical standards according to which these indicators can be easily adjusted to specific requirements of different countries and regions. By using publicly available indicators a comparison of innovativeness can be easier and more objective, with a lower risk assessment cost (Baczko, 2007a; 2007b). This can help rank companies, branches, etc. with respect to their innovativeness.

This work is based on corporate data from 2004-2006 on selected Polish companies which have been gathered via questionnaires containing both quantitative and qualitative data as well as public statistics, patents granted, stock data, firms reports, data concerning signed contracts of EU firms and experts judgments. Due to the very specifics of the problem considered, the use of some non-standard data analysis techniques was proposed which can capture characteristic features of companies and provide the results in a very human consistent form to facilitate further analysis and a broad dissemination of results. Basically, we propose the use a non-standard technique for the analysis of data via verbalization through so-called *linguistic summaries of data* we have been developing for.

In Section 2 we present some details on the proposed system for innovativeness evaluation. In Section 3 we briefly present the concept of a linguistic summary of data, in a broader perspective of natural language related data processing and generation. In Section 4 we show some examples of linguistic summaries of data concerning innovativeness, and then some conclusions.

2 Assessment and Evaluation of Innovativeness

The concept of an innovativeness assessment and evaluation of the innovation system is based on many results obtained within economics, inspired presumably by Schumpeterian approach (McCraw, 2007) which emphasizes a crucial role of companies, entrepreneurs and workers in the development and implementation of innovativeness. The system should attain several goals like making possible an in-depth study of the structure of business expenditures for research and development and innovativeness which is clearly an important element of the profile of an innovative company, and also an ultimate purpose of the project: to identify factors that have impact on the level of innovativeness of a company. This should both help manage with problems of insufficient innovativeness and undertake proper actions. An important task is also to stimulate companies to disclose their business expenditures for research and development, and then to increase these expenditures. Closely related to the former is also the optimization of the allocation of public and private expenditures in science, technology and innovations which should be undertaken, and this clearly requires tools to identify most promising companies, branches etc. This all is a prerequisite for strategic planning and control of the innovation process in the economy.

The approach presented in this paper has been applied in the system for evaluating the innovativeness of companies based on individual integrated indicators developed in 2005 by the Institute of Economics, Polish Academy of Sciences based on the collection of data from the companies obtained via electronic questionnaires. The evaluation performed using quantitative data has been supplemented by a qualitative evaluation by experts. This approach yields results that rank individual innovative companies and single out the most important characteristics of the innovativeness process to identify types of an innovative behavior, reference models, and characteristics of market innovativeness, process innovativeness, innovation related expenditures, patents obtained, contracts, etc. This research was continued and extended with an additional questionnaire developed by the Institute of Economics, Polish Academy of Sciences in cooperation with the European University Viadrina in Frankurt (Oder), Germany (Baczko, 2008), and using Dunn and Bradstreet Co.'s balance sheet analysis.

The project has shown the viability of developing criteria of innovativeness evaluation on the microeconomic level using indicators developed by representatives of companies, financial institutions, research centers, and public administration. The research concerned the innovativeness of large, medium-size, small, and micro companies, and involved over 2500 companies using individual integrated indicators based on micro data and experts evaluations. The indicators and rankings were made publicly available.

3 Verbalization of Data Analysis Results via Linguistic Summaries

In this paper, due to the specifics of both the problem considered and data, we use for a human consistent presentation of data analysis results linguistic (data) summaries which will now be briefly outlined followed by a presentation of their derivation.

3.1 Lingustic Data Symmaries

In the basic approach to linguistic summarization by Yager (1982), notably in its constructive and implementable form by Kacprzyk and Yager (2001), and Kacprzyk, Yager and Zadrożny (2000), implemented in Kacprzyk and Zadrożny (2001a,c,d, 2002, 2005), we have:

- V is a quality (attribute) of interest, e.g. salary in a database of workers,
- $Y = \{y_1, \ldots, y_n\}$ is a set of objects (records) that manifest quality V, e.g. the set of workers; hence $V(y_i)$ are values of quality V for object y_i,
- $D = \{V(y_1), \ldots, V(y_n)\}$ is a set of data (the "database" on question)

A *linguistic summary* of a data set (data base) is composed of:

- a summarizer S (e.g. young),
- a quantity in agreement Q (e.g. most),
- truth T - e.g. 0.7,
- a qualifier K (optionally), i.e. another linguistic term (e.g. well-earning), determining a fuzzy subset of Y.

as, e.g., "T(*most* of employees are *young*)=0.7". The truth T may be meant more generally as, e.g., validity.

The summarizer S (and qualifier K) is assumed to be a linguistic expression such as, e.g., "young", semantically represented as a fuzzy set in {1, 2, ..., 90}. One can easily extend this to a confluence of attribute values as, e.g, "*young* and *well paid*", and even to non-trivial, *human-consistent* summarizers (concepts) as, e.g.: *productive* workers, involving complicated *combinations of attributes*, e.g.: a hierarchy (not all attributes are of the same importance), ANDing and/or ORing, taking k out of n, *most*, etc. of the attribute values.

The calculation of truth (validity) of the linguistic summary is equated with the calculation of the truth value (from [0,1]) of a linguistically quantified statement (e.g., "*most* of the employees are *young*") that can be done using Zadeh's (1983) calculus of linguistically quantified propositions (Zadeh and Kacprzyk, 1999) or Yager's (1988) OWA operators (cf. Yager and Kacprzyk, 1997); for a survey, see also Liu and Kerre (1998).

Thus, basically linguistic summary is a linguistically quantified proposition, written

$$Qy\text{'s are } S \tag{1}$$

where Q is a (relative) linguistic quantifier (e.g., most), $Y = \{y\}$ is a set of objects (e.g., employees), and S is some property (e.g., well paid), as, e.g., in "most of the employees are well-paid".

By adding another property K, we obtain

$$QKy\text{'s are } S \tag{2}$$

e.g., "most (Q) of the highly qualified (K) employees (y's) are well paid (S)".

We seek the truth values of such statements which are equal to: if S and K are fuzzy sets in Y, and a (proportional, nondecreasing) Q is a fuzzy set in [0,1] exemplified by

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8 \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8 \\ 0 & \text{for } x \leq 0.3 \end{cases} \tag{3}$$

then, due to Zadeh (1983)

$$\text{truth}(Qy's \text{ are } S) = \mu_Q[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mu_S(y_i)] \tag{4}$$

$$\text{truth}(QKy's \text{ are } S) = \mu_Q[\textstyle\sum_{i=1}^{n}(\mu_K(y_i) \wedge \mu_S(y_i))/\sum_{i=1}^{n}\mu_K(y_i)] \tag{5}$$

One can use other criteria for the evaluation of the linguistic summaries, cf. Kacprzyk and Yager (2001), and Kacprzyk, Yager and Zadrożny (2000), like degrees of: imprecision, covering, and appropriateness, and the length of a summary. For more measures, see Kacprzyk, Wilbik and Zadrożny (2008, 2010).

The above linguistic summaries can be extended to the dynamic context, notably to the linguistic summarization of time series as proposed in a series of our papers Kacprzyk, Wilbik and Zadrożny (2008, 2010) in which the analysis of how trends concerning some numerical attributes evolve over time, how long some types of behavior last, how rapid changes are, etc. is provided. However, this will be considered in a next paper.

3.2 Generation of Linguistic Data Summaries

The problem is therefore how to generate the best summaries . The brute force solution, i.e. an exhaustive search, is clearly not a scalable choice, and some implicit enumeration type schemes should be used, possibly with some other tools.

We follow a general Kacprzyk and Zadrożny's (1998) interactive approach for the definition of elements of an intended linguistic summary via a graphical user interface of a fuzzy querying add-on that stems from their earlier papers on the use of fuzzy logic in querying databases (Kacprzyk and Ziółkowski, 1986; Kacprzyk, Zadrożny and Ziółkowski, 1989) via imprecise requests which led to Kacprzyk and Zadrożny's FQUERY for Access package, an add-in to Microsoft Acceess® that makes it possible to use fuzzy linguistic terms in database queries. Notably the following terms, which form a dictionary of the system, are available:

- fuzzy values as by *low* in "profitability is *low*",
- fuzzy relations as by *much greater than* in "income is *much greater than* spending", and
- linguistic quantifiers as by *most* in "*most* conditions have to be met".

The elicitation (definition) of fuzzy sets corresponding to the particular fuzzy values is usually done by using an interface with the user(s) who provides responses to some questions.

Linguistic quantifiers are key elements of both the linguistic data summaries considered here and in the fuzzy querying interface implemented through FQUERY for Access. They are defined in Zadeh's (1983) sense, as fuzzy sets on [0, 1]. They may be interpreted either using Zadeh's (1983) approach or via the OWA operators (Yager, 1988; or Yager and Kacprzyk, 1997).

The matching degree, $md(\cdot,\cdot)$, for the query "Q of N conditions are satisfied" for record t is equal to

$$md(Q,\text{condition}_i,\text{t}) = \mu_Q[\tau(\sum_i md(\text{condition}_i,\text{t}))] \tag{6}$$

and, if we add different importance values to the particular conditions, then the aggregation formula is equivalent to (5), and then

$$md(QK,\text{condition}_1,\text{t}) = \\ = \mu_Q[\tau(\sum_i (md(\text{condition}_1,\text{t}) \wedge \mu_K(\text{condition}_1))/ \\ /\sum_i \mu_K(\text{condition}_1))] \tag{7}$$

In FQUERY for Access fuzzy queries are transformed into syntactically correct queries in Microsoft Access which is done by using the parameters. The query set by the user is then automatically transformed and then run as a native query of Microsoft Access, cf. Kacprzyk and Zadrożny(1995, 1999, 2010) and Zadrożny and Kacprzyk (1995).

In our setting, the fuzzy queries directly correspond to the linguistic summaries, and therefore the derivation of a linguistic summary may proceed interactively within the fuzzy querying system making it possible to interactively define the summarizers (indication of attributes and their combinations). The possible queries are:

- *simple* as, e.g., "salary is *high*"
- *compound* as, e.g., "salary is *low* AND age is *old*"
- *compound with quantifier*, as, e.g., "*most* of {salary is *high*, age is *young*, ..., training is *well above average*}".

We use a "natural" granulation of the set of possible values by using some reasonable number (7±2!) of values like: very low, low, medium, high, very high, and also "comprehensible" and intuitively appealing quantifiers as: most, almost all, ..., etc.

The derivation of a linguistic summary proceeds therefore as follows:

- the user formulates a set of linguistic summaries of interest (relevance) using the fuzzy querying add-on described above,
- the system retrieves records from the database and calculates the validity of each summary by matching the records with the summaries, and
- a most appropriate linguistic summary is chosen.

As shown by Kacprzyk and Zadrożny (2005, 2010), the concept of a protoform in the sense of Zadeh (2002) is highly relevant here. Basically, a *protoform* is defined as an abstract prototype of a linguistically quantified proposition, that is, given as (1) and (2). These are the most abstract protoforms. Their less abstract forms are obtained via instantiation of their particular elements, i.e., qunatifer, summarizer and qualifier. In order to describe the process, it is convenient to consider the summarizer (and the qualifier) as an abstract fuzzy logic statement "*X* IS *A*", where *X* is a placeholder for an attribute of objects in *Y* and *A* is a placeholder for a fuzzy set (linguistic term) determining its (fuzzy) value as, for instance, "age IS young", "salary IS low", and also "salary IS *A*". Two former summarizers are fully instantiated, while the latter still contains an abstract form of the attribute value (*A*). Thus, the protoforms may be seen to form a natural hierarchy (tree) a

root of which is (2), the leaves are actual linguistic summaries with all elements (quantifiers, summarizers and qualifiers) fully instantiated and the intermediate nodes are partly instantiated linguistic summaries with some abstract elements. Zadeh's protoforms may conveniently be used as a fundamental element of the user interface in that the user selects a protoform of a linguistic summary and then the system instantiates the selected protoform in all possible ways, replacing abstract symbols denoting its elements with chosen fuzzy values and linguistic quantifiers stored in a dictionary of linguistic terms.

What concerns what we assume about the linguistic summaries sought, there are two extremes: (a) we assume a totally abstract protoform, or (b) we assume that all elements of a protoform are given on the lowest level of abstraction as specific linguistic terms. In case (a) data summarization is extremely time consuming but may produce interesting, unexpected results while in case (b) the user has to guess a good candidate for a summary but the evaluation is simple, equivalent to the answering of a (fuzzy) query, it is equivalent to ad hoc queries. This may be shown in Table 1 in which 5 basic types of linguistic summaries are shown (cf. Kacprzyk and Zadrożny, 2005, 2010); $S^{structure}$ denotes that the attributes and the connection of predicates referring to them in a summary are known, while S^{value} denotes the values of the attributes sought.

Table 1 A taxonomy of linguistic summaries

Type	Given	Sought	Remarks
1	S	Q	Simple summaries through ad-hoc queries
2	$S\ K$	Q	Conditional summaries through ad-hoc queries
3	$Q\ S^{structure}$	S^{value}	Simple value oriented summaries
4	$Q\ S^{structure}\ K$	S^{value}	Conditional value oriented summaries
5	Nothing	$S\ K\ Q$	General fuzzy rules

Type 1 summaries may be easily obtained by a simple extension of fuzzy querying. The user has to construct a query, a candidate summary, and it has to be determined what is the fraction of rows matching this query and what linguistic quantifier best denotes this fraction. A Type 2 summary is a straightforward extension of Type 1. Type 3 summaries require much more effort as their primary goal is to determine typical or exceptional -depending on the quantifier- values of an attribute. A Type 4 summary is meant to find typical (exceptional) values for some, possibly fuzzy, subset of rows (defined by the qualifier K). Computationally, Type 5 summaries represent the most general form considered by us: the fuzzy rules describing dependencies between specific values of particular attributes. The summaries of Type 1 and 3 have been implemented as an extension to Kacprzyk and Zadrożny's (1995,1999,2001b) FQUERY for Access. Two approaches to Type 5 summaries generation have been proposed. First, a subset of such summaries may be obtained by analogy with association rules concept and employing

their efficient algorithms (cf. Kacprzyk and Zadrożny, 2005 – 2012). Second, genetic algorithms may be used to search the space of summaries (cf. George and Srikant, 1996).

The derivation of linguistic summaries by *association rules* mining (Agrawal and Srikant, 1994) seems to be a promising approach due to a relatively wide availability of software. We will briefly show here an example of how to derive quite general a type of linguistic data summaries using an intrinsic similarity of Type 5 summaries and association rules.

An (basic) association rule is written as:

$$A_1 \wedge A_2 \wedge ... \wedge A_n \rightarrow A_{n+1} \tag{8}$$

and states that if in a database row all the attributes from $\{A_1, A_2,.., A_n\}$ take on value 1, then also attribute A_{n+1} is expected to take on value 1.

A row in a database (table) *supports* a set of attributes $\{A_i\}_{i \in I}$ if all attributes from the set take on value 1. The two main quality measures for the association rule (**Błąd! Nie zdefiniowano zakładki.**) are employed: the *support* which is the fraction of the number of rows supporting the set of attributes $\{A_i\}$, $i \in \{1,..,n+1\}$ in a database (table), and the *confidence* which is the fraction of rows supporting $\{A_i\}$, $i \in \{1,..,n+1\}$ among all rows supporting $\{A_i\}$, $i \in \{1,..,n\}$. While the support determines a statistical significance of a rule, the confidence measures its strength. Usually, we are interested in rules having values of the support above some minimal threshold and a high value of the confidence.

Many algorithms for finding all association rules possessing a required support measure were devised, see, e.g. (Agrawal and Srikant, 1994; Borgelt and Kruse, 2002).

Among many extensions of the above basic form of an association rule, we can mention: (a) the right-hand side, like the left-hand side, may contain a conjunction of the attributes, (b) many-valued scalar values and their hierarchies may be used, (c) numerical, real-valued attributes may be used leading to the *quantitative association rules*, and (d) some constraints may be imposed on combinations of attributes in rules. Therefore, we can rewrite (**Błąd! Nie zdefiniowano zakładki.**) as:

$$A_1=a_1 \wedge A_2=a_2 \wedge ... \wedge A_n=a_n \rightarrow A_{n+1}=a_{n+1} \wedge ... \wedge A_{n+m}=a_{n+m} \tag{1}$$

The association rules may naturally be interpreted as a special type of a linguistic summaries in which the antecedent and consequent of (**Błąd! Nie zdefiniowano zakładki.**) correspond to the qualifier K and summarizer S of (**Błąd! Nie zdefiniowano zakładki.**), respectively. The summarizer S is assumed to be a formula, atomic or complex. Clearly, the structure of the qualifier and the summarizer is somehow limited, this simplicity can be beneficial for the derivation of efficient algorithms for rule generation.

In our previous works (cf. Kacprzyk and Zadrożny, 2001a,d, 2012) we implemented the mining of linguistic summaries corresponding to the association rule (**Błąd! Nie zdefiniowano zakładki.**) within FQUERY for Access. First, we generalized (**Błąd! Nie zdefiniowano zakładki.**) to:

$$A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n \rightarrow A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m} \quad (10)$$

that is, *fuzzy values* f_i instead of crisp values can be used leading to a *fuzzy association rule*.

Then, first, we enriched the structure of an atomic condition A_i IS f_i, (an *item* meant in the terminology of the association rules) to:

$$A_i \text{ IS } (f_{j1} \vee ... \vee f_{jk}) \quad (11)$$

where f_{jl} are some fuzzy values defined over the domain of A_i. Second, we assumed the use of a flexible aggregation operator in the summarizer and/or qualifier formula as:

$$Q \text{ of } (A_1 \text{ IS } f_1, A_2 \text{ IS } f_2, \quad , A_n \text{ IS } f_n) \quad (2)$$

For mining the association rules, we use our FQUERY for Access. There is an apparent benefit of mining (fuzzy) quantitative association rules in such an environment. Namely, the quantitative association rules usually require discretization of the attributes obtained via a partition of a domain into a number of intervals. Then, each interval is treated as an additional binary attribute and then many known algorithms for the generation of classical association rules may be employed. If a fuzzy querying system is employed, the discretization is obtained "for free" as the linguistic terms from the dictionary are readily available for that purpose. They may be well designed by fuzzy querying interface developers or, even better, crafted by the users themselves which provides for their immediate and high interpretability by the user.

Our implementation of association rules mining is based on the Agrawal and Srikant's (1994) AprioriTID algorithm (cf. Borgelt and Kruse, 2002) which works in two steps: (1) find *frequent itemsets*, and (2) produce rules from each itemset. We will focus on the first step which is more difficult.

An itemset is a conjunction of items (**Błąd! Nie zdefiniowano zakładki.**) or (**Błąd! Nie zdefiniowano zakładki.**). A row in the database (table) *supports* an itemset if the corresponding conjunction "is true" (the degree of satisfaction exceeds some threshold) for this row. An itemset containing k items is called a k-itemset. The algorithm starts with the evaluation of 1-itemsets. These itemsets which are not supported by sufficient number (*minsup*) of rows are deleted. Previously, we assumed the 1-itemsets only as in (**Błąd! Nie zdefiniowano zakładki.**). To implement the items such as (**Błąd! Nie zdefiniowano zakładki.**) and (**Błąd! Nie zdefiniowano zakładki.**) we have to extend this step. First, only the "regular" 1-itemsets as in (**Błąd! Nie zdefiniowano zakładki.**) are counted, i.e., a full scan of the database (table) is done and the frequency of appearance of all items is calculated. Then, the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**) are constructed but only such f_{ij} are taken into account that have the support greater than some value (a parameter of the method, in addition to minsup and minconf) higher than 0 and less than minsup. For example, if a regular 1-itemset "salary IS high" gets a very low support, then we will construct neither "salary IS medium or high" nor "salary IS low or high" 1-itemsets. This helps reduce the time and memory complexity of the algorithm.

Such a reduction is even more important while implementing the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**). Basically, we should take into account all subsets of the regular 1-itemsets and all possible quantifiers Q. This would be computationally intractable and in fact require a recursive use of AprioriTID in the first step. Thus, we limit ourselves to just one, fixed quantifier. Moreover, for obvious reasons, we take into account only such subsets of regular items that: all refer to different attributes, and there is enough number of them to make quantification meaningful. Thus, we will, e.g., neither construct a 1-itemset of the form "*most* (*salary IS high*, *salary IS low*,…)" nor "*most* (*salary IS high*, *age IS high*).

When the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**) are constructed we calculate their support. We assume that no row supports two different fuzzy values for the same attribute. Then, the support for A_i IS ($f_{j1} \vee ... \vee f_{jk}$) is just the sum of supports for A_i IS f_{jl}, l=1,…,k, calculated earlier. Now the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**) are constructed. They may use both the regular 1-itemsets as well as the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**), e.g., "Most (A_1 IS ($f_{11} \vee ... \vee f_{1k}$), A_2 *IS* f_2,…)" are allowed. The support for the 1-itemsets of type (**Błąd! Nie zdefiniowano zakładki.**) is then calculated. However, due to the use of AprioriTID another full scan of the database is not needed. During the first scan we have recorded in some data structures the IDs of rows supporting the particular regular 1-itemsets and now it is enough to operate on these structures. Then, the algorithm proceeds as usual (Agrawal and Srikant, 1994) generating and evaluating the k-itemsets for k=2,3,…. We only need to guarantee that no itemset produced twice refers to the same attribute, e.g., the 2-itemset "*salary IS high AND salary IS medium*" has to be excluded. Finally, all frequent itemsets found are taken into account when producing association rules of the confidence of at least the required value (minconf).

We deal with the real valued attributes so that for each such an attribute and each fuzzy value we introduce a new item which may be treated as binary, i.e., appearing in a row or not. In this respect, practically only a limited number of fuzzy values per attribute (say 3) leads to computationally tractable mining tasks.

A brief outline of the algorithm for the mining of linguistic summaries via extended fuzzy association rules may be presented as follows:

Step 1. *Selection of the attributes and fuzzy values*

The user chooses the attributes to be used, i.e. builds a query referring to the attributes to be taken into account (this is done via "the navigation" of protoforms hierarchy, as described in Section III). Then, the user initiates the data summarization process, sets the parameters (minsup, minconf, minimam support, …) and the system automatically performs the rest of the steps.

Step 2. *Construction of the item*s

For each pair – of the selected attributes and fuzzy values - the system creates an item, as described earlier.

Step 3. *Forming the data set and starting external application for fuzzy linguistic rules mining*

The items constructed are numbered. Then, the data set is produced describing each row with numbers of items supported by it. The calculations proceed by the

fuzzy querying module. When the data set is ready, an external application is started with this data set given on input.

Step 4. *Calculation of the support for the regular 1-itemsets*

A module reads the input data set and immediately calculates support for regular 1-itemsets. It also records for each 1-itemset numbers (IDs) of rows supporting it.

Step 5. *Construction of the 1-itemsets of type (***Błąd! Nie zdefiniowano zakładki.***) and calculation of their support*

Only the regular 1-itemsets of the support higher than a user-specified threshold are taken into account. The number of 1-itemsets of this type produced for a given attribute depends on the number of fuzzy values defined for it. The support is obtained by summing up the support of the constituent regular 1-itemsets. All new 1-itemsets are numbered.

Step 6. *Pruning of the set of 1-itemsets*

All itemsets with the support lower than the *support threshold* (minsup) are discarded. Additionally, also itemsets with the support higher than another threshold, an *item omit threshold*, are discarded since the items present in almost all records contribute nothing interesting to the rules produced.

Step 7. *Construction of the 1-itemsets of type (***Błąd! Nie zdefiniowano zakładki.***), calculation of their support and pruning*

Both the regular and 1-itemsets added in Step 5 are considered; we refer to them jointly as simple 1-itemsets. The 1-itemsets constructed are identified with lists of the constituent simple 1-itemsets. The lists are ordered lexicographically which makes the process of generation more efficient. The support is computed for the itemsets generated and those below the minsup threshold are discarded.

All itemsets produced so far and passing the pruning constitute the collection of 1-itemsets.

SET k = 2

Step 8. *Generate the k-itemsets*

They are generated from the frequent (k-1)-itemsets as in AprioriTID. Pairs of the frequent (k-1) itemsets of the form $A_1 \wedge A_2 \wedge ... \wedge A_{k-1}$ and $B_1 \wedge B_2 \wedge ... \wedge B_{k-1}$, where $A_i = B_i$ for i=1,..,k-2, are sought. Then, a new k-itemset of the form $A_1 \wedge A_2 \wedge ... \wedge A_{k-1} \wedge B_{k-1}$ is generated. In the original algorithm, the rules generated in such a way are additionally tested and possibly eliminated before Step 7. On the other hand, we add another k-itemset generation limitation, namely the items A_{k-1} and B_{k-1} have to correspond to different original attributes. This is obvious if the items A_{k-1} and B_{k-1} are regular. Otherwise, by identifying an item of type (**Błąd! Nie zdefiniowano zakładki.**) or (**Błąd! Nie zdefiniowano zakładki.**) with a list (set) of attributes referred to within it, the intersection of these sets is to be empty.

Step 9. *Calculate the support for all the k-itemsets*

The calculation is based on the recorded numbers (ID's) of rows supporting the particular (k-1)-itemsets. The similar data on the supporting rows is produced for the k-itemsets.

Step 10. *Pruning of the set of k-itemsets (as in Step 6)*

As a result we obtain the frequent k-itemsets.

IF the set of k-itemsets is void THEN GOTO Step 11.

SET k = k + 1; GOTO Step 8.
Step 11. *Generate rules from the frequent l-itemsets, l=1,..,k-1.*
Step 12. *Display the results*

The number of the rules produced is usually huge. Some counter-measures have to be undertaken, notably some aggressive pruning schemes.

4 Results of Implementation

In experiments we have employed a subset of the attributes shown in Table **Error! Bookmark not defined.** which describe innovative companies under consideration. The selection of attributes was motivated by the availability of their values for most of the companies considered.

The set of linguistic terms considered (a dictionary) sets the *granularity level* at which the raw data are analyzed, and we have assumed for simplicity the set of values of each attribute to include three linguistic terms: *low*, *medium* and *high*. The definition of linguistic terms is supported by FQUERY for Access, i.e., it makes it possible to define membership functions of fuzzy sets in the domains of particular attributes which are then used to model these terms. We used the linguistic quantifier "*most*". due to its intuitive appeal and relevance in the so called usuality qualification in computing with words (cf. Zadeh and Kacprzyk, 1999). The set of transformed data has been processed by AprioriTID (Agrawal and Srikant, 1994), and the association rules mining has been done by using an implementation of the Apriori algorithm by Christian Borgelt, cf. http://www.borgelt.net/apriori.html.

Table 1 Attributes of companies used for the linguistic summarization

Attribute name	Description
totalAssets2006	Total assets of the company in 2006
totalAssets2005	Total assets of the company in 2005
totalAssets2004	Total assets of the company in 2004
equityCapital2006	Equity of the company in 2006
equityCapital2005	Equity of the company in 2005
equityCapital2004	Equity of the company in 2004
netSales2006	Net sales of the company in 2006
netSales2005	Net sales of the company in 2005
netSales2004	Net sales of the company in 2004
grossProfit2006	Gross profit of the company in 2006
grossProfit2005	Gross profit of the company in 2005
grossProfit2004	Gross profit of the company in 2004
netProfit2006	Net profit of the company in 2006
netProfit2005	Net profit of the company in 2005
netProfit2004	Net profit of the company in 2004
salesDynamics20062005	Sales dynamics in 2005-2006

Table 1 (*continued*)

salesDynamics20052004	Sales dynamics in 2004-2005
ROA2005	Return on assets (ROA) of the company in 2005
ROADynamics20062005	ROA dynamics in 2005-2006
PointsRTDTotal2006	Total points (given by experts) related to RTD related activities in 2006
PointsRTDTotal2005	Total points (given by experts) related to RTD related activities in 2005
PointsRTDTotal20052006	Total points (given by experts) related to RTD related activities in 2005-2006
PointsPatentsTotal2006	Total points (given by experts) related to patents obtained in 2006
PointsTotal	Total points given by experts

In our experiments we obtained a lot of very interesting linguistic summaries. Some of them concern the general standing of the companies analyzed. These may be exemplified by:

> *Most* companies having *high* net revenues from sales and equivalent in 2004 had *high* total assets in 2004

which may be of interest in its own and may be useful to carry out the analysis of innovativeness.

Another class of interesting summaries shows some dependencies between particular assessments of experts evaluating innovativeness of the companies, for instance:

> *Most* of the companies having *at least a few* points for their RTD related activities in 2006 had also some points for that in 2005

which indicates some persistency in the RTD related activities of the companies. The above summary also suggests that the number of those which get those points was not growing in the period analyzed.

The same regularity is not directly exhibited in the opposite direction (i.e., with 2005 and 2006 compared in the reversed order), but we obtain:

> A *majority* of companies having some points related to patents registered in 2006 AND some points for their RTD related activities in 2005 had also some points for RTD related activities in 2006

Thus, in general, companies active in RTD in 2005 do not necessarily continue to do so in 2006 but those which got some patents in 2006 usually also had RTD related activities in 2006.

Due to space limitations we can show just a few of the summaries obtained. In particular, we do not show very interesting, though somehow specific linguistic summaries in the dynamic context which concern the time evolution of the innovativeness related indicators.

5 Concluding Remarks

We have presented the use of a powerful method of linguistic data summarization to an extremely important problem of how to assess and evaluate innovativeness of companies; the results can be easily extended to an analogous analysis for industries, branches, regions, countries, etc.

The use of linguistic summaries, i.e. a natural and comprehensible verbalization of data analysis results, implies an extraordinary human consistency. This is very important in view of a necessity to closely collaborate with domain experts in microeconomics, regional economics, management, etc. who need not be familiar with traditional, not human consistent enough means of data analysis and mining.

Our results have been found very valuable, notably in the analysis of innovativeness of SMEs (small to medium companies), and it was for instance found that a vital role in this respect play their networking capabilities in the field of human resources, intellectual property and ability to get international contracts based on the EU framework programs. Moreover, importance of the regional dimension was found (Baczko, 2008). An interesting by product was that the results obtained have been found valuable for groups carrying out technological Foresight projects while selecting experts for the Delphi studies.

As for future research direction, first, an extension of the analysis to the dynamic case, i.e. by using linguistic summaries of time series (cf. Kacprzyk, Wilbik and Zadrożny, 2008, 2010) seems to be promising. Second, the use of a modern approach to the generation of linguistic summaries via tools and techniques of natural language generation as suggested by Kacprzyk and Zadrożny (2010) seems to be viable and highly promising. Third, an extension towards the use of broadly perceived knowledge-based related techniques can be of relevance. The ontologies may play an important role, making it possible to formalize the domain knowledge concerning the innovation system which can help experts in their qualitative assessment of values of some, in particular qualitative, indicators and to provide a common conceptual ground to operate in a group.

References

Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. of the 20th Int. Conf. on Very Large Databases, Santiago, Chile, pp. 487–499. Morgan Kaufmann, San Mateo (1994)

Archibugi, D., Howells, J., Michie, J.: Innovation systems in a global economy. Technology Analysis and Strategic Management 11, 527–539 (1999)

Baczko, T., et al.: The Report on Innovativeness of Polish Economy in 2007. Polish Academy of Sciences, Warsaw (2007a) (in Polish)

Baczko, T.: Integrated micro indicators of innovativeness-new market and public policy institutional solution. In: Jakubowska, P., Kukliński, A., Żuber, P. (eds.) The Future of European Regions, Warsaw: Ministry of Regional Development, pp. 326–335 (2007b)

Baczko, T., et al.: Standortbedingungen in Ostdeutschland und Polen aus Sicht der Unternehmen. Wochenbericht des DIW Berlin 9, 91–97 (2008)

Baczko, T., et al.: The Report on Innovativeness of Polish Economy in 2008. Polish Academy of Sciences, Warsaw (2009a) (in Polish)

Baczko, T. (ed.): The Future Science and Technology and Innovation In-dicators and the Challenges Implied. Polish Academy of Sciences, Warsaw (2009b)

Baczko, T., Kacprzyk, J., Zadrożny, S.: Towards Knowledge Driven Individual Integrated Indicators of Innovativeness. In: Józefczyk, J., Orski, D. (eds.) Knowledge-Based Intelligent System Advancements: Systemic and Cybernetic Approaches, pp. 129–140. IGI Global, Hershey

Baraczyk, H., Cook, P., Heidenreich, R. (eds.): Regional Innovation Systems. University of London Press, London (1996)

Belton, V., Stewart, T.J.: Multiple Criteria Decision Making. Kluwer, Dordrecht (2001)

Biggiero, L., Laise, D.: Outranking methods. Choosing and Evaluating Technology Policy: a Multicriteria Approach. Science and Public Policy 30, 13–23 (2003)

Borgelt, C., Kruse, R.: Induction of Association Rules: Apriori Implementation. In: 15th Conf. on Comp. Statistics, Berlin, Germany, pp. 395–400. Physica Verlag, Heidelberg (2002)

George, R., Srikanth, R.: Data summarization using genetic algorithms and fuzzy logic. In: Herrera, F., Verdegay, J.L. (eds.) Genetic Algorithms and Soft Computing, pp. 599–611. Physica-Verlag, Heidelberg (1996)

Howells, J.: Innovation and regional economic development: A matter of perspective? Research Policy 34, 1220–1234 (2005)

Kacprzyk, J., Pasi, G., Vojtaš, P., Zadrożny, S.: Fuzzy querying: issues and perspective. Kybernetika 36, 605–616 (2000)

Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems 159, 1485–1499 (2008)

Kacprzyk, J., Wilbik, A., Zadrożny, S.: An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. International Journal of Intelligent Systems 25, 411–439 (2010)

Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. International Journal of General Systems 30, 133–154 (2001)

Kacprzyk, J., Yager, R.R., Zadrożny, S.: A fuzzy logic based approach to linguistic summaries of databases. International Journal of Applied Mathematics and Computer Science 10, 813–834 (2000)

Kacprzyk, J., Zadrożny, S.: FQUERY for Access: Fuzzy Querying for a Windows-Based DBMS. In: Bosc, P., Kacprzyk, J. (eds.) Fuzziness in Database Management Systems, pp. 415–433. Physica-Verlag, Heidelberg (1995)

Kacprzyk, J., Zadrożny, S.: Data mining via linguistic summaries of data: an interactive approach. In: Yamakawa, T., Matsumoto, G. (eds.) Methodologies for the Conception, Design and Application of Soft Computing - Proceedings of IIZUKA 1998, Iizuka, Japan, pp. 668–671 (1998)

Kacprzyk, J., Zadrożny, S.: The paradigm of computing with words in intelligent database querying. In: Zadeh, L.A., Kacprzyk, J. (eds.) Computing with Words in Information/Intelligent Systems, Part 2. Foundations, pp. 382–398. Physica-Verlag (Springer-Verlag), Heidelberg and New York (1999)

Kacprzyk, J., Zadrożny, S.: On a fuzzy querying and data mining interface. Kybernetika 36, 657–670 (2000)

Kacprzyk, J., Zadrożny, S.: On linguistic approaches in flexible que-rying and mining of association rules. In: Larsen, H.L., Kacprzyk, J., Zadrożny, S., Andreasen, T., Christiansen, H. (eds.) Flexible Query Answering Systems. Recent Advances, pp. 475–484. Springer, Heidelberg (2001a)

Kacprzyk, J., Zadrożny, S.: Computing with words in intelligent database querying: standalone and Internet-based applications. Information Sciences 34, 71–109 (2001b)

Kacprzyk, J., Zadrożny, S.: Data mining via linguistic summaries of databases: an interactive approach. In: Ding, L. (ed.) A New Paradigm of Knowledge Engineering by Soft Computing, pp. 325–345. World Scientific, Singapore (2001c)

Kacprzyk, J., Zadrożny, S.: Fuzzy linguistic summaries via association rules. In: Kandel, A., Last, M., Bunke, H. (eds.) Data Mining and Computational Intelligence, pp. 115–139. Physica-Verlag, Springer, Heidelberg and New York (2001d)

Kacprzyk, J., Zadrożny, S.: Protoforms of linguistic data summaries: towards more general natural-language-based data mining tools. In: Abraham, A., Ruiz del Solar, J., Koeppen, M. (eds.) Soft Computing Systems, pp. 417–425. IOS Press, Amsterdam (2002)

Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Information Sciences 173(4), 281–304 (2005)

Kacprzyk, J., Zadrożny, S.: Protoforms of Linguistic Database Sum-maries as a Human Consistent Tool for Using Natural Language in Data Mining. International Journal of Software Science and Computational Intelligence 1(1), 100–111 (2009)

Kacprzyk, J., Zadrożny, S.: Modern data-driven decision support systems: the role of computing with words and computational linguistics. International Journals of General Systems 39(4), 379–393

Kacprzyk, J., Zadrożny, S.: Computing With Words Is an Implement-able Paradigm: Fuzzy Queries, Linguistic Data Summaries, and Natural-Language Generation. IEEE Transactions on Fuzzy Systems 18, 461–472 (2010)

Kacprzyk, J., Zadrożny, S.: Derivation of Linguistic Summaries Is Inherently Difficult: Can Association Rule Mining Help? In: Borgelt, C., Gil, M.Á., Sousa, J.M.C., Verleysen, M. (eds.) Towards Advanced Data Analysis. STUDFUZZ, vol. 285, pp. 291–304. Springer, Heidelberg (2012)

Kacprzyk, J., Zadrożny, S., Ziółkowski, A.: FQUERY III+: a 'human consistent' database querying system based on fuzzy logic with linguistic quantifiers. Information Systems 6, 443–453 (1989)

Kacprzyk, J., Ziółkowski, A.: Database queries with fuzzy linguistic quantifiers. IEEE Transactions on Systems, Man and Cybernetics, SMC-16, 474–479 (1986)

Llerena, P., Matt, M. (eds.): Innovation Policy in a Knowledge-Based Economy Theory and Practice. Springer, Heidelberg (2004)

Malerba, F.: Sectoral Systems of Innovation. Cambridge University Press, Cambridge (2004)

Malerba, F., Brusoni, S.: Perspectives on Innovation. Cambridge University Press, Cambridge (2007)

Malerba, F., Cantner, U.: Innovation, Industrial Dynamics and Structural Transformation. Springer, Heidelberg (2007)

McCraw, T.K.: Prophet of Innovation: Joseph Schumpeter and Creative Destruction. Harvard University Press (2007)

oECD, The Measurement of Scientific and Technological Activities. Frascati Manual, Proposed Standard Practice for Surveys on Research and Experimental Development. OECD Publishing (2003)

OECD, Oslo Manual: Guidelines for Collecting and Interpreting Innova-tion Data, 3rd edn. OECD Publishing (2005)

Yager, R.R.: A new approach to the summarization of data. Information Sciences 28, 69–86 (1982)

Yager, R.R.: On ordered weighted avaraging operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics, SMC-18, 183–190 (1988)

Yager, R.R., Kacprzyk, J. (eds.): The Ordered Weighted Averaging Operators: Theory and Applications. Kluwer, Boston (1997)

Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. Computers and Mathematics with Applications 9, 149–184 (1983)

Zadeh, L.A.: Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. IEEE Transaction on Systems, Man and Cybernetics, SMC-15, 754–763 (1985)

Zadeh, L.A.: A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform. In: BISC Seminar. University of California, Berkeley (2002)

Zadeh, L.A., Kacprzyk, J. (eds.): Computing with Words in Information/Intelligent Systems. 1. Foundations, 2. Applications. Physica-Verlag, Springer, Heidelberg and New York (1999)

Zadrożny, S., Kacprzyk, J.: Summarizing the contents of Web server logs: a fuzzy linguistic approach. In: Proc. 2007 IEEE Conf. on Fuzzy Systems, London, UK, pp. 1860–1865 (2007)

An Investigation of Computational Complexity of the Method of Symbolic Images

Giulia Rotundo

Department of Economics and Management, University of Tuscia, Viterbo, Italia
giulia.rotundo@uniroma1.it

Abstract. The widespread presence of maps in discrete dynamical models needs the usage of efficient algorithms for their investigation. The method of symbolic images is more efficient than exhaustive numerical simulation of trajectories because it transforms a map into a graph through a discretization of the state space, so it opens the way to the usage of graph algorithms and it provides a unified framework for the detection of system features. In this framework, a modification of the algorithm described by Osipenko et al. is proposed and its efficiency is analyzed. Issues on the convergence of the method raise when the dynamical system is described by a not-Lipschitzian nonlinear map in the plane. As case study it is shown the application of the method on an evolutionary model of boundedly rational consumer characterized by the presence of a denominator that can vanish.

1 Introduction

The analysis of the global structure of maps is at the core of the analysis of dynamical systems. The limits of pure numerical investigation are most due to computational time and numerical precision. Appropriate methods have been developed to overcome the impossibility to simulate all the trajectories of the system, and the two main general approaches divide methods into direct and indirect ones. The direct methods focus on the development of numerical techniques that find particular dynamical structures. Direct approaches most reduce the study of the system to a root finding problem, they can provide very exact solutions of a problem, but most used methods, like as the Newton-based ones, do not converge globally, so they are only useful for local analysis.

The indirect methods are closely related to the simulation of trajectories. Usually, the system's dynamics are approximated by the computation of parts of a limited number of trajectories. These methods have several advantages: they are always applicable, fast to compute, and no a priori knowledge about a system is needed. The drawback is that they can capture only those structures of a system which possess a stable long-term behaviour. Both approaches are essential tools for the numerical investigation, but for the global analysis of dynamical systems numerical methods are needed which do not require any a priori knowledge about the position of a structure and are also capable to reveal those structures which do

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 109–126.
springerlink.com

not possess stable long-term behavior. A successful approach to achieve this aim is given by methods which are based on the discretization of the phase space. Such methods can be considered as a combination of direct and indirect approaches.

