On the Application of Data Mining to Official Data

Hossein Hassani and Shahin Gheitanchi and Mohammad Reza Yeganegi

2008

On the Application of Data Mining to Official Data

Hossein Hassani[a,b*] , Shahin Gheitanchi[c] , Mohammad Reza Yeganegi[d]

[a] *Statistics Group, Cardiff School of Mathematics, Cardiff University, CF24 4AG, UK.*

[b] *Statistical Research and Training Center (SRTC), Tehran, Iran*

[c] *Department of Engineering and Design, University of Sussex, BN1 9QT, UK*

[d] *Faculty of Mathematical Science, Shahid Chamran University of Ahvaz, Ahvaz, Iran.*

Abstract

Retrieving valuable knowledge and statistical patterns from official data has a great potential in supporting strategic policy making. Data Mining (DM) techniques are well-known for providing flexible and efficient analytical tools for data processing. In this paper, we provide an introduction to applications of DM to official statistics and flag the important issues and challenges. Considering recent advancements in software projects for DM, we propose intelligent data control system design and specifications as an example of DM application in official data processing.

Keywords: Data mining, Official data, Intelligent data control system.

1 Introduction

In statistics, the term “official data” denotes the data collected in censuses and statistical surveys by National Statistics Institutes (NSI’s), 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. As statistical offices usually have many large data sets from various sources, it is important to assess the potential application of Data Mining (DM) for the re-use of all these data sets. Not only DM helps to find interesting and reliable results in the data that have been processed for many years, it can also help to reduce the number of separate data collection tasks that we now maintain.

It is not surprising that statistical offices have not previously utilized DM, as the main task of NSI’s is in data production and analysis is often out-sourced to different institutes. Furthermore, it seems that the idea of exploring a database with the objective of finding unexpected patterns or models is not familiar to official statisticians who have to answer precise questions and make forecasts. Statistical analyses are done generally if they can be repeated in a production framework. However, as far as NSI’s manage large databases on population, trades, agriculture and companies, there are certainly great potentialities

*Corresponding author. Tel: (029) 2087 4811; Fax: (029) 2087 4199.
E-mail addresses: hassanih@cf.ac.uk (H. Hassani)

in exploiting DM techniques on their data. Based on this reason some projects have been done and are being worked upon by several groups, which are supported by national or international statistical offices.

The public availability of large official data sets is considered as a new challenge for the DM community. Very few applications of data mining techniques in the discovery of new models or patterns in official data sets have been reported. Of course in NSI's there has always been used some of exploratory data analysis, or model choice algorithms. But it seems that there are few, if not none, known applications of DM techniques in the meaning of trying to discover new models or patterns in their databases by using the new tools described before (Saporta, 1998; Saporta, 2000).

The structure of this paper is as follows. Section 2 gives various definitions of data mining with respect to the literature. Applications of DM to official data, concepts, related topics, softwares and projects are discussed in sections 3 and 4. Intelligent data control system is presented in Section 5, and conclusions are drawn in Section 6.

2 Data Mining

Data mining can be considered as a set of automated techniques used to extract or previously unknown pieces of information from large databases. Data mining is not a business solution but simply the underlying technology. In technical terms, DM is described as an application of intelligent techniques such as neural networks, fuzzy logic, genetic algorithms, decision trees, nearest neighbor method, rule induction, and data visualization, to large quantities of data to discover hidden trends, patterns, and relationships (for more information, see for example, Han and Kamber, 2006; Wang, 2003; Hand *et al.*, 2001).

The term DM is not new to statisticians. It is a term synonymous with data dredging or fishing and has been used to describe the process of trawling through data in the hope of identifying patterns (Hand, 1998a). Some of the numerous definitions of DM, or knowledge discovery in databases (KDD) are as follows:

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information and structure from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analysing changes, and detecting anomalies (Mackinnon and Click, 1999).

Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects within, and if the database is a faithful mirror of the real world registered by it (Siebes, 1996). Data mining refers to using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous and it stands of low value, because of no direct application. It is the hidden information in the data that is useful.

