ECML/PKDD-2002 Workshop Programme

Workshop W5

Mining Official Data

Helsinki, August 20th, 2002

Workshop Notes

Edited by

Paula Brito
Faculty of Economics,
University of Porto,
Porto, Portugal
mpbrito@fep.up.pt

Donato Malerba
Dipartimento di Informatica
University of Bari
Bari, Italy
malerba@di.uniba.it

Program Commitee:

Timo Alanko, Statistical R&D Unit, Statistics Finland, Helsinki, Finland
Edwin Diday, CEREMADE, Paris-9 Dauphine University, France
Floriana Esposito, Department of Informatics, University of Bari, Italy
Paulo Gomes, National Institute of Statistics (INE), Lisbon, Portugal
Haralambos Papageorgiou, Department of Mathematics, University of Athens, Greece
Willi Klösgen, Fraunhofer Inst. for Autonomous Intelligent Systems, Sankt Augustin, Germany
Carlos Marcelo, National Institute of Statistics (INE), Lisbon, Portugal
Michael May, Fraunhofer Inst. for Autonomous Intelligent Systems, Sankt Augustin, Germany
Monique Noirhomme, Institut d'Informatique, University Notre-Dame de la Paix, Namur, Belgium
Mireille Summa, CEREMADE, Paris-9 Dauphine University, France
Ian Turton, Centre for Computational Geography, University of Leeds, UK

Home page: http://www.di.uniba.it/~malerba/activities/mod02/index.html

Acknowlegements:

KDNET: The Knowledge Discovery Network of Excellence

Preface

In statistics, the term "*official data*" denotes data collected in censuses and statistical surveys by National Statistics Institutes (NSIs), as well as administrative and registration records collected by government departments and local authorities. They are used to produce *"official statistics"* for the purpose of making policy decisions, and to facilitate the appreciation of economic, social, demographic, and other matters of interest to the governments, government departments, local authorities, businesses, and to the general public. For instance, population and economic census information is of great value in planning public services (education, fund allocation, public transport), as well as in private businesses (placing new factories, shopping malls, or banks, as well as marketing particular products). Moreover, survey data on specific topics, such as labour force, time use, household budget, are regularly collected by NSIs to keep updated information on some economic and social phenomena.

The application of data mining techniques to official data has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society. Nevertheless, it is not straightforward and requires a challenging methodological research, which is still in an initial stage. In particular, to develop successful applications of data mining techniques to official data, the following issues must be dealt with:

- *Aggregated data.* NSIs make a great effort in collecting census data, but they are not the only organizations that analyse them: data analysis is often done by different institutes. By law, NSIs are prohibited from releasing individual responses to any other government agency or to any individual or business, so data are aggregated for reasons of privacy before being distributed to external agencies and institutes. Data analysts are confronted with the problem of processing data that go beyond the classical framework, as in the case of data concerning more or less homogeneous classes or groups of individuals (second-order objects or macro-data), instead of single individuals (first-order objects or micro-data). The extension of classical data analysis techniques to the analysis of second-order objects is one of the main goals of a novel research field named "symbolic data analysis".
- *Data quality.* There are different ways to determine data quality, including the percentage of errors in data and the use of an unbiased sampling procedure. One problem with official data is the human error involved in inputting. Some outlier detection techniques developed in data mining can be used to find errors in collected data. Data can also be missing, which is the biggest problem with official data. In this case, clustering methods can be used to replace missing values. Finally, data mining techniques can also be used to control the sample bias, before disseminating official aggregated data used in further processing.
- *Timeliness.* This can be considered another aspect of data quality. Public and private institutions are currently urged to reduce the delay between the time of data collection and the moment in which decisions are made according to some statistical indicators. A typical example is the inflation rate computed by the European Institute of Statistics (Eurostat) and the decision made by the Central Bank of Europe (BCE) on the tax rate.

A timely delivery of data analysis results may involve the synthesis of new indicators from official data, the design of different infrastructures for timely data collection, or the application of 'anytime algorithms', which provide the data miner with a ready-to-use model at any time after the first few examples are seen and guarantee a smooth quality, increasing with time.

- *Geo-referenciation*. The practice of geo-referencing census data has increasingly spread over the last few decades and the techniques for attaching socio-economic data to specific locations have markedly improved at the same time. In the UK, for instance, household expenditure data are provided for each enumeration district (ED), the smallest areal unit for which census data are published. At the same time, vectorized boundaries of the 1991 census EDs enable the investigation of socio-economic phenomena in association with the geographical location of EDs. These advances cause a growing demand for more powerful data analysis techniques that can link population data to their spatial distribution.
- *Metadata.* In statistics, metadata concerns the information used for the most correct understanding of statistical data and their related analysis. They mainly refer to explanations-definitions-procedures that are followed from the designing phase up to the phase of a publication of survey's results. Examples of metadata are the various statistical populations, sampling techniques, definitions of nomenclatures, classifications, monetary units and so on. The basic use of metadata is in interpreting and validating data, as well as in finding and accessing relevant information. However, metadata can also be used for analysis purposes. This is notably the case of research studies performed on government and official statistics, since macro data are of little use to any data consumer if they are not accompanied by additional information, such as what they represent, how they were collected and manipulated and so on.

The workshop maintains a balance between theoretical issues and descriptions of case studies to promote synergy between theory and practice. It aims to be a highly communicative meeting place for researchers working on similar topics, but coming from different communities. Topics of interest include, but are not limited to:

- Methodologies and policies for the analysis of official data
- Confidentiality protection in data mining
- Mining aggregated data and Symbolic data analysis
- Data mining for quality control in data capture and transformation
- Data mining techniques for outlier detection
- Data mining techniques for qualitative comparison of statistics
- Infrastructures for timely collection/delivery of official data/statistics
- Anytime data mining algorithms for timely delivery of official statistics
- Spatial data mining of official/business data
- Infrastructures for the provision of metadata
- Use of meta-data in data mining techniques
- Application of meta-data to the validation of data mining results
- Case studies of mining official data
- Descriptions of official data sources and related data mining problems

We wish to thank the members of the Program Committee for their assistance in setting up this workshop and in reviewing submitted papers. We wish also to thank the ECML/PKDD 2002 Workshop Chairs Hendrik Blockeel and Jean-François Boulicaut, for supporting the organisation of this workshop and the Knowledge Discovery Network of Excellence (KDNET) for the economical support. Finally, we wish to thank authors and invited speakers for their excellent contributions in promoting discussion and the development of new ideas and methods on the workshop topics.

Paula Brito and Donato Malerba
July, 2002

Table of Contents

Invited Talks

Regular Papers

Schedule

8:30	–	9:00	Registration
9:00	–	9:10	*P. Brito and D. Malerba* Opening remarks
9:10	–	9:55	Invited talk: *E. Diday(CEREMADE, Univ. Paris-Dauphine)* An introduction to Symbolic Data Analysis and the SODAS software
9:55	–	10:20	*S. Frutos, E. Menasalvas, C. Montes,J. Segovia* Calculating economic indexes per household and censal section from offical Spanish databases
10:20	–	10:45	*C. Drummond, S. Matwin, C. Gaffield* Inferring and revising theories with confidence: data mining the 1901 Canadian census
10:45	–	11.00	Coffee break
11:00	–	11:25	*W. Klösgen, M. May* Census data mining – an application
11:25	–	11:50	*G. Paass, J. Kindermann* Bayesian regression mixtures of experts using MCMC
11:50	–	12:15	*D. Malerba, F.A. Lisi, A. Appice, F. Sblendorio* Mining spatial association rules in census data: a relational approach
12:15	–	12:40	*D. Rodrigues, F. Vala, J. Monteiro* Hinterlands delimitation of *Lisboa e Vale do Tejo* cities
12.40	–	13:05	*C. Soares, P. Brazdil, C. Pinto* Machine learning and statistics to detect errors in forms: competition or cooperation?
13:05	–	14:30	Lunch
14:30	–	15:15	Invited talk: *Ph. Nanopoulos & J. King (EUROSTAT)* Important issues on statistical confidentiality methods
15:15	–	15:40	*G. D'Angiolini* Developing a metadata infrastructure for official data: the ISTAT experience
15:40	–	16:05	*R. Sund* Utilization of administrative registers using statistical knowledge discovery
16:05	–	16:15	Closing remarks

An Introduction to Symbolic Data Analysis and the Sodas Software

Edwin Diday

University Paris 9 Dauphine
Ceremade. Pl. Du Mle de L. de Tassigny. 75016 Paris, FRANCE

Abstract. The data descriptions of the units are called "symbolic" when they are more complex than standard ones due to the fact that they contain internal variation and are structured. Symbolic data arise from many sources, for instance in order to summarize huge Relational Data Bases by their underlying concepts. "Extracting knowledge" means getting explanatory results, that why, "symbolic objects" are introduced and studied in this paper. They model concepts and constitute an explanatory output for data analysis. Moreover they can be used in order to define queries of a Relational Data Base and propagate concepts between Data Bases. We define "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis to symbolic data tables as input in order to find symbolic objects as output. Any SDA is based on four spaces: the space of individuals , the space of concepts, the space of descriptions modelling individuals or classes of individuals, the space of symbolic objects modelling concepts. Based on these four spaces, new problems appear such as the quality, robustness and reliability of the approximation of a concept by a symbolic object, the symbolic description of a class, the consensus between symbolic descriptions, and so on. In this paper we give an overview on recent development in SDA. We present some tools and methods of SDA and introduce the SODAS software prototype (issued from the work of 17 teams of nine countries involved in an European project of EUROSTAT) .

1 Introduction

As input, when large data sets are aggregated into smaller more manageable data sizes we need more complex data tables called "symbolic data tables" because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical value.
In a symbolic data table, a cell can contain a distribution (Schweitzer (1985) says that "distributions are the number of the future!"), or intervals, or several values linked by a taxonomy and logical rules. The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data table is

increasing in order to get more accurate information and summarize extensive data sets contained in Data Bases.

Since the first papers announcing the main principles of Symbolic Data Analysis (Diday 1987a, 1987b, 1989) much work has been done up to the most recent book published by Bock and Diday (2000) and the proceedings of IFCS'2000 (Kiers et al. 2000) which contains a large chapter devoted to this field. In factorial analysis, Cazes, Chouakria, Diday and Schechtman (1997) have defined a principal component analysis of individuals described by a vector of numerical intervals. In the same direction is the work by Verde and De Carvalho (1998) that takes care on given dependence rules (see also Lauro and Palumbo 1998). In the case where the individuals are described by symbolic data, Conruyt (1994) in the case of structured data, Ciampi et al. (1995), Périnel (1996), have developed an extension of standard decision trees. In the same direction is the work by Bravo and Garcia-Santesmases (1998) on "segmentation trees for stratified data" and Rasson and Lissoir (1998). See also (Auriol 1995) for a link with the domain of "Case Based Reasoning". In order to select the symbolic variables which distinguish at the best individuals or classes of individuals, several works have been done such as Vignes (1991) and more recently Ziani (1996). It is often useful to calculate dissimilarities between symbolic objects; in that direction mention should be made of Gowda and Diday (1992), De Carvalho (1994, 1998a). A complete review is reported in the work by Esposito et al. (2000). If each cell of the data table is a random variable represented by a histogram (for instance, the histogram of the inhabitant age of a town), a histogram of histogram can be calculated for instance, by taking care of rules between the variable values in De Carvalho (1998b), or by using the capacity theory (Diday & Emilion 1995, 1997, Diday et al. 1996). Noirhomme and Rouard (1998, 2000) give a way of representing multidimensional symbolic data (see also Gigout 1998).

Starting from standard data has been proposed a way for extracting symbolic objects from a factorial analysis (Gettler-Summa 1992), and a way for extracting symbolic objects from a partition (Stephan et al. 2000). Starting from time-series, Ferraris, Gettler-Summa, Pardoux, Tong (1995), have defined a way for providing symbolic objects (see also Gettler-Summa & Pardoux 2000).

More recently, several dissertations have been presented in the Paris 9 - Dauphine University. Mfoumoune (1998) for the sequential building of a pyramid where each node is associated to a symbolic object. Chavent (1997), in order to build a partition of a set of symbolic objects by a top-down algorithm which provides also a symbolic object associated to each obtained class (see chapter 11 in Bock , Diday (2000)). Stéphan (1998) for extracting symbolic objects from a data base (see also Stéphan et al. 2000). Hillali (1998) for describing classes of individuals described by a vector of probability distributions. Pollaillon (1998), for extending Galois lattices and extracted pyramid to symbolic data at input and "complete" symbolic objects at output (Pollaillon 2000). Tang (1998) for extending Factorial Correspondence Analysis and O. Rodriguez (2000) for extending regression and Multidimensional Scaling to interval data.

1.1 The Input of a Symbolic Data Analysis: a "Symbolic Data Table"

"Symbolic data tables" constitute the main input of a Symbolic Data Analysis. They are defined in the following way: columns of the input data table are «symbolic variables» which are used in order to describe a set of units called "individuals". Rows are called «symbolic descriptions» of these individuals because they are not as usual, only vectors of single quantitative or categorical values. Each cell of this «symbolic data table» contains data of different types:

(a) Single quantitative value: for instance, if «height» is a variable and w is an individual: height(w) = 3.5.
(b) Single categorical value: for instance, town(w) = London.
(c) Multi-valued: for instance, in the quantitative case height(w) = {3.5, 2.1, 5} means that the height of w can be either 3.5 or 2.1 or 5. Notice that (a) and (b) are special cases of (c).
(d) Interval: for instance height(w) = [3, 5], which means that the height of w varies in the interval [3, 5].
(e) Multi-valued with weights: for instance a histogram or a membership function (notice that (a) and (b) are special cases of (e) when the weights are equal to 1 or 0).

Variables can be:

(f) Taxonomic: for instance, «the colour» is considered to be "light" if it is "yellow", "white" or "pink".
(g) Hierarchically dependent: for instance, we can describe the kind of computer of a company only if it has a computer, hence the variable "does the company has computers? " and the variable " kind of computer " are hierarchically linked.
(h) With logical dependencies, for instance: «if age(w) is less than 2 months then height(w) is less than 10».

Many example of such symbolic data are given in the chapter 3 in (Bock & Diday 2000).

Sources of Symbolic Data. Symbolic data are generated when we summarize huge sets of data. The need of such summary can appear in different ways, for instance, from any query to a data base which induces categories and descriptive variables. These categories can be, for instance, simply the towns or in a more complex way, the socio-professional categories (SPC) crossed with categories of age (A) and regions (R). Hence, in this last case, we obtain a new categorical variable of cardinality $|\text{SPC}| \times |A| \times |R|$ where $|X|$ is the cardinality of X. The descriptive variables of the households can then be used in order to describe these categories by symbolic data.
Symbolic Data can also appear after a clustering in order to describe in an explanatory way (by using the initial variables) the obtained clusters.
Symbolic data may also be "native" in the sense that they result from expert knowledge (scenario of traffic accidents, type of emigration, species of insects, ...), from the probability distribution, the percentiles or the range of any random variable associated to each cell of a stochastic data table, from time series (in representing each time series by the histogram of its values or in describing intervals of time), from confidential data (in order to hide the initial data by less accuracy), etc. They result

also from Relational Data Bases, in order to study a set of units whose description needs the merging of several relations as is shown in the following example.

1.2 Output of Symbolic Data Analysis

Most of the symbolic data analysis algorithms give in their output the symbolic description "d" of a class of individuals (which are the partial or complete extent of a given concept), by using a "generalization" process. By starting with this description, symbolic objects model the underlying concept and give a way, to find at least, the individuals of this class.

Example: The age of two individuals $w1$, $w2$ which satisfy a given concept (for instance they leave in the same town), are age($w1$) = 30, age($w2$) = 35, the description of the class C = {$w1$, $w2$} obtained by a generalization process can be [30, 35]. The extent of this description contains at least $w1$ and $w2$ but may contain other individuals. In this simple case the symbolic object "s" is defined by a triple: $s = (a, R, d)$ where d = [30, 35], R = "$\in$" and "a" is the mapping: $W \rightarrow$ {true, false} such that $a(w)$ = the true value of "age(w) R d" denoted with [age(w) R d]. An individual w is in the extent of s if $a(w)$ = true.

More formally (see figure 1), let W be a set of individuals, D a set containing descriptions of individuals d_w or of a class of individuals d_C, "y" a mapping defined from W into D which associates to each $w \in W$ a description $d_w \in D$ from a given symbolic data table. We denote by R, a relation defined on D. It is defined by a subset W of $D \times D$. If $(x, y) \in W$ we say that x and y are connected by R and this is denoted by $x\ R\ y$. More generally we say that $x\ R\ y$ take its value in a set L. We can have L={true,false}, in this case [$d'\ R\ d$] = true means that there is a connection between d and d'. We can also have L= [0, 1] if d is more or less connected to d'. In this case, [$d'\ R\ d$] can be interpreted as the "true value" of $x\ R\ y$ or " the degree to which d' is in relation R with d". For instance, $R \in \{=, \equiv, \leq, \subseteq\}$ or R is an implication, a kind of matching taking care of missing values, etc. R can also use a logical combination of such operators.

2 Symbolic Objects

A «symbolic object» is defined by a description "d", a relation "R" for comparing d to the description d_w of an individual and a mapping "a" called "membership function". More formally: "a symbolic object is a triple $s = (a, R, d)$ where R is a relation between descriptions, d is a description and a is a mapping defined from W in L depending on R and d".
Symbolic Data Analysis concerns usually classes of symbolic objects where R is fixed, "d" varies among a finite set of coherent descriptions and "a" is such that: $a(w)$=[$y(w)\ R\ d$] which is by definition the result of the comparison of the description of the individual w to d. More generally, many other cases can be considered. If, for instance, the mapping "a" is of the following kind: $a(w) = [h_e(y(w))\ h_j(R)\ h_i(d)]$

where the mappings h_e , h_j and h_i are "filters" which will be discussed hereunder. There are two kinds of symbolic objects:

- «Boolean symbolic objects» if $[y(w)\ R\ d] \in L = \{\text{true, false}\}$. In this case, if $y(w) = (y_1,...,y_p)$, the y_i are of type (a) to (d), defined in section 1.
  Example: Let be $a(w) = [y(w)\ R\ d]$ with R: $[\ d'\ R\ d\ ] = \vee_{\ i=1,\ 2}\ [d'_i\ R_i\ d_i\ ]$ where $\vee$ has the standard logical meaning and $R_i = \subseteq$. If $y(w)$ = (colour(w), height(w)), d=({red, blue, yellow}, [10,15])=(d_1, d_2), colour(u)={red, yellow}, height(u)={21}, then $a(u)$=[colour(u) $\subseteq$ {red, blue, yellow}]$\vee$ [height(u) $\subseteq$ [10,15]]= true $\vee$ false = true.
- «Modal symbolic objects» if $[y(w)\ R\ d] \in L = [0,1]$.
  Example: Let be $a(u) = [y(u)\ R\ d]$ where for instance R: $[d'\ R\ d\ ]$ = Max $_{\ i=1,\ 2}$ $[d'_i\ R_i\ d_i\ ]$. The choice of the Max is among many other possible choices related to copulas theory (Diday 2000). The "matching" of two probability distributions is defined for two discrete probability distributions $d'_I = r$ and $d_i = q$ of k values by: $r\ R_i\ q = \sum_{j=1,k}\ r_j\ q_j\ e^{(r_j\ -\ min\ (r_j,\ q_j))}$. By analogy with the Boolean case we denote $[\ d'R\ d\ ]= \vee^*_{\ i=1,\ 2}[\ d'_i\ R_i\ d_i\ ]$ where $\vee^*$ = Max . With these definitions it is possible to calculate the mapping "a" of a symbolic object $s = (a, R, d)$ where SPC means «socio-professional-category» and d = ({(0.2)12, (0.8)[20 ,28]}, {(0.4)employee, (0.6)worker}) by: $a(u)$ = [age(u) R_1{(0.2)12, (0.8) [20 ,28]}] $\vee^*$ [SPC(u) R_2{(0.4)employee, (0.6)worker}].
  Notice that in this example the weights (0.2), (0.8), (0.4), (0.6) represent frequencies but more generally other kinds of weights may be used as "possibilities", "necessities", "capacities", etc. Notice that the R_i depends on this choice, (Diday 1995).

2.1 Syntax of Symbolic Objects in the Case of "Assertions"

If the initial data table contains p variables we denote $y(w) = (y_1(w),..., y_p\ (w))$, $D = (D_1,...,D_p)$, $d \in D$: $d = (d_1,..., d_p)$ and $R' = (R_1,...,R_p)$ where R_i is a relation defined on D_i. We call «assertion» a special case of a symbolic object defined by $s = (a, R, d)$ where R is defined by $[d'\ R\ d] = \wedge_{\ i=1,\ p}\ [\ d'_i\ R_i\ d_i]$ where "$\wedge$" has the standard logical meaning and "a" is defined by: $a(w) = [\ y(w)\ R\ d\ ]$ in the Boolean case. Notice that considering the expression $a(w) = \wedge_{\ i=1,\ p}\ [y_i\ (w)\ R_i\ d_i]$ we are able to define the symbolic object $s = (a, R, d)$. Hence, we can say that this explanatory expression defines a symbolic object called "assertion".

For example, a Boolean assertion is: $a(w)$ = [age(w) $\subseteq$ {12, 20 ,28}] $\wedge$ [SPC(w) $\subseteq$ {employee, worker}]. If the individual u is described in the original symbolic data table by age(u)={12, 20} and SPC (u) = {employee } then: $a(u)$ = [{12, 20 }$\subseteq$ {12, 20 ,28}] $\wedge$ [{employee}$\subseteq${employee, worker}]= true.

In the modal case, the variables are multi-valued and weighted, an example is given by $a(u) = [y(u)\ R\ d\ ]$ with $[d'\ R\ d\ ] = f(\{[y_i(w)\ R_i\ d_i]\}_{i=1,...,p})$ where for instance, $f(\{[y_i(w)\ R_i\ d_i]\}_{i=1,...,p}) = \Pi_{\ i=1,\ 2}\ [\ d'_i\ R_i\ d_i\ ]$ where in case of probability distributions, the "matching" is defined for two discrete density distributions $d'_i = r = (r_1, ...,r_k)$ and $d_i = q = (q_1, ... , q_k)$ of k values by: $r\ R_i\ q = \sum_{j=1,k}\ r_j\ q_j\ e^{(r_j\ -\ min(r_j,q_j))}$.

By analogy with the Boolean case we denote $[\, d' R\, d\,] = \wedge^*_{i=1,2} p_i [\, d'_i\, R_i\, d_i\,]$ where the meaning of "$\wedge^*$" is given by the definition of the mapping "*f*". For instance, with these choices, a modal assertion $I = (a, R, d)$ is completely defined by the equality: $a(w)$ = [age(w) R_1 {(0.2)12, (0.8) [20 ,28]}] $\wedge^*$ [SPC(w) R_2 {(0.4)employee, (0.6)worker}].

Extent of a symbolic object s. In the Boolean case, the extent of a symbolic object is denoted *Ext*(I) and defined by the extent of a, which is: $Extent(a) = \{w \in W / a(w) = \text{true}\}$. In the modal case, given a threshold α, it is defined by $Ext_a(s) = Extent_a(a) = \{w \in W / a(w) \geq \alpha\}$.

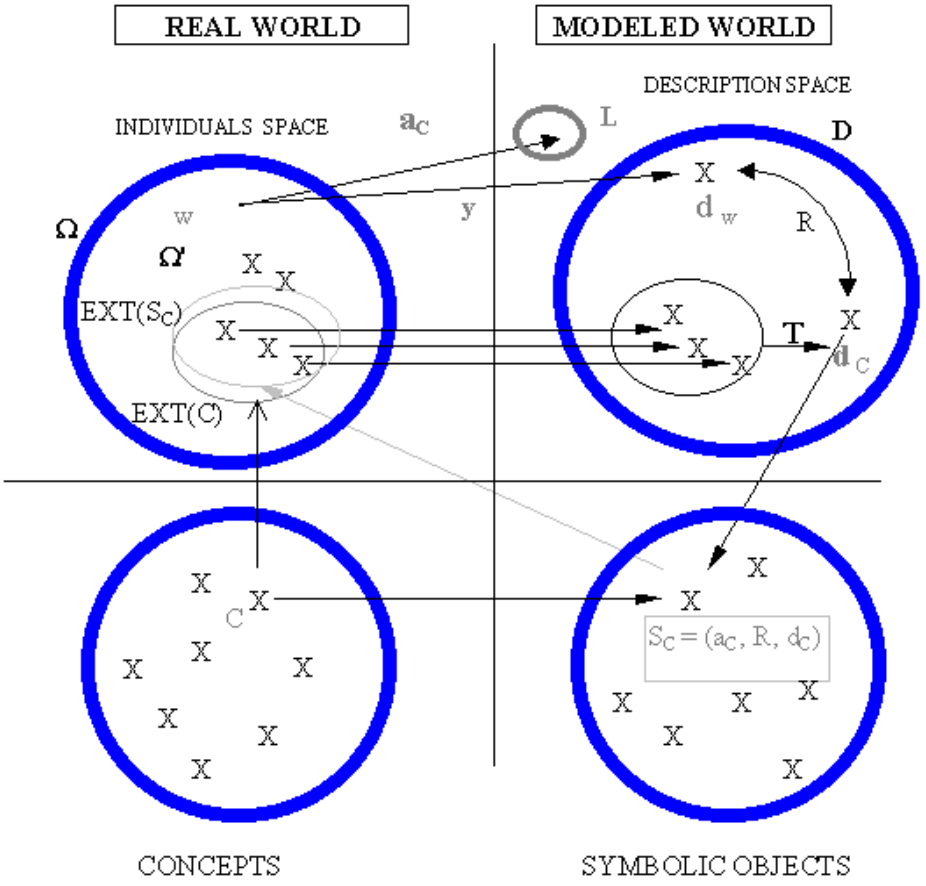


Fig. 1. Modeling by a symbolic object of a concept known by its extent

2.2 Underlying Structures of Symbolic Objects: a Generalized Conceptual Lattice

Under some assumptions on the choice of R and T (for instance $T \equiv$ Max if $R \equiv \leq$ and $T \equiv$ Min if $R \equiv \geq$) it can be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Diday 1991, Brito 1994, Diday & Emilion 1995, 1997), Polaillon & Diday (1997), Polaillon (1998), Bock & Diday (2000)), where the vertices are closed sets defined thereunder by «complete symbolic objects». More precisely, the associated Galois correspondence is defined by two mappings F and G:

- F: from $P(W)$ (the power set of W) into S (the set of symbolic objects) such that $F(C)= s$ where $s = (a, R, d)$ is defined by $d = T_{c \hat{I} C}\, y(C)$ and so $a(w) = [y(w)\ R\ T_{c \hat{I} C}\ y(C)]$, for a given R. For example, if $T_{c \hat{I} C}\, y(C) = \cup_{c \in C}\, y(C)$, $R \equiv$ «$\subseteq$», $y(u) = \{$pink, blue$\}$, $C = \{$c, c'$\}$, $y(C) = \{$pink, red$\}$, $y(c') = \{$blue, red$\}$, then $a(u) =[y(w)\ R\ T_{c \hat{I} C}\ y(C)] = [\{$pink, blue$\} \subseteq \{$pink, red$\} \cup \{$blue, red$\}) = \{$pink, red, blue$\}] =$ true and $u \in Ext\ (s)$.
- G: from S in $P(W)$ such that: $G(s) = Ext\ (s)$.

A «complete symbolic object» s is such that $F(G(s)) = s$. Such objects can be selected from the Galois lattice but also, from a partitioning, a hierarchical or a pyramidal clustering, from the most influential individuals in a factorial axis, from a decision tree, etc.

In order to see how much a given symbolic object is characteristic of a class A, an hypergeometric distribution can be used. Let N be the size of W', n the size of A, $p = Ext(s/W')/N$ the proportion in W' of individuals belonging in the extent of s, X a random variable whose value at each resample is the proportion in A of individuals belonging in the extent of s. Then, the hypergeometric law gives the probability of $X = x$ by: $Pr(X=x) = C^{x}_{Np}\, C^{n-x}_{N-Np} / C^{n}_{N}$ where $C^{n}_{N} = N! / n!(N-n)!$ is the number of possible samples of size n in N, $C^{x}_{Np} =$ Np!/(Np-x)!x! is the number of groups of x individuals belonging in the extent of s in W' and $C^{n-x}_{N-Np} = (N-Np)!/ (n-x)!(N-Np-n+x)!$ is the number of groups of $(n-x)$ individuals which are not belonging in the extent of s in W'. If the operator T produces k symbolic objects of extent in A with size $x_1,\ldots, x_k$ then the more $Y = \mathring{a}_{i\,=\,1,k}\, Pr(X = x_i)/k$ is small, the more these symbolic objects are characteristic of the class A. This happen for instance, when p is small and x/n large or p large and x/n small. Notice that in the case where s is a complete symbolic object the size of the extent is n and $p = n/N$, so $Pr\ (X = n) = \boldsymbol{C^{n}_{n}}\ x\ \boldsymbol{C^{0}_{N-n}} / \boldsymbol{C^{n}_{N}} = 1x1/\ \boldsymbol{C^{n}_{N}} = ((N-n)!\ n!)/\ N!$ which is the probability of a complete symbolic object of size n in a population of size N. When bootstrapping W', if the mean of the random variable Y is out of the chosen confidence interval, then the more its standard deviation is low the more the characterization is reliable. If we are interested by the variation of the characteristic of a specific symbolic objet, notice that at each resample we have to recognize each symbolic object. This can be done by the use of a dissimilarity measure between symbolic objects from one resample to the next (Esposito et al. 2000). The closest are considered to be the same.

A «complete symbolic object» s is such that $F(G(s)) = s$. Such objects can be selected from the Galois lattice but also, from a partitioning, a hierarchical or a pyramidal clustering, from the most influential individuals in a factorial axis, from a decision tree, etc.

2.3 Modeling Individuals, Classes of Individuals and Concepts

In figure 1 the "set of individuals" and the "set of concepts" is considered to be in the "real world", the "modeled world" is the "set of descriptions" which models individuals (or classes of individuals) and the "set of symbolic objects" which models concepts. We start with a "concept" C whose extent denoted $Ext(C/W')$ is known in a sample W' of individuals. For instance, if the concept is "insurance companies", for instance, 30 insurance companies among a sample W' of 1000 companies. Each individual w of the extent of C in W' is described by using the mapping Y such that $Y(w)$ describe the individual w. We generalize the set of descriptions of the individuals of $Ext(C/W')$ with the operator T in order to produce the description d_C (which can be a set of Cartesian products of intervals and (or) distributions).

i. The comparison relation R is chosen in relation with the T choice. For instance, if $T = \cup$ then $R =$ "$\subseteq$", it $T = \cap$, then $R =$ "$\supseteq$".
ii. The membership function is then defined by $a_C(w) = [y(w)\ R_C\ d_C]$ and then the symbolic object modelling the concept C is the triple $s = (a_C, R, d_C)$.

When we don't have concepts as input, we get them in the following way:

i. A clustering of W' by using the description of the individuals produces a set of classes.
ii. To each interesting class denoted A, we associate a concept C and a symbolic object $s_A = (a_A, R_A, d_A)$ with $a_A = [Y(w)\ R_A\ d_A]$ where d_A is obtained by using an operator T on the set of the descriptions of the individuals of A, as in the preceding case.
iii. The concept C is considered to be modeled by s_A.

2.4 Some Advantages in the Use of Symbolic Objects

We can observe at least five kinds of advantages in the use of symbolic objects.

1. They give a summary of the original symbolic data table in an explanatory way, (i.e. close to the initial language of the user) by expressing descriptions based on properties concerning the initial variables or meaningful variables (such as indicators obtained by regression or factorial axes).
2. They can be easily transformed in terms of a query of a Data base and so they can be used in order to propagate concepts between data bases (for instance, from one country to another country).
3. By being independent of the initial data table they are able to identify any matching individual described in any data table.
4. In the use of their descriptive part, they are able to give a new symbolic data table of higher level on which a symbolic data analysis of second level can be applied.
5. In order to characterize a concept, they are able to join easily several properties based on different variables coming from different relations in a Data Base and different samples of a population.
6. In order to apply exploratory data analysis to several data bases, instead of merging them in a huge data base, an alternative is to summarize each Data Base by symbolic objects and then to apply Symbolic Data Analysis to the whole set of obtained symbolic objects.

3 Some Symbolic Data Analysis Methods

Symbolic Data Analysis methods are mainly characterized by the following principle:

i. they start as input with a symbolic data table and they give as output a set of symbolic objects. These symbolic objects give explanation of the results in a language close to the one of the user and moreover have all the advantages mentioned in 5).
ii. They use efficient generalization processes during the algorithms in order to select the best variables and individuals.
iii. They give graphical descriptions taking account of the internal variation of the symbolic objects.

The following methods are developed in Bock & Diday (2000) and in the SODAS software:

- Principal Component and Discriminate Factorial Analysis of a symbolic data table. The output of these methods preserves the internal variation of the input data in the sense that the individuals are not represented in the factorial plane by a point as usual but by a rectangle which allows the definition of a symbolic object with explanatory factorial axes as variables;
- extension of elementary descriptive statistics to symbolic data (central object, histograms, dispersion, co-dispersion, etc. from a symbolic data table);
- extracting symbolic objects from the answers to queries of a relational data base;
- partitioning, hierarchical or pyramidal clustering of a set of individuals described by a symbolic data table such that each class be associated with a complete symbolic object;
- dissimilarities between Boolean or probabilistic symbolic objects;
- extension of decision trees on probabilistic symbolic objects;
- generalization by a disjunction of symbolic objects of a class of individuals described in a standard way;
- inter-active and ergonomic graphical representation of symbolic objects.

4 Symbolic Data Analysis in the SODAS Software

The general aim of SODAS can be stated in the following way: building symbolic data in order to summarize huge data sets and then, analyze them by Symbolic Data Analysis. For instance, if a set of households is characterized by its region, the number of bedrooms and of dining-living, its socio-economic group, we obtain a data table of the kind of table 1:

Table 1. Standard Data Table where the units are Households

Household number	Region	Bedroom	Dining-Living	Socio-Econ group
11404	Northern-Metropolitan	2	1	1
11405	Northern-Metropolitan	2	1	3
11406	Northern-Metropolitan	1	3	3
12111	Northern-Metropolitan			
12112	East anglia	1	3	3
12112	East anglia	2	2	1
12112	Greater London N-E	1	2	3

In census data there is a huge set of households. In order to compare the regions, we can summarize them by describing each region by the households of their inhabitants. In order to do so, we delete the first column of this table and we obtain the table 2:

Table 2. The first column of table 4 concerning the household number has been deleted

Region	Bedroom	Dining-Liv	Socio-Ec gr
Northern-Metropolitan	2	1	1
Northern-Metropolitan	2	1	3
Northern-Metropolitan	1	3	3
Northern-Metropolitan			
East-anglia	1	3	3
East-anglia	2	2	1
East-anglia	1	2	3
Greater London North-East			

We can now describe each town by the histogram of the categories of each variable. This is done in table 3 which is a symbolic data table as each cell contains a histogram and not a quantitative or categorical number as in the standard data tables. It is easy to see that standard data analysis methods will not apply in the same way with these kind of symbolic data. For instance that a decision tree will not be the same if the variables are categories and each cell of the associated data table contains a frequency and if the variable are symbolic and each cell contains a histogram. In the first case each branch of the decision tree represents an interval of frequency (for instance, “the frequency of the category [20, 30] years old is less then 0.3”), whereas in the second case it represents an interval of values (for instance, “the age is less then 50 years old”). For more details see in Bock & Diday (2000) the chapter 11.

Table 3. A symbolic data table where the units are now regions

Region	Bedroom	Dining-Living	Socio-Ec gr
Northern Metropolitan	(2\3) 2, (1\3) 3	(2\3) 1, (1\3) 3	(1\3) 1, (2\3) 3
East-anglia	(2\3) 1, (1\3) 2	(2\3) 2, (1\3) 3	(1\3) 1, (2\3) 3
Greater London			

The main steps for a symbolic data analysis in SODAS can then be defined as following:
If there is more than one data table, put the data in a relational data base (ORACLE, ACCESS, and so on). Then, define a context by giving: the units (individuals, households, and so on), the classes (regions, socio-economics groups,...), the descriptive variables of the units. Then, build a symbolic data table where the units are the preceding classes, the descriptions of each class is obtained by a histogram as in table 6 or by a generalization process applied to its members. This is done by a computer program of SODAS called "DB2SO" (from Data Bases Two Symbolic Objects). Finally, apply to this symbolic data table, symbolic data analysis methods (histogram of each symbolic variable, dissimilarities between symbolic descriptions, clustering, factorial analysis, discrimination of a symbolic data table, graphical visualization of symbolic descriptions, and so on).

5 Conclusion

The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data tables in order to extract new knowledge, is increasing due to the expansion of information technology, now able to store an increasing amount of huge data sets. This need, has led to a new methodology called "Symbolic Data Analysis" whose aim is to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination, decision trees,...) to new kind of data table called "symbolic data table" and to give more explanatory results expressed by real world concepts mathematically represented by easy readable "symbolic objects". The aim of the EUROSTAT European Community project called SODAS for a «Symbolic Official Data Analysis System» in which 17 institutions of 9 European countries are concerned was to produce a first software of Symbolic Data Analysis (fig. 2). Three Official Statistical Institutions was involved in this project: EUSTAT (Span), INE (Portugal) and ONS (England). An example of future application proposed on their Census data consists in finding clusters of unemployed people and their associated mined symbolic objects in a country, calculating its extent in the census of another country and describing this extent by new symbolic objects in order to compare the behaviour of the two countries. In that way, several new theoretical development are needed as the selection and the stochastic convergence of symbolic objects. Also, as the consensus between set of symbolic objects and their associated concepts extracted from different data bases. New software development are also needed as a tool in order to be able to transform a symbolic object extracted from a data base in a query of this data base or of another data base. This new tool may be called SO2DB as it is complementary to the actual DB2SO (Malerba et al,

2002). Moreover, the next steps will be to improve the actual SDA methods (robustness, validity of the results, extending standard tests to symbolic data, etc.) and extend the symbolic data analysis methodology to regression, multidimensional scaling, neural network etc. The SODAS software is free and available at http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

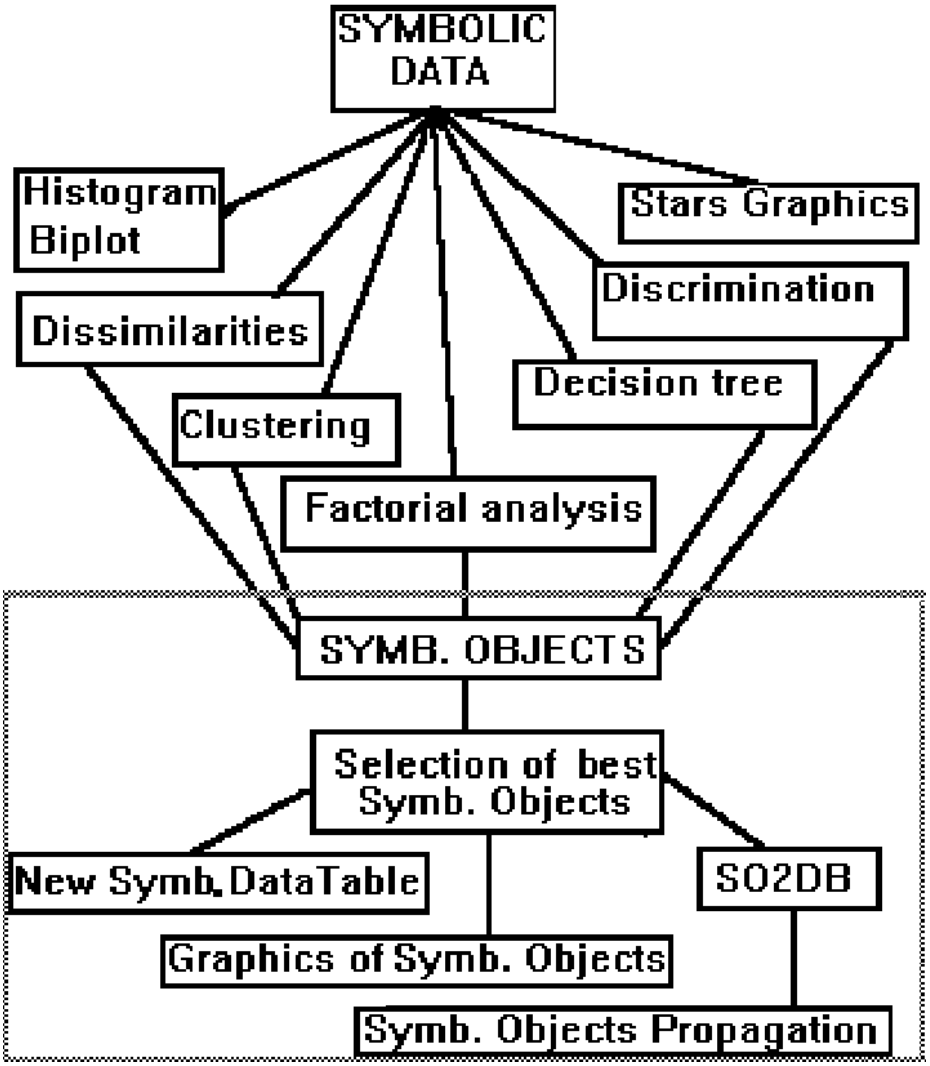


Fig. 2. Software development of the SODAS project

References

Auriol E. (1995) “Intégration d’approches symboliques pour le raisonnement à partir d’exemples” Thèse de doctorat, Université Paris 9 Dauphine.

Bock H.H., Diday E. (2000) “Analysis of Symbolic Data”. Study in Classification, Data Analysis and Knowledge Organization. Springer Verlag.

Bravo C., Garcia-Santesmas J. (1998) "Symbolic objects description of strata by segmentation trees". Proc. NTTS. Ph. Nanopoulos, Garonna, Lauro edit, Eurostat, Sorrento (Italy).

Brito P., Diday E. (1991) “Pyramidal representation of symbolic objects” NATO ASI Series, Vol. F 61. Proc. Knowledge Data and computer-assisted Decisions. Schader and Gaul edit. Springer-Verlag.

Brito P. (1994) “Order structure of symbolic assertion objects”. IEEE TR. on Knowledge and Data Engineering Vol.6, n° 5, October.

Cazes P., Chouakria A., Diday E., Schechtman Y.(1997) “Extension de l’Analyse en Composantes Principales à des données intervalles”. Revue de Statistiques Appliquée, vol. XXXVIII, n°3, 1990,pp 35-51.

Chavent M. (1997) “Analyse des Données symboliques. Une méthode divisive de classification”. Thèse de doctorat, Université Paris 9 Dauphine.

Ciampi A., Diday E., Lebbe J., Périnel E., Vigne (1995) R. “Recursive partition with probabilistically imprecise data”. OSDA’95. Editors: Diday, Lechevallier, Opitz Springer Verlag (1996).

Conruyt N. (1994) “Amélioration de la robustesse des systèmes d’aide à la description, à la classification et à la détermination des objets biologiques”. Thèse de doctorat, Université Paris 9 Dauphine.

De Carvalho, F.A.T. (1994) “Proximity coefficients between Boolean symbolic objects”. In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P. and Burtschy, B. (Eds.): *New Approaches in Classification and Data Analysis*, Springer-Verlag, Heidelberg, Germany, 387-394.

De Carvalho F.A.T. (1998a) “New metrics for constrained boolean symbolic objects” Proc. KESDA’98, Eurostat. Luxembourg.

De Carvalho F.A.T. (1998b) “Statistical proximity functions of boolean symbolic objects based on histograms” IFCS, Roma, Springer-Verlag.

Diday E. (1987a) “The symbolic approach in clustering and related methods of Data Analysis” in “Classification and Related Methods of Data Analysis”, Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.