The investigation of a dynamical system using symbolic images basically consists in a discretization of the state space and in the construction of a directed graph which represents the structure of the state space. The graph is the symbolic image of the system, and it is an approximation of the system flow. The method provides a unified framework for the detection of stable points as well as unstable limit cycles. The mathematical theory was presented in a series of papers coworked by Osipenko, and this work refers to the most recent ones (Osipenko and S. Campbell (1999), Avrutin et al. (2006), Fundiger (2006)) considering vector fields, and extending the method to maps.

The usage of such a graph speeds up the computational time, since, once it is constructed, the investigation of the map is a matter of graph analysis. In fact, each strongly connected part represents an invariant set of the flow. This allows to locate return trajectories of all types without any restrictions concerning their stability. The tuning of parameters plays a key role in the efficiency of the algorithms. This paper considers the non-Lipschitzian maps and the modification of the proof of convergence with respect to the Lipschitzian case shown in Avrutin et al. (2006), Fundiger (2006). An analysis of their computational efficiency is also given in this work. The extension to non-Lipschitzian maps was suggested by their occurrence in economic evolutionary models. In particular, as a case study it is shown the application of the above method to an evolutionary model of boundedly rational consumers already studied in Bischi and Tramontana (2005). The model is characterized by a two-dimensional map with a denominator that can vanish. The map describes the evolution of the consumption of the agent (x) and his/her preference for the good (α). In particular, the discrepancy between expected and realized utility causes an updating in the consumption decisions of the agent. Moreover, the preference for the good is assumed to be endogenously dependent on the consumption, differently from the neoclassical assumption of exogenous preferences. From a mathematical point of view, the map generates interesting dynamic phenomena related to the recently introduced concepts of *focal points*, *lobes* and *prefocal curves* (Bischi et al. 1999, 2003). In this work, the analysis is limited to the analysis of attractors, and the method is used on not-Lipschian functions.

The sections are organized as follows. The next sections sums up the symbolic image approach, and the problem of parameters tuning. Section 4 shows the economic model and its relevant features. Section 5 completes the analysis performs the application of the method to the map that rises from the economic problem under examination.

2 Symbolic Images

The method of symbolic images associates a map to a graph through a proper discretization procedure. Such approach has the advantage to offer a unified framework for the analysis of the global structure of vectors. Hence, this opens the way to the study of maps through graph theory analysis. In particular, the

investigation of invariant sets of the map corresponds to the detection of strongly connected components of the graph (Avrutin et al. (2006)). The task of usability of the method needs the development of efficient algorithms. Therefore, the theory of dynamical systems and graphs must be completed by the knowledge of data structures for efficient computer programming. The discretization allows to use the same procedure both for stable and unstable limit cycles. In the next subsection the theoretical background, and the implementation framework are reported. The result is that allowing some limited memory storage, the complexity may be lowered with respect to the results shown in (Avrutin et al, 2006). Then, sufficient conditions are given for the convergence of the method when the map is not Lipschitzian.

2.1 Theoretical Background

Let us consider a dynamical system which is generated by a vector field and defined on a C^{∞} -smooth manifold M :

$$\vec{f} : M \to M$$

Most theorems also assume that M is a compact set in R^n, because a covering could be not finite on not-compact sets, and $\vec{f}$ is an homeomorphism. The theoretical work of Osipenko (1994), which are the basis for numerical methods under examination, are restricted to dynamical systems generated by homeomorphisms. However, the numerical methods shown in this work are generally also applicable for noninvertible maps, describing a dynamical that is discrete in time, and considering a finite covering of the domain M by closed sets.

$$C = \{M_1, \ldots, M_n \mid \bigcup_{i=1}^{n} M_i = M\}$$

Each set $\forall i \in \{1, \ldots, n\}, M_i \subset M$ is named box of the covering C with respect to the investigation area M .

For each box M_i its image $\vec{f}(M_i)$ is defined as follows: $\vec{f}(M_i) = \left\{ \vec{y} \mid \vec{y} = \vec{f}(x), \vec{x} \in M_i \right\}$.

Then, for each box M_i, the covering C_i of its image is defined as $\vec{f}(M_i)$. Selecting all the boxes that have a not-empty intersection with $\vec{f}(M_i)$:

$$C_i = \{M_j \mid M_j \bigcap \vec{f}(M_i) \neq 0\}.$$

The correspondence between M_i and C_i is used to build a directed graph $G = (V, E)$ where V is the set of vertices, and E is the set of edges, i.e. a set of

ordered couples of vertices. Each box M_i corresponds to a vertex $c_i \in V$. A directed edge exists between c_i and c_j iff $M_j \in C_i$, so the number of vertices in V is equal to the number of vectors in C, i.e. $|V| = |C|$.

Definition 1. $G = (V, E)$ built as described above is named symbolic image of $\vec{f}$ with respect to the covering C.

The symbolic image can be considered as a finite approximation of $\vec{f}$, depending on the covering C.

In Avrutin et al (2006), the result is given for subsets as follows.

Definition 2. Let $L \subseteq V$ be a set of vertices which generates a subgraph G(L). Then the set of vertices $Ex(L) = \{c_i \in L \mid \exists c_j \in V \setminus L : (c_i, c_j) \in E\}$ is called the exit of L.

Remark 1. Let $D \subseteq M$ correspond to a subset of vertices $L \subseteq V$ in the symbolic image G of the map $\vec{f}$. If D is an attractor for $\vec{f}$, then $Ex(L) = 0$.

Further definitions are needed:

Definition 3. The diameter $\delta(M_i)$ of a box $M_i \in C$ is given by

$$\delta(M_i) = \max\left\{\left\|\vec{x} - \vec{y}\right\| \mid \vec{x}, \vec{y} \in M(i)\right\}.$$

Definition 4. The diameter of the partition is given by the largest diameter of its boxes: $\delta(M) = \max\{\delta(M_i) \mid M_i \in C\}$.

The definition of path in G is the usual one, and it is reported here for an easy reference.

Definition 5. A path in G is a countable sequence $\omega = \{c_{i_1}, \ldots, c_{i_m}\}$ of vertices c_{i_k} such that $\forall k \ (c_{i_k}, c_{i_{k+1}}) \in E$.

Remark 2. The previous definition is not considering the size of the set $\omega = \{c_{i_1}, \ldots, c_{i_m}\}$, that could also be not finite.

Definition 6. The length of the path $\omega = \{c_{i_1}, \ldots, c_{i_m}\}$ is the cardinality of ω, $|\omega|$, which coincides with m if ω is finite.

Definition 7. A path in G is named p-periodic if $\exists p \mid \forall k \in Z \ \ c_{i_k} \rightarrow c_{i_{k+p}}$ e $\forall p' < p \ \ c_{i_k} \neq c_{i_{k+p'}}$

Definition 8. $\forall \varepsilon > 0$, a sequence infinite in both directions $\{\vec{x}_k\}, k \in Z$ is said to be an ε-orbit or pseudo-orbit of $\vec{f}$ if $\forall k \in Z$ the distance between the image

$\vec{f}(\vec{x}_k)$ of the point $\vec{x}_k$ and the next point $\vec{x}_{k+1}$ is less then ε: $\left|\vec{f}(\vec{x}_k) - \vec{x}_{k+1}\right| < \varepsilon$

Definition 9. A pseudo-orbit $\left\{\vec{x}_k\right\}_{k\in Z}$ is p-periodic if $\forall k \in Z \ \vec{x}_k = \vec{x}_{k+p}$ and $\forall p' < p \ \ \vec{x}_k \neq \vec{x}_{k+p'}$.

Remark 3. There is a natural correspondence between a p-periodic orbit $\{\vec{x}_1, \ldots, \vec{x}_p\}$ and a path $\{c_{i_1}, \ldots c_{i_p}\}$ on its symbolic image.

In practice, a real orbit is seldom known exactly and, in fact, usually ε-orbits are found for ε sufficiently small positive.

Definition 10. A point $\vec{x}_k, k \in Z$ is called chain recurrent if, $\forall \varepsilon > 0$, a periodic ε-orbit passes through it.

Definition 11. A set of points chain-recurrent is named set chain-recurrent and denoted by Q. A chain-recurrent set is invariant, closed, and it contains trajectories periodic, homoclinic, and other singular trajectories. A chain-recurrent point may become periodic under a small C^0-perturbation of $\vec{f}$.

Definition 12. $c \in V$ is recurrent if there is a periodic path passing through it. $c_i \in V$ and $c_j \in V$ are called equivalent if there is a periodic path containing both them.

Denote the subset of recurrent vertices in G as $RV(G)$. The set $RV(G)$ decomposes itself into classes H_1, H_2, … of equivalent recurrent vertices. Each of them is a strongly connected component of G. The boxes M_i corresponding to vertices in $RV(G)$ are a neighbourhood of the chain recurrent set. The detection of this neighbourhood is the basic task for every symbolic image calculation.

Discretization is natural in numerical investigation, due to the finite size memory of any computer and physical device. A perfect correspondence between the evaluation of the function and its symbolic image may be got through a very small diameter of the partition.

Remark 4. Let A be the adjacency matrix that represents G. If the diameter of the partition is equal to the maximum precision used for floating point variables, then G corresponds to a tabular representation of $\vec{f}$, therefore the results of the evaluation of paths in G are equal to the results obtained through the evaluation of trajectories of $\vec{f}$.

Of course, such representation of $\vec{f}$ has no computational benefits, and a more coarse grouping of points into boxes is of main interest. The target is to find out the best covering through boxes, and to refine it through covering with boxes with

decreasing diameter. The basic technique for all the numerical methods shown in this paper is multilevel phase space discretization.

The procedure is described as follows in Avrutin et al. (2006). It is based on an iterated subdivision and selection of the investigation area. Let s be an index for the subdivision level, let $C^s = \{M_1^s, \ldots, M_n^s \mid M_i^s \subset M\}$ indicate a covering for that level, so $\delta(C^s)$ is the largest diameter of the covering C^s.

Let $G^s=(V^s,E^s)$ be its symbolic image. Then a subset of boxes belonging to C^s gets selected for the next subdivision. . Let S^s be such a selected area for level subdivision s. If fixed points are to be estimated, then $L \subseteq V$ should be selected such that $S^s = L^s$, $L^s = \{M_i \mid c_i \in L, Ex(L) = 0\}$. If the chain recurrent set should be approximated, then all the recurrent boxes of C^s get selected. Let $RV(G^s)$ be the set of recurrent vertices of G^s. Applying this technique, a neighbourhood of the chain recurrent set *Q* can be determined. Let us denote such a neighbourhood as $S^s = \left\{M_i \mid c_i \in RV(G^s)\right\}$ where s is the subdivision depth. This neighbourhood is an outer covering of the chain recurrent set.

After the selection, a new set C^{s+1} is built. C^{s+1} is a partition of C^s through the split of the boxes of C^s into smaller boxes. In particular, $M_i^s = \bigcup_k M_k^{s+1}$. The new symbolic image for the new covering is $G^{s+1} = (V^{s+1}, E^{s+1})$. The new vertices $c_{i,k}$ have a natural mapping from V^{s+1} to V^s and every path on G^{s+1} can be transformed on some path on G^s. This means that also the corresponding ε-orbits are mapped from G^{s+1} to G^s.

The subdivision procedure continues till the level that has the desired precision.

The following result holds for $S^s = \left\{M_i \mid c_i \in RV(G^s)\right\}$.

Theorem 1. (Osipenko 1994)

The sequence of sets S^0, S^1, … has the following properties

1. the neighbourhoods S^k are embedded into each other, i.e.: $S^0 \supset S^1 \supset S^2 \supset \cdots Q$
2. if $\delta(C^s) \xrightarrow[s\to\infty]{} 0$, then $\lim_{s\to\infty} S^s = \bigcap_s S^s = Q$

This theorem states that by increasing the subdivision level, the chain recurrent set can be approximated with an arbitrary precision, and that the boxes which have to be selected for further subdivision are the ones corresponding to recurrent vertices. The set of recurrent vertices can be detected through the calculation of the strongly connected components on the graph.

The analysis of the number of boxes that is maintained from a level to the next one plays a relevant role in the study of the efficiency of the method, that is going to be explored in the next sections.

3 Implementation of the Basic Framework and Performance Analysis

The method described in the previous section gives rise to the following algorithms: The first for the detection of attractors, through the selection $S^s = L^s$;the second for the detection of attractors and ε-orbits through the selection of $S^s = RV(G^s)$.

Algorithm 1.

a. inizialization of the area of investigation $S^0 = M^0$

b. iterations on the level of subdivision $s, s \in \{1, s_{\max}\}$:

1. $C^{s-1} = S^{s-1}$ and divided in cells with smaller diameter;
2. $\forall i \mid M_i^{s-1} \in C^{s-1}$, determine C_i^s;
3. construction of G^s;
4. selection of $L^s \subseteq V^s$ such that $Ex(L^s) = 0$;
5. selection of the new area of investigation $S^s = L^s$;
6. test on the termination conditions: $s < s_{\max}$ or $\delta(M^s) < \varepsilon$;
7. the iterations stops whether the termination conditions are satisfied, otherwise the next iteration starts with $s = s + 1$.

Algorithm 2

a. inizialization of the area of investigation $S^0 = M^0$;

b. iterations on the level of subdivision $s, s \in \{1, s_{\max}\}$:

1. $C^{s-1} = S^{s-1}$ and divided in cells with smaller diameter;
2. $\forall i \mid M_i^{s-1} \in C^{s-1}$, determine C_i^s;
3. construction of G^s;
4. selection of $L^s \subseteq V^s$ such that $Ex(L^s) = 0$;
5. selection of the new area of investigation $S^s = RV(G^s)$;
6. test on the termination conditions: $s < s_{\max}$ or $\delta(M^s) < \varepsilon$;
7. the iterations stops whether the termination conditions are satisfied, otherwise the next iteration starts with $s = s + 1$.

The algorithms only differ for step b.5. In Algorithm 1, the choice $S^s = L^s$ may exclude during iterations the boxes containing attractors if the diameter of cells is too high and not sufficiently tight to the attractor.

The outline of the next section follows the steps of the analysis of the performance shown in Avrutin et al (2006), and the results of this paper are compared with theirs.

3.1 Multilevel Space Discretization

A task that is not relevant for the mathematical proof, but that plays a key role for the design of the algorithm and for its efficiency is the shape of boxes.

Avrutin et al. (2006) and Fundiger (2006) discuss several implementation of the discretization of the space into boxes M. They show that if the shape of boxes is rectangular, then it is possible to detect boxes by using either a vector of pointers or a hash function for calculating the box which the point corresponds to. The usage of the vector has the drawback to grow exponentially with $\left|C^s\right|$, and to finish soon the computer memory, so they advice to use hash functions. Here it is used the hash function suggested by Avrutin et al. (2006), that recursively calculates the global positioning of the box, till to the desired level and embeds all those information into a unique number. Their warning about the possible overflow of the maximum length of integer numbers, has been overcome by the software, and this allow to reduce the computational time, as it is shown in the next section. Thus, the complexity of the calculus is $s \cdot d$, since the box number is calculated in $O(1)$ time for each level *s* and dimension *d*. The usage of such a hash function naturally maintains the ordering of cells, since it is automatically got through the calculus. The maintenance of the order allows faster cell detection and numbering.

3.2 Graph Building

The relevant task in building the graph is the creation of the correspondence between M_i and C_i. Following Avrutin et al. (2006) a fixed number k of scan points is used. It is possible to prove the following result.

Proposition 1. By using hash function described above, the correspondence between M_i and C_i requires $O(k \log k)$ in the mean case, and $O(k^2)$ in the worst case.

Proof. Due to the fact that k is fixed and finite, the computer memory will suffice for storing a temporary vector of fixed length k to be used for the calculus of each C_i^s, and to be deleted afterwards. Considering that for each scan point the cell whose image must be calculated, and that the calculation of each arrival cell is

$O(s \cdot d)$, the calculation of the image of the scan points is $O(k \cdot s \cdot d)$. The set of the C_i^s must be cleaned by duplicated elements, i.e. the vector must be sorted, and read once more for deleting the duplicated elements. Remembering that sorting algorithms, like as quicksort, use $O(k \log k)$ in the mean time and $O(k^2)$ in the worst time, the detection of C_i requires $O(s \cdot d \cdot k \log k)$ in the mean case and O(d s k^2) in the worst case.

Considering that the detection of arrival boxes must be done for each box in C^s, the construction of G^{s+1} is $O(s \cdot d \cdot |C^s| \cdot k \log k)$ in the mean case and $O(s \cdot d \cdot |C^s| \cdot k^2)$ in the worst case

For high enough $|C^s|$, the above result is better than the result O($|C^s|$ log $|C^s|$) shown in Avrutin et al. (2006) and Fundiger (2006). Moreover, the dependence from $|C^s|$ is reduced, unless a new dependence of k from $|C^s|$ is introduced.

3.3 Complexity of the Algorithm

3.3.1 Algorithm 1

For the calculation of $S^s = L^s$ a reading of vertices and edges is necessary, so the time complexity $O\left(|V^s| + |E^s|\right)$. Therefore, the completion of each iteration of Algorithm 1 needs $O\left(dsk \log k |V^s| + s\left(|V^s| + |E^s|\right)\right)$ in the mean case and $O\left(dsk^2 |V^s| + s\left(|V^s| + |E^s|\right)\right)$ in the worst case. Since the variable that risks to be most increased is $|V^s|$, the overall complexity is $O\left(|V^s|\right)$.

3.3.2 Algorithm 2

For the calculation of the strongly connected components this section is going to use the Tarjan algorithm, that has time complexity $O\left(|V^s| + |E^s|\right)$. Therefore, the completion of each iteration of Algorithm 2 needs $O\left(dsk \log k |V^s| + s\left(|V^s| + |E^s|\right)\right)$ in the mean case and

$O\left(dsk^2\left|V^s\right|+s\left(\left|V^s\right|+\left|E^s\right|\right)\right)$ in the worst case. Since the variable that risks to be most increased is $|V^s|$, the overall complexity is $O\left(\left|V^s\right|\right)$.

In spite of the similar computational complexity, Algorithms 1 and 2 tackle different tasks. It is worth remarking that the condition b.5 in Algorithm 1 strongly depends on the diameter of cells, and that the different selection of boxes may lead Algorithm 1 to violate the monotonic inclusion of boxes much more than it happens in Algorithm 2, where recursive boxes are always included in the selected covering, and the violation of the monotonicity could only rise from a poor selection of scan points.

3.4 Covering Selection in Case of Absence of Lipschitzianity

The proof of the convergence of the method is based on the monotonicity of the diameter and on the inclusion of cells. Therefore, the selected cells form an outer cover, that is going to shrink on the desired object. However, both Avrutin et al (2006) and Fundinger (2006) stress that the usage of scan point for calculating C^s may be not precise, and scan points could be mapped into boxes not selected at a previous stage. Therefore, they show how to select scan points of C^s so that C^{s+1} is an outer cover of $\bigcup_i \vec{f}(M_i^s)$ in case of Lipschitzian functions. If the Lipschitz condition does not hold, no method is provided. This hash function allow to calculate easily in which box falls the image of the scan point , without considering extra sets since the beginning of the procedure, but being able to add them easily, whether this is needed going through the next levels. This section shows sufficient conditions for Theorem 1 to hold, i.e. it shows which should be S^0 so to maintain the decreasing monotonicity of coverings.

Definition 13. Let us use the notation $\vec{f}^{\,s-s'} = \underbrace{\vec{f}\bullet\vec{f}\bullet\cdots\bullet\vec{f}}_{s-s' \text{ times}}$. Let $h^{-1}(\bullet)$ be the function that maps backward the box M_i^s to $C^{s'}$, $s<s'$, , i.e. $h^{-1}:(\wp(M),s')\to\wp(M)$ such that

$$h^{-1}(M_i^s\ ,s')=\bigcap\left\{M_j^{s'}\,\middle|\,\vec{f}^{\,s-s'}(M_j^{s'})\supset M_i^s\right\}$$

and, on subsets $M^s=\bigcup_i M_i^s$,

$$h^{-1}(M^s,s')=\bigcap\left\{\left\{M_i^{s'}\right\}_i\,\middle|\,\bigcup_i \vec{f}^{\,s-s'}(M_i^{s'})\supset M^s\right\}$$

Remark 5. Due to the design of the hash function, $h^{-1}(M_i)$ can be computed $O(s \cdot d)$.

Proposition 2. Let C^s the covering at level s. Then the results of Theorem 1 hold if the initial selection is $S^0 = h^{-1}(C^s, 0)$, and, for the next levels $S^{s'} = h^{-1}(C^s, s')$.

In fact, choosing this selection the sequence of $S^0 \supset S^1 \supset S^2 \supset \cdots Q$ and the diameter is decreasing, i.e. $\delta(C^s) \xrightarrow[s\to\infty]{} 0$. Q is included in C^s since all the cells with $S^s \subset C^{s'}$, $s' < s$,

Therefore, if the insertion of a cell M_j^{s+1} such that $M_j^{s+1} \not\subset \bigcup_i M_i^s$ is needed during the iterations, Theorem 1. can be proved starting from $\bigcup_i h^{-1}\left(M_i^{s+1}, 0\right)$.

4 The Evolutionary Economic Model

The model starts considering a utility function $U(x, \bar{x})$ that depends of the consumption of a given good x and the rest of the goods (denoted by $\bar{x}$. The price of the good x is equal to p, while the price of x is normalized to 1. Given the income (m) the agent has to choose one of the combinations of goods satisfying the budget constraint:

$$px + \bar{x} = m$$

A rational agent maximizes the utility function $U(x, \bar{x})$ under the budget constraint. The optimal choice is such that:

$$\begin{cases} S = \dfrac{\partial U / \partial x}{\partial U / \partial \bar{x}} = p \\ px + \bar{x} = m \end{cases}$$

The consumer is a boundedly rational one in the sense that is not able to compute the maximizing choice and he/she adjusts the demanded quantity of good following a discrete time rule of thumb given by $x_{t+1} = x_t + \mu\left[S(x_t) - p\right]$

where μ measures the speed of adjustment. The discrepancy between $S(x_{t-1})$ and p (which is 0 for the optimum) is interpreted by the consumer as a

proxy of the distance between the last consumption choice and the rational one. In order to obtain an explicit map a utility function is needed. Let us consider the well known Cobb-Douglas utility function $U(x,\bar{x}) = x^{\alpha}\left(\bar{x}\right)^{1-\alpha}$

That has the nice property that the exponent α can be considered a measure of the preference of the agent for the good x. Under this assumption the adjustment process becomes

$$x_{t+1} = x_t + \mu\left[\frac{\alpha}{1-\alpha}\frac{m}{x_t} - \frac{p}{1-\alpha}\right]$$

This is the way the consumer adjusts consumption for any given value of the preference parameter. A further assumption assumes that the preference are not exogenously given but it changes with the consumption of the good according to the following S-shaped function

$$\alpha_{t+1} = \frac{1}{k_1 + k_2 k_3^{x_t}}$$

under the restrictions: $k_1 > 1$, $k_2 > 0$ and $0 < k_3 < 1$, due to the necessity to maintain the value of α between 0 and 1.

The dynamical system arising from these assumption is

$$\begin{cases} x' = x + \mu\left[\dfrac{\alpha}{1-\alpha}\dfrac{m}{x} - \dfrac{p}{1-\alpha}\right] \\ \alpha' = \dfrac{1}{k_1 + k_2 k_3^{x}} \end{cases} \tag{1}$$

where $'$ denotes the unit time advancement operator.

The steady states of the adaptive model are the fixed points of the above map, i.e. the solutions of the system

$$\begin{cases} \alpha = \dfrac{p}{m}x \\ \alpha = \dfrac{1}{k_1 + k_2 k_3^{x}} \end{cases}$$

The following proposition has been shown in (Bischi and Tramontana (2005)):

Proposition 1. At least one positive fixed point of the map (1) exists for each set of parameters such that $k_1 > 1$, $k_2 > 0$ and $0 < k_3 < 1$. Three positive steady states exist provided that $\frac{m}{p} > -4$ and $k_2^{\min} < k_2 < k_2^{\max}$, where

$$k_2^{\min} = \frac{\frac{2mpk_1}{m - \sqrt{m^2 + \frac{4mpk_1}{\ln k_3}}} - k_1}{k_3^{\frac{m - \sqrt{m^2 + \frac{4mpk_1}{\ln k_3}}}{2mpk_1}}} \; ; \; k_2^{\min} = \frac{\frac{2mpk_1}{m + \sqrt{m^2 + \frac{4mpk_1}{\ln k_3}}} - k_1}{k_3^{\frac{m + \sqrt{m^2 + \frac{4mpk_1}{\ln k_3}}}{2mpk_1}}}$$

Bischi and Tramontana (2005) determine the attractors and their basins. Two sets of parameters are relevant. The first set gives rise to an attractor *A=(4.3602646, 0.0872)*, shown in Fig. 1(a) together with its basin of attraction. Parameters used for drawing the figure are reported in the table here below. With the second set of parameters the steady state denoted by A is a saddle point that generated through a flip bifurcation a stable cycle of period two denoted by $A_2^{(1)}$ and $A_2^{(2)}$. It results that for a set of parameter. Table 1. reports the set of parameters and the figures

Table 1 Parameters of Fig. 1

Fig.	*m*	*P*	μ	k_1	k_2	k_3
1.a	500;	10	0.537	10	57	0.432
1.b	1677	10	0.917	2.4	78	0.9

5 Numerical Results

The modified method proposed in section 3 is used for the solution of model (1). The map has a denominator that vanishes, therefore it is not Lipschitzian. The the symbolic image method is applied on both cases outlined in Fig. 1a and Fig.1b,c.

5.1 First Set of Parameters

Firstly, the case of the set of parameters used for drawing Fig. 1(a) is analyzed, considering both $S^s = Ex(L)$ and $S^s = RV(G^s)$. Fig. 2 shows that Algorithm 1 correctly locates attractor *A*, but monotonicity is often violated. Fig. 3 shows that, in the example, the usage of $C^s = RV(G^s)$ verifies the monotonicity inclusion of selected boxes at each level.

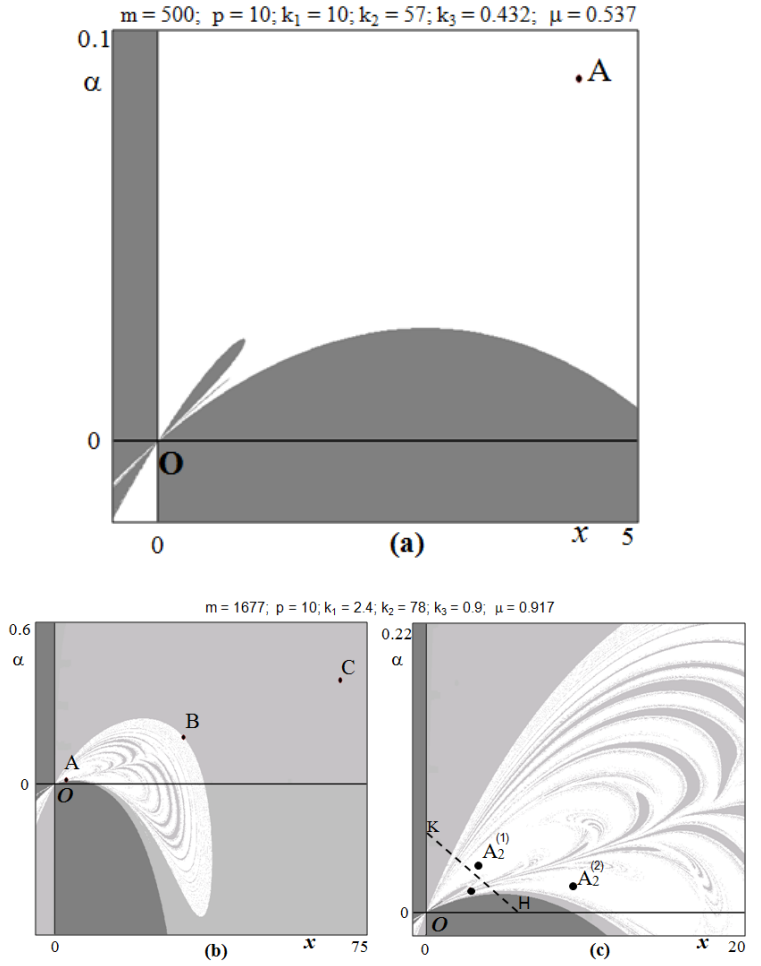


Fig. 1 The two behaviors of the map are shown. (a) A=(4.3602646, 0.0872) (b) A=(2.963 - 0.016), B=(30.703 - 0.183) and C=(68.195 - 0.40665). (c) A flips into a 2-cycle ciclo-2 A2(1) (3.309 - 0.033) A2(2) (9.885 - 0.017).

5.2 Second Set of Parameters

This section reports results used using both Algorithms 1 and 2.

Algorithm 1 correctly locates the attractor C also using a low number of boxes and number of internal scan points, as it can be seen from Fig. 4. Point B, and points A1 and A2 are never discovered whether using $S^s = Ex(L)$. In fact, any box surrounding them does not belong to the set for which Ex(L)=0.

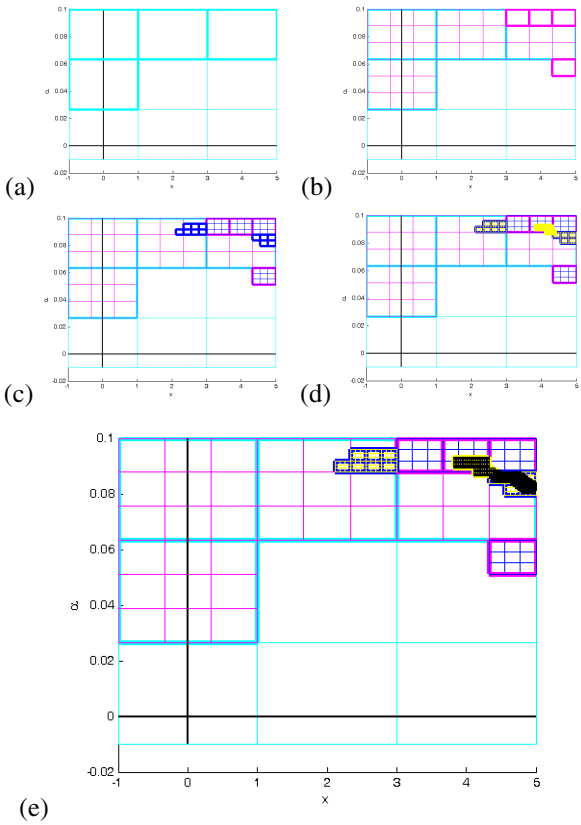


Fig. 2 The same parameters of Fig.1.a were used for the map. $S^s = Ex(L)$. The effect of a low number of internal scan points is shown. There are 3×3 boxes, 4 scan points close to vertices of each box, and 5 levels of boxes division. The initial range is (-1,5) × (-0.01,1), the size of the smallest box is 0.0027× 0.0001. (a) shows the first grid C^0 (light blue boxes) and the S^1 (bold blue boxes). (b) shows also C^1 (light magenta) and S^2 (bold magenta). It is easy to see that the low number of internal scan points violates the monotonic inclusion of S^s, and that the initial partition should have contained an extra box. (c) At level 3 starting from C^2 (light blue boxes inside the magenta ones). The resulting S^3 (bold dark blue bordered cells) is selected, completely outside C^0, but inside the covering C^1. (d) S^4 (yellow) is contained in partition 1, but it is only partially overlapping with partitions at levels 2 and 3 (e) Also the partition at level 5 does not verify the monotonic inclusion of selected cells.

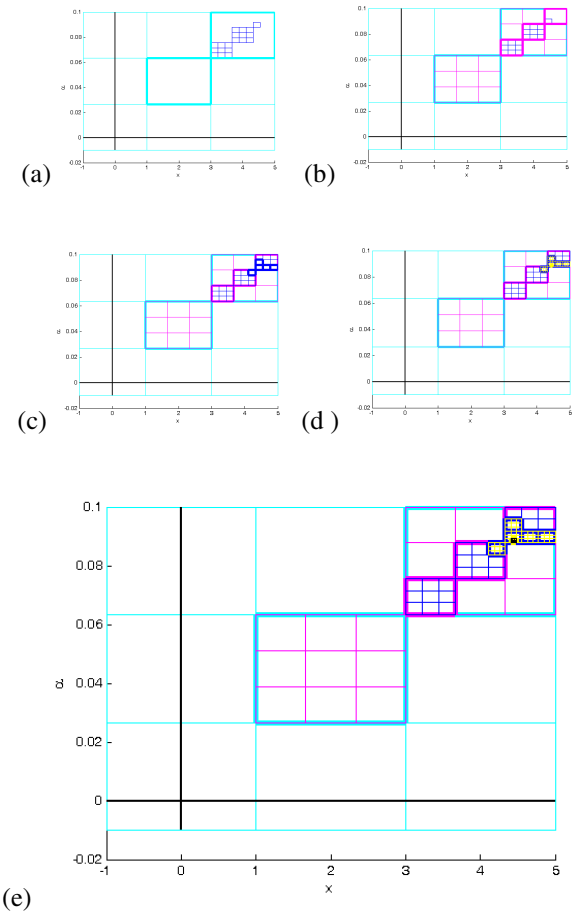


Fig. 3 The same parameters of Fig.1.a were used for the map. $S^s = RV(G^s)$. There are 3×3 boxes, 4 scan points close to vertices of each box, and 5 levels of boxes division. The range is (-1,5)×(-0.01,1), the smallest box size is 0.0027×0.0001. (a) shows C^0 (light cyan lines) and S^1 (bold cyan lines) (b) C^1 (thin magenta boxes) and S^2 (bold magenta boxes). (c) shows C^2 (light blue lines internal to the magenta boxes) and S^3 (bold bluelines). (d) C^3 (thin yellow lines) and S^4 (bold yellow lines). (e) C^5 (black boxes).

Fig. 5 reports the results of Algorithm 2, where $S^s = RV(G^s)$. In this case, C is found for any selection of parameters, but B, A1 and A2 are traced only for either a good enough number of scan points, or a high number of cells. Fig. 5.(a) shows the area located by the first iteration 100×10 boxes, 2 scan points, and fig. 5.(b) 3×3 boxes, 30×30 scan points. Therefore, the selection of parameters for the discretization is relevant for non stable attractors.

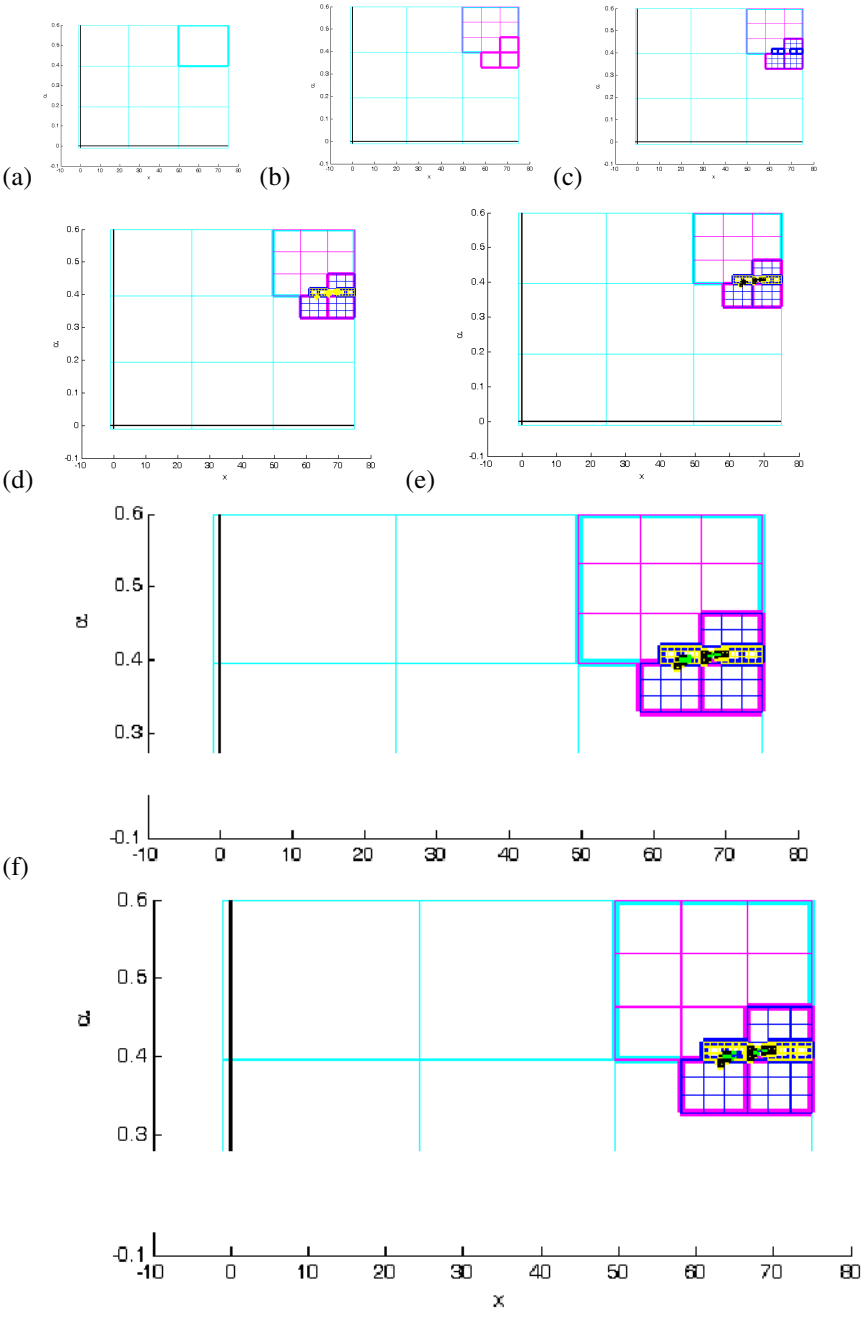


Fig. 4 Colors define the levels of the covering as outlined for the previous figures till level 5. (f) level 6 is defined through green boxes (g) level 6 is defined through blue boxes

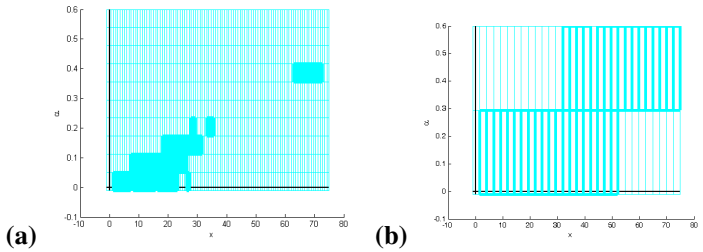


Fig. 5 C^0 (thin light blue cells) and S^1 (bold light blue cells). (a) 100×10 boxes, 2 scan points (b) 3×3 boxes, 30×30 scan points.

6 Conclusions

This work adds insights in the theory and algorithm concerning the method of symbolic images. The method is based on a discretization of maps that allows to investigate maps through graph algorithms. This work shows sufficient conditions for the convergence of the method is the hypotesis of Lipschitzianity does not hold. The complexity of the algorithm and proper data structure that allow to reduce the computational time of the algorithm using a fixed amount of memory are shown. The method is applied to the solution of an economic problem, that is described by a map with a vanishing denominator, and its performance is discussed.

References

[AV06] Avrutin, V., Levi, P., Schanz, M., Fundinger, D., Osipenko, G.S.: Investigation of dynamical systems using symbolic images: efficient implementation and applications. International Journal of Bifurcation and Chaos 16(12), 3451–3496 (2006)

[BGM99] Bischi, G.-I., Gardini, L., Mira, C.: Plane maps with denominator. Part 1: some generic properties. International Journal of bifurcation and Chaos 9(1), 119–153 (1999)

[BGM03] Bischi, G.-I., Gardini, L., Mira, C.: Plane Maps with Denominator. Part II: Non-invertible maps with simple focal points. International Journal of Bifurcation and Chaos 13(8), 2253–2277 (2003)

[BT05] Bischi, G.-I., Tramontana, F.: Basins of attraction in an evolutionary model of boundedly rational consumers. Pure Math. Appl. 16(4), 345–363 (2005)

[FU06] Fundinger, D.: Investigating dynamics by multilevel phase space discretization, unpublished doctoral dissertation, university of Stuttgart (2006)

[OC99] Osipenko, G.S., Campbell, S.: Applied symbolic dynamics: Attractors and filtrations. Discrete and Continous Dynamical Systems 5(1, 2), 43–60 (1999)

Spatial Database Quality and the Potential Uncertainty Sources

Šárka Hošková-Mayerová[1], Václav Talhofer[2],
Alois Hofmann[3], and Petr Kubíček[4]

[1] Department of Mathematics and Physics, University of Defence, Faculty of Military Technology, Kounicova 65, 662 10 Brno, Czech Republic
sarka.hoskova@unob.cz
[2,3] Department of Military Geography and Meteorology, University of Defence, Faculty of Military Technology, Kounicova 65, 662 10 Brno, Czech Republic
{vaclav.talhofer,alois.hofmann}@unob.cz
[4] Department of Geography, Masaryk University, Faculty of Science, Kotlarska 2, 611 37 Brno, Czech Republic
kubicek@geogr.muni.cz

Abstract. One of the most significant features of geo-information systems is the creation of geospatial analyses. The analyses are based on fundamental geospatial data which model the landscape in the certain territory of interest. The analyses themselves are often described by a mathematical apparatus which uses a wide range of branches of mathematics, especially vector analysis, differential geometry, statistics, probability, fuzzy logic, etc. The classical mathematical description of analysis is clear and precisely defined. Complex geospatial analyses, however, work above geospatial data that do not have to be homogeneous from the point of view of quality. With respect to the capacity and technological possibilities of the data supplier, the input data can have different level of geometric and thematic accuracy, their thematic attributes can remain unfulfilled or the data can be obsolete to a certain extent. Also the location of objects (e.g. forested areas, soil, water area, etc.) can be uncertain concerning the impossibility to define them accurately (e.g. areas of different soil kinds are blended together) or change with time (coast line of watercourse that changes depending on rainfall). The stated imprecision and uncertainty then influence the result of the complete geospatial analysis. This influence gets bigger with the number of input data objects.