There are two kinds of structures which are sought in DM activities: models and patterns (Hand, 1998b). Building models is a major activity of many statisticians and econometricians, especially in NSI's and it will not be necessary to elaborate too long on this. A model is a global summary of relationships between variables, which both helps to understand phenomena and allows predictions. A model is generally chosen on an a priori basis, based upon a simplifying theory. Exploration of alternative models is made feasible by DM algorithms. Data mining appears as a collection of tools that are presented usually in one package and apply several techniques on the same data set. DM algorithms offer extensive possibilities of finding models relating variables together. In contrast to the global description given by a model, a pattern is often defined as a characteristic structure exhibited by a few number of points. For instance, a small subgroup of customers with a high commercial value or conversely highly risked (Hand 1998a).

3 Application of Data Mining to Official Data

To develop successful applications of DM techniques to official data, the following issues must be dealt with:

3.1 Aggregated data

NSIs make a great effort in collecting census data, but they are not the only organizations that analyze 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).

It should be noted that a fruitful application of DM is to use the DM techniques that incorporate privacy concerns. There has been extensive research in the area of statistical databases motivated by the desire to be able to provide statistical information without compromising sensitive information about individual (see, for example, comprehensive surveys in Adam and Wortmann (1998) and Shoshani (1982))

3.2 Data Quality

According to Total Quality Management (TQM), data quality can be considered as consistently meeting customers' expectations. There are several aspects of data quality, like integrity, validity, consistency, and accuracy but we shall not go into details here. Data Quality Mining (DQM) can be defined as the deliberate application of DM techniques for the purpose of data quality measurement and improvement (Hassani and Anari, 2005; Hassani and Haeri Mehrizi, 2006a). The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases. There are many starting points to employ today's common DM methods for the purposes of DQM. Methods for deviation and outlier detection seem promising (Hassani and Haeri Mehrizi, 2006b). But it is also straight forward to employ clustering approaches and dependency analysis for data quality purposes. In addition, if we are able to supply training data prepared by a human then also classifiers might do a good job. It is even conceivable that neural networks and artificial intelligent utilized to recognize data deficiencies (Yeganegi *et al.*, 2006; Hipp *et al.*, 2000). Basically, we can classify the application of DM to improve data quality in these four important aspects (Hipp *et al.*, 2001):

1- measuring and explaining data quality deficiencies,
2- correcting deficient data,
3- extension of KDD process models to reflect the potentials of DQM,
4- development of specialized process models for pure DQM.

3.3 Timeliness

Timeliness 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.

3.4 Confidentiality

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. Therefore, in any public discussion it would be important to emphasize that DM is looking for patterns and statistical relationships in data, not for individuals and their details and their relationships. Thus DM 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 maximize the use of the data while keeping the original agreement with the data source (Nanopoulos and King, 2002).

3.5 Metadata

Mining official data implies retrieving knowledge 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 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 centralized metadata management. By means of homogeneously documenting data from different sources in a unique environment, a centralized metadata system provides the rough material for data integration (D'Angiolini, 2002).

4 Projects and Softwares

4.1 Analysis System of Symbolic Official data (ASSO)

ASSO started in January 2001 and ran for 36 months as part of the European Union fifth framework Research and Development program in the Information Society Technology strand. ASSO offers methods, methodology and software tools for the analysis of multidimensional complex data (numerical or non-numerical) coming from databases in statistical offices and administrations using symbolic data analysis. ASSO resulting software is called SODAS2 (SODAS2 software). It is an improved version of the SODAS software developed in the previous SODAS project following users requests. This new software is more operational and attractive. It also proposes new innovative methods and demonstrates that the underlying techniques meet the needs of statistical offices [1].

4.2 Knowledge Extraction for Statistical Offices (KESO)

KESO is an ESPRIT-IV project under Eurostat/DOSIS. The project was scheduled for three years and started on 1st Jan. 1996. The goal of the KESO project was to construct a versatile, efficient, industrial strength DM system prototype that satisfies the needs of providers of large-scale databases. The development will be guided by a continuous assessment of the participating public and private Statistical Offices both on the applicability and the added value for their complex datasets [2].