Diday E. (1987b) “Introduction à l’approche symbolique en Analyse des Données ”. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.

Diday E. (1989) “Introduction à l’approche symbolique en analyse des données”. RAIRO (Revue, d’Automatique, d’informatique et de Recherche Opérationnelle), vol. 23, n°2.

Diday E. (1991) “Des objets de l’Analyse des Données à ceux de l’Analyse des Connaissances” in “Induction symbolique et numérique”. Y. Kodratoff and E. Diday edit. CEPADUES-EDITIONS, Toulouse, France.

Diday E. (1995) “Probabilist, possibilist and belief objects for knowledge analysis”. Annals of Operations Research . 55, 227-276.

Diday E., Emilion R. (1995) “Lattices and Capacities in Analysis of Probabilist Objects”. Proceed. of OSDA’95 (Ordinal and Symbolic Data Analysis). Springer Verlag Editor (1996).

Diday E., Emilion R. (1997) “Treillis de Galois maximaux et Capacités de Choquet” Compte rendu à l’Académie des Sciences. Analyse Mathématique, t. 324, série 1.

Diday E., Emilion R., Hillali Y. (1996) “Symbolic data analysis of probabilist objects by capacities and credibilities. XXXVIII Societa Italiana Di Statistica. Rimini, Italy.

Diday E.(1998) “L’Analyse des Données Symboliques: un cadre théorique et des outils”. Cahiers du CEREMADE n° 9821.

Diday E. (2000) “Partitioning concepts described by distributions with copulas modelling” OSDA ‘2000. Bruxelles.

Esposito, F., Malerba, D., & Tamma, V.(2000) “Dissimilarity Measures for Symbolic Objects” (Section 8.3), In Bock, H. H. & Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 165-185.

Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995) “Knowledge extraction using stochastic matrices: Application to elaborate a fishing strategy” Proc. Ordinal and Symbolic Data Analysis. Paris ; Diday, Lechevallier, Opitz edit.Springer Studies in Classification.

Gettler-Summa M. (1992) “Factorial axis interpretation by symbolic objects”. Journées - Symbolique - Numérique. Université Paris IX- Dauphine. Lise-Ceremade.

Gettler-Summa, M., Pardoux, C. (2000) Noirhomme-Fraiture, Rouard M. (2000) “Symbolic Approaches for Three-way Data” (Chapter 12), In Bock, H. H. & Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 342-354.

Gigout E. (1998) “Graphical interpretation of symbolic objects resulting from data mining”. Proc. KESDA’98, Eurostat. Luxembourg.

Gowda K.C., Diday E. (1992) “Symbolic clustering using a new similarity measure”. IEEE Trans. Syst. Man and Cybernet. 22 (2), 368-378.

Hillali, Y. (1998) “Analyse et modélisation des données probabilistes: capacités et lois multidimensionnelles”, Thèse de doctorat, University Paris 9 Dauphine.

Kiers, H., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.) (2000) Data Analysis, Classification, and Related Methods. Series: Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Berlin.

Lauro C., Palumbo F. (1998) “New approaches to Principal Component Analysis of Interval Data”. Nanopoulos, Ph., Garonna, Lauro, C. (eds.) Proc. NTTS’98 Sorrento, Italy. Eurostat, Luxembourg.

Malerba, D., Esposito, F., Monopoli M. (2002). Estrazione e matching di oggetti simbolici da database relazionali. Atti del Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati SEBD’2002, 265-272.

Mfoumoune E.-M. (1998) “Analyse de données symbolique incrémentale et apprentissage”, Thèse de doctorat, University Paris 9 Dauphine.

Michalski R., Diday E., Step R.E. (1982) “A recent advances in Data Analysis: clustering objects into classes characterized by conjonctive concepts”. Progress in Pattern Recognition, vol 1. L; Kanal and A. Rosenfeld Eds.

Noirhomme-Fraiture, Rouard M. (1998) “Representation of Sub-Populations and Correlation with Zoom Star”. Proc. NTTS’98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).

Noirhomme-Fraiture, Rouard M. (2000) “Visualizing and Editing Symbolic Objects” (Chapter 7), In Bock, H. H. & Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 125-138.

Périnel E. (1996) “Segmentation et Analyse de Données Symboliques: Application à des données Probabilistes Imprécises”. Thèse de doctorat, Université Paris 9 Dauphine.

Pollaillon G., Diday E. (1997) “Galois lattices of symbolic objects” Rapport du Ceremade University Paris9- Dauphine (February).

Pollaillon G. (1998) “Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme”. Thèse de doctorat, Université Paris 9 Dauphine.

Pollaillon G. (2000) “Pyramidal Classification for Interval Data Using Galois Lattice Reduction” (Section 11.4), In Bock, H. H. & Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 324-341.

Rasson J.P., Lissoir S. (1998) “Symbolic Kernel discriminante analysis” Proc. NTTS’98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).

Rodriguez O. (2000) “Classification et modèles linéaires en Analyse des Données Symboliques”. Thèse de doctorat, University Paris 9 Dauphine.

Stéphan (1998) “Construction d’objets symboliques par synthèse des résultats de requêtes SQL”. Thèse de doctorat, Université Paris 9 Dauphine.

Stéphan, V., Hébrail, G., Lechavallier, Y. (2000) “Generation of Symbolic Objects from Relational Database” (Chapter 5), In Bock, H. H. & Diday, E. (Eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag: Berlin, pp. 78-105.

Tang Ahanda, B. (1998) “Extensions des méthodes d’analyse factorielle sur des données symboliques. Thèse de doctorat, Université Paris 9 Dauphine.

Vignes (1991) “Caractérisation automatique de groupes biologiques”. Thèse de doctorat, Université Paris 9 Dauphine.

Verde R., F.A.T. De Carvalho (1998) “Dependance rules influence on factorial representation of boolean symbolic objects”. Proc. KESDA’98, Eurostat. Luxembourg.

Ziani D. (1996) “Sélection de variables sur un ensemble d’objets symboliques” Thèse de doctorat, Université Paris 9 Dauphine.

Important Issues on Statistical Confidentiality Methods

Ph. Nanopoulos and John King

Eurostat, European Commission, L-2920 Luxembourg
photis.nanopoulos@cec.eu.int; john.king@cec.eu.int

Abstract. This paper sets out, in the context of official statistics, some of the key issues of confidentiality and the methods developed to maintain confidentiality. The relevance of the issues and methods to data mining of official data are discussed. Recent developments that will increase the availability of microdata for scientific research are outlined.

1 Introduction

The title of this workshop is "Mining Official Data". So what is the connection between data mining and confidentiality of official statistics? One link is that there may be lessons for the data-mining practitioner to be learned from the practices developed by statistical organisations in order to protect the confidential nature of their data. But to mine data, there must be access to data. So perhaps the presumption of the Workshop is that data used for the creation of official statistics could be made available for mining. This paper looks at the proposition from the viewpoint of an official statistical organisation. It therefore looks at the constraints and issues that statistical offices live with in their regular work on official economic and social data. Researchers, including data miners, who wish to have access to the datasets underlying official statistics will need to understand and respect these issues and principles.

These issues apply in particular to data obtained through surveys and censuses, either voluntary or compulsory, from individuals and households, and from companies or enterprises. Some of these issues may not arise where data have been obtained by different methods, for example from administrative records, but even then other constraints may exist in the form of specific laws protecting those records or from general Data Protection laws. An example of the latter constraint is the European Union (EU) Directive 95/46 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, and also the consequential laws implementing the directive in the EU Member States. But that is presumably a topic well known to data miners, and well understood by them, and so this paper will concentrate on confidentiality issues related to official economic and social statistics and the data on which they are based.

Data mining may seem to be the antithesis of protecting the confidentiality of official statistics. The former seems to have a carefree, heroic and exploratory image that is the opposite of the conservative approach taken in official statistics. So in any public discussion it would be important to emphasise that data mining is looking for *patterns* and *statistical* relationships in data—not for individuals and their details and their relationships. So data mining is *pattern* recognition, not *people* recognition, or company recognition.

The principle of informed consent is the basis of much of official statistics. A guarantee of confidentiality for the information provided is often the basis of obtaining the data. It also imposes a constraint on what can be done with the data and by whom. In this context, confidentiality issues and statistical disclosure methods have been developed to maximise the use of the data while keeping the original agreement with the data source.

This paper sets out some of the issues, principles and methods that have been developed, particularly in relation to traditional outputs—tabulations—of statistical offices. Then, some of the issues relating to providing access to confidential data are indicated. Finally, some recent developments, at the European level, that will provide better access to anonymised micro-datasets for scientific research are described.

2 Tabular Data Disclosure Control

Tables (including cross-tabulations) are the simplest and most common form of official statistical output. It was recognised early on that tables could be disclosive, particularly in the case of economic data. The important criterion for tables is that dissemination is only possible if data subjects (those providing the data) cannot be identified directly or indirectly. The problem is then one of checking whether a table has sensitive cells; and then of dealing with the sensitive cells.

For example, a table might show production levels of a particular commodity. If only one company makes this commodity then the table would disclose its output. A more complex table, perhaps of production by region and by size category of the production unit, might make it easy to recognise a company (from several making that commodity) and show its output of the particular commodity.

Early methods for protection of tabular information included suppressing cell values in tables and forming broader categories of the classifying variables. These methods are still in use, sometimes implemented by hand. Other methods include rounding values in cells and adding noise (e.g. adding +1, 0, or -1 to the cells in a random pattern. Some of these methods are not very satisfactory in that marginal or grand totals may not be the exact sum of the constituent cells as shown.

Over time, and following interesting research in the subject, a more sophisticated understanding of the problem has developed. And the development of more sophisticated methods for statistical disclosure control for tables has followed this understanding. For example, it was recognised that, even if a table appeared non-

disclosive to a casual reader, it might be disclosive if the reader had some additional relevant information. If a table cell had an aggregate from two observations, then one of the members of the cell could deduce, by subtraction, the contribution of the other member. Hence, the requirement, quite common in official statistics, that cells should contain at least three members before any information can be released about the cell. Various rules have developed over time, often based on the number of units in the cell, including the "dominance" rules, now commonplace.

Cell suppression itself has developed into a sophisticated art. We now consider the suppression of "secondary" cells to protect the "primary" cells, and also the degree (or interval) of protection afforded by different suppression patterns.

Recent developments include methods for the simultaneous protective treatment of tables with several dimensions that have one or more overlapping (i.e. shared) dimensions. These methods have been implemented in software by statistical offices and continue to be the subject of research. An example is the *GHMITER* engine developed at the Statistisches Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen. A user-friendly graphical interface for this, *CIF*, has been developed by Eurostat; and the engine is also being incorporated into the software suite *t-Argus.*

Another issue that is becoming recognised as very important is the constraint imposed by published tables on the to-be-published tables. There is active research in this area. Karr and Sanil indicate that the whole tabulation plan for a particular dataset should be considered at the outset. Otherwise publication of a particularly interesting table (or at least the interesting parts of it) may be prevented because of the information already published. In the Eurostat context, this could mean that an aggregate for the 15 Member States (MSs) might not be publishable because of the aggregates already published (or suppressed) at MS level.

3 Dissemination of Anonymised Records

The dissemination of anonymised records is an issue that has become increasingly important and will undoubtedly continue to be so. It is probably the issue discussed in this paper that will be of greatest interest because the records could be the raw material for data mining.

3.1 Research Interest

Anonymised records, or anonymised micro-datasets, are becoming important because of increasing interest from researchers in access to them. This interest has two related drivers. An aspect of modern life is the increasing interest in and demand for evidence-based policy, policy analysis, and monitoring policies and their impact. This kind of activity requires timely, detailed information and frequently requires more detailed analyses than are presently published by statistical organisations. Sometimes these analyses are seen as outside the remit of national statistical organisations (NSIs)

or even as activities that could compromise the perceived independence of NSIs. Indeed, these analyses are performed often by academic institutions or independent research institutions.

Detailed data are needed for these types of analyses. The obvious and most relevant source is often identified as the data collected and held by NSIs. Hence there is an increasing pressure on NSIs and other statistical organisations to provide detailed data on a wide range of topics. In particular, for the European Union (EU), pan-EU analyses and research are becoming more and more important. The same could also be said for the Euro-zone. So the need is for access to pan-EU datasets for this research. Eurostat holds many such datasets, and so it is seen, by analogy with the national situation, as the natural, simple and direct potential source for these datasets.

The second driver here is the changing nature of research itself. Much modern research cannot be satisfied with aggregate data—micro-data are needed for fine analysis and model building. Hand-in-hand with this there has been an evolution (perhaps revolution would be a more appropriate description) of research computing capacity—both hardware and software tools. This has considerably increased the demand for access to micro-data records for computing correlation matrices, estimating models and other analyses, depending on the context of the research topic.

3.2 Providing Data while Maintaining Confidentiality

At the same time, statistical organisations, both NSIs and supra-national and international institutions, are increasingly seeing making more use of the data held by them as an important contribution to society and as part of an obligation to make better use of their resources (data). But there are constraints on what statistical organisations, particularly NSIs, can do and on how they can do it. The role of researchers and research organisations is thus an important one, and it is an increasing one too. NSIs can assist in this development by providing some form of access to some form of data to these researchers and research organisations.

But there are problems in providing micro-datasets to researchers—the main one being the confidential nature of the data themselves. The question is, how can the confidentiality of the data be maintained while at the same time providing researchers with the information they would like.

There are two key issues here. The first is the basis on which the data has been obtained; and the second, related to the first, is the perception of the data supplier. Most information used for official statistics has been obtained against a pledge of confidentiality and a guarantee that it will be used for statistical purposes only. The ideas of informed consent and confidentiality underlie official statistical data collection. The principle is that the data subject has a right to know what the information will be used for and who will see their information. The argument here is that if there is a new kind of use of the information, then the data subject should be made aware of it.

The perception of the data supplier is very important. Many statistical offices feel that this is a major factor affecting response rates to surveys and data collections. Damage to perceptions, or a loss of confidence through even inadvertent disclosure of confidential data, would result in worse data in the future.

Nevertheless, several statistical offices have created anonymised micro-datasets for access by researchers. These are intended to provide researchers with a dataset in which the information content has been reduced sufficiently for the risk of identification of a record or of disclosure to be acceptably small. Some datasets are available, with little bureaucratic procedures and at minimum cost, to academic researchers. Other statistical offices have a more conservative approach. Differences reflect conditions, attitudes, legal issues and past practices in different countries. There are differences, too, in the meaning of "anonymisation" and in the degree of risk that would be acceptable in releasing anonymised micro-datasets.

3.3 Creating Anonymised Records

For the creation of anonymised micro-datasets, various techniques have been proposed—broadly the same as those used to protect tabular information but also micro-aggregation techniques, suppressing variables and ensuring there is a delay between collection and reference period of the data and its release as microdata.

Some methods used, often in combination, at present include:

- reducing the geographic coverage;
- rounding;
- grouping or combining categories;
- adding noise or perturbations;
- micro-aggregation;
- data swapping;
- top- (or bottom-) coding;
- imputing values from a model;
- suppressing fields or cells;
- suppressing variables;
- time delay.

An innovative approach suggested recently by Abowd and Woodcock is the creation of synthetic datasets—retaining the internal structure of the variables but containing no real records. Although these synthetic datasets might satisfy some research purposes, it is not yet clear whether they would also be of use in situations where the relationship being searched for may exist in the original dataset but not have been specified in the creation rules for the synthetic dataset.

3.4 Micro-aggregation

Micro-aggregation may need a little explanation as it is relatively unknown and unused. It is one method (or, rather, a set of methods) for anonymising potentially disclosive data and thus for creating anonymised micro-datasets. In essence, a

variable that is sensitive or potentially disclosive is perturbed in a particular way with the intention of retaining as much information and pattern in the data as possible but at the same time reducing the risk of the information about that variable being disclosive. Records are clustered into small groups on one or more variables. Then the aggregate or average value of the variable over the group is assigned to each member of the group. In an early presentation of the approach, Defays and Nanopoulos proposed fixed same sized groups, but more recent research has considered variable sized groups.

For example, a variable may be of the amount of local taxes paid by a household. Depending on the way this amount is determined or calculated, this amount may enable a researcher to identify the local authority in which the household resides. This information, particularly in conjunction with other information about the household in the record, may raise the risk of identification of the household to an unacceptable level. So a simple anonymisation process would be to suppress this variable as it provides information enabling the local authority of the household to be identified. But suppose that the amount of local taxes paid could be averaged (for similar households) over several similar local authorities. This average (the micro-aggregate) could replace the actual amount paid in the records for the households. This would mean that the particular local authority of a household could not be identified unambiguously, thus reducing the risk of identification. But the patterns in the data would, to a large extent, be unchanged.

Different approaches have considered different notions of "similarity" for forming the groups, and these have lead to different ways of creating the groups of records to be aggregated. What they have in common is first the ranking of a variable according to a defined criterion and then the grouping of successive records. Some approaches have used values of a variable; others, the first principal component of a set of variables; yet others, the sum of z-scores.

4 Confidential Data at Eurostat—the Legal Context

The principle of statistical confidentiality is effectively the contract connecting the statistician with all those providing their individual data, either voluntarily, as is frequently the case, or by legal obligation, with a view to producing the statistical data essential for the society as a whole. From the formal legal point of view most of the European countries have established legal provisions for statistical confidentiality a long time ago. At the European level, the principle has been enshrined in Article 285 of the Treaty establishing the European Community as a fundamental principle for Community statistics. Article 285 provides that the production of Community statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality. The confidentiality principle is therefore part of the European basic constitutional charter and has thus acquired the highest status in legal terms.

The principle has been further specified and data received, held, used and disseminated by Eurostat are controlled by a set of laws that have developed since the Treaty founding the European Communities. In 1990, Council Regulation 1588/90 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Community set out basic rules and safeguards for the handling of confidential data. Subsequently, in 1997, the "Statistical Law"—EU regulation 322/1997 on Community Statistics—expanded on these basic rules. In particular, a legal definition of statistical disclosure was introduced. Article 13 states:

"1. Data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.

To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit."

This definition has replaced the former definition laid down in Regulation 1588/90 where confidential data were defined as "data declared confidential by the Member States in line with national legislation or practices governing statistical confidentiality." The notion of confidential data has consequently become an objective notion with a clear Community dimension.

This definition uses explicitly five different concepts: individual information; a third party; identification (direct or indirect) of a statistical unit; disclosure of individual information to a third party; and means that can be reasonably used by a third party to identify the said unit.

The concept of "individual information" is not explicitly defined in the European statistical law and so interpretation should be as wide as possible thus making confidential any information concerning a statistical unit. Nevertheless, article 13 goes on to state:

"2. By derogation from paragraph 1, data taken from sources which are available to the public and remain available to the public at the national authorities according to national legislation, shall not be considered confidential."

So, in conclusion, we can say that "individual confidential information is any individual information which is not normally publicly available".

The Statistical Law also states that confidential data must be used exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes (article 15). The law also makes provision for access to confidential data for scientific purposes (article 17).

5 Recent developments on access to datasets for scientific research

At the European level, there have been developments recently which will help to meet the demand for access to microdata and to open up datasets for scientific research purposes. A new Commission Regulation, 831/2002, concerning access to confidential

data for scientific purposes, was adopted on 17 May and came into force on 7 June 2002. This was the culmination of a long process of discussion, negotiation and drafting with 15 Member States (MSs). The Regulation sets out procedures and conditions under which access to confidential data for scientific purposes may be granted. The regulation refers to four important sources:

- European Community Household Panel (ECHP);
- Labour Force Survey (LFS);
- Community Innovation Survey (CIS);
- Continuing Vocational Training Survey (CVTS).

In summary, researchers must belong to research institutions and organisations within the MSs. A detailed proposal must be prepared stating the purpose of the research, methods to be used and details of the data to be used. Safeguards for the secure holding of the datasets will be necessary and controls on access by individuals will be required. Agreement to conditions and safeguards will be through a contract with the researchers' institution. There is no right of access to confidential data under the Regulation. MSs can withhold the data of their country from any particular research request. Access to confidential datasets can be on the premises of Eurostat with checks on the output and results to maintain confidentiality; or access can be through anonymised micro-datasets. Work is now proceeding in Eurostat, in close collaboration with the NSIs of MSs and with the research community, in putting this into practice.

Incidentally, the new Regulation 831/2002 now provides a legal definition of anonymised micro-datasets. " "anonymised microdata" shall mean individual statistical records which have been modified in order to minimise in accordance with current best practice the risk of identification of the statistical units to which they relate."

For some of the data sources mentioned in the Regulation, the first step will be the creation of anonymised micro-datasets. Although there is a wealth of knowledge of these datasets in all the NSIs, and some experience of creating anonymised micro-datasets, the work will proceed through discussion with, and agreement from, the NSIs of the MSs, using established methods as described above. The creation of anonymised micro-datasets for household and individual data seems to be relatively simple. At present work is proceeding on investigating whether satisfactorily anonymised micro-datasets can be created for records for business and enterprise data as well.

In the longer term, it is hoped that access could be provided to other datasets on the same, or a very similar, basis.

The data miner interested in accessing these anonymised datasets will need to study the Regulation carefully. The purpose must be statistical and the activity and output must be scientific research. This would seem to exclude any activities for commercial purposes. The datasets may not be brought together with, or compared with, any other datasets. The credentials of the researcher and the university or research organisation for the research proposed will need to be established.

Some of this may seem off-putting. But for genuine scientific research, the Regulation offers new opportunities to access large, respected and comprehensive

datasets covering the MSs of the European Union. The potential is great, as is the opportunity and the challenge to researchers.

References

Abowd, J.M. and Woodcock, S.D. Disclosure limitation in longitudinal linked data. In Confidentiality, Disclosure, and Data Access, (pp.215-78), North-Holland, 2001.

Defays, D. and Nanopoulos, P. "Panels of Enterprises and Confidentiality: The Small Aggregates Method" in Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada, pp.195-204.

European Union. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995, pp.0031-0050.

Karr, Alan F. and Sanil, Anish P. Web systems that disseminate information but protect confidential data. Proceedings of 53rd Session of the International Statistical Institute, 2001, Bulletin of the ISI, 53rd Session, volume LIX, book 1 pp. 265-8.

Developing a Metadata Infrastructure for Official data: the ISTAT experience

Giovanna D'Angiolini

e-mail dangioli@istat.it
Istituto Nazionale di Statistica (ISTAT), via C. Balbo 16
00186 Roma, ITALY,

Abstract. Mining official data implies retrieving data from different surveys or administrative sources and properly interpreting them as measures of observed phenomena. In order to support such an activity, ISTAT is developing a metadata infrastructure which is based on two centralized metadata management system, and implies the development of tools for exploiting metadata in the data manipulation activities. Such a strategy is supported by the definition of a conceptual metadata framework together with proper documentation models for each class of metadata.

1 The ISTAT Strategy for Metadata Management: choices and involved problems

Mining official data implies retrieving data from different surveys or administrative sources and properly interpreting them as measures of observed phenomena. Such an activity requires the availability of several classes of metadata concerning the characteristics and the information content of each exploitable source of information. To ensure the dissemination of such metadata to the data users is a primary task for National Statistical Institutes (NSI), nevertheless to introduce metadata management practices in the official data production is often a challenge. Most NSIs consider the development of a metadata infrastructure a long-term goal, which requires a carefully devised strategy. Moreover the increasing need for integrating data from several sources obliges the NSIs to pursue a policy of centralised metadata management. By means of homogeneously documenting data from different sources in a unique environment, a centralised metadata system provides the rough material for data integration.

Therefore the core of the ISTAT strategy is the development of two centralised systems for metadata management, SIDI and SDOSIS, which manage metadata concerning the production processes of surveys and the information content of surveys and SIS, respectively. They will disseminate such metadata to both data users and survey designers. Moreover they are conceived as metadata servers for those data management systems and software tools which are exploited in the data production and dissemination activities. Another important experience which we are carrying on

is the development of ESPLORIS, a system which exploits metadata for assisting the user in extracting data from data collections produced by several sources. As analysts of real world phenomena, the end users of official data have a basic requirement: to search the large data collections which are produced by NSIs for retrieving those data which are better suited to the analysis of a given class of phenomena, and properly transform them. This is an exploratory activity, which requires the availability of proper metadata. Many OLAP/DW tools exist which are well suited to perform data exploration inside an organisation. Inside a single organisation the role of the different classes of users constrains their information requirements. On the contrary the official data users have unpredictable information requirements, therefore a richer amount of metadata is required for steering them in retrieving and transforming data. This is the reason why NSIs should invest in developing specialised tools for supporting the data users' exploratory activities. Moreover tools for metadata-based data retrieving are an important component of a general metadata infrastructure because a good metadata quality is ensured only if metadata are actually used in data manipulation activities.

Last but not least, a NSI's metadata strategy requires a sound conceptual foundation. This implies answering such questions: which are the relevant classes of metadata for properly retrieving, transforming, analysing official data? How should we model such metadata?

In sections 2 and 3 we outline our general conceptual framework for metadata specification and present our model OSI, which we use for modelling the information content of surveys and statistical information systems. In section 4 we discuss the main characteristics of the systems which compose our metadata infrastructure, namely SIDI, SDOSIS and ESPLORIS.

2. Classes of Metadata for Documenting Official Data

There is a general agreement on the need for a general conceptual framework which specifies relevant classes of metadata, starting from an in depth understanding of the role of metadata in the data production and analysis activities (see for example [13], [14], [18]) In our opinion, a general conceptual framework for metadata specification should specify classes of metadata and relationships among them according to several dimensions, and provide the metadata managers with documentation models for each singled out class of metadata.
Moreover, the different contexts in which metadata are used and the different ways of communicating and using metadata should be analysed.

We propose to single out relevant metadata classes on the basis of such dimensions: the content of metadata, their level of abstraction, their scope, that is "what metadata describe", "how we look at metadata", "which extents of data are described", respectively .

Our specification of the content of metadata, that is, "what metadata describe", is based on the following main concepts:

a SOURCE of statistical information is any process which is activated to observe real world phenomena so as to produce statistical information. A survey is a source,

an administrative data collection is a source too. A source produces data collections by means of applying proper data production techniques and procedures;

a STATISTICAL INFORMATION SYSTEM (SIS) is an integrated collection of pieces of statistical information which concern related phenomena and are issued by different sources. SIS are built for satisfying various and/or unpredictable information requirements. They are typically produced by NSIs for public usage.

These concepts define the contexts which are described by metadata: the documented data come from observing real world objects by means of specific techniques and procedures, for properly using data the analysts need to know what has been observed and the way it has been observed.

Therefore two main classes of metadata are singled out, concerning the information content of a source or a SIS and the characteristics of each source as an observation process, respectively. These are the basic metadata classes. Other relevant metadata concern the quality and the characteristics of the issued data when regarded as the result of a particular repetition of a production process. Such metadata have a different explanation and specification, therefore we separately analyse them. An outcome of our work on such other classes of metadata is our implemented system SIDI, which manages quality indicators together with metadata about the survey production process, in an integrated way (see [8]). Further analysis of the two main metadata classes can be developed along two other dimensions: the level of abstraction and the scope.

A level of abstraction defines a particular “way of looking at" the objects which we want to represent. By means of considering sources and SIS at different levels of abstraction we single out three main metadata layer: the conceptual layer, the organisational layer and the operational layer.

Sources and SIS as knowledge bases are the main objects which are defined in the conceptual layer. Proper classes of conceptual metadata describe the production process which is associated with each source as well as the information content of sources and SIS.

We single out two components in the specification of the information content of sources and SIS. The first component encompasses the specification of the observed part of the real world, which is expressed in terms of elementary statistical concepts such as statistical unit, variable, as well as the specification of structured objects which are derived from such elementary objects. Each concept has a definition and a set of links with the other concepts. We call such a specification a terminology. The second component is the specification of the issued data: their meaning is documented by means of representing them as associations of terms of a terminology. We denote such associations by the name Information Frames. Each source or SIS has its own terminology, however sources can share production procedures and terminology concepts, but they produce their own information frames.

At lower levels of abstraction we describe the distribution of SIS among organisations and the implementation of SIS and sources as data production, transformation and dissemination systems.

In the organisational layer we distinguish among data producers, data users and other organisations, in the operational layer we document the physical data repositories and the data manipulation procedures, that is, those data management and

manipulation systems which are built for supporting the production and the usage of statistical data.

Along the dimension "scope" we distinguish between local and global metadata (see [14]). In the conceptual layer, local metadata are those metadata which describe the information content and the characteristics of each source of information. Global metadata are those metadata which are obtained by means of conceptually integrating or standardising local metadata. The specification of the information content of a SIS is an example of global metadata. Conceptual integration (or, equivalently, harmonisation) is the activity which is performed in order to produce global metadata concerning the information content. It is a complex activity, which requires comparing and matching the conceptual objects which belong to different source terminologies, by means of analysing their definitions; for this purpose, the object definitions must be expressed in a structured shape, as combinations of formally defined constructs. This is the reason why inside an NSI's collection of sources there is never complete conceptual integration of the terminologies and generally only delimited areas of integrated concepts are defined, corresponding to SIS or other sets of partially integrated surveys such as general standards or area standards. On the contrary it is easier to attain a conceptual layer standard description of the surveys' production processes, by means of exploiting shared descriptions of operations (see [8]). In the organisational and operational layer, global metadata should describe those interactions among organisations, processes, data and agents which are involved in statistical data production and usage.

In our metadata framework each metadata class plays the role of a homogeneous viewpoint with which we can associate a documentation model. A documentation model is the specification of a set of meta-objects which have their own meta-properties and are linked by meta-relationships. It allows for specifying in a standard way those metadata which belong to a given metadata class.

3 Modelling metadata

As a consequence of our activity we have especially studied how to model the production process from a documentation viewpoint, for SIDI, and how to model the information content of surveys and other sources, for SDOSIS and ESPLORIS.

SIDI is based on a conceptual layer model which allows for associating each survey with a set of OPERATIONS (see [8]). An operation is a high-level description of survey procedures, such as Data capturing by means of CATI techniques. Each operation is associated with a set of CONTROL ACTIONS, namely particular operations which are performed for monitoring the production procedures. Operations and control actions are performed by AGENTS, moreover they produce and exploit DATA REPOSITORIES.

We have defined a conceptual layer model, called OSI (Objects-Information Frames), for specifying terminologies and information frames of surveys and SIS (see [7] and [3]). OSI aims to support analysing concepts for integration as well as performing complex data manipulation activities. Therefore it strongly differs from the OLAP data model (see [6], [11]), based on the notions of cube, fact, dimension,

which is only aimed to let the analyst perform simple operations on pre-defined data marts. Unfortunately, to support non planned data manipulation is not an objective of the OLAP/DW research, even if many researchers in this area (see for example [1], [4], [10]) stress the need for richer conceptual specification languages. OSI borrows some concepts from those conceptual models which were proposed in the statistical database research area, whose main goal, however, was to support an operational layer activity, namely statistical database design activity (see for example [15], [16], [17]).

The main particular features of our model depend on its underlying analysis-oriented conceptualisation (for similar approaches see for example [2], [5], [9], [13]). From this viewpoint each source is regarded as a different way of observing the reality of interest, data from distinct sources are evaluated and used differently. This is the reason why our model enforces a clear distinction between the specification of the observed part of the real world, which can be shared by different sources, and the description of the data issued by each source. Our modelling approach has two other important features. The first one is the explicit representation of the observed sets of individual objects as STATISTICAL UNITs. The other one is the distinguished specification for the observed qualitative properties of individual objects, on one side, and the set of admissible values for such properties, on the other side. They are represented as CLASSIFICATION VARIABLEs and CLASSIFICATIONs, respectively. In our opinion the lack of such a distinction is an important limit of the Fact/Dimension data model which is used in most of the existing OLAP tools. As an example, let us consider two data such as Total number of University Students by Sex and University Location and Total number of Persons by Age-Class, Sex and Person Residence. Let us simply define Sex and University Location as dimensions for the fact table Total number of University Students by Sex and University Location, and Sex and Person Residence as dimensions for the fact table Total number of Persons by Age-Class, Sex and Person Residence. In this case, the name of the variable is the name of the dimension. Let us suppose that we want to calculate the derived indicator Number of University Students/ Number of 19-25Aged Persons, by Sex and Region. Given the above definition of dimensions, we cannot state if this indicator may be obtained, because we do not know if the two fact table share the classification Region. A different choice is possible, that is, to define Region as a shared dimension for the two fact tables instead of University Location and Person Residence respectively: in such a way however, the exact meaning of the data is not conveyed to the user. In fact, the core of the statistical activity is to observe and measure homogeneous sets of objects: this implies singling out the basic sets of objects of interest (the STATISTICAL UNITs) and partitioning them according to pre-defined sets of values (CLASSIFICATIONs) for their observed qualitative properties (CLASSIFICATION VARIABLEs). The available classifications are established on the basis of the analysts' goals. Therefore, the latter ones are basic meta-objects when the data are actually modelled for satisfying analytical purposes. It is worth noting that in our vision the ultimate goal of the metadata representation is to enable the data user to directly build new data by means of performing meaningful data manipulations. From this viewpoint the problem of the poor definition of the Fact/Dimension model cannot be solved by means of simply adopting proper naming conventions for the modelled

objects because the data user may need to manipulate variables and classification as distinct objects, as in the above example.

The terminology of a survey or SIS is a collection of concepts which describe the information content of the survey or SIS. A set of basic meta-objects is provided for describing the observed part of the real world, namely STATISTICAL UNIT, CLASSIFICATION VARIABLE, NUMERICAL VARIABLE, CLASSIFICATION, IDENTIFIER_SET, ASSOCIATION. The observed part of the real world is described by specifying a network of elementary concepts, each one belonging to one of these meta-objects.

A STATISTICAL UNIT is a set of observable real world individual objects. The notion of statistical unit describes observable populations such as Household, Business, as well as sets of observable events which involve instances of observable populations, such as Household-vacation, Person-hospitalisation.

A CLASSIFICATION VARIABLE is a qualitative property of observable real world individual objects, such as Sex, Economic Activity, which can be used for classifying statistical units. *Identifier* is a special CLASSIFICATION VARIABLE, which is used to name the single items of a statistical unit.

A NUMERICAL VARIABLE is a quantitative property of observable real world individual objects, such as Family Income, Turnover, on which simple summarising function (such as SUM, AVERAGE) can be applied. Any numerical variable has a numerical domain (such as INTEGER, REAL) and a unit of measure if its domain is Real. *Weight* is the special NUMERICAL VARIABLE which is used for counting the number of items of a statistical unit.

A CLASSIFICATION is a set of states which can be observed for some qualitative property of observable real world individual objects. Each classification is associated with an extension which is a list of names of states; as an example, Sex-classification is associated with the set {male, female}. It is well-known that classifications can be organised in CLASSIFICATION SYSTEMS, which establish a set of hierarchical relationships between classifications. An example of official classification system is the NACE classification for the economic activity.

An IDENTIFIER_SET is a name for a set of names of observable real world individual objects. Each IDENTIFIER_SET is associated with an extension which is a list of names of individual objects.

A CLASSIFICATION VARIABLE may be associated with one or more CLASSIFICATIONs, and in particular cases with IDENTIFIER_SETs; a CLASSIFICATION may be associated with one or more CLASSIFICATION VARIABLEs. *Identifier* may only be associated with one or more IDENTIFIER_SETs.

An ASSOCIATION is a one-to-many or a one-to-one relationship between statistical units. It is worth noting that for the analytical purposes it is convenient to regard any many-many relationship between statistical units as a statistical unit too. PART_OF, GROUP are names for special associations.

A SIS or survey terminology specifies relationships among real world objects belonging to one of the above meta-concepts (see at the end of the paper for a diagram which presents the main relationships among OSI meta-objects). STATISTICAL UNITs are associated with CLASSIFICATION and NUMERICAL

VARIABLEs; two STATISTICAL UNITs may be connected by an IS_A (subset) relationship; ASSOCIATIONs connect statistical units; two CLASSIFICATIONs may be connected by an AGGREGATION relationship, inside one or more classification systems. The properties of the documented real world objects encompasses the NAME of the represented object with its SYNONIMS, the OBJECT-DEFINITION, which is expressed in terms of other objects, one or more relationships with other objects of the above described types. An important feature of the OSI model is the availability of a set of CONCEPT DEFINITION CONSTRUCTS, which are used for specifying the definitions of terminology objects in a formal way, so as to support their analysis and conceptual integration. These constructs are also used for deriving new elementary objects from the available ones.

OSI encompasses two other classes of meta-concepts, which are used to describe CONSTRAINTS on the terminology objects and their relationships, and for defining STRUCTURED OBJECTS which are built by associating terminology objects, respectively.

Example of CONSTRAINTS concerning terminology objects are STATISTICAL UNIT PARTITION RELATIONSHIP, NUMERICAL VARIABLE SUM RELATIONSHIP.

A STATISTICAL UNIT PARTITION RELATIONSHIP may be established between a statistical unit and a vector of statistical units which are connected by IS_A (subset) relationships with the given statistical unit. Such a constraint states that the vector of statistical units is a partition of the given statistical unit. A vector of couples CLASSIFICATION VARIABLE/CLASSIFICATION may have been used for defining the partition. A simple example is the partition of Person into Male with Age_class≥14, Male with Age_class<14, Female with Age_class≥14, Female with Age_class<14, which is based on the vector [Sex/SexClassification, Age_class/Age_classGroups].

A NUMERICAL VARIABLE SUM RELATIONSHIP is established between a numerical variable and a vector of numerical variables, when the given numerical variable corresponds to the sum of the vector components.

The most important example of a STRUCTURED OBJECT is the STATISTICAL TABLE meta-object.

The meta-object STATISTICAL TABLE describes a common outcome of a basic data manipulation operation, namely the result of applying a summarising function such as SUM, AVERAGE on the values of a numerical variable or a vector of numerical variables which are associated with the instances of one or more sets. In a statistical context the sets of interest for a summarising operation are statistical units or components of a vector of statistical units, in the latter case, the vector of statistical units is defined by means of establishing a partition relationship with a given statistical unit. As a consequence, generally a statistical table is a collection of elementary components, where each component is the result of applying the given summarising function on one of the given numerical variable for one of the defined subsets of a partitioned statistical unit. An example is the object Number and total Turnover of Businesses by Dimension and Economic Activity. In its definition it is implicit that we have a set of Businesses with their Dimension, Economic Activity, Turnover and Weight, and two classifications for Business Dimension and Economic Activity, for instance Dimension Groups and NACE Groups. We make two

operations: a) partitioning Businesses according to the vector [BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups], b) for each component of the attained vector of statistical units, which is a partition of Businesses, applying an operator SUM on the values of the vector [Turnover, Weight] which are associated with its instances. The components of this statistical table are numbers which represents the total turnover or the number of businesses for one of the subsets which we have obtained by partitioning Businesses in the described way.

Obviously statistical tables may be used to specify the content of the issued data, but issued data are not the only context in which statistical tables occur. In fact, representing statistical table in a terminology is mandatory because often a statistical source directly collects statistical tables instead of, or in addition to, information related to individual real world objects; moreover, components of statistical tables are often involved in the definition of elementary objects such as numerical variables. Another example of a STRUCTURED OBJECT is the RATIO meta-object, which is used for describing those statistical indicators, such as Number of Students for each Professor, which are obtained as ratios between two couples of the kind statistical unit/numerical variable (in the example, the ratio is between Student/Weight and Professor/Weight).

Having defined a terminology for a SIS or a single survey, we can describe the meaning of its issued data. Among such data we consider those data collections which are the direct output of the data capturing and editing procedures as well as those data collections which are the output of further transformation procedures. At a conceptual layer, we specify the content of the issued data as a set of interrelated INFORMATION FRAMES. An information frame is a conceptual object which is specified as a tuple of terms of a terminology. Moreover an information frame refers to a TIMESET, which is a list of temporal references, representing the set of its observation occasions. In most cases TIMESET is the same for all those data which have been issued by a survey which has been repeated several times in a given period. We single out two basic kinds of information frames, SET_OF_INDIVIDUALS and SUMMARY. The former models collections of individual items (so-called microdata), such as List of Students with Sex and Age-class, for each Degree Course, the latter models data which have been obtained by means of summarising pre-existing collections of individual items. More precisely, SUMMARYs are used for modelling so called macrodata, such as Total number of Students by Sex and Age-class in Italy, as well as pre-aggregated data, such as Total number of Students by Sex and Age-class, for each Degree Course, which have links with microdata, in the example List of Degree Courses. It is worth noting that the conceptual relationship linking two information frames is the same which links the statistical units to which they refer (Enrolled in the example). It is worth noting that another kind of information frame, which we do not analyse in this paper, is used for modelling those data which are obtained by means of combining macrodata, such as indicator tables.

Both SET_OF_INDIVIDUALS and SUMMARY information frames are specified according to a template in which a STATISTICAL UNIT is mandatory, together with a TIMESET and the special numerical variable *Weight*. The other components of an INFORMATION FRAME definition may be NUMERICAL VARIABLEs as well as CLASSIFICATION VARIABLES/CLASSIFICATION couples. An

Identifier/IDENTIFIER_SET couple is a mandatory component in a SET_OF_INDIVIDUALS definition. For SUMMARYs, an operation is associated with each NUMERICAL VARIABLE, such as COUNT, SUM, AVERAGE.

A SET_OF_INDIVIDUALS denotes a set of tuples whose components are single instances of the specified component variables, that is, modalities of classifications or values in the domain of a numerical variable. Each tuple contains the instances of the specified concepts which have been collected for an observed individual instance of the specified statistical unit. As an example, let us consider the datum List of Businesses with their Dimension, Economic Activity, Turnover and assume that it has been obtained by means of observing those businesses whose names are listed in List of Businesses Identifiers in those occasions which are listed in List of observation occasions, and that we have recorded the business dimension according to a Business Dimension Classification and the business economic activity according to NACE Groups. Such a datum is specified as

[BUSINESS,
IDENTIFIER*LIST_OF_IDENTIFIER,
BUSINESS_DIMENSION*BUSINESS_DIMENSION_CLASSIFICATION,
ECONOMIC_ACTIVITY*NACE_GROUPS,
TURNOVER, WEIGHT,
LIST_OF_OBSERVATION OCCASIONS].

A SUMMARY denotes a set of tuples which arranges the components of a statistical table. As an example, let us consider the datum Number and total Turnover of Businesses by Dimension, and Economic Activity. It corresponds to the statistical table which is defined for the statistical unit Business in the above described way, by means of partitioning Businesses on the basis of the vector [BusinessDimension/DimensionGroups, EconomicActivity/NACEGroups], and properly applying the SUM function on the vector [Turnover, Weight], for each subset of Business in the obtained partition. Such a datum is specified as

[BUSINESS,
BUSINESS_DIMENSION*DIMENSION_GROUPS,
ECONOMIC_ACTIVITY*NACE_GROUPS,
SUM(TURNOVER),
SUM(WEIGHT),
LIST_OF_OBSERVATION OCCASIONS].

This SUMMARY denotes a set of tuple, in which each tuple is referred to a component of the partitioning vector of the specified statistical unit, Business. Therefore each tuple is composed by that particular combination of modalities of the specified couples BusinessDimension/DimensionGroups and EconomicActivity/NACEGroups which uniquely identifies the referred component of the partitioning vector, together with the values of SUM(TURNOVER) and SUM(WEIGHT) which have been calculated for such a component.

It is worth noting that, despite their different meaning, the two kinds of information frames can be processed in a very similar way. This is a feature that OSI shares with the Fact/Dimension model which is generally used in OLAP applications.

In order to describe the whole set of data issued by SIS or by single sources of statistical information we need to specify relationships among information frames.

Two SET_OF_INDIVIDUALS may be connected by an IS_A (subset) relationship, which is the same which connects their statistical units.