The aim of the presented paper is to find a relation between the quality of input data and the reliability of the geospatial analysis result. The authors approach is based on mathematical models of analyses, models of vagueness and uncertainty, as well as from models for quality evaluation. In the research were used real data from the territory of the Czech Republic - current as well as historical - and specific methods of space evaluation used in decision-making processes of commanders and staff of the army.

Keywords: uncertainty, geospatial analyses reliability, mathematical modelling, models of vagueness and uncertainty.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 127–142.
springerlink.com

1 Modeling of Uncertainty in Geospatial Data

Not even the field of spatial analyses is free of uncertainty. The uncertainty in source data in connection with the simplifying approach of used geospatial models results in spreading of the original errors, or even their increasing. The quality of decision-making based on geospatial analyses (i.e. their results) is significantly influenced by the quality of the input data, the quality of used geospatial models, and the level of uncertainty that spreads or is created during a geospatial analysis. In accordance with Shi (2010) and the above stated approach, the following section is divided into chapters outlining the possibilities of dealing with uncertainty in fields:

- Modeling of uncertainty in geospatial data
  - Modeling of positional uncertainty (coordinate)
  - Modeling of attribute uncertainty (characteristics)
  - Modeling of connected positional and attribute uncertainty
- Modeling of uncertainty in geospatial models
- Modeling of uncertainty in geospatial analyses

Especially conceptual approaches and possibilities with relevant references of chosen work done in the appropriate field are mentioned.

1.1 Modelling of Positional (Coordinate) Uncertainty

Within the scope of GIS, it is possible to define a geographical object on the earth surface as a geospatial feature that is characterized by its position, corresponding attributes and relations with the surrounding objects. The position of geographical object can be represented in different geospatial models - objects, fields, and analytical functions with parameters (Shi 2010). The positional uncertainty of geospatial features is best expressed with the help of object model and that is why the attraction from now on will only be paid to this way of representation of objects. It is also probably the most widespread model used in GIS when the point coordinates are used for the entry of position of geospatial feature.

Positional uncertainty can be set based on the difference between the measured position of geospatial feature and its "real" value in the real world. As a result of quite imperfect recognition of the real world and restrictions due to imperfection of the measuring technology, the positional uncertainty is present practically with all geospatial features. The positional uncertainty then significantly influences the reliability of results and analyses and in a broader sense also the decisions based on the support of geospatial data.

In order do describe the positional uncertainty of a geospatial object, it is useful to classify objects by their primitive parts and by the types of movement they support under uncertainty (Shekhar and Xiong 2008):

- objects that are single points (point objects);
- objects that comprise multiple points whose internal geometry cannot change under uncertainty (rigid objects, lines as well as polygons);

- objects that comprise multiple points whose geospatial relations can vary under uncertainty (deformable objects, lines as well as polygons).

1.2 Methodology of Definition of Probability Layout of Positional Uncertainty

One of the possible approaches of solution of positional uncertainty in geographical objects is to use the theory of probability (Shekhar and Xiong 2008). The proposed solutions differ in their complexity from a simple " epsilon band" (ε), when around every point or line there is a buffer created with the radius ε, up to the estimate with the help of *probability distribution functions* (PDF) for all basic vertexes (points) of the object. For two-dimensional vertex object, the pdf of coordinates *x* and *y* is defined as follows:

$$F_{XY}(x,x) = \text{Prob}(X \le x, Y \le y),$$

where *X* and *Y* are random variables representing the real, however not known, position of the particular point (Fig. 1). The diameter and standard deviation of coordinates *x* and *y* are important parameters of PDF. While the diameter defines the position of the point in every direction, the standard deviation defines its uncertainty.

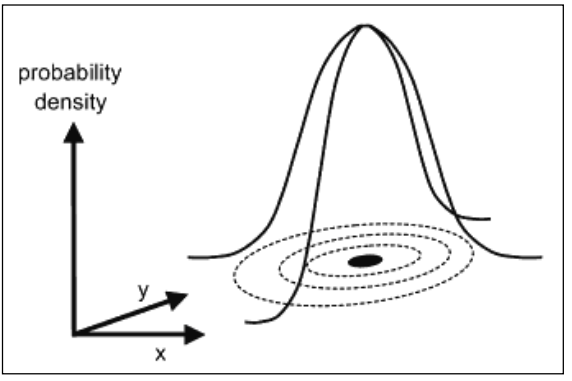


Fig. 1 Sketch of bivariate probability distribution of a point object's position (edited according to Shekhar a Xiong 2008)

1.2.1 Probability Layout for Positional Uncertainty

For solid objects consisting of more connected points it is possible to define shifting and rotation of the object around one referential point. In a Fig. 2a there is shown an example of 4 random projections of the positional vague solid object, whose uncertainty in direction *x* is higher than in direction *y* and for which the rotational uncertainty is quite low. Buildings can be mentioned as a good example.

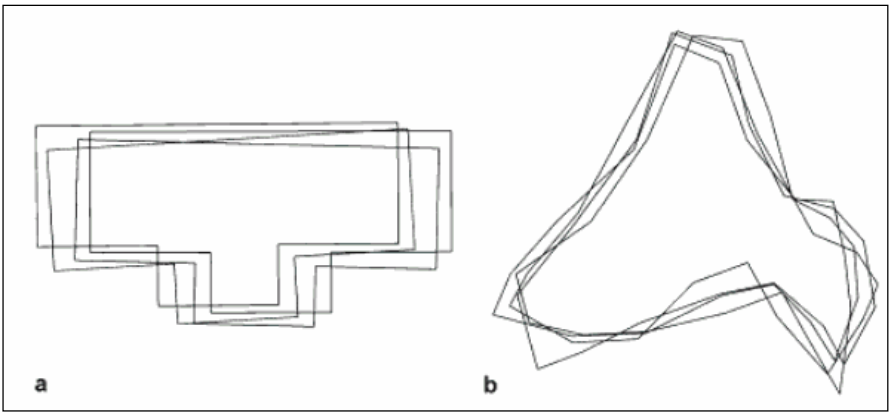


Fig. 2 The demonstration of positional uncertainty modeling for uncertain rigid objects (a) and uncertain deformable objects (b) Edited according to Shekhar a Xiong (2008)

For changing objects (for example level of water surface during flood, atmospheric phenomena, etc.) it is necessary to define connected pdf for all basic points from which the object is made. From this stems the need to model relations, or correlations among vague points. The neighboring points will rarely change independently as the mistakes in measurement, georeferencing or cartographic generalization are often similar in neighboring positions.

Semivariogram is an example of correlational model dependent on the distance of point. In Fig. 2b it is possible to see an example where a strong geospatial relation (correlation) among neighboring points is supposed. The correlation is not complete as 4 done realizations of the object are deformed (Shekhar a Xiong 2008).

1.3 Modeling of Attribute Uncertainty

Attributes can be defined as descriptive characteristics or the essence itself of geographical objects (Shi 2010). An example is the name of the owner, which is the attribute of the cadastral land. Within GIS attributes describe certain characteristics, features or values of geospatial features.

The essence of the real world can be described in two basic ways - with the help of coherent or incoherent representation. Based on the differing essence, the attribute data can also be characterized either as coherent or discreet (categorical). In certain cases also other divisions are used, such as qualitative, quantitative or hierarchical division into other subordinate categories. For the purposes of this text, only simplified division will be used - categorical (e.g. types of usage of earth - get the final number of values from pre-defined list) and coherent (e.g. values of flow rate or temperature - get any number values within the frame of defined interval or domain).

1.3.1 Attribute Uncertainty

Uncertainty of attributes is defined as closeness of attribute value to the real value in the real world, where individual objects are usually of a complex character and where the material of which the object is made is geospatially heterogeneous and the borders of the object are not usually sharply defined (compare section I.). The essence and level of attribute uncertainty is influenced by many factors:

a) Attributes of the geospatial object themselves
b) The way of definition and identification of the attributes
c) Method and technology used for getting or measuring of the attribute value
d) Mathematical grounding of geospatial analyses and model procedures used for attribute data

Attribute uncertainty can consequently influence the quality of the geospatial decision-making. In many GIS applications there are theoretical requirements for accuracy of the attributes even higher than requirements for positional accuracy. It is also necessary to emphasize that attribute uncertainty requires further study as it has attracted less attention than positional uncertainty (Wang et al. 2005).

1.3.2 Attribute Uncertainty and Other Types of Uncertainty

For better understanding of the attribute uncertainty it is necessary to more precisely describe its relation to other parts of uncertainty, especially (Wang et al. 2005):

- Attribute uncertainty and positional uncertainty - both types are mutually interconnected. The foundations of geospatial information are a combination of space-time location and a set of characteristics (attributes), all of which together create sort of covering uncertainty in geospatial data which is then possible to divide to positional, attribute and their mutual combination.
- Attribute uncertainty and quality of data. Quality of geodata can be measured from two points of view - internal and external. The internal view determines correspondence of theoretical specification and obtained geodata. The external view indicates how the specifications correspond to the current and perhaps even potential needs of the users. Attribute uncertainty is one of the basic indicators of geodata quality (Guptil a Morrison 1995)
- Attribute uncertainty and data error. As it was said before, an error in data attributes is a difference between the value recorded in geodatabase and the real value existing in the real world. Attribute uncertainty can be considered to be a sort of extension of attribute error in data when except the error; the uncertainty involves also elements of randomness, unclearness, inaccuracy, interference, etc.

1.3.3 Sources of Attribute Uncertainty

Despite the fact that the basic sources of uncertainty were already mentioned in section I., in relation to the attribute uncertainty according to Wang and coll. (2005) it is possible to specifically mention especially the following sources and causes of creation of uncertainty (Fig. 3):

- *Objective reality* - real world represents a complex system which is made of many inaccurately defined entities. Two of these entities of the same classification class can show different characteristics (e.g. spectral) and on the contrary, two entities showing similar characteristics can be of different classification classes. Traditionally it was assumed that model of the real world saved in geospatial databases is sharply limited, described in a quality way and accurately measured in the form of computer model in GIS. In many cases, however, the used geometric models (point, line, area) do not correspond to the real form and some of the attributes of geofeatures are inaccurate or even not available.
- *Subjective recognition* - as individual entities of the real world are complex and changeable in time, people must choose only the most important geospatial aspects for just description of objective reality. All attribute data are gained with the help of certain theoretical approaches, techniques and methods that implicitly or explicitly specify the required level of abstraction or generalization. The obtained data are only a selection of all available attributes of a feature and describe only a certain, however mostly significant, part of a real range of characteristics of a feature. As a result of this, computer-modeled entities lose certain aspects of real entities and bring uncertainty into the processing.
- *Computer processing*. Data GIS in computer model interpret the real world thanks to the binary code of zeros and ones. Part of uncertainty then certainly arises during computer processing (physical model, logical model, coding of data, data processing, data analysis, optimization of algorithms, etc.), and thus differences between the real and coded value of a certain attribute are created. Individual uncertainty is cumulated and spread in the computer processing, there can also arise new types of uncertainty connected to the used methods of computer processing.
- *Increase of heterogeneity*. Attribute uncertainty becomes even more complex when we connect (integrate) various kinds of attribute data, usually from different sources and with different level of reliability. If such heterogeneous local databases are joined together in a global context, just as it is true with the realization of directions INSPIRE, then there are conflicts and uncertainty with consequence that is so far difficult to describe.

Individual sources of uncertainty and their accumulation are documented in Fig. 3.

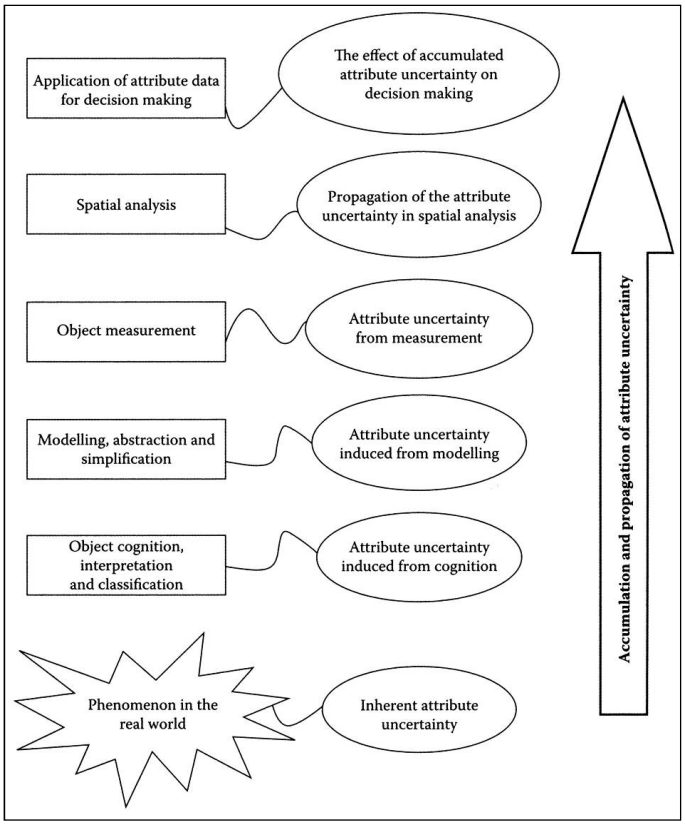


Fig. 3 Individual sources of uncertainty and their accumulation (Shi, 2010)

2 Uncertainty and Quality of Geospatial Data

The system of quality of geospatial data evaluation goes mainly from industry, where quality is considered as a desirable goal to be achieved through management of the production process. Statistical quality control has a relatively long history in manufacturing, where it is used to ensure conformity in products by predicting the performance of manufacturing processes. Quality is more difficult to define for data. Unlike manufactured products, data do not have physical characteristics that allow quality to be easily assessed. Quality is thus a function of intangible properties such as 'completeness' and 'consistency' (Veregin, 1999). The uncertainty of positional information and thematical properties is necessary to consider too.

Many authors dealt with application of common quality definition into evaluation of digital geographic data and information (DGI). Finally, recommendations of International Organisation for Standardisation (ISO), Open Geospatial Consortium (OGC), and Defence Geospatial Information Working Group (DGIWG) were developed from their results. The base quality components

are defined in various sources (DGIWG-103, 2008), (Jacobsson & Giversen, 2007), (Joos, 2006), (Sanderson, Stickler, & Ramage, 2007). In the next table (**Table 1**) there is the list of elements and subelements as described in ISO 19113 resulting from several years' discussion:

Table 1 Elements and subelements of data quality according to ISO 19113

Element of quality	Subelement of quality
Completeness	Commission
	Omission
Logical consistency	Conceptual consistency
	Domain consistency
	Format consistency
	Topological consistency
Positional accuracy	Absolute or external accuracy
	Relative or internal accuracy
	Gridded data position accuracy
Temporal accuracy	Accuracy of a time measurement
	Temporal consistency
	Temporal validity
Thematic accuracy	Classification correctness
	Non-quantitative attribute correctness
	Quantitative attribute accuracy

The elements and subelements of quality, listed in the Table 1, express mainly technical properties. Their quality parameters are usually expressed mathematically (e.g. per cent, standard error, probability of occurrence etc.).

The whole quality model incorporates not only technical components, but also components describing legislative, technology, market and own production (see Fig. 4).

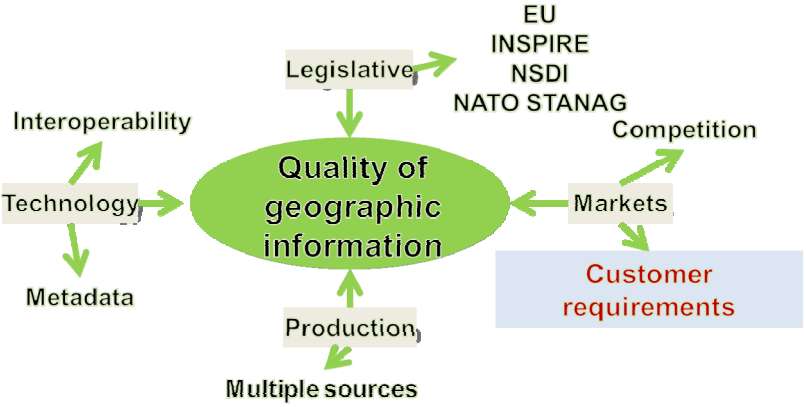


Fig. 4 Components of the quality of geographic information (Jacobsson & Giversen, 2007)

Each product including DGI has to be made for the specific user (or group of users) and only his satisfaction with the product is the final criterion for quality of this product evaluation. Usability as an expression for a product's potential to accomplish the goal of the user is often mentioned term. Usability can be described as results of technical quality parameters evaluation added with textual remarks related to customer requirements. The other approach is to apply some system which enables to combine different possibilities of expression of quality parameters. The application of the Value Analyse Theory (VAT) (Miles, 1989) is one of possibilities.

3 Pilot Study

To verify the impact of data quality and their uncertainty on a result of a geospatial analyse the task of Cross Country Movement (CCM) was chosen as an example. CCM can be solved as a common problem or with consideration of certain types of vehicles (the most frequent or the weakest in the unit, but in case of armed forces usually off road vehicles). The detailed theory of CCM can be found in (Rybanský, 2009).

The solution can offer to the commander not only one possibility (alternative), but the variants from which he can choose according to his intentions and the current situation at the given area.

3.1 Cross Country Movement

The main goal of CCM theory is to evaluate the impact of geographic conditions on a movement of vehicles in a terrain. For the purpose of classification and qualification of geographic factors of CCM, it is necessary to determine:

- particular degrees of CCM
- typology of terrain practicability by kind of military (civilian) vehicles
- geographic factors and features with significant impact on CCM

As a result of the geographic factors impact evaluation we get three known degrees of CCM:

- GO - passable terrain
- SLOW GO - passable terrain with restrictions
- NO GO – impassable terrain

Geographic factors determining CCM and the selection of the access routes are follows:

- gradient of terrain relief and micro relief shapes
- vegetation cover
- soil conditions
- meteorological conditions
- water sheets, water courses

- settlements
- communications
- other natural and manmade objects

Common impact of geographic factors on deceleration of vehicle movement at given section of route can be expressed as follows:

$$v_j = f\left(v_{\max}, C_1, C_2, ..., C_n\right),\ j = 1, \ldots, k \tag{1}$$

where:

- v_j is vehicle speed at *j*- section of vehicle path [kph]
- v_{max} is maximum vehicle speed on the road [kph]
- C_i is coefficient of deceleration
- n is a number of geographic factors effecting at given section of terrain
- k is a number of sections on vehicle path

The impact of given geographic factor can be evaluated as a *coefficient of deceleration 'Ci'* from the scale of 0 to 1. The coefficient of deceleration shows the real (simulated) speed of vehicle v_j in the landscape in the confrontation with the maximum speed of given vehicle v_{max}. The impact of the whole n geographic factors can be expressed as the formula:

$$v_j = v_{\max} C_j, \quad \text{where} \quad C_j = \prod_{i=1}^{n} C_i, \quad n = 1, \ldots, \mathbb{N}, j = 1, ..., k \tag{2}$$

The main coefficients of deceleration are listed in the next table (see **Table 2**).

Table 2 Main coefficients of deceleration

Basic coefficient	Geographic signiffication and impact
C_1	Terrain relief
C_2	Vegetation
C_3	Soil and soil cover
C_4	Weather and climate
C_5	Hydrology
C_6	Buil-up area
C_7	Road network

Each coefficient consists of several coefficients of 2nd grade which are used for more detailed expression of complex impact on vehicle deceleration. For example C_1 is express as:

$$C_1 = C_{11} C_{12},$$

where:

- C_{11} is deceleration coefficient by impact of gradient factor,
- C_{12} is deceleration coefficient by impact of microrelief factor.

The value of given coefficient is evaluated from information of feature position and its thematical properties.

The calculations of coefficient C_i are based on geometrical and thematic characteristic of objects which in compliance with Chapter 1 have certain uncertainty. These uncertainties have an impact on calculations and on final result of analyses as well. For example during the calculation of coefficient C_{11} the following effects occur:

1. Degree of slope is calculated from Digital Terrain Elevation Model (DTED) within pixel size of 20 metres. However, resulting pixel in „cost map" is 5 meters.
2. In particular pixel only the maximal slope is considered regardless its orientation of the future vehicle movement. This cause the uncertainty of longitudinal or transversal tilt determination.
3. In case, the precise value of coefficient c_4 is not known (climate and weather impacts), then the uncertainty of vehicle slipping on the slope due to the surface moisture has to be considered.

Similarly, it is possible to evaluate the impact of uncertainty of all objects characteristics on resulting coefficients C_i due to formula (2).

3.2 Geospatial Database Utility Value Evaluation

The master DGI database is usually utilised as a base for geospatial data analyses. The master database can be very detailed, carefully maintained and used in many applications. But nobody can suppose that this database contains all information he could need.

The task of CCM solution could require more information that is available in the master database. Geographer-analyst has to consider which information and in what quality can he obtain from master database. E.g. for mentioned C_{12} coefficient it is necessary to select all micro relief obstacles in the area of interest (road and railway embankments, excavations, terrain steps, trenches etc.). Further he has to find out all their properties and their accuracy or count how many characteristics are missing, In other words he must describe both positional and attribute uncertainty in a broader sense.

3.3 CCM of TATRA 815 Analyses

The common army vehicle TATRA 815 ARMAX was chosen for a particular vehicle evaluation.

Table 3 The technical characteristics of TATRA 815 ARMAX (Tatra, 2010)

Parameter	Value
Length (m)	7,87
With (m)	2,5
High (m)	3,01
Maximum climbing capability at 26000 kg of cargo	36°
Maximum climbing capability at 41000 kg of cargo	18°
Maximum climbing capability up to rigid step (m)	0,5
Maximum width of trench (m)	0,9
Maximum road speed (kph)	85
Maximum depth of wade without water streaming (m)	1,2

Two processes were accomplished to verify theory of utility value evaluation. Only data with full information (all attribute properties had to be filled) were considered *in the first case*. If some information was missing in any geographic object system did not consider it and one reliable path was analysed. The first process was in progress according to next schema (Fig. 5):

- creation temporally geospatial database for CCM solution from appropriate data and information going from master database (Digital Land Model 25)
- CCM evaluation - only reliable information are considered
- final cost map calculation
- minimum cost path calculation

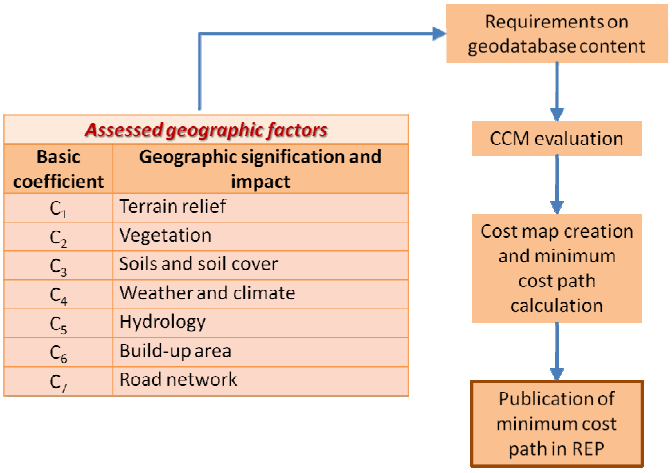


Fig. 5 Geospatial analyse without database quality evaluation

Only one solution is offered to commander. This solution seems to be appropriate, but geographer-analyst generally doesn't know details about situation on area of responsibility and commander's intentions. Problems should appear when the tactical (or other) situation does not make the published path possible to use. Commander usually requires a new solution in such a case to miss prohibited area. The quality characteristics of temporally database are then to be considered by geographer to be sure, where are the weakest points of a new analysis. The weak points of analysis have to be sent to commander together with own analysis and it is up to him what will be the final decision. Two tasks for commander appear in CCM example:

1. Use less reliable path and consider that vehicles could stay in front of some obstacle
2. Wait and order to GEO team to improve geospatial database (e.g. required properties) as soon as possible and then use new reliable path

Described *the second case* is possible to express as a progress:

- Creation temporary geospatial database for CCM solution from appropriate data and information going from master database (Digital Land Model 25).
- Temporary geospatial database quality evaluation and eventually comparing with the etalon quality (U^n = 0.8830 in the example, etalon functionality is F = 1.0068).
- CCM evaluation – only obstacles with fully added properties are considered as obstacles and their properties are compared with the technical parameters of vehicle. The other obstacles are considered only as potential obstacles.
- Final cost map calculation.
- Minimum cost path (Path 1 in the picture Fig. 7) and reliable path (Path 2) calculation. Minimum cost path assumes all obstacles without properties are probably passable. Reliable path is the same as in the first case.

If commander decides to wait for a new solution, geographer has to consider which features are the most important for given task and tries to improve their quality. The VAT is a suitable tool that can help him to make an appropriate solution under pressure of common time and personal restrictions. Detailed description is available in (Talhofer & Hofmann, 2009).

In case of mentioned task of CCM, the percentage of filled attributes of obstacles were increased of 65% in comparison with the previous stage and then the utility value of temporally database increased to 0.9132. Improved temporary database served as a new source for the cost map calculation and new reliable path (Path 3 in the picture Fig. 7) calculation.

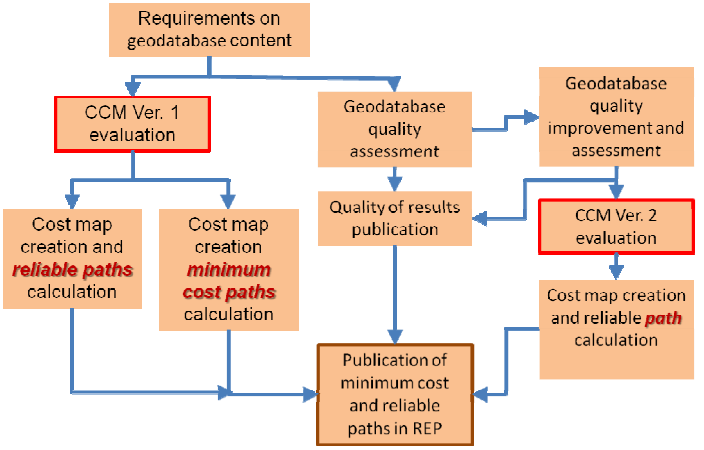


Fig. 6 Geospatial analysis with database quality evaluation and alternative results creation

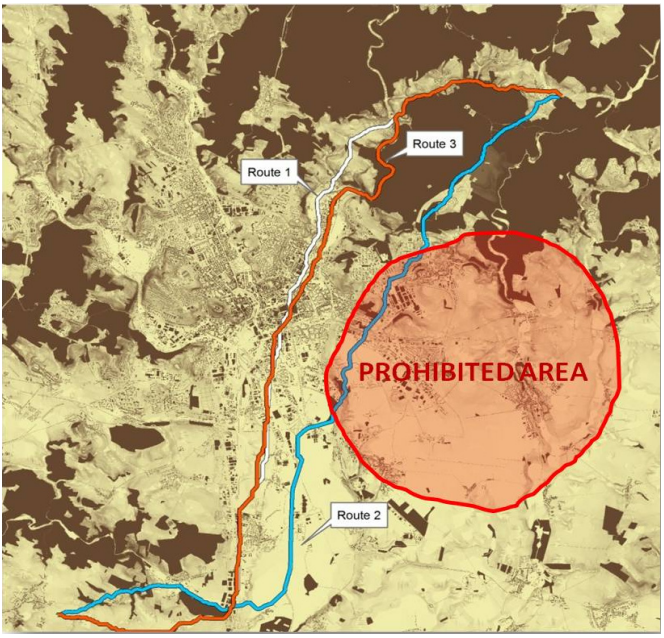


Fig. 7 Different results of CCM

3.4 CCM Modelling with Uncertainty

All the above described geographic factors (topography, vegetation, soils, weather, hydrology, build-up area, road network) can be treated as certain

information with crisp boundaries and varying levels of (attribute) quality. However, the process of geographical abstraction and generalization about the real world phenomena imposes a vast degree of approximation and uncertainty (Zhang & Goodchild, 2002). In order to overcome this situation we can employ two basic approaches – probabilistic or fuzzy sets theories. For the chosen background data model the fuzzy set approach seemed to be more appropriate. Fuzzy approach was postulated by (Zadeh, 1965) and further developed by Burrogh and Frank, (1996) within the field of geospatial data modelling. It assumes that the traditional Boolean set theory, in which the assignments are crisp, fails to represent uncertain (vague) entities or categories. Under fuzzy set theory, the transition between membership and non-membership is gradual, and any location (pixel in case of raster representation) belongs to fuzzily defined classes valued within the unit interval <0, 1>. Further, fuzzy membership values for a location typically sum to 1.0 across all classes (Zhang & Goodchild, 2002).

Positional and thematical uncertainty propagates with the number of layers which are used for CCM evaluation. To precise the whole impact of uncertainty on geospatial analysis results is the task for next development of our project.

4 Conclusion

It is obvious; the geospatial data quality has the significant impact on geospatial analyses. The authors tried to present a relationship between geospatial data, type of task which has to be solved, and the quality and the reliability of the final result. To present the whole system, one task was chosen, described, and some necessary simplifications were applied. To develop our ideas in a more complicated analysis and to consider its uncertainty will involve a future project.

Acknowledgements. Research results presented above were kindly supported by the project „The evaluation of integrated digital geospatial data reliability“ funded by the Czech Science Foundation (Project code 205/09/1198) and also by the research task „Dynamic geovisualization in crises management “supported by the Czech Ministry of Education (grant MSM0021622418).

References

Burrogh, P.A., Frank, A.U. (eds.): Geographic Objects with Interminate Bounaries. Taylor & Francis Inc., Bristol (1996)

Guptill, S., Morrison, J.: Elements of spatial data quality. Internatioanal Cartographic Association. Elsevier Science, Tokyo (1995)

Jacobsson, A., Giversen, J.: Eurogeographics (2007), http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf (retrieved 2009)

Joos, G.: Data Quality Standards. XXIII FIG Congres - Shaping the Change. Munich, Germany (2006)

Miles, L.D.: Techniques Of Value Analysis Engeneering, 3rd edn. Eleanor Miles Walker, USA (1989)

Rybanský, M.: Cross-Country Movement, The Imapct And Evaluation Of Geographic Factors, 1st edn. Akademické nakladatelství CERM, s.r.o. Brno, Brno, Czech Republic (2009)

Sanderson, M., Stickler, G., Ramage, S.: Spatial Data Quality and the Open Source Community. OSGeo Journal-The Journal of the Open Source Geospatial Foundation, 1–5 (September 2007)

Shekhar, S., Xiong, H.: Encyclopedia of GIS. Springer (2008)

Shi, W.: Principles of Moddeling Uncertainties on Spatial Data and Spatial Analysis. CRC Prass, Taylor and Francis Group, Boca Raton (2010)

Talhofer, V., Hofmann, A.: Possibilities of evaluation of digital geographic data quality and reliability. In: 24th International Cartographic Conference, The World's Geo-Spatial Solutions. ICA/ACI, Santiago de Chilie (2009)

Tatra, A.: Tatra is the solution, TATRA (2010), http://partners.tatra.cz/exter_pr/vp/new/typovy_listprospekt.asp?kod=341&jazyk=CZ (retrieved May 17, 2010)

Veregin, H.: Data quality parametr. In: Longley, P.A. (ed.) Geographic Information System, 2nd edn., pp. 179–189. John Wiley & Sons, INC., Chichester (1999)

Wang, S., Shi, W., Yuan, H., Chen, G.: Attribute Uncertainty in GIS Data. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 614–623. Springer, Heidelberg (2005)

Zadeh, I.: Fuzy Sets. Information and Control 8, 338–353 (1965)

Zhang, J., Goodchild, M.: Uncertainty in Geographical Information. Taylor and Francis, Inc., London (2002)

Mathematical Model Used in Decision-Making Process with Respect to the Reliability of Geodatabase

Šárka Hošková-Mayerová[1], Václav Talhofer[2], Alois Hofmann[3], and Petr Kubíček[4]

[1] Department of Mathematics and Physics, University of Defence, Faculty of Military Technology, Kounicova 65, 662 10 Brno, Czech Republic
sarka.mayerova@unob.cz
[2,3] Department of Military Geography and Meteorology, University of Defence, Faculty of Military Technology, Kounicova 65, 662 10 Brno, Czech Republic
{vaclav.talhofer,alois.hofmann}@unob.cz
[4] Department of Geography, Masaryk University, Faculty of Science, Kotlarska 2, 611 37 Brno, Czech Republic
kubicek@geogr.muni.cz

Abstract. The first aim of the article is to show how it is possible - thanks to the use of sophisticated analytical tools for evaluation of data quality - to better understand geospatial data. Another aim is to assess the impact of data quality on the results of space analyses that are made of them and that are the basis for such decision-making processes, in which it is necessary to take into account the impact of geographical environment.

Organizations that are engaged in creating geospatial databases usually define the content of these databases (i.e. listing of geographical objects and their features) and quality of the data being saved (e.g. geometric, topological and thematic accuracy, level of standardization etc.). As the area of the land that is described with the use of geospatial data is usually significantly larger than the capacity and technological possibilities of the responsible organization, it is not possible to keep the defined content and its quality in the entire secured area on the same level. When creating the geospatial analysis it is therefore necessary to take into account the immediate quality level of data in the particular area and to have the technologies for finding out the reliability of the result of the particular analysis available. From the real practice a request of commanders is known, that is to have not only the result of their own analysis available as basics for their qualified decision (decision-making process) but also relevant information about its reliability.

The authors of the article have quite good experience from the preparation of digital geospatial data for decision-making processes in the armed forces and within the scope of the described research they have available a large quantity of real geospatial data (current as well as historical), on which they are doing their own research focused on the mathematical modeling and evaluation.

Keywords: reliability, decision making process, mathematical modelling, geospatial data, GIS, quality assessment, utility value.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 143–162.
springerlink.com

1 Maps and Geospatial Data: The Base for Command and Control

1.1 Command and Control in a Historical Context

Command and control was in the past connected especially to the activities of armed forces. The system of command and operation management in a certain territory in history of mankind changed depending on used weapon technologies. If there were especially cold weapons used, the decisive activities leading to victory or defeat took place in several hours or maximum within a few days. Direct command on the battlefield was then realized with the help of messengers, signals and pre-agreed procedures. At the same time the commander had the overview of the whole battlefield and could make decisions how to proceed.

With the development of weapons and weapons systems and with the creation of big regular armies, changes have been made to command and control systems. The commander then lost the possibility to watch the battlefield directly and to control several-thousand soldiers in the combat units. That is why headquarters staffs who served the commander to support his decisions were formed. The coordination of the staff then necessarily required working on common base. As the military activities very often took place in the landscape, maps, especially *topographic maps*, were as one of the most important base more and more used for command and operation management. They were used for planning the future actions as well as for operational solution of a current situation during a fighting. Their content corresponded to commander and soldier requirements and maps allowed to orientate oneself in the field and to control the fighting activities as well.

Working procedures including the system of using detailed maps for planning and operation management transformed piecemeal also into the civil affairs, especially into fields where it is necessary for a higher number of units to cooperate, e.g. fire brigade, police, etc.

Work of the headquarters and its system developed depending on ways of command as well as on technical and technological development of military technical equipment and used sources. From the point of view of using the information about the territory, digitization of the landscape has slowly started. There is a new expression - *electronic battlefield*. Within the scope of electronization vast infrastructures of space data - that to a certain extent substitute classical paper maps - are currently being created.

1.2 Common Operational Picture and Recognized Environmental Picture

As a basic working environment of staff work in an electronic environment of communication and information systems is a *Common Operational Picture (COP)*, which all authorized personnel of headquarters share and which creates a

united platform for solutions of specialized tasks and cooperation of individual units. Basic topographic data are usually put into this system with the help of *Web Mapping Services (WMS*). COP contents visualized picture of landscape which to certain extent corresponds to classic topographic maps.

Geographic units, except providing *basic topographic database (geospatial database)*, create a lot of *geospatial analyses* to support the decision-making process, while using other information and data, that are not a regular part of COP. Results of those analyses are presented as a *Recognized Environmental Picture (REP)*. Examples of these analyses can be for example the calculation of visibility and hidden spaces, calculation of steepness of slopes, analysis of immediate conditions of movement of technical equipment in free space on roads as well as off roads, analysis of climatic conditions with prediction of development of several hours or a day ahead with respect to the possibility of used technical equipment, etc. REPs are created based on standard offer within the frame of geographic support system or they are created based on direct order of the commander. In the next picture there is presentation of one possibility of *Communication and Information Technologies (CIT)* application for COP and REP.

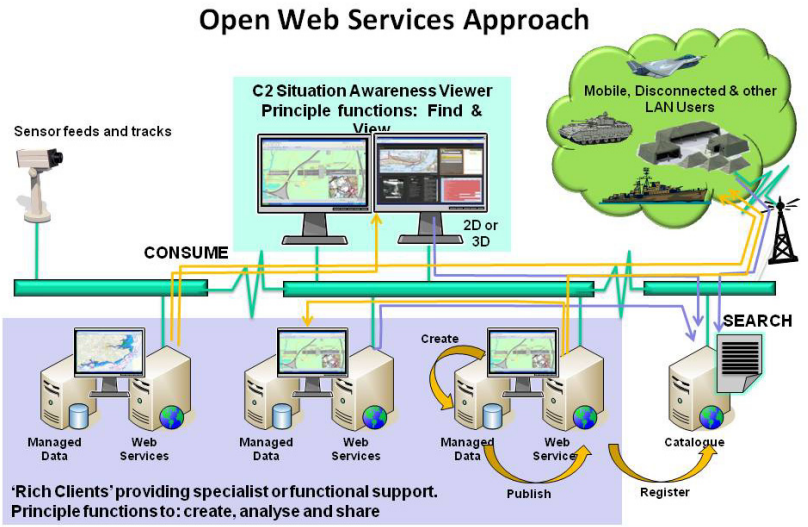


Fig. 1 Application of CIT in a command and control system (source (Tate, 2010))

As source data for REP are used data that are a product of geographic services and about which all necessary information is known, as well as other sources originating from outside the department of defense. In many tasks all the source data are mutually combined and based on mathematically or procedurally described processes new data are created. The user of source and newly created data should always, besides the space information, get also information about its

quality. In case of source primary data, this fact is usually provided by the producer. However, if new data are being created, it is necessary to have tools so that it was possible to determine the quality from the relevant quality information of the source data.

2 Quality of Geospatial Data

Systems of data and information evaluation that are positionally determined vary depending if technical parameters of data or technological influences when collecting them are evaluated, or if their final utility value given by the information quality is assessed. According to recommendation of ISO 19113, measurable qualities of provided data are evaluated as components of quality, as it is shown in the following table.

Table 1 Elements and subelements of data quality according to ISO 19113

Element of quality	Subelement of quality
Completeness	Commission
	Omission
Logical consistency	Conceptual consistency
	Domain consistency
	Format consistency
	Topological consistency
Positional accuracy	Absolute or external accuracy
	Relative or internal accuracy
	Gridded data position accuracy
Temporal accuracy	Accuracy of a time measurement
	Temporal consistency
	Temporal validity
Thematic accuracy	Classification correctness
	Non-quantitative attribute correctness
	Quantitative attribute accuracy

As technical parameters of data, the accuracy of positional information can be evaluated, it is often given as a standard positional error or a standard error in individual coordinate axes. The accuracy of thematic information is also evaluated this way. In this respect, however, the evaluation itself can be more complicated because thematic information may be various (e.g. characteristics of construction material given by the selection of one value from offered options in contrast to setting number of inhabitants of a certain settlement, etc.) and also because not all evaluated parameters must always be completed with the given object.

Generally, when formulating a problem of quality evaluation of spatial data and resulting geospatial information it is necessary to follow the recommendation of international organizations, such as ISO, OGC and DIGIWG, which pursue the development of geoinformatics, and recommendation of direction INSPIRE in the

long term (Konečný, Kubíček, & Staněk, 1998). These organizations and consortiums develop systems for quality evaluation. For instance, according to Guidebook for Implementation of Standards ISO 19100 for Quality Evaluation of Geographical Information (Jacobsson & Giversen, 2007) it is necessary to evaluate quality as a complex production as well as user's problem (see Fig. 2)

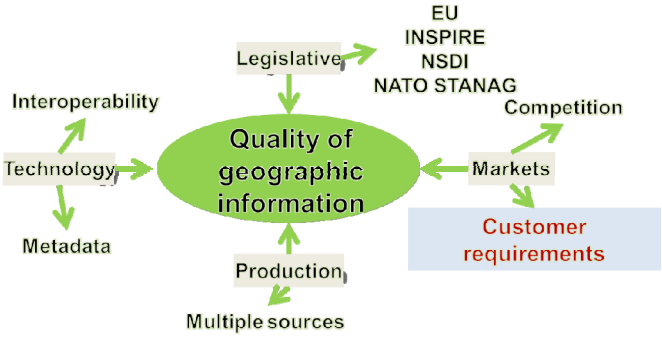


Fig. 2 Reasons for introduction of standards for quality evaluation of geospace information - according to (Jacobsson & Giversen, 2007)

The quality evaluation itself comes from general scheme components of quality, where production-technological aspects are evaluated, as well as operational and safety aspects, and in relation to a specific use of the product or service also aspects of reliability (see Fig. 3).

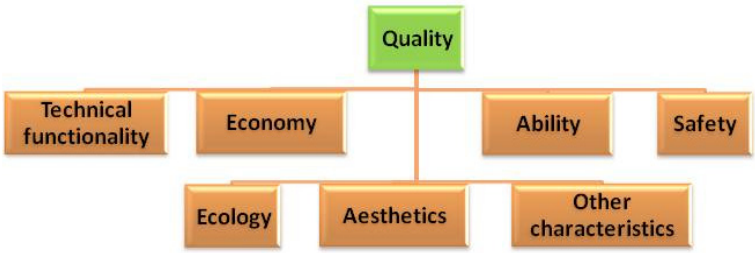


Fig. 3 General components of quality evaluation

While technical and technological parameters are possible to be evaluated generally without the necessity to know a given task, procedure or use of spatial information, reliability is necessary to be evaluated - with respect to the specific use - in a concrete process. The following text briefly explains how is possible to work with geospatial data quality in geospatial analysis.