4.3 Spatial Mining for Data of Public Interest (SPIN)

The project was scheduled for three years and started on 1st Jan. 2000. The main objective of the SPIN project is to support statistical offices in their timely and cost effective dissemination of statistical data and to offer exciting new possibilities for the scientific analysis of georeferenced data. To this end a Spatial Data Mining System (SPIN) is developed. It integrates state of the art Geographic Information Systems (GIS) and Data Mining Systems (DMS) functionality in an open, highly extensible, internet-enabled plug-in architecture. By adapting methods from machine learning and Bayesian

[1] http://www.info.fundp.ac.be/asso/

[2] http://db.cwi.nl/projecten/project.php4?prjnr=77

Statistics to spatial data analysis the state of the art in DM will be advanced. The state of the art in GIS will also be advanced by developing new methods for the visualization of spatial and temporal information [3].

5 Intelligent Data Control System: design and specifications

In this section, we propose the design of an Intelligent Data Control System (IDCS) as an application of DM to the official statistics[4]. In order to improve the quality of the official data we follow a layered approach. A layered system benefits from the followings:

- Low complexity layers where result in solving complex problems.
- Higher adaptation and flexibility in order to achieve optimum performance.
- Enabling coexistence of current techniques and algorithms in different layers, provided having common interfaces among the layers.
- Robustness and stability because having easy debugging layers.
- High security because of providing layered access for different role players in the system.

The provided design is a generic model for improving the quality of official data. For specific problems and implementations, detailed study of the specification of the problem and the goals of the solution are required. Using artificial intelligence technique, the system is able to learn, behave and self organize the objectives of each task. Neural networks is a well-known technique which can benefit the IDCS model. This technique is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing. In most cases an neural network is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms neural networks are non-linear statistical data modelling or decision making tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

It is important to minimize the human supervision in order to reduce cost of data collection and save human resources while leaving the heavy processes to the advanced computerized systems. The proposed generic model enables optimization of layer in order to reduce the human supervision over the system while improving the quality of official data by minimizing the errors. Figure 1 shows the relation of human supervision and quality of official data over N layers of IDCS system.

[3] http://www.ais.fraunhofer.de/KD/SPIN/index.html

[4] The labor intensive work on the using this system are underway, and we hope to present those results in future in another paper.

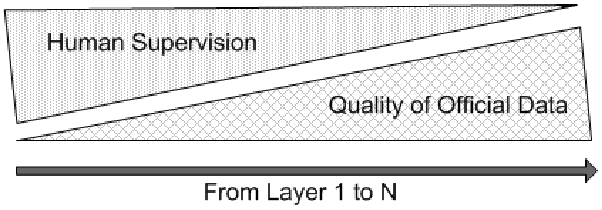

Figure 1: Relation between human supervision and quality of official data over N layers.

5.1 Design an intelligent system

Based on the mentioned properties for intelligent data processing system, we propose the following multi-layer architecture to reduce human supervision while minimizing the data errors throughout the system. Figure 2 shows general model of IDSC system.

Layer one:

The first layer is the input of the system and all information are entered to the system through layer one. This layer consists of all algorithms and techniques for revision and conceptual editing of the data. This layer, based on revision method and conceptual editing table, processes the data of each unit and finds the possible errors in the data set. The output of this layer is a queue of units data sets which may not be correct because of imperfectness of the intelligent processing of the system. We provide an example of an artificial neural network with very simple node to demonstrate how the intelligent system might be designed. Assuming a binary system, an artificial neuron with three inputs can be modelled as following

$$\text{Output of the neuron} \begin{cases} 1 & \text{if} \quad w_0 I_0 + w_1 I_1 + w_b > 0 \\ 0 & \text{if} \quad w_0 I_0 + w_1 I_1 + w_b = 0 \end{cases}, \tag{1}$$

where I_i and w_i $(i = 1, 2)$ are the i-th actual inputs and the weights of each input to the neuron and also w_b is the bias factor for controlling purposes. The neuron processes the weighted inputs and generates the output. By having multi-layer of cascaded neurons, complex behaviours and processes are performed.

Layer two:

This layer processes the output of the first layer. Based on the uncertainty about the correctness of the data in the queues (fed from first layer), a decision is taken to correct the data. After the processes of the layer, the questioners of some units may be returned for reconsideration. The data which are corrected in this way are collected and re-processed by the first layer. Furthermore, in case of using neural networks in the first layer, the errors could be used to adaptively modify the weights and bias factor to improve the decision making process of the first layer. Adaptive updating of the weight will increase the fault tolerance of the system by providing a neural network at each time which is formed based on the inputs. For example, in a single layer neural network, if most of the errors are originated from a particular input, the weight of that input can be modified accordingly.