Two SET_OF_INDIVIDUALS may be connected by an ASSOCIATION, which is the same which connects their statistical units. A SUMMARY may be connected with a SET_OF_INDIVIDUALS by an ASSOCIATION, in the role of son. In these cases, special CLASSIFICATION VARIABLE/IDENTIFIER_SET couples occur in the information frame which has the role of son, where IDENTIFIER_SET is referred to the father information frame. In particular, this is the case when a SUMMARY is used for modelling pre-aggregated data.

The data which are observed and released by each source are thoroughly described as associations of terms, when their meaning is what matters.

However, such data are also characterised by the operations which produced them as well as by the operations which can modify them. OSI specifies all the admissible TRANSFORMATIONS which can be applied on information frames. Generally speaking an information frame is obtained by another information frame by means of applying transformations which belong to one of these classes:

a) simply ruling out components: CLASSIFICATION VARIABLE/CLASSIFICATION couples and NUMERICAL VARIABLES for SET-OF-INDIVIDUALS, only NUMERICAL VARIABLES for AGGREGATES;

b) summarising by means of selecting a subset of CLASSIFICATION VARIABLE/CLASSIFICATION couples and then applying a summarising function;

c) summarising by means of choosing, for any CLASSIFICATION VARIABLE/CLASSIFICATION couple, another existing classification which has the role of father in an AGGREGATION relationship with the given classification and then applying a summarising function;

d) selecting a subset of observation occasions;

e) applying elementary transformations to the information frame components which implies deriving new elementary objects: new CLASSIFICATIONs, new CLASSIFICATION VARIABLEs, new NUMERICAL VARIABLEs, or a new reference STATISTICAL UNIT. In all these transformations the OSI concept definition constructs are involved. In particular, a new statistical unit is built by means of specifying selection criteria which involve the given CLASSIFICATION VARIABLEs, or derived CLASSIFICATION VARIABLEs.

OSI provides the data user with a set of information frame transformations which include a rich set of CONCEPT DEFINITION CONSTRUCTS for deriving new objects from the observed ones and therefore is richer than the set of OLAP operations.

4. The ISTAT metadata infrastructure

SIDI (see [8]) is the component of the ISTAT metadata infrastructure which is dedicated to the specification and maintenance of metadata concerning the survey production processes. Indeed SIDI has been designed as a tool for monitoring the

quality of the ISTAT surveys from both a qualitative and a qualitative viewpoint. Therefore it is not only a metadata management system, it also allows for calculating and disseminating standard quality indicators for each ISTAT survey. As a metadata management system, SIDI warrants a standard specification of the survey production processes, which is ensured by means of a network of thesaura. For each meta-concept in our model we have built in the SIDI database a thesaurus of admissible descriptions. In particular, thesaura have been defined for OPERATIONS and CONTROL ACTIONS and other auxiliary concepts. Thesaura have been defined for STATISTICAL UNITS and for the observed phenomena, too, which will be shared with SDOSIS. The most complex thesaura have been given a hierarchical structure, so as to steer the user in choosing the most suitable description, starting from general descriptions and navigating towards more precise concepts. The conceptual links among thesaura which are established by our model are represented in the database. For describing a particular feature of the survey production process the survey manager may choose a description in the thesaurus or insert a new description; in the latter case, the new thesaurus item must be validated by a particular system user, the quality manager, whose role is to keep a good level of standardisation by means of properly managing the content of the thesaura network. This features of SIDI ensures a meaningful concept-based inquiry. The end user chooses one or more operations, one or more control actions, one or more statistical units or phenomena, and the system select those surveys whose production process specification matches such user-defined search criteria; then the end user can select a single survey in this list and navigate across its metadata and quality indicators, or select several surveys and compare their quality indicators.

At present SIDI is implemented and manages metadata describing the majority of the ISTAT surveys, we are now designing the first version of SDOSIS.

SDOSIS is aimed to document the information content of ISTAT surveys as well as the results of any integration activity. Future versions of SDOSIS will directly support the integration activity, by means of offering functionalities for the analysis of terminologies. The present version of SDOSIS is equipped with functionalities for specifying terminologies and information frames of surveys and SIS. In order to warrant a good quality of metadata SDOSIS manages some classes of operational metadata concerning input and output data repositories. In particular, the survey managers can describe the survey questionnaires and store their image, moreover they can specify the physical characteristics of those administrative archives which they exploit as data sources and document the way by which the survey data are disseminated: print tables, files, database relations, data marts. Moreover, the SDOSIS database encompasses a classification repository, in which the set of modalities of each documented classification is stored, together with correspondence tables for linking modalities of different classifications. Unlike the production process, the information content of surveys and SIS cannot be specified in a homogeneous way by means of pre-defined thesauri. A standard specification would require conceptual integration, which is only obtained by means of in-depth analysing the information content of the involved sources. SDOSIS manages a standard terminology, based on official standards, which is represented by means of a network of thesaura which store standard terms for statistical units, variables and classifications. However, the survey managers are not obliged to adopt such a terminology nor to define compatible terms.

They define the survey's own terminology and may declare, for each term, a correspondence with a standard term. As an alternative choice, they may declare a correspondence with a term in another survey's terminology, or with an area standard term, which is shared by a set of similar surveys. In such a way, SDOSIS documents all those situations in which a partial integration of surveys have been established. Because of the more complex context which it documents SDOSIS provides the end user with more inquiry functionalities. In particular, it offers two distinguished concept-based inquiry functionalities which exploit standard and non-standard terms, respectively. Both of them enable the user to choose terms in a network of term repositories concerning statistical units, numerical variables, classification variables, classifications and search for surveys whose description matches the specified criteria. However the first functionality allows the user to choose terms in the standard terminology thesaura, while the other one allows the user to choose terms in repositories of non-standard terms. For the purpose of warranting meaningful inquiries, such repositories include those terms which are shared by several surveys as well as those survey terms which have no correspondence with other terms. After having selected a single survey of interest, the user can navigate across its terminology as well as view its information frames, and view the characteristics of the input and output data repositories. Proper inquiry functionalities are provided for the other system entities: SIS, standard terminologies, local standard terminologies.

As we discussed in the foregoing sections, we have decided to implement ESPLORIS after having observed that most OLAP/Data warehousing tools are not suitable for the requirements of the SIS users (see [3] and [13] for a similar approach). ESPLORIS is a specific tool for implementing multi-source SIS, which employs metadata for steering the users in selecting sources of information and extracting new data from data collections produced by several sources, through navigation and manipulation. ESPLORIS is built around a knowledge base (KB in the following) in which the information content of the implemented SIS is described in terms of the OSI model. The interaction with the data users is based on the conceptual metadata specification stored in the KB. The user interface represents the relevant classes of metadata and their relationships by means of graphs. Operational metadata which describe logical and physical structures and their correspondence with conceptual metadata are represented, too, for the use of several system components.

In the ESPLORIS knowledge base, each source of information is associated with its own information frames, but all information frames are defined by employing shared sets of statistical units, numerical variables, classification variable/classification couples. These sets of elementary concepts describe the real world which is observed by the SIS as a whole. Moreover, a unique network of classification systems is implemented and represented inside the knowledge base.

The system allows the data user to explore the whole set of conceptual metadata in the ESPLORIS KB. Such an activity produces a query on the Data Base component of the system, which stores the data issued from the information sources. To build such a query the user employs a graphical interface, which assists him/her in a step-by-step fashion. Use cases have been employed to model all phases of the user's activity. An interface panel corresponds to each phase.

Exploration of the conceptual domain and definition of the statistical unit and variables of interest: this is the user activity in the first interface panel, called *Navigation Panel*. The *Navigation Panel* presents the network of Statistical Units which are connected by means of IS-A relationships, together with the corresponding network of Information Frames. When the user selects a statistical unit, the Information Frames which have the selected Statistical Units as a component are enlightened. Once the end user has chosen a reference Information Frame, the system presents a star-shaped graph, in which the Association links between the selected Information Frame and other Information Frames are showed. In such a way the user can choose variables of interest among the Classification and Numerical variables of the selected Information Frame as well as build new variables by means of navigating across the linked Information Frames, therefore he/she is enabled to carry out a non-planned data manipulation. As an example, let us consider the variable Type of the Degree Course in which he/she is Enrolled, which is referred to the statistical unit Student. Such a kind of a variable is obtained by means of navigating along the association Enrolled between Student and Degree Course. The existing OLAP tools require to have it previously defined by a data warehouse administrator, in order to include it in a data mart. Thanks to its richer metadata specification and user-friendly interface, ESPLORIS allows the end user to on-line build such a variable when needed. All the components which have been selected in the *Navigation Panel* define the *Conceptual Query*, the first structured object built by the user.

Query building: this is the user activity in the second interface panel, called *Query Panel*. Here the user, starting from a *Conceptual Query*, defines the logical structure of target data, by means of performing classical OLAP operations, such as defining a new Statistical Unit on the basis of a selection condition, choosing more aggregated Classifications, summarising and ruling out variables. In such a way the second structured object is built, called *Logical Query*. On user request, the *Logical Query* is transformed in a set of SQL commands which are required for data retrieval, called *Physical Query*. Data retrieval: this is the user activity in the third interface panel, called *Presentation Panel*. The *Physical Query* is executed and the *Presentation Panel* returns the data set, which is finally presented to the user. In the current version, the available options are data visualisation and export in standard formats (Excel, text files).

Future versions of ESPLORIS will provide the end user with a richer set of transformation operations, in the *Information Frame Transformation* panel. Moreover it will possible to store the new created Information Frames, thus enabling the end user to dynamically build his/her own data marts.

5. Future work

At present, to implement our devised metadata infrastructure by means of developing a metadata management infrastructure is our main activity. A related theoretical work aims to define a complete conceptual framework and a well established methodology for metadata specification, in particular for the information content metadata integration. This also implies documenting the relationships among

metadata at different levels of abstraction, in particular how to make the conceptual specification of data and processes correspond to the description of the same objects in operational terms, as sets of concrete procedures which have input and output datasets. There is another requirement for a metadata specification methodology: to define which views of the conceptual layer specification of data and processes should be communicated and used in different concrete contexts. Finally, what is most important is to promote the extensive usage of metadata in practical data manipulation activities: this is the only way to warrant a good quality of the defined conceptual metadata.

References

[1] R. Agrawal, A. Gupta, S. Sarawagi, “Modeling multidimensional databases,” IBM Research Report, 1995

[2] G. Appel, “Metadata Driven Statistical Information Systems”, Statistical Metainformation Workshop, Luxembourg 1993

[3] P. Barboni, G. D'Angiolini, L. Fanfoni, M. Paolucci, G. Sulsenti "ESPLORIS. Exploring Multi_source Statistical Information Systems through Metadata", NTTS-ETK Conference Crete 2001

[4] L. Cabibbo, R. Torlone, "A logical approach to multidimensional databases", EDBT98

[5] T. Catarci, G. D’Angiolini, M. Lenzerini, “Concept Language for Statistical Data Modeling”, Data and Knowledge Engineering, 1995

[6] E. F. Codd, “Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate”, Technical Report, Codd and Associates, 1993.

[7] G. D’Angiolini, L. Fanfoni, M. Paolucci, "Modelling and Managing Metadata: the ISTAT experience", METANET Project Conference, Voorburg (Netherlands) 2001

[8] G. D’Angiolini, M. Fortini, M. Signore, (1996) “Metainformation Management Systems in the Survey production Process: A System for Survey Quality Control”, Proc. 2nd ASC

[9] G. De Giacomo, P. Naggar, “Conceptual Data Model with Structured Objects for Statistical Databases”, SSDBM 1996

[10] M. Jarke, M. Lenzerini, Y. Vassilou, P: Vassiliadis "Fundamentals of Data Warehouses" Springer- Verlag 1999

[11] R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, "The Data Warehouse Lifecycle Toolkit", Wiley & Sons, 1998

[12] SJ. P. Kent, M. Schuerhoff (1997) "Some Thoughts about a Metadata Management Systems", SSDBM 1997

[13] S. I. Mc Clean, W. Grossman, K. A. Froeschl, (1998) "Towards Metadata-Guided Distributed Statistical Data Processing" NTTS '98

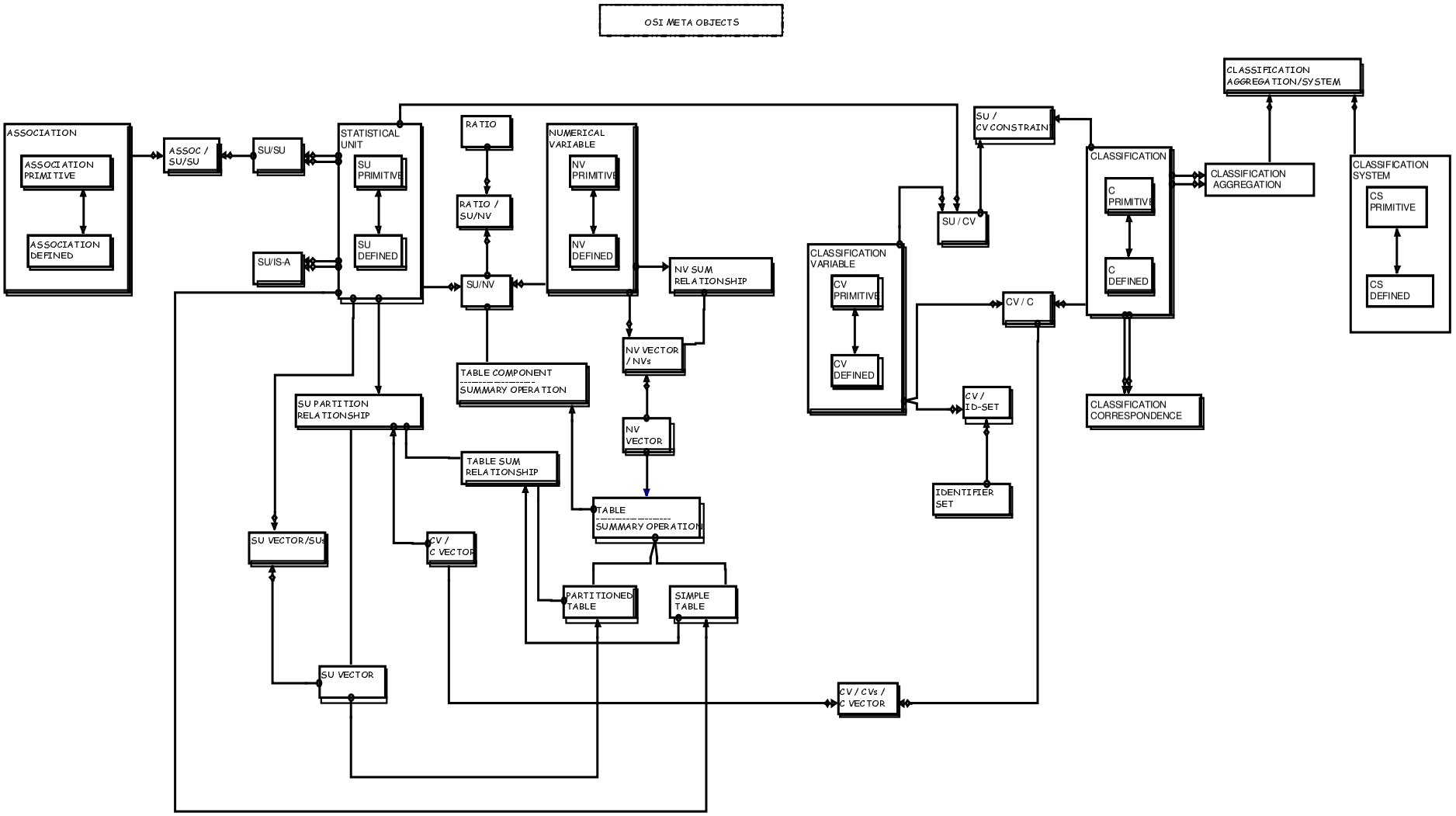
[14] ONU-ECE - "Guidelines for the modelling of Statistical Data and Metadata", 1995

[15] M. Rafanelli, A. Shoshani, “STORM: A Statistical Object Representation Model”, SSDBM 1990

[16] A. Shoshani, "Statistical databases and OLAP: similarities and differences", International Conference on Information ad Knowledge Management, 1996

[17] S.Y.W. Su “SAM*: A semantic association model for corporate and scientific/statistical databases”, Information Sciences, vol.29, No 2-3, May-June 1983

[18] B. Sundgren, “Some properties of statistical information: Pragmatic, Semantic and Syntactis”, Statistics Sweden 1991


OSI META OBJECTS
ASSOCIATION
ASSOCIATION PRIMITIVE
ASSOCIATION DEFINED
ASSOC / SU/SU
SU/SU
SU/IS-A
STATISTICAL UNIT
SU PRIMITIVE
SU DEFINED
RATIO
RATIO / SU/NV
SU/NV
NUMERICAL VARIABLE
NV PRIMITIVE
NV DEFINED
NV SUM RELATIONSHIP
SU / CV CONSTRAIN
SU / CV
CLASSIFICATION VARIABLE
CV PRIMITIVE
CV DEFINED
CV / C
CV / ID-SET
IDENTIFIER SET
CLASSIFICATION
C PRIMITIVE
C DEFINED
CLASSIFICATION CORRESPONDENCE
CLASSIFICATION AGGREGATION/SYSTEM
CLASSIFICATION AGGREGATION
CLASSIFICATION SYSTEM
CS PRIMITIVE
CS DEFINED
NV VECTOR / NVs
NV VECTOR
TABLE COMPONENT
SUMMARY OPERATION
SU PARTITION RELATIONSHIP
TABLE SUM RELATIONSHIP
TABLE
SUMMARY OPERATION
SU VECTOR/SUs
CV / C VECTOR
PARTITIONED TABLE
SIMPLE TABLE
SU VECTOR
CV / CVs / C VECTOR

Inferring and Revising Theories with Confidence: Data Mining the 1901 Canadian Census

Chris Drummond[1], Stan Matwin[1], and Chad Gaffield[2]

[1] School of Information Technology and Engineering,
University of Ottawa,
Ontario, Canada, K1N 6N5
{cdrummon,stan}@site.uottawa.ca
[2] Institute of Canadian Studies,
University of Ottawa,
Ontario, Canada, K1N 6N5
gaffield@uottawa.ca

Abstract. This paper shows how data mining can help historians analyze and understand important social phenomena. Using data from the Canadian census of 1901, we discover the influences on bilingualism in Canada at beginning of the last century. Our approach, based around a decision tree, not only infers theories directly from data but also evaluates existing theories and revises them to improve their consistency with the data. One novel aspect of this work is the use of confidence intervals to determine which factors are both statistically and practically significant, and thus contribute appreciably to the overall accuracy of the theory. When inducing a decision tree directly from data, confidence intervals determine when new tests should be added. If an existing theory is being evaluated, confidence intervals also determine when old tests should be replaced or deleted to improve the theory. Our aim is to minimize the changes made to an existing theory to accommodate the new data. To this end, we propose a semantic measure of similarity between trees and demonstrate how this can be used to limit the changes made.

1 Introduction

The aim of this research is to develop a data mining tool that will help historians explore the influences on the languages spoken in Canada at the beginning of the last century. At the time of Confederation in 1867, language was a secondary issue to other concerns, most notably, religion. By the turn of the century, however, language was becoming an increasingly significant concern in Canada as in other western countries, and during the following decades, it came to be seen as a principal indicator of an individual's identity. While much research has focused on the changing official views of language in Canada, little is known about the actual linguistic abilities of the Canadian population before the later twentieth century.

To address this problem, we apply a data-mining algorithm to the 1901 Canadian census. For the first time, the census asked all residents in Canada three

language questions: mother-tongue, ability to speak English, and ability to speak French. Our research investigates a random five-percent sample of the 1901 enumeration that has been created by the Canadian Families Project. The sample is composed of all individuals living in households that were randomly selected from each microfilm reel of the census enumeration for that year. Households were selected to permit analysis of individuals with relevant social units. The resulting sample is a cluster sample but given the nature and large size of the sample, the design effect is not a concern is this study. For a detailed analysis of this question, see Ornstein [7]. The sample includes data on 231,909 individuals over the age of five, and it allows us to explore how factors such as ethnic origin, mother-tongue, place of birth and residence, age and sex influenced the frequency of bilingualism across Canada. We build upon research that focused on the interpretive implications of how the census questions were posed, and how the actual enumeration was undertaken [3]. We now focus on the responses to these questions written down by the census officials at the doorsteps of individuals and families across the country.

The data mining algorithm we use is the decision tree. Decision trees are easy to understand, even by non-specialists, and have been used by domain experts in many diverse applications [6]. In decision tree learning, an important issue is over-fitting avoidance. A complex tree that fits the training data well typically has unnecessary structure that does not contribute to the accuracy of the tree and may even degrade it. To make the trade-off between accuracy and tree size more principled, we use confidence intervals to prune the tree rather than one of the existing methods. Using confidence intervals allows the determination of not only a statistically significant improvement in the accuracy of the tree, but also to quantify the size of the improvement. A test then will only be added to the tree if the expected accuracy gained is sufficiently large to justify it.

We are interested not only in inferring theories directly from the data but also in testing existing theories, such as those representing the views of politicians of that era, to see if they are confirmed, or indeed contradicted, by the data. Confirmation is likely to be a matter of degree and not all parts of the theory will be affected equally. In this paper, we use a measure of the semantics of a tree to minimize the amount a theory is changed to bring it into accordance with the data. This should help historians not only evaluate an existing theory but also to identify any erroneous assumptions on which it was based.

In the following sections, we first will show how confidence intervals are used to prune a tree grown directly from the data. We then show how our semantic measure combined with confidence intervals and new data is used to evaluate and revise an existing theory on the influences on bilingualism in 1901.

2 Inducing A Decision Tree

A binary tree is used to represent the theories induced from the data. Although sometimes deeper than a tree with a greater branching factor, binary tests should help historians determine not only what are the important attributes but also

the critical values of those attributes. The tree is grown in the standard greedy manner, the best test, according a splitting criterion, is selected to be added to the tree. The main difference is that a test is actually added only when there is a high confidence that a worthwhile increase in accuracy will result.

$$f(a,v) = \max_{a,v} |P(L_{a,v}|+) - P(L_{a,v}|-)| \tag{1}$$

Using the splitting criterion of equation 1, the best split has the greatest difference in the estimated probability of a positive instance going left $P(L_{a,v}|+)$ and a negative instance going left $P(L_{a,v}|-)$ [10]. The criterion is applied to each attribute and each value and the attribute-value with the greatest difference is selected. This value becomes the left branch of the split and the right branch represents the remaining values of the attribute. The difference in likelihood provides a measure of the probability that positive and negative examples come from different distributions. A large difference tends to produce branches with a large difference in class ratios and ultimately leads to better accuracy.

Our aim is to only add tests that improve the accuracy of the tree by a useful amount. But when greedily growing a decision tree adding a single test may not improve accuracy at all. This is often due to the strong imbalance in classes away from the root node. Modifying the class distribution to reduce this imbalance produces a measure that is more likely to show improvement when a single test is added but produces negative values if there is unlikely to be any advantage in adding the test. Based on the training data, the side of split where the positive likelihood is greater than the negative likelihood is labeled positive and the other side negative. Equation 2 gives the accuracy of the split if the left and right hand sides are labeled positive and negative respectively. Here, the role of the probability of each class, $P(-)$ and $P(+)$, is evident. To make a statistic less sensitive to class distribution, the values are replaced by ones closer to 0.5, by applying the squashing function $P'(a) = (P(a) + 1)/(1 + 2)$ to the class probabilities. The resultant statistic can be viewed either as accuracy with a modified class distribution or as the linear combination of two statistics, accuracy and likelihood difference, with the numbers in the squashing function controlling each statistic's influence. The statistic is divided by the fraction of instances reaching the test, and thus estimates the overall improvement in performance.

$$Acc = P(L|+)P(+) + P(R|-)P(-) \tag{2}$$

In decision tree learning, the complexity of the tree is controlled by pruning. In post-pruning, the tree is first grown until it fits the training set well and then extraneous tests, not expected to improve accuracy, are pruned away. In pre-pruning, new tests are only added if they are likely to improve accuracy. Frank [2] experimentally compared the two techniques based on significance tests and found little performance difference. Here, we use pre-pruning but based on confidence intervals rather than significance tests. To generate confidence intervals, we follow the basic procedure proposed by Margineantu and Dietterich

[4]. We apply the same bootstrapping technique [1] (but for a different purpose) as each new test is added to the tree. Rather than discuss the method in detail, we refer to their paper [4]. If two tests are being compared, a three dimensional confusion matrix is used. If we are considering adding or deleting a test, we use the confusion matrix and its row marginals. We apply our test statistic to 500 randomly generated matrices. After sorting the resulting values in ascending order, the fiftieth element will be the lower bound of a 90% one-sided confidence interval.

If this lower bound is greater than zero, we are confident that the gain is statistically significant. In Figure 1 a), H0 is the null hypothesis that the difference is less than or equal to zero, H1 is the alternative hypothesis that adding the test improves accuracy. Not only is the lower bound greater than H0, it is also greater than 0.5%. We can be confident that this test would improve the accuracy of the tree by 0.5%, so the test would be added. If the bound is smaller than the chosen percentage or smaller than H0, see Figure 1 b), the test would not be added. When starting with an existing theory, we are also interested is deleting structure. Applying the same test allows us, see Figure 1 c), to determine that we are confident that removing structure does not degrade performance.

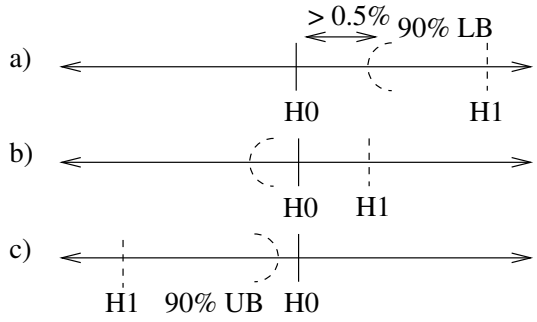


Fig. 1. Using Confidence Intervals for Pruning

The values used to decide when a test should be added where chosen by the authors to represent a reasonable confidence in a useful increase in accuracy. Future work will investigate the effect of varying these values and changing the test statistic used to estimate the increase in accuracy.

3 Theories Induced from Data

In this section, we explore theories generated directly from the data. We use eight attributes from the 1901 census data felt to be potentially relevant to the issue of bilingualism. Some of the nominal attributes have had their values combined into groups and the continuous attribute *age* has been divided into three intervals. To generate the class label *Bilingual*, we combined the attributes *Can speak English*

and *Can speak French* but removed instances where one or both of the attributes were unknown. For the rest of this paper, unilingual will mean can speak French or English and bilingual will mean can speak both. To decide whether a new test should be added, an increase in performance of 0.5% is needed at a confidence level of 90%. Attributes will only be added if the number of instances on each side of the split is greater than 10. Numbers less than 10 might belong to a single family or a related group and be therefore of little interest. The instances are randomly split into a test and training set, 75% of the instances going to the training set. A pruning set is produced from a random 25% of the training set. The splits are all stratified to maintain the class ratios in each set.

Figure 2 is the tree representing the factors that affected bilingualism throughout Canada in 1901. At each leaf the classification is shown: bilingual is labeled "Y" and unilingual is labeled "N". The most important attribute, at the root of the tree, is *mother-tongue*. The split is between those that have French as their mother-tongue "MTONGUE=FR", and those that do not (divided into English, German, Gaelic and Others) "MTONGUE=oth". Notably, for this latter category the tree terminates at a leaf immediately below the root. This classifies all people that do not have French as their mother-tongue as unilingual. The former category is further divided by *birth place*, those born in urban communities "BPLACE=UR" and can write are mostly bilingual. For rural communities "BPLACE=RU", this is only true for males aged 20 to 49. The accuracy gained by adding each attribute is shown to the left of the vertical line. To the right of the line, the total accuracy (80.17%) is labeled "A", the majority classifier accuracy (72.89%) is labeled "MC", the total gain in accuracy (7.28%) is labeled "G" and its 90% lower bound (7.00%) is labeled "LB". The lower bound is generated using bootstrapping on the overall confusion matrix.

```
MTONGUE=FR                          2.91|
|  BPLACE=RU                        1.66|
|  |  AGE=20-49                     1.79|
|  |  |  SEX=F: N                   0.54|A 80.17 MC 72.89
|  |  |  SEX=M: Y                       |G  7.28 LB  7.00
|  |  AGE=oth: N                       |
|  BPLACE=UR                          |
|     CANWRITE=N: N                   0.37|
|     CANWRITE=Y: Y                       |
MTONGUE=oth: N                            |
```

Fig. 2. Decision Tree for Canada

We next explore how the factors that affected bilingualism varied across Canada. Figure 3 shows a map [3] of Canada in 1901 when the census was taken.

[3] ©2001. Government of Canada with permission from Natural Resources Canada

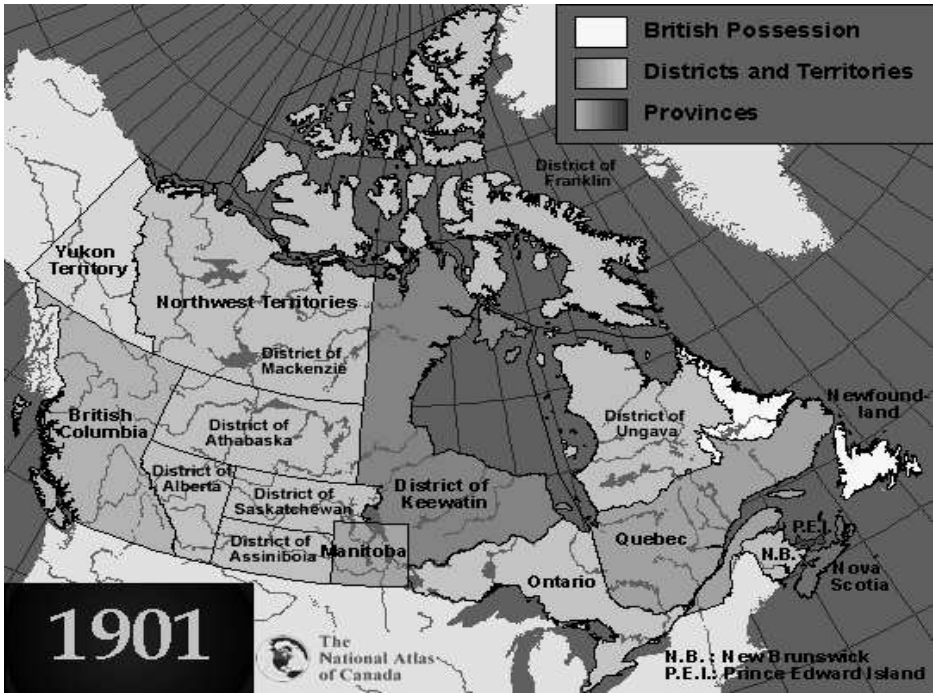


Fig. 3. Map of Canada in 1901

The territories and districts were very sparsely populated at this time. So we combine them into a single region, with a population size more in accordance with other regions. We also make a single region out of the eastern provinces; New Brunswick, Nova Scotia and PEI. We grow decision trees for each of the regions as shown in Figure 4. For British Columbia, the tree consists of the single attribute *mother-tongue* classifying all individuals with a mother-tongue of French as bilingual and all others as unilingual. The majority classifier is already quite accurate, see Figure 4, due the large preponderance of unilingual people in British Columbia. But using the attribute *mother-tongue* correctly predicts a bit over a third of the bilingual people without sacrificing much accuracy on the unilingual ones. Adding extra attributes produces no appreciable improvement. For the territories, the tree has the same root node, but an additional attribute *can read* improves accuracy when the mother-tongue is French. For Manitoba, the tree also has the same root node, but the additional attribute is now *can write*. For Ontario, as for British Columbia, only the single attribute of *mother-tongue* is used. The Eastern provinces have a tree which is similar to Manitoba. *Mother-tongue* is again the most important attribute, adding the attribute *can write* is useful, although it does not improve accuracy on its own. However with an additional attribute excluding children "AGE=5-19" , accuracy is improved.

```
British Columbia
MTONGUE=FR: Y                    4.58|A 91.88 MC 87.30
MTONGUE=oth: N                       |G  4.58 LB  3.48

Territories
MTONGUE=FR                       4.52|
|  CANREAD=N: N                  0.64|A 93.54 MC 88.38
|  CANREAD=Y: Y                      |G  5.16 LB  4.24
MTONGUE=oth: N                     |

Manitoba
MTONGUE=FR                       7.66|
|  CANWRITE=N: N                 0.63|A 89.51 MC 81.22
|  CANWRITE=Y: Y                     |G  8.29 LB  6.72
MTONGUE=oth: N                     |

Ontario
MTONGUE=FR: Y                    7.06|A 94.28 MC 87.22
MTONGUE=oth: N                     |G  7.06 LB  6.69

Eastern Provinces
MTONGUE=FR                       8.64|
|  CANWRITE=N                    0.00|
|  |  AGE=5-19: N                1.76|A 90.66 MC 80.26
|  |  AGE=oth: Y                    |G 10.40 LB  9.55
|  CANWRITE=Y: Y                    |
MTONGUE=oth: N                     |

Quebec
BPLACE=RU                        7.15|
|  AGE=5-19: N                   0.00|
|  AGE=oth                           |
|     SEX=F: N                   1.52|
|     SEX=M                          |
|        CANWRITE=N: N           0.82|
|        CANWRITE=Y                  |
|           MTONGUE=FR: Y        0.27|
|           MTONGUE=oth: N           |A 67.05 MC 53.96
BPLACE=UR                          |G 13.09 LB 12.34
   SEX=F                         0.15|
   |  CANWRITE=N: N               1.56|
   |  CANWRITE=Y                       |
   |     MTONGUE=FR: Y            0.78|
   |     MTONGUE=oth: N                |
   SEX=M                               |
      CANWRITE=N: N                0.84|
      CANWRITE=Y: Y                    |
```

Fig. 4. Regional Decision Trees

For Quebec, a quite different tree is produced. Although the attribute *mother-tongue* is used, it appears much further down the tree, close to the leaves. The most important attribute is *birth place*, indicating if the person was born in a rural or urban community. The attributes used on both sides of this split are very similar. Although for people born in rural communities, children are immediately classified as unilingual. The overall tree is much less accurate than those of the other regions. But as there was a nearly equal number of bilingual and unilingual speakers in Quebec, it still a considerable improvement over the majority classifier.

From an algorithmic perspective, attributes seem generally to be added if, and only if, they result in an increase in accuracy at the leaves of a practically significant amount. For the larger trees this is not always the case. This might be due to using a 90% confidence limit, 10% of the time this limit will not be met. It might also be due to the test statistic not being a direct measure of accuracy. In the latter case, postpruning using accuracy might address the problem, but this remains the subject of future work. For Quebec, it was possible to increase accuracy by about 0.7%, by reducing the confidence interval to 50% and removing the requirement for any gain. But the number of tests went from 9 to 32, so is of debatable merit. With the statistic we use, it is possible to produce a split where the majority class for each branch is the same. This is makes no difference in accuracy and can be removed to make the tree smaller. In fact, for most of the trees this was unnecessary as there was no additional structure.

From a historical perspective, the decision trees are in keeping with some, though not all, of the ways in which politicians, census officials, and other observers at the time discussed the question of bilingualism. The general assumption was that English was becoming an international language of commerce, and that if Canada were to continue developing, everyone in the country should be able to speak it. In contrast, no public figure stressed the importance of learning French. In this sense, the question of bilingualism was directed to two groups: French-language residents and immigrants who did not speak either French or English. The decision trees confirm that the mother-tongue francophones accounted for much of the bilingualism in Canada. Similarly, individuals who were more likely to be involved in commerce were more bilingual. The importance of economic factors is also seen in the greater tendency of middle-aged males in rural areas in Quebec (more likely to be working in rural industries or in the forest economy) to be more bilingual. At the same time, this rural pattern shows how the decision trees diverge from the theories that underlay the contemporary public debate. Specifically, the trees reveal an extent of diversity in language patterns that is inconsistent with how observers characterized Canadian society. For the most part, for example, Quebec was assumed to be a quite homogeneous society especially in the countryside. The general picture was of a unilingual French-language rural world in Quebec that contrasted with the bilingual urban communities of Montreal and to a lesser extent Quebec City. The decision trees reveal that Quebec was indeed a quite distinct part of Canada in terms of bilingualism but that within this distinction there was still considerable diversity.

4 Revising an Existing Tree

In this section, we show how an existing tree is revised so as to minimize the change to the underlying semantics of the theory it represents. The main difference with other forms of theory revision [5, 9] lies in how we quantify changes to the theory and how we use confidence intervals to decide when those changes are worth making. Our notion of the semantics of a decision tree is based on how the tree partitions the attribute space. We capture this semantics by generating instances consistent with the tree. To limit the number of instances, we generalize the definition of an instance so that the probability of an attribute having a particular value is specified. This is similar to the treatment of unknown values in C4.5 [8]. By adding a weight to the instance we can simulate the effect of multiple examples without incurring the additional processing cost. In our approach, the user constructs a decision tree to classify a specified number of imaginary instances, say 1000. An example of what such a tree might look like is shown in Figure 5. Each leaf is marked with the number of individuals from the original thousand that are bilingual and unilingual.

```
MTONGUE=FR
|   CANWRITE=N: N Y(17) N(25)
|   CANWRITE=Y: Y Y(120) N(21)
MTONGUE=oth: N Y(52) N(765)
```

Fig. 5. Simple Domain Theory

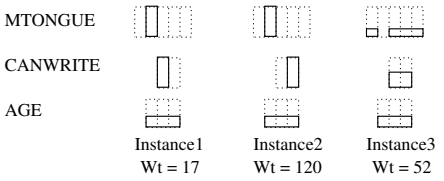


Fig. 6. The Positive Instances

To generate instances consistent with the tree, each path through the tree is represented by as many instances as there are classes at the leaf. Six instances are needed; three for the positive class, bilingual and three for the negative class, unilingual. An instance following a left hand branch has the probability of the attribute value associated with each specified test set to one. For the right hand side branch, the probability is a uniform distribution over the remaining values. Figure 6 shows the probability values for some of the attributes for the positive instances. The negative instances will be identical except for the weights shown at the bottom of Figure 6. The attribute *mother-tongue* has five possible values, indicated by the dashed rectangles. The first two instances travel down the topmost branch of the decision tree. They have the probability of the mother-tongue being French set to one, indicated by the bold continuous rectangle. The third instance, which travels down the bottommost branch, has the probability of the mother-tongue being French set to zero and all other values of mother-tongue are set to a probability of one quarter. The first two instances travel different branches of the attribute *can write*. The first instance has a one for the "N" value, the second instance a one for the "Y" value. All unused attributes on a specific path, such as *age*, have a uniform distribution across all values.

Using these instances, it is now possible to change the order of the tests, or indeed to add a new test, and produce the same partition of the attribute space into classes. Figure 7 shows the effect of changing the root node from *mother-tongue* to *can write*. The same number of instances are classified as bilingual and unilingual. The distribution on the center branch is the same, but the top and bottom most branches have changed. As these two branches are a mixture of instances where the majority class was unilingual, they still classify instances as unilingual. The topmost branch is made up of the first instance in Figure 6 plus half the third instance. The third instance had a uniform probability for *can write*. As this attribute is now the root node, this instance must be sent down both branches. This is achieved by making an additional copy of the instance. For the original instance, the probability of value "N" for *can write* is set to one, the same as the first instance. For the copy the probability for value "Y" is set to one, the same as the second instance. As there are only two values, the weight for both instances is set to half the original weight. If there were more, the weight is the original weight times the fraction of values represented by the branch. There is no longer a uniform distribution for the attribute *mother-tongue*, which was different for the first and third instances. The splitting criterion would choose this attribute as a possible additional test. This would not, however, change the classification of instances. A linear scan across the instances indicates that the classification will not change if new tests are added, so no split is made.

```
CANWRITE=N: N :- Y(43) N(407)
CANWRITE=Y
   MTONGUE=FR: Y :- Y(120) N(21)
   MTONGUE=oth: N :- Y(26) N(383)
```

Fig. 7. Changing the Root Node

To update the tree at each existing test, the splitting criterion is applied to a combination of the data generated to be consistent with the tree and the new data. If the original theory preferred certain attributes, any changes to the theory will tend to use those attributes, rather than introducing new ones, say by promoting them higher up the tree. New tests will only be introduced if the new data has a strong preference for them. To achieve this, the splitting criterion is applied separately to the old and new data. The values returned are combined linearly to form a single value. The coefficients are determined by the number of instances, or weight, of the old data versus the number of new instances.

There are four possibilities that might occur. A new test might be added where the original tree had a leaf. The original test might be replaced by a different test. The original test might deleted altogether, or the old test maintained. To determine which takes place, confidence in the new best test is determined. If the original tree had a leaf at this node, a new test will be added to the tree if the lower bound of the confidence interval is greater than 0.5. This is the same as growing the tree directly from the data. If the new test is the same as the

old test nothing will change. If the new test is different and its confidence interval exceeds the threshold it is compared to the new test. If the lower bound of the confidence interval for the difference exceeds the threshold, the test will be changed. If the new test does not exceed the threshold and the upper bound of the confidence interval on the difference does not include zero, the test is deleted.

The old and new data might also differ in how an instance should be classified at a leaf. A confidence interval can be used to decide which classification should be used. Again a bootstrapping technique is used, this time based on just the binomial ratios. At the leaf we can use lower bound of accuracy directly rather than our test statistic.

5 Evaluating an Existing Theory

In this section, we present an experiment showing how the method discussed in section 4 is applied to a theory representing the views held by politicians in Canada in 1901. The theory has been developed from analyses of debate in the House of Commons and newspaper coverage of political discussion about the language questions posed in the 1901 census. For a comprehensive analysis of the political debate about language, see Gaffield [3]. The decision tree representing the theory, see Figure 8, was designed to classify an imaginary 1000 people. The design exercise began by ranking attributes according to how politicians of that era expected them to influence bilingualism. Each branch of the tree was then assigned some proportion of the 1000 people, indicated by the numbers in parentheses. Next, each attribute was considered for its effect on the proportion of bilingual speakers, and the appropriate ratio of bilingual to unilingual individuals was assigned to each branch. Politicians certainly did not all agree on the importance of various factors and their perceived influence on reported bilingualism, and therefore the experimental parameters represent a distillation of somewhat divergent views.

Ethnic origin was assessed to be the most important attribute, only those of French origin were expected to be bilingual, most other individuals were expected to be unilingual. The next most important attribute was assessed to be *birthplace*, being urban born was more strongly associated with bilingualism than being rural born. Attributes *sex*, *age* and *can write* were then added in that order. Once the tree was constructed, instances consistent with it were generated. The proportions of the classes at the leaves, indicated by the "Y()" and "N()" in the figure, were then adjusted so that the order of attributes was maintained. The tree is reasonably accurate (78.960%), only 1.2% less accurate than the tree grown directly from the data (80.169%).