3 Geospatial Analyses

Spatial geographic analyses (geospatial analyses) are mostly created based on a demand of the commander who wants to know specific current impacts of geographical environment and current weather conditions on his intended or planned activity, perhaps also stated impacts on the current activity in progress. Geographers and meteorologists are responsible for preparing these kinds of analyses. Most analyses that are created are usually provided to commanders, without telling them any information about the quality of the source data and their impact on reliability of the resulting analysis. This situation is documented in the following picture.

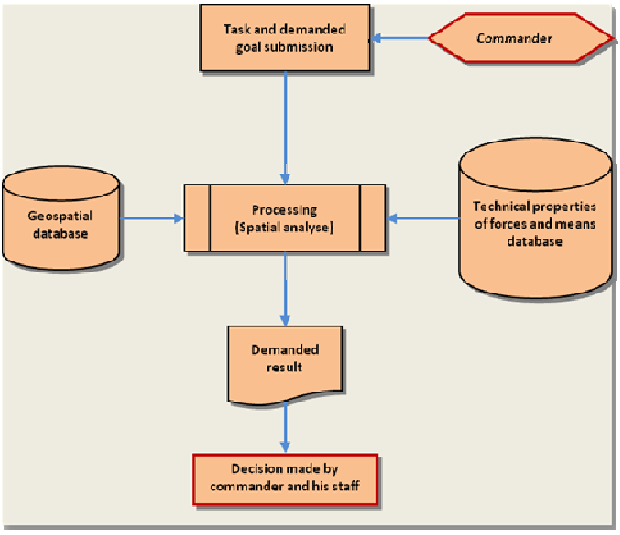


Fig. 4 Usual process of spatial analyses for the decision making process

This usual procedure has several advantages and disadvantages. The advantages are:

- relatively simple solution of the task,
- usually unambiguous result,
- commander does not have to think about the reliability characteristics of the obtained result.

Geospatial analyses are always created according to mathematical models or created schemes and given spatial data are used for them. Parameters of data quality can significantly affect the obtained results and that is why the stated procedure of geospatial analyses has certain disadvantages:

- total dependence on geospatial data,
- quality of the analysis result is not known,
- without additional information about the quality of required result the commander is left to use only one solution and has no possibility of choice among various options.

So that the above-stated disadvantages were minimized, we suggest involving also quality evaluation of source data into the geospatial analyses results.

If also level of geospatial data quality is evaluated in the process of geospatial analysis, it is possible to create several solutions. Geographer-analyst can prepare more options of solutions according to how he/she deals with quality characteristics. It is possible to create reliable, less reliable or very risky options. The commander then gets not only the result itself but also other information of riskiness of obtained solutions and it is then his responsibility what steps to take. He is either satisfied with all information he has received or he asks the geographer to increase the quality of spatial data and to prepare a new version of the space analysis. Certainly the appropriate conditions must be created to be able to increase the quality - time, personal as well as materialistic. The following picture illustrates this procedure.

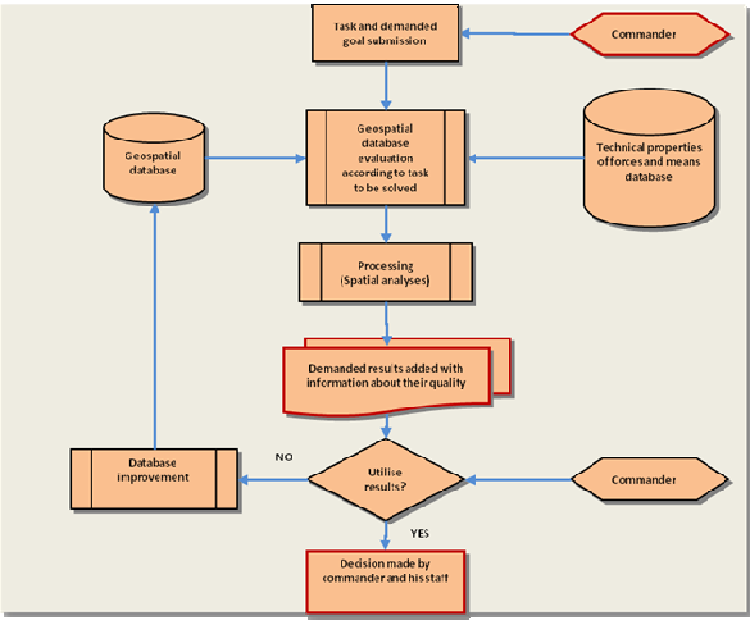


Fig. 5 Process of geospatial analyses for decision making process with the data quality consideration

4 Geospatial Data Quality Control

The process illustrated in the Fig. 5 assumes to have any complex system for level of spatial data quality evaluation which enables us not only determinate level of data quality but also to express a possibility how to increase the quality and what costs should be expect for this improvement.

4.1 Geospatial Database Utility Value Evaluation

Five essential criteria imply from geospatial data quality review (Talhofer & Hofmann, 2009). Their assessment gives the baseline for relatively reliable determination of each product utility value:

- *Database content* expresses mostly compliance of its definition and users' needs, i.e. concord of the "real modelled world" and its model represented by objects and phenomena stored in the database.
- *Database quality* defines the quality of stored data.
- *Database timeliness* explains how frequently the entire database or its elements updated is.
- *Importance of the area* is determined by users' needs so that it meets the requirements of processed or supported area range.
- *User friendliness.* This criterion defines data usability in various software environment types of GIS nature reflected mostly in compliance to standard principles. By the help of this criterion data standardization, independence, and security is revised.

Each of the criteria is mathematically assessable through independent tests and can be described as a quality parameter. In the next table there is a list of all used criteria.

4.2 General Assessment of Geospatial Data Utility

The product or a part of the product resultant function utility value may be assessed based on the above mentioned criteria using a suitable aggregation function (Talhofer, Hoskova, Hofmann, & Kratochvil, 2009):

$$F = p_3 k_3 p_4 k_4 (p_1 k_1 + p_2 k_2 + p_5 k_5) \quad (1)$$

The value of F function expresses the degree of usability of geospatial database for a given task. The chosen form of the aggregation function concerns also the case if the user gets to obsolete data or data are from an area beyond his interest so that their usage could seriously affect or even disable the geospatial data functions. Therefore k_3 and k_4 are multiple terms in the formula (1). The weight of each criterion is marked as p_i, where $i = 1,\ldots,5$. The mentioned aggregation function proves the product status at the questioned instant and its utility rate. It is applicable also to experiments to find the ways of how to increase product utility at minimum cost increment.

Table 2 List of criteria for the geospatial geodatabase utility value evaluation

Main characteristics - main criteria	**Sub-criteria characteristics**	**Definition**	**Quality parameter**
Data model content – k_1	Complexity of conceptual landscape model	Concord of conceptual model and user requirements.	Percentage of incomplete information - k_{11}
	Compliance of required resolution of geometric and thematic data – k_{12}	Concord of required geometric and thematic resolution.	Percentage of objects without required level of geometric and thematic resolution – k_{121}, k_{122}
Technical functionality – k_2	Transparency of data sources and methods for secondary data derivation - k_{21}	Transparency of source materials on primary data collection	Level of availability of information about used sources - k_{211}
		Transparency of used methods and model for secondary data derivation	Level of availability of information about used methods – k_{212}
	Position accuracy – k_{22}	Compliance with declared horizontal accuracy	Percentage of objects with unsatisfied conditions of declared horizontal accuracy – k_{221}
		Compliance with declared vertical accuracy	Percentage of objects with unsatisfied conditions of declared vertical accuracy – k_{222}
	Thematic accuracy – k_{23}	Compliance with declared accuracy of thematic data	Percentage of objects with unsatisfied conditions of declared thematic accuracy – k_{23}
	Logical consistency – k_{24}	Degree of adherence of geographic data (data structure, their features, attributes and relationships) to the models and schemas (conceptual model, conceptual schema, application schema and data model)	Percentage of objects with topological inconsistence – k_{241}
			Percentage of objects with thematic inconsistence – k_{242}
			Percentage of objects with time inconsistence – k_{243}
	Data completeness – k_{25}	Degree of adherence of the entirety of geographic data (features, their attributes and relationships) to the entirety of the modelled part of landscape	Percentage of missing objects or objects there are surplus – k_{251}
			Percentage of incomplete thematic properties of objects – k_{252}
Database timeliness – k_3	Degree of adherence geographic data to the time changes in the landscape – k_3	Value of the time function describing process of the landscape changing	Number of changes
			Time since the last up-date
			…
Landscape importance –	Value of inverse distance to objects	Landscape importance for subserved task or	Position of evaluated objects

Table 2 (*continued*)

Main characteristics - main criteria	Sub-criteria characteristics	Definition	Quality parameter
k_4	of interest – k_4	functions	
Techniques of application and safety – k_5	Data standardization – k_{51}	Declared standards adherence	Percentage of non-compliance objects with declared standard
	Independency on application software– k_{52}	Degree of independency on application software	Independency/dependency on software used for data editing
	Data protection – k_{53}	Degree of the data protection system and its level	Level of protection of user access right– k_{531}
			Level of protection of copyright– k_{532}
			Level of physical data protection– k_{533}

4.3 Individual Benefit Cost Assessment Structure

The organisation, such as the Geographic Service of the Army of the Czech Republic or the Czech Office for Surveying, Mapping, and Cadastre, are usually responsible for *geospatial databases (GDB)* development continuously covering all the Czech Republic area or some parts of the World. Digital Landscape Model (DLM25 or DMU25 in the Czech language), Multinational Geospatial Co-Production Program (MGCP) or Vector Map Level 1 (VMap1) can be mentioned as examples from military branch.

The GDB are usually developed and maintained by individual partial components of the complete database, such as save units, measurement units, map sheets etc. Therefore, it is quite a good idea to assess their utility value in the above-described system within the established the storing units introducing *individual utility value*. Similarly the individual utility value can be applied for the selected part of master databases from given *area of interest* which is used for certain task.

When assessing database utility, it is useful to define *ideal quality level* at first. The ideal level is used as a *comparison standard* to express each criterion compliance level. Using the comparison standard the individual criteria compliance level and consequently aggregate utility may be assessed.

The compliance level of each individual criterion $u_{n,s}$ is given as follows:

$$u_{n,s} = \frac{k_s}{k_s^*} \tag{2}$$

where

- k_s is for the value of s^{th} criterion compliance,
- k_s^* is for the level of compliance of s^{th} criterion or its group criterion of the comparison standard.

Than the aggregate individual utility value (*individual functionality – U_n)* of the n^{th} save unit is defined by the aggregation function of the some type as (1). Therefore:

$$U_n = p_3 u_{n,3} p_4 u_{n,4} (p_1 u_{n,1} + p_2 u_{n,2} + p_5 u_{n,5}) \tag{3}$$

Particular criteria usually consist of several sub-criteria (see Table 2). The authors took 20 criteria into their consideration; hence the equation for calculation the aggregate individual utility value is therefore a function of 20 variables that characterise the levels of compliance for each individual criterion.

Any modification of selected criterion has an impact on the value of U_n. Individual variables are independent one to another, so the derivation of the function can model the changed utility values or individual utility values.

$$dU = \frac{dU_n}{du_{n,i}} \tag{4}$$

where $i = 1,\ldots, 5$, $n = 1,\ldots, N$, and N is number of all saved units in the database.

Determination of dU value is thus feasible in two ways regarding the desired information structure. When assessing *individual variables effects* on the individual functionality value, while the other variables keep constant values, it is necessary to differentiate U function as follows:

$$dU = \frac{dU_n}{du_{n,i}} \frac{du_{n,i}}{dx} \tag{5}$$

where x is one of the 20 mentioned variables.

In practice, however, such situations may arise that multiple factors may change at the sometime, e.g. the technical quality of database changes in all its parameters—the secondary data derivation methods will improve location and attribute accuracy and the data integrity will increase, and moreover the data are stored in a geodatabase accessible to all authorised users. In this database the data are maintained properly with respect to all topologic, thematic and time relations. In such a case it is suitable to define dU value as a total differential of all variables describing the modified factors.

Database functionality degree is comparable to the cost necessary for provisions—direct used material, wages, other expenses (HW, SW, amortisation, costs for co-operations, tax and social payments etc.), research and development cost, overhead cost and others. Functionality and cost imply *relative cost efficiency (RCE)* calculated as follows:

$$RCE = \frac{F}{\sum_{i=1}^{n} E_i} \tag{6}$$

where $i = 1,\ldots, N$.

Similarly to individual utility value, it is possible to consider the impact of particular variables of expenses E_i on final *RCE*. The goal is to find such solution as the functionality will be maximised and the expenses will be minimize.

The GDB benefit cost assessment including individual benefit cost is a task for a data manager or a geographer-analyst which is responsible to provide demanding project. The system enables him to consider which quality parameters are possible to improve in given time, with given technological conditions, with given sources, with given co-workers etc.

5 Pilot Study

In order to verify our methodology the task of Cross Country Movement (CCM) was chosen as an example. CCM can be solved as a common problem or with consideration of certain types of vehicles (the most frequent or the weakest in the unit, but in case of armed forces usually off road vehicles). The detailed theory of CCM is in (Rybanský, 2009).

The solution can offer to the commander not only one possibility, but the variants from which he can choose according to his intentions and the current situation at the given area.

5.1 Cross Country Movement

Let us recall the basis of the CCM theory. The main goal of CCM theory is to evaluate the impact of geographic conditions on of movement of vehicles in terrain. For the purpose of classification and qualification of geographic factors of CCM, it is necessary to determine:

- particular degrees of CCM
- typology of terrain practicability by kind of military (civilian) vehicles
- geographic factors and features with significant impact on CCM

As a result of the geographic factors impact evaluation we get three known degrees of CCM:

- GO - passable terrain
- SLOW GO - passable terrain with restrictions
- NO GO – impassable terrain

Geographic factors determining CCM and the selection of the access routes are follows:

- gradient of terrain relief and micro relief shapes
- vegetation cover
- soil conditions
- meteorological conditions
- water sheets, water courses

- settlements
- communications
- other natural and manmade objects

The impact of given geographic factor can be evaluated as a *coefficient of deceleration 'C$_i$'* from the scale of 0 to 1. The coefficient of deceleration shows the real (simulated) speed of vehicle v_j in the landscape in the confrontation with the maximum speed of given vehicle v_{max}. The impact of the whole n geographic factors can be expressed as the formula:

$$v_j = v_{\max} \prod_{i=1}^{n} C_i, \qquad n = 1, \ldots, N. \tag{7}$$

The main coefficients of deceleration are listed in the next table (see Table 3).

Table 3 Main coefficients of deceleration

Basic coefficient	Geographic signification and impact
C_1	Terrain relief
C_2	Vegetation
C_3	Soils and soil cover
C_4	Weather and climate
C_5	Hydrology
C_6	Build-up area
C_7	Road network

Each coefficient consists of several coefficients of 2nd grade. For example C_2 is express after simplification as:

$$C_2 = C_{21} C_{22}$$

where:

- C_{21} is deceleration coefficient by impact of trunks spacing,
- C_{22} is deceleration coefficient by impact of trunks diameter.

The values of deceleration coefficients are counted for given vehicle (its technical properties) from ascertained properties of geographic objects stored in the spatial geodatabase. Using formula (7) it is possible to create a cost map in which the value of each pixel is the final (modelled) speed. The cost map can be as a source for the fastest path, reliable path etc. calculation.

Example for C_2 coefficient evaluation

Let us to define a task 'Looking for the appropriate place for hidden command point and find out an appropriate route there'. If the place is found in a forest, its properties is necessary to evaluated concerning to given army vehicles.

The forest properties are saved in the DLM25 database where the trunks spacing is specified as *TSC* parameter and trunks diameter as *SDS* parameter measured in the high of 1.2 metres above the surface. The vehicle can pass forest up to TSC_{max} spacing and SDS_{max} thickness without some serious problems. The terrain vehicle can pass forest with reduced speed, if the trunks spacing is smaller but passable (TSC_{min}) and vehicle has to turn among trees, or trees thickness enables to break them (SDS_{min}). If the size of obstacle is bigger, the vehicle velocity is 0. Properties TSC_{min} and SDS_{min} are given by the technical description of given vehicles validated at the field tests, and comparative values are read from spatial geodatabase. In the mathematical formula the condition can be express:

$$C_{21} = \begin{cases} 1 & for \quad TSC > TSC_{max} \\ 0{,}5 & for \quad TSC = (TSC_{min}; TSC_{max}) \\ 0{,}25 & for \quad TSC < TSC_{min} \wedge SDS < SDS_{max} \\ 0 & for \quad TSC < TSC_{min} \wedge SDS > SDS_{max} \end{cases} \tag{8}$$

$$C_{22} = \begin{cases} 1 & for \quad SDS < SDS_{min} \\ 0{,}5 & for \quad SDS = (SDS_{min}\,; SDS_{max}) \\ 0 & for \quad SDS > SDS_{max} \wedge TSC < TSC_{min} \end{cases} \tag{9}$$

5.2 Geospatial Database Utility Value Evaluation

The master DGI database is usually utilised as a base for spatial data analyses. The national or international databases as DLM25, VMAP1or MGCP are very detailed, carefully maintained and used in many applications. But nobody can suppose that those databases contain all information he could need.

The task of CCM solution could require more information that is available in the master database. Geographer-analyst has to consider which information and in what quality can he obtain from master database. E.g. all forests in the area of interest are necessary to select for mentioned C_{2i} coefficients. Further he has to find out all their properties and their accuracy or count how many characteristics are missing. The system presented in the Table 2 serves as a manual. Next step is the individual utility value of given part of master database evaluation.

Attributes are usually defined as the characteristics or nature of objects. In geospatial sense, an attribute is regarded as a property inherent in a spatial entity (Shi, 2010). In our case an attribute is a characteristics or variable constituting the base for the computation of basic coefficients C_1 - C_7. These attributes differ in their nature according to the real world phenomena they represent.

Not all attributes are available within the used thematic spatial databases. So far the incompleteness of attributes has been omitted. Thus the real state-of-the-art has not been taken into account and the resulting CCM path has been considered as ‘certain’. One of the possibilities to make the resulting path closer to reality is to take the data attribute incompleteness into account and inform the decision maker (commander) about the uncertain parts of the path.

Two variants of the DLM25 database were utilised for the pilot project. The feature properties were defined according to the Feature Attribute Coding Catalogue (FACC) adapted as Catalogue of the Topographic Objects (CTO) (MTI,

2005) in the first variant updated in 2005. The missing values of object's attributes were marked as 0 in given domains. The 4th edition of CTO was transformed in accordance with the DGIWG Feature Data Dictionary (DGIWG-500, 2010) in 2010 and transformed edition (updated in 2010) was used in the second variants (MoD-GeoS, 2010). The missing properties were marked in several attribute categories as:

- -32767 for unknown variables,
- -32766 for unpopulated variables,
- -32765 for not applicable variables,
- -32764 for other variables, and
- -32768 for No or Null values.

The smaller personal database was created in the area of interest round Brno of the size approximately 400 km^2 and all objects and phenomena necessary for CCM evaluating was selected from DLM25 master databases of both variants. The individual utility value was counted for both variants, but with a small simplification. At the first step we didn't do any independent tests for position accuracy determination, further we didn't consider the software independency, and landscape importance. Then we suppose the whole database is complete, the position and thematical resolution corresponds to our task, and the data are properly protected.

On the base of statistical analyse 12.65% objects have any problems mainly incomplete attributes in the first variant of DML25 while 3.45% objects have any similar problems in the second one. The time difference is 5 years between both variants. Hence the individual utility value was calculated by the use of the formula (3) as 0.6887 for the 2005 variant and 0.8825 for the 2010 variant. The ideal quality level is 1.0068. Both variants were used for CCM of TATRA 815 evaluation.

5.3 CCM of TATRA 815 Analyses

The common army vehicle TATRA 815 ARMAX was chosen for a particular vehicle evaluation.

Table 4 The technical characteristics of TATRA 815 ARMAX (Tatra, 2010)

Parametr	Value
Length (m)	7,87
With (m)	2,5
High (m)	3,01
Maximum climbing capability at 26000 kg of cargo	36°
Maximum climbing capability at 41000 kg of cargo	18°
Maximum climbing capability up to rigid step (m)	0,5
Maximum width of trench (m)	0,9
Maximum road speed (kph)	85
Maximum depth of wade without water streaming (m)	1,2

For further simplification only data with full information (all attribute properties had to be filled) were considered *in both variants*. If some information was missing in any geographic object system didn't consider it and one reliable path was analysed. Authors knew that e.g. narrow streams are probably passable for TATRA lorry, but their passing was allowed only over bridges if no information about depth and banks characteristics were in the database.

The process was in progress according to next schema (Fig. 6):

- CCM evaluation - only reliable information are considered,
- final cost map calculation according to equation (7),
- minimum cost path calculation from one initial point to three destinations placed in the forests as some hidden position.

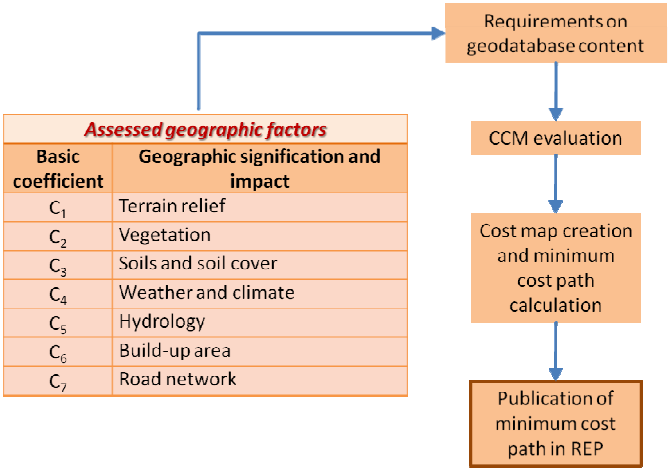


Fig. 6 Spatial analyse without database quality evaluation

Only one solution is offered to commander. This solution seems to be appropriate, but geographer-analyst generally doesn't know details about situation on area of responsibility and commander's intentions. Problems should appear when the tactical (or other) situation doesn't make the published path possible to use. Commander usually requires a new solution in such a case to miss prohibited area. The quality characteristics of temporally database are than to be considered by geographer to be sure, where are the weakest points of a new analysis. The weak points of analysis have to be sent to commander together with own analysis and it is up to him what will be the final decision. Two tasks for commander appear in CCM example:

1. Use less reliable path and consider that vehicles could stay in front of some obstacle
2. Wait and order to GEO team to improve spatial database (e.g. required properties) as soon as possible and then use new reliable path

The second case was simulated by the second variant of database in which the quality parameters were improved.

ArcGIS 9.3 was used for all calculations and analyses. The main analyses were described using ModelBuilder to have simpler possibility to change the input parameters. In the next figures there are the main results – cost maps. The cost of each pixel is symbolized in the gray scale where darker tone signifies higher cost, higher speed in this case.

Fig. 7 The cut cost map created from DLM25 2005 version

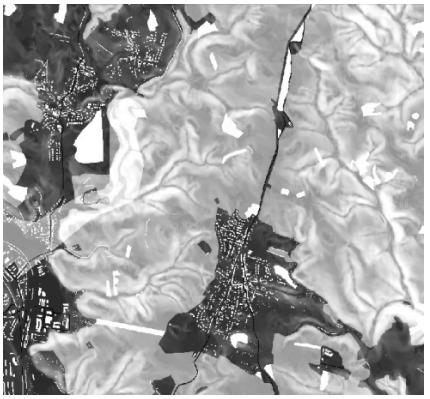


Fig. 8 The cut cost map created from DLM25 2010 version

The minimum cost paths were evaluated using both cost maps and the same process created in ModelBuilder were applied. The results are in the next figures.

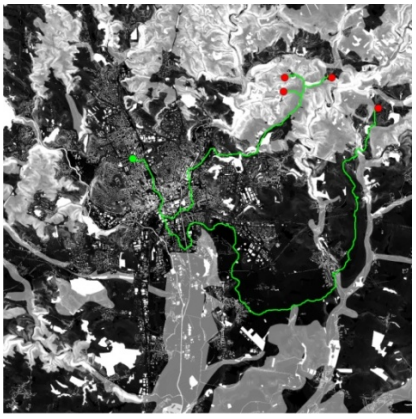


Fig. 9 The minimum cost paths in CM of 2005 version. The initial point is green, the destinations are red

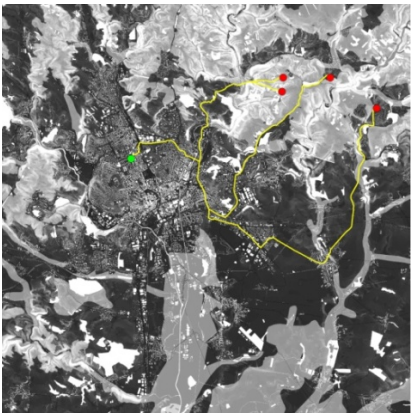


Fig. 10 The minimum cost paths in CM of 2010 version. The initial point is green, the destinations are red

The comparing of both results presented over the topographic situation is shown in the next picture (Fig. 11).

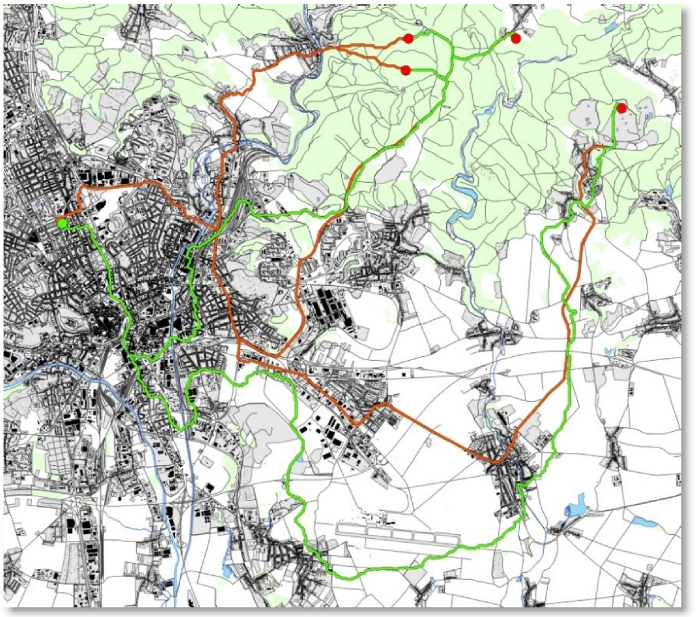


Fig. 11 Comparing of two variant of minimum cost paths. Red ones answer to 2005 version and the green ones to 2010 version

5.4 Discussion

The obtained results proved that the results of spatial analyses are highly dependent on overall quality of digital spatial data. Relatively small changes in the database can cause significant differences in received results.

Authors purposely considered only original stages of selected objects from master databases in both variants where no uncertainty is evident. But the real situation in the landscape is quite different. Some thematic properties are uncertain and they can change in shorter or longer time (trees growth, soil moisture, river with and depth etc.), some are uncertain because of no possibility to measure them in the required precision (e.g. soil types borders). There exist several methods how to handle the attribute uncertainty and incompleteness (Shi, 2010).

Slightly other situation is when the missing properties of data are considered. Some properties are missing on purpose because their application has no sense or is impossible because of some restrictions. Some properties are missing because of lack of time to add all declared object properties and from the geographer point of view the object importance seems not to be very high. Moreover, the geographer can assume that the user could know the area of interest and that he is able to add

missing properties (with respect to his experience). E.g. Small streams, round Brno, have in the summer its width up to 2 meters and depth up to 0.5 meters, so they are easily passable. On the other hand, banks of small streams can be high, muddy etc. and so impassable. To complete all declared properties is costly and time demanding, so sometimes the question is whether to add the missing data and spend money for it or not to add the data and rely on the user‘s knowledge about the environment.

Using the cost map it is possible to create not only reliable path, but also risky path if some ‘unimportant’ obstacles are sign as passable. If the commander obtains the risky path, the additional information about obstacles which were not included into consideration is necessary to be added to him. Then it is up to commander to consider the level of risk decision taken.

6 Conclusion

The pilot project has demonstrated a strong relationship between quality data and the results of spatial analysis. Likewise, it pointed to the problem of defining quality. It is not possible to assess only the technical properties of the spatial database, but it is necessary to consider the quality of the entire complex. For example, when changing standard used, which at first glance may not be large, its influence on the resulting analysis can be substantial. Specifically, during the solution of the CCM task, we faced a problem of classification of thematic properties. In an earlier version, blank data were marked only to the value 0, while in the new version, the blank data disintegrate into several groups identified by the values of -32765, -32766, etc. according to specifications (DGIWG-500, 2010).

In the pilot project, we have dealt only marginally with uncertainty in setting the boundaries of geographic objects and phenomena and with the uncertainty of their thematic properties. The authors are aware of the fact that this uncertainty, given by the natural conditions, can significantly affect the results of spatial analysis. For example, the width of the water flow is usually stored in a database as the value of the width of the flow in the normal state level. If it spill over due to the heavy rains, this width may be much larger, or vice versa during prolonged drought it may diminish. As a result, the conditions of negotiability of this flow changes significantly. The problem of implementation of the principles of uncertainty and their mathematization will be the task of the solution of our future project.

The actual pilot project showed that the proposed way to address the relationship between the quality of spatial data and spatial analysis is possible and should continue.

Acknowledgements. Research results presented above were kindly supported by the project „The evaluation of integrated digital geospatial data reliability“ funded by the Czech Science Foundation (Project code 205/09/1198) and also by the research task „Dynamic geovisualization in crises management “supported by the Czech Ministry of Education (grant MSM0021622418).

References

DGIWG-500. Implementation Guide to the DGIWG Feature Data Dictionary (DFDD) (2.2.2 - July 19, 2010 ed.). DGIWG (2010)

Jacobsson, A., Giversen, J.: Eurogeographics (2007), http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf (retrieved 2009)

Konečný, M., Kubíček, P., Staněk, K.: Mezinárodní standardizační aktivity v oblasti prostorově orientovaných informačních systémů. In: Sborník příspěvků Konference Geomatika, vol. 98, pp. 35–41. CAGI, Praha (1998)

MoD-GeoS, Catalogue of the Topographic Objects DLM25 (7.3 ed.). Ministry of Defence of the Czech Republic, Geographic Service, Dobruska (2010)

MTI. Catalogue of the Topographic Objects DLM25 (4 ed.). Military Topographic Institute, Dobruska (2005)

Rybanský, M.: Cross-Country Movement, The Imapct And Evaluation Of Geographic Factors, 1st edn. Akademické nakladatelství CERM, s.r.o. Brno, Brno, Czech Republic (2009)

Shi, W.: Principles of Moddeling Uncertainties on Spatial Data and Spatial Analysis. CRC Prass, Taylor and Francis Group, Boca Raton (2010)

Talhofer, V., Hofmann, A.: Possibilities of evaluation of digital geographic data quality and reliability. In: 24ht International Cartographic Conference, The World's Geo-Spatial Solutions, ICA/ACI, Santiago de Chilie (2009)

Talhofer, V., Hoskova, S., Hofmann, A., Kratochvil, V.: The system of the evaluation of integrated digital spatial data realibility. In: 6th Conference on Mathematics and Physics at Technical Universities, pp. 281–288. University of Defence, Brno (2009)

Tate, J.: Open standards the Gateway to Interoperability? GEOINT Summit 2010. Jacobs Flaming Conferences, Vienna (2010)

Tatra, A.: Tatra is the solution, TATRA (2010), http://partners.tatra.cz/exter_pr/vp/new/typovy_listprospekt.asp?kod=341&jazyk=CZ (retrieved May 17, 2010)

Complex Networks Topology: The Statistical Self-similarity Characteristics of the Average Overlapping Index

Francisco O. Redelico[1,2] and Araceli N. Proto[3]

[1] Facultad de Ciencias, Fisicomatemáticas e Ingeniería, Universidad Católica Argentina
francisco.redelico@uca.edu.ar
[2] Departamento de Ingeniería Industrial, Instituto Tecnológico Buenos Aires
[3] Laboratorio de Sistemas Complejos, Facultad de Ingenieria , UBA
aproto@fi.uba.ar

Abstract. In this paper some statistical properties of the Average Overlapping Index (AOI) are quantified. The AOI can be interpreted as a measure of local clustering properties of a node, indicating the node robustness against external perturbation. It has been considered in many different disciplines such as computer science, macroeconomics, nonlinear dynamics and opinion formation. The AOI values reflect the networks topology, in the way that according the complex network generation mechanism, some AOI values became forbidden. For that reason the corresponding AOI set for each network has multifractal properties. This multifractal property is capable to grasp the generation mechanism of the respective network. The support of the multifractal is also a fractal set.

Keywords: Multifractal analysis, Complex Networks, Average Overlapping Index.

Introduction

Complex networks have its roots in graph theory and are intensively used in different fields. An *unweighted and undirected graph* can be represented with two sets $\{N, A\}$, $N = \{1, \ldots, M\}$ representing the nodes and $A = \{a_{1,1}, a_{1,2}, \ldots, a_{2,1}, a_{1,2}, \ldots, a_{M,1}, \ldots, a_{M,M}\}$ containing the edges, or arcs, linking each node. A graph, and a complex network also, admits a matrix representation, i.e. the adjacency matrix $\mathbf{A}$ where $a_{i,j} = 1$ iff the node i is linked the node j and $a_{i,j} = 0$ otherwise.

Among the several measures utilized in the literature in order to characterize complex networks, the Node Degree appears as the more representative, being defined as the number of edges incident (outgoing or ingoing) to the

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 163–174.
springerlink.com

node. Another proposed measure in the literature is the *Average Overlapping Index* (AOI). It was first introduced and then analyzed and utilized in very different disciplines such macroeconomics [2, 3], nonlinear dynamics [4], opinion formation [5] and computer science [6, 7]. It can be interpreted as a measure of local clustering properties of a node.

In this paper a relationship between the AOI and the node degree is introduced for each of the most popular complex networks, i.e. Erdös-Rényi model [13], Small-world model [16] and the scale free model [9]. In addition, the statistical self-similarity property of the AOI is quantified among these complex networks models using the multifractal Legendre coordinate spectrum [20].

The paper reads as follows, Section 1 and 2 a brief review of the measures analyzed in this paper as well as the types of complex networks used for simulations and the singularity spectrum, are outlined in order to fix nomenclature; Section 3 numerical results are shown, and Section 4 is devoted to conclusions.

1 Statistical Properties of Complex Networks

In this section, both, the models generating complex networks, and measures for describing these networks and the measures mentioned in the Introduction section are described.

1.1 Statistical Measures of Complex Networks

The Node Degree for a binary no directed network is defined as

$$k_i = \sum_{j=1}^{N} a_{i,j} \tag{1}$$

where $a_{i,j}$ is the element $i, j-th$ of the adjacency matrix of the network and N is the total number of nodes in the network.

Let us also remind the AOI mathematical formulation and its properties in the case of a unweighted network made of M nodes linked by edges (l, m) $(1 \leqslant l \leqslant M;\ 1 \leqslant m \leqslant M)$. The *overlapping index* is defined as

$$O_{lm} = \frac{N_{lm}\,(k_l + k_m)}{4\,(M-1)\,(M-2)}, \qquad l \neq m \tag{2}$$

where N_{lm} is the measure of the common number of neighbors of l and m nodes, (in a fully connected network, $N_{lm} = M-2$) and k_l is the *vertex degree* or the number of vertex connections of node l. Eq.(2) can be expressed in matrix form as

$$\mathbf{O} = \mathbf{a}_l \mathbf{a}_m^t\,(\mathbf{a}_l + \mathbf{a}_m)\,\mathbf{1}^t \tag{3}$$

where $\mathbf{O}$ is a $M \times M$ dimension matrix containing the AOI values, $\mathbf{a_i}$ is the corresponding adjacency matrix row for the node i, $\mathbf{1}$ is a M dimensional column vector of 1's and ^{t} denotes transpose.

The *Average Overlap Index* for the node l is defined as:

$$\langle O_l \rangle = \frac{1}{M-1} \sum_{m=1}^{M} O_{lm} \tag{4}$$

and in matrix form

$$\langle \mathbf{O} \rangle = \mathbf{O}_l \frac{\mathbf{1}}{M-1} \tag{5}$$

where $\langle \mathbf{O} \rangle$ is a vector containing the AOI values, $\mathbf{O}_l$ is the $l-th$ row of the $\mathbf{O}$ matrix, $\mathbf{1}$ is a M dimensional vector of ones and M is the number of nodes in the networks.

These measures $\langle O_i \rangle$ can be interpreted as a kind of clustering measure: the higher the number of nearest neighbors is, the higher the $\langle O_i \rangle$ is. Since the summation is made over all possible j sites connected to i (thus over all sites in a fully connected graph), $\langle O_i \rangle$ expresses a measure of the local density near the i site. The $\langle O_l \rangle$ can thus be interpreted as a measure of the stability of the node against perturbations due to an external cause. Thus a high $\langle O_l \rangle$ reflects how much a node is strongly attached to its community.

1.2 Statistical Models for Complex Networks

There are three basic types model for complex networks, called the Erdös-Rényi, the Watts-Strogatz and the Albert-Barabasi model.

- *Erdös-Rényi model*: This model was proposed by Erdös and Rényi [13] in order to study statistical properties of graph meanwhile the number of connections between nodes increase. In this kind of model two nodes are linked with an edge with a probability p_c, $p_c \in (0, 1)$. Statistical properties of these models are well setup in various papers [9, 14].
- *Small-world model*: The experimental psychologist Stanley Milgram coined the concept of *small world* in [15]. Inspired by this concept, Watts and Strogatz [16] developed a type of complex network with small world properties making interpolations between regular graph and random graphs. The Watts-Strogatz model is based on a rewiring with a probability p_c. through the procedure:

  1. Starting with a ring with N nodes, each connected to its $2m$ nearest neighbors, with a total of $K = mN$ edges.
  2. Each node is rewiring to its neighbors with a probability p_c.

- *Scale free model*: Several complex networks are scale free, i.e. its degree distribution follows a *power law* instead of a Poisson distribution [9]. Barabasi and Albert [17] postulated that this power law distribution is generated

by two mechanisms, i.e. *Growth* and *Preferential attachment*. Growth mechanics means that the model captures the essential features of an open system from a *small* group of nodes; the network evolves by adding new nodes during its life-span. Preferential attachment mechanism means that the probability of two nodes being connected depends on the nodes degree. The algorithm for generating this kind of networks is as follows,

1. In the first step there are m_0 isolated nodes.
2. Adding a new node n_j
3. The probability of a node n_j links with an existing node n_i is proportional to the degree of the node n_i
4. Repeat 2 and 3 until the network has $N - m_0$ nodes.

More detailed can be found in [9, 10, 18] where a discussion of complex networks generation mechanisms is exposed.

1.3 Brief Review of Multifractal Analysis

The dimension of a fractal object, like the Cantor Set [19], could be determined by the similarity dimension. In this case its value is $log(2)/log(3)$ [20]. However, although this measure is appropriated for describing a fractal set it is not so appropriated for multifractal one where even an infinite number of dimensions can be present. For that reasons a spectrum of singularities is a useful for describing such sets. We will look for fractal and multifractal properties of the AOI. This multifractal analysis will give a framework in order to describe the robustness of the network.

Multifractals can be described by the spectrum of the Holder exponents and its fractal dimension $f(\alpha)$ [20]. Although there are several ways to compute the spectrum $(\alpha, f(\alpha))$in this paper a *Legendre transformation* of $q$and $D_q$variables are applied in order to compute it, as it follows,

$$\alpha = \frac{d}{dq}\left[(q-1)\, D_q\right] \tag{6}$$

and

$$f(\alpha) = q\frac{d}{dq}\left[(q-1)\, D_q\right] - (q-1)\, D_q \tag{7}$$

where D_q is the generalized dimension [20], $D_q = \lim_{\delta \to 0} \frac{1}{q-1} \frac{d \log[S_q(\delta)]}{d \log(\delta)}$ obtained using a generalization of the correlation sum as, $S_q(\delta) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{N-1} \sum_{i=1, i \neq j}^{N} \Theta(\delta - \delta_{i,j}) \right]^{q-1}$ where $\Theta(x)$ is the Heaviside function $\Theta(x) = \begin{cases} 0 & \text{for} \quad \text{x ¡ } 0 \\ 1 & \text{for} \quad \text{x} \geqslant 0 \end{cases}$ and $\delta_{i,j} = \sqrt{\sum_{k=0}^{N-1} (X_{i-k} - X_{j-k})^2}$ where N is the number of elements within the F set.

The interested reader is referred to classical books like [20, 24] where this subject is widely covered.

2 Numerical Results

In this Section the corresponding results are presented. In order to relate to Average Overlapping Index and the node degree (Eq. 1) we made the following *ansatz* whose validity will be proved:

$$AOI = \alpha k_i^{\beta} \tag{8}$$

where α and β are parameters to be estimated. This assumption implies that a power law relationship may exists between the AOI and the Node degree (k_i).

The R^2 estimator will be used to measure the quality of simulations and Eq. 8, defined as

$$R^2 = \frac{\left(\langle xy\rangle - \langle x\rangle \langle y\rangle\right)^2}{\left(\langle xx\rangle - \langle x\rangle \langle x\rangle\right)\left(\langle yy\rangle - \langle y\rangle \langle y\rangle\right)} \tag{9}$$

where x and y are the variables under analysis. In order to quantified the uncertainty related to the estimation of the free parameters in Eq. 8 a t-student confidence interval will be used, according the following,

$$\left(\langle AOI\rangle - M^{-1/2}\hat{\sigma}t_{M-1,1-\alpha/2}, \langle AOI\rangle + M^{-1/2}\hat{\sigma}t_{M-1,1-\alpha/2}\right) \tag{10}$$

where $\langle AOI\rangle$ is the mean of the AOI-set, M the number of nodes, $\hat{\sigma}$ is the variance estimation of the AOI set and $t_{M-1,1-\alpha/2}$ is a t-Student variable with $M-1$ degree of freedom with a α probability [26]. Finally, the multifractal spectra $(\alpha, f(\alpha))$ (Eq. 6,7) is evaluated in order to find whether or not the AOI is able to capture complex networks selfsimilarity properties.