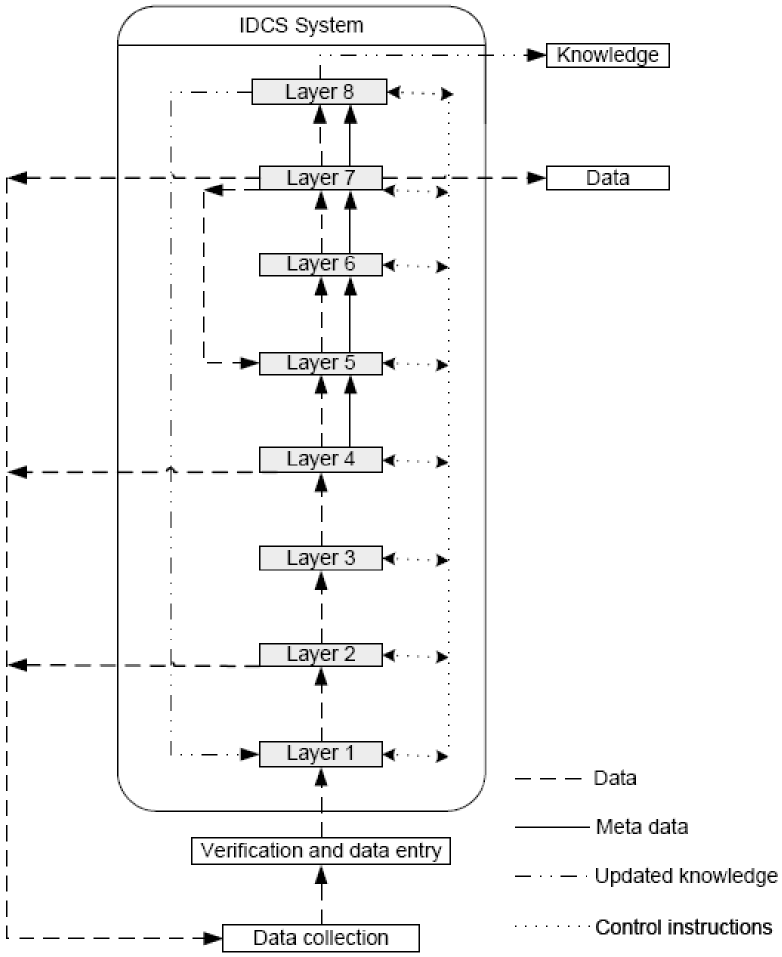

Figure 2: General model of IDCS.

Layer three:

In the previous two layers all the errors which were detectable by the current system were corrected. This layer is to make sure about the correctness of the data from the previous sections. It samples the data and processes them using simple techniques. The main technique used in this layer is combination of OLAP and a method to describe the structure of the current data set. OLAP is an approach to quickly provide answers to analytical multi-dimensional queries. Some parts of the errors which may have passed through the first two layers will appear by analyzing OLAP tables and the output of the data description method. It is obvious if there is any problem in registering the data, the output the mentioned combination technique will not be a logical structure. In short, this layer will report the data which the outputs of description statistics on OLAP tables don't have logical structure.

Layer four:

In this layer, by using the information from the third layer (which was for reducing the structural errors of the data sets) some of the errors are corrected. To correct the errors which could not be solved in this layer, the original data need to be collected again. The output of this layer is the corrected data, meta-data and information about the structure of data. Meta-data are the information which describe the actual data and are used to help the processes in other layers. Therefore, the format of meta-data should be understandable to the other layers which may use different systems other than the layer which has generated the meta-data. Extensible Mark-up Language (XML) is a common technique to implement meta-data flow. It should be noted the corrected data by the person who handles the original data, may also be wrong and the system can't detect the errors. Therefore, the error will remain in the data.

Layer five:

The fifth layer finds the hidden irregularities in the data. In this layer a DM program will process the database to find the errors which were not detected in the previous layers. This layer, by analyzing the situation of data vectors in each unit in the current data space tries to find the units which may have a high density data space. To perform this process, the layer uses clustering method for units. In clustering method, similar records in the data are grouped together in cluster forms which will result in having a high level view of data. The output of this layer is the values of units which they have a high density data and are in a single cluster. For this purpose, considering the common queues of units in each data-space state is very useful and informative. Alternative advanced DM techniques such as trees and networks (Pujari, 2001) could be used.