Figure 9 shows the politicians' theory after revision using data for the whole of Canada. This revised theory is more accurate than the politicians' theory. It is slightly more accurate (80.204% lower bound 79.926%) than the decision tree grown directly from the data (80.169%), see Figure 2, although the base of the tree is identical. Much of the structure from the theory has been deleted, but quite a lot remains, indicated by the "o" and "+" in figure 9. The most significant

```
ORIG=FR :- (400)
|  BPLACE=RU :- (212)
|  |  SEX=F : N :- 0.235 Y(20) N(65)
|  |  SEX=M :- (127)
|  |     AGE=20-49 :- (72)
|  |     |  CANWRITE=N : N :- 0.444 Y(12) N(15)
|  |     |  CANWRITE=oth : Y :- 0.556 Y(25) N(20)
|  |     AGE=oth : N :- 0.364 Y(20) N(35)
|  BPLACE=oth :- (188)
|     SEX=F :- (78)
|     |  AGE=20-49 :- (48)
|     |  |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|     |  |  CANWRITE=Y : Y :- 0.667 Y(20) N(10)
|     |  AGE=oth :- : N :- 0.333 Y(10) N(20)
|     SEX=M :- (110)
|        AGE=>=50 :- (40)
|        |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|        |  CANWRITE=Y : Y :- 0.545 Y(12) N(10)
|        AGE=oth : Y :- 0.714 Y(50) N(20)
ORIG=oth :- : N :- Y(50) N(550)
```

Fig. 8. The Politicians' Theory

change to the theory is the first test, *mother-tongue* replaces *ethnic origin* and accounts for most of the improvement in the revised theory. The additional structure, indicated by the "+'s" in figure 9, is the part of the politicians' theory which was not deleted when the tree was revised. It identifies two bilingual groups for people whose mother-tongue is not French. Urban males (labeled "+1") of French origin are predominantly bilingual, as are urban females (labeled "+2") of French origin, aged 20 to 49 who can write. This branch accounts for the slight increase in the accuracy of the tree. These groups were identified in the original theory. As the data supports this division, albeit very weakly, they have not been deleted. The additional structure, indicated by the "o's", is not supported by the data. It was not deleted, however, as the tests did not indicate a statistically significant increase in accuracy. This structure does not change the classification of the tree and so could easily be deleted.

From an algorithmic perspective, it seems that attributes were modified and deleted when there was a clear advantage in doing so. But when the data did not support such deletion, the semantics of the original theory was maintained. From a historical perspective, the Canadian politicians of 1901 used mother-tongue to help clarify ambiguities among the labels used for ethnic groups; they did not see language as being a good identifier in and of itself. These theory revision experiments suggest that mother-tongue was more important that politicians believed at the time. But they were aware that times were changing, but probably not to the extent to which the data seems to suggest, and this led to addition of language questions to the census.

```
   MTONGUE=FR :- 0.541
   |  BPLACE=RU :- 0.467
   |  |  AGE=20-49 :- 0.575
   |  |  |  SEX=F : N :- 0.451 Y(1547) N(1886)
   |  |  |  SEX=M : Y :- 0.674 Y(2892) N(1399)
   |  |  AGE=oth  : N :- 0.386 Y(3946) N(6278)
   |  BPLACE=UR :- 0.674
   |     CANWRITE=N :- 0.409
o  |     |  ORIG=FR :- 0.408
o  |     |  |  SEX=F : N :- 0.303 Y(240) N(550)
o  |     |  |  SEX=M : N :- 0.499 Y(451) N(452)
o  |     |  ORIG=oth :- : N :- 0.437 Y(35) N(45)
   |     CANWRITE=Y :- 0.732
o  |        SEX=F : Y :- 0.647 Y(2570) N(1405)
o  |        SEX=M : Y :- 0.813 Y(3399) N(782)
   MTONGUE=oth :- 0.101
+     ORIG=FR :- 0.404
+     |  BPLACE=RU : N :- 0.343 Y(348) N(666)
+     |  BPLACE=UR :- 0.506
+     |     SEX=F :- 0.448
+     |     |  AGE=20-49 :- 0.581
+     |     |  |  CANWRITE=N : N :- 0.344 Y(8) N(16)
+2    |     |  |  CANWRITE=Y : Y :- 0.621 Y(90) N(55)
+     |     |  AGE=oth : N :- 0.295 Y(43) N(103)
+1    |     SEX=M :- : Y :- 0.570 Y(162) N(122)
+     ORIG=oth :- : N :- 0.089 Y(3823) N(38942)
```

Fig. 9. The Revised Theory

6 Limitations and Future Work

From a historical perspective, the census was designed to provide evidence of the learning of English by French-language individuals. The trees, indeed, show this but they also show that a constellation of factors underlay the language patterns including age, sex, and rural-urban differences and this was not uniform across the country. It is for this reason more research is needed on specific geographic areas such as the so-called Bilingual Belt as well as on other data from the census including economic variables. From an algorithmic perspective, the test statistic and other design choices have proven effective in practice on this data set but need to be experimentally validated on other data sets. Confidence in an existing theory might not constant for all parts of the theory. The existing theory determined the old tests and influenced the choice of new tests but did not affect the confidence value. An alternative would be to take a more Bayesian approach, perhaps using credible intervals rather confidence intervals, allowing locally defined confidence values.

7 Conclusions

From a historical perspective, the most compelling conclusions concern the extent to which the Quebec patterns appear to differ from those of the other regions of Canada, and the complexity in the patterns of bilingualism at the turn of the century. From an algorithmic perspective, this work has demonstrated how confidence intervals can be used to identify factors that are both statistically and practically significant. It has also shown how combining a semantic measure of similarity between trees with confidence intervals can be used to evaluate and modify an existing theory.

References

1. B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.
2. E. Frank. *Pruning decision trees and lists*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2000.
3. C. Gaffield. Linearity, non-linearity, and the competing constructions of social hierarchy in early twentieth century canada: The question of language in 1901. *Historical Methods*, 33(4):255–260, 2000.
4. D. D. Margineantu and T. G. Dietterich. Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 582–590, 2000.
5. R. J. Mooney. Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning*, 10(1):79–110, 1993.
6. S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.
7. M. D. Ornstein. Analysis of household samples: The 1901 census of canada. *Historical Methods*, 33(4):195–198, 2000.
8. J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
9. G. G. Towell and J. W. Shavlik. Knowledge-based artificial neural networks. *Artificial Intelligence*, 70:119–165, 1994.
10. P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, pages 5–44, 1997.

Calculating economic indexes per household and censal section from official Spanish databases

Sonia Frutos[(1)], Ernestina Menasalvas[(1)], Cesar Montes[(1)], Javier Segovia[(1)]

[(1)]Facultad de Informática, Campus de Montegancedo, UPM. 28660 Boadilla del Monte, Madrid, Spain

Abstract.

In the competitive environments, in which all sorts of organisations move it is of utmost importance to have information about clients. Public databases offer information about households and families. However, the non-crossed and non-georeferenced format of these databases often makes it difficult to extract typologies and information.

There are only two public databases from which to get information at the household or family level in Spain: Population and Housing Censuses, which provide aggregated and georeferenced information, and the Family Expenditure Surveys, which provide information on household consumption, both published by the National Statistics Institute. The two databases cannot be directly cross-referenced, because the Family Expenditure Surveys offer a detailed description of the families, whereas the Census provides the same data but aggregated without cross-references.

In this paper, we define a procedure for cross-referencing these DBs and calculating the economic household indexes for Spanish censal sections that define the average quarterly economic behaviour of the households located in each censal section. The necessary *Symbolic Data Analysis* procedure is based on neural networks and provides an estimate of the trend in these indexes over a series of years. The procedure can be easily extrapolated to similar problems with official data sources from other countries.

1. Introduction

It is of utmost importance in the competitive environments in which all shorts of organisation operate to have geographical, social and economic information about their customers. The National Statistics Institute's public databases offer information about households and families generally (second-order objects or

macro data) from which behaviours can be extrapolated that can be transferred to real customers, but the format and content of these databases is limited by the data protection laws, which often makes it difficult to extract, for example, household typologies and information by geographic location.

In Spain there are only two public databases from which information at the household and family level can be obtained:

- The Population and Housing Censuses (PHC), conducted approximately every 10 years, which provides aggregated information at the censal section level. A **Censal Section** is the smallest administrative unit of information about which there is National Statistics Institute's censal information and is composed of about 400 households over a total of approximately 32000 sections into which the country is divided.
- The quarterly Family Expenditure Surveys (FES) with information on consumption by household of over 300 products.

Both databases are published by the National Statistics Institute (INE) and would yield some economic household indexes by Spanish censal sections of utmost importance today in customer relationship management (CRM) applications. There are many applications that can be obtained from these indexes, for example:

- Evaluate what censal sections in the country are more predisposed (higher index) to expenditure on a given consumer product, so that a company engaged in the sale of this product can easily locate its best points of sale.
- Use income or financial interest indexes for a bank to do a mailing to the households of the censal sections with highest income and likelihood to use financial services.
- Study the trend and periodicity of expenditure on a product to help to plan marketing strategies depending not only on the censal section but also the quarter and year in which it is launched.

The two above-mentioned databases cannot be directly crossed because whereas the Family Expenditure Surveys (FES) offer a detailed description of the families/households surveyed about their expenditure and income, explaining for each one of these the socio-economic condition of the principal earner, the household type, etc., the Census provides the same data but aggregated, without cross-referencing, indicating, for example, how many families there are in a censal section of a given socio-economic level and how many of a type of household, but not their cross-reference, that is, it does not indicate how many there are of a given socio-economic level and at the same time of a type of household. One possible shortcut for overcoming this problem would be to situate the families surveyed in the FES by censal section and then extrapolate, but the FESs do not indicate the source relative to the censal sections of each surveyed family, which means that it is impossible to directly locate the results of a FES in the censal sections.

In this paper, we provide a neural-network based procedure for estimating quarterly household economic indexes for Spanish Censal Sections, based on the

above-mentioned DBs, as well as a linear forecasting model for estimating the trend of each of the defined indexes.

The remainder of the article is organised as follows. Section 2 summarises other approaches taken in the same direction. Section 3 presents the definition of the household economic indexes to be calculated and section 4 details a procedure for estimating these values, as well as their trend over this set of years. Section 5 presents a practical application of the approach presented to validate the technique. Finally, section 6 presents the conclusions and future lines of work.

2. Related Work

The first standard micromarketing tool, called MOSAIC [1, 2, 3, 4] and launched by the British company Experian, appeared in Spain in the early 90s. Others, like the German Bertelsman's REGIO or Equifax's Microvision [7], soon followed. These were first-generation micromarketing tools, oriented at new customer conquest strategies.

These tools are based on having classified the Spanish population by life style typologies, using for this purpose statistical classification techniques. The classification criteria used are descriptive socio-demographic variables, which means that each group or typology is very similar with regard to the variables studied (e.g., age, socio-economic condition, educational level, …) and, at the same time, very different from the members of the other groups or typologies. By means of this typification process, each of the roughly 32,000 censal sections in Spain is associated with a given life style typology.

This type of tools are characterised by:

- They do not explain behaviours; they only describe the socio-demographic characteristics of each typology.
- They use statistical classification techniques (cluster analysis)
- Only one typology is associated with a censal section.

In the late 90s, Bertelsmann-Direct published its Family Expenditure Items Indexes within the Habits® [8] product. These indexes are different from other tools like Mosaic or Regio and even from the Life Style and Consumption indexes of Habits® itself, in that what they express is a distribution of family consumption in a series of products for each censal section, by means of indexes whose philosophy follows the idea to be explained in section 3. However, the statistical technique used to find the model that explains the household variable is Generalized Linear Models, which very much limits the predictive capability of the indexes, as the underlying models are clearly non-linear. Moreover, a measure or application of these techniques as a benchmark with which to compare the technique proposed in this paper, unlike what we do in section 5, is not publicly available.

3. Definition of Household Economic Indexes by Spanish Censal Sections

The household economic indexes by Spanish censal sections to be calculated define the average quarterly economic behaviour of the households located in each Spanish censal section. These indexes are expressed in absolute (euros or pesetas) and relative (percentage) terms with respect to the national, autonomous community and municipal mean.

These economic indexes are divided into three categories, expenditure, earnings and financial:

- **Expenditure indexes per quarter and household**: these are 259 indexes that indicate the expenditure on different budget items that a family living in a Spanish censal section would have and are divided into:
  1. Food, Drink and Tobacco (86 items)
  2. Clothing and Footwear (24 items)
  3. Housing, Heating and Lighting (26 items)
  4. Furniture, Furnishings, Fittings and Current Household Maintenance Costs (29 items)
  5. Medical Services and Health Expenditure (11 items)
  6. Transport and Communications (18 items)
  7. Leisure, Entertainment, Education and Culture (35 items)
  8. Other Goods and Services (20 items)
  9. Other Expenditure (10 items)

  The budget items cover the Household Expenditure totalling 259 products. Their definition coincides with the description in the Continuous Family Expenditure Survey (ECPF'92), Base = 1985, published by the Spanish National Statistics Institute, up to the variety level of (Cod.Ecpf).
- **Quarterly Income by Household** in each censal section. Specifically, the following 7 indexes are defined according to the definition of these income found in the ECPF'92:
  1. Income from employment: monetary and non-monetary
  2. Income from self-employment: monetary and non-monetary
  3. Income from capital and property: monetary and non-monetary
  4. Income from pensions
  5. Income from unemployment benefits
  6. Income from other regular transfers
  7. Other income: monetary and non-monetary
- **Quarterly Financial Indexes per Household** in each censal section of financial interest, the following are derived from the above:
  1. Income: the estimated average income from the sum of the 7 income concepts per quarter and household.
  2. Expenditure: the estimated average consumption of the sum of the 259 items per quarter and household.
  3. Debt: the average quarterly debt per household generated in the censal section.

4. Savings: the average quarterly saving per household generated in the censal section.

4. Data Analysis Procedure for Calculating Indexes and Trend

The procedure followed to estimate the average value of each of the household economic indexes by Spanish censal sections, as well as their trend, is not simple as it involves cross-referencing information from the two DBs, one with georeferenced information -the PHC- and the other with consumption information –the FES-, whose content and fields are different. On the one hand, the Population and Housing Census (PHC) offers an aggregated description of the socio-economic composition of a censal section with parameters like:

- Sex (Male, Female)
- Age (5 ranges)
- ...

But with the constraint that the data are not cross-referenced. For example, the PHC indicates how many households are headed by males, how many by females, how many by people aged from 25 to 30 years, etc., but does not indicate how many there are headed by males who are **also** aged from 25 to 30 years. Why would cross-referencing be useful? Because the FES offers information on consumption by household, also providing information on the socio-economic composition of the household surveyed. If we knew the composition of the households by censal section in the PHC, all we would have to do is to go to the FES and look for surveyed families with a similar composition, look at their expenditure and extrapolate to the families of a similar profile of the censal section. Unfortunately, this is not possible and we have to follow a procedure like the following:

1. Group families surveyed in the FES. The aim is to form groups with surveyed families who are known to live close to each other. These family groups may or may not belong to different censal sections. This is one of the critical points of the methodology. Fortunately, the FES usually include a field that indicates such closeness, either because it is so explicitly, or for other reasons such as indicating that they have been surveyed by the same interviewer on the same day, which leads to think that they are not very far away to prevent he interviewer wasting time travelling.
2. Calculate the socio-economic composition of each family group. The aim is to get the socio-economic description of the non-cross-referenced composition of each family group (percentage of families whose principal earner is male, percentage of families whose principal earner is female, percentage of 1-, 2-, 3-member families, etc., percentage of families whose principal earner has higher education,…) The content of this description must be obtained from the

information available in the survey used, must be as broad as possible and, above all, must be information also available in the PHC.

3. Get the average indexes per family group. For each family group, the average value of each index must be calculated.
   3.1. The income, expenditure, investment and property indexes are calculated by summing their components at the surveyed family level and then calculating the average per group.
   3.2. The saving index is calculated by first getting the savings at the level of surveyed family, where saving is equal to income minus expenditure. The goal of this index is to find out the average amount of saving of the group families, that is, the amount of money the families of the group in question may have available to purchase new goods. For this purpose, the index must be adjusting in its calculation using the value 0 (no saving) in families whose saving is negative. All the saving indexes at family level must be 0 or positive. After this adjustment, the average per group is calculated.
   3.3. The debt index is calculated by first getting the debt at the level of surveyed family, where debt is equal to income minus expenditure. This index is adjusted using the value 0 (no debt) for families whose debt is positive. All the debt indexes at family level must be 0 or negative. After this adjustment, the average is calculated for the group.

It is important to note that the calculated indexes are an average per family and quarter estimated for the period of years covered by the Family Expenditure Survey used, as the family groups used have been selected by geographic proximity irrespective of the time at which the survey was conducted.

4. Get estimation models for each index. The aim is to get estimation models that would take the non-cross-referenced socio-economic description of the composition of a family group as an input and would output an estimate of the value of the average index for all the group families obtained in the above point. These models will then apply to other family groups whose composition is in principle unknown, which means that it is important that these models satisfactorily generalise the solution. The description of the models used here is found in section 4.1 and 4.2.
5. Calculate the socio-economic composition of each censal section. For each censal section of the PHC, the same socio-economic description as calculated in step 2 has to be calculated. This is why it was specified then that the socio-economic description used should be available in both databases.
6. Get indexes per censal section applying the models. Apply the non-linear models calculated in point 4 to each censal section, taking the descriptors obtained in step 5 as input. We would get then the 270 expenditure, income and financial interest indexes for each censal section.
7. To get the indexes relative to nation, autonomous community and municipality, first we calculate the indexes for each censal section, which are averaged out for the nation, autonomous community and municipality, and then the ratio is established with respect to these measures.

8. Get the temporal evolution of the indexes. For each of the 272 indexes, their evolution must be calculated separately for a set of years that cover at least the years during which the Family Expenditure Survey used was conducted.

4.1 Method for estimating each index

Due to the required characteristics, the model to be used must be a **non-linear** technique, which is why a neural net was chosen.

The socio-economic description of the household composition of each group will be the neural net input x_i and the output s_j will be the values of the indexes estimated, by means of a non-linear transformation.

The chosen non-linear transformation is characterised by the use of the transformation function

$$f(a) = 2.5ae^{-a^2}$$

For this purpose, one model per index is created rather than a single model for all the indexes to be estimated. All the models are identical expect for a series of constants that are later adjusted. The non-linear transformation model suited for the procedure follows the following mathematical formulation: le xi be the inputs and I their number and let s be the output to be calculated. The output or index will be calculated using the following non-linear formula,

$$y_j = f(\sum_{i=1}^{I} w_{ji} x_i); \quad s = \sum_{j=1}^{20} \varpi_j y_j \quad where\ f(a) = 2.5ae^{-a^2}$$

The constants to be adjusted for the model to achieve a correct estimation of each index are wji and ϖj. These constants will be adjusted using the well-known back propagation method described in [9]. This method calls for a measure of square error that has to be calculated for all the family groups created. The error will be measured as the square root of the difference between the value of the real index and the value of the estimated index.

4.2 Method for estimating temporal evolution

In a first step, the surveys are grouped according to a time unit, years, quarter, month or web, and the average of the index for each unit of all the surveys conducted during these periods is calculated. This grouping is totally independent and different from the family groupings carried out in point 1, as whereas the latter were based on a criterion of geographic neighbourhood, these are based on a criterion of temporal and geographic proximity jointly, as the data obtained are not statistically significant.

The second step involves adjusting a time forecasting model for each index and applying it in the period of time to be studied. For this purpose, we propose the

following linear prediction model of the index s where Y is an integer that indicates the year, Q the quarter, M the month and W the week in question:

$$s(Y,Q,M,W) = \alpha Y + \beta Q + \chi M + \delta W + \varepsilon$$

The constants α, β, χ, δ, and ε are adjusted by techniques of linear regression techniques on the data obtained in the previous sep. Varying the values of year, quarter, month or week, we will be able to get predictions on the indexes in time periods outside what are included in the survey used.

5. An example of methodology validation: calculation of the income per household in the Autonomous Community of Madrid

The symbolic data analysis methodology has been tested on the 1991 Population and Housing Census (PHC) and the Continuous Family Expenditure Survey (CFES), Base 1985, published by the Spanish INE. The CFES covers 100% of the variables, shares with the PHC many descriptive variables of the households and includes a field with geographic information. The CFES is quarterly and covers the years from 1992 to 1996.

The composition of the household groups has been done using the following 5 variables

- Socio-economic category of the principal earner (7 levels)
- Educational level of the principal earner (7 levels)
- Sex (2 types)
- Age (5 levels)
- Municipality size (4 levels)

Totalling 25 descriptors per family group.

In the CFES, we managed to form 584 family groups that are known to live close geographically. The CFES surveys the same number of families several times during the quarters of the years 1992 to 1996, leading to approximately 106 different families. That is, the 584 groups correspond to grouping of these 106 families surveyed at different times.

Using the database of 584 groups of the CFES, with their 25 socio-economic descriptors and their 270 indicators of income and expenditure, as a training set, the neural models are calculated and applied to the PHC. Additionally, the temporal evolution of the indexes is obtained on the temporal basis of the quarter and calculating its prediction for the years 1997 to 2000.

The model has been extensively validated at private companies, but here we are going to present a validation with an official index. In particular, the Statistics Institute of the Autonomous Community of Madrid annually produces a Municipal

Household Income (MHI indicator). The Autonomous Community of Madrid includes about 1,400,000 households distributed in 148 municipalities.

This indicator is produced mainly on the basis of the Income Tax Returns, particularly, on the basis of the assessment basis of this tax, corrected by other indirect indicators that can be used to estimate this variable and the available family income. The aim is to estimate this important macromagnitude that measure the real expenditure capability of families through the family income obtained in net terms (that is, having removed taxes, deductions and withholdings). The important thing for our purpose is that it is an index produced by direct means and absolutely unrelated to the DBs used in our methodology.

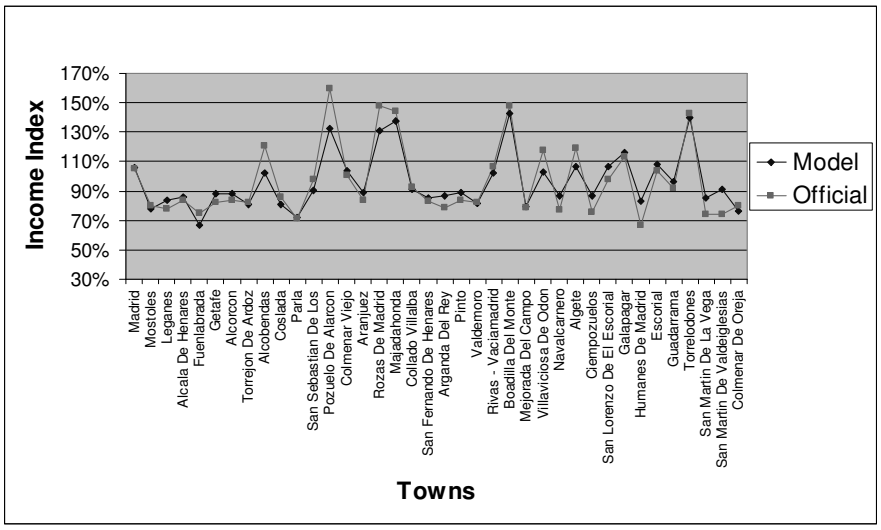


Figure 1. Comparison between the official income index for 1997 and the predicted by the model.

The MHI index used is the 1997 index and is provided at municipality level, which means that index calculated with our methodology must be aggregated to this level, the municipality level. Figure 1 shows a comparison of the indexes for the biggest municipalities in the Autonomous Community of Madrid. The index is a percentage, where 100% represents the average expenditure per household of the Autonomous Community of Madrid. The correlation level of the indexes is high, 0.92.

Deviations can be appreciated in the figure, which we put down to two problems:

- The PHC provides the distribution of the households in 1991, a distribution that has varied considerably in municipalities in expansion during the 1991-1997 period, like Alcobendas and Pozuelo de Alarcon.

- The income index of the model has been calculated by means of the CFES, averaged out for the 1992-1996 period.

It is to be assumed that the degree of adjustment would be greater with more updated databases.

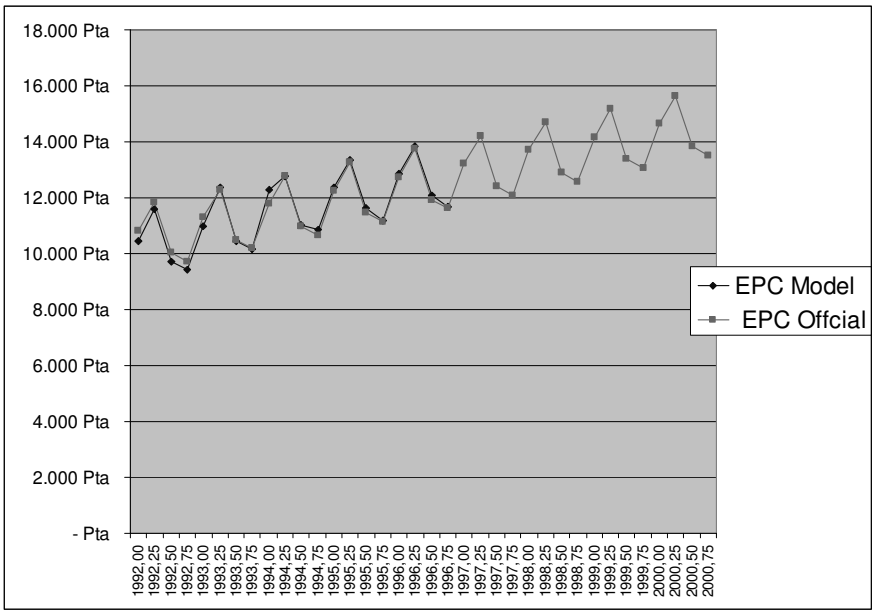


Figure 2. Averaged Electric Power Consumption per home (EPC), in pesetas. Comparison between official and model data. Model includes prediction for period 1997-2000

On the other hand, it is impossible to validate the estimate of the evolution of the indexes, but some results are encouraging. Figure 2 shows the comparison between the estimated expenditure and real consumption of electric power in principal residence per household (EPC). The calculation was done as a means value per family. The official EPC, calculated on families surveyed in the CFES includes the period of 1992 to 1996, whereas the model estimates in the range 1992 to 2000. It is clear how well the model grasps the trend and seasoning features of the index.

6. Conclusions

In this paper, we have presented the definition of the household indexes by Spanish censal sections characterised because they define the average quarterly behaviour of households located in each of the Spanish censal sections, and are expressed in absolute terms (euros or pesetas) and relative terms (percentage) with respect to the national, autonomous community or municipal average.
Additionally, we have established a procedure for estimating the average value of each of the economic indexes of the household by censal section within a set of years preventing the problems of having to work with aggregated data based on official statistics.
The results are promising and easily transferable to DBs in other countries.

References

[1] MOSAIC http://www.micromarketing-online.com/play.htm

[2] Gabbot, M.and Sutherland, E. (1993). 'Marketing information systems in universities', Marketing Intelligence & Planning, Vol.11 (7).

[3] Halsley, A. (1992), 'Opening Wide the Doors of Higher Education', NCE briefing (6), National Commission on Education, London.

[4] Mitchell, V. W., McGoldrick, P., (1993) 'The Role of Geodemographics in segmenting and targeting consumer markets: A Delphi study' European Journal of marketing, Vol 28 No. 5, 1994, pp 54-72, MCB University Press

[5] Sleight, P., (1997), 'Targeting customers, Second edition – How to use Geodemographic and Lifestyle data in your Business', NTC Publications, Oxford.

[6] Tonks, D. and Farr, M. W. (1995) 'Market Segments for higher education', Market Intelligence and planning, Volume 13 (4).

[7] MICROVISION http://www.asnefequifax.es/micromkt.htm

[8] HABITS http://www.bertelsmann-direct.com/Webbertelsmann/

[9] D. Rumelhart, G. Hinton y R. Williams, "Learning representations by back-propagating errors". Nature, 323 pp. 533-536, 1986

Census Data Mining – An Application

Willi Klösgen and Michael May

Fraunhofer Institute for Autonomous Intelligent Systems
Knowledge Discovery Team
D-53757 Sankt Augustin, Germany
{kloesgen, may}@ais.fhg.de

Abstract. Because of data privacy regulations, census data are available for analysis only in aggregated form. Primary data (responses of persons) are aggregated in many cross tabulations for small geographical units. Thus the target objects of secondary analysis are small areas (enumeration districts or wards). Any cell or marginal of a cross tabulation can be used as variable on these target objects. The target objects can be linked with other spatial objects (e.g. rivers, roads, railway lines) for spatial analyses. In this paper we discuss the special requirements that occur for this type of aggregate data mining including spatial analyses. We show an application of SubgroupMiner, which is an advanced subgroup mining system supporting multirelational hypotheses, efficient data base integration, discovery of causal subgroup structures, and visualization based interaction options.

1 Introduction: Mining Spatial Subgroups

The goal of spatio-temporal data mining is to discover attributive, spatial and temporal patterns and to analyse their potential interactions. The patterns describe hypotheses about spatially and timely referenced data. Spatial patterns additionally include variables that do not only refer to properties of the analysis objects themselves (attributive patterns), but also to spatially neighbored objects and their properties. Temporal patterns include analyzing change and trend. In this paper we focus on spatial patterns from the perspective of the subgroup mining paradigm.

Subgroup Mining [Klösgen 1991, 1996, 2002] is used to analyze dependencies between a target and a large number of explanatory variables. The approach can be applied for exploration, classification, or optimization. Interesting subgroups with some designated type of deviation, change, or trend pattern are searched, e.g. subgroups with an over proportionally high target share (mean) for a value of a discrete (continuous) target, or subgroups for which the target share (mean) has significantly changed between two time points, or shows a trend pattern for a sequence of time points. Subgroups are subsets of analysis objects described by selection expressions of a query language, e.g. simple conjunctional attributive selections, or multirelational selections joining several tables representing different (spatial) objects. Interestingness aspects include statistical significance, interpretability, and non-redundancy of subgroups.

A spatial query language that includes operations on the spatial references of objects describes spatial subgroups. A spatial subgroup, for instance, consists of the young children that live near a nuclear power plant of type boiling water reactor. A spatial predicate (nearby) operates on the coordinates of the spatially referenced objects persons and power plants. Further some attributive selectors (age = young, type = boiling_water_ reactor) define which objects belong to the subgroup.

While the spatial dimension is covered by spatial description languages for subgroups, the temporal dimension is represented by change or trend patterns that determine the evaluation criteria for an interesting or statistically significant subgroup.

The subgroup-mining paradigm provides the main components for these approaches: description languages for subgroups, search strategies in hypothesis spaces, hypothesis evaluation, scaling, visualization, and causality analysis.

This paper describes an application of *SubgroupMiner* on census data. The goal of the system is to provide a spatial and temporal mining tool. The system improves all stages of the knowledge discovery cycle:

- Data Access: Subgroup Mining is partially embedded in a spatial database, where analysis is performed. No data transformation is necessary and the same data is used for analysis and mapping in a GIS. This is important for the applicability of the system since pre-processing of spatial data is error-prone and complex.
- Pre-processing and analysis: SubgroupMiner handles both numeric and nominal target attributes. For numeric explanatory variables on-the-fly discretization is performed. Spatial and non-spatial joins are executed dynamically.
- Post-processing and Interpretation: Similar subgroups are clustered according to degree of overlap of instances to identify multicollinearities. A Bayesian network between subgroups can be inferred to support causal analysis.
- Visualisation. SubgroupMiner is dynamically linked to a GIS, so that spatial subgroups are visualized on a map. This allows to bring in background knowledge into the exploration process, performing several forms of interactive sensitivity analysis and exploring relations to further variables and spatial features.

The paper is organized as follows. Section 2 introduces the context of census data. In section 3, the representation of spatial data and spatial subgroups is discussed. The analysis framework is presented in section 4.

2 Census Data

We discuss an application example to illustrate the special requirements of census data mining and especially show the interaction between spatial subgroup mining and GIS mapping. The UK Census, undertaken every ten years, collects population and other statistics essential to those who have to plan and allocate resources. Major customers include departments of national and local government, and providers of services such as health and education.

In the example, we analyse UK 1991 census data for North West England, one of the twelve regions in UK. The basic geographical units used in our analyses are the 1011 wards situated in the 43 local authorities of NW England. Deprivation indices that are the focus of our analysis are given for these wards. The next geographical level below wards are enumeration districts.

Census data can be aggregated to any level of spatial unit. The appropriate level for an analysis depends of the problem and especially the available secondary data (e.g. on deprivation). Lower levels ensure a higher homogeneity of aggregated variables thus providing a higher potential to identify and evaluate hypotheses on individuals (persons). On the other side, lower levels require scalable methods, since the number of the main analysis objects can get very large when the overall region (as North West England) is not strongly limited.

For the 2001 Census England and Wales had 116,895 EDs with an average size close to 200 households (450 people). Census data are available as aggregated cross tabulations for each geographical unit (wards). Table 1 is one (small) cross tabulation of the about 100 tabulations that are provided for different dimensions (economic position, ethnic group, gender in Table 1). Each of the cells of the cross tabulations (e.g. 54 cells of Table 1) can be used as a variable on the geographical units. Thus some 10.000 variables are available for the main analysis objects (wards). Typically a small subset of these variables is selected for a special analysis.

Table S09 Economic position and ethnic group: Residents aged 16 and over							
			Ethnic group				
				Indian			
	TOTAL		Black	Pakistani and	Chinese and other	Persons born in	
Economic position	PERSONS	White	groups	Bangladeshi	groups	Ireland	
TOTAL PERSONS	1	2	3	4	5	6	
Males 16 and over	7	8	9	10	11	12	
Economically active	13	14	15	16	17	18	
Unemployed	19	20	21	22	23	24	
Economically inactive	25	26	27	28	29	30	
Females 16 and over	31	32	33	34	35	36	
Economically active	37	38	39	40	41	42	
Unemployed	43	44	45	46	47	48	
Economically inactive	49	50	51	52	53	54	

Table 1: An aggregated cross tabulation available e.g. for all wards

Also available are detailed geographical layers, among them streets, rivers, buildings, railway lines, shopping areas. Table 2 shows these layers including subtypes for some layers. These layers have own attributes such as featcode (indicating the subtype of the spatial object) or length (of line). Only a few of the many point layers on sports and tourist facilities are included in our analyses, because most of them seem not relevant for the selected target variables.

Layer name	Description	Type	Objects
Motorway	Motorway	Line	494
	Motorway (over), Motorway tunnel		
PrimRoad	Primary route, dual carriageway	Line	3945
	Primary route, dual carriageway (over)		
	Primary route, single carriageway		
	Primary route, single carriageway (over)		
	Primary route, narrow		
	Primary route, narrow (over)		
	Primary route tunnel		
A_Road	A road, dual carriageway	Line	3882
	Other subtypes: see PrimRoad		
B_Road	B road, dual carriageway	Line	4368
	Other subtypes: see PrimRoad		
Mnr_Rd4o	Minor road over 4 meters wide	Line	9705
	Minor road over 4 meters wide (over)		
	Minor road over 4 meters wide tunnel		
Mnr_Rd4u	Minor road under 4 meters wide / over / tunnel	Line	8756
Railway	Railway, standard gauge	Line	4231
	Railway, standard gauge (over)		
	Railway, narrow gauge / narrow gauge (over)		
	Railway tunnel / Railway station		
UrbAreaL	Large Urban Area (outer limit)	Line	384
	Large Urban Area (inner limit)		
UrbAreaS	Small Urban Area (outer limit) / (inner limit)	Line	2235
Water	Inland water (inner limit)	Line	438
	Inland water (outer limit)		
River	River (primary), source / middle / lower	Line	12103
	River (secondary), source / middle / lower		
	River (other and drains)		
Canal	Canal	Line	968
	Canal tunnel / Canal (over)		
Wood	Wood/Forest (inner limit)	Line	859
	Wood/Forest (outer limit)		
Foreshor	Foreshore (sand, inner limit)	Line	209
	Foreshore (other) and offshore rocks (il)		
	Foreshore (sand, outer limit)		
	Foreshore (other) and offshore rocks (ol)		
National	National boundary	Line	12
County	County boundary	Line	88
District	District boundary	Line	61
Park	National park/forest park	Line	11
CampCara	Camping and caravanning combined sites	Point	212
…			

Table 2: Geographic Layers (spatial objects of type line / point)

Deprivation indices are selected as target variables, i.e. the analysis goal is to gain some information on attributive and spatial dependencies of these variables and their interactions. Information from the Census (sometimes in combination with other variables) is often combined into a single index score (Table 3) to show the level of deprivation in an area. Over the years a number of different such indices have been developed for different applications. In general, these measures show a strong correlation between the level of deprivation and a variety of health indicators.

Variable	Jarman	Townsend	Carstairs	DoE
Unemployment	X	X	X(males)	X
Low social class	X		X	
Overcrowded households	X	X	X	X
Households lacking basic amenities				X
Single parent	X			
Under age 5	X			
Lone pensioner	X			
residents who have changed address in the previous year	X			
head of household born in the new commonwealth	X			
Households with no car		X	X	X
Not owner occupied		X		
Children living in flats				X
Children in low earning households				X
Low educational participation				X
Low educational attainment				X
Standard Mortality Ratios				X
Male long term unemployment				X
Income Support recipients				X
Home Insurance Weightings				X

Table 3: Variables used in the calculation of four deprivation indices

Individual variables are usually weighted before they are combined. One of the simplest approaches is to normalize the scores around a mean of zero and express individual components in terms of the number of standard deviations. As a result, the measures are ordinal, hence they are often accompanied by ranking.

3 Representation of Spatial Data and of Spatial Subgroups

Census data, deprivation indices, and the data for the other geographic layers are loaded into a spatial database system (Oracle Spatial). Before analysing the data, a special view is constructed by selecting a subset of the very many census variables and their normalization.

Most modern Geographic Information Systems (GIS) use an underlying Database Management System (DBMS) for data storage and retrieval. In object-relational databases spatial data is represented as follows:

A **spatial data base S** is a set of relations $R_1,...,R_n$ such that each relation R_i in S has a geometry attribute G_i or an identifier A_i such that R_i can be linked (joined) to a relation R_k in S having a geometry attribute G_k.

A **geometry attributes G_i** consists of ordered sets of x,y-pairs defining points, lines, or polygons.

Different types of spatial objects are organized in different relations R_i, e.g. roads, rivers, enumeration districts or wards, buildings. Each such relation is called a **geographical layer.** Each layer can have its own set of attributes A_1, … A_n, called **thematic data**, and at most one geometry attribute G. The attributes $A_1,..., A_n$ are the usual numeric or nominal attributes found in a relational database.

For querying multirelational spatial data, a major extension a spatial database adds is the efficient implementation of a **spatial join**. A spatial join links two relations each having a geometry attribute based on distance or topological relations (inside, covers, adjacent, touches). For supporting spatial joins efficiently, special purpose indexes like KD-trees or Quadtrees are used.

Preprocessing vs. dynamic approaches

The above description shows that a GIS representation is *multi-relational.* A **relation graph** is shown in Figure 1 for seven tables. A link in this graph connects two tables. Foreign keys are simple links: e.g. from diagnoses to persons. Implicit spatial links are given by the spatial references of objects. E.g. a spatial predicate relates persons and industrial plants: a person lives near a plant (either precalculated and materialized as in Figure 1, or dynamically calculated during analysis).

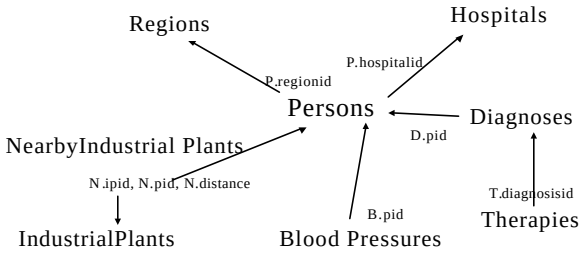


Figure 1: Object classes (tables) of a multirelational spatial application

While the relation graph of Figure 1 has a maximal depth of 3 (e.g. persons, diagnoses, therapies), the relation graph for the census application is a simple star shaped graph. The spatial objects (table 2) are arranged around the target analysis units (wards) such that the maximal depth is only 2.

There are different strategies to deal with multi-relational data in data mining. One possibility is to **preprocess** the data and join the relevant variables from secondary

tables to a target table. The resulting table has a rectangular form and can be analysed using standard methods like decision trees or regression. Multirelational analysis approaches are not necessary in this case.

However the preprocessing approach has disadvantages. First, the set of relations S and the set of possible joins L between tables in S constrain the hypothesis space. Each type of object can spatially interact with any other type of spatial object in numerous ways according to the topological relations. When all these joins are meaningful, the set L is prohibitively large. It would be desirable to set up the problem in such a way that at least in principle all hypotheses are in the search space of an algorithm, or, at least, that it is not the preprocessing that prevents this.

The extended target table generated by preprocessing will include only a small part of the information available in the original tables. But often, it is not known before, in which parts of the data the interesting results can be found. Thus it may be difficult to select the potentially relevant variables and aggregations. When e.g. the number of diagnoses of a person is aggregated from the diagnoses table to the persons table, the correlation of this number with other attributes from the diagnoses table is lost (e.g. number of diagnoses of a special type). Thus it would be necessary to aggregate the diagnosis numbers for each value of (some) attributes of the diagnoses table and possibly also for combinations of values. Thus the number of derived attributes will easy explode for complex relation graphs.

Thus secondly, there is an obvious trade-off between the computing time needed for pre-processing, and the required space for redundant storage on one hand, and the computational complexity of the analysis run. Preprocessing may take a long time, and much of the preprocessing may turn out to be unnecessary since a certain part of the hypothesis space will be pruned away by the data-mining algorithm anyway. It would be desirable to perform expensive spatial joins only for that part of the hypothesis space that is really explored during search.

Thirdly, a further disadvantage occurs in applications where the data can change, either due to adding, deleting or updating. Since pre-processing leads to redundant data storage, we suffer the usual problems of non-normalized data storage, thoroughly investigated in the database literature.

An advantage of preprocessing with respect to a dynamic approach is that once the data is preprocessed the calculation has not to be done again. But here a dynamic approach could cache search results to improving efficiency.

The general approach we apply for multi-relational data mining relies on dynamically joining the tables. The joins that are arranged during search follow paths in a prespecified relation graph. The relation graph includes the edges between table nodes. In the multirelational model we have k object classes O_1, ..., O_k that are represented by tables. They are the nodes of the relation graph. Further we have a set of links where each link is a relation between two object classes and is represented by a prespecified link condition that defines a subset of the product of the two object classes. These links are the edges in the relation graph.

When deciding on dynamic versus static spatial joins for the census application, the following characteristics are important: structure of the relation graph, number

of thematic attributes in geographical layers, size of the relations, and data dynamics. The number of attributes that can be induced for the primary table (e.g. wards) depends on the depth of the relation graph and the number of thematic attributes. E.g. for discrete thematic attributes, an own attribute can be induced for each value and for each combination of values of several attributes (when combinations of attributes are included). Such an attribute could hold the information that a ward is intersected by a road of type A (or of type A and length L). The number of potential induced attributes exponentially grows with the depth of the relation graph. With each additional layer joined, the number of induced attributes is multiplied by the number of combinations of attribute values of the additional layer. Since the structure of the relation graph is simple (depth = 2; only joins between wards and geographic layers and no joins between geographic layers) and the number of thematic attributes is small, the number of induced attributes is manageable for this application.

Also the data dynamics are extremely low. The UK census is undertaken every ten years and also the other layers are faily stable, such that no data update problems occur. Therefore the generation of an overall ward relation extended by all the possible induced attributes from the geographical layers would be preferable, because efficiency (computation time) of analyses will be higher avoiding expensive joins during analyses. This would especially be necessary for finer levels of target objects resulting in large tables (enumeration districts instead of wards). We apply a dynamic join approach (no precalculated universal join) for the analyses (section 4), because the joins are performed for these table sizes (1011 wards and e.g. 9705 roads) within some few seconds such that an interactive analysis is still possible.

In general, an extended dynamic strategy could be useful. This strategy would not require an universal target relation constructed in a preprocessing step, but would dynamically store the induced attributes, which are generated during a multirelation search by joining several tables, into the target table. In a subsequent analysis, it would not be necessary any more to construct the join again, but the stored induced attributes could be accessed from the target table.

Spatial subgroups

A multirelational subgroup is a subset of target objects that is defined by conditions on variables including variables induced from secondary tables. These conditions are described by a query that consists of conjunctive selectors. The query language of SubgroupMiner is described in (Klösgen and May 2002) and is only summarized here referring to the main options of a multirelational subgroup language.

The first option determines which links (joins) between the various (spatial) object classes are selected, i.e. which links are used to construct a (next) conjunctive selector. SubgroupMiner exploits a predefined Relation Graph (Figure 1), that includes the possible links and their details (which attributes and aggregations).