Fig. 1 and Table 1 show the results for Eq. 8 for the Erdös-Rényi type networks. In Fig. 1, different connection probabilities are plotted; they increases as the AOI decreases. In Table 1 the estimated parameters for different connection probabilities are shown. The goodness of the fit (Eq. 9) runs from 1 to 0,8718 as the connection probability increases or in other words the fitting is better for low connection probabilities. This fact can be explained as AOI reflects how much a node is strongly attached to its community. So, low AOI values mean that the node is not strongly connected, having, then, a low connection probability.

In Fig. 2 the estimated Legendre spectrum for the AOI over 10000 realizations of the Erdös-Rényi model is shown with the standard error bars. In this figure only the spectrum for a network with connection probability 0.7 is shown for sake of clarity. It is shown not only that the AOI has multifractal

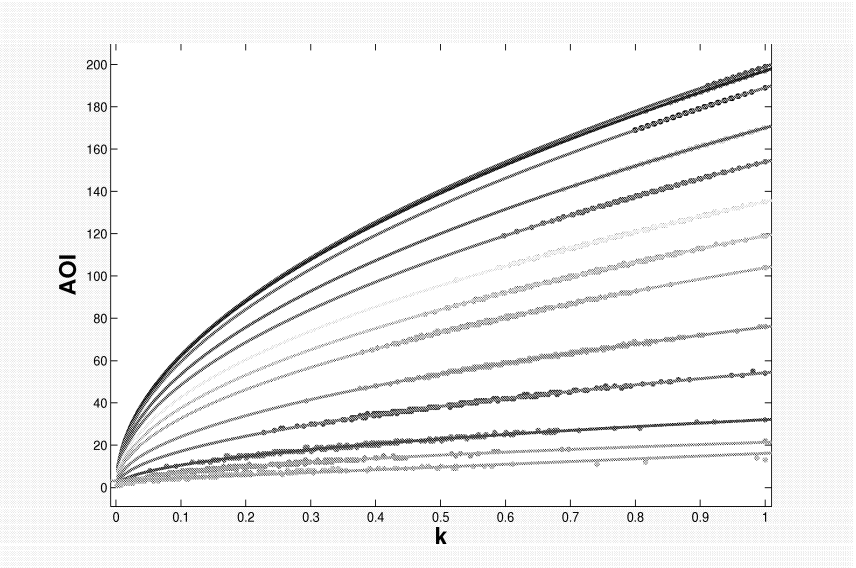


Fig. 1 Node degree, (k), Eq. 1, versus Average Overlapping Index, (AOI) Eq. 4 for the Erdös-Rényi model. The estimated parameters for Eq. 8 as well as other fitting statistics are shown in Table 1. The connection probability of the networks increases as the AOI decreases.

properties but also its support has a fractal dimension $D_0 = 0.9802$. For the Watts-Strogatz type networks, Fig. 3 and Table 2 shows the results for Eq 8. In this power law the goodness of fit, as it is shown in Table 2, increasing as the rewiring probability also increases. As before, this fact can be explained as the AOI reflects how much a node is strongly attached to its community. So, low AOI value means that the node is not highly connected, and then it has a low rewiring probability, as is explained in Section 1. Fig 4 shows the estimated Legendre spectrum for 1000 realization of the Watts-Strogatz model with 10000 nodes each one and a rewiring probability equal to 0.7. In contrast to the former (Fig. 2), it is not so symmetric, this fact will be interpreted in the Conclusion section.

In Fig. 5 and Table 3 the estimated parameters for Eq. 8 for the Albert-Barabasi type network are presented. The goodness of fit remains almost constant through the node degree, as in shown in Table 3. Fig 6 shows the estimated Legendre spectrum for 1000 realization of the Albert-Barabasi model with 10000 nodes each one for a node degree equal to 2. It is the most asymmetric spectra; this point will be noted in the Conclusion Section.

Table 1 Several connection probabilities for the Erdös-Rényi model (Column 1). Parameter estimated for Eq. 8 with its t-student confidence levels (Ec. 10) (Columns 2 y 3). Values for the R^2 (Ec. 9) for each fit (Column 4). The higher the connection probability, the poorer the goodness of fit is.

p_c	α	β	R^2
0.02	199 (199, 199)	0.5022 (0.5018, 0.5025)	1
0.05	197 (197, 197)	0.5013 (0.5008, 0.5019)	0.9999
0.1	189 (189, 189)	0.5022 (0.5013, 0.5031)	0.9998
0.2	170 (170, 170.1)	0.5024 (0.5011, 0.5037)	0.9997
0.3	153.9 (153.9, 154)	0.502 (0.5002, 0.5037)	0.9994
0.4	135.1 (135.1, 135.2)	0.5006 (0.4985, 0.5026)	0.9991
0.5	119 (118.9, 119.1)	0.5007 (0.4985, 0.5029)	0.999
0.6	103.9 (103.7, 104.1)	0.5024 (0.4995, 0.5053)	0.9983
0.7	75.89 (75.76, 76.02)	0.4996 (0.4959, 0.5034)	0.9973
0.8	54.25 (54.09, 54.42)	0.4996 (0.495, 0.5043)	0.9955
0.9	32.07 (31.83, 32.31)	0.4808 (0.4731, 0.4885)	0.9881
0.95	21.38 (21.09, 21.67)	0.4799 (0.4703, 0.4895)	0.9832
0.98	12.94 (12.25, 13.64)	3.233 (3.046, 3.421)	0.8718

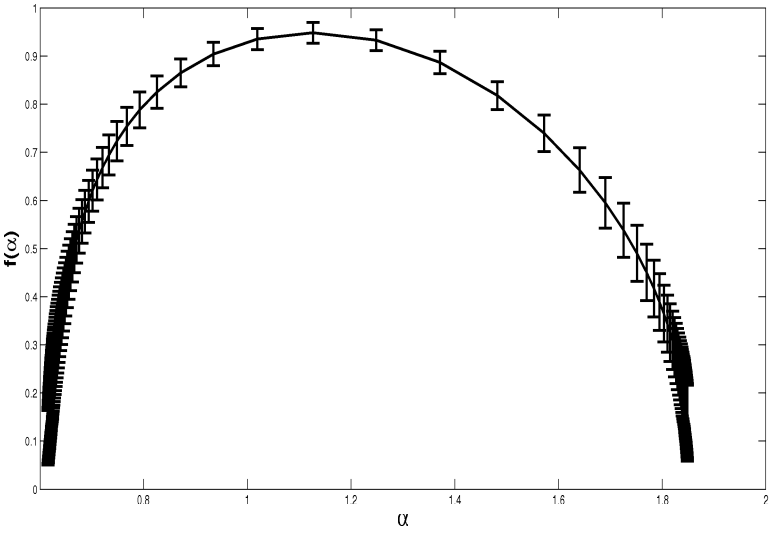


Fig. 2 Legendre spectrum for the AOI estimated over 1000 realization of the Erdös-Rényi model with 10000 each one and connection probability equal to 0.7. The support of the measure has fractal dimension as $D_0 = 0.9802$. The spectrum is quite symmetric, indicating there is not a privileged node regarding its connections.

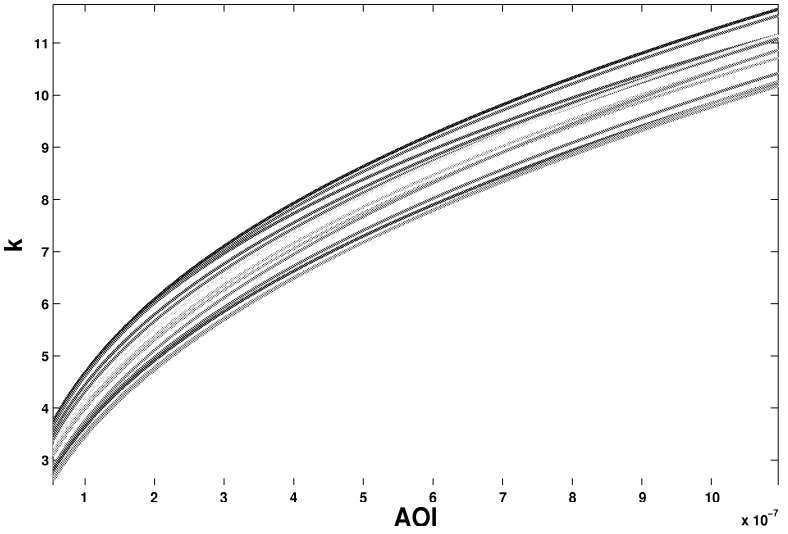


Fig. 3 Node degree, (k), Eq. 1, versus Average Overlapping Index, (AOI) Eq. 4 for the Watts-Strogatz model . The estimated parameters for Eq. 8 as well as other fitting statistics are shown in Table 2. The rewiring probability of the networks increases as the AOI decreases.

Table 2 Several rewired probabilities for the Watts-Strogatz model (Column 1). Parameter estimated for Eq. 8 with its t-student confidence levels (Ec. 10) (Columns 2 y 3). Values for the R^2 (Ec. 9) for each fit (Column 4). The higher the rewired probability, the poorer the goodness of fit is.

p_c	α	β	R^2
0.02	1960 (1738, 2182)	0.3743 (0.3674, 0.3811)	0.905
0.05	1848 (1634, 2062)	0.3713 (0.3643, 0.3784)	0.9002
0.1	2185 (1949, 2421)	0.3828 (0.3761, 0.3894)	0.9158
0.2	2179 (1942, 2417)	0.384 (0.3772, 0.3908)	0.9174
0.3	2697 (2403, 2990)	0.3994 (0.3925, 0.4062)	0.9244
0.4	2656 (2379, 2932)	0.4004 (0.3938, 0.4071)	0.9301
0.5	3297 (2967, 3627)	0.4159 (0.4095, 0.4224)	0.9385
0.6	3481 (3141, 3821)	0.4213 (0.415, 0.4277)	0.9422
0.7	3457 (3107, 3807)	0.4222 (0.4155, 0.4288)	0.9386
0.8	4600 (4159, 5041)	0.4427 (0.4364, 0.4491)	0.9463
0.9	4680 (4261, 5099)	0.446 (0.44, 0.452)	0.9513
0.95	4694 (4264, 5124)	0.4466 (0.4405, 0.4528)	0.9502
0.98	4792 (4358, 5226)	0.4483 (0.4422, 0.4544)	0.9499

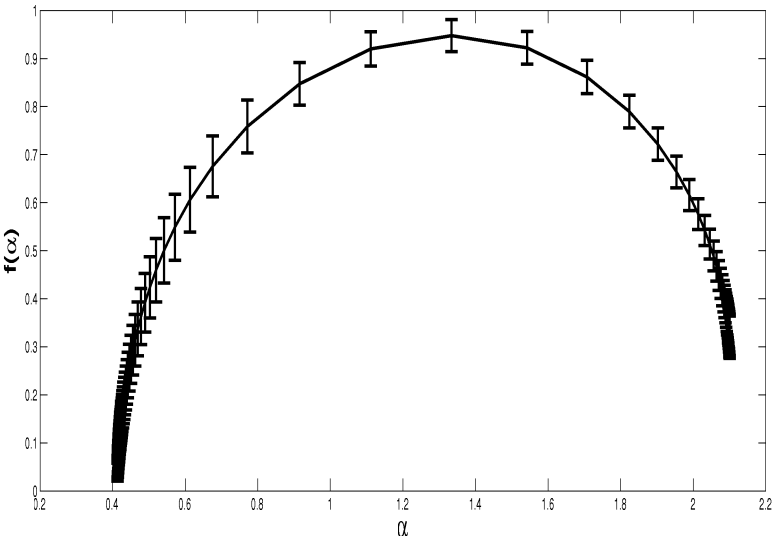


Fig. 4 Legendre spectrum for the AOI estimated over 1000 realization of the Watts-Strogatz model with 10000 each one and rewiring probability 0.7. The support of the measure has fractal dimension as $D_0 = 0.9609$. The spectrum is not quite symmetric, indicating the generation mechanism, which privileges the connection of a new node according its own rewiring probability as it was explained in Section 1.

Table 3 Several average node degree probabilities for the Albert-Barabási model (Column 1). Parameter estimated for Eq. 8 with its t-student confidence levels (Ec. 10) (Columns 2 y 3). Values for the R^2 (Ec. 9) for each fit (Column 4).The goodness of fit seems to remain constant while varying the average node degree.

m	α	β	R^2
2	4937 (4718, 5156)	0.4811 (0.4771, 0.4851)	0.9585
3	3826 (3660, 3992)	0.4641 (0.4601, 0.4682)	0.9413
4	4736 (4496, 4975)	0.4762 (0.4718, 0.4807)	0.948
5	4514 (4319, 4709)	0.4719 (0.468, 0.4757)	0.9555
6	3706 (3572, 3840)	0.4674 (0.4638, 0.4711)	0.9534

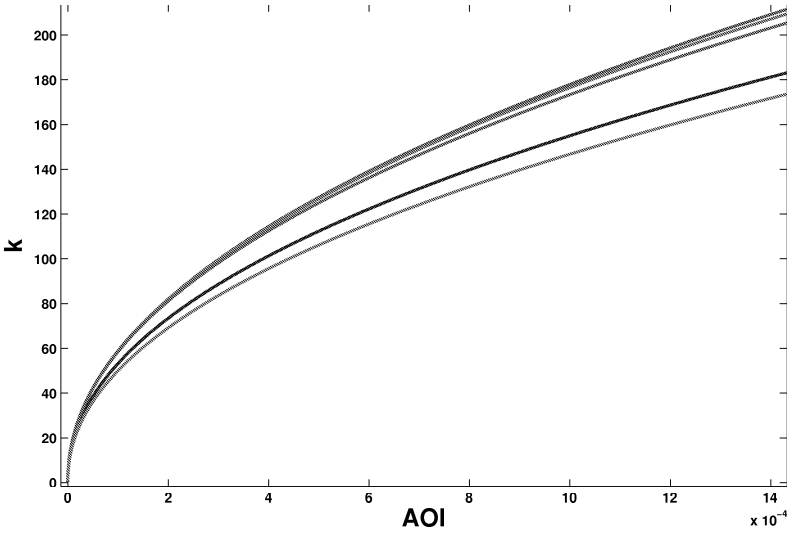


Fig. 5 Node degree, (k), Eq. 1, versus Average Overlapping Index, (AOI) Eq. 4 for the Albert-Barabási model. The estimated parameters for Eq. 8 as well as other fitting statistics are shown in Table 3. The node degree of the networks increases as the AOI decreases.

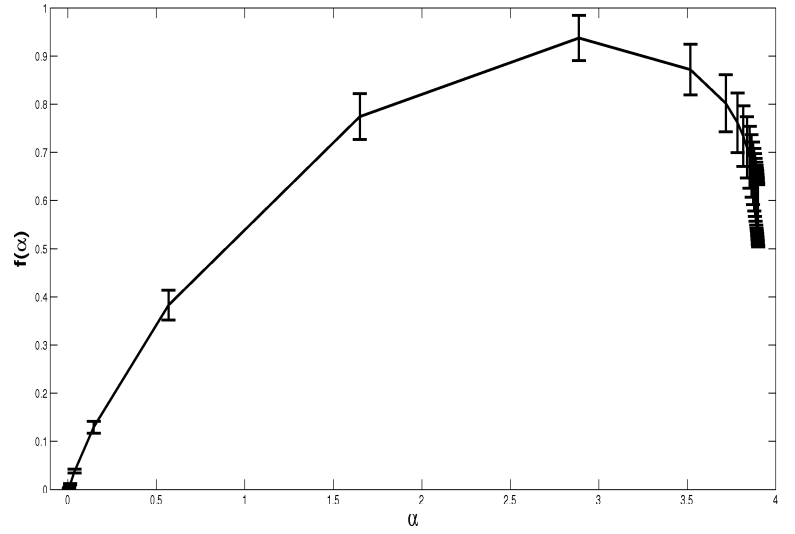


Fig. 6 Legendre spectrum for the AOI calculated over 1.000 realization of the Albert Barabási model with 10.000 nodes. The support of the measure has fractal dimension as $D_0 = 0.9391$. The spectrum is not symmetric at all and it is because there is a connection pattern privileging the growth of the AOI for the most connected nodes.

3 Conclusions

In Fig. 1,3 and 5 as well as in Tables 1,2 and 3 it is shown that there is a power law relationship (Eq. 8) between the AOI (Eq. 4) and the Node Degree (Eq. 1). The strength of this relationship can be quantified by the R^2 (Eq. 9 and the 4-th column of each Table 1,2 and 3), the higher R^2 appears in the Erdös-Rényi model and the weakest is in the Albert and Barabasi model. It is also shown that the parameters α and β in Eq. 8 depend on the network type. These facts suggests that although the Average overlapping index may be used to characterize complex network in the same way as the node degree, as the R^2 values may suggest, its relation make weaker when the networks model generation becomes more complex.

The fractal and multifractal properties of the AOI set have been also quantified by means of the Legendre spectrum. The AOI set for all simulated networks presents multifractal properties (Fig. 2,4 and 6) independently of the parameters used for the respective network generation model as was explained in the subsection 1.2. The fact the AOI set within each type of network is similar and between each type of networks is different indicates that the AOI values can grasp the internal generation mechanism of a complex network. As a network generation evolves, several regions of the AOI set domain becomes forbidden, and this phenomenon seems to be the same within each network model, as their corresponding Legendre spectrum shown. For example, in the Albert-Barabasi model a new node has much more probability to be linked to a node with a high node degree than to a node with a low node degree; so the average overlapping index of the nodes with high degree grows and remains constant for the nodes with low degree, for that reason the average overlapping index of the node having a high degree behaves different from the nodes having a low degree. Similar situation holds for the W-S network, in that case the rewiring probability dictates the growth dynamics and the average overlapping index behavior and in the Erdös-Rényi model the connection is the responsible for the average overlapping index ongoing.

The support of the average overlapping index in the networks simulated is also fractal since its $D_0 < 1$, i.e. 0.9802, 0.9609 and 0.9391 for Erdös-Rényi , Watts-Strogatz and Albert-Barabasi model, respectively.

In summary, the proposed *ansatz* (Eq. 1) was verified using simulated complex networks and for that reason a power law, relating the average overlapping index and the node degree was laid. The multifractal properties of the average overlapping index was established for the Erdös-Rényi , Watts-Strogatz and Albert-Barabasi model. This multifractal characteristic indicates that the AOI can grasp the internal generation mechanism of each complex network model.

References

1. Gligor, M., Ausloos, M.: Eur. Phys. J. B 63, 533–539 (2008)
2. Redelico, F.O., Proto, A.N., Ausloos, M.: Physica A
3. Redelico, F.O., Clippe, P., Proto, A.N., Ausloos, M.: Journal of Physics Series C (2010)
4. Redelico, F.O., Proto, A.N.: Int. J. Biffurcation and Chaos (2010)
5. Rotundo, G., Ausloos, M.: Physica A. 389(23), 5479–5494 (2010)
6. Maharaj, E.A., D'Urso, P.: Physica A 389(17), 3516–3537 (2010)
7. Shen, Z., Wang, Q., Shen, Y., Jin, J., Lin, Y.: Proceedings of IEEE International Instrumentation and Measurement Technology Conference, I2MTC 2010, art. no. 5488210, pp. 600–604 (2010)
8. Newman, M.: SIAM Review 45, 167 (2003)
9. Albert, R., Barabasi, A.-L.: Reviews of Modern Physics 74, 47–97 (2002)
10. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Physics Reports 424, 4–5, 175–308 (2006)
11. Wang, B., Tang, H., Guo, C., Xiu, Z.: Physica A 363(2), 591–596 (2005)
12. Demetrius, L., Manke, T.: Physica A 346(3-4), 682–696 (2004)
13. Erdös, P., Rényi, A.: On random graphs. Publ. Math. Debrecen 6, 290–297 (1959)
14. Bollobas, B.: Random Graphs. Academic, London (1985)
15. Milgram, S.: Psychology Today 2, 60–67 (1967)
16. Watts, D.J., Strogatz, S.H.: Nature 393, 440–442 (1998)
17. Barabási, A.-L., Albert, R.: Science 286, 509–512 (1999)
18. Batagelj, V., Brandes, U.: Phys. Rev. E 71, 036113 (2005)
19. Cantor, G.: On the Power of Perfect Sets of Points (De la puissance des ensembles parfait de points). Acta Mathematica 4, 381–392 (1884); English translation reprinted in Classics on Fractals, Edgar, G.A.(ed.) Addison-Wesley (1993) ISBN 0-201-58701-7
20. Falconer, K.: Fractal geometry: mathematical foundations and applications. Wiley (2003)
21. Paladin, G., Vulpiani, A.: Physics Reports 156, 147 (1987)
22. Benzi, R., Paladin, G.M., Parisi, G., Vulpiani, A.: J. Phys. A: Math. Gen. 19, 823 (1983)
23. Halsey, T.C., Jensen, M.H., Kadanoff, L.P., Procaccia, I., Shraiman, B.I.: Phys Rev. A 33, 1141 (1986)
24. Clinton Sprott, J.: Chaos and time series analysis. Oxford University Press (2008)
25. Ghanen, R., Spanos, P.D.: Stochastics Finite Element: A Spectral approach. Springer, NY (1991)
26. Casella, G., Berger: Statistical Inference Duxbury. Thomson Learnig (2002)
27. Falconer, K., Wiley, J.: Fractal Geometry - Mathematical Foundations and Applications, 2nd edn. (2003)

Taxicab Non Symmetrical Correspondence Analysis for the Evaluation of the Passenger Satisfaction

Biagio Simonetti and Antonio Lucadamo

Dipartimento di Studi dei Sistemi Economici, Giuridici e Sociali
Università del Sannio
{simonetti,antonio.lucadamo}@unisannio.it

Abstract. Taxicab Non Symmetrical Correspondence Analysis (TNSCA) is a technique which is more robust than the ordinary Non Symmetrical Correspondence Analysis (NSCA). TNSCA is a variant of the classical Correspondence Analysis (CA) for analyzing the two-way contingency table with a structure of dependence between two variables. In order to overcome the influence due to the presence of the outlier, TNSCA gives uniform weights to all points based on the taxicab singular value decomposition. The visual map constructed by TNSCA offers a clearer perspective than that obtained by correspondence analysis and it may be very useful in evaluating the satisfaction of public transportation passengers.

Keywords: Taxicab non symmetrical correspondence analysis, taxicab singular values decomposition, L_1 norm, robustness, passenger satisfaction.

1 Introduction

The diffusion of ISO 9001, its' certification and the adoption of the principles defined in the Charter of the services of Transport sector led to an increasing number of Public Transport companies, activating procedures for the assessment of quality control. Furthermore as people are more mobile, they expect a higher level of performance and quality of service regarding public transportation. For these reasons, public transportation companies must tailor their service to the desires and needs of their actual or potential customers. It is then necessary to transition from a system oriented to the production of services conditioned in the achievement of wrong objectives, toward a system which is oriented in the sale of the service.

The immediate consequence gives greater attention to the quality of services and relative customer satisfaction. The concept of quality over the years however has been considerably modified as it has been closely bound that of customer satisfaction. For example, the American Society for Quality defines quality as:

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 175–184.
springerlink.com

"the characteristics of a product or service that bear on its ability to satisfy stated or implied needs". From that statement it is easy to see how some people believe that perception is reality: i.e. the perception of service versus the actual service. It is then necessary to turn attention to what is perceived, rather than what is supplied.

In recent years many authors have focused on studying the customers perception of the service provided by local public transportation. The new role of a company that evaluates the service it provides by means of customer opinion should be maintained by that company. The company should dedicate itself to the acceptance of customer criticism by using said data to continuously improve its service. (Negro G., 1996). Therefore, in the context of TPL, studying passengers' criticism (PS) about the quality of the service offered is a strategically important.

The problem is that measuring satisfaction is complex due to its subjective nature. Since the evaluation of passenger satisfaction depends on cognitive emotional and psychological aspects (Oliver, 1993), it is a latent variable and can be evaluated only through a set rubric. Consequently, the analysis of satisfaction is only achieved through appropriate statistical techniques. In this paper we will show how the Taxicab Non Symmetrical Correspondence Analysis is a useful technique to evaluate the findings given to some parts of a survey and the overall satisfaction perceived by the passengers or customers. The paper is organized as follows: section 2 consists of a technical review of NSCA and section 3 introduces the Taxicab Non Symmetrical Correspondence Analysis. Section 4 demonstrates how the TNSCA provides more interpretable results than the NSCA in evaluating passenger satisfaction and section 5 introduces other considerations.

2 Non Symmetrical Correspondence Analysis

Non Symmetrical Correspondence Analysis (D'Ambra, Lauro 1989) is a method for visualizing contingency tables, with an asymmetric relationship between two variables. For our purposes we consider it as a particular kind of reduced rank matrix approximation method derived from generalized singular value decomposition (SVD).

Consider a two-way contingency table *N* of dimension $I \times J$ according to *I* and *J* categories of variables **Y** (response) and **X** (predictor), respectively.

Denote the matrix of joint relative frequencies by $\mathbf{P} = \left(p_{ij}\right)$ so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$.

Let $p_{i\bullet} = \sum_{j=i}^{J} pij$ and $p_{\bullet j} = \sum_{i=i}^{I} pij$ be the i[th] marginal row proportion and the j[th] marginal column proportion respectively.

Suppose that the relationship between these two variables is such that the J columns are predictor variables and are used to predict the outcome of the I rows response.

Categories. Furthermore, let $\mathbf{\Pi} = \left(\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} \right)$ be the matrix of differences between the unconditional marginal prediction $p_{i\bullet}$ and the conditional prediction $\frac{p_{ij}}{p_{\bullet j}}$.

If, for all of the (i, j)[th] cells, there is a perfect lack of predictability of the rows given the column categories then $\pi_{ij} = 0$. This is equivalent to concluding that there is complete independence between the two variables. A more formal and more global measure of predictability can be made by calculating the tau index (Goodman and Kruskal, 1954; Light and Margolin, 1971):

$$\tau = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} \right)^2}{1 - \sum_{i=1}^{I} p_{i\bullet}^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\pi_{ij}^2}{1 - \sum_{i=1}^{I} p_{i\bullet}^2} = \frac{N_\tau}{1 - \sum_{i=1}^{I} p_{i\bullet}^2} \quad (1)$$

Here the tau numerator, N_τ, is the overall measure of predictability of the rows given the columns, while the denominator measures the overall error in prediction. This does not depend on the predictor categories. When all distributions of the predictor variable are identical to the overall marginal distribution, there is no relative increase in predictability so τ is zero. Therefore, the NSCA of two-way contingency tables involves decomposing N_τ to obtain optimal measures of dependance. This is achieved by applying singular value decomposition (SVD) on π_{ij} so that:

$$\pi_{ij} = \sum_{s=1}^{S} \lambda_s a_{is} b_{js} \quad (2)$$

where $S = \min(I, J) - 1$. The SVD (2) involves obtaining the measure λ_s which is the s[th] singular value of π_{ij}. Similarly, $\mathbf{a}_s$ and $\mathbf{b}_s$ are the orthonormal singular vectors in both not-weighted and weighted metric, respectively:

$$\sum_{i=1}^{I} a_{is} a_{is'} = \begin{cases} 1 & s = s' \\ 0 & s \neq s' \end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j} b_{is} b_{is'} = \begin{cases} 1 & s = s' \\ 0 & s \neq s' \end{cases}$$

Furthermore, it follows that N_τ can be expressed in terms of the singular values such that

$$N_\tau = \left\| \mathbf{\Pi} \right\|^2 = \sum_{s=1}^{S} \lambda_s^2 .$$

For the graphical representation of the asymmetric dependence between the variables, define the coordinates of the i[th] response category (row) and j[th] predictor category (column) for the s[th] dimension of a correspondence plot by

$$f_{is} = a_{is}\lambda_s \qquad g_{is} = b_{is}\lambda_s$$

Therefore, a predictor profile coordinate, g_{is} that is situated close to the origin indicates that the j[th] predictor category does not contribute to the predictability of the response variables. Similarly a predictor coordinate that lies at a distance from the origin will indicate that category as important for predicting the row categories. For more details on the theory and application of the classical technique of NSCA, refer to D'Ambra and Lauro(1989) and Kroonenberg and Lombardo (1999).

3 Taxicab Non Symmetrical Correspondence Analysis

The Taxicab Singular Values Decomposition (TSVD, Choulakian 2006) is a particular orthogonal decomposition based on L1-norm distance. This is also referred to as Manhattan or City-Block; or more colloquially as the taxicab norm because it is the distance that a car would drive in a city mapped out in square blocks (if there are no one-way streets).

Taxicab geometry, is essentially the study of an ideal city with all roads running horizontal or vertical as the roads must be used to get from point A to point B; thus, the normal Euclidean distance function in the plane needs to be J modified. The shortest distance from the origin to the point (1,1) is now 2 rather than $\sqrt{2}$. Therefore, taxicab geometry is the study of the geometry consisting of Euclidean points, lines, and angles in $\Re^2$ with the taxicab metric: $d[(x_1, y_1),(x_2, y_2)] = |x_2 - x_1| + |y_2 - y_1|$.

A good discussion of the properties of this geometry is given by Krause (1986) and Kay (2001).

TSVD shall be applied to the matrix $\mathbf{\Pi}$ containing the differences between the unconditional marginal prediction $p_{i\bullet}$ and the conditional prediction $\frac{p_{ij}}{p_{\bullet j}}$, as previously defined. Therefore, let the rank of $\mathbf{\Pi} = k$.

Denote an *I*-dimensional vector by $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_I)'$, and the quantity $\mathbf{\Pi v}$ to be the projection of the *I* row points of $\mathbf{\Pi}$ on **v.** Let $\mathbf{T}_J$ be the collection of all vectors of length *J* with coordinates ± 1.

The first principal axis $\mathbf{v}_1$ of the row points of $\mathbf{\Pi}$ is an element of $\mathbf{T}_J$ such that the taxicab norm of $\mathbf{\Pi v}$ is maximized:

$$\max_{v \in T_J} \left\| \mathbf{\Pi v} \right\|_1 = \left\| \mathbf{\Pi v}_1 \right\|_1$$

The first row factor score is

$$f_1 = \mathbf{\Pi v}_1 \tag{3}$$

to which is related the first taxicab measure of dispersion $\lambda_1 = \left\| \mathbf{\Pi v}_1 \right\|_1$

Following the same procedure, the first column scores are obtained by

$$\mathbf{g}_1 = \mathbf{\Pi}' \mathbf{u}_1 \tag{4}$$

defining with $\mathbf{u}_1 = \operatorname{sgn}(\mathbf{f}_1) \in \mathbf{T}_I$ where $\operatorname{sgn}(\cdot)$ is the coordinates wise sign function that assigns 1 if $(\cdot) > 0$ and -1 if $(\cdot) \leq 0$ and $\mathbf{T}_I$ is the collection of all vectors of length *I* with coordinates ± 1.

The relationship between $\mathbf{v}_1$ and $\mathbf{u}_1$ expresses the measure of dispersion λ_1 in the following way:

$$\lambda_1 = \left\| \mathbf{\Pi v}_1 \right\|_1 = \left\| \mathbf{f}_1 \right\|_1 = \mathbf{u}_1 \mathbf{f}_1 = \left\| \mathbf{\Pi}' \mathbf{v}_1 \right\|_1 = \left\| \mathbf{g}_1 \right\|_1 = \mathbf{v}_1 \mathbf{g}_1$$

To compute the second axis, a sequential procedure may be applied considering the residual data set:

$$\mathbf{\Pi}^{(1)} = \mathbf{\Pi} - \mathbf{f}_1 \mathbf{g}_1' / \lambda_1$$

where $\mathbf{f}_1 \mathbf{g}_1' / \lambda_1$ is the rank 1 reconstruction of $\mathbf{\Pi}$ matrix on the first principal axis.

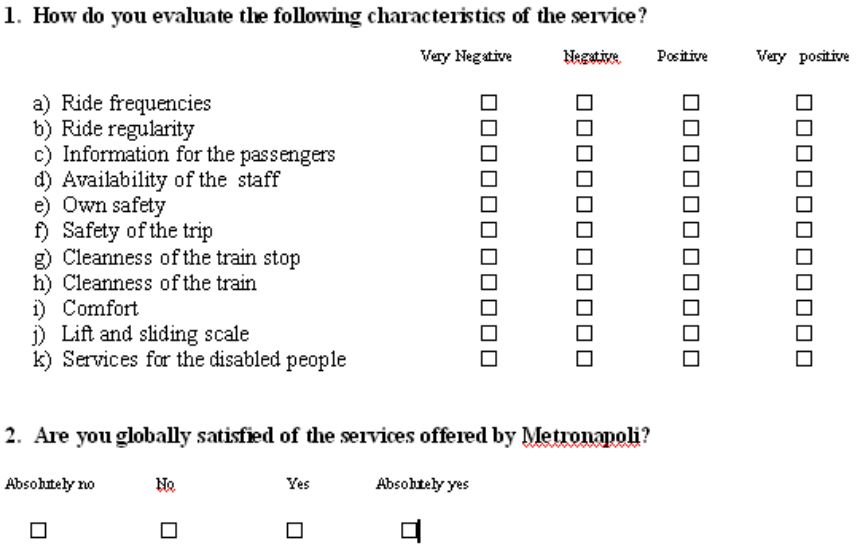
The described procedure can be applied for *k* times in order to compute the *k* principal axes.

It can be shown that after *k* iterations, the residual data matrix $\mathbf{\Pi}^{(k)}$ becomes zero.

4 Taxicab Non Symmetrical Correspondence Analysis for the Evaluation of the Passenger Satisfaction

The data that will be analyzed has been collected in a survey for studying the satisfaction of the passengers of Metronapoli, the firm that manages the metro of Naples. The collection has been carried out through a questionnaire given to travelers, including questions structured for the evaluation of the different features of the service (see layout in figure 1). Such services include racing frequency, racing regularity, staff society, travel security, personal and patrimonial safety, services for disabled and overall satisfaction for the service provided. The

questionnaire was filled out by approximately 2000 service users. For each question, there were four possible answers: very negative evaluation, negative evaluation, positive evaluation and very positive evaluation.



1. How do you evaluate the following characteristics of the service?

	Very Negative	Negative	Positive	Very positive
a) Ride frequencies	☐	☐	☐	☐
b) Ride regularity	☐	☐	☐	☐
c) Information for the passengers	☐	☐	☐	☐
d) Availability of the staff	☐	☐	☐	☐
e) Own safety	☐	☐	☐	☐
f) Safety of the trip	☐	☐	☐	☐
g) Cleanness of the train stop	☐	☐	☐	☐
h) Cleanness of the train	☐	☐	☐	☐
i) Comfort	☐	☐	☐	☐
j) Lift and sliding scale	☐	☐	☐	☐
k) Services for the disabled people	☐	☐	☐	☐

2. Are you globally satisfied of the services offered by Metronapoli?

Absolutely no	No	Yes	Absolutely yes
☐	☐	☐	☐

Fig. 1 Questionnaire filled out by the passengers of Metronapoli

The first theory is to study the relationships that may exist between some aspects of the service with regard to complete satisfaction. Attention here has been focused on a particular aspect: the perceived security on the metro.

Of course the Correspondence Analysis is the first analysis that can be applied. This permits the visualization of the association between the modalities of the two considered variables. The contingency table may be useful as a snapshot of the situation as well as to build the results of the analysis.

Table 1 Contingency table (considered variables: security satisfaction and overall satisfaction)

	ov_1	ov_2	ov_3	ov_4	Total
sec_1	9	38	30	6	83
sec_2	19	102	167	13	301
sec_3	13	135	835	103	1086
sec_4	12	55	579	352	998
Total	53	330	1611	474	1994

Interesting here is the fact that many passengers who give a rating of "4" to the perception of security have a lower overall satisfaction rating. However many users that were "moderately" satisfied with security had an higher overall satisfaction rating. In order to visualize whether or not there is a good association between the two variables, it is wise to consider the following graph:

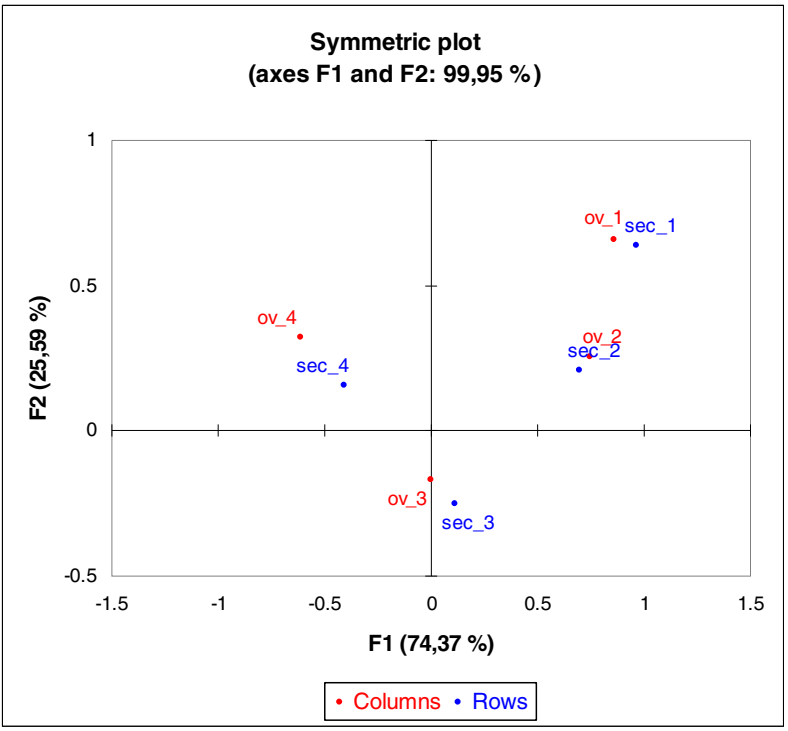


Fig. 2 First two factors of the Correspondence Analysis (security and overall satisfaction)

It is easy to see that there is indeed a strong relationship between the opinions rendered regarding overall customer satisfaction (ov_) and those regarding the evaluation of metro security (sec_). The problem is that the Guttman effect is clear. In fact if the first quadrant is considered, then it is clear that the customers who gave negative responses to the subway survey rendered negative or very negative ratings for both security and overall satisfaction. However, the fourth quadrant contains positive evaluations and the second quadrant contains very positive ratings. Therefore, if the information contained in the first axis is considered, then the difference between satisfied and unsatisfied customers is clear. If however, the information in the second axis is considered, then the analysis lacks evidence. For this reason the NSCA may serve as a solution.

It is possible to note the asymmetric relationship between the two variables in the data analyzed: the judgment with regard to a single aspect of service and the influence of that one item on the overall customer satisfaction. The use of this technique overcomes the problem underlined in the previous graph. There is, however, a very strange situation that presents itself in the following figure. Here, the overall ratings are all located very close to the origin of the axes. In this case it is not possible to underline the links between the two modalities.