Layer six:

This layer, after receiving information from the fifth layer, analyses the reasons for units vector aggregation in each section of data. This analysis is possible based on the information provided from the previous layer. Number of units in a cluster, common queues of these units, number and type of queues are some examples of the information that can be processed. By knowing the reason for aggregation of data, we can divide the data to two categories. The first category is the data which are aggregated in a specific place of the data space for a logical reason. And the second category is the data which there is no logical reason for aggregation of them in the data space. In the second category we can consider the data which the reason of their aggregation is a common error. In the process of the reason for aggregation of data, error sources and common errors are available. This layer can generate meta-data for the category of data which have logical reason for aggregation. The meta-data is used in future processes. The output of this layer is categorization of the clusters from the fifth layer and a report for reasons of occurrence of the clusters. In case of the clusters which the error has occurred in the gathering of data, the report includes resources for occurrence of errors and the problems which are caused in the data by these errors. Also it includes the meta-data for clusters with logical reasons for their existence.

Layer seven:

This layer is a control and correction layer to monitor the generated clusters in previous layers. Two sources of error may introduce false data to the system which result in generating incorrect cluster groups. The first source of data is the mistakes which are made in the data collection process and may happen by the person who collects data. The other source of errors is the imperfectness of the intelligent system in previous layers which could be reducing by adaptively training the system over time. The algorithm in the layer tried to correct the errors based on the reports and meta-data from the previous layers. At this point some questioners may be returned for correction.

Layer eight:

In the layers introduced so far, meta-data and new data has been generated. The last layer makes use of these data to monitor and correct possible errors from previous layers. The most important feature of this layer is to make use of meta-data produced in the process to correct the system. To make the system independent and self-correcting, we need to convert the information to knowledge and meta knowledge. The most important information which can be used to reduce the probability of not detecting a wrong data are the information and meta-data generated in layer five. This information includes reasons for an error, mistakes in the data set which is made by errors and reasons for aggregation of errors in a special situation. With the use of this information the previous knowledge and the questioners are reviewed. Furthermore, the data enables to monitor the behaviour of system itself for improvement and optimization. Therefore, this layer has two important roles: First is to consider the changes within the system to keep it optimized (process control). Second is to generate knowledge based on the meta-data and information which is generated during the process and use the generated knowledge in the lower layers.

Finally, some important points about the described intelligent system are listed as following:

1. In this system after passing through the fifth layer, each time some parts of data are corrected. All the data (even the parts which did not need any correction) are required to return to the fifth layer. The reason is the fifth layer considers the data with their relational situation and then generates the output. Therefore every change in the data may lead to change of the situation. In other words, after any changes in the data, all the data are required to be returned back to its previous layer.

2. The mentioned system is an intelligent organization which achieves a distributed intelligence by distributing the knowledge among all the steps of the process. In other words, the intelligence of IDCS relies on the aggregation of the distributed intelligence. Therefore, data flow and distribution of knowledge in the system are important features of the system. By improving the flow of data and knowledge, the system will become more intelligent.

3. According the system architecture, dynamic nature of IDCS shell enables the system be a combination of the human resources and the intelligent processing techniques. This will also inter-operability of the proposed architecture with potential available systems used for this purpose.

4. By implementing the system based on intelligent techniques, lots of time and energy are saved by reducing number of checks on questioners while keeping the quality and automating the system. It is obvious that the IDCS saves time and resources.

5. In the first layer, knowing that inputting data using keyboard may cause errors (although the data are verified after they are entered), using data entry software can improve the efficiency and precision of the system. Based on the accuracy of the data entry software, verification of data may not be needed. One of the recommended methods of data entry is to use computer-added data collection systems.

6. In the first look IDCS looks to be not a practical system. In fact, IDCS introduces an intelligent platform which can be implemented using the current systems. However, since the last layer is designed for ideal situation, it needs to be changed according to the implementation requirements. But the other layers remain un-changed because the algorithms for those layers already exist and they just need to be revised to be used for IDCS system.