As a basic aggregation option, SubgroupMiner uses the existential quantifier, e.g. the subgroup *Wards.male=high and Rivers.type=primary* is a condensed description of the set of wards with a high share of males and intersected by at least one primary river.

A next option includes aggregate functions such as count, average, max, min (Knobbe et al 2001). The subgroup *Wards.area=large and Rivers.max(length)<l* describes wards with a large area and only intersected by rivers with a limited length.

Another option includes variables to distinguish several objects of one class for applying a predicate on these objects, e.g. *wards situated near two industrial plants with special conditions* (e.g. distance between two plants is small). Such selections are typically included in ILP approaches such as Malerba and Lisi (2001).

The type of refinement is another option. There are two possibilities how a further thematic attribute can refine a subgroup. E.g. the *wards intersected by at least one primary river* can be refined (introducing an additional conjunctive condition) by *wards intersected by at least one primary river and intersected by at least one polluted river*. Another refinement are *wards intersected by at least one primary and polluted river.* The type of refinement is important for aggregation functions.

Details on how these options are applied (e.g. which aggregation functions on which variables, the number of objects to be distinguished and the predicate(s) to be applied on the objects) are prespecified in the Relation Graph.

4 Applying Subgroup Mining to Deprivation Indices

After loading census, deprivation, and geographical data, an Oracle Spatial database holds a table for each census cross tabulation and each geographical layer. As a next preprocessing step, a tool is used to select variables from the very many census tables and to normalize them. Generally we select variables from the margins of the cross tabulations and not so often the inner cells (e.g. for cross table 1: *total Chinese persons* and not *Chinese unemployed males*). Especially for cross tabulations with very deep classifications, the cells are correlated providing (too much) redundancy. Normalization is necessary to adjust the different sizes of wards, thus not the number of Chinese persons, but the number of Chinese persons divided by the total number of persons is included in the resulting ward table. Several normalization options can often be useful, e.g. unemployed males wrt males or total persons. The selection and normalization tool will typically be used many times during an analysis process to include additional variables or to modify normalizations.

With this preprocessing step performed, we can analyse a target table including numerical variables derived from the census (shares such as white persons in a ward related to all residents) and join the target table with geographic layers. Selectors of subgroup descriptions (section 3) need discretizations for numerical variables. Subgroup Miner can automatically discretize the numerical variables during an analysis or rely on predefined discretizations. We at first use the simplest automatic option that generates only two selectors for a numerical variable, e.g. *Wards.males=high* and *Ward.males=low* comparing the percentage of males in a ward to the average percentage over all wards.

In a first experiment, we select *carstairsidx* as target variable and include all selected census variables as well as the other deprivation variables into search to build subgroups. The target variable is numeric and the system uses the mean pattern as a

default. Thus subgroups are searched for which the mean of the target variable is significantly higher than the total mean (over all wards). The found subgroup *low_social=high and married=low and unempl_male=high* (subset of wards with above average value of low_social and below average value of percentage of married persons and above average value of unempl_male) has e.g. an average value for the Carstairs index of 6,24 compared to the overall average of 0,94.

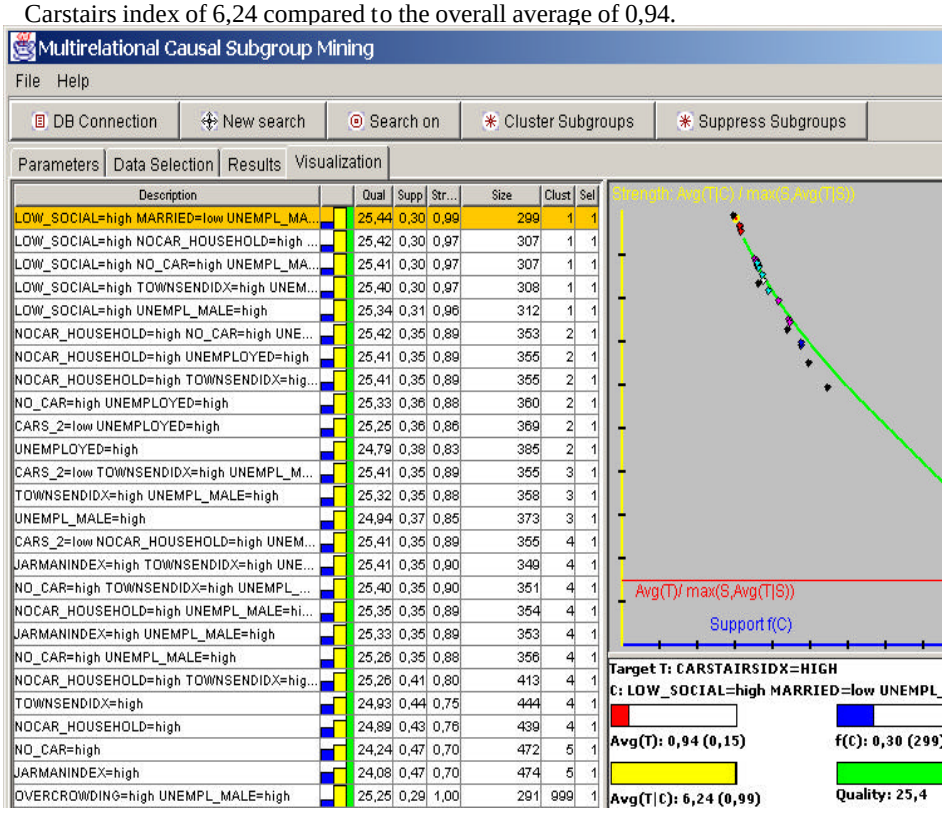


Figure 2: Subgroups with high Carstairs index (all selected census attributes included)

The found subgroups (shown in Figure 2) are ordered by clusters after applying a clustering option (complete linkage method, similarity measure for pair of subgroups based on their overlapping). Five clusters and five remaining single element clusters (999) have been identified. The first cluster includes subgroups with refinements of *low_social=high* and the second cluster subgroups with *no-car_household= high.* The third and fourth cluster include subgroups with *unempl_males=high* (third cluster consists of refinements of fourth cluster). These clusters fairly well reproduce the definition of the Carstairs index (compare section 2). The fourth variable included in the definition of the Carstairs index (overcrowding) occurs in two single element clusters. The results also show the high correlation between the four deprivation indices.

This first experiment has been performed to check the validity of subgroup results. Although some very simple default options have been used such as discretization by average value and mean pattern for ordinal target variable (compare section 2), the results reproduce the definition of the target variable (similar results are found for the negative index, i.e. subgroups with a significantly low average value of the index). More detailed subgroup analyses (not shown here) study the relative importance of the contributing variables of the indices (an index is constructed as a weighted sum of not independent variables that strongly correlate). Other analyses compare the four deprivation indices (e.g. subgroups with a difference of indices).

Since the distribution of many census variables is very skew for the wards, we next apply a more profound discretization based on clustering. Then dense discretization intervals are constructed. When there are e.g. very many small values of a variable and some middle and a few high values, two or three homogeneous intervals are identified based on optimizing the boundary points where e.g. the first interval includes all the small values. This clustering method can especially exclude variables that are not useful to build subgroups (e.g. one cluster includes nearly all values due to the extremely skew distribution of the variable).

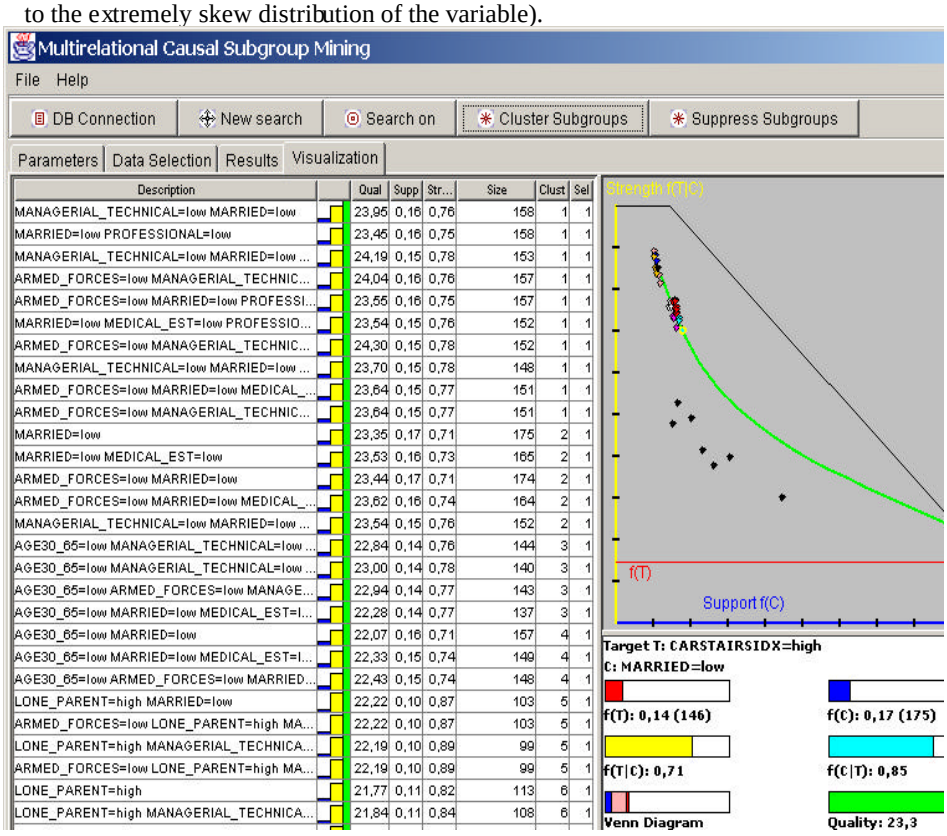


Figure 3: Subgroups dependent of high Carstairs index

Further we exclude all variables that are included in the definition of the Carstairs index from search and also the other deprivation indices. To avoid problems of the mean pattern with only ordinally scaled target variables, we select as target variable the binary variable *carstairsidx=high* (high interval identified with discretization).

Seven subgroup clusters and some single subgroups are identified by the subgroup clustering method (Figure 3). To reduce these results, a redundancy elimination algorithm is run suppressing subgroups that are conditionallly independent of the target group given another subgroup. This Bayesian Network based causality approach (Klösgen 2002) suppresses 24 of the 40 subgroups resulting in the following "causal" subgroups (in the ordering of Figure 3).

Causal Subgroups: 5 7 13 14 23 25 27 28 30 31 33 34 36 37 38 40

A summary of the main "causal" factors (including the dual problem: wards with low Carstairs index) is shown in Table 4.

Variable	Category	Subgroup Size	TargetRate T in Subgroup T = High Carstairs index (14 % of all wards)	TargetRate T in Subgroup T = Low Carstairs index (52 % of all wards)
lone_parent	high	11 %	82 %	
	low	50 %		86 %
age0-4	high	19 %	50 %	
	low	35 %		87 %
unskilled	high	16 %	53 %	
	low	45 %		84 %
long_term_illness	high	22 %	42 %	
	low	34 %		90 %
partly-skilled	high	15 %	49 %	
	low	43 %		89 %
married	low	17 %	71 %	
	high	46 %		91 %
managerial_technical	low	44 %	31 %	
age6-29	low	34 %		88 %
cohabit	low	41 %		75 %

Table 4: Main single factors causing high / low Carstairs index

Underpriviledged wards (e.g. high Carstairs index defined by a combined high rate of unemployed males, low social status, overcrowded households, households with no car) tend to be populated by lone parents, families with young children, unskilled and partly skilled persons, long term ill persons, unmarried persons. The dual properties characterize priviledged wards. Since data are only given as aggregates thus characterizing wards and not individual persons, it can not be concluded that these subgroups (e.g. lone parents or unmarried persons) hold the Carstairs properties on the individual level. Lone parents have not necessarily a low social status, but tend to live in areas with a high rate of persons with a low social status. Using a lower aggregation level (enumeration units or the still more homogeneous output areas) will increase the possibility to infer individual hypotheses. Discussing these problems of aggregate data analysis are beyond the scope of this paper.

Next we analyse the dependence of the Carstairs index of the spatial objects. We include all line objects listed in Table 2 and two thematic attributes for each spatial object class (*featcode* represents subtypes and *length* is a numerical attribute holding the length of the line object). Table 5 summarizes the results.

Subgroups with high average Carstairs index **overall carstairs average for all wards = 0.94**	**Quality** (significance)	**Support** (wards#)	**Carstairs Average**
DISTRICT.DISTRICT_ID=6	6.23	36	5.32
DISTRICT.DISTRICT_ID=22	3.99	35	3.79
DISTRICT.LENGTH=high	3.48	174	1.97
DISTRICT.ALL	3.16	240	1.71
COUNTY.COUNTY_ID=5	4.38	12	6.34
RIVER.ALL	3.26	857	1.13
MNR_RD4U.ALL	1.22	734	1.05
PARK.PARK_ID=2	1.20	76	1.51
Subgroups with low average Carstairs index			
WOOD.LENGTH=high	6.20	215	-0.67
WOOD.FEAT=inner limit	5.90	48	-2.62
WOOD.ALL	5.74	344	-0.13
WATER.LENGTH=high	6.06	128	-1.20
WATER.ALL	4.70	263	-0.12
WATER.FEAT= inner limit	2.23	6	-2.95
PRIMROAD .FEAT=dual carriageway, over other feature	5.14	44	-2.31
PRIMROAD.FEAT= dual carriageway	4.74	152	-0.58
RIVER.FEAT=secondary, source	4.99	301	-0.09
RIVER.FEAT=secondary,middle	4.81	160	-0.55
RIVER.FEAT=primary,lower	2.80	35	-1.05
RIVER.FEAT= primary,source	2.33	45	-0.51
MOTORWAY.LENGTH=low	4.77	145	-0.66
MOTORWAY.ALL	4.60	210	-0.27
MOTORWAY.FEAT=over other feature	4.35	125	-0.62
RAILWAY.FEAT=standard gauge, over other feature	4.68	248	-0.16
RAILWAY.FEAT=tunnel	4,09	34	-2.02
B_ROAD .FEAT=single carriageway, over other feature	4.24	162	-0.36
MNR_RD4O .FEAT=over other feature	4,12	220	-0,11
URBAREAL.LENGTH=low	4,07	247	-0,02
URBAREAL.FEAT=large , inner limit	2,99	79	-0,44
NATIONAL.ALL	3.94	21	-2,71
CANAL.FEAT=over other feature	3,77	38	-1,63
CANAL.LENGTH=high	2,86	141	-0,01
COUNTY.LENGTH=low	3,65	87	-0,66
COUNTY.COUNTY_ID=31	3,02	8	-3,62
PARK.PARK_ID=1	3,54	26	-1,99
DISTRICT.DISTRICT.ID=2	3,47	22	-2,19
MNR_RD4U.FEAT=over other feature	3,47	134	-0,25
A_ROAD .FEAT=single carriageway, over other feature	3,13	105	-0,29
A_ROAD.FEAT=dual carriageway	2,41	102	-0,02

Table 5: Spatial Subgroups with high / low average Carstairs index

(Underprivilidged) wards with a high Carstairs index are situated near (large) district boundaries or boundaries of special single districts or counties, near rivers, and

near main roads under 4 m wide. However, wards situated near the source or middle part of a secondary river or near the source or lower part of a primary river have a low Carstairs index. Also main roads under 4 m wide with this feature dominating another feature (over other) have a low Carstairs index.

There are more spatial characteristics for wards with a low Carstairs index (privilidged wards). They are e.g. located near woods (especially large woods or inner areas of woods), near waters (especially large waters or inner areas), near dual carriageway primeroads, motorways, tunnels of railways, inner parts of large urban areas, national boundaries, long canals.

The way these data mining results are presented to the user is essential for their appropriate interpretation. We use a combination of cartographic and non-cartographic displays linked together through simultaneous dynamic highlighting of the corresponding parts.

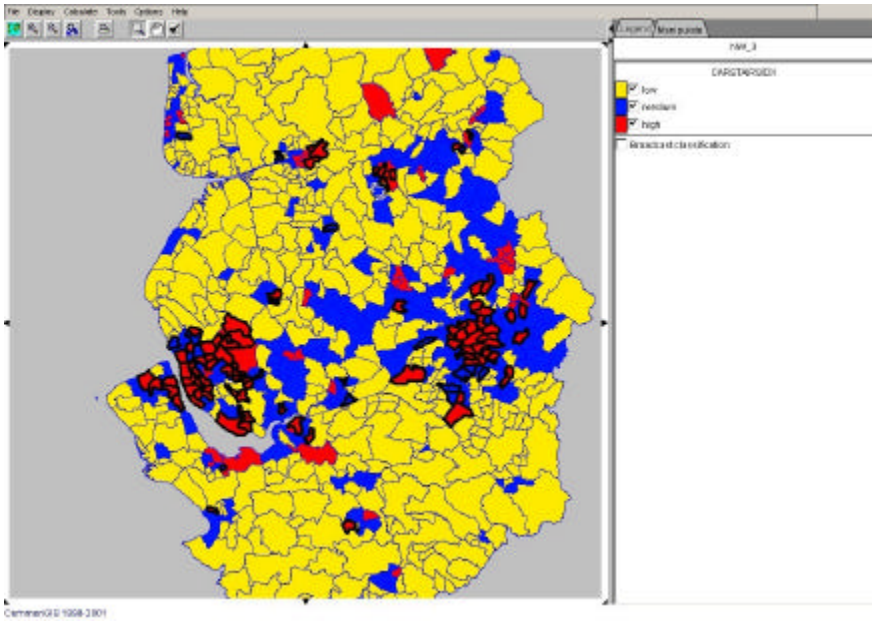


Fig. 5. Wards satisfying the subgroup description C (lone_parent=high) are highlighted with a thicker black line. Wards also satisfying the target (high Carstairs index) are in a lighter color.

The user navigates in the list of subgroups (Fig. 3), which are dynamically highlighted in the map window (Fig. 5). As mapping tool, the the CommonGIS system [Andrienko and Andrienko 1999] is integrated, whose strengths lie in the dynamic manipulation of spatial statistical data.

The application has been developed within the IST-SPIN!-project, that integrates a variety of spatial analysis tools into a spatial data mining platform based on Enter-

prise Java Beans [May and Savinov 2001]. Besides Subgroup Mining these are Spatial Association Rules [Malerba and Lisi 2001], Bayesian Markov Chain Monte Carlo and the Geographical Analysis Machine GAM [Openshaw et al. 1999]. Data are provided by the partners Manchester University and Metropolitan University.

Conclusion and Future Work

Two-layer database integration of multirelational subgroup-mining search strategies has proven as an efficient and easy portable architecture. Scalability of subgroup mining for large datasets has been realized for single relational and multi-relational applications with a not complex relation graph. The complexity of a multirelational application mainly depends of the number of links, the number of secondary attributes to be selected, the depth of the relation graph, and the aggregation operations. Scalability is also a problem, when several tables are very large. Some spatial predicates are expensive to calculate. Then sometimes a grid for approximate (quick) spatial operations can be selected that is sufficiently accurate for data mining purposes. When several large tables are spatially joined, it is advantageous to precalculate the spatial operations. We are currently investigating options to combine static and dynamic links; links can e.g. be declared as static in the relation graph definition. The specification of textual link conditions and predicates in the relation graph that are then embedded into a complex SQL query has proven as a powerful tool to construct multirelational spatial applications. While the analyses of deprivation indices described in this paper treat very general problems with fairly obvious results, a more detailed study on the differences and problems of the various indices is performed as a pilot application within the SPIN! Project.

Acknowledgments

Work was partly supported by SPIN! – Spatial Mining for Data of Public Interest (IST Program, IST-1999-10563, 2000-2002).

References

G. Andrienko, Andrienko, N. Interactive Maps for Visual Data Exploration, *International Journal of Geographical Information Science* 13(5), 355-374, 1999

W. Klösgen 1991. Visualization and Adaptivity in the Statistics Interpreter EXPLORA. In *Proceedings of the 1991 Workshop on KDD*, ed. Piatetsky-Shapiro,G., pp. 25-34.

W. Klösgen 1996. Explora: A Multipattern and Multistrategy Discovery Assistant. *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Cambridge, MA: MIT Press, 249–271.

W. Klösgen 2002. Subgroup Discovery. Chapter 16.3 in: *Handbook of Data Mining and Knowledge Discovery*, eds. W. Klösgen and J. Zytkow, Oxford University Press, New York.

W.Klösgen 2002. Causal Subgroup Mining. To appear.

W.Klösgen, M. May 2002. Spatial Subgroup Mining. In *Proceedings of Sixth European Symposium on Principles of KDD* (PKDD 2002), Berlin:Springer.

A. Knobbe, M. de Haas, A. Siebes 2001. Propositionalisation and Aggregates. In *Proceedings of Fifth European Symposium on Principles of KDD* (PKDD 2001), Berlin:Springer.

D. Malerba, F. Lisi 2001. Discovering Associations between Spatial Objects: An ILP Application. *ILP 2001*, LNAI 2157, Berlin: Springer, 156-163.

M. May, Savinov, A. An Architecture for the SPIN! Spatial Data Mining Platform, *Proc. New Techniques and Technologies for Statistics, NTTS 2001*, 467-472, Eurostat, 2001

Openshaw, S., Turton, I., Macgill, J. and Davy, J. Putting the Geographical Analysis Machine on the Internet, in Gittings, B. (ed.) *Innovations in GIS 6*, 1999

Mining Spatial Association Rules in Census Data: A Relational Approach

Donato Malerba, Francesca A. Lisi, Annalisa Appice, Francesco Sblendorio

Dipartimento di Informatica, Università degli Studi di Bari,
via Orabona 4, 70126 Bari, Italy
{malerba, lisi, appice, sblendorio}@di.uniba.it

Abstract. In this paper we propose a method for the discovery of spatial association rules, that is, association rules involving spatial relations among (spatial) objects. The method is based on a multi-relational data mining approach and takes advantages of the representation and reasoning techniques developed in the field of Inductive Logic Programming (ILP). In particular, the expressive power of predicate logic is profitably used to represent spatial relations and background knowledge (such as spatial hierarchies and rules for spatial qualitative reasoning) in a very elegant, natural way. The integration of computational logics with efficient spatial database indexing and querying procedures permits applications that cannot be tackled by traditional statistical techniques in spatial data analysis. The proposed method has been implemented in the ILP system SPADA (Spatial Pattern Discovery Algorithm). We report the preliminary results on the application of SPADA to Stockport census data.

1 Introduction

Censuses make a huge variety of general statistical information on society available to both researchers and the general public. Population and economic census information is of great value in planning public services (education, funds allocation, public transportation) as well as in private businesses (locating new factories, shopping malls, or banks, as well as marketing particular products).

The application of data mining techniques to census data, and more generally, to official data, has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society [22]. Nevertheless, it is not straightforward and requires challenging methodological research, which is still in the initial stage.

One of the research issues related to mining census data is geo-referenciation. The practice of attaching socio-economic data to specific locations has increasingly spread over the last few decades. In the UK, for instance, household expenditure data are provided for each enumeration district (ED), the smallest areal unit for which census data are published. At the same time, vectorized boundaries of the 1991 census EDs

enable the investigation of socio-economic phenomena in association with the geographical location of EDs. These advances cause a growing demand for more powerful data analysis techniques that can link population data to their spatial distribution.

Spatial data mining methods and techniques have been proposed for *the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases* [13]. In this paper we focus our attention on the specific task of discovering *spatial association rules*, that is, association rules involving spatial objects and relations.

The problem has already been tackled by [12], who implemented the module Geo-associator of the spatial data mining system GeoMiner [10]. This method, however, suffers from severe limitations due to the restrictive data representation formalism, known as *single-table assumption*. More specifically, it is assumed that data to be mined are represented in a single table (or relation) of a relational database, such that each row (or tuple) represents an independent unit of the sample population and columns correspond to properties of units.

In spatial data mining applications this assumption turns out to be a great limitation. Indeed, different geographical objects may have different properties, which can be properly modeled by as many data tables as the number of object types. In addition, attributes of the neighbors of some spatial object of interest may influence the object itself, hence the need for representing object interactions. From a database perspective, this means that two relations are required, one for the *reference* EDs, that is, the EDs whose socio-economical factors are the subject of investigation, and one for the neighboring EDs, which are considered *task relevant*, because they are spatially adjacent to some reference EDs.

The recently promoted *relational* approach to data mining [6], looks for patterns that involve multiple relations of a relational database. Thus data taken as input by these approaches typically consists of several tables and not just a single one, as is the case in most existing data mining approaches. Patterns found by these approaches are called *relational* and are typically stated in a more expressive language than patterns defined in a single data table.

The following is an example of a *relational association rule*:

$$male\text{-}full\text{-}time\text{-}employee\%(X,low) \wedge male\text{-}part\text{-}time\text{-}employee\%(X,low) \wedge neighbor(X,Y) \wedge comm\text{-}activities(Y,high) \rightarrow male\text{-}self\text{-}employed\%(X,high) \quad (32\%,70\%)$$

which states that in 70% of the cases, the low percentage of full-time and part-time male employees in some reference ED *X*, adjacent to another task relevant ED *Y*, with many commercial activities, implies a high percentage of self-employed males in *X*. The relational *pattern*

$$male\text{-}full\text{-}time\text{-}employee\%(X,low) \wedge male\text{-}part\text{-}time\text{-}employee\%(X,low) \wedge neighbor(X,Y) \wedge comm\text{-}activities(Y,high) \wedge male\text{-}self\text{-}employed\%(X,high)$$

occurs in 32% of reference EDs.

It is noteworthy that in this example, and more generally in relational association rules, the items are first-order logic *atoms*, that is, *n*-ary *predicates* applied to *n terms*.

In this example terms can be either *variables*, such as *X* and *Y*, or *constants*, such as *low* or *high*. In other words, subsets of *first-order logic*, which is also called predicate calculus or relational logic, are used to express relational patterns and relational association rules.

Considering this strong link with logics, it is not surprising that many algorithms for multi-relational data mining originate from the field of *inductive logic programming* (ILP) [19, 5, 14, 20]. Extending a single table data mining algorithm to a relational one is not trivial. Efficiency is also very important, as even testing a given relational pattern for validity is often computationally expensive. Moreover, for relational pattern languages, the number of possible patterns can be very large and it becomes necessary to limit their space by providing explicit constraints (*declarative bias*).

However, mining *spatial* association rules is a more complex task than mining *relational* association rules, whose solutions have already been reported in the literature [4]. Two further degrees of complexity are:

1. the implicit definition of spatial relations and
2. the granularity of the spatial objects.

The former is due to the fact that the location and the extension of spatial objects *implicitly* defines spatial relations such as topological, distance and direction relations. Therefore, complex data transformation processes are required to make spatial relations explicit (see the application of machine learning techniques to topographic map interpretation [16]).

The latter refers to the fact that spatial objects can be described at multiple levels of granularity. For instance, UK census data can be geo-referenced with respect to the following hierarchy:

$$\text{ED} \rightarrow \text{Ward} \rightarrow \text{District} \rightarrow \text{County}$$

based on the *inside* relationship between locations. Interesting rules are more likely to be discovered at low granularity levels (ED and ward) than at the county level. On the other hand, large support is more likely to exist at higher granularity levels (District and County) rather than at low levels.

In the next section, a new algorithm for mining spatial association rules is reported. The algorithm, named SPADA (Spatial Pattern Discovery Algorithm), is based on an ILP approach to relational data mining and permits the extraction of multi-level association rules, that is, association rules involving spatial objects at different granularity levels. SPADA has been implemented in Sictus Prolog and is interfaced to an Oracle8i™ database, empowered by an Oracle Spatial cartridge, which enables spatial data to be stored, accessed, and analyzed quickly and efficiently. The system also performs the appropriate data transformation by extracting spatial features (FEATEX module) and by discretizing numerical attributes (RUDE module). The application of SPADA to two data mining tasks involving UK census data is reported in Section 3.

2 Mining spatial association rules with SPADA

The discovery of spatial association rules is a descriptive mining task aiming to detect associations between *reference objects* and some *task-relevant objects*. The former

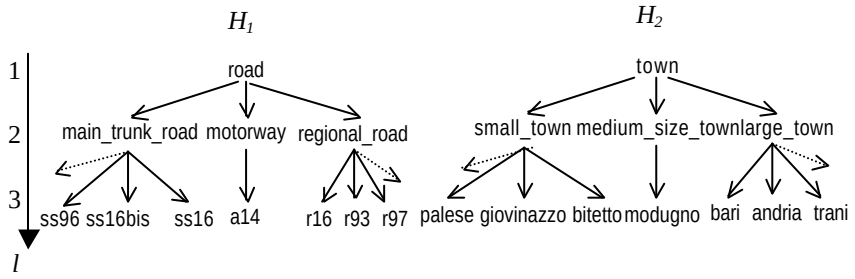


Fig. 1. Two spatial hierarchies and their association to three granularity levels (*l*).

are the main subject of the description, while the latter are spatial objects that are relevant for the task in hand and are spatially related to the former. For instance, we may be interested in describing a given area by finding associations between large towns (reference objects) and spatial objects in the road network, hydrography, and administrative boundary layers (task-relevant objects). The following is an example of spatial association rule that can be generated:

$$is\_a(X,large\_town) \wedge intersects(X,Y) \wedge is\_a(Y,road) \rightarrow$$
$$intersects(X,Z) \wedge is\_a(Z, road) \wedge Z^1 Y \qquad (91\%, 85\%).$$

It states that "***If*** *a large town* X *intersects a road* Y, ***then*** X *intersects a road* Z *distinct from* Y ***with*** *91%* ***support and*** *100%* ***confidence***".

Since some kind of taxonomic knowledge on task-relevant objects may also be taken into account to obtain descriptions at different granularity levels (*multiple-level association rules*), finer-grained answers to the above query are also expected, such as:

$$is\_a(X,large\_town) \wedge intersects(X,Y) \wedge is\_a(Y,regional\_road) \rightarrow$$
$$intersects(X,Z) \wedge is\_a(Z, main\_trunk\_road) \wedge Z^1 Y \qquad (45\%, 90\%)$$

which provides more insight into the nature of the task relevant objects *Y* and *Z*, according to the spatial hierarchy reported in Fig. 1. It is noteworthy that the support and the confidence of the last rule changed. Generally, the lower the granularity level, the lower the support of association rules. Therefore, we follow Han and Fu's [9] proposal to use different thresholds of support and confidence for different granularity levels.

The problem of mining association rules can be formally stated as follows:

Given

- a spatial database (SDB),
- a set of reference objects *S*,
- some sets R_k, $1 \leq k \leq m$, of task-relevant objects
- some spatial hierarchies H_k involving objects in R_k
- *M* granularity levels in the descriptions (1 is the highest while M is the lowest) (see Fig. 1)
- a set of granularity assignments ψ_k which associate each object in H_k with a granularity level
- a *domain specific knowledge DK*
- a *declarative bias DC*

- a couple of thresholds *minsup*[*l*] and *minconf*[*l*] for each granularity level

Find strong multi-level spatial association rules.

An ILP approach to mining spatial association rules has already been reported in [17]. Representation problems, and algorithmic issues related to the application of our logic-based computational method are discussed in the next two sub-sections.

2.1 The representation

The basic idea in our proposal is that a spatial database boils down to a deductive relational database (DDB) once the spatial relationships between reference objects and task-relevant objects have been extracted. The expressive power of first-order logic in databases also allows us to specify background knowledge (BK), such as spatial hierarchies and domain specific knowledge expressed as sets of *rules*, which are stored in the intensional part of the DDB and can support, amongst other things, spatial qualitative reasoning.

Henceforth, we denote the DDB in hand $D(S)$ to mean that it is obtained by adding the spatial relations extracted from SDB regarding the set of reference objects S to the previously supplied *BK*. The ground facts in $D(S)$ can be grouped into distinct subsets: Each group, uniquely identified by the corresponding reference object $s \in S$, is called *spatial observation* and denoted $O[s]$. It is given by:

$$O[s] = O[s|s] \cup \{ O[r \mid s] \mid \text{a spatial relation } q(s,r) \text{ exists in } D(S) \}$$

It contains not only spatial relations between s and some task-relevant object $r \in R_k$ but also spatial relations between r and some $s' \in S$. It is noteworthy that a spatial observation refers to one and only one reference object $s \in S$. The unique reference object associated to a spatial observation allows us to define the support and the confidence of a spatial association rule (see definition below).

Let $A=\{a_1, a_2, \ldots, a_t\}$ be a set of atoms whose terms are either variables or constants (Datalog atoms [2]). Predicate symbols used for A are all those permitted by the user-specified declarative bias, while the constants are only those defined in DDB. Conjunctions of atoms on A are called *atomsets* [3] like the itemsets in classical association rules. In our framework, a language of patterns $L[l]$ at the granularity level l is a set of well-formed atomsets generated on A. Necessary conditions for an atomset P to be in $L[l]$ are the presence of the *key atom* defining a reference object ω at level l, the linkedness [11], and safety. To a pattern P we assign an existentially quantified conjunctive formula $eqc(P)$ obtained by turning P into a Datalog query.

Definition A pattern P covers an observation $O[s]$ if $eqc(P)$ is true in $O[s] \cup BK$.

Definition Let O be the set of spatial observations in $D(S)$ and O_P denote the subset of O containing the spatial observations covered by the pattern P. The *support* of P is defined as $\sigma(P) = |O_P| / |O|$.

Definition A spatial association rule in $D(S)$ at the granularity level l is an implication of the form

$$P \rightarrow Q \ (s\%, c\%)$$

where $P \cup Q \in L[l]$, $P \cap Q = \emptyset$, P includes the key atom and at least one spatial relationship is in $P \cup Q$. The percentages s% and c% are respectively called the support and the confidence of the rule, meaning that s% of spatial observations in $D(S)$ is covered by $P \cup Q$ and c% of spatial observations in $D(S)$ that are covered by P is also covered by $P \cup Q$.

Definition The support and the confidence of a spatial association rule $P \rightarrow Q$ are given by $s = \sigma(P \cup Q)$ and $c = \varphi(Q|P) = \sigma(P \cup R) / \sigma(P)$.

In multi-level association rule mining, an *ancestor* relation between two patterns at different granularity levels $PL[l]$ and $P' \in L[l']$, $l < l'$, exists if and only if P' can be obtained from P by replacing each spatial object $h \in H_k$ at granularity level $l = \psi_k(h)$ with a spatial object $h' < h$ in H_k, which is associated with the granularity level $l' = \psi_k(h')$.

The frequency of a pattern depends on the granularity level of task-relevant spatial objects.

Definition Let $minsup[l]$ and $minconf[l]$ be two thresholds setting the minimum support and the minimum confidence respectively at granularity level l. A pattern P is *large* (or frequent) at level l if $\sigma(P) \geq minsup[l]$ and all ancestors of P with respect to the hierarchies H_k are large at their corresponding levels. The confidence of a spatial association rule $P \rightarrow Q$ is high at level l if $\varphi(Q|P) \geq minconf[l]$. A spatial association rule $P \rightarrow Q$ is *strong* at level l if $P \cup Q$ is large and the confidence is high at level l.

2.2 Method

The task of mining spatial association rules itself can be split into two sub-subtasks:

1. Find large (or frequent) spatial patterns;
2. Generate highly-confident spatial association rules.

Algorithm design for frequent pattern discovery has turned out to be a popular topic in data mining. The blueprint for most algorithms proposed in the literature is the levelwise method [18], which is based on a breadth-first search in the lattice spanned by a generality order $\geq$ between patterns. The space is searched one level at a time, starting from the most general patterns and iterating between the candidate generation and candidate evaluation phases. The algorithm SPADA implements the aforementioned levelwise method.

The pattern space is structured according to the θ-subsumption [21]. Many ILP systems adopt θ-subsumption as the generality order for clause spaces. In this context we need to adapt the framework to the case of atomsets. More precisely, the restriction of θ-subsumption to *Datalog queries* (i.e. existentially quantified conjunctions of Datalog atoms) is of particular interest.

Definition Let Q_1 and Q_2 be two queries. Then Q_1 *q-subsumes* Q_2 if and only if there exists a substitution θ such that $Q_1 \supseteq Q_2\theta$.

We can now introduce the generality order adopted in SPADA.

Definition Let P_1 and P_2 be two patterns. Then P_1 is more general than P_2 under θ-subsumption, denoted as $P_1 \geq_\theta P_2$, if and only if P_2 θ-subsumes P_1.

It is noteworthy that $\geq_\theta$ on patterns represented as Datalog queries is monotone with respect to support, which is the criterion for candidate evaluation in SPADA. The quasi-ordered set spanned by $\geq_\theta$ can be searched by a *refinement operator*, namely a function which computes a set of refinements of a pattern. In particular, we need a refinement operator under θ-subsumption that enables the bottom-up search of the pattern space from the most specific to the most general patterns.

Definition Let $\langle G, \geq_\theta \rangle$ be a pattern space ordered according to $\geq_\theta$. An *upward refinement operator under q-subsumption* is a function ρ such that $\rho(P) \subseteq \{Q \mid Q \geq_\theta P\}$.

Such a refinement operator drives the search towards patterns with decreasing support, therefore all refinements $\rho(P)$ of an infrequent pattern P are infrequent. This is the first-order counterpart of one of the properties holding in the family of the Apriori-like algorithms [1], on which the pruning criterion is based.

For each granularity level ℓ(), SPADA generates and evaluates candidates by searching the pattern space. The *candidate generation* phase consists of a refinement step followed by a pruning step. The former applies the refinement operator under θ-subsumption to patterns previously found to be frequent by preserving the property of linkedness [11]. The latter mainly involves verifying that candidate patterns do not θ-subsume any infrequent pattern. Further pruning criteria have been implemented in SPADA. In particular, the system checks that candidates are not alphabetic variants of previously discovered patterns. The complexity of this test is $O(n^2)$, where n is the number of atoms in the two patterns to be compared. The *candidate evaluation* phase is performed by comparing the support of the candidate pattern with the minimum support threshold set for the level being explored. If the pattern turns out not to be a large one, it is rejected.

2.3 Integrating SPADA with other software components

The application of the ILP approach to spatial databases is made possible by a middle-layer module for feature extraction, as shown in Fig. 2. This layer is essential to cope with one of the main issues of spatial data mining, namely the requirement of complex data transformation processes to make spatial relations explicit.

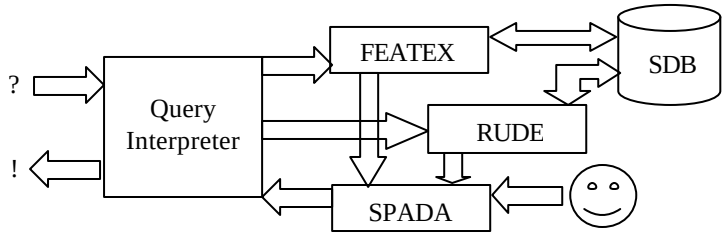


Fig. 2. Integration of SPADA with other software modules which support spatial feature extraction (FEATEX) and discretization of numerical features (RUDE). Additional input to SPADA, such as declarative bias and background knowledge, is directly provided by the user.

This function is partially supported by the spatial database (SDB), which offers spatial data types in its data model and query language and supports them in its implementation, providing at least spatial indexing and efficient algorithms for spatial join [8]. Thus spatial databases supply an adequate representation of both single objects and spatially related collections of objects. In particular, the abstraction primitives for spatial objects are point, line and region. Among the operations defined on spatial objects, spatial relationships are the most important because they make it possible, e.g., to ask for all objects in a given relationship with a query object. The Oracle Spatial cartridge implements the 9-intersection model [7] to support the computation of some topological relations.

Many spatial features (relations and attributes) can be extracted from spatial objects stored in SDB. They can be categorized as follows:

1. *geometric*, that is, based on the principles of Euclidean geometry;
2. *directional*, that is, regarding relative spatial orientation in 2 or 3D;
3. *topological*, that is, binary relations that preserve themselves under topological transformations such as translation, rotation, and scaling;
4. *hybrid*, that is, features which merge properties of two or more of the previous three categories.

This variety requires the development of a feature extractor module, named FEATEX, which also enables the coupling of SPADA with the SDB. FEATEX is implemented as an Oracle package of procedures and functions implemented in the PL-SQL language. In this way, it is possible to formulate complex SQL queries involving both spatial and aspatial data (e.g., census data). The set of spatial features that can be extracted by this module is reported in Table 1.

Table 1. Spatial features extracted by the feature extractor module.

Feature	Meaning	Type	Values
almost_parallel(Y, Z)	Parallelism relation between Y and Z	Hybrid relation	{true, false}
almost_perpendicular(Y,Z)	Perpendicularity relation between Y and Z	Hybrid relation	{true, false}
density(Y, Z)	AREA(Y)/AREA(Z)	Hybrid relation	Real
direction(Y)	Geographic direction of object Y	Directional attribute	{north, east, north_west, north_east}
distance(Y,Z)	Distance between Y and Z	Geometrical relation	Real
layer_name(Y)	Object Y type	Aspatial attribute	Layer name
line_shape(Y)	Object Y shape	Geometrical attribute	{Straight, curvilinear}
relate(Y,Z)	Topological Relation between Y and Z	Topological attribute	Type of topological relation

Since SPADA, like many other association rule mining algorithms, cannot process numerical data properly, it is necessary to perform a discretization of numerical features with a relatively large domain. For this purpose we have implemented the relative unsupervised discretization algorithm RUDE [15] which proves to be suitable for dealing with numerical data in the context of association rule mining. At the end of all this data processing, query results are stored in temporary database tables. An ad-hoc PL-SQL function transforms these tuples into ground Datalog facts of *D(S)*.

3 Application to Stockport census data

In the context of the SPIN! project we investigated the application of spatial data mining techniques to some issues reported in the Unitary Development Plans (UDP) of Stockport, one of the ten Metropolitan Districts of Greater Manchester, UK.

3.1 The data

Spatial analysis is made possible by the use of the Ordnance Survey's digital maps of the district, where several interesting layers are available, namely ED/ward/district boundaries, roads, bus priority lines, and so on. In particular, Stockport is divided into twenty-two wards for a total of 589 EDs. By joining UK 1991 census data available at the ED summarization level with ED spatial objects it is possible to investigate socio-economic issues from a spatial viewpoint. In total 89 tables, each having 120 attributes on average, have been made available for policy analysis. Census attributes provide statistics on the population (resident at the census time, ethnic group, age, marital status, economic position, and so on), on the households in each ED (number of households with *n* children, number of households with *n* economically inactive people, number of households with two cars, and so on) as well as on some services available in each ED (e.g., number of schools).

For the application of our spatial association rule mining method we have focused our attention on transportation planning, which is one of the key issues in UDP.

3.2 Characterizing the area crossed by the M63 motorway

One of the problems is a decision-making process concerning the M63 motorway. More precisely, we are asked to describe the area of Stockport served by the M63 (i.e. the wards of Brinnington, Cheadle, Edgeley, Heaton Mersey, South Reddish) from the sociological viewpoint, in order to provide some hints for transport planners. The data considered in this analysis concerns census statistics on commuters. The description of the area is expressed by some spatial association rules at two levels of granularity. A hierarchy for the Stockport ED layer has been obtained by grouping EDs on the basis of the ward they belong to (see Fig. 3) and expressed as Datalog facts in BK.

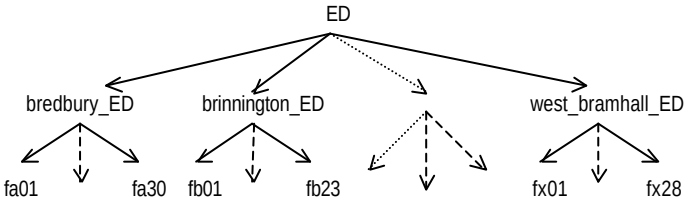


Fig. 3. An is-a hierarchy for the Stockport ED layer

Spatial association rules should relate EDs crossed by the M63 (reference objects) to EDs in the area served by the M63 (task relevant objects). The relations of intersection (EDs-motorways) and adjacency (EDs-EDs) have been extracted for the area of interest and transformed into Datalog facts of $D(S)$. The following census attributes have selected for this experiment:

- *s820161*, persons who work outside the district of usual residence and drive to work;
- *s820213*, employees and self-employed workers who reside in households with 3 or more cars and drive to work;
- *s820221*, employees and self-employed workers who reside in households with 3 or more cars and work outside the district of usual residence.

