Intelligent Data Analysis 7 (2003) 497–500
IOS Press

Guest editorial

Mining official data

Paula Brito[a] and Donato Malerba[b]
[a]*School of Economics, University of Porto, Porto, Portugal*
E-mail: mpbrito@fep.up.pt
[b]*Dipartimento di Informatica, University of Bari, Bari, Italy*
E-mail: malerba@di.uniba.it

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 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 challenging methodological research, which is still in the initial stages.

This special issue includes six papers which constitute updated and extended versions of papers selected from those presented at the Workshop on Mining Official Data, chaired by the guest editors of this issue in Helsinki in August 2002. The workshop was organized under the auspices of the European project KDNet (The Knowledge Discovery Network of Excellence) and within the framework of the 13th European Conference on Machine Learning (ECML'02) and the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02).

Different directions can be distinguished in the approach of the problem of mining official data. In this issue, emphasis is placed on the following topics:

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. In this context, a European project, SPIN, has been developed to address problems concerning geo-referenciation. SPIN's

main objective is to offer new possibilities for the analysis of geo-referenced data. To this end a Spatial Data Mining system is developed which integrates a state of the art Geographic Information System and a Data Mining functionality in an open, highly extensible, Internet-enabled plug-in architecture. In this issue, three papers are related to the topic of geo-referenciation, and constitute contributions from the SPIN project. The first one (by Klösgen, May and Petch) addresses the problem of subgroup mining and presents an application of SubgroupMiner, which is an advanced subgroup mining system, partially embedded in a spatial database and dynamically linked to a GIS. The second paper is concerned with discovering spatial association rules, and proposes a method based on a multi-relational data mining approach (by Appice, Ceci, Lanza, Lisi and Malerba). The third paper (by Paaβ and Kindermann) uses Bayesian models to represent complex relations between geo-referenced variables and to allow the estimation of the inherent uncertainty of predictions.

Aggregated data. 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 a data processing problem that goes 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 new research field named "symbolic data analysis". Symbolic data analysis improved greatly with the European project "Symbolic Official Data Analysis System (SODAS)"; as the result of this project a software package SODAS was developed [1]. The ASSO project (2001–2003) aims to developing further the methodology and tools for symbolic data analysis. Diday and Esposito's paper presents an introduction to symbolic data analysis and introduces the reader to the Sodas Software.

Both SPIN and ASSO are projects on "data analysis and statistical modelling"[1] of the EPROS (European Plan of Research in Official Statistics) EU programme promoted by Eurostat, the European Institute of Statistics, which has the support of research in official statistics as integral part of its activities [2].

The two remaining papers (one by R. Sund and the other by Frutos, Mensalvas, Montes and Segovia) focus on methodological aspects of official data analysis and make propositions on how to successfully explore the information from different kinds of sources.

To give the reader a better insight of this issue, we summarize the main ideas behind the individual papers in the following.

The first paper, "Utilisation of Administrative Registers Using Scientific Knowledge Discovery", by R. Sund, presents a methodological framework for the utilisation of administrative registers in the creation of scientifically valid information. The main methodological issues are discussed, the emphasis being on the connections between problem, data and analysis in the case of massive secondary data sets. A structure based on an event-history framework is proposed to represent information, which is showed to be useful for handling dynamic phenomena. Sund refers to censoring and discusses preprocessing tasks. The ideas presented are illustrated on the basis of a case study within health services research.

In "Mining Census Data for Spatial Effects on Mortality", W. Klosgen, M. May and J. Petch describe a system for spatial data mining, *SubgroupMiner*, illustrating its features by an application to UK census data, in order to explain high mortality rates. It is emphasized that the analysis applies to aggregated data, at the ward level. The system is partially embedded in a spatial database where analysis is performed, so that no data transformation is necessary: the same data are used for analysis and mapping in a GIS. The

[1] NORIS (Nomenclature on Research in Official Statistics) classification.

system handles both numeric and nominal target variables, discretizing explanatory numeric variables and executing dynamically spatial and non-spatial joins. Similar subgroups are clustered together and a Bayesian network can be inferred. The spatial subgroups can be visualized on a map, due to the dynamic link to a GIS. The application shows how the system uses the links between attributive data, spatial objects and geographic data to find subgroups relevant for the target variable.