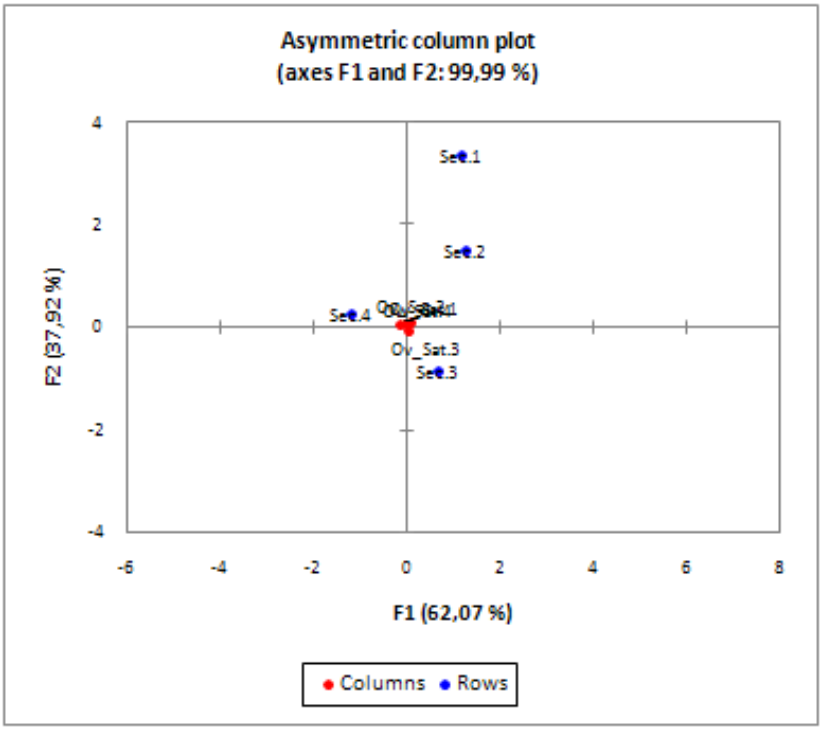


Fig. 3 First two factors of the Non Symmetrical Correspondence Analysis (security and overall satisfaction)

Wherefore as the NSCA is a more appropriate technique for this type of study rather than the CA, it still yields negative results. In such circumstances, it is our opinion that the TNSCA is the better procedure to use in order to visualize the relationship of the data:

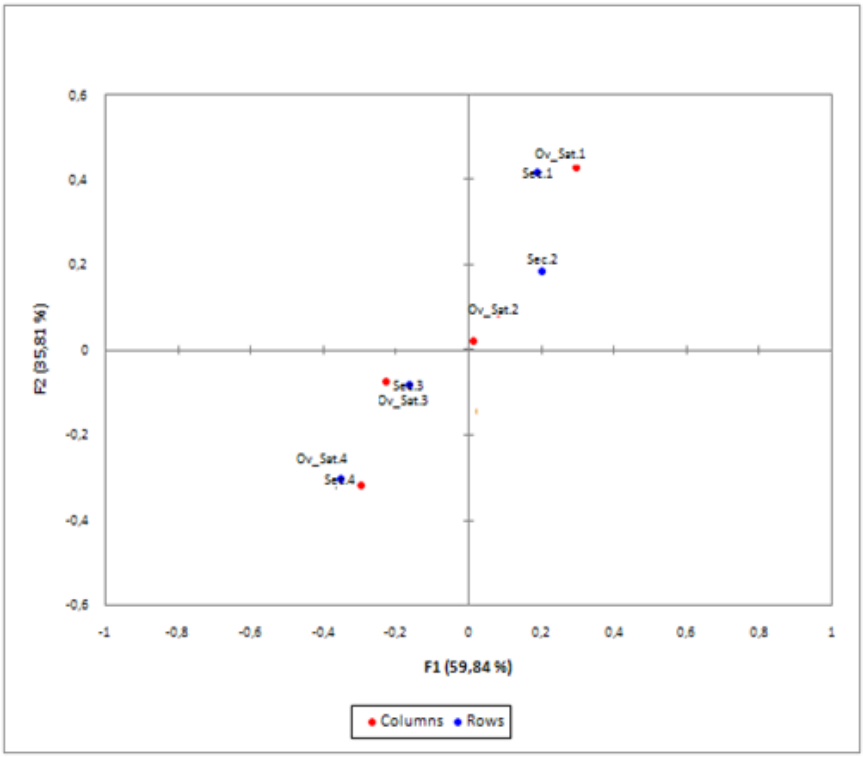


Fig. 4 First two factors of the Taxicab Non Symmetrical Correspondence Analysis (security and overall satisfaction)

Here it is clear that there is a perfect relationship between the customer response given regarding security to the customer overall service satisfaction data received. Furthermore it is possible to underline how all the modalities are positioned in a linear fashion. There is a total separation between satisfied and unsatisfied passengers. In the first quadrant are the people that gave low scores for both security and overall service. In the third quadrant there are people that are satisfied with said service. Considering the two axes separately, it is evident that on the left of the first axis are those people who are satisfied while those who were not satisfied are on the right. The same result exists for the second axis from bottom to top.

5 Conclusions

Overall customer satisfaction and customer satisfaction in specific cases are both important aspects that public transportation firms have recently studied. They understood that it was necessary to focus on quality control and this is measured against the level of perceived customer satisfaction. In order to correctly analyse

customer satisfaction and simultaneously find means of improvement, it is necessary to use appropriate statistical techniques. This paper has demonstrated how the TNSCA can be a useful tool to highlight the relationships between specific services and overall customer satisfaction. The CA (while giving some insight) does not provide results which are readily understood. The NSCA (although seemingly offering correct results) falls short of the proper solution. Only TNSCA provides solid reliable results.

References

Benzécri, J.-P., et al.: Analyse des Données, vol. 2. Dunod, Paris (1973)

Bradley, R.A., Katti, S.K., Coons, I.J.: Optimal scaling for ordered categories. Psychometrika 27, 355–374 (1962)

Choulakian, V.: Taxicab correspondence analysis. Psychometrika 71(2), 1–13 (2006)

D'Ambra, L., Lauro, N.C.: Non-Symmetrical Correspondence Analysis for three-way contingency table. In: Coppi, R., Bolasco, S. (eds.) Multiway Data Analysis, pp. 301–315. Elsevier, Amsterdam (1989)

D'Ambra, L., Lombardo, R.: Normalized Non Symmetrical Correspondence Analysis for three-way data sets. Bulletin of The International Statistical Institute, 49th Session Book 1, 301–302 (1993)

Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Journal of the American Statistical Association 49, 732–764 (1954)

Kay, D.C.: College Geometry: A Discovery Approach, 2nd edn. Addison Wesley Longman Inc., Boston (2001)

Krause, E.F.: Taxicab geometry: An adventure in non-Euclidean geometry. Dover, New York (1986)

Kroonenberg, P., Lombardo, R.: Non symmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research Journal 34, 367–397 (1999)

Light, R., Margolin, B.: An analysis of variance for categorical data. J. Am. Statisti. Assoc. 66, 534–544 (1971)

Negro, G.: Organizzare la qualità nei servizi, Il Sole 24 ore, Milano (1996)

Oliver, R.L.: A conceptual model of service quality and service satisfaction: compatible goals, different concepts. In: Advances in Service Marketing and Management: Research and Practice, vol. 2. JAI Press (1993)

Simonetti, B.: Taxicab Non-Symmetrical Correspondence Analysis. In: Proceedings of "MTISD 2008, Methods, Models and Information Technologies for Decision Support Systems", Università del Salento, Lecce, September 18-20, pp. 257–260 (2008)

Simulation Study for PLS Path Modelling with High-Order Construct: A Job Satisfaction Model Evidence

Enrico Ciavolino and Mariangela Nitti

Abstract. The aim of the paper is to present a study on the high-order latent variables for the partial least squares path modelling (PLS-PM).

A Monte Carlo simulation study is proposed for comparing the performances of the two best-known methods for modelling higher-order constructs, namely the *repeated indicators* and the *two-step* approaches. The simulation results, far from covering all the potential uses of the two approaches, could provide some useful suggestions to those researchers who are intending to introduce a further level of abstraction in modelling the phenomenon of interest.

An illustrative case study on the job satisfaction is reported in order to show how theoretical and statistical instances have to be taken into consideration when modelling higher-order constructs.

Keywords: Partial least squares, high-order latent variables, structural equation model, Monte Carlo simulation, job satisfaction.

1 Introduction

Partial least squares (PLS) enables researchers in many fields of social sciences to investigate models at a high level of abstraction. As a matter of fact, many concepts, in the psychological as well as the economic field, are defined as being composed of several different facets. Although each of these facets could be seen as separate, they are all different aspects of the same abstract concept. To exemplify, the *job satisfaction* construct is frequently defined as summarising different dimensions, including satisfaction with relationships with colleagues, salary, supervisors, advancement opportunities and so on. All these dimensions, rather than being separate, are all integral parts of a person's job satisfaction.

The dimensions of a higher-order construct could be then conceptualised under an overall abstraction, and it is theoretically meaningful to use this abstraction for the representation of the dimensions, instead of merely interrelating them.

In this paper, a high-order latent variables model for PLS path modelling (PLS-PM) is presented. Two model-building approaches (named repeated indicators and two-step approach) are compared through a Monte Carlo simulation study for

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 185–207.
springerlink.com

determining which method better represents the relationships among construct levels. The paper is organised as follows: in Section 2, the PLS estimation method for structural equation models and its extension to higher-order construct modelling are described; Section 3 presents the simulation study and draws conclusions on the performance of the two approaches presented; in Section 4, a job satisfaction model and its analysis show the application of the high-order modelling to a real case study; conclusions and remarks are given in Section 5.

2 PLS-PM for Higher-Order Constructs

Due to its ability of estimating complex models, PLS-PM can be used to investigate models with a high level of abstraction. The basic PLS design was completed for the first time in 1966 by Herman Wold [19] for the use in multivariate analysis, and subsequently extended for its application in the structural equation modelling (SEM) in 1975 by Wold himself [20]. An extensive review on PLS approach is given in [8]. The model-building procedure can be thought of as the analysis of two conceptually different models. A measurement (or outer) model specifies the relationship of the observed variables with their (hypothesised) underlying (latent) constructs; a structural (or inner) model then specifies the causal relationships among latent constructs, as posited by some theory. The two sub-models' equations are the following:

$$\boldsymbol{\xi}_{(m.1)} = \mathbf{B}_{(m,m)} \cdot \boldsymbol{\xi}_{(m.1)} + \boldsymbol{\zeta}_{(m,1)}$$
$$\mathbf{x}_{(p,1)} = \boldsymbol{\Lambda}_{(p,m)} \cdot \boldsymbol{\xi}_{(m.1)} + \boldsymbol{\delta}_{(p,1)}$$

where the subscripts m and p are the number of the latent variables (LV) and the manifest variables (MV) respectively in the model, while the letters $\boldsymbol{\xi}$, $\mathbf{x}$, $\mathbf{B}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ indicate LV and MV vectors, the path coefficients linking the LV, the factor loading linking the MV to the LV, and the error terms of the model.

2.1 The PLS Algorithm

The parameters estimation [5] is based on a double approximation of the LVs $\boldsymbol{\xi}_j$ (with j=1, ..., m). The *external estimation* $\mathbf{y}_j$ is obtained as the product of the block of MVs $\mathbf{X}_j$ (considered as the matrix units for variables) and the *outer weights* $\mathbf{w}_j$ (which represent the estimation of measurement coefficients, $\boldsymbol{\Lambda}$). The *internal estimation* $\mathbf{z}_j$ is obtained as the product of the external estimation of $\boldsymbol{\xi}_j$, $\mathbf{y}_j$ and the *inner weights* e_j.

According to the relationship among MVs and LVs hypothesised, outer weights are computed as:

$$\mathbf{w}_j = \mathbf{X}_j' \, \mathbf{z}_j$$

for Mode A (reflective relationship), and:

$$\mathbf{w}_j = (\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j' \cdot \mathbf{z}_j$$

for Mode B (formative relationship).

The *inner weights* $e_{j,i}$, in the centroid scheme, are defined as the sign of the correlation between the connected estimated $\mathbf{y}_j$ and $\mathbf{y}_i$, with $i \neq j$.

The PLS algorithm starts by initialising outer weights to one for the first MV of each LV; then, the parameters estimation is performed, until convergence, by iteratively computing:

- *external estimation*, $\mathbf{y}_j = \mathbf{X}_j \ \mathbf{z}_j$;
- *internal estimation*, $\mathbf{z}_j = \sum_{j \neq i} e_{j,i} \mathbf{y}_i$;
- *outer weights estimation*, with Mode A or B.

The causal paths among LVs (the coefficients in the **B** matrix) are obtained through ordinary least squares (OLS) method or PLS regression.

2.2 Modelling Higher-Order Constructs

Wold's original design of PLS-PM does not consider higher-order LVs; each construct has to be necessarily related to a set of observed variables in order to be estimated. On this basis, Lohmöller proposed a procedure for the case of hierarchical constructs, the so-called *hierarchical component model* or *repeated indicators approach* [13], which is the most popular approach when estimating higher-order constructs through PLS. The procedure is very simple: 'a second-order factor is directly measured by observed variables for all the first-order factors. While this approach repeats the number of MVs used, the model can be estimated by the standard PLS algorithm' [17]. The manifest indicators are repeated in order to represent the higher-order construct. A prerequisite for the repeated indicators approach is that all indicators of the first-order and the second-order factors should be reflective [15].

This approach is the most favoured by researchers when using PLS for modelling higher-order constructs. A disadvantage of this approach is a possible biasing of the estimates by relating variables of the same type together through PLS estimation, as shown in the remainder of the paper

Another way of building a higher-order model is the *two-step approach*: the LV scores are initially estimated in a model without second-order constructs [1]. The LV scores are subsequently used as indicators in a separate higher-order structural model analysis. It may offer advantages when estimating higher-order models with formative indicators [7], [17]. The implementation is not one simultaneous PLS run. A clear disadvantage of such a two-stage approach is that any construct which is investigated in stage two is not taken into account when estimating LV scores at stage one. The two approaches are illustrated in Figure 1.

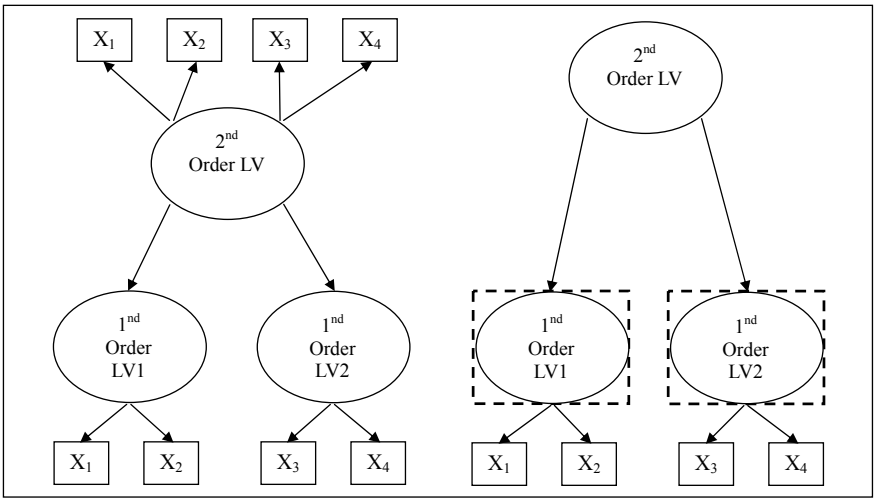


Fig. 1 Model building: Repeated indicators and two-step approaches

3 The Simulation Study

The object of this simulation is to compare the performance of different designs of models, using several number of items, with a balanced or unbalanced design of items. The study takes into account also different sample sizes, in order to understand the effect of the sample dimension. The performances are evaluated by means of the prediction accuracy, the estimate bias and the efficiency of the considered approaches. The following paragraphs report the simulation plan and the comments on the results obtained.

3.1 The Simulation Plan

The Monte Carlo simulation is structured along the following steps: at the *first step*, we define the structure of the model and the parameters of the population. In the *second step*, we generate randomly the second-order LV and given the parameters and the error terms we estimate the first-order LVs. According to the outer parameters and error terms, in the *last step*, we generate the first and second-order MVs.

The underlying population model used for the simulation consisted of one second-order LV (denoted by ξ_1^{II}) and three first-order LVs (denoted by ξ_1^{I}, ξ_2^{I} and ξ_3^{I}), as shown by Figure 2.

In the figure, for simplicity, are reported only the LVs, because the MVs will change according to the different simulation hypothesis, but in all the cases they are linked to the LVs in a reflective way. The relationship between first and second-order LVs is also modelled as reflective, so that the LVs of the lower level could be seen as generated by the construct at the higher level.

The two approaches performances have been compared on the basis of sample size (n = 50, 100, 300, 1000), model structure (balanced and unbalanced) and number of indicators per construct. The items are set to 2, 4 and 6 for each construct in the balanced designs; the two unbalanced models, called U_1 and U_2, have respectively 2, 4, 6 and 6, 8, 10 items per construct. The study design considers 500 replications for each condition. Both unbalanced designs have been simulated by changing the sequence of the constructs, in order to detect any differences due to their position in the model. No difference was found in the simulation results depending on the LVs sorting.

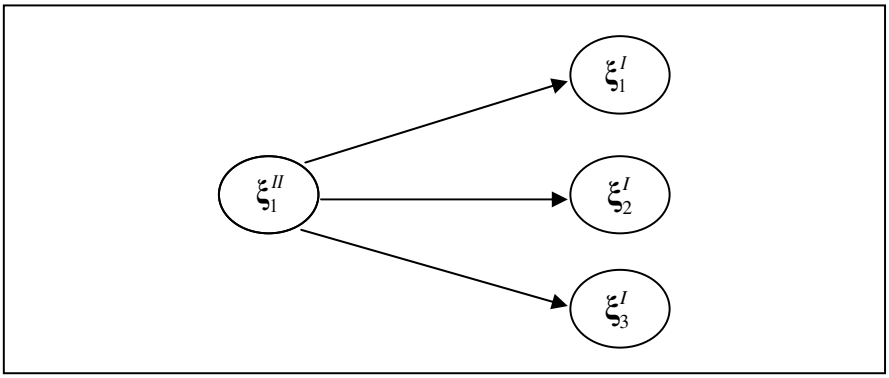


Fig. 2 Path diagram for high order

The path coefficients vector ($\boldsymbol{\beta}$) of the structural model is assumed to have the elements equal to 0.8, while all the elements of the vector loadings ($\boldsymbol{\lambda}^I$ and $\boldsymbol{\lambda}^{II}$) for the measurement models, are set at 0.7 with the exception of $\boldsymbol{\lambda}^{II}$ for the repeated indicators, as explained later in this section.

The value 0.7 leads to reliable measurement models (see Table 1): the median value of Cronbach's alpha for each construct ranges from 0.592 (for 2 indicators per construct) to 0.880 (for 10 indicators within the second unbalanced model), composite reliability from 0.824 (2 indicators) to 0.908 (for the second-order LV within the second unbalanced model) and average variance extracted (AVE) from 0.488 (10 indicators within the second unbalanced model) to 0.714 (2 indicators). Moreover these values allow comparisons with previous Monte Carlo simulations ([4], [9]).

Table 1 Reliability measures

ITEMS PER CONSTRUCT	CRONBACH'S ALPHA				COMPOSITE RELIABILITY				AVE			
	ξ_1^I	ξ_2^I	ξ_3^I	ξ_1^{II}	ξ_1^I	ξ_2^I	ξ_3^I	ξ_1^{II}	ξ_1^I	ξ_2^I	ξ_3^I	ξ_1^{II}
2	0.596	0.592	0.595	0.708	0.833	0.831	0.832	0.824	0.714	0.712	0.713	0.534
4	0.745	0.745	0.745	0.801	0.840	0.840	0.841	0.872	0.596	0.595	0.596	0.554
6	0.815	0.813	0.814	0.841	0.867	0.866	0.867	0.890	0.526	0.522	0.524	0.564
2;4;6	0.595	0.744	0.813	0.798	0.832	0.840	0.866	0.869	0.713	0.571	0.523	0.550
6;8;10	0.814	0.854	0.880	0.868	0.867	0.887	0.903	0.908	0.525	0.501	0.488	0.571

For each design condition, simulated data sets were built up with 500 replications; the starting point was the generation of the second-order LV (ξ_1^{II}) as a random variable $\xi_1^{II} \sim N(0,1)$. The generated data were re-scaled in the interval [1, 5], in order to reproduce a Likert-type scale.

Let us consider a sample size equal to *n* and the matrix formulation of the LVs ($\xi_{(n,1)}^{II}$ and $\xi_{(n,3)}^{I} = [\xi_1^I \mid \xi_2^I \mid \xi_3^I]$); the first-order LVs $\xi_{(n,3)}^{I}$ have been computed as the product of $\xi_{(n,1)}^{II}$ by the path coefficients vector $\boldsymbol{\beta}_{(1,3)}$, with the addition of an error component $\zeta_{(n,3)}$, according to the following equation:

$$\xi_{(n,3)}^{I} = \xi_{(n,1)}^{II} \cdot \boldsymbol{\beta}_{(1,3)} + \zeta_{(n,3)}$$

Each vector of error component ζ_j is drawn from a univariate normal distribution [10] with mean equal to zero and standard deviation, var(ζ_j), chosen for satisfying, per each j^{th} first-order LV, the equation:

$$R_j^2 = \frac{\text{var}(\text{model}_j)}{\text{var}(\text{total}_j)} = \frac{\text{var}(\text{model}_j)}{\text{var}(\text{model}_j) + \text{var}(\text{error}_j)},$$

where var(total$_j$) is the variance of $\xi_{(n,j)}^{I}$, given that:

$$\xi_{(n,j)}^{I} = \xi_{(n,1)}^{II} \cdot \beta_{(1,j)} + \zeta_{(n,j)} = \text{model}_j + \text{error}_j, \text{ (per each } j\text{)},$$

and var(ζ_j) is:

$$\text{var}(\zeta_j) = \frac{\text{var}(\xi_j^I) \cdot (1 - R_j^2)}{R_j^2}$$

The R^2 values for the three first-order LVs were set to 0.8. The values for the observed items were computed from the first-order LVs, with, as already defined above, the coefficients all set equal to 0.7 plus an error term distributed as a continuous uniform: $\boldsymbol{\delta} \sim U(-1,1)$. MVs are generated starting from the LVs, given the lambda coefficients, following the formula:

$$\mathbf{X}_{(n,k)} = \boldsymbol{\xi}^{I}_{(n,1)} \cdot \left(\boldsymbol{\lambda}^{I}\right)^{-1}_{(1,k)} + \boldsymbol{\delta}_{(n,k)},$$

where k is the number of MVs related to each LV. All the items $\mathbf{X}_{(m,n)}$ were standardised, having a mean of zero and a standard deviation of one.

Finally, for the special case of the repeated indicator approach, where items of the first-order are replicated at the second level, the population weights were obtained according to the formula:

$$\boldsymbol{\lambda}^{II}_{(1,q)} = (\boldsymbol{\xi}^{II})^{-1}_{(1,n)} \cdot (\mathbf{X}_{(n,q)} - \boldsymbol{\delta}_{(n,q)}),$$

where q is the number of the MVs at the first level, replicated for constructing the second-order LV.

Per each run and in each simulation condition, both the repeated indicators and two-step approaches were used to estimate inner and outer parameters of the model. The inner weighting scheme used is the factor scheme and the outer weights are estimated by assuming reflective relationships (Mode A).

Per each replication the following outcomes are measured:

- Inner coefficients, with the standard deviation and t-values.
- R^2 and R^2 adjusted for the first-order LVs.
- The squared correlation between the predicted LV scores and the defined true scores.
- The relative bias.
- The relative efficiency.

These outcomes are reported and discussed in the next paragraphs in order to summarise and compare the performances of the two methods in terms of prediction accuracy and parameter efficiency and bias.

3.2 The Path Coefficients

Table 2 reports the simulation results about the coefficients β (computed as the average of the 500 replications), the standard errors and the statistical significance, given by the t-test. Results are grouped according to the estimation approach used (reported in the first column), simulation sample size (second column), and number of items per construct (the last five columns). For each of these combinations, path coefficients, standard errors and resulting t-statistics of the three parameters (β_1, β_2, β_3) are reported in rows.

Table 2 Path coefficients, standard error and t-test for the inner model

APPROACH	SAMPLE SIZE	VALUES	ITEMS PER CONSTRUCT														
			2			4			6			2;4;6			6;8;10		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Repeated Indicators	50	path	0.808	0.811	0.806	0.847	0.849	0.850	0.863	0.863	0.862	0.724	0.839	0.916	0.823	0.868	0.902
		SE	0.052	0.051	0.053	0.040	0.039	0.038	0.036	0.034	0.036	0.069	0.039	0.022	0.043	0.033	0.024
		t-statistic	15.461	16.050	15.211	21.290	21.589	22.371	23.772	25.238	23.948	10.446	21.380	41.513	19.036	26.267	37.722
	100	path	0.792	0.794	0.792	0.836	0.834	0.835	0.854	0.852	0.854	0.706	0.827	0.906	0.815	0.860	0.898
		SE	0.041	0.040	0.040	0.032	0.032	0.030	0.027	0.029	0.028	0.052	0.033	0.019	0.034	0.026	0.019
		t-statistic	19.468	19.891	19.645	25.904	26.034	27.527	31.617	29.093	31.033	13.478	25.426	48.355	24.037	33.118	47.317
	300	path	0.773	0.773	0.775	0.819	0.820	0.819	0.841	0.839	0.840	0.682	0.811	0.897	0.804	0.850	0.890
		SE	0.028	0.029	0.028	0.021	0.022	0.021	0.017	0.019	0.017	0.035	0.023	0.014	0.020	0.016	0.012
		t-statistic	27.540	27.065	27.927	39.061	37.904	38.483	48.172	43.652	48.299	19.751	35.348	62.510	39.647	53.442	72.522
	1000	path	0.754	0.755	0.756	0.805	0.806	0.805	0.828	0.828	0.828	0.663	0.798	0.887	0.793	0.841	0.888
		SE	0.019	0.019	0.019	0.015	0.015	0.014	0.012	0.013	0.013	0.022	0.015	0.010	0.009	0.006	0.012
		t-statistic	40.137	38.818	40.770	53.154	53.699	55.778	68.299	64.783	65.613	30.674	52.607	92.343	88.419	148.230	72.044
Two-step approach	50	path	0.804	0.804	0.807	0.841	0.839	0.841	0.855	0.857	0.854	0.816	0.839	0.847	0.857	0.865	0.867
		SE	0.055	0.052	0.053	0.044	0.039	0.042	0.035	0.035	0.035	0.048	0.042	0.037	0.034	0.032	0.032
		t-statistic	14.639	15.329	15.135	19.124	21.335	19.912	24.197	24.580	24.385	16.910	19.800	22.957	25.253	27.039	27.238
	100	path	0.792	0.789	0.790	0.833	0.833	0.830	0.852	0.849	0.850	0.802	0.832	0.842	0.852	0.860	0.864
		SE	0.040	0.042	0.041	0.031	0.029	0.032	0.024	0.025	0.026	0.037	0.030	0.027	0.027	0.023	0.023
		t-statistic	19.934	19.008	19.140	26.680	28.645	26.139	35.385	33.796	33.306	21.582	27.448	30.667	31.177	36.651	37.861
	300	path	0.771	0.773	0.772	0.818	0.818	0.818	0.838	0.839	0.839	0.783	0.818	0.827	0.842	0.850	0.855
		SE	0.028	0.027	0.029	0.021	0.020	0.021	0.018	0.018	0.018	0.026	0.020	0.019	0.018	0.016	0.015
		t-statistic	27.804	28.935	26.762	38.248	39.943	38.129	46.802	47.437	47.652	29.714	40.236	43.932	47.489	53.801	58.879
	1000	path	0.755	0.756	0.755	0.805	0.805	0.805	0.828	0.828	0.828	0.798	0.830	0.842	0.832	0.841	0.847
		SE	0.019	0.018	0.019	0.014	0.014	0.014	0.012	0.012	0.012	0.015	0.011	0.009	0.012	0.011	0.010
		t-statistic	40.247	40.878	40.290	56.056	58.137	57.851	69.957	69.232	68.859	42.642	57.079	62.568	67.287	79.039	84.745

The estimated path coefficients are all significant, as we expected by the hypotheses made for defining the simulation plan. Some consideration can be made about t-values. The following Figures 3 and 4 report the t-test on the y-axis, and the sample size on the x-axis, respectively for the repeated indicators and the two-step approach.

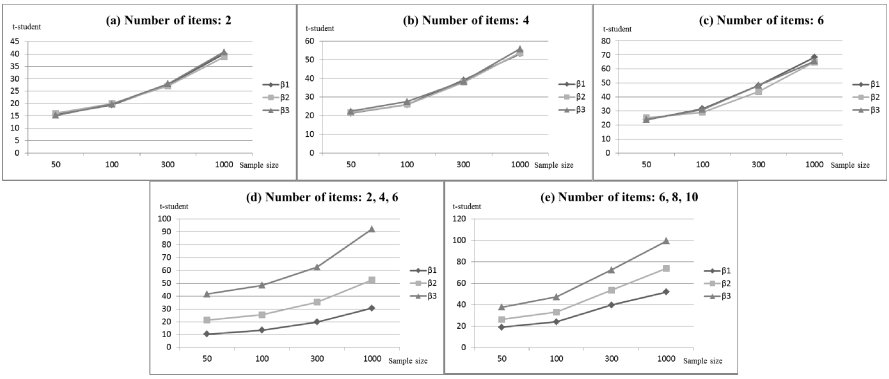


Fig. 3 T-test for the repeated indicator

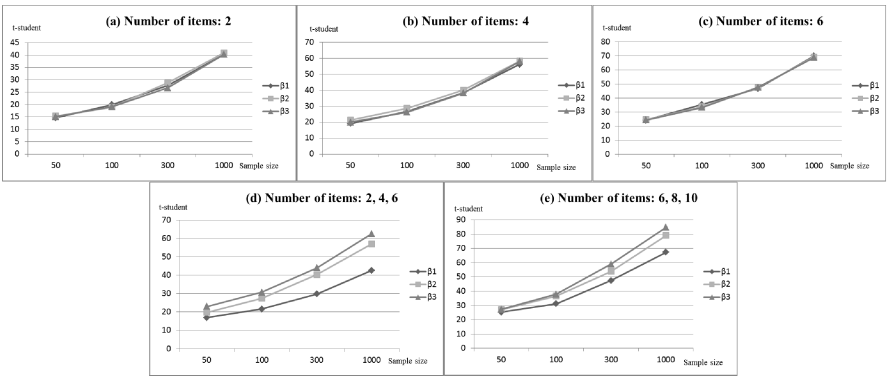


Fig. 4 T-test for the two-step approach

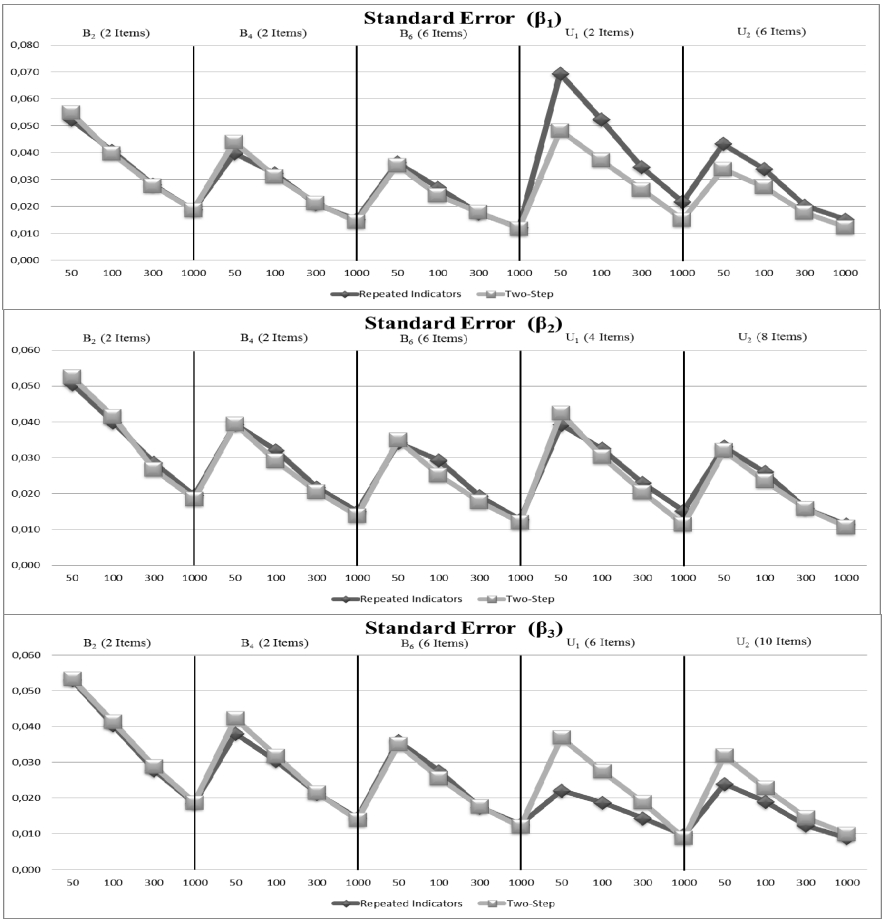


Fig. 5 Standard errors of the path coefficients

The five patterns of measurement models, with the number of items equal to 2 (a), 4 (b), 6 (c) and the two unbalanced designs (d) and (e) are the sections of the figure. It is interesting to notice that increasing the dimension of the sample, from 50 to 1000, the significance increase in a proportional way. Moreover, in the unbalanced designs, higher significances are obtained with a bigger number of items per LV, that is, 6 items, for β_3 in (d) and 10 items for β_3 in (e) lead to the highest t-tests.

Figure 4 illustrates the trend of the t-test for the two-step approach. The results can bring us to the same considerations made for the repeated indicators approach.

In both cases, significance rises more sharply when the design is unbalanced than when constructs have the same number of items.

The standard error of each estimated β is shown in Figure 5. When items per construct are equal in the same model (B_2, B_4, and B_6 designs), no differences are found between the two approaches. In the case of both unbalanced designs, variability in the esteems is lower when using the two-step approach, with the minimum number of items per constructs. With respect to the construct with the bigger number of items, the repeated indicators model produces more robust estimations. The two approaches are identical with the medium number of items in the unbalanced designs.

3.3 LVs Prediction Accuracy

To evaluate the prediction accuracy of both methods, we have calculated the amount of variance explained for the first-order LVs, and the squared correlation between predicted LVs score and the defined true scores. Table 3 reports in the first three columns respectively the first-order LVs, the sample size dimension and the approach used. The last section of the table reports the number of items per construct. The compared values are the $\overline{R}^2$, for the repeated indicators, and the $\overline{\lambda}^2$ for the two-step approach. Throughout the paper we have always talked about the path coefficients as β, just to have a similar definition of the path coefficients and to avoid confusion of terms. In truth, the coefficient between the first-order LVs and the second-order LV, for the two-step approach, are lambdas (λ). The square of these values, in the case of standardised data, reflects the percentage of explained variance.

In this way we can compare the average of the explained variance ($\overline{R}^2$ and $\overline{\lambda}^2$), over the 500 replications, of two methods. It is possible to notice that there is no relevant difference between the methods, except for the cases of the unbalanced designs.

In these cases, reported in detail in Figures 6 and 7, the two-step approach has a percentage of explained variance higher than the repeated indicator for the LV with the smallest number of items (2 items for U_1 and 6 items for U_2). Increasing the number of items per construct, the explained variance tends to be higher for the repeated indicators (specifically, when items per LV are 6 and 10 within, respectively, U_1 and U_2).

It is worth a reminder that the percentage of explained variance concerns the inner model, which means the structural relationships between LVs.

Table 3 $\overline{R}^2$ and $\overline{\lambda}^2$ for the first-order LVs

LV	SAMPLE SIZE	APPROACH	ITEMS PER CONSTRUCT				
			2	4	6	2;4;6	6;8;10
ξ_1^I	50	*Repeated Indicator*	0.648	0.713	0.741	0.519	0.673
		Two-step approach	0.647	0.708	0.730	0.582	0.735
	100	*Repeated Indicator*	0.626	0.697	0.728	0.496	0.662
		Two-step approach	0.627	0.694	0.726	0.643	0.725
	300	*Repeated Indicator*	0.597	0.669	0.669	0.464	0.673
		Two-step approach	0.595	0.669	0.703	0.613	0.709
	1000	*Repeated Indicator*	0.569	0.648	0.686	0.439	0.628
		Two-step approach	0.570	0.648	0.685	0.582	0.692
ξ_2^I	50	*Repeated Indicator*	0.648	0.717	0.740	0.700	0.749
		Two-step approach	0.647	0.705	0.735	0.704	0.748
	100	*Repeated Indicator*	0.628	0.694	0.723	0.682	0.737
		Two-step approach	0.622	0.694	0.720	0.692	0.740
	300	*Repeated Indicator*	0.597	0.672	0.672	0.658	0.749
		Two-step approach	0.597	0.669	0.704	0.670	0.723
	1000	*Repeated Indicator*	0.571	0.649	0.686	0.637	0.707
		Two-step approach	0.571	0.647	0.685	0.646	0.707
ξ_3^I	50	*Repeated Indicator*	0.646	0.718	0.739	0.837	0.811
		Two-step approach	0.652	0.707	0.730	0.717	0.751
	100	*Repeated Indicator*	0.626	0.696	0.727	0.820	0.804
		Two-step approach	0.623	0.689	0.723	0.709	0.746
	300	*Repeated Indicator*	0.600	0.670	0.670	0.804	0.811
		Two-step approach	0.596	0.669	0.704	0.684	0.731
	1000	*Repeated Indicator*	0.571	0.648	0.686	0.786	0.788
		Two-step approach	0.570	0.648	0.685	0.709	0.717

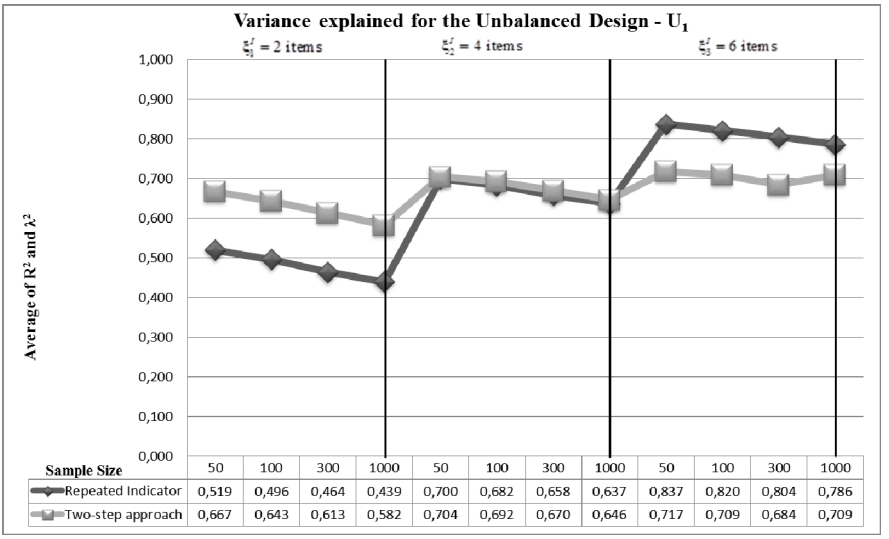
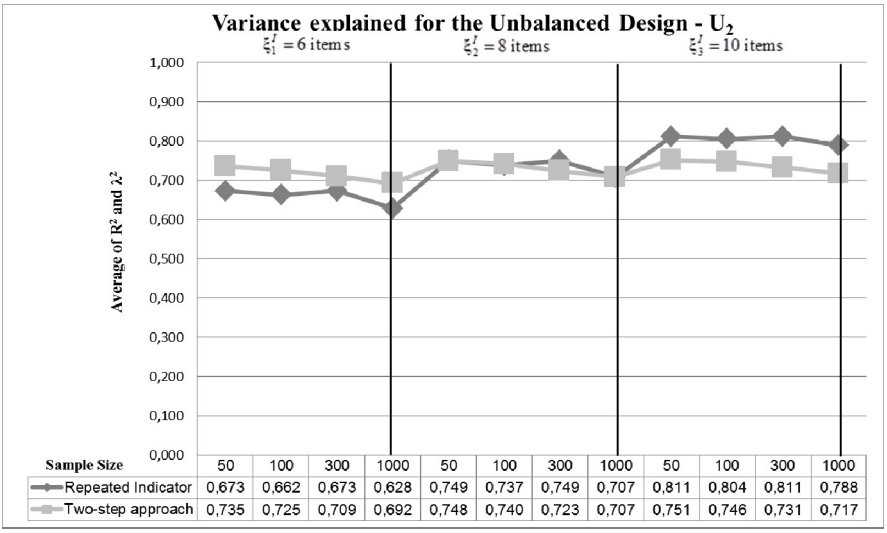


Fig. 6 $\overline{R}^2$ and $\overline{\lambda}^2$ for the unbalanced design U_1

The average of the squared correlations, over 500 replications, between predicted LV scores and the defined true scores are reported in Table 4. In this case we evaluate the ability of the approach to predict/replicate the true value of the defined LV, by considering the relationships between the MVs and the LVs. Reading the table at a glance, a first consideration can be made about the number of items: it is quite clear how, in the balanced design, increasing the number of items, regardless of the approach considered, the R^2 between the predicted and the true scores increases.



Sample Size	50	100	300	1000	50	100	300	1000	50	100	300	1000
Repeated Indicator	0,673	0,662	0,673	0,628	0,749	0,737	0,749	0,707	0,811	0,804	0,811	0,788
Two-step approach	0,735	0,725	0,709	0,692	0,748	0,740	0,723	0,707	0,751	0,746	0,731	0,717

Fig. 7 $\overline{R}^2$ and $\overline{\lambda}^2$ for the unbalanced design U_2

In the case of balanced design, with a high number of items (6 items in our case), the repeated indicators gives the best results in terms of r^2; with a sample size between 50 and 300, the repeated indicators produces higher values of r^2; the two-step approach gives the best results in the case of high sample size and with few items (2 and 4 in our case).

In the case of the unbalanced design U_1, the differences are due to the sample size dimension: with a little sample size (50), the best approach is the repeated indicator, while, with higher sample sizes (≥100), the best approach is the two-step.

Contrarily, differences in accuracy within the U_2 design seem to be imputed to the constructs composition: provided that prediction accuracy increases as the number of items increases, the repeated indicator approach is more accurate than the two-step in estimating LVs with a higher number of items.

Table 4 The average of squared correlation (r^2) between the predicted LV scores and the defined true scores

LV	SAMPLE SIZE	APPROACH	ITEMS PER CONSTRUCT				
			2	**4**	**6**	**2;4;6**	**6;8;10**
ξ_1^1	50	*Repeated Indicator*	**0.680**	**0.808**	**0.863**	**0.679**	**0.861**
		Two-step approach	0.673	0.801	0.858	0.674	0.860
	100	*Repeated Indicator*	0.628	**0.772**	0.836	0.627	**0.834**
		Two-step approach	**0.630**	0.768	**0.837**	**0.630**	0.833
	300	*Repeated Indicator*	**0.567**	0.722	**0.795**	0.564	**0.797**
		Two-step approach	0.561	**0.722**	0.794	**0.566**	0.795
	1000	*Repeated Indicator*	0.507	0.672	**0.754**	0.505	**0.773**
		Two-step approach	**0.508**	**0.672**	0.753	**0.507**	0.752
ξ_2^1	50	*Repeated Indicator*	**0.678**	**0.811**	**0.862**	**0.806**	**0.891**
		Two-step approach	0.673	0.805	0.861	0.799	0.890
	100	*Repeated Indicator*	**0.627**	**0.773**	**0.836**	0.769	**0.872**
		Two-step approach	0.625	0.768	0.834	**0.772**	0.868
	300	*Repeated Indicator*	**0.565**	**0.721**	**0.794**	0.719	**0.839**
		Two-step approach	0.563	0.719	0.794	**0.722**	0.836
	1000	*Repeated Indicator*	0.505	0.672	**0.754**	0.673	**0.817**
		Two-step approach	**0.507**	**0.673**	0.752	**0.672**	0.804
ξ_3^1	50	*Repeated Indicator*	**0.678**	**0.811**	**0.863**	**0.863**	**0.912**
		Two-step approach	0.672	0.803	0.858	0.858	0.910
	100	*Repeated Indicator*	**0.629**	**0.771**	**0.837**	0.834	**0.894**
		Two-step approach	0.623	0.769	0.833	**0.835**	0.893
	300	*Repeated Indicator*	**0.565**	**0.723**	**0.794**	0.794	**0.866**
		Two-step approach	0.562	0.722	0.793	**0.796**	0.864
	1000	*Repeated Indicator*	0.506	0.672	**0.755**	0.754	**0.854**
		Two-step approach	**0.507**	**0.674**	0.754	**0.754**	0.836

Summarising, the repeated indicator approach proves more accuracy with smaller sample sizes and a higher number of indicators; on the contrary, the two-step estimation approach requires bigger sample sizes (at least 100 units) and few indicators to be accurate.