Since they refer to residents aged 16 and over, they have been normalized with respect to the total number of residents aged 16 and over (s820001). Moreover, they have been discretized by RUDE, since they are all numeric (more precisely, integer valued). At the end of this transformation process, each ED is described by three ground atoms in $D(S)$, namely *dr_out*(X, [a..b]), *cars3_dr*(X, [a..b]), *cars3_out*(X, [a..b]), where X denotes an ED, while [a..b] is one of the intervals returned by RUDE.

The key atom defining the reference objects in S is *ed_on_M63*(X), which is intensionally defined in the BK by means of the following rule:

```
ed_on_M63(X) :- intersect(X, m63).
```

The BK also includes the declarative specification of some rules for spatial qualitative reasoning, namely

```
can_reach(X, Y) :- intersect(X, m63), intersect(Y, m63), Y\=X.
close_to(X, Y) :- adjacent_to(X, Z), adjacent_to(Z, Y), Y\=X.
```

Finally, the following thresholds for support and confidence were defined: *min_sup*[1]=0.7 and *min_conf*[1]=0.9 at the first level, and *min_sup*[2]=0.5 and *min_conf*[2]=0.8 at the second level.

SPADA was run on the $D(S)$ obtained. The runtime was 331 secs for association rules at granularity level 1, and 310 secs for level 2 (data refers to a PC Pentium III 1GHz with 256 Mb RAM).

Initially, the system returned 12,925 frequent patterns out of 74,338 candidate patterns, for a total of 12,466 strong rules. By analyzing them we observed that some were actually useless, since they did not relate spatial data to census data. In other words, some association rules were pure spatial patterns, such as the following:

ed_on_M63(X) , can_reach(X,Y) → *is_a(Y,ward_on_m63_ED)* (90.0%, 100.0%)

which states that if an ED (*Y*) in the area served by the M63 can be reached from an ED crossed by the M63, then that ED is certainly (100% confidence) an ED of a ward crossed by the M63. Despite the high support and confidence, this pure spatial pattern is of no interest for transport planners.

In a second run, we decided to declare a bias for patterns containing at least one of the census attributes *dr_out*(X, [a..b]), *cars3_dr*(X, [a..b]) and *cars3_out*(X, [a..b]). The system generated 10,513 strong association rules in 1520 secs (time increased because of constraint checking for each generated pattern). Some of them have a very high support and confidence and provide the expert with some hints on the habits of commuters, such as the following association rule discovered at level 2:

(rule 27) *ed_on_M63(X), close_to(X,Y), is_a(Y,Bedgeley_ED)* →
cars3_out(X,[0.0..0.037]) , cars3_dr(X,[0.0..0.037]) *(100%,100%)*

which states that "if an ED crossed by the M63 (*X*) is close to another ED of the ward of Bedgeley (*Y*), then in that ED the percentage of people living in households with 3 or more cars and going/driving out of the district to work is very low (less than 4%)". It is important to point out that this is simply an association and does not define any kind of cause-effect relationship between the place where people live and their social habits. Another interesting spatial association rule at the same granularity level is the following:

(rule 177) *ed_on_M63(X), can_reach(X,Y)* → *is_a(Y,heaton_mersey_ED) ,*
dr_out(Y,[0.2857..0.4782]) , cars3_out(Y,[0.0..0,037]) *(80.0%,88.88%)*

which states that "if an ED *Y* in the M63 area can be reached from another one crossed by the M63 motorway (*X*), then it is in the Heaton Mersey ward and has quite a high percentage of people that drive to work but don't live in households with 3 ore more cars".

Finally, we decided to constrain the search space further, by asking only for those spatial patterns involving EDs where people have the same commuting habits. This time SPADA found only 345 strong rules (79 for level 1 and 266 for level 2) in about 833 secs. The following is an example of association found by the system at the granularity level 2:

(rule 76) *ed_on_M63(A)* → *can_reach(A,B), is_a(B,cheadle_ED), can_reach(A,C),*
C ¹ B, is_a(C,edgeley_ED), cars3_dr(C,[0.0..0.037]), cars3_dr(B,[0.0..0.037])
(90%, 90%)

which states that from an ED crossed by the M63 it is possible to reach (by the same motorway) two EDs, one in Cheadle and one in Edgley, with the same low percentage of people living in families with three or more cars and driving out of the district to work.

3.3 Accessibility of the Stepping Hill Hospital

Another problem concerning transport planning is the accessibility of the Stepping Hill Hospital in Stockport. To study this problem we decided to mine association rules relating five EDs close to the Stepping Hill Hospital (*task relevant* objects) with EDs

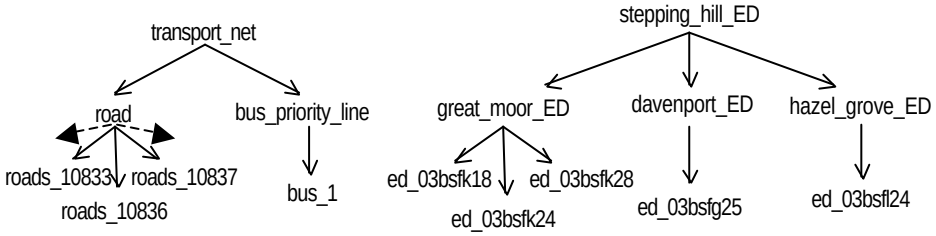


Fig. 4. Two spatial hierarchies defined for the mining task concerning the accessibility of the Stepping Hill Hospital.

within a distance of 10 Km from the hospital (reference objects). The goal is that of understanding which reference EDs have direct access to the task relevant EDs. To define the accessibility we used the Ordnance Survey data on transport network (roads and bus priority line). In the domain knowledge we defined a predicate *can_reach(X,Y)* stating that ED Y can be reached from ED X if one of the two following conditions hold:

1. Both are crossed by the same road or bus priority line;
2. From X it is possible to reach Z and from Z it is possible to reach Y (transitivity property)

This is the only spatial relation used in the spatial association rules. Our observation is that the accessibility of an area cannot be defined on the basis of the transport network alone. Even though some roads connect a reference ED X with a task relevant ED Y, people leaving in X might have problems to reach Y because they do not drive. This means that sociological data available in the census data tables can be profitably used to give an improved definition of accessibility. We selected four attributes on the percentage of households with zero, one, two, and three or more cars, we discretized them with RUDE and generated the following four binary predicates for SPADA: *no_car*, *one_car*, *two_cars*, *three_more_cars*. The first argument of the predicate refers to an ED, while the second argument is an interval returned by RUDE.

In this task we have two spatial hierarchies mapped into three granularity levels (Fig. 4). The declarative bias requires that the spatial association rules contain at least one of the four predicates above. SPADA generated 63 rules in 12 secs. Two of the rules returned by SPADA are the following:

ed_around_stepping_hill(A), can_reach(A,B), is_a(B,stepping_hill_ED) → *two_cars(A,[9.0e-003..0.179])*

(11.84%, 66.66%)

ed_around_stepping_hill(A), can_reach(A,B), is_a(B,stepping_hill_ED) ® *no_car(A,[0.266..0.653])*

(13.15%, 74.07 %)

They state that if from an ED it is possible to reach the area of the Stepping Hill Hospital, then the percentage of households without car can be between 26.6% and 65.3% while the percentage of households with two cars is between 9% and 17.9%. These association rules are interesting for urban planners, since they relate data on the transport network with data on sociological factors. In the future work, this task will be more deeply investigated.

4 Conclusions

In the above application we have seen that some of the discovered rules actually convey new knowledge, however the search for these "nuggets" requires a lot of tuning and efforts by the data analyst in order to constrain the search space properly and discard most of the obvious or totally useless patterns hidden in the data. This is typical of exploratory data analysis, and SPADA can be considered one of the most advanced tools that data analysts currently use in their iterative knowledge discovery process.

One of the main limitations of SPADA, which is also a problem of many other relational data mining algorithms, is the requirement of some expertise in data and knowledge engineering. Indeed, the user should know how data are organized in the spatial database (e.g., layers and physical representation of objects), the semantics of spatial relations that can be extracted from digital maps, the meaning of some parameters used in the discretization process and in the generation of spatial association rules, as well as the correct and most efficient way to specify the domain knowledge and declarative bias. Finally, in future work, we will investigate some "interestingess measures" of rules for presentation purposes, so that the user can browse the output XML file of spatial association rules as simply as possible.

Acknowledgments

The authors thank Jim Petch, Keith Cole and Mohammed Islam (MIMAS, University of Manchester, England) and Chrissie Gibson (Department of Environmental and Geographical Sciences, Manchester Metropolitan University, England) for providing access to census data and digital OS maps of Stockport, Manchester. The work presented in this paper is in partial fulfillment of the research objectives set by the IST European project SPIN! (Spatial Mining for Data of Public Interest) and by the MURST COFIN-2001 project on "Methods for the extraction, validation and representation of statistical information in a decision context". Thanks to Lynn Rudd for her help in reading the paper.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of the Twentieth VLDB Conference, Santiago, Chile (1994)
2. Ceri, S., Gottlob, G., Tanca, L.: What you Always Wanted to Know About Datalog (And Never Dared to Ask). IEEE Transactions on Knowledge and Data Engineering 1,1, (1989) 146-166
3. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations. In: Lavrac, N., Dzeroski, S. (eds.): Inductive Logic Programming. LNCS 1297, Springer-Verlag, Berlin

(1997) 125-132
4. Dehaspe, L., Toivonen, H.: Discovery of frequent Datalog patterns. Data Mining and Knowledge Discovery 3(1) (1999) 7-36
5. De Raedt, L.: Interactive Theory Revision. Academic Press, London (1992)
6. Dzeroski, S., Lavrac, N. (eds.): Relational Data Mining, Springer-Verlag, Berlin (2001)
7. Egenhofer, M.J, Herring, J.R.: Categorizing binary topological relations between regions, lines, and points in geographic databases. In M.J. Egenhofer, D.M. Mark, and J.R. Herring (eds.): The 9-intersection: Formalism and ist Use for Natural-language Spatial Predicates, pages (1994) 183-271
8. Güting, R.H.: An introduction to spatial database systems. VLDB Journal, 3,4 (1994) 357-399.
9. Han, J., Fu, Y.: Discovery of multiple-level association rules from large databases. In U. Dayal, P.M.D. Gray, S. Nishio (eds.): VLDB'95, Proceedings of the 21st International Conference on Very Large Data Bases, Morgan-Kaufmann (1995) 420-431.
10. Han, J., Koperski, K., Stefanovic, N.: GeoMiner: A System Prototype for Spatial Data Mining. In Peckham, J. (ed.): SIGMOD 1997, Proceedings of the ACM-SIGMOD International Conference on Management of Data. SIGMOD Record 26, 2 (1997) 553-556.
11. Helft, N.: Inductive generalization: a logical framework. In: Bratko, I., Lavrac, N. (eds): Progress in Machine Learning. Sigma Press (1987) 149-157
12. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Egenhofer, M.J., Herring, J.R. (Eds.): Advances in Spatial Databases. LNCS 951, Springer-Verlag, Berlin (1995) 47-66.
13. Koperski, K., Adhikary, J., Han, J.: Spatial Data Mining: Progress and Challenges. In Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada (1996)
14. Lavrac, N., Džeroski, S.: Inductive Logic Programming: techniques and applications. Ellis Horwood, Chichester (1994)
15. Ludl, M.-C., Widmer, G.: Relative Unsupervised Discretization for Association Rule Mining. In D.A. Zighed, H.J. Komorowski, J.M. Zytkow (Eds.): Principles of Data Mining and Knowledge Discovery, LNCS 1910, Springer-Verlag (2000) 148-158.
16. Malerba, D., Esposito, F., Lanza, A., Lisi, F.A.: Machine learning for information extraction from topographic maps. In H. J. Miller & J. Han (Eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, London, UK, (2001) 291-314
17. Malerba, D., Lisi, F.A.: An ILP method for spatial association rule mining. Working notes of the First Workshop on Multi-Relational Data Mining, Freiburg, Germany (2001) 18-29.
18. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3) (1997) 259-289
19. Muggleton, S. (ed): Inductive Logic Programming, Academic Press, London (1992)
20. Nienhuys-Cheng, S.-H., deWolf, R.: Foundations of inductive logic programming. Springer, Heidelberg, Germany (1997)
21. Plotkin, G.: A note on inductive generalization. Machine Intelligence, 5 (1970) 153-163
22. Saporta, G.: Data Mining and Official Statistics. Atti della Quinta Conferenza Nazionale di Statistica, Rome (2000) 15-17

Bayesian Regression Mixtures of Experts for Geo-Referenced Data

Gerhard Paaß and Jörg Kindermann

Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
53754 St. Augustin, Germany
{Paass, Kindermann}@ais.fraunhofer.de

Abstract. Politicians, planners and social scientists have an increasing need for tools clarifying the spatial distribution of relevant features. Special interest is in what-if analyses: what would happen if we change some features in a specific way. To predict future developments requires a statistical model with inherent modelling uncertainty. In this paper we investigate Bayesian models which on the one hand are able to represent complex relations between geo-referenced variables and on the other hand estimate the inherent uncertainty in predictions. For solution the models require Markov-Chain Monte Carlo techniques.

1 Introduction

Spatial interpolation and extrapolation is an essential feature of many Geographic Information Systems (GIS). It is a procedure for estimating values of a variable at un-sampled locations. Based on Tobler's Law of Geography, which stipulates that observations close together in space are more likely to be similar than those farther apart, these procedures try to separate spatial correlation from random noise. They can, however, be divergent and lead to very different results if the underlying structural assumptions are not fulfilled. As a consequence, an understanding of the initial assumptions and methods used is key to the spatial interpolation process.

Bayesian statistics offers a way to mitigate these problem. It describes the uncertainties inherent in a statistical analysis by means of probability distributions, which capture the degree of belief that a quantity is located in some interval. This applies to observable quantities like the variables of interest as well as to unobservable quantities as the parameters of models, and their structural properties. During the last decade a number of new computation strategies have been developed which allow the solution of large scale problems for very complex models by means of stochastic simulation.

In this paper we describe the Bayesian variant of a flexible semi-parametric model, a *mixture of experts*, which is able to represent a wide variety of complex dependencies. It is composed of a series of localized component models called *experts,* which cover local properties of the relation in question.

In the next chapter we will describe spatial data and their specific properties. In chapter three we shortly describe classical statistical inference procedures like

least squares and in chapter four its Bayesian counterparts. Chapter five compiles some ensemble methods which use collections of possible models to describe the inherent variability or to get better predictions by forming a committee. Chapter six describes the classical methods of spatial statistics, which mostly are derived from linear least squares approaches. The last chapter is central to the paper as it analyses different advanced nonlinear procedures and assesses their potential in the spatial domain, especially in a Bayesian framework.

2 Bayesian Statistics

2.1 Basic Setup

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model. In addition probability distributions on unobserved quantities such as predictions for new observations may be derived. Assume we have independent observations $(z_1, \mathbf{x}_1), \ldots, (z_n, \mathbf{x}_n)$ of the inputs $\mathbf{x}_i \in \Re^k$ and outputs $z_i \in \Re$. We may arrange the observed inputs in the matrix $\mathbf{X} = \mathbf{X}_{(n,k)} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and the outputs in a vector $\mathbf{z} = \mathbf{z}_{(n,1)} = (z_1, \ldots, z_n)'$. Bayesian inference assumes the existence of a joint distribution $p(\theta, \mathbf{z}, \mathbf{X})$. We are especially interested in the conditional distribution $p(\theta, \mathbf{z}|\mathbf{X}) = p(\theta|\mathbf{X})p(\mathbf{z}|\theta, \mathbf{X})$. Let the *prior distribution* $p(\theta) = p(\theta|\mathbf{X})$ describe the information about the parameter θ before the data $\mathbf{z}$ is available.

Then *Bayes' rule* yields the *posterior density*

$$p(\theta|\mathbf{z}, \mathbf{X}) = \frac{p(\theta, \mathbf{z}|\mathbf{X})}{p(\mathbf{z}|\mathbf{X})} = \frac{p(\theta|\mathbf{X})p(\mathbf{z}|\theta, \mathbf{X})}{p(\mathbf{z}|\mathbf{X})} = \frac{p(\theta)p(\mathbf{z}|\theta, \mathbf{X})}{\int p(\theta)p(\mathbf{z}|\theta, \mathbf{X})\, d\theta} \tag{1}$$

which describes the distribution of parameters after $\mathbf{X}$ and $\mathbf{z}$ have been observed. To make predictive inferences about an unknown observable and a new input $\mathbf{x}_0$ we calculate the prediction $p(z|\mathbf{x}_0, \theta)$ for each θ and $\mathbf{x}_0$ and weight them according to the posterior $p(\theta|\mathbf{z}, \mathbf{X})$ of parameters

$$p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) = \int p(z|\mathbf{x}_0, \theta)p(\theta|\mathbf{z}, \mathbf{X})d\theta \tag{2}$$

This gives us the complete distribution of z for a new input $\mathbf{x}_0$ in the light of the data $\mathbf{z}, \mathbf{X}$. We can evaluate any characteristics of this distribution, for instance its expected value $E(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X})$, or a *highest posterior density region* which is the smallest region covering the output with a prescribed probability, e.g. 90%. It may – of course – no longer be contiguous but consist of a set of contiguous subset.

2.2 Prediction and Markov Chain Monte Carlo

The predictive distribution for the output of interest conditional on the new input $\mathbf{x}_0$ and the observed data $\mathbf{z}, \mathbf{X}$ was $p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) = \int p(z|\mathbf{x}_0, \theta)p(\theta|\mathbf{z}, \mathbf{X})d\theta$.

We may approximate the integral by a sum

$$p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) \approx \frac{1}{N} \sum_{j=1}^{N} p(z|\mathbf{x}_0, \theta_j) \qquad \theta_j \sim p(\theta|\mathbf{z}, \mathbf{X}) \tag{3}$$

If the θ_j are independently generated according to the posterior then the sum converges to the desired density by the law of large numbers. Subsequently we may describe $p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X})$ by different features, e.g. expectation, variance or posterior intervals.

The *Metropolis-Hastings algorithm* allows to generate a sample of parameter values θ_j distributed according to the posterior density. This involves the construction of a Markov chain $\theta(0), \theta(1), \ldots$ designed to be distributed according to the posterior density $p(\theta|\mathbf{z}, \mathbf{X})$. If the chain is currently at $\theta = \theta(t)$, the *Metropolis-Hastings algorithm* [Tie94] requires a *proposal density* $\boldsymbol{q}(\theta, \tilde{\theta})$, which is the conditional distribution of proposing a move from θ to $\tilde{\theta}$. The *acceptance probability* is defined as

$$p_{\mathrm{acc}}(\theta, \tilde{\theta}) = \min \left\{ 1, \frac{p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, \boldsymbol{q}(\tilde{\theta}, \theta)}{p(\theta|\mathbf{z}, \mathbf{X})\, \boldsymbol{q}(\theta, \tilde{\theta})} \right\} \tag{4}$$

With probability $p_{\mathrm{acc}}(\theta, \tilde{\theta})$ the candidate $\tilde{\theta}$ is accepted and the chain moves to $\theta(t+1) = \tilde{\theta}$. Otherwise the candidate is rejected and $\theta(t+1)$ takes the old value θ. For the actual transition probability $\boldsymbol{p}(\theta, \tilde{\theta}) := \boldsymbol{q}(\theta, \tilde{\theta}) p_{\mathrm{acc}}(\theta, \tilde{\theta})$ the *detailed balance* condition holds for all $\theta, \tilde{\theta}$

$$p(\theta|\mathbf{z}, \mathbf{X})\, \boldsymbol{p}(\theta, \tilde{\theta}) = p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, \boldsymbol{p}(\tilde{\theta}, \theta) \tag{5}$$

If the resulting Markov chain is aperiodic and irreducible (i.e. reaches all states with positive probability) then its distribution converges to an invariant stationary limit distribution, which is just the posterior distribution $p(\theta|\mathbf{z}, \mathbf{X})$ [Tie94].

If we have several candidate models, where the number and the interpretation of parameters is different, the approach cannot be used. [Gre95] has proposed an MCMC-scheme for varying dimension problems, termed *reversible jump MCMC*. When the current state is θ and $p(\theta|\mathbf{zX})$ is the target probability measure (the posterior density) we consider a countable number of different moves m. Depending on the state θ a move m and a destination $\tilde{\theta}$ is proposed with $q_m(\theta, \tilde{\theta})$ as joint distribution. $q_m(\theta, \tilde{\theta})$ may be a sub-probability measure, with probability $1 - \sum_m \int_{\tilde{\theta}} q_m(\theta, \tilde{\theta}) d\tilde{\theta}$ no move is attempted.

For the case that θ and $\tilde{\theta}$ have the same dimension, the procedure reduces to the Metropolis-Hastings algorithm (4). Now suppose that starting from θ a move of type m is proposed that yields a higher-dimensional $\tilde{\theta}$. This can be implemented by drawing a vector $\mathbf{u}$ of continuous variables distributed according to a known density $p_m(\mathbf{u})$ independent of θ. It is required that the sum of the dimensions of θ and $\mathbf{u}$ is equal to the dimension of $\tilde{\theta}$. Then the new state $\tilde{\theta}$ is defined by an invertible deterministic function $\tilde{\theta} = h_m(\theta, \mathbf{u})$. The reverse of

the move can be accomplished by using the inverse transformation, so that the proposal is deterministic. Then we get the acceptance probability

$$p_{\mathrm{acc}m}(\theta, \tilde{\theta}) = \min\left(1, \left|\frac{\partial h_m(\mathbf{u}, \theta)}{\partial(\mathbf{u}, \theta)}\right| * \frac{p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, j_m(\tilde{\theta})}{p(\theta|\mathbf{z}, \mathbf{X})\, j_m(\theta)\, p_m(\mathbf{u})}\right) \tag{6}$$

Here $j_m(\theta)$ and $j_m(\tilde{\theta})$ are the probabilities of selecting move m or its inverse in states θ and $\tilde{\theta}$ respectively. [Gre95] shows that the detailed balance condition 5 holds and consequently the equilibrium distribution of the resulting Markov chain is the posterior distribution $p(\theta|\mathbf{z}, \mathbf{X})$. Similar to the usual Metropolis-Hastings formula 4 the densities have to be known only up to a factor, which cancels out in 6.

The reversible jump algorithm is a major improvement in the Markov Chain Monte Carlo approach. It allows to explore complete model classes instead of a single model with a given structure. Note, however, that for the different classes prior probabilities are required.

Instead of specifying all priors explicitly we may use mixtures between priors of different shapes, so called hierarchical models [GCSR95, p.119], to introduce the prior information in a less restrictive way. The final weighting of different priors then is determined by the data.

3 Mixtures of Experts

Modular and hierarchical systems allow complex learning problems to be solved by dividing the problem into a set of subproblems, each of which may be simpler to solve than the original problem. In spatial statistics it is natural to assume that the data can be well described by a collection of functions, each of which is defined over a relatively local region of the input space. A *modular architecture* can model such data by allocating different modules to different regions of the space. *Hierarchical architectures* arise when we assume that the data are well described by a multi-resolution model – a model in which regions are recursively divided into subregions. An example is the decision tree model.

The learning algorithm simultaneously has to determine a partition of the input space into regions as well as the local models (experts) within each region. The *mixture of experts* approach developed by [JJNH91] uses different sub-models for partitioning (*gating models*) the input space as well as local prediction (*expert models*). In contrast to the decision tree the regions are not disjoint but there is a gradual change between regions. For each input point the predictions of the different experts are computed and used with weights determined by the gating network.

If we have m expert networks $z = f_j(\mathbf{x}, \theta_j)$, $j = 1, \ldots, m$, we need a gating network $g(\mathbf{x}, \phi)$ with one output $w_j = g_j(\mathbf{x}, \phi)$ for each expert network. To arrive at normalized weights these outputs are transformed by the 'softmax' function

$$\alpha_j(\mathbf{x}, \phi) = \frac{\exp(g_j(\mathbf{x}, \phi))}{\sum_{l=1}^{m} \exp(g_l(\mathbf{x}, \phi))} \tag{7}$$

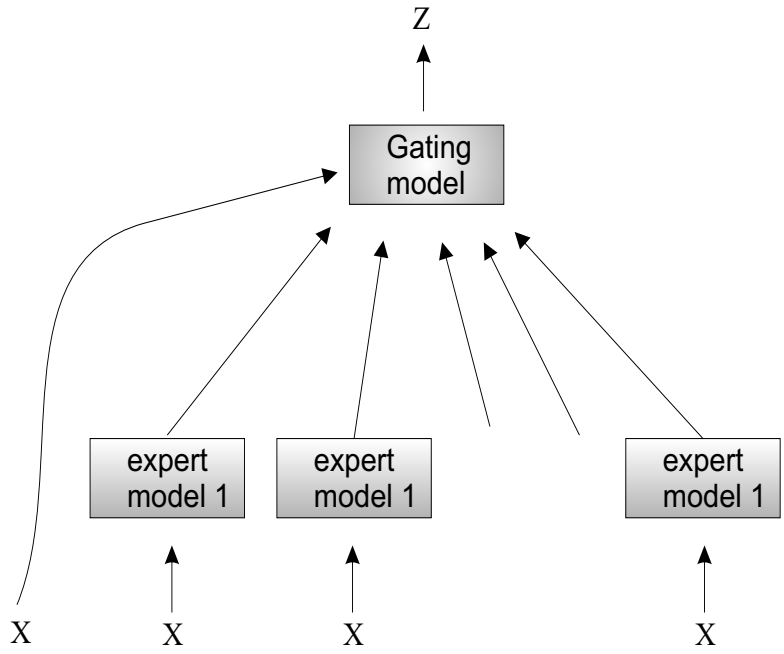


Fig. 1. In a mixture of experts a gating model defines probabilities for the different experts. The outputs of the expert models are weighted by these probabilities.

and the final output is the mixture of experts

$$z = \sum_{j=1}^{m} \alpha_j(\mathbf{x}, \phi)\, f_j(\mathbf{x}, \theta_j) + \varepsilon_j \qquad E(\varepsilon_j) = 0 \tag{8}$$

We may use virtually any model as expert model as long as it fits to the data $(z, \mathbf{x})$. Note that for Bayesian analysis a complete specification of the related distributions is required.

As an arbitrary number of experts may be combined we may use computationally simple models, whose combination may represent arbitrary complex dependencies. Candidates for continuous $z \in \Re$ are

- constants $z = c_j$. The gating network generates convex combinations of these constants.
- linear regression models $z = \sum_{i=1}^{k} x_i \theta_i + \varepsilon$ with normal error $\varepsilon \sim N(0, \sigma^2)$.
- quadratic or nonlinear regression models $z = \sum_{j=1}^{m} h_j(\theta_i) + \varepsilon$ with normal error $\varepsilon \sim N(0, \sigma^2)$ and fixed basis functions.
- Arbitrary generalized linear models [JPT97].

For discrete $z \in \{1, \ldots, r\}$ we may use any Bayesian classifier, and – in combination with the softmax function – arbitrary models with values in $\Re$. Simple examples are

- linear logistic model $f_j(\mathbf{x}, \theta) = \exp(\mathbf{x}'\theta_j) / \sum_{l=1}^{m} \exp(\mathbf{x}'\theta_l)$
- radial basis function models $f_j(\mathbf{x}, \theta) = h_j(\mathbf{x}, \theta_j, \sigma_j) / \sum_{l=1}^{m} h_l(\mathbf{x}, \theta_l, \sigma_l)$ with $h_j(\mathbf{x}, \theta_s, \sigma_s) = \prod_{j=1}^{k} \left(2\pi\sigma_j^2\right) \exp\left(-\frac{1}{2\sigma_j^2}\left(x_j - \theta_{sj}\right)^2\right)$

As gating network we may select any models $g_j(\mathbf{x}, \phi)$ with outputs in $\Re$ or any "probability model" $\alpha(\mathbf{x}, \phi)$ which generates a probability vector with m components, i.e. classifier models.

3.1 Prior Distributions

The choice of priors for a model is an important one in Bayesian inference. Priors embody the assumption about such aspects as the generative processes of the data and form of the model. The priors on a model are typically placed either on the structure (number of models) or the parameters of gate and expert models. The parameters of the gate and expert models are assumed to be mutually independent $p(\theta, \phi) = p(\theta)p(\phi)$. They may depend themselves on hyper-parameters, which themselves may be varied during the MCMC analysis. For the mean of radial basis functions as well as the means of regression models we use Normal priors with diagonal covariance matrix

$$p(\theta_s) = \prod_{j=1}^{k} \left(2\pi\rho_j^2\right) \exp\left(-\frac{1}{2\rho_j^2}\left(\theta_{sl} - \bar{\theta}_{sl}\right)^2\right) \tag{9}$$

For the variance σ^2 we use a Gamma prior on the inverse variance $\beta = 1/\sigma^2$

$$p(\log\beta) = \frac{1}{\Gamma(\tau)} \left(\frac{\beta}{v}\right)^{\tau} \exp\left(-\frac{\beta}{v}\right) \tag{10}$$

were $2/v$defines the prior sum of squared error that we might expect and 2τ defines the prior number of observations that we might expect an expert to see.

3.2 Comparison to other Models

It is instructive to visualize the regions defined by different types of experts. As shown in figure 2 logistic units $\exp(\mathbf{x}'\theta_j)$ put a "soft" threshold into the input space where they change their value from 0 to 1. Combined with the softmax function this results in mainly straight boundaries that partition the input space. It is important that each unit affects the whole partition. On the other hand radial basis function units $h_j(\mathbf{x}, \theta_j, \sigma_j)$ assign the region around the mean value θ_j to the corresponding unit. This leads to a Voronoi tesselation of the input space with linear boundaries between units, as long as the covariance terms for all units are identical.

Earlier Mixture of experts approaches therefore used logistic gating models but in a hierarchical fashion [JPT97], [Wat97]. In the bhighest layer two regions were defined, which were recursively partitioned by other gates of lower

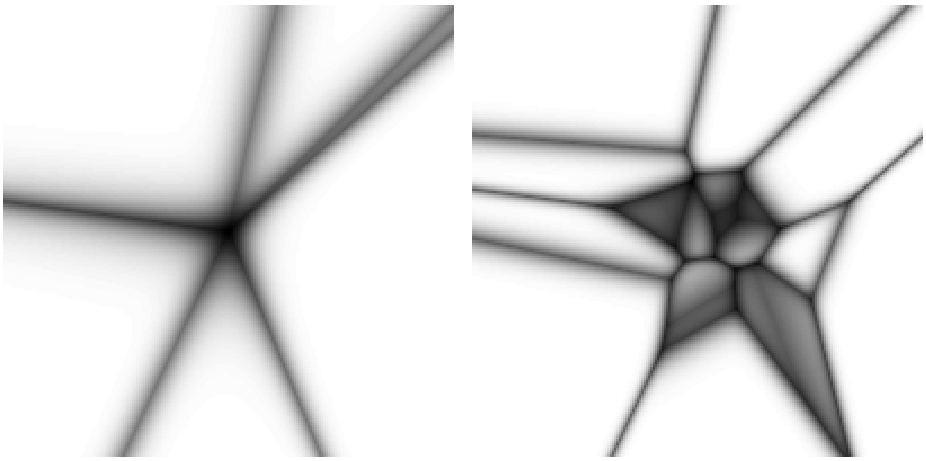


Fig. 2. Maximal probability of membership in a single region if the gates are logistic (left) or radial basis functions (right).

regions. For the Bayesian analysis these hierarchical mixtures of experts have a definite disadvantage: it is nearly impossible to change high-level gates in a MCMC analysis as this means that the whole tree of gates has to be deleted and rebuilt. Therefore Bayesian analyses tend to concentrate in a local minimum of the posterior density.

If we use non-hierarchical radial basis functions gates the changes only affect neighboring points. The MCMC algorithm can generate all plausible structures and effectively explore the posterior density. Therefore we prefer radial basis function units in our analysis.

There are a number of advanced statistical methods which may be applied to spatial problems in a similar way like the mixture of experts. They may be used in a semi-parametric fashion, i.e. they should be able to fit a wide set of functional relations in a nearly automatic way. They all can be evaluated in the framework of Bayesian statistics. This allows the flexible introduction of prior knowledge and the calculation of the uncertainty of statistical inference. Generalized additive models [VR97, p.281] and projection pursuit regression [FS82] define models on marginal variables and therefore are not able to fit arbitrary distributions. Local regression models are [CGJ95] are an attractive competitor of mixture of experts. Neural networks in the form of multilayer perceptrons [Nea95] also use logistic units and have the same convergence problems as hierarchical mixtures of experts. Similar problems occur for decision trees [PK98][CGM98] and multivariate adaptive regression splines (MARS) [Fri91][DMS97] which generate recursive partitions of the input space.

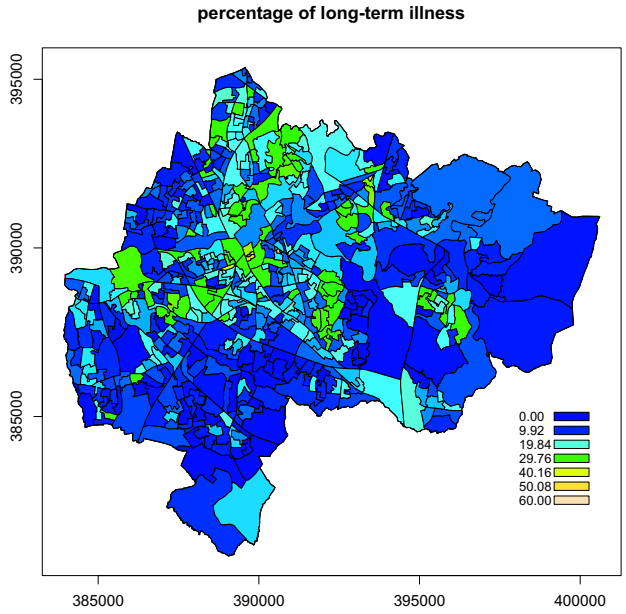


Fig. 3. Mean predicted value of long-term illness in Stockport using the Bayesian model.

4 Markov Chain Monte Carlo

The Markov Chain Monte Carlo analysis uses the following proposals to modify a model:

- Change the mean values of one gate unit.
- Change the variance of one variable for all gate units.
- Change the regression parameters of one expert model.
- Change the error variance of one expert model (not for classifier experts).
- Split one expert model into two (with different parameters).
- Merge two randomly selected expert models into one, whose parameters are the mean values of the components.

After an initial phase of several thousand iterations the MCMC algorithm reaches the stationary distribution of mixture of expert models. After this burn-in phase the models with all their parameters are stored for later use. We use the coda-package of R to determine the convergence to stationarity [BR98].

5 Application to Geodata

The mixture of experts model was implemented in the SPIN! system developed during the SPIN! project of the European community. It is a general tool for

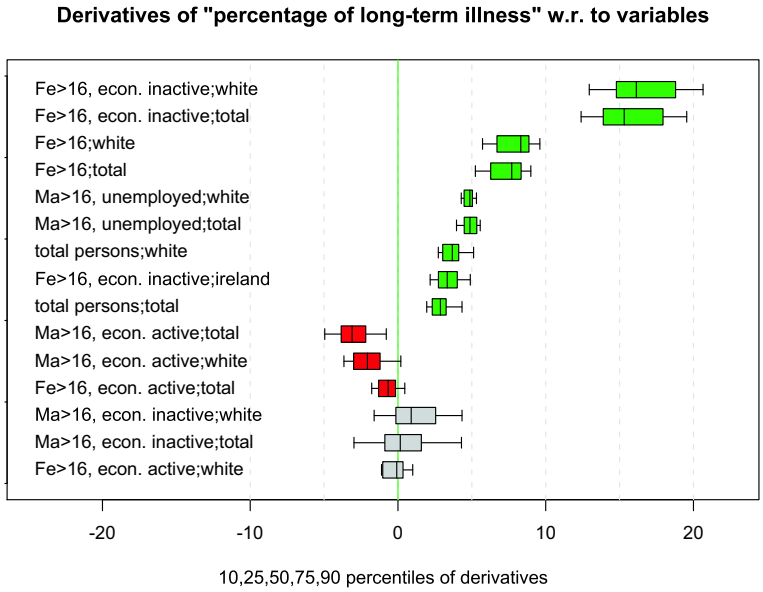


Fig. 4. Distributions of derivatives of long-term illness in Stockport for a specific ward. The boxes indicate 25% and 75% percentiles with the median in between. The outside "whiskers" are the 10% and 90% percentiles.

simulation Bayesian models by Markov Chain Monte Carlo. The system is implemented in Java to avoid compatibility problems.

As an introductory example we use data from Stockport, a town near Manchester, U.K. For small units of about 100 households (wards) we have statistics from the 1991 census including the basic demographic features as well as employment, car ownership, etc.

The Bayesian model was used to predict long-term illness from these figures. Hence the model is forced to adapt to the relation between the values of the input variables within the individual wards and the corresponding output variable variable long-term illness. The derivative of the output variable with respect to an input variable describes, how many units the output variable probably will increase if we increase the input variable for one unit. As the model is non-linear, the derivative will depend on the specific location, i.e. the input variables of the ward.

This figure may be important for planners if the want to check the stochastic relation between variables. It does not, however, imply, that the input variable actually may be changed, as many variables may not be controlled.

As our Bayesian model explicitly captures uncertainties the derivative is uncertain too. In figure 4 the resulting distribution of derivatives for a specific ward is shown. The graph can be generated interactively by clicking on a ward in the map above. The derivatives show that long-term illness in wards like the current

ward usually grows with the fraction of females aged higher than 16 years, which are economically inactive. This probably mainly applies to female pensioners. On the other hand long-term illness decreases if the number of economically active men increases.

On the workshop we will apply the approach to other data of North-West England.

References

[BR98] SP. Brooks and G.O. Roberts. Assessing the convergence of markov chain monte carlo algorithms. *Statistics and Computing*, pages 319–335, 1998.

[CGJ95] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. In Tesauro et al. [TTL95], pages 705–712.

[CGM98] H. Chipman, E. George, and R. McCulloch. Bayesian CART model search. *JASA*, 93:935–960, 1998.

[DMS97] D. Denison, B. Mallick, and A. F. M. Smith. Bayesian mars. Technical report, Imperiag College, London, 1997.

[Fri91] Jerome H. Friedman. Adaptive spline networks. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, volume 3, pages 675–683. Morgan Kaufmann Publishers, Inc., 1991.

[FS82] J. H. Friedman and W. Stuetzle. Projection pursuit methods for data analysis. *Modern Data Analysis*, pages 123–147, 1982.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[Gre95] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–713, 1995.

[JJNH91] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[JPT97] R. A. Jacobs, F. C. Peng, and M. A. Tanner. A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, 10(2):231–241, 1997.

[Nea95] R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, Dep. of Computer Science, Univ. of Toronto, 1995.

[PK98] G. Paass and J. Kindermann. Bayesian classification trees with overlapping leaves applied to credit scoring. In X. Wu, R. Kotagiri, and K. B. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining*, pages 234–245. Springer Verlag, 1998.

[Tie94] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[TTL95] G. Tesauro, D. Touretzky, and T. Leen, editors. *Advances in Neural Information Processing Systems 7*. The MIT Press, 1995.

[VR97] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer Verlag, New York, 3rd edition, 1997.

[Wat97] S. R Waterhouse. *Classification and Regression using Mixtures of Experts*. PhD thesis, Cambridge University Engineering Dept., October 1997.

Hinterlands delimitation of *Lisboa e Vale do Tejo* cities

Duarte Rodrigues
duarte.rodrigues@ine.pt

Francisco Vala
francisco.salvador@ine.pt

José Monteiro

jose.monteiro@ine.pt

Serviço de Estudos
Direcção Regional de Lisboa e Vale do Tejo
Instituto Nacional de Estatística
Av. António José de Almeida nº 2 – 1000-043 Lisboa
Telef: +351 21 842 61 00 Fax: +351 21 842 63 65

Abstract

For a long time that regional and urban science has interpreted territory through a systemic analysis where cities network are the fundamentals of spatial organization. Differentiation established between cities results from distinct levels of centrality, which allows setting an urban hierarchy.

The present discussion aims to study *Lisboa e Vale do Tejo* urban structure through cities rank position as well as defining theirs hinterlands and understanding the interactions between them.

The empirical application of this study is supported by information about services provided to the population that was compiled in Inventário Municipal (INE, 1998). City delimitation at level of *freguesia* (minimum territorial unit of *Inventário Municipal* information) was a necessary previous step for this analysis.

1 Theoretical Framework

The main purpose of present work is to study the system of *Lisboa e Vale do Tejo* (LVT) cities, through definition of cities hierarchy and analysis of flows between them. The Central Place Theory (CPT), developed by Christaller (1933) and Lösch (1940)[1], provides the theoretical fundamentals to this analysis.

In Portugal, there are few works on this area, which can be explained by the lack of information sources that allow empirical applications[2]. The delimitation of Évora influence area by Gaspar (Gaspar, 1981) represents one of the most important investigation in this area.

Before a brief presentation of Central Place Theory it's necessary to explain the meaning of following terms, that will be used during this work:

[1] Lopes (Lopes, 1987) and Alves et al (Alves et al, 1999) present a detailed version of this theory.

[2] Gaspar (Gaspar, 1981) presents a survey of international studies on Central Place Theory applications.

Central function - activity that provides goods or services, located on central position within its market area (e.g. hospital, driving school, video club)[3]. As more specialised and rare is the function, more central it will be.

As explained by Polèse (Polèse, 1998) the more specialised functions, located on the top of functions hierarchy, have the following features:

- Important scale economies that implies high minimum dimension of demand;
- Low frequency of consumption and, as a consequence, low transport costs associated with consumers movements;
- Its consumption implies larger movements by the population.

Functional Unit – each unit that provides a central function. Several functional units can provide the same central function.

Central Place – urban place that provides central functions for its peripheral region, that represents its **area of influence** or its hinterland.

Centrality – represents the level of central functions supplied by a certain urban place.

Area of influence (hinterland) of central function (for a certain urban place) – geometric place where the consumers of central function are.

The central place theory was developed with the purpose to explain why cities arise and was centred on the study of economic activity location, more specifically the activities of services sector. This theory tries to explain the size and the spatial distribution of urban places and also the relation between them.

According to this theory the centrality of an urban place is proportional to the specialisation of functions that are supplied and, as a consequence, is also proportional to the dimension of its area of influence. The more central urban places area, more population they have.

The centrality of urban place depends on the level of specialisation of the functions that it provides. As a result of this direct relation between hierarchy of functions and hierarchy of urban places, the relation among urban places strictly happens in a hierarchic way. The flows between urban places only happen in a vertical/upward way.

2 The Lisboa e Vale do Tejo cities

City delimitation is part of a greater INE project – "Urban Statistics" – which has dissemination of statistical information at city level as the main goal. This new statistical spatial unit - the city – will be the support of compilation and dissemination of statistical information in the near future.

This study only considers the agglomerations that were legally created as cities under the terms of Law nº11/82, second of June. This law defines "the rules for creation and extinction of local governments, and designation and determination of settlements category". On its 13rd article establishes that: "a village can only get a city title when has 8000 inhabitants living on a continuous built area, and with, at least,

[3] At this work we will make no difference between central function and central good or service, for instance between hospital (the function) and several specialities provided by it (the central services). This option is due to the limitation of information that will be analysed.

half of the following public equipment: a) hospital with 24 hours service; b) pharmacy; c) fireman corporation; d) theatre and cultural centre; e) museum and library; f) Tourist accommodation; g) Preparatory and secondary Schools; h) Kindergartens; urban and suburban public transports; Public gardens."