The problem of discovering spatial association rules, that is, association rules involving spatial relations on spatial objects, is addressed by A. Appice, M. Ceci, A. Lanza, F. Lisi and D. Malerba in the paper "Discovery of Spatial Association Rules in Georeferenced Census Data: A Relational Mining Approach". The authors overcome the limitation of Geo-Associator, which used a single-table representation model, by using a multi-relational approach, which takes into account object interactions. A new algorithm using ILP concepts is presented, which aims at extracting multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels, considered in census data. Data from a spatial database are pre-processed so as to be represented in a deductive database, with two main objectives: first, if data are pre-processed, calculations need not be repeated; second, the possibility of specifying background knowledge and domain specific knowledge, which will allow for the searching of patterns which could not be found otherwise. The method was successfully applied to UK census data on a problem concerning accessibility of an urban area. It allowed the discovery of human interpretable patterns constituting new knowledge for urban planners.

The paper "Bayesian Regression Mixtures of Experts for Geo-Referenced Data", by G. Paaβ and J. Kindermann, presents a study where Bayesian models are used to represent relations between geo-referenced variables, which allow for the estimation of the uncertainty of predictions. It constitutes a Bayesian variant of the *mixture of experts* model, which uses different submodules for partitioning the input space and local prediction (*expert models*). A Markov Chain Monte Carlo analysis is used to generate the models. The proposed methodology is illustrated by an application to data from Stockport (UK), in order to predict long-term illness from census statistics. It is shown how Bayesian models are able to capture relations in spatial data. The analysis of the derivatives of the outputs with respect to the inputs allows the evaluation of the influence of input variables. Two versions of mixture models are investigated, using "hard splits" and "soft splits"; they both produce the same predictive surface, but "hard splits" yield extreme values of the derivatives, corresponding to abrupt changes between regions, which makes this version less plausible.

In the paper "An Introduction to Symbolic Data Analysis and the Sodas Software" E. Diday and F. Esposito give the reader a comprehensive presentation of the principles and objectives of Symbolic Data Analysis and present the Sodas Software. Symbolic Data Analysis extends data analysis methods to more general data than those permitted in the classical tabular model, which are recorded in "symbolic data tables". The reader is introduced to the notion of "symbolic object" and its underlying structure, and it is shown how symbolic objects can be used to model individuals, classes of individuals or concepts. It is made clear that symbolic data analysis provides a natural framework for the analysis of aggregated data. The authors present some Symbolic Data Analysis methods and focus on the study of dissimilarity measures for symbolic objects. The Sodas software, developed within the framework of the European project "Symbolic Official Data Analysis System (SODAS)", is presented. The general aim of Sodas is to build symbolic data that summarize huge data sets, and then analyse the resulting symbolic data tables by symbolic data analysis methods. It is clear that this is still an open field with many possibilities for future research.

The paper "Calculating economic indexes per household and censal section from official Spanish databases", by S. Frutos, E. Menasalvas, C. Montes and J. Segovia proposes a methodology to combine

information from different official databases, which cannot be cross-referenced, in order to produce economic indexes. It constitutes a case study from Spanish official data, using Population and Housing Census Data, together with Family Expenditure Surveys (FES). The proposed methodology consists in grouping families surveyed in the FES, who are known to live close to each other, then determining a socio-economic description for each group and the average values of the economic indices per group. Based on this information, estimation models are determined for each index, using a non-linear technique, which estimate the index from the socio-economic descriptions. The models are then applied to censal sections for which socio-economic compositions can be calculated, so as to determine index values per censal section. The calculations are then extended to larger units, and their temporal evolution is studied. The methodology seems to be promising since the results obtained in the presented case study are quite good.

The papers in this special issue constitute only a small sample of recent research related to official data mining. However, they give the reader a sense of some of the most challenging problems that arise when tackling this kind of data. There is no doubt that this is a very promising research area, which will offer researchers a wide range of topics. We hope that this issue will draw added attention to this field and stimulate further research.

Acknowledgements

We gratefully acknowledge the work of T. Alanko, E. Diday, F. Esposito, P. Gomes, H. Papageorgiou, W. Klösgen, C. Marcelo, M. May, M. Noirhomme, M. Summa and I. Turton, who acted as program committee of the ECML/PKDD'03 workshop on "Mining Official Data" and helped us in reviewing the papers of this special issue.

References

[1] H.H. Bock and E. Diday (Eds), *Analysis of Symbolic Data*, Exploratory methods for extracting statistical information from comflex data, Series Study in Classification, Data Analysis and Knowledge Organization. Vol. 15, Springer Verlag, Berlin.

[2] *Epros: Progress Report R&D in statistics 1999–2000*, Monographs of Official Statistics. Luxembourg: Office for Official Publications of the European Communities, 2001