6 Conclusion

Data mining (DM) is a growing discipline which originated outside statistics in the database management community, mainly for commercial concerns. Data mining can be considered as the branch of exploratory statistics where one tries to find new and useful patterns, through the extensive use of classic and new algorithms. To avoid drawing wrong conclusions and statisticians, considering uncertainty and risk, an automated process of knowledge extraction, validation and correction, called IDCS, is introduced. However, even with very efficient software, human expertise and interventions are necessary. Official statistics benefits from using DM, however, it may imply a redefining the mission of NSI's. By introducing IDCS, we have shown that DM can be beneficial for official statistics which requires data processing over huge data bases.

Acknowledgments

The authors would like to thank the editor and the referees for their constructive comments which have led to substantial improvements in this paper. We extend our appreciation and gratitude to our colleagues in the Statistical Centre of Iran, particulary Mr M. Anari and the Bureau of Statistical Definitions and Standards, for their help during the research.

References

[1] Adam, N. R. and Wortmann, J. C. (1998). Security-control methods for statistical databases: A comparative study, *ACM Computing Surveys*, 21(4), pp. 515–556.

[2] D'Angiolini, G. (2002). Developing a Metadata Infrastructure for Official data: the ISTAT experience.
http://www.di.uniba.it/ malerba/activities/mod02/pdfs/dangiolini.pdf

[3] Han, J. and Kamber, M. (2006). Data Mining *Concepts and Techniques*, Second Edition. Elsevier Inc.

[4] Hand, D. J., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining.* MIT Press.

[5] Hand, D. J. (1998a). Data mining: statistics and more?.*The American Statistician*, **52**, pp. 112–119.

[6] Hand, D. J. (1998b). Data mining-reaching beyond statistics, *Research in Official Statistics*, **2**, pp. 5–17.

[7] Hassani, H. and Haeri Mehrizi, A. (2006a). Data Mining and official statistics, *Journal of Statistical Centre of Iran* , vol **67**, No 4, pp. 21–34.

[8] Hassani, H. and Haeri Mehrizi, A. (2006b). Data Mining & Official Data, *In Proceeding of the 8th Iranian International Statistics Conference*, Shiraz, Iran, pp 61–68.

[9] Hassani, H. and Anari, M. (2005). Using Data Mining for Data Quality Improvement. *In Proceedings of the 55th session International Statistical Institute (ISI)*, Sydney, Austeralia.

[10] Hipp, J. Gntzer, U. and Grimmer, U. (2001). Data Quality Mining - Making a Virtue of Necessity. *Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001)*, Santa Barbara, California, pp. 52–57.

[11] Hipp, J. Guntzer, U. and Nakhaeizadeh, G. (2000). Mining association rules: Deriving a superior algorithm by analysing today's approaches. *In Proceedings of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '00)*, pp 159–168, Lyon, France.

[12] Mackinnon, M. J. and Glick, N. (1999) Data Mining and Knowledge Discovery in Databases - An Overview, *Australian* & *New Zealand Journal of Statistics* , **41** (3), pp. 255-275.

[13] Nanopoulos, Ph. and King, J. (2002). Important Issues on Statistical confidentiality.
http://www.di.uniba.it/ malerba/activities/mod02/pdfs/nanopoulos.pdf

[14] Pujari, A. K. (2001). Data Mining Techniques, Universities Press, Hyderabad, India.

[15] Saporta, G. (2000). Data Mining and Official Statistics. *Quinta Conferenza Nationale di Statistica, ISTAT, Roma.* http://cedric.cnam.fr/PUBLIS/RC184.pdf

[16] Saporta, G. (1998). The Unexploited Mines of Academic and Official Statistics, in Academic and Official Statistics Co-operation, *Eurostat*, pp. 11–15.

[17] Shoshani A. (1982). Statistical databases: characteristics, problems, and some solutions. *In Proc. of the international conference on very large data bases (VLDB)*, pp. 208-222.

[18] Siebes, A. (1996). Data Mining: What it is and how it is done, *SEBD*, pp. 329–344.

[19] Wang, J. (2003). *Data Mining: Opportunities and Challenges.* Idea Group Publishing.

[20] Yeganegi, M. R. Hassani, H. and Haeri Mehrizi A. (2006). Artificial Intelligence and Its Application to Official Statistics , *In Proceeding of the 8th Iranian International Statistics Conference*, Shiraz, Iran, pp 120–132.