In spite the dimension and functional criteria, article 14th defines a wide exception possibilities: "important historical, cultural and architectonic may justify a different weight of the eligibility criteria's".

However, neither this law, nor recent laws which establish a specific city elevation, present precise spatial limits of cities.

The *Lisboa e Vale do Tejo* Region accounts a total of 30 cities, from which 16 are located on Lisbon Metropolitan Area (LMA).

City delimitation was based on three main principles:

1 – City delimitation should be taken with the smallest territory level – *subsecção estatística* (a block of houses in urban areas);

2 – Local Governments should be involved on delimitation process, as they are the most important space actors on space management and transformation;

3 – Delimited cities should represent the city at the present time.

The delimitation criteria followed was:

- Morphological criteria – analysis of population and dwellings density at *subsecção* territorial desegregation and analysis of topographic and aerophotography[4];
- Planning criteria – analysis of local governments physical planning instruments, specially Urban, Urban Growth, Existent Industry, Projected Industry, Projected Equipment, Urban Green Area[5] and urban perimeter[6];
- Functional. Employment areas, specially industrial areas, as these areas tend to assume peripheral locations and may represent an important part of city labour market.

Due to diversity of city appearance at a national level (from historical cities to those that result of post-suburbanization evolutions), local authorities end up playing a major role on delimitation process, not only by the legal planning instruments provided, but also by meetings with planning technicians.

Globally city limits are a compromise between local government definitions and *subsecções* territorial division.

Inventário Municipal information structure led to city definition at level of *freguesia*[7].

---

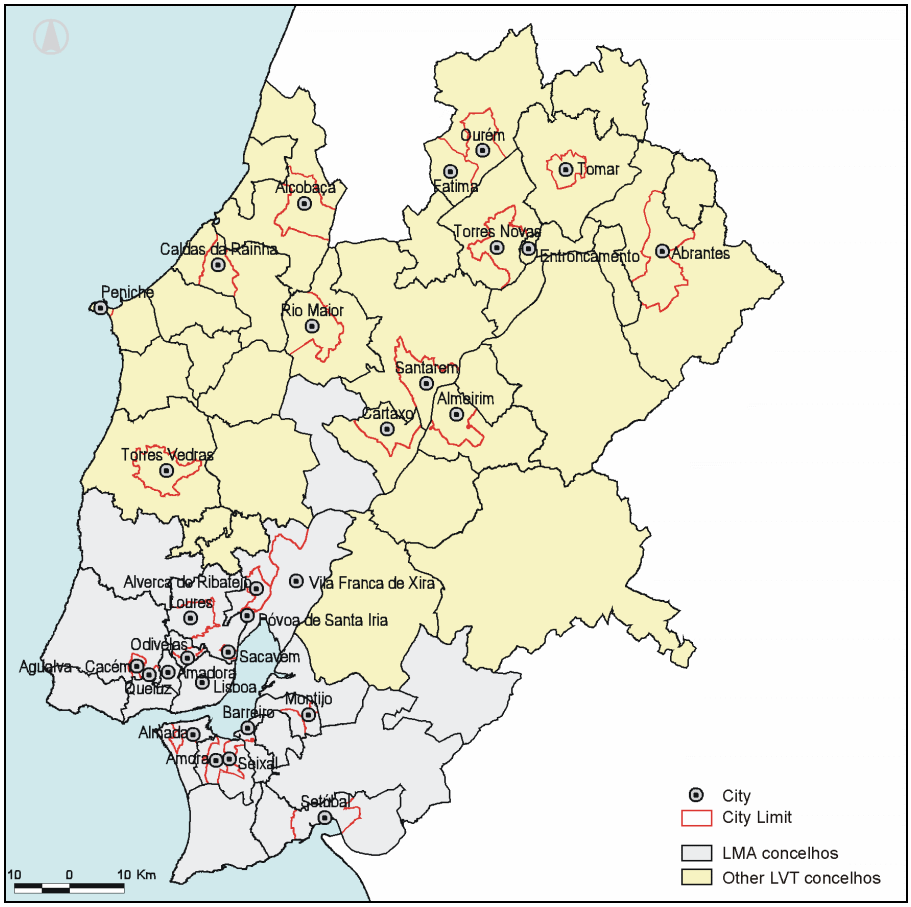
[4] Ortofocartografia from Intituto Português da Cartografia e Cadastro, série 1:10 000, 1998; Vector Cartography form Associação de Municípios do Oeste., 1:10 000, 2000; Vector Cartography from Instituto Geográfico do Exército, 1:25 000; other vector high scale Cartography from Câmaras Municipais.

[5] The Space Classes denomination presented corresponds to the Class/Category Type defined by Direcção Geral do Ordenamento do Território e Desenvolvimento Urbano (1998).

[6] The geographical information used for this analysis were digitised by Direcção Geral do Ordenamento do Território e Desenvolvimento Urbano and Local governments cartography from legal planning instruments.

[7] *Freguesia* represents the Portuguese NUTS V. NUTS IV – *Concelho* – is formed by a group of *freguesias.*

Figure 1 – The cities of Lisboa e Vale do Tejo



3 The functions hierarchy

The current empirical application is based on a set of 126 central functions that cover the following areas: trade and service, health, education and social security. For these functions we have information from Inventário Municipal[8], not only about availability of the functions, but also about where population goes to consume a specific function if it isn't available in the inquired *freguesia*.

So, the sources of information are the five Inventários Municipais 1998 for each Portuguese NUTS II (*Norte, Centro, Lisboa e Vale do Tejo, Alentejo* and *Algarve*). They have information about what are the functions provided by cities and what are the interactions among cities and among them and their hinterlands. In this survey is

[8] Inventário Municipal is a Portuguese survey about functions available in *freguesias* and is answered by the presidents of the *freguesias*.

asked the number of functional units available to provide each function and if function is unavailable it is asked where[9] population goes to consume.

The hierarchy of functions is supported by the diversity of goods and services provided by central functions. On the top of it, representing the most central functions, are placed the most specialised functions with lowest demand frequencies.

Hierarchy is based on the number of functional units for each central function, at a national level. This variable is used as a proxy of function rarity. Therefore the function *hospital* is in a higher position of the hierarchy than the function *video club*, because there are 89 functional units for the former and 1446 for the other.

Thus, in the first position of the central functions hierarchy is the *alcoholism treatment clinic* (42 functional units) and in the last position is the *café, bar tavern* (34606 functional units) (see appendix 1).

There is a need of grouping the 126 functions in a limited number of classes, because the hierarchy will be used for segmentation of forthcoming analysis output.

The construction of specialisation classes of functions was done according to the following steps:

i. Application of statistical method – natural breaks – to the variable number of functional units. This method identifies breakpoints between classes using a statistical formula (Jenk's optimization). Jenk's method is rather complex, but basically it minimizes the sum of the variance within each of the classes.

   This method was used in an iterative way increasing in each step the number of classes. The iteration was stopped when the good segmentation on the top of hierarchy was achieved (happened with 11 classes). This process allowed getting a first class with a small number of functions.

ii. Afterwards, a casuistic grouping of sequential classes was done. The goal of this grouping was to get a small number of classes with an equilibrated number of functions among them and fewer functions in the first and the last classes.

So the result is the following five classes of central functions hierarchy:

More specialisation ↑	**Class**	**Class name**	**Number of functions**	Original class(es)
	1	Highly specialised functions	12	1
	2	Specialised functions	33	2+3
	3	Medially specialised functions	42	4+5+6
	4	Lowly specialised functions	31	7+8
	5	Not specialised functions	8	9+10+11

[9] By *freguesia.*

4 The cities hierarchy – centrality

The centrality represents the functions range provided by central places. Central places that provide more specialised functions will have higher centrality indexes.

In the theory, the most central place will be the one that provides more functions[10]. However, empirically this assumption is not valid: a city that provides a function of *n* level (degree of specialisation) does not always provide all functions of lowers levels; there is no city that provides all the 126 functions (Lisboa provides the maximum number of functions – 122).

The construction of a centrality index is based on an essential assumption - cities that provide more functions will be more central. In addition, functions will be weighted according to the following principles:

- More central functions located in higher positions of hierarchy will be considered more important. So functions will be weighted by their degree of specialisation (S).
- Cities with more functional units providing a specific function will be considered more relevant. Between two cities that supply the function *hospital*, will be considered more important the one that has more functional units. So the number of functional units (FU) will weight functions in cities.

The centrality index (CI) will be calculated according to the next formula:

$$CI_j = \sum_{i=1}^{126} \left[\exists F_{ij} * \left(\frac{S_i + FU_{ij}}{2} \right) \right], \text{ where:}$$

- i represents the central functions and j the cities.
- $\exists F_{ij}$ represents a binary variable (0,1) that assumes the value 1 when the city j provides the function i and 0 in the opposite situation.
- S_i means the degree of specialisation of function i and it is inversely proportional to the number of functional units that exist in Portugal. It is equal to the inverse of number of functional units of the function i.
- FU_{ij} represents the dimension of function i in city j and it is equal to the number of functional units.
- E_i and UF_{ij} were normalised, making their maximums equal to one. In UF_{ij} this normalisation is made by function. By this way, both factors have the same weight on centrality index construction.

Table 1 shows the cities hierarchy resulting from the methodology presented above.

[10] For Christaller one central place that provides a function of *n* level (degree of specialisation) also provides all functions of lower levels.

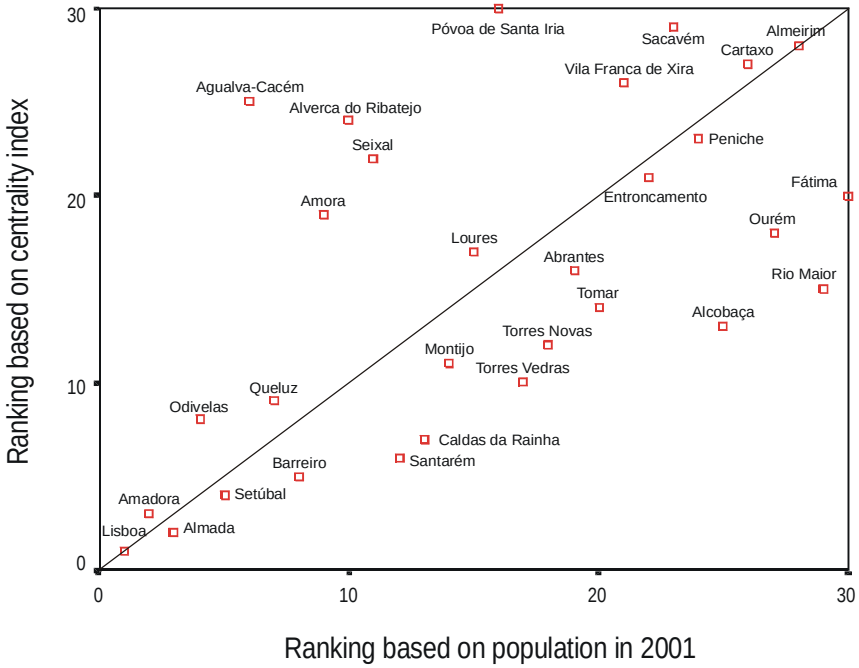
Table 1 – Cities hierarchy

Cities	Centrality Index	N^{er} of functions	Population 2001	Area (km^2) 2001
Lisboa	64,93	122	564 657	84,6
Almada	16,08	117	111 933	25,0
Amadora	15,85	114	175 872	23,8
Setúbal	12,48	114	91 319	53,3
Barreiro	11,18	112	53 909	8,3
Santarém	10,60	116	30 537	78,7
Caldas da Rainha	9,50	113	29 511	46,7
Odivelas	9,22	104	92 175	10,9
Queluz	9,14	108	78 123	6,7
Torres Vedras	8,70	109	23 831	62,5
Montijo	8,38	111	29 173	41,2
Torres Novas	8,34	112	22 405	72,8
Alcobaça	8,22	108	15 451	82,2
Tomar	7,95	108	18 904	31,2
Rio Maior	7,45	107	11 532	90,0
cities of LMA				

Cities	Centrality Index	N^{er} of functions	Population 2001	Area (km^2) 2001
Abrantes	7,34	104	22 028	121,3
Loures	6,60	103	28 429	47,9
Ourém	6,46	103	11 919	61,3
Amora	6,26	93	50 991	24,5
Fátima	6,26	95	10 302	71,9
Entroncamento	6,21	106	18 173	13,8
Seixal	6,03	96	31 116	13,7
Peniche	6,03	103	15 595	7,7
Alverca do Ribatejo	5,98	97	40 065	23,4
Agualva-Cacém	5,78	98	81 843	10,4
Vila Franca de Xira	5,45	98	18 442	212,1
Cartaxo	5,05	96	14 501	61,6
Almeirim	4,83	95	11 607	68,9
Sacavém	3,93	84	17 659	3,8
Póvoa de Santa Íria	3,76	82	24 277	4,8

As it was concluded by Central Place Theory it is possible to find a strong relation between centrality of cities and their number of inhabitants (Figure 2). Exceptions of this rule are the cities Agualva-Cacém, Alverca do Ribatejo, Póvoa de Santa Iria, Amora and Seixal. These cities are located in the periphery of Lisboa. The intense growth of their population, as a result of suburbanisation process, was not kept up with decentralisation of functions from Lisboa city.

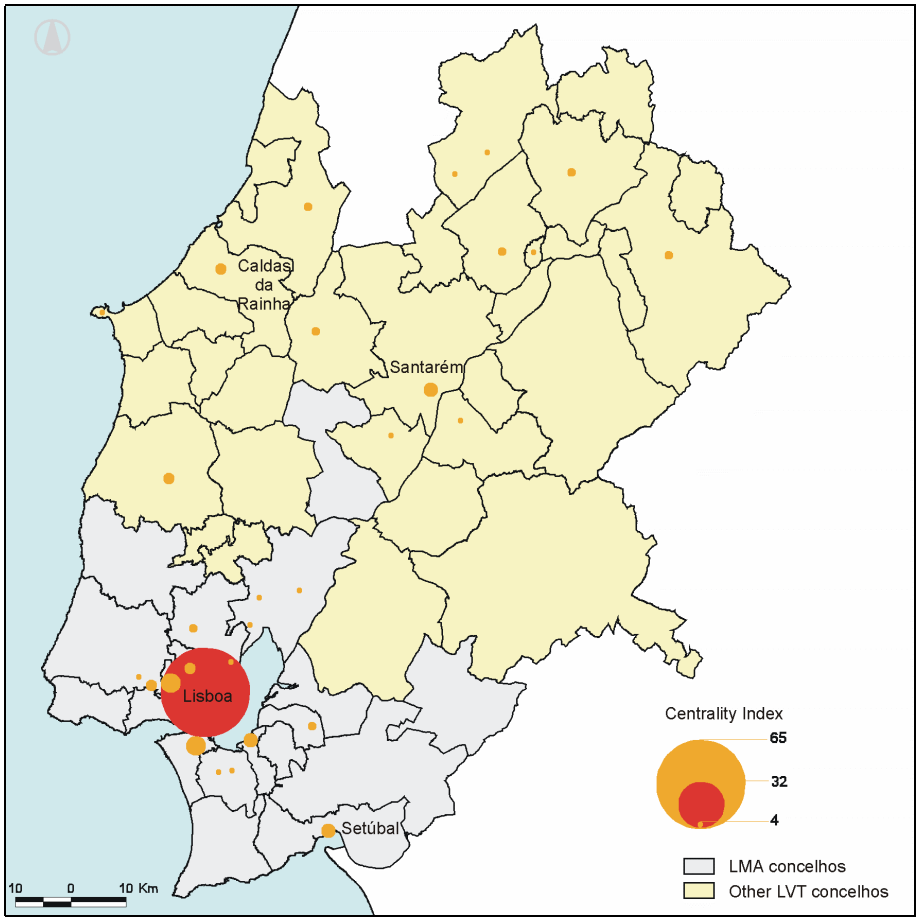
Figure 2 – Cities according to their position in ranking based on centrality index and ranking based on population in 2001



Analysing the cities network of *Lisboa e Vale do Tejo* (Figure 3) it is possible to identify the following features:

- There is a set of important cities located in the core of LMA – Lisboa, Almada, Amadora and Barreiro – with high geographic proximity and solid integration, that can be observed through the high commuting movements among them.
- Also in the LMA we found less important cities – Odivelas, Agualva-Cacém, Queluz, Loures, Sacavém, Póvoa de Santa Iria, Alverca do Ribatejo, Vila Franca de Xira, Seixal, Amora e Montijo - forming an external ring of the LMA core.
- Setúbal appears as a city with high centrality index (4th in the ranking) and represents an important centre of the South area of LMA;
- Outside LMA, Santarém and Caldas da Rainha are the most central cities. Santarém appears as the most important city of the urban subsystem of *Tejo* region (NUTS III of *Lezíria do Tejo* and *Médio Tejo*) and Caldas da Rainha as the most important of the *Oeste* region subsystem.

Figure 3 – Centrality index of *Lisboa e Vale do Tejo* cities



5 The cities hinterlands

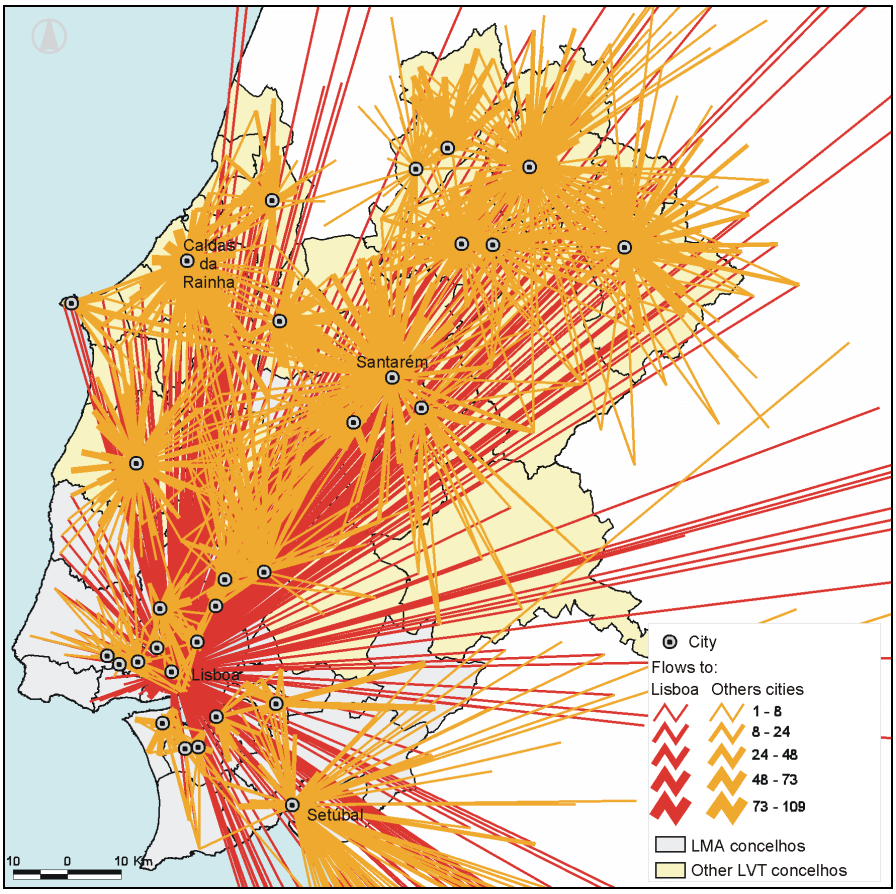
Cities hinterlands analysis allows a better understanding of how urban system of *Lisboa e Vale do Tejo* works. City hinterland is formed by the set of *freguesias* whose population goes that city to consume central functions.

Based on presentation of flows due to central functions consumption (all 126 functions) (Figure 4) we can observe the following features of urban system of *Lisboa e Vale do Tejo*:

- The huge hinterland of its major city – Lisboa – that goes beyond LVT borders;
- The importance of Setúbal as destination for Alentejo region;
- The intense interaction among cities located around Lisboa, that makes their hinterlands extremely diffuses;

- Santarém and Caldas da Rainha are the cities located outside LMA with the greatest hinterlands.

Figure 4 – Flows of central functions consumption in LVT cities (all functions)



The less specialised functions are available in most of the cities and, as a consequence, their consumption flows are less and with smaller extensions (Figure 5).

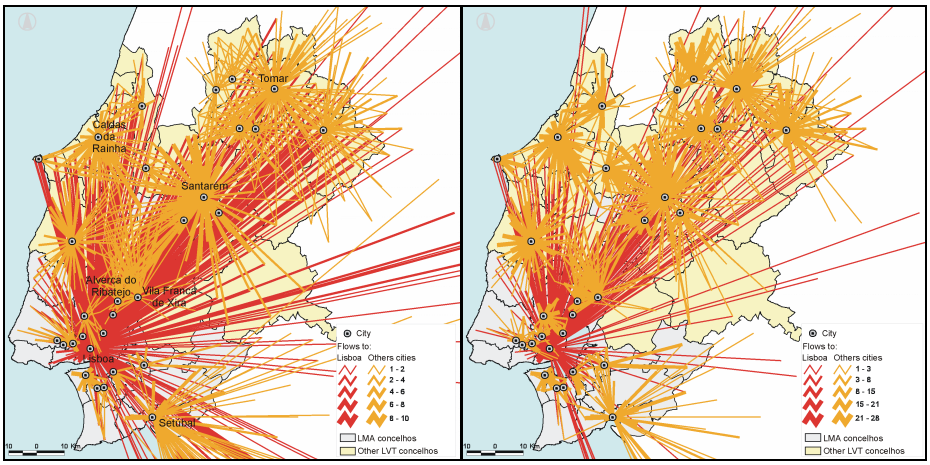
Due to their rarity, the most specialised functions (both highly specialised and specialised functions) tend to resume the most important flows. Hinterlands based on this sort of functions go beyond *concelho* borders (Figure 5).

The analysis of Figure 5 emphasises the features of LVT urban system described above and allows to add the following characteristics:

- Cities of Vila Franca de Xira *concelho*, mainly Alverca do Ribatejo and Vila Franca de Xira, have important hinterlands, specially towards North.
- Tomar appears as the principal city of Médio Tejo region, mainly in highly specialised functions;

Figure 5 – Flows of central functions consumption in LVT cities

Highly specialised functions **Specialised functions**



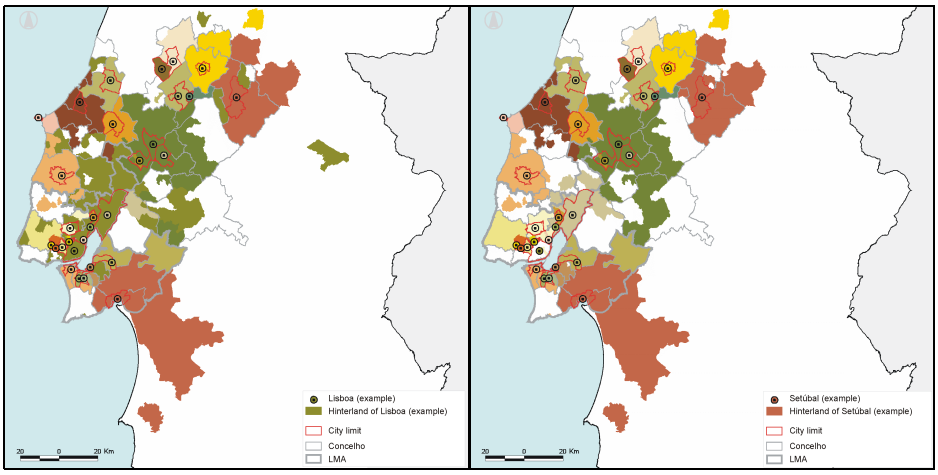
In order to measure the dimension of hinterlands, for example based on its population, it is necessary to link each *freguesia* to just one city. The most frequent destination for each *freguesia* was the base of this linkage process.

The next analysis will be segmented by the inclusion (or not) of Lisboa in the system. This procedure leads to a better understanding of the dimension of each city hinterland, because Lisboa, due to its high centrality, hides other cities hinterlands. Therefore, hinterland dimension is calculated without Lisboa, as a possible destination, for all cities, excepted for Lisboa itself. By this way, in Table 2, the population of Lisboa hinterland may be also part of other city hinterland.

Freguesias without a link to any city may have two different meanings: they provide most of the functions in analysis or the flow patterns are so diffuse that *freguesia* itself becomes the most frequent destination.

Figure 6 – Hinterlands of LVT cities (most frequent destination for highly specialised and specialised functions)

With Lisboa **Without Lisboa**



The cities with bigger hinterlands, measured by population, area mainly located inside LMA. Outside LMA, Santarém, Torres Vedras and Caldas da Rainha have the bigger hinterlands.

Table 2 – Dimension of LVT cities hinterlands

City	Population 2001	Area (km2) 2001	Dwellings 2001
Lisboa	1 225 580	2 167	599 722
Amadora	348 610	301	162 385
Almada	223 847	114	124 646
Loures	215 519	168	98 520
Setúbal	196 157	2 080	96 017
Barreiro	145 408	83	67 860
Santarém	143 598	1 995	69 123
Odivelas	116 953	19	51 109
Torres Vedras	110 805	600	58 138
Queluz	109 159	32	48 169
Vila Franca de Xira	108 506	584	49 636
Caldas da Rainha	82 774	582	45 701
Agualva-Cacém	81 843	10	36 765
Sacavém	58 891	14	24 150

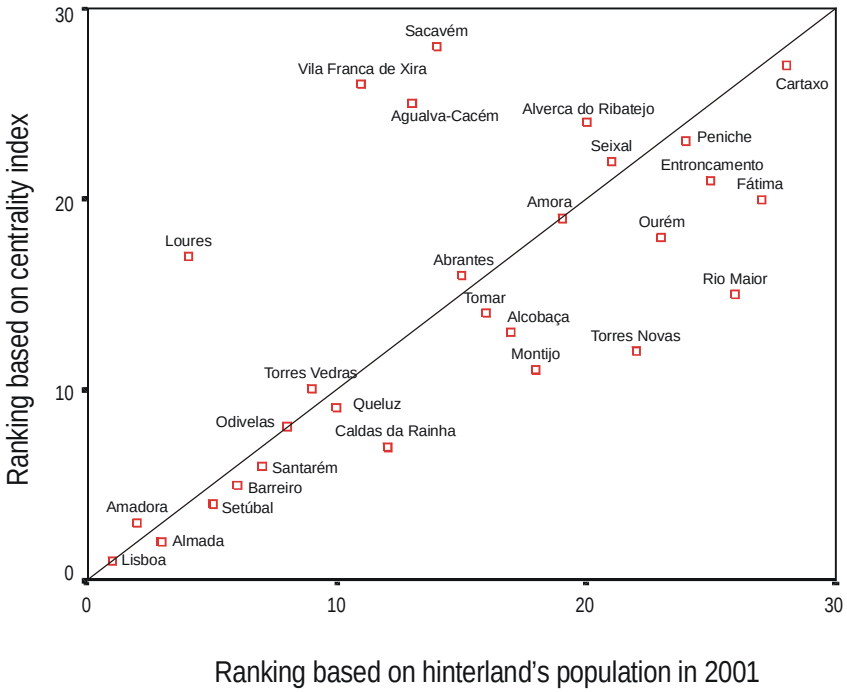
LMA cities

City	Population 2001	Area (km2) 2001	Dwellings 2001
Abrantes	57 720	1 302	33 499
Tomar	56 818	607	33 035
Alcobaça	56 320	347	29 257
Montijo	52 178	474	25 827
Amora	50 991	24	22 259
Alverca do Ribatejo	45 077	35	19 454
Seixal	42 053	30	19 344
Torres Novas	41 247	280	20 195
Ourém	33 587	311	19 260
Peniche	27 316	78	16 729
Entroncamento	25 783	64	12 494
Rio Maior	20 686	266	10 159
Fátima	10 302	72	5 068
Cartaxo	10 115	19	4 854

6 Conclusion

Despite of *Lisboa e Vale do Tejo* urban system does not present a strict hierarchical organisation, its structure can be substantially explained by Central Place Theory. The more central cities have more population either in city limits (Figure 2) or in its hinterlands (Figure 7).

Figure 7 – Cities according to their position in ranking based on centrality index and ranking based on its hinterland's population in 2001



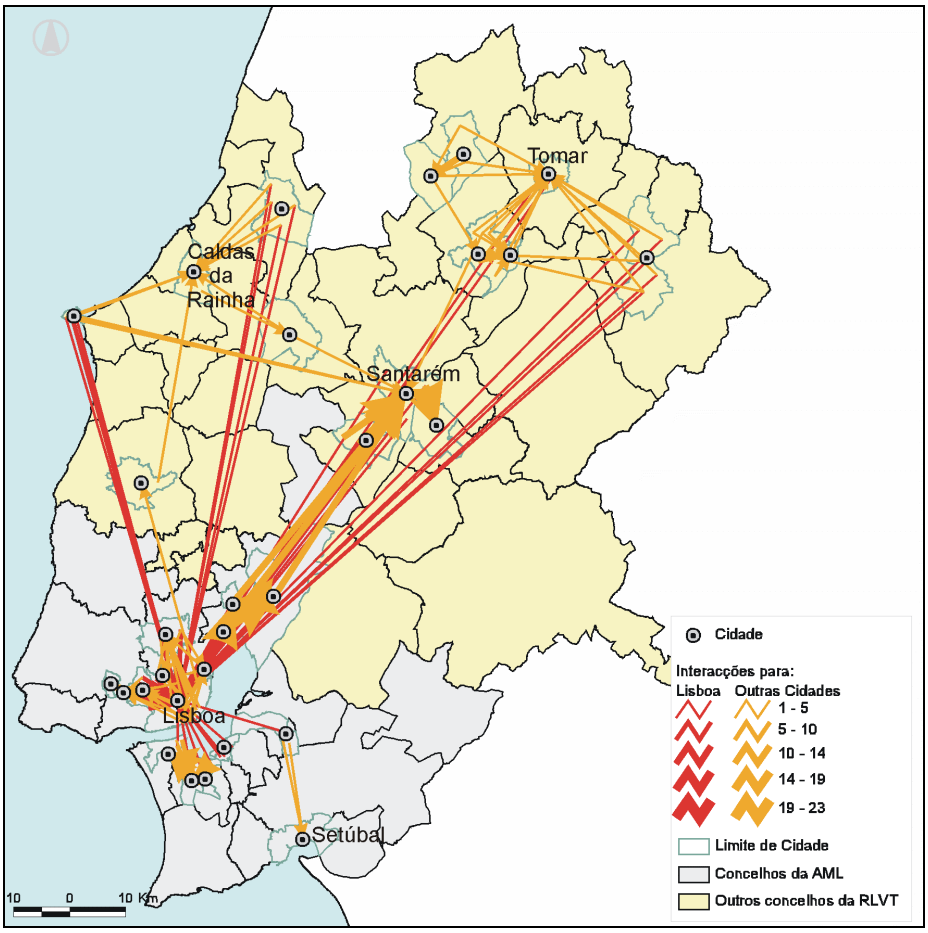
One can identify four distinct city levels on LVT region:

i. Lisboa as the main city, not only of LVT region, but also of Portugal, with a hinterland that covers almost all of the Portuguese territory;
ii. A group of cities located around Lisboa, with high centrality indexes but smaller hinterlands. This contrast can be explained by high geographic proximity among them;
iii. Setúbal, Santarém, Caldas da Rainha, Tomar and Torres Vedras are the principal cities at a regional level. They work as important destinations, principally at the level of most specialised functions;
iv. The rest of the cities are located in hinterlands of the former cities. However these cities provided a great part of 126 functions analysed in this study.

The flows of central functions consumption (only) between cities (Figure 8) are mainly from less central cities to higher central cities. Again, Lisboa, Santarém, Caldas da Rainha and Tomar appear as the cities that have bigger flows. Setúbal role

in this node approach is quite low, because territory outside LVT region represents a major part of its hinterland (Figure 6).

Figure 8 – Flows of central functions consumption between LVT cities (all functions)



7 References

1. Alves, M. Brandão et AL: *Formação de Centros e Sistemas Urbanos (Tópicos).* CIRIUS, série didática, documento de trabalho nº 5/98-99 (1999)
2. Benko, Georges: *A Ciência Regional.* Celta (1999)
3. Gaspar, Jorge: *A área de Influência de Évora: sistema de funções e lugares centrais*, 2ª edição. Centro de Estudos Geográficos (1981)
4. INE e DGOTDU: *Tipologia de Áreas Urbanas.* INE (1998)

5. Lopes, A. Simões: *Desenvolvimento Regional*, 4ª edição. Fundação Calouste Gulbenkian (1995)
6. Polèse, Mario: *Economia Urbana e Regional: Lógica espacial das transformações económicas.* APDR (1998)

8 Appendix - Central function hierarchy

Function	N^{er} of functional units	*Ranking*	Classes	
			with natural breaks (11)	Final (5)
Alcoholism treatment clinic	42	1	1	1
Health care centers with interning service	84	2	1	1
Slaughter house	85	3	1	1
Hospital	89	4	1	1
Drug addict treatment clinic	97	5	1	1
Secondary school - private	101	6	1	1
Hypermarket	112	7	1	1
Clinic with interning service	114	8	1	1
Health care center to AIDS	117	9	1	1
TAC service	126	10	1	1
Employment agency	130	11	1	1
Kennel	178	12	1	1
Professional formation center	196	14	2	2
Car inspection center	196	15	2	2
Rehabilitation center for locomotion handicap	196	13	2	2
Lower secondary school - private	201	16	2	2
Health care center to drugs addict	218	17	2	2
Primary school (5-6 years) - private	227	18	2	2
Commercial registry office	231	20	2	2
Autonomous service of ambulances	231	19	2	2
Tribunal	250	21	2	2
Professional school	256	22	2	2
Predial registry office	292	23	2	2
Civil registry office	295	24	2	2
Bi-weekly market	305	25	2	2
Weekly market	320	26	2	2
Tourism office	322	27	2	2
Secondary school - public	330	28	2	2
Notary's office	347	29	2	2
Treasury	348	30	2	2
Finance bureau	356	31	2	2
Health care centers without interning service	357	32	2	2
Echography service	370	33	2	2
Radiology service	405	34	3	2
Musical equipment store	413	35	3	2
School by TV	451	36	3	2
Monthly market	467	37	3	2
Car rent	471	38	3	2
Primary school (4 years) - private	515	39	3	2
Children's boarding homes	522	40	3	2
Fire corporation	524	41	3	2
Fire corporation with ambulance service	533	43	3	2
Fuel station (24hours)	533	42	3	2
Veterinary clinic	592	44	3	2
Police-station	642	45	3	2
Pets shop	701	46	4	3
Sweet herbs store	735	47	4	3
Travel agency	753	48	4	3
Market	753	49	4	3
Driving school	765	50	4	3
Exchange office	837	51	4	3
Shopping center	860	52	4	3
Primary school (5-6 years) - public	863	53	4	3
Annual market	892	54	4	3
Consultancy office	936	55	4	3
Lower secondary school - public	951	56	4	3
Nursing center	995	57	4	3
Record store	1001	58	4	3
Keys store	1070	59	4	3
Scrap dealer	1141	60	5	3
Free times occupation for young people	1189	61	5	3
Homes for the elderly people	1279	62	5	3
Copy center	1332	63	5	3

Function	N^{er} of functional units	*Ranking*	Classes	
			with natural breaks (11)	Final (5)
Optician	1370	64	5	3
Video Club	1446	65	5	3
Funeral agency	1449	66	5	3
Dye-house/Laundry	1480	67	5	3
Computing equipment store	1485	68	5	3
Homes for the elderly people (only day)	1528	69	5	3
Clinical analysis service	1562	70	5	3
Perfume's shop	1626	71	6	3
Free times occupation for children	1635	72	6	3
Transport agency	1650	73	6	3
Haberdasher's shop	1654	74	6	3
Parochial center	1676	75	6	3
Bicycle stand	1687	76	6	3
Nursery - baby unit	1713	77	6	3
Sport goods store	1731	78	6	3
Health care center delegation	1747	79	6	3
Motorcycle stand	1765	80	6	3
Estate agent's	1873	81	6	3
Nursery - private	1881	82	6	3
Tyre store	1916	83	6	3
Supermarket	1940	84	6	3
Building office	1986	85	6	3
Photographic goods store	2051	86	6	3
News stand	2113	87	6	3
Wholesale dealer	2276	88	7	4
Fuel station	2410	89	7	4
Beauty institute	2413	90	7	4
Fish shop	2468	91	7	4
Jewellery/Watch-maker's	2551	92	7	4
Pharmacy	2558	93	7	4
Repair-shop for agriculture equipment	2749	94	7	4
Farmer goods store	2813	95	7	4
Insurance broker	2834	96	7	4
Florist	2872	97	7	4
Drugstore	3056	98	7	4
Electric material store	3147	99	7	4
Trucks	3438	100	8	4
Motorcycle/bicycle garage	3512	101	8	4
ATM	3544	102	8	4
Bank	3545	103	8	4
Accounting office	3547	104	8	4
Lawyer office	3646	105	8	4
Bookshop/Stationer's	3661	106	8	4
Repair-shop for domestic equipment	3892	107	8	4
Car stand	3944	108	8	4
Repairing articles of personal use	4022	109	8	4
Fruit store	4050	110	8	4
Material store of construction	4063	111	8	4
Nursery - public	4106	112	8	4
Foot-wear store	4632	113	8	4
Domestic equipment store	4638	114	8	4
Furniture store	4958	115	8	4
Gas station	5020	116	8	4
Confectioner's	5668	117	8	4
Baker's shop	5729	118	8	4
Butcher's/Sausage shop	6708	119	9	5
Cars garage	7833	120	9	5
Primary school (4 years) - public	8322	121	9	5
Hairdresser	9956	122	9	5
Clothes store	10144	123	9	5
Restaurant	13667	124	10	5
Grocery	20330	125	10	5
Café, bar, tavern	34606	126	11	5

Machine Learning and Statistics to Detect Errors in Forms: Competition or Cooperation?

Carlos Soares[1], Pavel Brazdil[1], and Cláudia Pinto[2]

[1] LIACC/Faculty of Economics, University of Porto, R. Campo Alegre 823, 4150-180 Porto, Portugal, {csoares,pbrazdil}@liacc.up.pt
[2] INE/DRN, Edíficio Scala, R. de Vilar 235, 9º andar, 4050, Porto, Portugal claudia.pinto@ine.pt

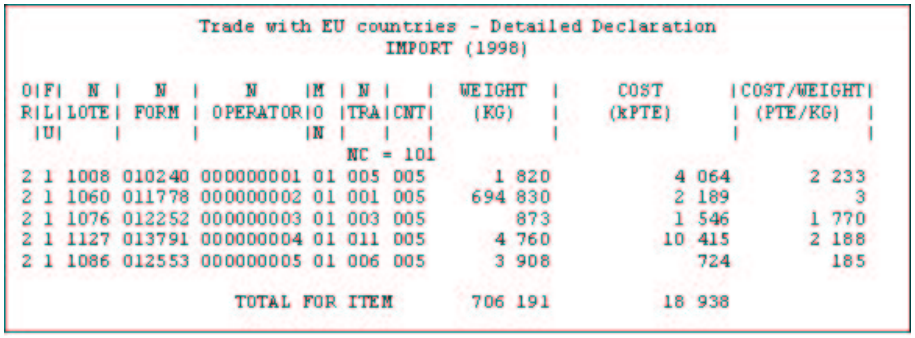
Abstract. We address the problem of detecting errors in foreign trade forms, which are submitted by companies to the Portuguese Institute Statistics (INE). In previous work, we have compared statistical techniques for outlier detection with an inductive learning method, with the latter obtaining the best results. Here, we present more recent results on that problem. We also hypothesize that the combination of outlier detection methods with inductive learning algorithms might be a better approach to dealing with this problem and we propose ways of doing so.

1 Introduction

This paper is concerned with the problem of detecting errors in foreign trade data collected by the Portuguese Institute of Statistics (INE). The objective is to identify the transactions that most likely contain an error. These will then be manually analyzed by specialized staff and corrected if an error really exists. Previous work on this problem has compared statistical methods for outlier detection with C5.0, a decision tree induction algorithm [1]. The latter technique obtained the best results, achieving the minimum goals that were established by the domain experts. This approach, however, only addressed part of the problem. As will be explained in Section 2, errors may be detected by looking into several fields while, in the work mentioned and which will be summarized in Section 3, only one field was taken into account. Here we extend the work to deal with all the relevant fields. We address this problem using a combination of models [2], which is currently a very popular approach in Machine Learning and Data Mining (Section 4). The model combination approach is then proposed as a way of enabling "cooperation" rather than "competition", as was done in [1], between learning and outlier detection methods (Section 5).

2 The Problem

The transactions made by portuguese companies with organizations from other EU countries are communicated to the Portuguese Institute of Statistics using the INTRASTAT form. In this form the company provides information about the transaction, namely:



Trade with EU countries - Detailed Declaration
IMPORT (1998)

O R	F L U	N LOTE	N FORM	N OPERATOR	M O N	N TRA	CNT	WEIGHT (KG)	COST (kPTE)	COST/WEIGHT (PTE/KG)
					NC = 101					
2	1	1008	010240	000000001	01	005	005	1 820	4 064	2 233
2	1	1060	011778	000000002	01	001	005	694 830	2 189	3
2	1	1076	012252	000000003	01	003	005	873	1 546	1 770
2	1	1127	013791	000000004	01	011	005	4 760	10 415	2 188
2	1	1086	012553	000000005	01	006	005	3 908	724	185
				TOTAL FOR ITEM				706 191	18 938	

Fig. 1. An excerpt of the INTRASTAT database. The data presented was modified to preserve confidentiality.

- item id
- weight of the traded goods
- total cost
- type (import/export)
- source, indicating whether the form was submitted using the digital or paper versions of the form
- form id
- company id
- stock number
- month
- destination or source country, depending on whether the type is export or import, respectively.

At INE, the data are inserted into a database. Figure 1 presents an excerpt of a report produced with data concerning import transactions for a month in 1998 of item with id 101, as indicated by the field labelled "NC", below the row with the column names.

Given that both form filling and its transcription to the database are manual processes, errors often occur. For instance, an incorrectly introduced item id will associate a transaction with the wrong item. Another common mistake is caused by the use of incorrect units like, for instance, declaring the cost as PTE instead of kPTE. Some of these errors have no effect on the final statistics while others can affect them significantly.

The number of transactions declared monthly is in the order of tens of thousands. When all of the transactions relative to a month have been entered into the database, they are manually verified with the aim of detecting and correcting as many errors as possible. In this search, the experts try to detect unusual values on a few attributes. One of these attributes is Cost/Weight, which represents the cost per kilo and is calculated using the values in the Weight and Cost columns. In Figure 1 we can see that the values for Cost/Weight in the

second and last transactions are much lower than in the others. If we analyze the corresponding forms we can conclude that the second is, in fact, wrong, due to the weight being given in grams rather than kilos, while the last one is correct.

Our goal was to reduce the time spent on this task by automatically selecting a subset of the transactions that includes almost all the errors that the experts would detect by looking at all the transactions. To be acceptable by the experts, the system should select less than 50% of the transactions containing at least 90% of the errors. Additionally, the experts would like to be able to understand the decisions made by the system, and thus, it should provide an interpretable model. Note that efficiency is not important because the automatic system will hardly take longer than half the time the human expert does.

The automation of this task is not easy. Firstly, although it is not shown in Figure 1 for the sake of simplicity, quantity and cost may be declared in two fields each. In same goods the quantity may be measured in a way other than weight, which is called *supplementary units*. For instance, in a transaction of cars, the supplementary units are the number of units. A transaction may also have two values, the one which is declared, the *invoice value*, and a *statistical value*, calculated using an appropriate formulae. Therefore, we may have up to four combinations of Cost/Quantity to analyze. This adds more complexity to the problem since, on one hand, different combinations are the best error predictors for different groups of transactions. On the other hand, none of the combinations is present in all the transactions. Therefore, a complete solution to the problem requires all the combinations to be treated.