3.4 Bias and Efficiency of the Parameters

In order to evaluate the estimation accuracy, the deviation of the estimated parameters from the true values has been assessed. A suitable measure of the accuracy is the relative bias (RB), obtained as the ratio of the difference between the average of the parameters estimated, with 500 replications, and true value, over the true value, according to the formula:

$$RB = \frac{E(\hat{\theta}) - \theta}{\theta}$$

The formula is equivalent to the mean RB [16]. Positive RB indicates an overestimation of the true parameter, negative RB an underestimation. RBs of the three path coefficients obtained by each approach under all conditions are reported in Table 5:

Table 5 RB of the path coefficients

PATH	SAMPLE SIZE	APPROACH	ITEMS PER CONSTRUCT				
			2	4	6	2;4;6	6;8;10
β_1	50	*Repeated Indicator*	0.0098	0.0588	0.0787	-0.0949	**0.0288**
		Two-step approach	**0.0051**	**0.0515**	**0.0682**	**0.0206**	0.0834
	100	*Repeated Indicator*	**-0.0095**	0.0452	0.0677	-0.1179	**0.0187**
		Two-step approach	-0.0103	**0.0413**	**0.0648**	**0.0027**	0.0369
	300	*Repeated Indicator*	**0.0098**	0.0231	0.0507	-0.1477	**0.0048**
		Two-step approach	-0.0361	**0.0227**	**0.0347**	**-0.0213**	-0.0218
	1000	*Repeated Indicator*	-0.0569	0.0063	**0.0351**	-0.1713	**-0.0111**
		Two-step approach	**-0.0567**	**0.0061**	0.0480	**-0.0464**	-0.0734
β_2	50	*Repeated Indicator*	0.0133	0.0613	0.0786	0.0490	0.0846
		Two-step approach	**0.0052**	**0.0493**	**0.0715**	**0.0490**	**0.0783**
	100	*Repeated Indicator*	**-0.0078**	0.0426	0.0644	**0.0338**	0.0745
		Two-step approach	-0.0139	**0.0410**	**0.0610**	0.0400	**0.0339**
	300	*Repeated Indicator*	**0.0133**	0.0253	0.0487	**0.0143**	0.0628
		Two-step approach	-0.0343	**0.0227**	**0.0344**	0.0230	**-0.0229**
	1000	*Repeated Indicator*	-0.0558	0.0074	**0.0355**	**-0.0021**	**0.0505**
		Two-step approach	**-0.0554**	**0.0058**	0.0487	0.0048	-0.0725
β_3	50	*Repeated Indicator*	**0.0079**	0.0625	0.0772	0.1454	0.1278
		Two-step approach	0.0090	**0.0508**	**0.0678**	**0.0587**	**0.0707**
	100	*Repeated Indicator*	**-0.0094**	0.0443	0.0671	0.1331	0.1221
		Two-step approach	-0.0131	**0.0379**	**0.0628**	**0.0524**	**0.0278**
	300	*Repeated Indicator*	**0.0079**	0.0238	0.0495	0.1212	0.1125
		Two-step approach	-0.0347	**0.0223**	**0.0349**	**0.0340**	**-0.0369**
	1000	*Repeated Indicator*	**-0.0553**	**0.0061**	**0.0351**	0.1083	0.1020
		Two-step approach	-0.0563	0.0063	0.0488	**0.0523**	**-0.0906**

For each condition, the best performing approach is marked in bold. A general overview gives the following evidence:

- the two approaches give after all equivalent accuracy when the measurement model has 4 or 6 indicators per construct;
- the repeated indicator approach is the most accurate when the MVs per construct are 2, especially with greater sample sizes;
- the two-step approach is generally more accurate than the repeated indicator approach in the first unbalanced design, with sample sizes lower than 300 and when the number of items per construct is equal to 2 and 6;
- the two-step approach outperforms the repeated indicator when there are 8 and 10 items per construct within the second unbalanced design; on the contrary, repeated indicators is better than two-step with items per construct equal to 6;
- the repeated indicators approach heavily underestimates, within the unbalanced design U_1, the path coefficient β_1, linking the second-order LV with the construct with few items (2), in proportion to sample dimension;
- on the other hand, the repeated indicators approach leads to a serious overestimation of the structural coefficient affecting the first-order LV with the highest number of items (both in the unbalanced designs U1 and U2).

To compare the two approaches in terms of means of squares error (MSE), where MSE is defined as $MSE = n^{-1} \cdot \sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i \right)^2$, the relative efficiency (RE) is

used. Considering two estimators, *A* and *B*, the ratio between the MSE($\hat{\theta}^A$) and MSE($\hat{\theta}^B$) defines the RE:

$$RE = \frac{MSE(\hat{\theta}^A)}{MSE(\hat{\theta}^B)}$$

RE<1 means that the first estimator (estimator *A*) is preferred (estimator *B* is inefficient relative to estimator *A*).

Table 6 Relative efficiency of the two approaches

SAMPLESIZE 50															
APPROACH	ITEMS PER CONSTRUCT 2			ITEMS PER CONSTRUCT 4			ITEMS PER CONSTRUCT 6			ITEMS PER CONSTRUCT 2;4;6			ITEMS PER CONSTRUCT 6;8;10		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Repeated Indicator	0.003	0.003	0.003	0.004	0.004	0.004	0.005	0.005	0.005	0.011	0.003	0.014	0.002	0.006	0.011
Two-step approach	0.003	0.003	0.003	0.004	0.003	0.003	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.005	0.005
RE	**0.920**	**0.962**	**0.984**	1.044	1.271	1.149	1.249	1.142	1.226	4.065	**0.924**	3.936	**0.540**	1.079	2.028
SAMPLESIZE 100															
APPROACH	ITEMS PER CONSTRUCT 2			ITEMS PER CONSTRUCT 4			ITEMS PER CONSTRUCT 6			ITEMS PER CONSTRUCT 2;4;6			ITEMS PER CONSTRUCT 6;8;10		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Repeated Indicator	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.004	0.004	0.012	0.002	0.012	0.001	0.004	0.010
Two-step approach	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.001	0.002	0.003	0.003	0.004	0.005
RE	1.042	**0.884**	**0.930**	1.134	1.139	1.130	1.120	1.165	1.145	8.411	**0.923**	4.656	**0.401**	1.004	2.162
SAMPLESIZE 300															
APPROACH	ITEMS PER CONSTRUCT 2			ITEMS PER CONSTRUCT 4			ITEMS PER CONSTRUCT 6			ITEMS PER CONSTRUCT 2;4;6			ITEMS PER CONSTRUCT 6;8;10		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Repeated Indicator	0.002	0.002	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.015	0.001	0.010	0.000	0.003	0.008
Two-step approach	0.002	0.001	0.002	0.001	0.001	0.001	0.002	0.002	0.002	0.001	0.001	0.001	0.002	0.003	0.003
RE	**0.941**	1.056	**0.866**	**0.992**	1.173	1.050	1.083	1.032	1.018	15.422	**0.874**	8.785	**0.202**	1.004	2.536
SAMPLESIZE 1000															
APPROACH	ITEMS PER CONSTRUCT 2			ITEMS PER CONSTRUCT 4			ITEMS PER CONSTRUCT 6			ITEMS PER CONSTRUCT 2;4;6			ITEMS PER CONSTRUCT 6;8;10		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Repeated Indicator	0.003	0.003	0.003	0.000	0.000	0.000	0.001	0.001	0.001	0.019	0.000	0.008	0.000	0.002	0.007
Two-step approach	0.002	0.002	0.002	0.000	0.000	0.000	0.001	0.001	0.001	0.002	0.000	0.000	0.001	0.002	0.002
RE	1.149	1.121	1.235	1.109	1.220	1.056	1.030	1.075	1.027	11.345	1.094	16.986	**0.298**	**0.934**	2.949

Bold numbers indicate that the first estimator (repeated indicator) is more efficient than the second (two-step approach); at first sight, the two-step approach might seem preferable to the repeated indicator, except in the case of 2 items per construct. The difference between the two approaches is indeed relevant just in the first unbalanced design, for the first and the third estimated parameters (β1 and β3), those linking the second-order LV to first-order constructs having, respectively, 2 and 6 items. In the second unbalanced design, RE<1 indicates that the repeated indicator approach is more efficient than the two-step in estimating the path coefficient influencing the construct with 6 items, while two-step is better

when the construct has 10 items. For all other conditions, the MSEs of the two esteems are quite identical, and their ratio is consequently close to 1.

3.5 Suggestions for the Choice of the High-Order Approach

The results obtained with the Monte Carlo simulations can provide some useful suggestions to those researchers who are intending to introduce a further level of abstraction in modelling the phenomenon of interest.

Table 7 reports a summary of the simulation results in order to give a schematic representation of the two-step and repeated indicator performances, referred to the estimation of the LVs and the parameters. Indifference in the choice of one approach rather than the other is marked with '-', a slight preference with '+', a strong preference with '++'.A further differentiation between the approaches is made for the unbalanced designs, with respect to the number of items of the construct at issue.

As the table reports, none of the high-order approaches excel in all simulation conditions; they are rather equivalent according to several criteria.

Table 7 Summary matrix

	Sample Size/Methods		Balanced						Unbalanced		Unbalanced	
			2		4		6		2;4;6		6;8;10	
			RI	TS	RI	TS	RI	TS	RI	TS	RI	TS
LATENT VARIABLES	Explained Variance	50	-	-	-	-	-	-	6 items	2 items	10 items	6 items
		100	-	-	-	-	-	-	6 items	2 items	10 items	6 items
		300	-	-	-	-	-	-	6 items	2 items	10 items	6 items
		1000	-	-	-	-	-	-	6 items	2 items	10 items	6 items
	Accuracy	50	+		+		+		++		+	
		100	+		+		+			++	+	
		300	+		+		+			++	+	
		1000		+		+	+			++	+	
PARAMETERS	Relative Bias	50	-	-		++		+		2-6 items	6 items	8-10 items
		100	-	-		+		+		2-6 items	6 items	8-10 items
		300	-	-		+		+		2-6 items	6 items	8-10 items
		1000	-	-		+	+		++		++	
	Relative Efficiency	50	+			+		+		2-6 items	6 items	10 items
		100	+			+		+		2-6 items	6 items	10 items
		300	+			+		+		2-6 items	6 items	10 items
		1000		+		+		+		2-6 items	6 items	10 items

Note: ++ = highly recommendable; + = Recommendable; - = Indifferent choice

The choice of the best approach clearly depends on the type of design:

- In the case of balanced design, with all sample sizes and number of items, the two methods lead to the same amount of explained variance (which grows depending on the number of items, see Table 3). For the accuracy of the LVs, the best choice is the repeated indicators, with some exceptions for big sample size with few items. When the main concern is the bias of the parameters, then the two-step approach gives the best Monte Carlo simulations, except for the case of few variables.

- In the case of unbalanced design, the suggestions are more elaborated/detailed. As regards the amount of explained variance each approach has strengths and weaknesses. The differences reveal that, when using the repeated indicators, the relationship between second and first-order LVs is much more explanatory whether the lower-order LV is measured by the bigger number of items (6 and 10 for, respectively, U_1 and U_2). On the contrary, the two-step approach produces better explained relationships between the two orders of the model in correspondence of smaller constructs.

  For the parameters estimation, in general, the optimal approach is the two-step, in terms of bias and relative efficiency. The exceptions are for the case of big sample size (1000), in terms of bias; or for smallest number of items (U_2 design), both for the bias and the efficiency.

The conclusions based on the simulation results can help the researcher in the choice of the best approach, given the constraints on the model, as the number of items, the sample size dimension and the type of design. The above-mentioned points have to be completed by some theoretical considerations on the relationships between the first and the second-order LVs.

In the repeated indicators approach, the second-order LV does not necessarily require the uni-dimensionality. It is indeed composed by a heterogeneous set of MVs underlying conceptually distinct first-order LVs. Moreover, the second-order LV is hierarchically superior since it includes all first-order LVs and has a causal effect on them.

In the two-step approach, the first-order LVs have to be related among them, requiring uni-dimensionality, since they measure the second-order LV. The two-step approach, then, defines a measurement rather than a causal relationship between the first and the second-order level.

In the end, we can conclude that for the repeated indicators, the second-order LV, being hierarchically superior, could be seen as a context variable and the focus is on the impact of the context/second-order variable on the first-order LVs.

In the two-step approach, the second-order variable is measured by the first-order variable and the aim is to what extent each first-order LV reflects (in terms of covariance) the composition of the second-order level.

4 The Empirical Evidence: A Job Satisfaction Model

In this paragraph we apply both high-order approaches experimented, to a job satisfaction case study. We discuss the results obtained on the real data set, and relate them to the conclusions drawn via Monte Carlo simulation.

4.1 The Data and the Model

The data used from the ICSI[2007] concern 320 social cooperatives spread across Italy, sampled from the Istat Census 2003 database[1].

[1] The Istat Census 2003 counts 5093 operative social cooperatives with 153284 paid workers.

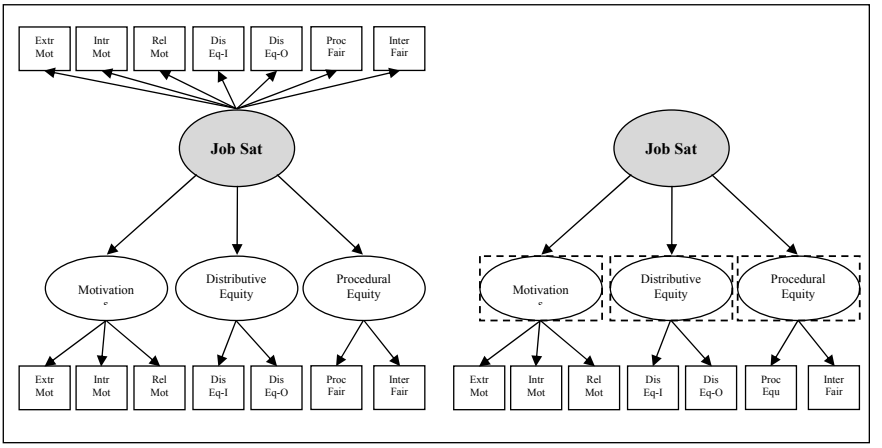


Fig. 8 The second-order job satisfaction path model: repeated indicators and two-step approach

4134 paid workers answered the questionnaire designed by academic experts in economic and organisation fields. Some descriptive analyses on the quality of these data are already developed in [2].

The second-order model here proposed uses a subset of variables from the whole ICSI^{2007} questionnaire. Taking into account the several Likert-type scales used, the Rasch rating scale model has been computed to construct the MVs of the following constructs [3]: *motivations* of workers; *distributive equity* and *procedural equity* of work.

Motivations to work has been assessed through three measurable variables: *intrinsic motivations*, such as the interest in the sector in which the organisation operates and the values shared between worker and organisation; *extrinsic motivations*, such as the pay or the working hours flexibility; *relational motivations*, related to the opportunity to increase the quality of the relations with colleagues, users and others.

Distributive equity relies on the perceived fairness of one's outcome with respect to the input/output ratio. It can be distinguished into *individual* (stress, responsibility, effort, training and loyalty) and *others* (wages of colleagues, superiors and economic resources of the cooperative).

Procedural equity concerns the fairness of the formal allocation process, measured by *procedural fairness*, expressed by the availability of information on worker activity and organisation goals, and *interactional fairness*, as the attention to workers' needs and proposals.

In this frame, *job satisfaction* has been conceived as higher-order LVs underlying the three dimensions (motivations, procedural and distributive equity) modelled as first-order LVs. In other words, workers' answers on specific aspects of their job are seen as reflecting an overall perception of satisfaction about their job.

The conceptual representation of the second-order job satisfaction model, for both the approaches, is reported in Figure 8. Given the theoretical model design,

the number of items per LV and the sample size dimension, the best approach to adopt, based on the above simulation results, should be the two-step. In this paper, for descriptive needs, we analyse both the approaches.

4.2 Results

The job satisfaction model has been estimated through a tailor-made algorithm developed in MATLAB, by adopting both the repeated indicators and two-step approach. A first check for constructs reliability, reported in Table 8, shows that all LVs in the model are reliable: both Cronbach's alpha and composite reliability coefficients are above the conventionally accepted threshold.

Table 8 Reliability measures

REPEATED INDICATORS		
	Cronbach's α	**C.R.**
MOTIVATIONS	0.694	0.829
DISTRIBUTIVE EQUITY	0.746	0.887
PROCEDURAL EQUITY	0.685	0.863
JOB SATISFACTION	0.609	0.748
TWO-STEP		
	Cronbach's α	**C.R.**
MOTIVATIONS	0.694	0.831
DISTRIBUTIVE EQUITY	0.746	0.887
PROCEDURAL EQUITY	0.685	0.864
JOB SATISFACTION	0.329	0.680

PLS performs the estimation of regression coefficients in the structural equation model; the bootstrap procedure approximates the sampling distribution of the estimator by re-sampling from the original sample, in order to test the parameters significance. The analysis used 200 replications, with a bootstrap sample equal to 1000. Results are shown in Table 9.

The main difference between the two approaches concerns the path coefficients linking the second-order LV, *job satisfaction*, with the first-order construct *motivations*: in the case of the repeated indicators, the strength of the association (β=0.590) is twice the path coefficient estimated by the two-step approach (λ=0.277). It should be noted that simulation results showed that, at large samples in the unbalanced design, path coefficients linking the second-order LV and the first-order construct with the highest number of items are overestimated (see Table 5). Further simulations on a model with the same design as the job satisfaction model confirm this result.

As regards the variability of the parameters estimation, the repeated indicators approach gives more robust measures with all three path coefficients, especially in the case of the β linking job satisfaction with motivations. As already seen (Figure 5), the

standard error for the parameters produced by the two-step approach is higher within the unbalanced design U_1 for the construct with the highest number of items.

Table 9 Path coefficients, standard error and t-statistics

REPEATED INDICATORS				
JOB SATISFACTION				
	Path coefficient	**SE**	**t-value**	**G.o.F.**
MOTIVATIONS	0.590	0.028	21.094	***0.475***
DISTRIBUTIVE EQUITY	0.618	0.026	23.947	
PROCEDURAL EQUITY	0.766	0.016	48.456	
TWO-STEP				
JOB SATISFACTION				
	Path coefficient	**SE**	**t-value**	**G.o.F.**
MOTIVATIONS	0.277	0.087	3.166	***0.530***
DISTRIBUTIVE EQUITY	0.624	0.037	17.021	
PROCEDURAL EQUITY	0.725	0.010	73.901	

Furthermore, according with Cohen's categorisation of the effect size [6], and fixing a communality to 0.5, the goodness of fit (G.o.F.) thresholds for small, medium and large effect size are 0.29, 0.41 and 0.47. For both approaches, G.o.F. coefficients indicate a good fit of the model to the data.

The explained variance with the two approaches is shown in Table 10. Consistently with the simulation study, higher amounts of explained variance are reached with the repeated indicators when we deal with the largest construct (see Figure 6); conversely, the two-step approach leads to better-explained relationships in case of smaller constructs, such as distributive and procedural equity in the job satisfaction model.

Table 10 Explained variance with the two approaches

EXPLAINED VARIANCE		
	REPEATED INDICATORS	**TWO-STEP**
	JOB SAT	**JOB SAT**
MOTIVATIONS	0.348	0.113
DISTRIBUTIVE EQUITY	0.382	0.522
PROCEDURAL EQUITY	0.59	0.701

The interpretation of the case study results can help to understand the differences of the two approaches in terms of application perspective. The job satisfaction, for the repeated indicators, defines an LV of context, which expresses the total amount of satisfaction of the workers. The path coefficients reported in Table 9 (0.59; 0.618; 0.766) define the intensity of the causal relationships

between the job satisfaction and its three sub-dimensions, represented by first-order LVs. Which means, for instance, keeping the other parameters constant, if we increase the satisfaction of a quantity equal to 1, the perception about procedural equity will increase by 0.766.

In the two-step approach, the job satisfaction defines again a global measure of workers' satisfaction, but in this case the relationships with the motivations, distributive and procedural equities are structural coefficients of a measurement model. The path coefficients (0.277; 0.624; 0.725) reflect the composition of the high-order job satisfaction; they indeed do not represent how much the high-order dimension affects the first-order sub-dimensions, but the extent to which the first-order constructs reflect the higher level of abstraction.

5 Conclusions

In social as well as in business and organisational sciences, the importance of the analysis of high-order LVs, and their relationships with lower level dimensions steadily increases. The correct specification of a considerable variety of concepts often requires a multidimensional definition: job satisfaction, beliefs, negative and positive emotions and several other well-known constructs (see [12], [14], [18]) can hardly be seen as a unique dimension and raise the matter of describing a further level of abstraction.

The purpose of this paper was to provide researchers using PLS-PM for analysing high-order LVs with an overview of the two main approaches present in the literature, as well as an evaluation of their suitability. The two approaches presented are the repeated indicators and the two-step approach.

By means of a Monte Carlo simulation study the different approaches have been compared in terms of variance explained by the second-order LV, prediction accuracy, estimation accuracy and efficiency. Simulation results are intended to define some recommendations in the choice of the most suitable method depending on the structure of the model hypothesised. The models tested in the study were varied for sample sizes, number of items per construct and design structure (balanced or unbalanced).

Generally speaking, simulation results suggest that the main differences between the two approaches are due to the model design, which can be balanced or unbalanced. In the former case, results indicate a better prediction accuracy of the repeated indicators approach; conversely, parameters estimation accuracy is higher with the two-step approach. In terms of explained variance, the two approaches are substantially equal. In the latter case, the repeated indicators approach outperforms two-step in LVs reconstruction when the design is composed of 6, 8 and 10 items, while the two-step is preferable in terms of parameter accuracy with the bigger constructs in the two designs. Finally, the repeated indicators approach gives higher explained variance for the bigger first-order LVs in the model while the two-step approach is better for the smaller LVs.

As an illustrative example, the last section reported a case study on job satisfaction survey, in order to show how theoretical and statistical aspects have to be considered in higher-order PLS-PM.

The evaluation of the model needs an integration of simulation results with the case study analysis, which allows the clarification of the role of the second-order LV in the model. In the case of repeated indicators, the compliance with construct uni-dimensionality is not required for the second-order LV as it consists of heterogeneous MVs. The path coefficients estimate the effect of the second level on the first level. In the two-step approach, uni-dimensionality is required: here, the second-order LV is measured by the first-order LVs. Then, path coefficients measure the composition of the second-order LV.

These considerations have to be taken into account in the interpretation of the job satisfaction model. It is remarkable to notice that the first-order LV motivations, which is strongly influenced (β=0.590) by the *job satisfaction* when the estimation approach used is that of repeated indicators, turns out to have little importance in reflecting the *job satisfaction* (λ=0.277) when the approach used is the two-step.

In the first case, workers' motivations can be improved by acting on the general level of satisfaction; in the second case, motivations are not decisive in the composition of the job satisfaction, which is mostly reflected by distributive and procedural equities. It can be concluded that the same variable plays two different roles. We suggest adopting the two-step approach when the aim of the research is to measure a concept at a high level of abstraction; on the other hand, if the goal to be reached is the evaluation of the impact of an overall abstraction on the first-order facets, the most suitable choice is the repeated indicators.

This work is not without limitations. A Monte Carlo study can provide suggestions about the population model under investigation, but researchers often select structural equation models of greater complexity. Formative relationships, both at the first and the second level, could be hypothesised according to the theoretical definition of the construct involved in the model. Different numbers of indicators could also be used for measuring the LVs as well as true parameters could be different from those of the model used in this study. All higher-order model structure combinations [11] need to be investigated.

The introduction of categorical variables through the approach of the external information is another issue to be addressed.

Furthermore, as this work is limited to PLS-based approaches, other SEM techniques (such as LISREL) have not been considered. This approach could be the object of a future exploration through a Monte Carlo design in order to have another basis for a comparison.

From the empirical perspective, future study will be focused on the analysis of the customer satisfaction.

References

[1] Agarwal, R., Karahanna, E.: Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS Quarterly 24(4), 665–694 (2000)

[2] Carpita, M.: The Quality of Work in Social Cooperatives, Measurements and Statistical Models, Franco Angeli, Milan, Italy (2009)

[3] Carpita, M., Vezzoli, M.: Measures and drivers of the job satisfaction. In: MTISD 2010, Book of short papers (2010)
[4] Chin, W.W., Marcolin, B.L., Newsted, P.R.: A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic mail adoption study. Information Systems Research 14(2), 189–217 (2003)
[5] Ciavolino, E., Al-Nasser, A.D.: Comparing generalized maximum entropy and partial least squares methods for structural equation models. Journal of Nonparametric Statistics 21(8), 1017–1036 (2009)
[6] Cohen, J.: Statistical Power Analysis for the Behavioural Sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale (1988)
[7] Diamantopoulos, A., Winklhofer, H.M.: Index construction with formative indicators: An alternative to scale development. Journal of Marketing Research 38(2), 269–277 (2001)
[8] Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H.: Handbook of Partial Least Squares. Concepts, Methods and Applications, New York. Springer Handbooks of Computational Statistics (2010)
[9] Henseler, J., Chin, W.W.: A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. Structural Equation Modeling: A Multidisciplinary Journal 17(1), 82–109 (2010)
[10] Hulland, J., Ryan, M.J., Rayner, R.K.: Modeling customer satisfaction: a comparative performance evaluation of covariance structure analysis versus partial least squares. In: Handbook of Partial Least Squares, New York. Springer Handbooks of Computational Statistics (2010)
[11] Jarvis, C.B., MacKenzie, S.B., Podsakoff, P.M.: A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. Journal of Consumer Research 30, 199–218 (2003)
[12] Judge, T.A., Hulin, C.L.: Job satisfaction as a reflection of disposition: A multiple-source causal analysis. Organizational Behavior and Human Decision Processes 56, 388–421 (1993)
[13] Lohmöller, J.B.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)
[14] Murry, J.P., Dacin, P.A.: Cognitive Moderators of Negative-Emotion Effects: Implications for Understanding Media Context. Journal of Consumer Research 22, 439–446 (1996)
[15] Rajala, R., Westerlund, M.: Antecedents to consumers' acceptance of mobile advertisements – a hierarchical construct PLS structural equation model. In: The 43rd Hawaii International Conference on System Sciences (HICSS), January 5-8 (2010)
[16] Reinartz, W.J., Echambadi, R., Chin, W.W.: Generating non-normal data for simulation of structural equation models using Mattson's method. Multivariate Behavioural Research 37(2), 227–244 (2002)
[17] Reinartz, W., Krafft, M., Hoyer, W.D.: The customer relationship management process: Its measurement and impact on performance. Journal of Marketing Research 41(3), 293–305 (2004)
[18] Shimp, T.A., Kavas, A.: The theory of reasoned action applied to coupon usage. Journal of Consumer Research, 795– 809 (1984)
[19] Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnajah, P.R. (ed.) Multivariate Analysis, pp. 391–420. Academic Press, New York (1966)
[20] Wold, H.: Path models with latent variables: The NIPALS approach. In: Blalock, H.M. (ed.) Quantitative Sociology, pp. 307–357. Seminar Press, NewYork (1975)

Log-Ratio and Parallel Factor Analysis: An Approach to Analyze Three-Way Compositional Data

Michele Gallo

Department of Human and Social Sciences, University of Naples – L'Orientale,
Largo S.Giovanni Maggiore, 30 I-80134 Naples, Italy
mgallo@unior.it

Abstract. For the exploratory analysis of three-way data, Parafac/Candecomp model (CP) is one of the most applied models to study three-way arrays when the data are approximately trilinear. It is a three-way generalization of PCA (Principal Component Analysis). CP model is a common name for low-rank decomposition of three-way arrays. In this approach, the three-dimensional data are decomposed into a series of factors, each relating to one of the three physical ways. When the data are particular ratios, as in the case of compositional data, this model should consider the special problems that compositional data pose. The principal aim of this paper is to describe how an analysis of compositional data by CP is possible and how the results should be interpreted.

Keywords: Centered log-ratios, Aitchison geometry, Parafac, Candecomp, three-mode analysis, one-mode plot, pre-components plot.

1 Introduction

Compositional data are commonly present in many disciplines. Nevertheless, the constant-sum constraint of compositional data is often either ignored or improperly incorporated into statistical modeling, and a misleading interpretation of the results is given. There are different approaches to incorporate compositional data into a statistical modeling, when it is not realistic to assume a multinomial distribution of the data. An approach was proposed by Aitchison (1986) and it is characterized by a log linear transformation of original data. Following Aitchison's approach, more multidimensional techniques have been adapted to analyze compositional data, for example, Principal Component Analysis (Aitchison, 1982), Partial Least Squares (Hinkle and Rayens, 1995; Gallo, 2003), Discriminant Partial Least Squares (Gallo, 2010), Hierarchical Cluster (Martìn-Fernàndez *et al.*, 1998) are only some of multivariate techniques proposed in literature.

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 209–221.
springerlink.com

Compositional data can be arranged into three-way matrices, for instance, when each composition corresponding to *J* parts, and *I* compositions is collected in *K* occasions. The Parafac (PARAllel FACtor analysis) is a multi-way method which was simultaneously originated in 1970 by Harshman and Caroll and Chang. The latter named it Candecomp (CANonical DECOMPosition). This method was proposed for interval and ratio scale variables and by a constrained version to enable nominal scale variable in a quantitative analysis (Bro, 1998). It is proposed to analyze a particular ratio scale measurement called compositional data, which create special problems. The first one is the interpretability of covariance structures. Every row of the different slices of the three-way matrix has one-sum. Therefore the correlation between each pair of variables is not free to range over the usual interval (−1, 1), with relative difference from the standard interpretation of correlation between the variables.

Like other multivariate methods, CP can be used to analyze the compositional data but only after an adequate transformation and considering that this data can only give relative information.

The main approach to the analysis of compositional data, when assuming a multinomial distribution of the data is not realistic, is characterized by a log-ratio transformation of the original data. Aitchison (1986) proposed three different kinds of data transformations.

Following this approach, in this paper we examine the problems that potentially occur when a CP analysis on compositional data is performed. In Section 2, after a brief review of CP model, we define the properties of compositional data and the preprocessing required. In Section 3, the CP for compositional data is proposed and how the results can be read. Finally, a data set containing 26 countries, 8 variables and 5 years is analyzed in Section 4.

2 Compositional Data and Relative Geometry

To define a compositional data, let $\tilde{v}_1 \ldots \tilde{v}_J$ be positive quantities with the same measurement scale $\tilde{\mathbf{v}} = \tilde{v}_1,\ldots,\tilde{v}_J$ with $\tilde{v}_1 > 0,\ldots,\tilde{v}_J > 0$. The vector $\tilde{\mathbf{v}}$ is the basis of compositional data and $\mathbf{v} = \tilde{\mathbf{v}} / \left\|\tilde{\mathbf{v}}\right\|$ is a composition vector, where $\|\,.\,\|$ is the norm of vector. Let $\Re^J$ and S^J be a *J*-dimensional real space and a *J*-dimensional simplex, respectively. The constraining operator defines a transformation $\Re_+^J \to S^J$ where $\Re_+^J$ is the positive orthant of $\Re^J$ defined by $\Re_+^J = \left\{\left(\tilde{v}_1,\ldots,\tilde{v}_J\right):\ \tilde{v}_1 > 0,\ldots,\tilde{v}_J > 0\right\}$ and S^J is defined by $S^J = \left\{\left(v_1,\ldots,v_J\right):\ v_1 > 0,\ldots,v_J > 0; \sum_{j=1}^{J} v_j = 1\right\}$. In this framework, two bases $\tilde{\mathbf{v}}$ and $\dot{\mathbf{v}}$ are compositional equivalents if there exist a positive constant *h* such that $\tilde{\mathbf{v}} = h\dot{\mathbf{v}}$. In this case the two bases have the same compositional class $\mathbb{C}$, and $\mathbf{v} = \tilde{\mathbf{v}} / \left\|\tilde{\mathbf{v}}\right\| = \dot{\mathbf{v}} / \left\|\dot{\mathbf{v}}\right\|$ is a unique standardized composition. By

considering a geometrical point of view, any compositional class is represented by a ray from the origin in $\Re_+^J$ and the intersection with the simplex S^J defined by its unit-sum. The natural sample space for compositional data is the unit simplex S^J, and the basic operations required for this vector space structure are perturbation and powering (for more detail see Aitchison , 1986; Pawlowsky-Glahn *et al.*, 2007; Aitchison and Ng, 2005). Perturbation of any two compositions $\mathbf{v}_i, \mathbf{v}_{i'} \in S^J$ is $\mathbf{v}_i \oplus \mathbf{v}_{i'}$ and its inverse $\mathbf{v}_i \ominus \mathbf{v}_{i'}$, are defined as follows: $\mathbf{v}_i \oplus \mathbf{v}_{i'} = \mathcal{C}\left[v_{i1}v_{i'1}, \ldots, v_{ij}v_{i'j}, \ldots, v_{iJ}v_{i'J}\right]$ and $\mathbf{v}_i \ominus \mathbf{v}_{i'} = \mathcal{C}\left[v_{i1} / v_{i'1}, \ldots, v_{ij} / v_{i'j}, \ldots, v_{iJ} / v_{i'J}\right]$.

Powering of any composition $\mathbf{v}_i \in S^J$ and any constant $\alpha \in \Re$ is $\alpha \odot \mathbf{v}_i = \mathcal{C}\left[v_{i1}^{\alpha}, \ldots, v_{ij}^{\alpha}, \ldots, v_{iJ}^{\alpha}\right]$.

The perturbation operation and power transformation into the simplex correspond to the translation and multiplication into real space; with the aim of extending the structure of S^J to a linear sample space the following inner product $\langle ., . \rangle_a$, norm $\| . \|_a$ and distance $d_a(.,.)$, for two compositions $\mathbf{v}_i, \mathbf{v}_{i'} \in S^J$, are defined by Aitchison (1992) as

a. $\left\langle \mathbf{v}_i, \mathbf{v}_{i'} \right\rangle_a = \sum_{j=1}^{J} \sum_{j'=1}^{J} \log\left(\mathrm{v}_{ij} / \mathrm{v}_{ij'}\right) \log\left(\mathrm{v}_{i'j} / \mathrm{v}_{i'j'}\right) / 2J$

b. $\left\|\mathbf{v}_i\right\|_a = \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} \log\left(\mathrm{v}_{ij} / \mathrm{v}_{ij'}\right)^2 / 2J \right]^{1/2}$

c. $d_a\left(\mathbf{v}_i, \mathbf{v}_{i'}\right) = \left[\sum_{j=1}^{J} \sum_{j'=1}^{J} \left(\log\left(\mathrm{v}_{ij} / \mathrm{v}_{ij'}\right) - \log\left(\mathrm{v}_{i'jk} / \mathrm{v}_{i'j'}\right)\right)^2 / 2J \right]^{1/2}$

This geometric structure of the simplex, known as Aitchison geometry, satisfies standard properties, as compatibility of the distance with perturbation and powering transformation, and subcompositional coherence (Pawlowsky-Glahn and Egozcue, 2001; Egozcue *et al.*, 2011). Thus, working in simplex space is analogous to working in real space.

In case of the same compositional object, e.g., *i*th is observed in *k*th occasions, we can arrange it into a single vector of juxtaposed composition. Thus, let $\mathbf{v}_{ik}$ be an *i*th object observed in *k*th occasion where $\mathbf{v}_{ik} \in S_k^J$, then it is possible to define $\underline{\mathbf{v}}_i$ as $\underline{\mathbf{v}}_i = \left[\mathbf{v}_{i1} | \ldots | \mathbf{v}_{ik} | \ldots | \mathbf{v}_{iK}\right]$ as the vector of the *i*th object coordinate

into K simplex spaces. Gallo (2011) has shown the following properties for the vector of juxtaposed compositions $\underline{\mathbf{v}}_i$:

a. $\underline{\mathbf{v}}_i \oplus \underline{\mathbf{v}}_{i'} = \left[\mathbf{v}_{i1} \oplus \mathbf{v}_{i'1} \middle| \mathsf{K} \middle| \mathbf{v}_{ik} \oplus \mathbf{v}_{i'k} \middle| \mathsf{K} \middle| \mathbf{v}_{iK} \oplus \mathbf{v}_{i'K} \right]$

b. $\alpha \mathsf{e}\ \underline{\mathbf{v}}_i = \left[\alpha \mathsf{e}\ \mathbf{v}_{i1} \middle| \mathsf{K} \middle| \alpha \mathsf{e}\ \mathbf{v}_{ik} \middle| \mathsf{K} \middle| \alpha \mathsf{e}\ \mathbf{v}_{iK} \right]$

c. $\left\langle \underline{\mathbf{v}}_i, \underline{\mathbf{v}}_{i'} \right\rangle_a = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{j'=1}^{J} \log\left(\mathrm{v}_{ijk} / \mathrm{v}_{ij'k}\right) \log\left(\mathrm{v}_{i'jk} / \mathrm{v}_{i'j'k}\right) / 2JK$

d. $\left\| \underline{\mathbf{v}}_i \right\|_a = \left[\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{j'=1}^{J} \log\left(\mathrm{v}_{ijk} / \mathrm{v}_{ij'k}\right)^2 / 2JK \right]^{1/2}$

e. $d_a\left(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_{i'}\right) = \left[\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{j'=1}^{J} \left(\log\left(\mathrm{v}_{ijk} / \mathrm{v}_{ij'k}\right) - \log\left(\mathrm{v}_{i'jk} / \mathrm{v}_{i'j'k}\right) \right)^2 / 2JK \right]^{1/2}$

Thus the juxtaposed vector of compositions satisfies the standard properties of compositional vector.

2.1 Preprocessing of Compositions

According to the principle that compositional data give information about relative, not absolute values of components, it is appropriate to consider a logarithms of the ratios, called log-ratios, to transform the compositional data before using standard unconstrained multivariate analysis. Aitchison (1982) proposed that the logarithm of the ratio, known as log-ratio, between all pairs of the parts of a composition could be considered. Through this transformation, a direct association between the simplex sample space and the real space is found. In this way, it is possible to work in real space, where it is easier working, and later, through the inverse function, the results can be taken back into simplex space.

This transformation can be used for compositional data arranged in juxtaposed vector too. Thus, the vector $\underline{\mathbf{v}}_i$ is transformed in a new vector with $(J-1)JK/2$ pairwise log-ratios (*plr*), where the generic element is $\log\left(v_{ijk} / v_{ij'k}\right)$ with $(j < j')$ and $k = 1, \mathsf{K}, K$.

Other log-ratio transformations proposed in literature are additive log-ratios (*alr*), centered log-ratios (*clr*) and the isometric log-ratios (*ilr*). The *alr* and *clr* were introduced by Aitchison (1986), while *ilr* was introduced by Egozcue et *al.* (2003), and they can be defined by the juxtaposed vector of compositions as follows:

1. $\widehat{\underline{\mathbf{z}}}_i = \left[\widehat{\mathbf{z}}_{i1} \,|\, \cdots \,|\, \widehat{\mathbf{z}}_{ik} \,|\, \cdots \,|\, \widehat{\mathbf{z}}_{iK}\right] = \left[alr(\mathbf{v}_{i1}) \,|\, \cdots \,|\, alr(\mathbf{v}_{ik}) \,|\, \cdots \,|\, alr(\mathbf{v}_{iK})\right]$ where $\widehat{\mathbf{z}}_{ik} = \left[\log(v_{i1k} / v_{iJk}), \cdots, \log(v_{iJ-1k} / v_{iJk})\right]$
2. $\underline{\mathbf{z}}_i = \left[\mathbf{z}_{i1} \,|\, \cdots \,|\, \mathbf{z}_{ik} \,|\, \cdots \,|\, \mathbf{z}_{iK}\right] = \left[clr(\mathbf{v}_{i1}) \,|\, \cdots \,|\, clr(\mathbf{v}_{ik}) \,|\, \cdots \,|\, clr(\mathbf{v}_{iK})\right]$ where $\mathbf{z}_{ik} = \left[\log(v_{i1k} / g(\mathbf{v}_{ik})), \cdots, \log(v_{iJk} / g(\mathbf{v}_{ik}))\right]$ and with $g(\mathbf{v}_{ik}) = \left[\log(v_{i1k}) * \log(v_{i2k}) * \cdots * \log(v_{iJk})\right]^{1/J}$
3. $\breve{\underline{\mathbf{z}}}_i = \left[\breve{\mathbf{z}}_{i1} \,|\, \cdots \,|\, \breve{\mathbf{z}}_{ik} \,|\, \cdots \,|\, \breve{\mathbf{z}}_{iK}\right] = \left[ilr(\mathbf{v}_{i1}) \,|\, \cdots \,|\, ilr(\mathbf{v}_{ik}) \,|\, \cdots \,|\, ilr(\mathbf{v}_{iK})\right]$ where $\breve{\mathbf{z}}_{ik} = \left[\log(v_{i1k}), \cdots, \log(v_{iJk})\right]\Psi$

with $\Psi^t\Psi = \mathbf{I}_{J-1}$ and $\Psi\Psi^t = \left(\mathbf{I}_J - \hat{\mathbf{1}}_J\hat{\mathbf{1}}_J^t / J\right)$, $\mathbf{I}$ is the identity matrix and $\hat{\mathbf{1}}$ is a unit vector.