Furthermore, the error in the example shown above was easy to detect, but many of the errors spotted by the experts are not so obvious and their experience is essential to detect them. This information cannot be used directly in this work because we have no indication of which errors are "obvious" and which ones are not. Other difficulties are the small proportion of errors relative to the number of transactions (<0.5%), the patterns of normal transactions, which differ from item to item, and items with very few transactions.

3 Previous Results

In [1], a first approach to this problem only took the Cost/Weight attribute into account. The data concerned transactions from five months in 1998. It was provided in the form of two files per month, one with the transactions before being analyzed and corrected by the experts, and the other after that process. A considerable amount of time was spent preparing the data, for instance, to eliminate transactions that existed in one of the files but not in the other.

Four very different methods were applied. Two come from statistics and are univariate techniques: box plot [3] and Fisher's clustering algorithm [4]. The third one, Knorr & Ng's cell-based algorithm [5], is an outlier detection algorithm which, despite being a multivariate method, was used only on the Cost/Weight attribute. The last is C5.0 [6], a multivariate technique from the field of machine learning, based on induction of decision trees.

Decision tree induction is not an outlier detection technique and so we need to help the system to learn the notion of outlier. We created a new attribute that captures the notion of distance between points, i.e. different values of the Cost/Weight variable. Thus, NCost/WeightDistance represents the distance between a point (a value of Cost/Weight) and the average, in number of standard deviations:

$$\text{NCost/WeightDistance} = \frac{Cost/Weight - \overline{Cost/Weight}}{\sigma_{Cost/Weight}}$$

Note that for each transaction we use the average and standard deviation of the corresponding item. Finally, we added the attributes average and standard-deviation of cost weight and number of transactions, all calculated for each item id. From the set of original attributes, we selected only Cost, Weight and Cost/Weight. All the others are not expected to be predictive in their original form and, thus, were discarded. Following advice from the domain experts, import and export transactions were handled separately because their errors are quite different.

Given that C5.0 provides a confidence score associated with each prediction, it enables a compromise between the manual effort required and the number of errors detected to be made. This can be achieved by analyzing the positive predictions (i.e. the transactions that are identified as potential errors) in descending order of confidence and stopping when enough transactions have been analyzed. The results obtained based on this approach indicate that it is possible to detect 90% of the errors by analyzing just 40% of the transactions. Therefore, we were able to detect the required number of errors with less 10% of transactions than the domain experts would be willing to analyze.

Furthermore, the top nodes of the decision tree induced by C5.0 contain knowledge which makes sense to the domain experts. The first node, as would be expected, does outlier detection using the NCost/WeightDistance attribute. The second node, identifies small values of Cost/Weight. The experts confirmed that they pay more attention to smaller values of this attribute, because when an error generates a small value of Cost/Weight it usually affects the statistics more significantly than if it generates a large value. Finally, the third node recommends full manual verification of items with few transactions.

4 Dealing With All Cost/Quantity Fields

As mentioned in Section 2, a complete solution to the problem requires that all four combinations of the Cost/Quantity attributes be handled and not just one, like was done in the work described in the previous section. There are a number of ways to do that. We have opted to follow a model combination approach, which is currently very popular in Machine Learning and Data Mining [2]. We create a different data set for each problem (i.e. each combination), containing the attributes corresponding to the particular Cost/Quantity combination, the target attribute and other general attributes, like item id, etc, and then we

induce a model from each of those data sets. The next issue is how to combine the prediction of each of these models. Again there are several alternatives. We have opted for a voting scheme, which is also a common approach [7]. However, rather than using a uniform vote, where the most frequently voted class wins for each transaction, we have used weighted voting. The weight of each prediction is the corresponding confidence which enables us to explore the strength of each prediction [7].

Given that we are now dealing with all four combinations of the Cost/Quantity attributes, the number of transactions is larger than it was in the data used in the previous section because some of the transactions did not have value for the corresponding Cost/Quantity attributes combination. Therefore, our goal is to investigate whether the results in this new setting are better or worse than the ones obtained before. The evaluation method used is the hold-month-out approach, an adaptation of the well-known hold-out approach. The model is induced using data from all but one month and evaluated on the latter. The results are slightly worse than the ones obtained above although well within the limits established by the domain experts: 47% of the transactions were selected, containing 91% of the errors. This performance reduction was expected because the combinations of Cost/Quantity attributes which are now handled have much less data and very few errors. Therefore, the corresponding models are expected to have less generalization power, thus, hurting the general performance of the system.

5 Combining Inductive Learning and Outlier Detection

As summarized in Section 3, a few methods for the task of detecting potential errors in foreign trade data have been compared before [1]. These methods included not only an inductive learning algorithm but also outlier detection methods. As an alternative to the competition approach followed in that work, we believe that a collaboration approach, combining these methods, could obtain better results. As mentioned above, the combination approach is quite popular in Machine Learning [2] and, to the best of our knowledge, has never been applied to the problem of outlier detection.

As in the previous section, combination of models could be made by voting. An alternative approach is inspired in stacked generalization [8], where the predictions of base models are fed to a second-level inductive algorithm. This algorithm generates a model that predicts the same target as the base models but using their outputs as inputs, i.e. attributes. Cascade generalization [7] is an extension of stacked generalization, which feeds the original attributes used by the base models to the second-level algorithm, together with the predictions of the base models.

We have investigated a few outlier detection methods to use at the base level. Two methods that have been traditionally used in statistics for outlier detection are discordancy tests and principal component analysis [9]. More recent approaches that have been developed in the context of data mining are distance-

based [5], density-based [10] and unsupervised [11, 12] methods. Some of the more recent methods have been developed with large amounts of data in mind, as in the present case. We will select some of these methods to be used in our work. With this approach, we expect to improve our previous results. That is, we expect to select less than 40% of the transactions containing more than 90% of the errors.

6 Conclusions

We have presented recent results on the problem of detecting errors in foreign trade data that is collected by the Portuguese Institute Statistics (INE). We have used the structure of the problem, which was ignored in previous work, to divide it into four different sub-problems. The models generated for each of those problems were combined using weighted voting. We have also proposed the combination of outlier detection methods with inductive learning techniques, and investigated some outlier methods that could be used for that purpose. We will implement and empirically evaluate this approach.

Acknowledgments The authors wish to thank the INTRASTAT team at INE, without whom this work would not have been possible: Armindo Carvalho, Vítor Cortez, Óscar Alves, Fernanda Sengo, Ilda Alves and Natércia Ferreira.

References

1. Soares, C., Brazdil, P., Costa, J., Cortez, V., Carvalho, A.: Error detection in foreign trade data using statistical and machine learning methods. In: Proceedings of the 3rd International Conference and Exhibition on the Practical Applications of Knowledge Discovery and Data Mining (PADD99). (1999)
2. Diettrich, T.: Ensemble learning. In Arbib, M., ed.: The Handbook of Brain Theory and Neural Networks, MIT Press (2002)
3. Milton, J., McTeer, P., Corbet, J.: Introduction to Statistics. McGraw-Hill (1997)
4. Fisher, W.: On grouping for maximum homogeneity. Journal of the American Statistical Association (**53**) 789–798
5. Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th VLDB Conference. (1998)
6. Quinlan, R.: C5.0: An Informal Tutorial. RuleQuest. (1998) http://www.rulequest.com/see5-unix.html.
7. Gama, J., Brazdil, P.: Cascade generalization. Machine Learning **41** (2000) 315–343
8. Wolpert, D.: Stacked generalization. Neural Networks **5** (1992) 241–259
9. Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley & Sons (1978)
10. Breunig, M., Kriegel, H.P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In: Proceedings of the MOD 2000, ACM (2000)
11. Yamanishi, K., Takeuchi, J., Williams, G.: On-line usupervised outlier detection using finite mixtures with discounting learning algorithms. In: Proceedings of the KDD 2000, ACM (2000)

12. Lauer, M.: A mixture approach to novelty detection using training data with outliers. In Flach, P., de Raedt, L., eds.: Proceedings of the 12th European Conference on Machine Learning, Springer (2001) 300–311

Utilization of Administrative Registers using Statistical Knowledge Discovery

Reijo Sund

National Research and Development Centre for Welfare and Health (STAKES),
P.O.Box 220, 00531 Helsinki, Finland
Reijo.Sund@stakes.fi

Abstract. The increase in the volume of data collected for administrative purposes has introduced new difficulties in extracting useful information to support decision making. The aim of this study is to present some new methodological approaches for the utilization of administrative registers in research and in production of statistical information by combining fruitful ideas from the fields of statistics and data mining. Since the most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes, the ideas are first described as steps in that process and then also illustrated with a practical real-world example.

1. Introduction

The resources of providers of health and social services are limited. The effective use of these resources needs administrative planning and political willingness. Moreover, political issues and the increasing role of cost-effectiveness considerations have recently, among other things, introduced new pressures to improve health and social service systems (see e.g. [1]).

To support decision making and quality control there is a demand for scientifically valid and comparative information. On the other hand, the information revolution has made it possible to effectively collect and store huge data sets (see e.g. [2], [3], [4]).

For example, the administrative health registers in Finland contain massive amounts of detailed data of good quality. However, raw data as such are of little value since in the context of problem solving the results are important and not the actual data. To fulfil the need of information, an important aim in the development of administrative information systems development has recently been the improved utilization of register based data (see e.g. [5]).

In principle, data can be converted to useful information, if answers to relevant problems can be extracted from the data by appropriate means (see e.g. [6]). However, the growing size of data sets has brought out challenges for the traditional statistical approaches (see e.g. [7]). In addition, from the statistical point of view it is important to notice that administrative registers are so called second hand data sets, i.e. the data have already been collected and there is no more possibilities to design

optimal data collection strategies to solve some particular problem. It is also unrealistic to assume that the understanding of complex methodological details is the only way to get results in the practical data analysis. The need of information together with the lack of statisticians has lead to the development of alternative, more concrete, ways to analyze data (see e.g. [8], [9]). In other words, statistics is no more alone in the field of data analysis as Huber [10] points out: "The most data analysis is done by nonstatisticians, and there is much commonality hidden behind diversity of languages. Rather than to try to squeeze the analysis into a too-narrow view what statistics is all about, statisticians ought to take advantage of the situation, get involved interdisciplinary, learn from the experience, expand their own mind, and thereby their field, and act as catalysts for the dissemination of insights and methodologies. Moreover, the larger the data sets are, the more important the general science aspects of the analysis seem to become relative to the 'statistical' aspects."

1.1 Aim of this study

The aim of this study is to present some new methodological perspectives for the utilization of administrative registers in research and in production of statistical information by combining ideas mainly from the fields of statistics and data mining. The most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes.

First the concept of statistical knowledge discovery process is defined in order to allow the embedding of ideas to a suitable framework. After that certain parts of statistical knowledge discovery process are examined from the point of view of administrative register data. A systems approach and an event history framework are described in the 'understanding the problem' part. In the 'data understanding' part generalized event sequence as well as data structures, censoring and practical and statistical interpretations related to it are presented. The 'data preprocessing' part reviews common preprocessing techniques: data abstraction, cleaning, integration and reduction. Also some preprocessing tasks and methods suitable for event history framework, including an 'interesting pattern remapping' technique, are suggested. After that the ideas are illustrated with a practical real-world example and finally the key points of this paper are discussed in the conclusions.

1.2 Previous research

There is a huge amount of research related to knowledge discovery in databases (see e.g. [9]). Essentially the starting point here is the mining of administrative data with scientific interpretation [11] related to certain phenomena [12]. Practical example used here is from the field of health services research, which is closely connected to health care and medicine (see e.g. [13], [14], [15], [16], [17], [18]). Other relevant references are given in the text while the specific issues are presented.

2. Statistical Knowledge Discovery Process

The increase in the volume of collected data has presented new difficulties in extracting useful information to support decision making. The traditional manual data analysis has become insufficient, and methods for efficient analysis strategies are needed. In order to find sensible solutions to real-world problems, the whole problem solving must be done in a systematic way. So called knowledge discovery process is a formalization of a complex problem solving process which fundamentally requires human participation (see e.g. [19], [20]). There is even a proposed global standard for knowledge discovery process available [21]. Fundamental ideas in knowledge discovery process seem to be closely related to traditional statistical thinking and statistical practice (see e.g. [22]). As a consequence, statistical knowledge discovery process can be thought to consist of the following interactive steps: understanding the problem, understanding data, data preprocessing, modeling, evaluation and reporting. In this paper the first three steps are examined more thoroughly.

2.1 Understanding the problem

For problem solving, the problem has to be "matched" against the corresponding domain knowledge and data. Usually a statistician or data analyst is not an expert on the field where the problem arises. It means that the effective co-operation with domain experts is a fundamental part of knowledge discovery process. Here the characterization of the problem is formalized in a way which has been found to be easily understandable also for non-statisticians and, as a consequence, gives a common language for the experts of different fields.

A Systems Approach. Very often the starting point for a statistical modeling of any phenomenon is to characterize the important properties of the phenomenon, i.e. operationalize it as a system. A system is defined as a 'group of things or parts working together or connected in some way as to form the whole'. This definition of a system needs that the objects and restrictions related to the system are adequately described.

Time is essential factor in many problem domains. For example, disease processes evolve in time and patient records give the history of patients. For such dynamic phenomena, time should be explicitly considered in the system formulation.

Event History Framework. In fact, the importance of the longitudinal event history approach has long been recognized in many areas, especially in social sciences, econometrics and medical research (see e.g. [23], [24], [25], [26], [27]). In the analysis of dynamic phenomenon the focus is on the sequences of events which occur in time. Usually these event histories are described with the individuals' transitions across a set of discrete states in time. In the simplest case - known as survival analysis (see e.g. [28], [29]) - only one qualitative change is possible, for example the transition from a state 'alive' to the state 'dead'. However, in many cases one transition can not describe the dynamics of a phenomenon in a realistic way. For example the slow development of a disease, which finally reaches the state 'death', can not be modeled accurately using only the time between the first diagnosis and death.

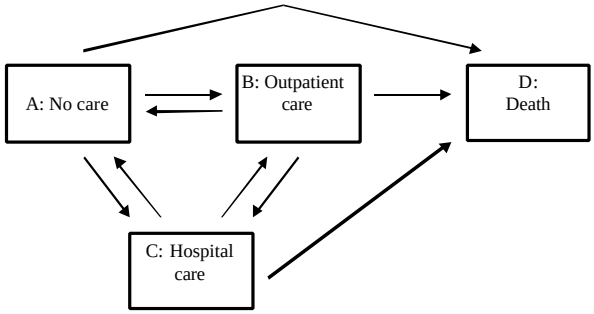


Fig. 1. A system as a directed graph

If the system is constructed in a reasonable way, the individuals' event histories can be considered as 'paths' through the system. Graphically the system can be presented as a directed graph whose nodes and edges have characteristics, which describing the properties of possible states and transitions in a system (Fig. 1).

2.2 Understanding data

Administrative data sets are massive second hand data sets. In general, data must have an appropriate form for storage and analysis purposes and for intuitive interpretation, as well. Here the essential parts of the traditional event history data structures including censoring (see e.g. [24], chapter 2) and event sequences originating from data mining (see e.g. [30], chapter 4) are combined to a generalized event sequence. Also some statistical models suitable for the analysis of generalized event sequence type data are presented, since with second hand data sets there are no more possibilities to design optimal data collection strategies to solve some particular problem, i.e. the data give the restrictions to analyzing possibilities.

Data structures. Event history data consist of observations of the form $(\tau_k,\mathbf{D}_k)$, where τ_k is an 'occurrence time' and $\mathbf{D}_k$ is an 'explanation' for the event (and n is the number of observations and $\tau_1 \le \tau_2 \le \ldots \le \tau_n$ and $\tau_i < \tau_j$ for at least one observation $i \ne j$ and i,j,k = 1,2,…,n).

Usually $\mathbf{D}_k$ consists of a set of attributes (variables). However, all the attributes do not necessarily contain important or interesting information. Some attributes can be unrelevant in relation to the solving of particular problem or can be easily derived from other attributes. It is often reasonable to divide up the relevant information into two subsets of attributes. The first subset defines an event type $\mathbf{E}_k$ and the other includes important 'covariate' information $\mathbf{i}_k$. For generality it is useful to allow a transformation f_k for the occurrence time τ_k. If not stated otherwise, $f_k(\tau_k) = \tau_k$ (k = 1,2,…,n).

Let m be the number of distinct occurrence times, t_i be the ith distinct occurrence time (i=1,2,…,m) and the event set $\mathbf{A}_i$ be the set of relevant information of observations occurred at the same time, i.e. $\mathbf{A}_i = \{(t_i,\mathbf{E}_k,\mathbf{i}_k)\}$, where i=1,2,…,m and for every i, k is over the observations for which $t_i = f_k(\tau_k)$.

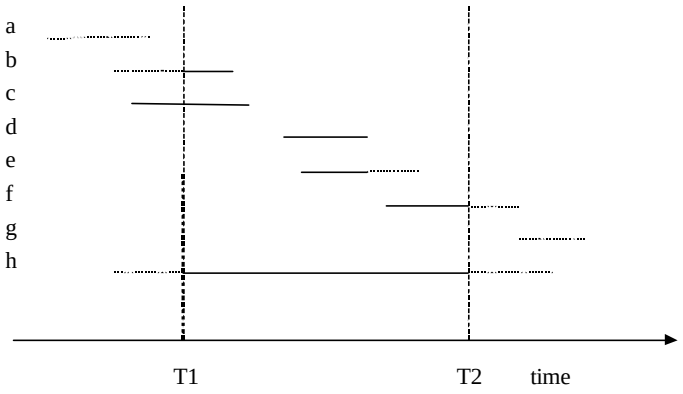


Fig. 2. Censoring

Censoring. In practice a data set is only a narrow window to the dynamics of a phenomenon. In other words the observations in the data set fulfill the condition $a < t_i < b$, where a and b are finite constants and i=1,2,…,m. The problems induced by a limited observation window are illustrated in Fig. 2, which shows examples of different types of 'censoring' for an individual's length of stay in some particular state (see also [27], 3-9). It is useful to include the type of censoring into data as an attribute, since the censoring must be taken care of in the analyses.

In the cases a and g there is no observed transitions to or from the state. This kind of censoring can be very problematic if the very first (or last) occurrence of some event type is considered more important than other occurrences (for example the first diagnosis of schizophrenia or the first back surgery operation). In the case b the transition to the state is not observed, but the transition from the state is known. In the case c both transitions are observed, but there is a potential problem, because the data outside the observation window are not complete for all individuals (typical situation for hospital discharge data). The case d corresponds to the uncensored observation. In the case e there has been an unknown or 'wrong' transition from the state (drop-outs or 'competing risk'). It is also possible that the follow up is ended before the occurrence of the event of interest (case f) or the transitions to and from the state are outside of the observation window (case h).

Generalized Event Sequences. The generalized event sequence **S** is defined to be a queue of event sets sorted by the (transformed) occurrence time, i.e. $\mathbf{S} = < \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m >$. In addition, let the conditional event sequence, to be called an event subsequence, be a sequence $\mathbf{S}_\theta = < \mathbf{A}_i \mid \mathbf{A}_i \in \mathbf{S}$ and condition θ is true $>$, where i=1,…,m.

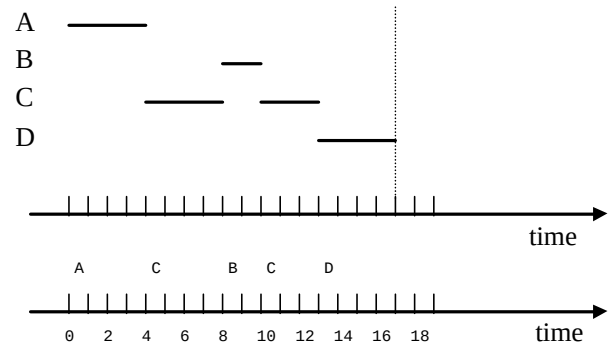
The definition of the generalized event sequence given above is very flexible and it can be used without any knowledge about the event history framework. However, using this framework generalized event sequences get a constructive and an intuitively clear interpretation. Moreover, transformations allowed in generalized event sequences make it possible to use the same data structure in the implementations of different statistical and data mining methods.

An Example of a Generalized Event Sequence. Fig. 3 shows an example of event history data and the corresponding event sequence, which could be a production of a system, let it be system P, presented in Fig. 1. There are also two graphical presentations for this particular event history. The first one is an event history description of transitions in a system where an individual actually stays in a current state until there is a transition to the another state. In the second one only the occurred events are marked to the figure. Since the time between two consecutive events is also the length of stay in a particular state, it is often very useful to include a 'length of stay' attribute into the covariate attributes $\mathbf{i}_k$, even though it can be easily calculated from the corresponding occurrence times.

In the definition of a generalized event sequence, parallel occurrences of events are allowed. In principle the systems approach can be made valid in this case by defining each combination of event sets to be a 'new' event type. In other words, it can be thought that the explanation for an event type is a combined description of events occurred at the same time. However, in many cases it is reasonable to classify the parallel events to different event types, since it is possible to find 'natural' interpretations for these parallel events.

Data in Fig. 3 is a description of movements of individual X in a system P. Let Q be a system with two states (state F: 'married' and state G: 'not married'). As a consequence, data corresponding to the movements of individual X in a system Q has the same form as the data in Fig. 3. In Fig. 4a is an example of that kind of situation. Now there are two event histories for individual X, one corresponding to the path through the system P and the other through the system Q. In this case it is known that the systems which 'generate' the data are parallel, but the both event histories can be combined into one event sequence (Fig. 4b).

T_k	$\mathbf{E}_k$
0	A
4	C
8	B
10	C
13	D



$$\mathbf{S} = < \{(0,A)\}, \{(4,C)\}, \{(8,B)\}, \{(10,C)\}, \{(13,D)\} >$$

Fig. 3. An example of event history data and the corresponding event sequence

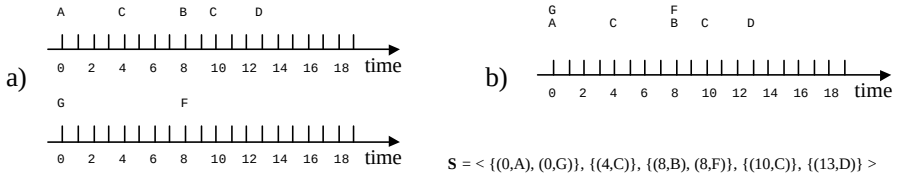


Fig. 4. An example of a situation with parallel occurrence times

In the case of two parallel systems the information concerning the 'source system' of an observation is a very valuable covariate. In general, there can, of course, be more than two data generating systems. If there is a need to restrict analyses to the observations from some particular system, that can be easily done using an event subsequence conditional to the corresponding system.

In practice there are always event histories for more than one individual. In that case it is trivial to include a covariate which identifies the individual, for example the social security number, while the whole data set still have the form of a generalized event sequence. Again it is possible to restrict the analyses to the arbitrary set of observations using event subsequences with suitable conditions.

Statistical Interpretations of a Generalized Event Sequence. From the statistical modeling point of view, a generalized event sequence can be interpreted as a sample path of a marked point process, if the occurrence times for all events are distinct. This leads to a very general family of statistical hazard-rate models suitable for censored data (see e.g. [31]).

In practice there is a lot of situations where so called calendar time of event occurrence is not important, since the 'real' information is the time between two consecutive events. In other words, the 'starting time' of a follow up can vary between individuals. A traditional solution, very common in survival analysis, is to transform the time axis from calendar time to the 'failure time'. This can be done using a transformation function $f_k(\tau_k) = \tau_k - b_k$, where b_k is the occurrence time of 'starting event' of a corresponding individual and k = 1,2,…,n (see e.g. [28], [29]).

Assuming that the probability of the next event type (state) depends only on time stayed in the current state, an appropriate choice for a model is a semi-Markov-model (see e.g. [32], [33]). Choosing the probabilities to change only on discrete time points, the semi-Markov-model can be formulated using the Markov chain, where the state space is expanded in a proper way (see e.g. [34]). In the traditional Markov chain the probability of the next event type (state) depends only on the current state (first order Markov property) and the probabilities are time-homogeneous (see e.g. [35], 372-427; [36], 61-84). As a matter of fact, the Markov chain interpretation corresponds to the situation, where the exact occurrence time is not important and only the order of observations matters. A generalized event sequence can be transformed to an event type sequence of this type using the transformation $f_k(\tau_k) = k$ (k = 1,2,…,n).

Moreover, choosing the transformation function f_k to be of the form $f_k(\tau_k) = c$, where c is an arbitrary constant and k = 1,2,…,n, the time dimension of the event sequence can be eliminated and the sequence reduces to data miners' classical market basket model (see e.g. [9], chapter 6) .

2.3 Data preprocessing

Usually massive second hand data sets contain so much information, domain specific features, inaccuracies and problems that raw data as such are not usable. To use data in problem solving, there must be understanding about the connections between the problem and data. In register based analyses the problem, domain knowledge and data also determine the most suitable model for the final problem solving. Moreover, most analyzing techniques are feasible to use only on moderately small data sets, since those typically need for access to the whole data set and the processing time will be directly proportional to the physical file size.

All in all, it can be stated that a sophisticated preprocessing incorporating non-technical domain knowledge in order to scale things down to a size fit for statistical analyses is the most important and time consuming part in register based data analysis.

Data abstraction. Often the connections between highly specific raw data and the highly abstract domain knowledge are so complicated that it is not possible to find any direct links between data and knowledge. In other words, an intelligent interpretation of raw data must be embedded into analyses, so that the resulting derived data set is at the level of abstraction corresponding to the current problem. Since noise is an unavoidable phenomenon also some kind of data validation and verification which makes use of knowledge should be performed. This kind of "intelligent" action is called data abstraction (see e.g. [17], chapter 2). For example, in the discovery of medical knowledge data is usually patient specific, while medical knowledge is patient independent and consists of generalizations that apply across patients.

Data cleaning. Real world data are very often more or less incomplete, noisy and inconsistent. Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data (see e.g. [37]). Compared to data abstraction, cleaning is more data driven and technically oriented. In other words, corrections of erroneous and inconsistent codes as well as missing values can be usually made to the whole database, but data abstraction always results in problem specific derived data sets.

Data integration and reduction. Two other common types of preprocessing are called as data integration and data reduction. The idea in data integration is to include data from multiple sources in analyses and it is also known as record linkage (see e.g. [38], [39]).

Data reduction obtains a reduced representation of a data set which is much smaller in volume than the original data set, yet produces the same (or almost the same) analytical results (see e.g. [9], chapter 3). As a matter of fact, data reduction can consist of anything from simple database queries to very complicated modeling.

Preprocessing tasks and methods in event history framework. Since preprocessing is highly domain and problem specific, it is difficult to give any general suggestions concerning tasks and methods. Moreover, sometimes it can be difficult to distinguish what is preprocessing and what is modeling.

However, some non-trivial preprocessing tasks in the event history framework could be for example: "define" the state space for some system (what happens to patient after surgical operation?), find interesting and frequent combinations of patterns (which are frequent combinations of diagnosis and operation codes?), classify patterns to adequate hierarchies (which diagnoses relate to complications of a surgical operation?) and confirm the expert's "hypotheses" about the phenomenon from data.

Data mining tools from the “frequent pattern family” (see e.g. [40]) seem to be useful to these kinds of preprocessing purposes. The extensions of the basic ideas including templates (see e.g. [41]), multiple levels (see e.g. [42]) and measures of interestingness (see e.g. [43], [44]) and similarity (see e.g. [30]) are also useful in certain situations.

One very straightforward, but extremely useful, technique is so called interesting pattern remapping. The idea is to first to "forget" the time dimension of the generalized event sequence by using appropriate transformation. Since the whole data then reduces to market basket case, it is possible to use a levelwise search (see e.g. [45]) to extract the frequent patterns from data regardless of occurrence times. In the remapping phase interesting patterns can be classified to appropriate event types simply by "giving a new name" for each pattern. Finally the time dimension is restored and records without interesting event types are removed. This abstraction usually results in a notably reduced data set with an interpretation corresponding to the current problem.

3. Practical example

This example is a simplified extraction from a study, which aimed 1) to develop and implement register based performance indicators to measure the effectiveness of surgical treatment of hip fracture and 2) to evaluate and compare the effectiveness of health care providers. The complete results are reported elsewhere ([46], [47]). The study is a part of a larger project which aims to develop register based methods for measurement of effectiveness in specialized health care.

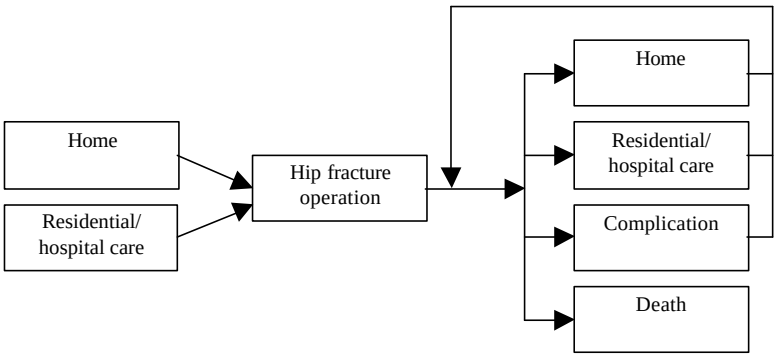


Fig. 5. System characteristics of remarkable life events related to hip fracture surgery

Understanding the problem. The first task in the project was to build up a research group consisting of experts of different fields. The group defined the actual problem more meticulously: the idea was to identify all hip fracture patients from the Finnish Health Care Register and follow the life events they encountered after hip fracture surgery according to register data.

A simplified system related to this particular problem is shown in Fig. 5. As it can be seen, the first hip fracture operation performed on an individual patient has a key role in the characterization of this phenomenon. Actually the state preceding the first hip fracture operation also matters from the clinical point of view, since the patients coming from home are usually in better condition than patients who are already in residential or hospital care. However, the major interest is in the events and pathways of care following the first hip fracture operation. In this case these events are classified into four categories. If everything goes well, the patient should get back home. Hip fracture is a serious condition for the elderly and it can be a starting point or catalyst also to other problems which may result in the need of residential or hospital care. Even fatal complications may follow hip fracture operation. This categorization was chosen, since the event types comprised the events of interest from the point of view of the original problem and even more importantly these life events are recorded in various registers.

For generality the system formulation also allows that there are multiple events after hip fracture surgery. As a matter of fact the pathways of care are more interesting than single events in situations, such as the cost-effectiveness evaluation. Only death in system formulation is an absorbing state, all other states can be followed by any other state, i.e. only the death state ends a path in the system. Moreover, even though these four states are distinct, they are not necessarily independent. For example death may be more probable after complication or it may not be very likely to get home if there is a decision for long term residential care. The absorbing death state causes even more problems, because it is not possible to say for any individual that there is an increased risk of complication after death. However, it is possible to speculate what would have happened, if death had not occurred. All in all, this kind of quite simple system definition seems to result in an extremely complicated competing risks model and more simplifications are needed in the actual modeling.

Understanding and preprocessing data. A cohort of patients with hip fracture in 1998 or 1999 were identified in the Finnish Hospital Care Register using a simple diagnosis group abstraction (all patients with at least one ICD-10: S72 diagnosis in 1998 or 1999). Using the unique person identity codes of the patient cohort, data on all inpatient and outpatient hospital care and deaths for this cohort were obtained from the Finnish Health Care Register, data warehouse of the Finnish Hospital Benchmarking Project and the National Causes of Death Register. The results of these straightforward database queries were integrated into a new data set containing 167 952 records for 17 099 patients.

Each record in this data set corresponds to one care episode in hospital (or death), not any actual event of the system, i.e. each observation includes information, such as patient and hospital ID-numbers, age, sex, area codes, diagnose and operation codes as well as dates of admission and discharge (or death). Data cleaning was performed

in order to correct impossible simultaneous hospital episodes, systematic errors in the use of symptom vs. cause diagnoses and some missing or erroneous attribute values in area codes.

Many types of censoring occurs in this data set: the first hip fracture operation (or other important event) can be outside the observation window, some hospital episodes may have begun before 1998, follow-up ends in the end of 1999 (there is census data available for the last day of every year in the Finnish Health Care Register) and the death may end the follow-up of a patient.

Operation codes corresponding to hip fracture surgery were abstracted into two different operation types. Using this and the diagnosis group abstraction of hip fracture, hip fracture operations were identified from the data. Since the state preceding hip fracture operation was also important, the histories of the patients were traced backwards and the preceding state was classified to be home or residential/hospital care using more complicated temporal data abstraction.

The forward direction abstractions were even more complicated and all event types needed special abstractions and techniques. For example, the acute complication events were identified using interesting pattern remapping, in which all clinically relevant complication diagnoses were remapped to one event type.

Modeling, evaluation and reporting. For the statistical modeling purposes it was assumed that any acute complication event after hip fracture operation is an outcome which reflects the effectiveness of the surgical treatment. In addition, deaths and the upper limit of the observation window were considered to cause censoring to the event of interest. With these assumptions the modeling reduced to standard survival analysis (see e.g. [29]) where the variables of interest are time between hip fracture operation and a complication or censoring event, and the censoring indicator. These variables were calculated for all patients over 60 years, who had been living at home before surgery. The final preprocessed data set had 8824 records, each containing relevant variables for one patient.

Since the outcome was an acute complication, the hazard function of acute complication occurrences was estimated. According to the hazard function, the probability of acute complications was higher during the first 30 day period after surgical operation. This finding based on the data corresponded to the domain knowledge and gave some evidence that the data abstraction was done in a proper way.

In spite of the fact that time dimension includes a lot of information, the actual effectiveness indicator related to acute complication was defined to be the rate of acute complications during the 30 day period after the surgical operation. The 30 day cumulative risk of acute complications obtained from the distribution function of censored event times was found to be suitable estimator for the rate (see e.g. [29], 48-52). The estimated 30 day acute complication rate was 13 % (95% CI: 12.3%-13.7%).

The most useful information was obtained, when rates were evaluated for health care providers, such as hospitals or hospital districts. This kind of profiling analyses allowed comparisons of effectiveness between the providers. In order to adopt more realistic, medically based criteria for judging the performance and improved case-mix adjustments, more sophisticated profiling analysis methods, such as control charts [48] and hierarchical multilevel Bayes-models [49], were used [46], [47].

4. Conclusions

In this paper, some new methodological approaches for the utilization of administrative registers in research and in production of statistical information were presented. This was done by combining ideas mainly from the fields of statistics and data mining. Since the most important contribution of these ideas lies in understanding of phenomena, problems and data in knowledge discovery processes, the ideas were described as steps in the statistical knowledge discovery process.

The first part in this process was the understanding of the problem. There were three main points in that part: the problem solving always needs domain knowledge, a systems approach generates a common language for the experts of different fields and event history framework is a well developed choice for outlining the characteristics of dynamic phenomena.

The second part was about the understanding of data. The key ideas were: Second hand register data has a form of a generalized event sequence, which is identical to the structure of event history data, and implicitly defines possible model types for problem solving. In practice, data are always censored, but they still give a glance to the data generating process, which have an intuitive interpretation in the event history framework. Moreover, a wide variety of suitable data structures for traditional models can be yielded from the generalized event sequence as special cases using the presented time transformation property.

In the third part, data preprocessing was discussed. Some suggestions of suitable tools were given and a pattern remapping technique was formulated. The ultimate message was that sophisticated preprocessing incorporating the non-technical domain knowledge in order to scale things down to a size fit for statistical analyses is the most important and time consuming part in the register based data analysis.

Finally the ideas were illustrated with a practical example. It was seen that register based data analysis becomes very complicated and challenging even in simple situations. The perspectives presented in this paper propose many practically useful approaches to solve fundamental problems encountered in the analysis of massive second hand data sets.

All in all, every data analyst should remember that "statistics (data) are not collected, but produced; research results are not findings, but creations" [50].

References

1. Delasie, L., Kastelein, A., van Merode, F, Vissers, J.M.H. (eds.): Managing Health Care under Resource Constraints. European Journal of Operational Research 105 (1998)
2. Shortliffe, E.H., Perreault, L.E. (eds.): Medical Informatics : Computer Applications in Health Care and Biomedicine, Second Edition. Springer-Verlag, New York (2001)
3. Nykänen, P.: Decision Support Systems from a Health Informatics Perspective. PhD thesis, Department of Computer and Information Sciences, University of Tampere, Acta Electronica Universitatis Tamperensis 55, Tampere (2000)
4. Shani, M.: The impact of information on medical thinking and health care policy. International Journal of Medical Informatics 58-59 (2000) 3-10

5. Nenonen, M., Nylander, O.: A Theoretical Framework for Health Information Systems. Themes 3/2001. National Research and Development Centre for Welfare and Health (STAKES), Helsinki (2001)
6. Dudewicz, E.J., Mishra, S.N.: Modern Mathematical Statistics. Wiley Series in Probability and Mathematical Statistics, John Wileys & Sons, New York (1988)
7. Friedman, J.H.: The Role of Statistics in the Data Revolution. Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Book 1 (1999) 121-124
8. Hand, D.J.: Data Mining : New Challenges for Statisticians. Social Science Computer Review 18 (2000) 442-449
9. Han, J. & Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Fransisco (2001)
10. Huber, P.J.: Massive Datasets Workshop: Four Years After. Journal of Computational and Graphical Statistics 8 (1999) 635-652
11. Fayyad, U., Haussler, D., Stolorz, P.: Mining Scientific Data. Communications of the ACM 39 (1996) 51-57
12. McCarthy, J.: Phenomenal Data Mining: From Data to Phenomena. SIGKDD Explorations 1 (2000) 24-29
13. Cios, K.J. (ed.): Medical Data Mining and Knowledge Discovery. Studies in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg (2001)
14. Goodall, C.R.: Data Mining of Massive Datasets in Healthcare. Journal of Computational and Graphical Statistics 8 (1999) 620-634
15. Katz, B.P. (ed.): Measuring Quality, Outcomes, and Cost of Care Using Large Databases. Annals of Internal Medicine 127 (1997) Number 8, Part 2
16. Lavrac, N.: Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16 (1999) 3-23
17. Lavrac, N., Keravnou E., Zupan B.: Intelligent data analysis in medicine. In: Encyclopedia of computer science and technology, Vol. 42 (Supp. 27), 113-157, Marcel Dekker, New York (2000)
18. Øhrn, A.: Discernibility and Rough Sets in Medicine: Tools and Applications. PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, NTNU report 1999:133, IDI report 1999:14, Trondheim (1999)
19. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39 (1996) 27-34
20. Brachman, R.J., Anand, T.: The Process of Knowledge Discovery in Databases. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)
21. Chapman ,P., Clinton, J., Kerber, R., Khabaza, T. Reinartz, T., Shearer, C. Wirth, R.: CRISP-DM 1.0 : Step-by-step data mining guide. The CRISP-DM consortium (2000) http://www.crisp-dm.org/
22. Phannkuch, M., Wild, C.J.: Statistical Thinking and Statistical Practice: Themes Gleaned from Professional Statisticians. Statistical Science 15 (2000) 132-152
23. Allison, P.D.: Event History Analysis - Regression for Longitudinal Event Data. Quantitative Applications in the Social Sciences series 46, Sage Publications, Beverly Hills (1984)
24. Blossfeld, H.-P., Rohwer, G.: Techniques of Event History Modeling: New Approaches to Causal Analysis. Lawrence Erlbaum Associates, Mahwah (1995)
25. Clayton, D.G.: The Analysis of Event History Data: A Review of Progress and Outstanding Problems. Statistics in Medicine 7 (1988) 819-841
26. Lancaster, T.: The Econometric Analysis of Transition Data. Econometric Society Monographs 17, Cambridge University Press, Cambridge (1990)
27. Yamaguchi, K.: Event History Analysis. Applied Social Research Methods Series 28, Sage Publications, Newbury Park (1991)
28. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York (1980)

29. Cox, D.R. & Oakes, D.: Analysis of Survival Data. Chapman and Hall, London (1984)
30. Moen, P.: Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining. PhD Thesis, University of Helsinki, Department of Computer Science, Series of Publications A, Report A-2000-1, Helsinki (2000)
31. Arjas, E.: Survival Models and Martingale Dynamics. Scandinavian Journal of Statistics 16 (1989) 177-225.
32. Janssen, J. (ed.): Semi-Markov Models: Theory and Applications. Plenum Press, New York and London (1986)
33. Janssen, J., Limnios, N. (eds.): Semi-Markov Models and Applications. Kluwer Academic Publishers, Dordrecht (1999)
34. Howard, R.A.: Dynamic Probabilistic Systems, Volume I: Markov Models, Volume II: Semi-Markov and Decision Processes. John Wiley & Sons, New York (1971)
35. Feller, W.: An Introduction to Probability Theory and Its Applications, third edition. John Wiley & Sons, New York (1968)
36. Ross, S.M.: Applied Probability Models with Optimization Applications. Dover Publications, New York (1970/1992)
37. Rahm, E., Do, H.H.: Data Cleaning: Problems and Current Approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 23 (2000) 3-13
38. Winkler, W.E.: The State of Record Linkage and Current Research Problems. U.S. Bureau of the Census, Technical Report 4/1999 (1999)
39. Alvey, W., Jamerson, B. (eds.): Record Linkage Techniques -- 1997. Proceedings of an International Workshop and Exposition, March 20-21, 1997, Arlington, VA. Federal Committee on Statistical Methodology, Office of Management and Budget, Washington DC (1997)
40. Mannila, H., Toivonen, H.: Knowledge Discovery in Databases: The Search for Frequent Patterns. Book Draft, Department of Computer Science, University of Helsinki (1998)
41. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A.I.: Finding Interesting Rules from Large Sets of Discovered Association Rules. In: Nabil, R.A., Bhargava, B.K., Yesja, Y. (eds.): Proceedings of 3th International Conference on Information and Knowledge Management (CIKM '94), ACM Press (1994) 401-407
42. Han, J., Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases. In: Proceedings of International Conference on Very Large Data Bases (VLDB'95), pages 420-431, Zurich (1995)
43. Sahar, S.: Interestingness via what is not interesting. In: Proceedings of KDD '99, San Diego, Ca, USA (1999)
44. Silberschatz, A., Tuzhilin, A.: What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Transactions on Knowledge and Data Engineering 8 (1996) 970-974
45. Mannila, H., Toivonen, H.: Levelwise Search and Border of Theories in Knowledge Discovery. Data Mining and Knowledge Discovery 1 (1997) 241-258
46. Rissanen, P., Sund, R, Nordback, I., Rousi, T., Idänpään-Heikkilä, U.: Lonkkamurtuman hoidon vaikuttavuuden rekisteriperusteinen mittaaminen ja vertailu (Register based measurement of effectiveness of surgical treatment of hip fracture, in Finnish). Aiheita sarja 2002. National Research and Development Centre for Welfare and Health (STAKES), Helsinki (2002)
47. Sund, R. et al.: Health Care Provider Profiling: Effectiveness of Surgical Treatment of Hip Fracture. Nordic Health Econometric Workshop, August 22, Helsinki, Finland (2002)
48. Adab, P., Rouse, A., Mohammed, M.A., Marshall, T.: Performance league tables: the NHS deserves better. British Medical Journal 324 (2002) 95-98
49. Marshall, E.C., Spiegelhalter, D.J.: Institutional Performance. In: Leyland, A.H., Goldstein, H. (eds.): Multilevel Modelling of Health Statistics. Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester (2001)
50. Irvine, J., Miles, I., Evans, J.: Introduction: Demystifying Social Statistics. In: Irvine, J., Miles, I., Evans, J. (eds.): Demystifying Social Statistics. Pluto Press, London (1979)