All these transformations have a role to play in the analysis of compositional data and the choice is determined by the kind of application or the results that we are interested to highlight. Here, just the *clr* transformation will be considered.

3 Three-Way Data

When compositional vectors are observed in several occasions they can be arranged in a three-way matrix as rows, columns or tubes. Here, the three-way matrix $\underline{\mathbf{V}}$ $(I \text{ x } J \text{ x } K)$ has $(I \text{ x } K)$ rows, where each row $\mathbf{v}_{ik}$ $(i = 1,...,I;\ k=1,...,K)$ is a composition vector, the J columns corresponding to the variables or parts of compositions. While $\underline{\mathbf{v}}_i = \left[\mathbf{v}_{i1} \,|\, \cdots \,|\, \mathbf{v}_{ik} \,|\, \cdots \,|\, \mathbf{v}_{iK}\right]$ is a *JK*-dimensional vector where an *i*th object is observed in *K* different simplex spaces, one for each occasion: $\mathbf{v}_{ik} \in S_k^J$ ($k = 1,...,K$). In other words, $\mathbf{v}_{ik}$ is a compositional *J*-dimensional vector observed on *k*th occasion, and $\underline{\mathbf{v}}_i$ is a vector of juxtaposed compositions. And finally, we define $\mathbf{V}_k$ $(I \text{ x } J)$ as a matrix obtained by fixing the third mode of $\underline{\mathbf{V}}$ at *k* ($\mathbf{V}_k$ is the *k*th frontal slice of $\underline{\mathbf{V}}$).

By considering the *clr* transformation discussed in Section 2.1, the logarithmic transformation can be applied on three-way matrix $\underline{\mathbf{V}}$, thus $\underline{\mathbf{L}}$ $(I \text{ x } J \text{ x } K)$ is an array with typical element $\log(v_{ijk})$, and $\mathbf{L}_k$ is the *k*th frontal slice $(k = 1,...,K)$. The *k*th frontal slice of the *clr* can be written as $\mathbf{L}_k\mathbf{P}_J^{\perp}$, where $\mathbf{P}_J^{\perp} = \left(\mathbf{I}_J - \hat{\mathbf{1}}_J\hat{\mathbf{1}}_J^t / J\right)$ is a symmetric and idempotent centering matrix

($\mathbf{I}$ is the identity matrix and $\hat{\mathbf{1}}$ is a vector of units). To ensure that log-ratios are centered with respect to columns, it is necessary that each frontal slice is premultiplied by the symmetric and idempotent centering matrix $\mathbf{P}_I^{\perp} = \left(\mathbf{I}_I - \hat{\mathbf{1}}_I \hat{\mathbf{1}}_I^t / I\right)$. Thus, $\underline{\mathbf{Y}}$ is the three-way matrix of *clr* with $\mathbf{Y}_k = \mathbf{P}_I^{\perp} \mathbf{L}_k \mathbf{P}_J^{\perp}$ the *k*th frontal slice.

There are two different approaches to study this data set. The first approach is to treat it as a set of matrices and apply a two-mode analysis, as singular value decomposition, to each frontal slice of the three-way array. In the second approach, we assume that the loadings for compositions and parts are the same across all the conditions, so a three-mode analysis, as CP, is applied. The latter approach is recommended from the point of view of parsimony, and in order to get information on what the solutions have in common (Kroonenberg, 2008 pp.50).

3.1 CP Analysis of Log-Ratio Data

Defining $\mathbf{A}$ $(I \text{ x } F)$ as the loadings matrix for compositions, $\mathbf{B}$ $(J \text{ x } F)$ as the loadings matrix for variables or parts, and $\mathbf{C}$ $(K \text{ x } F)$ as the loadings matrix for occasions, the CP model for log-ratios can be written as

$$\mathbf{Y}_A = \mathbf{A}\dot{\mathbf{I}}\left(\mathbf{C} \otimes \mathbf{B}\right)^t + \mathbf{E}_A$$

where $\otimes$ is the Kronecker product, F is the number of components used to approximate the data, $\mathbf{Y}_A = \left[\mathbf{Y}_1 \middle| \mathbf{K} \middle| \mathbf{Y}_k \middle| \mathbf{K} \middle| \mathbf{Y}_K\right]$ and $\mathbf{E}_A = \left[\mathbf{E}_1 \middle| \mathbf{K} \middle| \mathbf{E}_k \middle| \mathbf{K} \middle| \mathbf{E}_K\right]$ are the juxtaposition of the frontal slices of three-way data and residuals, respectively; and $\dot{\mathbf{I}}$ $\left(F \text{ x } F^2\right)$ denotes the matricized version of the unit superdiagonal array $\underline{\mathbf{I}}$ $(F \text{ x } F \text{ x } F)$ which has unit elements in the position (f, f, f), $f = 1, \mathrm{K}, F$, and zeros elsewhere.

To surmount the difficulties described in Section 2 and avoid resulting in failure, the constraints $\hat{\mathbf{1}}_I \mathbf{A} = \hat{\mathbf{0}}_F$ and $\hat{\mathbf{1}}_J \mathbf{B} = \hat{\mathbf{0}}_F$ ($\hat{\mathbf{0}}_F$ is a vector of zero F-dimension) should be taken into account in CP model. Another possibility is to express the CP model in a slice-wise form as:

$$\mathbf{Y}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k$$

where $\mathbf{D}_k$ is a diagonal matrix containing the *k*th row of $\mathbf{C}$.

Following the slice-wise form, fitting the CP model to $\underline{\mathbf{Y}}$ with the constraints $\hat{\mathbf{1}}_I \mathbf{A} = \hat{\mathbf{0}}_F$ and $\hat{\mathbf{1}}_J \mathbf{B} = \hat{\mathbf{0}}_F$ in the least squares sense can be expressed as:

$$\underset{\hat{\mathbf{1}}_I \mathbf{A}=\hat{\mathbf{0}}_F;\, \hat{\mathbf{1}}_J \mathbf{B}=\hat{\mathbf{0}}_F}{argmin} \sum_{k=1}^{K} \left\| \mathbf{Y}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t \right\|^2 \quad (1)$$

In Equation 1 the constrained $\hat{\mathbf{1}}_I \mathbf{A} = \hat{\mathbf{0}}_F$ and $\hat{\mathbf{1}}_J \mathbf{B} = \hat{\mathbf{0}}_F$ will automatically be respected. In other words, the double-centered of each $\mathbf{Y}_k$ $(k = 1,\mathsf{K}\ , K)$ assure loadings $\mathbf{A}$ and $\mathbf{B}$ will always be centered:

$$\underset{\hat{\mathbf{1}}_I \mathbf{A}=\hat{\mathbf{0}}_F;\, \hat{\mathbf{1}}_J \mathbf{B}=\hat{\mathbf{0}}_F}{argmin} \sum_{k=1}^{K} \left\| \mathbf{Y}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t \right\|^2 = argmin \sum_{k=1}^{K} \left\| \mathbf{Y}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t \right\|^2 \quad (2)$$

To proof Equation 2, we have to consider that $\mathbf{Y}_k = \mathbf{P}_I^{\perp}\mathbf{L}_k\mathbf{P}_J^{\perp}$ with $\mathbf{L}_k$ the *k*th frontal slice of $\underline{\mathbf{L}}$ $(I \text{ x } J \text{ x } K)$, $\mathbf{P}_I^{\perp}$ and $\mathbf{P}_J^{\perp}$ are symmetric and idempotent centering matrices as defined into Section 3.

Assuming that $\mathbf{L}_k$ is describe by

$$\mathbf{L}_k = \mathbf{G}\mathbf{D}_k\mathbf{H}^t + \mathbf{E}_k \quad (3)$$

To remove the overall mean of all elements of $\mathbf{L}_k$ the projection matrices $\mathbf{P}_I = \hat{\mathbf{1}}_I\hat{\mathbf{1}}_I^t / I$ and $\mathbf{P}_J = \hat{\mathbf{1}}_J\hat{\mathbf{1}}_J^t / J$ could be used and the matrix $\mathbf{Y}_k$ could be written as

$$\mathbf{Y}_\mathrm{k} = \mathbf{L}_\mathrm{k} - \mathbf{P}_I\mathbf{L}_\mathrm{k}\mathbf{P}_J \quad (4)$$

Considering Equation (3) in Equation (4)

$$\begin{aligned} \mathbf{Y}_\mathrm{k} &= \mathbf{G}\mathbf{D}_k\mathbf{H}^t + \mathbf{E}_k - \mathbf{P}_I\mathbf{G}\mathbf{D}_k\mathbf{H}^t\mathbf{P}_J - \mathbf{P}_I\mathbf{E}_k\mathbf{P}_J \\ &= \mathbf{P}_I^{\perp}\mathbf{G}\mathbf{D}_k\mathbf{H}^t\mathbf{P}_J^{\perp} + \mathbf{P}_I^{\perp}\mathbf{E}_k\mathbf{P}_J^{\perp} \end{aligned}$$

The solution shows how the matrix of the residuals is double-centered and the loss function be

$$argmin \sum_{k=1}^{K} \left\| \mathbf{Y}_k - \mathbf{P}_I^{\perp}\mathbf{G}\mathbf{D}_k\mathbf{H}^t\mathbf{P}_J^{\perp} \right\|^2$$

where the solutions $\mathbf{A} = \mathbf{P}_I^{\perp}\mathbf{G}$ and $\mathbf{B} = \mathbf{P}_J^{\perp}\mathbf{H}$ are given by means of the traditional CP algorithm. Therefore, the CP analysis of log-ratio data assures centered loadings matrices for $\mathbf{A}$ and $\mathbf{B}$, but not all loadings matrices are assured to be column-wise orthonormal. These properties suggest using special precaution when the coefficients of these matrices are to be interpreted.

3.2 Procedures for Displaying CP Results of Log-Ratio Data

In the literature, CP results are usually given in the form of tables or plots. Here, for the representation of the CP results of log-ratio data, we propose to use one-mode plots and pre-component plot.

We refer to one-mode plots in which the components of a single mode are plotted against each other in scatter plot, i.e. the first and the second columns of $\mathbf{A}$. To provide some adequate representation of the low-dimensional CP configuration of log-ratio data, it is necessary to carefully choose the axes with respect to which the compositions, or parts or occasions are to be displayed, and to compute coordinates with respect to these axes. In fact, using the coefficients of the loadings matrix as coordinate may lead to misleading plots, providing a distorted representation of the original Aitchison distances. The components for each single mode have not unit-norm and the orthogonality between them are not assured. So if the coefficients of $\mathbf{A}$ are just plotted against each other, without justifying the choice of orthogonal axes, an inadequate representation of the higher-dimensional space $\Re^{JK}$ is given. Following Kiers 2000, it is possible computing an auxiliary orthonormal base for $\mathbf{A}$ and projecting the coefficients onto a plot in which the columns of $\mathbf{W}_A = (\mathbf{C} \otimes \mathbf{B})\dot{\mathbf{I}}^t$ are orthonormal. In other words, one searches for a transformation matrix $\mathbf{T}_A$ for the matrix $\mathbf{W}_A$, such that $\mathbf{W}_A^* = \mathbf{W}_A \mathbf{T}_A$ is column-wise orthonormal, then the loadings matrix $\mathbf{A}$ is transformed by the inverse of $\mathbf{T}_A$, that is, $\mathbf{A}^* = \mathbf{A}\mathbf{T}_A^{-1}$, which are the coefficients to be plotted. Plotting the elements of $\mathbf{A}^*$ assures that the high-dimensional Aitchison distances are correctly represented in the low-dimensional plot within the accuracy of the approximation.

In the same way, an orthonormal base for $\mathbf{B}$ can be given. Given $\mathbf{W}_B = (\mathbf{C} \otimes \mathbf{A})\dot{\mathbf{I}}^t$ and let $\mathbf{T}_B$ be the transformation matrix for $\mathbf{W}_B$, such that $\mathbf{W}_B^* = \mathbf{W}_B \mathbf{T}_B$ is column-wise orthonormal. Then the loadings matrix $\mathbf{B}$ for the parts of compositions is transformed by the inverse of $\mathbf{T}_B$, that is, $\mathbf{B}^* = \mathbf{B}\mathbf{T}_B^{-1}$, and the coefficients of $\mathbf{B}^*$ are plotted. Given the accuracy of the approximation, the distances between the variables into low-dimensional plot of $\Re^{IK}$ indicate the relative variation between them. Of course, an analogous procedure can be used for the loadings matrix of the third mode.

Pre-component plots are used to investigate the relationships between the elements of different modes. It is only across-mode comparisons and it is constructed by plotting the components of the three modes one-by-one. Thus, the first component of the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ is plotted along a single line, where the coefficients are usually placed before unit-mean-square-scaling to compensate for the differences in the number of levels in the components of the three modes (Kroonenberg, 2008, pp.271).

4 Application and Discussion

4.1 Dataset and Analysis

The data are relative to the use of land for different type of agricultural production in 26 European countries: Belgium (be), Bulgaria (bg), Czech Republic (cz), Denmark (dk), Germany (de), Estonia (ee), Ireland (ie), Greece (gr), Spain (es), France (fr), Italy (it), Cyprus (cy), Latvia (lv), Lithuania (lt), Luxembourg (lu), Netherlands (nl), Austria (at), Poland (pl), Portugal (pt), Romania (ro), Slovenia (si), Slovakia (sk), Finland (fi), Sweden (se), Croatia (hr), Norway (no). The agricultural production has been divided in 7 macro-categories: cereals (Cer), fallow and green manure (Fal), fodder from arable land (Fod), industrial crops (Ind), root crops (Roo), vegetables (Veg), other field products (Oth). The data were observed over the years 2001–2005.

If we are interested in analyzing the size of the hectares of land that are used in the above-mentioned countries for different products over the years, it is possible to apply the CP analysis where the information gathered by the first component regarding the absolute magnitude. Differently, when we are interested in studying how those counties have parceled out the hectares of land for the different products during the years, we are recognizing that the size of the hectares of land that each country has used for the agricultural productions is irrelevant. In the latter case, we have interested in the relative, not the absolute magnitudes, thus our original data, which have all positive quantities, are being transformed in compositional data and preprocessed though *clr*-transformation. Afterwards, the CP analysis on these data is carried out. Finally, the results of the analysis are represented through graphic procedures such as one-mode and pre-component plots.

In Figure 1.a it is possible to observe that Austria and Germany have a similar compositional structure in the use of land for the different types of production. On the contrary, Cyprus and Bulgaria are widely different as to management for different kinds of production. With regard to the parts, it is possible to observe that the lowest relative variation in the use of land across the countries and during the years is in the ratio ‘cereals/root crops’, while the ‘industrial crops/vegetables’ ratio presents the highest relative variation across the data (Figure 1.b), followed by 'industrial crops/other field products'. Finally, it is possible to observe that the highest relative variation in the use of land is between the years 2005 and 2002, while the lowest relative variation is between the years 2003 and 2004 (Figure 1.c).

In Figure 2, it is possible to observe that the mean scores per parts of components and per countries have been removed at each time point. Moreover, all the years have positive elements. This means that when the product between the coefficient of one country and one variable is positive, the country has a high percentage of hectares of land used for this kind of production. Otherwise, when the product is negative the country has used a percentage of land lower than

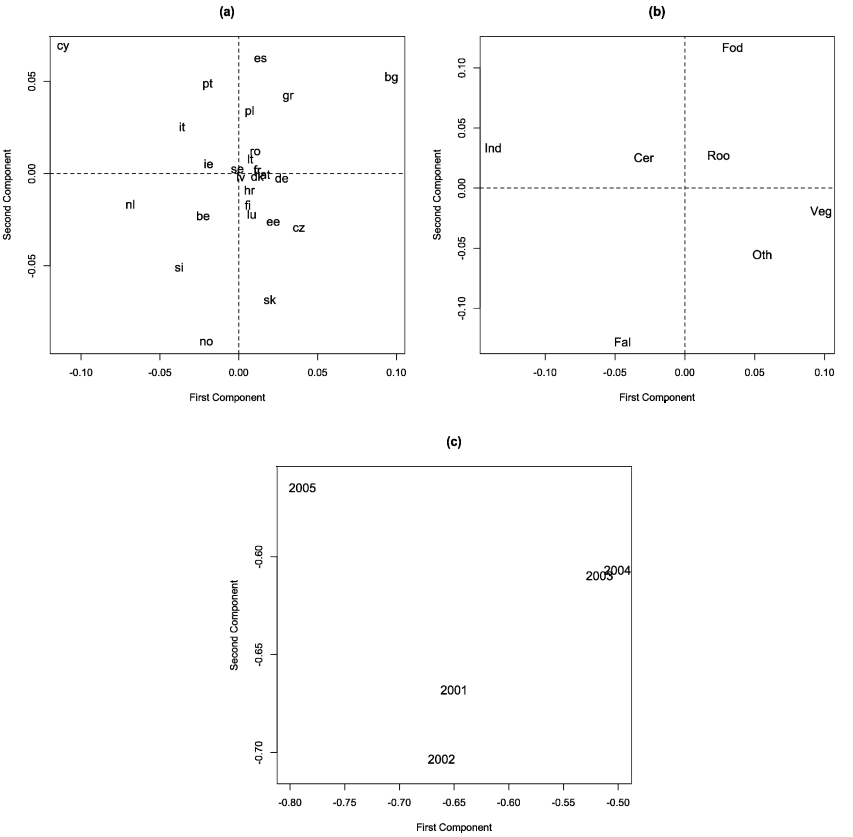


Fig. 1 (a) First-mode plot (countries): Belgium (be), Bulgaria (bg), Czech Republic (cz), Denmark (dk), Germany (de), Estonia (ee), Ireland (ie), Greece (gr), Spain (es), France (fr), Italy (it), Cyprus (cy), Latvia (lv), Lithuania (lt), Luxembourg (lu), Netherlands (nl), Austria (at), Poland (pl), Portugal (pt), Romania (ro), Slovenia (si), Slovakia (sk), Finland (fi), Sweden (se), Croatia (hr), Norway (no). (b) Second-mode plot (parts of components): cereals (Cer), fallow and green manures (Fal), fodder from arable land (Fod), industrial crops (Ind), root crops (Roo), vegetables (Veg), Other field products (Oth). (c) Third-mode plot (years): 2001–2005.

average. Thus, Cyprus and Netherlands have the highest percentages of land that is used for 'vegetables' and 'other field productions' production, as well as the lowest value for 'industrial crops'. Furthermore, Bulgaria has the highest percentage of land used for 'industrial crops'. And finally, countries as Sweden and Latvia have a compositional use of land for agriculture production, which accounts for the average of that of all countries analyzed.

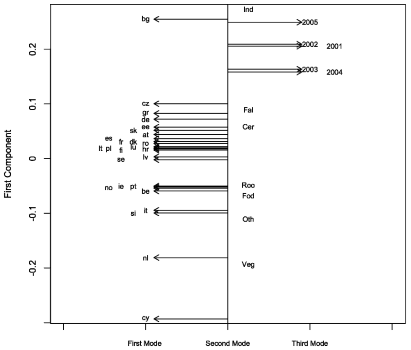


Fig. 2 Per-component plot for the first CP-component for each of the three modes. First mode (countries): Belgium (be), Bulgaria (bg), Czech Republic (cz), Denmark (dk), Germany (de), Estonia (ee), Ireland (ie), Greece (gr), Spain (es), France (fr), Italy (it), Cyprus (cy), Latvia (lv), Lithuania (lt), Luxembourg (lu), Netherlands (nl), Austria (at), Poland (pl), Portugal (pt), Romania (ro), Slovenia (si), Slovakia (sk), Finland (fi), Sweden (se), Croatia (hr), Norway (no); Second mode (parts of components): cereals (Cer), fallow and green manures (Fal), fodder from arable land (Fod), industrial crops (Ind), root crops (Roo), vegetables (Veg), Other field products (Oth); Third mode (years): 2001–2005.

5 Discussion

What do we have to expect when the CP analysis is carry out on a set of positive quantities? Undoubtedly, the first dimension gathers the information regarding the magnitudes. Thus, through the application that has been proposed we were able to observe that all variables stay on the same side of the one-mode plot and the countries are plotted according to the total land used for the agricultural production, therefore just a ranking of countries from the largest to the smallest is given on the first component. At the same time, this dimension explains the highest variability too, so the other information could not be gathered because they are hided by size effect. Furthermore, this analysis yields subcompositional coherence results. Thus, when the CP analysis of log-ratios is run onto a subset of all variables, the Aitchison distance measured between two full compositions is greater, or at least equal to, than the distance between them when considering any subcomposition. It is equal only in the case of non-informative parts that are erased from analysis. For example, an economist, who is interested only in a parts, such as, 'fallow and green manure', 'fodder from arable land', 'industrial crops', 'root crops', commonly forms the subcomposition based on these parts of composition. The CP analysis of log-ratios onto four parts gives, for these four parts, the same results of the CP analysis of log-ratios run on all seven variables. This property is not assured by a traditional PC analysis on compositional data.

One-mode and pre-component plots are very powerful tool for visualizing the PC results of log-ratio data. That is because a very large set of information can be extracted from these kinds of plots. Moreover, the pre-component plot has the property that the three-way array $\underline{\mathbf{Y}}$ can be approximately reconstructed, and in this graphical tool, the coefficients of three modes are, at the same time, represented for each component, so if we have one pre-component plot for the first *F*-components we could reconstruct the loadings matrices for compositions, parts and occasions, that are: $\mathbf{A}\ (I \text{ x } F)$, $\mathbf{B}\ (J \text{ x } F)$ and $\mathbf{C}\ (K \text{ x } F)$, so $\hat{\mathbf{Y}}_A = \mathbf{A}\dot{\mathbf{I}}\left(\mathbf{C} \otimes \mathbf{B}\right)^t$ is the flatted matrix of $\underline{\mathbf{Y}}$, that is, approximately, $\mathbf{Y}_A = \hat{\mathbf{Y}}_A + \mathbf{E}_A$.

To reconstruct the whole compositional data array, we need to know the means of the centered log-ratios as well as the geometric means of each composition and by the inverse *clr*-transformation we can approximately get the original dataset.

Other kinds of graphical display can be used for CP results of log-ratio data, such as joint biplot and trajectory plot (Kiers, 2000). Moreover, we should use the transformations shown in Section 2.1, and consider that the CP components are uniquely determined. These properties must be strictly taken into account before any attempt of interpretation.

References

Aitchison, J.: The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) 44(2), 139–177 (1982)

Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)

Aitchison, J.: On criteria for measures of compositional difference. Mathematical Geology 22(4), 487–511 (1992)

Aitchison, J., Ng, K.W.: The role of perturbation in compositional data analysis. Statistical Modelling 5, 173–185 (2005)

Bro, R.: Multi-way analysis in the food industry: models, algorithms and applications. Ph.D. Thesis. University of Amsterdam and Royal Veterinary and Agricultural University, Denmark (1998)

Caroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via N-way generalization of 'Eckart-Young' decomposition. Psychometrika 35, 283–319 (1970)

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279–300 (2003)

Egozcue, J.J., Barceló-Vidal, C., Martín-Fernández, J.A., Jarauta-Bragulat, E., Díaz-Barrero, J.L., Gallo, M.: Partial least squares for compositional data: an approach based on the splines. Italian Journal of Applied Statistics 15, 349–358 (2003)

Gallo, M.: Discriminant Partial Least Squares analysis on compositional data. Statistical Modelling 10(1), 41–56 (2010)

Gallo, M.: Compositional data and three-mode analysis (2011) (submitted)

Harshman, R.A.: Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-mode factor analysis. UCLA Working Papers Phonet 16, 1–84 (1970)

Hinkle, J., Rayens, W.: Partial least squares and compositional data: problems and alternatives. Chemometrics and Intelligent Laboratory Systems 30, 159–172 (1995)

Kiers, H.A.L.: Some procedures for displaying results from three-way methods. Journal of Chemometrics 14(3), 151–170 (2000)

Kroonenberg, P.M.: Applied Multiway Data Analysis. Wiley, Hoboken (2008)

Martìn-Fernàndez, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V.: Measures of difference. for compositional data and hierarchical clustering. In: Buccianti, A., Nardi, G., Potenza, R. (eds.) IAMG 1998. De Frede, Napoli (1998)

Mateu-Figueras, G.: Elements of Simplicial Linear Algebra and Geometry. In: Pawlowsky, V., Buccianti, A. (eds.) Compositional Data Analysis: Theory and Applications. Wiley, Chichester (2011)

Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment 15(5), 384–398 (2001)

Pawlowsky-Glahn, V. Egozcue, J.J. Tolosana-Delgado, R.: Lecture Notes on Compositional Data Analysis. Girona: Universitat. [Consultat 2 oct 2007]. (2007)

Analyzing AHP Matrix by Robust Regression

Gabriella Marcarelli, Biagio Simonetti, and Viviana Ventre

Università del Sannio
{marcarel,simonetti,ventre}@unisannio.it

Abstract. The Analytic Hierarchy Process (AHP) is a powerful process to help people to express priorities and make the best decision when both qualitative and quantitative aspects of a decision need to be considered. In this paper, in order to eliminate the influence of outliers, we use an approach based on Robust Partial Least Squares (R-PLS) regression for the computation of the values for the weights of a comparison matrix. A simulation study to compare the results with other methods for computing the weights proposed to analyze comparison matrix.

Keywords: Analytic Hierarchy Process, Robust Regression, Simulation Study.

1 Introduction

The Analytic Hierarchy Process (AHP) is a systematic procedure for representing the elements of a multicriteria decision maker (MCDM) problem, hierarchically. By means of AHP a decision problem is broken into smaller parts and then decision makers lead through a series of pairwise comparison judgments to express the relative intensity of the impact of the elements in the hierarchy. The AHP procedure shows how to use judgments and experience to analyze a complex decision problem by combining its qualitative and quantitative aspects into a single framework and generating a set of priorities for alternative courses of action. The process has inherent flexibilities in structuring a problem and in taking different judgments from people. These judgements are converted into numbers. Detailed exposition of the AHP is found in Saaty [10]. Saaty and Vargas [11], illustrate applications of the AHP in various real-life systems. A fundamental problem of decision theory is to derive weights for a set of activities according to the importance. The object of this approach is to use the weights, which we call priorities, to allocate a resource among the activities or, if precise weights cannot be obtained, to rank the most important activities. The AHP falls into the broad category of mathematical and behavioral science interests. The dominance matrices play a central role in the AHP [12]. The aim of our work is to investigate the consequence of changes in the judgments through perturbations on the entire set of judgments. This type of approach leads to the criterion of consistency. Thus, obtaining solutions in our method is not a statistical procedure. However, if we want

A.N. Proto et al. (Eds.): Adv. Dynamic Modeling of Economic & Social Systems, SCI 448, pp. 223–231.
springerlink.com

to compare any solution with the criterion of consistency, we refer to statistical reasoning and perturbations over the entire matrix of judgments. Different approaches are developed in order to evaluate the weights, using statistical procedures. Tucker [15] presents a method for the determination of parameters of a functional relation by factor analysis. Saaty and Vargas [11] have studied the eigenvalue approach with the least squares and logarithmic least squares methods. With inconsistency, just the eigenvalue method is applied. Laininen and Hämäläinen [7], starting from the consideration that the presence of outliers can have a significant influence on the weight estimates given by the eigenvector method and the logarithmic least squares regression, propose a robust version of logarithmic least squares regression in the case early mentioned.

2 Methods of Estimating and Analysis of Priority Vectors

Methods of estimating the vector of priorities were introduced along with clustering analysis to make the process more efficient and consistent. In the AHP the weight ratios are asked for all pairs of attributes. Usually the weights are derived by the principal eigenvector (EV) of the comparison matrix. Other estimation methods, such as the logarithmic least squares method, can be used to derive the weights as well.

In Saaty [10] is described how the eigenvector associated with the principal eigenvalue of a matrix *A* can be obtained as:

$$\lim_{k\to\infty}\frac{\sum_{m=1}^{k}\mathbf{A}^m e}{\sum_{m=1}^{k} e^T\mathbf{A}^m e}=\lim_{k\to\infty}\frac{\mathbf{A}e}{e^T\mathbf{A}e}=Cw$$

Where *e* is the column vector unity, e^T its transpose, and *C* a positive constant and $A^m = n^{m-1}A$ (for consistent matrices). This result allows us to approximate λ_{max} and *w* as accurately as desired but within computational capabilities.

In situations in which accuracy is not the most important factor, the vector of priorities can be approximated by one of three methods: average of normalized columns (ANC), normalization of row averages (NRA) and normalization of the geometric mean of rows (NGM) [11].

Let $\hat{w}_i$ be the priority estimate of the i^{th} activity. According to the first method we have:

$$\hat{w}_i(ANC)=\frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}}$$

The second method yields:

$$\hat{w}_i(NRA) = \sum_{i,j=1}^{n} \frac{a_{ij}}{\sum_{k=1}^{n} a_{ij}}$$

and

$$\hat{w}_i(NGM) = \frac{\left(\prod_{j=1}^{n} a_{ij}\right)^{1/n}}{\sum_{k=1}^{n}\left(\prod_{j=1}^{n} a_{kj}\right)^{1/n}}$$

A drawback of this method is that there is no practical statistical theory behind it.

If the matrix *A* is consistent, the three methods reproduce the original vector of priorities. However, in the inconsistent case ANC is more accurate than the other two methods. The NGM method provides a good estimate of the priorities if accuracy is not of extreme importance.

Other ways of approximating the vector of priorities are discussed in [10].

3 Methods Based on Regression Approaches

The standard method to calculate the weights from an AHP-matrix is to take the eigenvector corresponding to the largest eigenvalue of the matrix, standardizing the sum of the component equal to 1. This is not-statistical approach. A statistical approach is needed if there are random fluctuations in the ratios of the relative importance. All the human decisions are affected by random variations [4]. The use of Regression Methods for analyzing AHP-matrix is not new [2] [1]. The regression techniques resolves a classical statistical problem, to estimate the linear relationship between two sets of variables, $\mathbf{X}_{n,p}$ (explicative variables) and $\mathbf{y}_{n,1}$ (dependent variable) where n is the number of statistical units and q the number of the explanatory variables. The technique which is largely used to solve this problem is the multivariate regression model $\mathbf{y} = a + \mathbf{Xb} + \mathbf{e}$, where *a* is the intercept term, *b(p,1)* the gradient and *e(n,1)* the error term. The parameters are generally computed using least squares criteria. This techniques is well note as Least Squares Regression (LS). Laininen, R.P. Hämäläinen [7] instead of use of LS regression methods use a variant Logarithmic Regression (LOG) that is lesser sensitive in presence of value that are different from corresponding consistent value. In the statistical theory these values are called outliers. Such technique, Robust Logarithm Regression (RLOG) is more stable under random occurrences of outliers compared to EV method.

When $\mathbf{X}'\mathbf{X}$ does not exist (or is unstable) therefore the classical least squares regression model cannot be applied. The solution for overcoming this problem is offered by Partial Least Squares (PLS) Regression [16]. Garthwaite [5] proposed an alternative approach of the PLS based on simple linear regression showing that linear combinations of the explanatory variables can be formed sequentially and related to **y** variable by ordinary least squares regression. Since PLS regression is very sensitive to the presence of outliers in the data, different algorithms of robust version have been proposed. Rousseeuw [9] proposed the Least Median of Squares (LMS) estimator which is obtained from the following optimization set up

$$Minimize\left\{\underset{i}{median}\left(e_i^2\right)\right\}$$

where e_i is the residual of observation *i*.

Simonetti et al. [14], propose a robust version of PLS regression (R-PLS), based on least median square regression [9]. This median square regression is substituted for the least square regression. The R-PLS regression is lesser sensitive to presence of outliers that is in AHP context when the comparisons matrix is non consistent.

4 Inconsistency

A fundamental property of the pairwise comparison matrix is the consistency. If the decision maker is perfectly consistent in making estimates, the matrix *A* will satisfy the following consistency condition:

$$a_{ij} = a_{ik}a_{kj},$$

for each *i,j,k=1,2,. . . n* [10].

In the case of perfectly consistency, the reciprocal matrix *A* has unit rank since every row is a constant multiple of the first row.

If *A* is consistent then all its eigenvalues except one are zero. Since the sum of the eigenvalues of a matrix is equal to its trace, then *n* is the only non zero eigenvalue of *A*.

In the inconsistent case, which is the most common, the pairwise comparisons are not perfect, that is, the entry a_{ij} might deviate from the ratio of the real membership values w_i and w_j [3].

When the entries a_{ij} change slightly then the eigenvalues change in a similar fashion. The maximum eigenvalue is close to n, greater than *n*, while the remaining eigenvalues are close to zero.

In order to find the priority vector of *A* one should find the eigenvector corresponding to the maximum eigenvalue.

4.1 Measures of Consistency

Since small changes in the entries a_{ij} imply a small change in λ_{max}, the deviation of the greatest eigenvalue of A from the dimension of this matrix, n, represents a deviation from consistency. The following expression

$$CI = \frac{\lambda_{\max} - n}{n-1}$$

called Consistency Index, is proposed by Saaty as a measure of the consistency of comparisons. If the coefficient is not within certain acceptable limits, then the comparisons have to be repeated until an acceptable consistency level is reached [10].

Another measure of consistency is the Consistency Ratio:

$$CR = \frac{CI}{RI}$$

The *Random Index* (*RI*) is the average value of *CI* derived from a sample of size 500 of randomly generated reciprocal matrices with entries from the set [1/9,1/8,. . . .,1,2,. . . ,8,9].

The ratio *CR* should be about 10 percent or less for acceptable overall consistency. Otherwise, the quality of the judgmental data should be improved, perhaps by revising the manner in which questions are posed to make the pairwise comparisons.

Another measure of consistence is the so called Consistency Measure (*CM*), defined as

$$CM = \frac{2}{n(n-1)} \sum_{i>j} \frac{\bar{r}_{ij} - \underline{r}_{ij}}{\left(1+\bar{r}_{ij}\right)\left(1+\underline{r}_{ij}\right)}$$

where $r_{ij} = \max a_{ik} a_{kj},\ k = (1,....,n)$ stands for the extended bound of the comparison matrix element a_{ij} and r_{ij} is the inverse of r_{ji}. This measure ranges from 0 to 1 and its value increases as the inconsistency of the comparison matrix increases [13].

5 Influence of the Outliers on the Estimates of the Weights

The AHP is a method for formalizing decision making where there are limited number of choices but each has a number of attributes and it is difficult to formalize some of those attributes.

The AHP has been used in a large number of applications to provide some structure on a decision making process. Note that the system is somewhat ad-hoc (for example a scale 1-9 range) and there are number of “hidden assumption”.

Furthermore, it is possible to manipulate the rankings to get a preferred outcome (by "lying"). In particular, the method can be influenced by random errors and the values of the ratios of the weights may be exceptionally different from the corresponding consistent value: in this case the statement is called an outlier [8].

This kind of an outlier has powerful influence on the values of the estimated weights given by the standard eigenvector method and the least square regression method. It has been demonstrated that the robust regression technique is a method which is lesser sensitive to outliers [7] than the eigenvector method.

6 Example: Kangas Data

Consider the following consistent (CR=0) 4x4 AHP-matrix from a real case study [6].

Table 1 A consistent AHP-matrix

1	5	7	9
1/5	1	7/5	9/5
1/7	5/7	1	9/7
1/9	5/9	7/9	1

For such table, all methods considered in this paper, give the same results: 0.6878, 0.1376, 0.0983, 0.0764. Laininen, R.P. Hämäläinen [7] call these the correct weights (CW). Let us suppose that the judgment in comparison of the entities number three and four is not the consistent relation 9/7 but it is 7/9 as shown in [7].

Table 2 An inconsistent AHP-matrix

1	5	7	9
1/5	1	7/5	9/5
1/7	5/7	1	7/9
1/9	5/9	9/7	1

The CR for the data in table 2 is CR=0.12. The results of the estimate with the five methods considered here are shown in table 3.

Table 3 Estimates comparison

EV	LS	LS	LOG	RLOG
0.6878	0.6878	0.6878	0.6878	0.6878
0.1376	0.1376	0.1376	0.1377	0.1376
0.0872	0.0872	0.0864	0.0868	0.0983
0.0874	0.0874	0.0882	0.0877	0.0764

We can note that the RLOG and R-PLS methods give the same results and are exactly the same of the correct weights.

7 A simulation Study

From the previous example seems that the RLOG and R-PLS regression are the more stable methods in presence of outliers, reaching the same value of the CW. In this paragraph, we compare the two methods using a simulation study. Let consider the data of Table 1. Then we generate from this matrix, adding random errors, new matrices with fixed range of CR (0.00–|0.02; 0,02–|0,06; 0,06–|0,1). We have considered CR value lesser or equal than 0.1 because if CR is greater than 0.10 then a re-examination of the pairwise judgments is recommended until a CR less than or equal to 0.10 is achieved. For each range we generate 200 matrices. We show, by introducing a simulation study, how the estimates of the weights calculated by the R-PLS regression is lesser sensitive to the occurrence of the outliers than the RLOG. For each range of CR we compute the mean and the standard deviation of the 200 samples, as shown in tables 4, 5, 6.

Table 4 Comparison of estimated weights between RLOG and R-PLS for simulated matrices with a 0 <CR≤ 0.02

0 <CR≤ 0.02			
Mean RLOG	*St. Dev. RLOG*	*Mean R-PLS*	*St. Dev. R-PLS*
0.691	0.084	0.688	0.045
0.128	0.044	0.133	0.024
0.080	0.041	0.099	0.052
0.099	0.023	0.078	0.011

From the data in tables 4, 5, 6, is very clear that the mean of the estimate computed with R-PLS is more closed to the CW than RLOG with a very low variance.

Table 5 Comparison of estimated weights between RLOG and R-PLS for simulated matrices with a 0.02 <CR≤ 0.06

0.02 <CR≤ 0.06			
Mean RLOG	*St. Dev. RLOG*	*Mean R-PLS*	*St. Dev. R-PLS*
0.699	0.091	0.698	0.055
0.110	0.085	0.135	0.029
0.078	0.062	0.094	0.063
0.111	0.041	0.071	0.022

Table 6 Comparison of estimated weights between RLOG and R-PLS for simulated matrices with a 0.06 <CR≤ 0.1

0.06 <CR≤ 0.1			
Mean RLOG	*St. Dev. RLOG*	*Mean R-PLS*	*St. Dev. R-PLS*
0.699	0.091	0.698	0.055
0.110	0.085	0.135	0.029
0.078	0.062	0.094	0.063
0.111	0.041	0.071	0.022

8 Conclusions

The use of statistical theory is needed in the estimate of the weights for AHP-matrices if there are random fluctuations in the ratios of the relative importance. We have shown, using a simulation study, that R-PLS regression is a more robust technique compared to the other regression methods applied in AHP context. Further investigations on the influence of the outliers on the estimates, considering different range of coefficient ratio index, are in progress.

References

[1] Alho, J.M., Kangas, J.: Analyzing Uncertainties in Experts' Opinions of Forest Plan Performance. Forest Science 43(4) (1997)
[2] Crawford, G., Williams, C.: A note on the analysis of subjective judgments matrices. Journal of Mathematical Psychology 29, 387–405 (1985)
[3] D'Apuzzo, L., Marcarelli, G., Squillante, M.: Analysis of Qualitative and Quantitative Rankings in Multicriteria Decision Making. In: Lux, T., Faggini, M. (eds.) Economics from Tradition to Complexity, pp. 157–170. Springer, Heidelberg (2009) ISBN: 978-88-470-1082-6, doi:10.1007/978-88-470-1083-3_10
[4] Fischoff, B., Slovic, B., Lichtenstein, S.: Knowing what you want: measuring labile values. In: Wallsten, T.S. (ed.) Cognitive Processes in Choice and Decision Behavior. Lawrence Erlbaum Associates, Hillsdale (1980)
[5] Garthwaite, P.H.: An Interpretation of Partial Least Squares. Journal of the American Statistical Association 89, 122–127 (1994)
[6] Kangas, J., Matero, J., Pukkala, T.: Using the analytic hierarchy process in planning of multiple-use forestry. A case study. In: Finnish Forest Research Institute Research, Notes 412 (1992)
[7] Laininen, P., Hämäläinen, R.P.: Analyzing AHP-matrices by regression. European Journal of Operation Research 148, 514–524 (2003)
[8] Marcarelli, G., Simonetti, B.: Estimation of priorities in the AHP through Taxicab decomposition. In: Advances and Applications in Statistical Sciences (to appear) ISSN 0974-6811

[9] Rousseeuw, P.J.: Least median of squares regression. Journal of the American Statistical Association 79, 871–888 (1984)
[10] Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
[11] Saaty, T.L., Vargas, L.G.: The Logic of Priorities. In: Applications in Business, Energy, Health, and Transportation. Kluwer-Nijhoff, Boston (1982)
[12] Shepard, R.N.: A Taxanomy of Some Principal Types of Data and Multidimensional Methods for their Analysis. In: Shepard, R.N., Romney, A.K., Nerlove, S.B. (eds.) Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, New York, vol. 1, pp. 21–47 (1972)
[13] Salo, A.A.: Inconsistency analysis by approximately specified priorities. Math. Comput. Modelling 17, 123–133 (1993)
[14] Simonetti, B., Mahdi, S., Camminatiello, I.: Robust PLS Regression Based on Simple Least Squares Regression. In: Proceedings of the Conference, MTISD 2006, Procida, September 28-30 (2006)
[15] Tucker, L.R.: Determination of Parameters of a Functional Relation by Factor Analysis. Psychometrika 23, 1 (1958)
[16] Wold, H.: Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. In: Gani, J. (ed.) Perspectives in Probability and Statistics, pp. 520–540. Academic Press, London (1975)

Author Index