# Data Mining, Machine Learning and Official Statistics

Gilbert Saporta and Hossein Hassani

**Abstract** We examine the issues of applying Data mining and Machine Learning techniques to official data, and try to explain their underuse in National Statistical Institutes.

**Key words:** data mining, machine learning, official statistics, NSI

## 1 Introduction

Data mining (as statisticians call it) or knowledge discovery (as computer scientists prefer to label) has developed rapidly over the last two decades and is becoming increasingly significant in the assemblage of official statistics.

Despite the fact that data mining is being utilized and introduced in many different fields ranging from astronomy to chemistry, there is little or no evidence to suggest that it is being fully exploited in the analysis of official statistics for identifying new patterns or models. Saporta (1996) and Saporta (2001) stated that there exist only a few if not any reported applications of data mining in official statistics in the meaning of trying to discover new models or patterns in their databases. Since that, even by the year 2010 there was very little change in this regard as only a minimal application of data mining techniques in official data were reported (Hassani et al; (2010)), and as Letouzé (2012) emphasizes, it is indeed opportune to use data mining to supplement official statistics in order to gain richer and deeper insights.

Gilbert Saporta, CEDRIC-CNAM, gilbert.saporta@cnam.fr
Hossein Hassani, Bournemouth University, hhassani@bournemouth.ac.uk

The minimal applications of data mining for official statistics are not entirely surprising for the following reasons: first, National Statistical Institutes (NSIs) are tasked with data collection, while a common practice for many NSIs has been to outsource the analysis; second, the objective of official statisticians is to answer precise questions and make forecasts as opposed to finding unexpected patterns or models.

Witnessed today is an augment in the recognition of the prolific importance underlying the application of data mining to official statistics. A sound example is the introduction of a workshop on data mining in official statistics to the program at the 2012 SIAM International Conference on Data Mining. Through this conference the organizers endeavored to create synergies by bringing together statisticians working with official data and data mining specialists who have expressed an interest in this field.

In addition, the NTTS 2013 Conference on Research in Official Statistics, which is organized by Eurostat, offered a workshop on big data and data mining. These two conferences are evidence that assessing the application of data mining in official statistics is now considered imperative. It is also worth mentioning that in 2002 a workshop on mining official data was held within the framework of the 13th European Conference on Machine Learning and the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases. Nonetheless, there still remains a scope for vast amount of research work to be conducted in this field.

Big data is now a major concern for a large number of industries dealing with huge amounts of data and data streams (consumer analytics, health care, retail, etc.). No doubt it will last for a decade and official data will play an important role in this trend.

The aim of this article is to essentially review the work relating to the application of data mining in official statistics and to analyze why DM and Machine Learning techniques are underused in NSIs.

A large part of the material presented here comes from Saporta, Hassani and Silva (2014)[1] which provides a review of almost all published articles associated with the application of data mining in official statistics.

The remainder of this article is organized as follows: the section titled “Definitions” takes a look at the definitions of data mining, and its evolution with the growth of Machine Learning methodologies.The section “Why should Data Mining be applied to Official data” describes the need for applying data mining to official statistics and gives a short review of successful applications over the last two decades. Section 4 lists the issues that must be addressed when applying data mining in official statistics.The following section is about factors explaining the underuse of Data Mining and Machine Learning at NSIs. Section 6 is about Big Data, and the article wraps up in the Conclusion section.

[1] ‘‘Big Data. Copyright 2013 Mary Ann Liebert, Inc. http://liebertpub.com/big, used under a Creative Commons Attribution License: http://creativecommons.org/licenses/by/3.0/us/’’

## 2 Definitions

Prior to addressing the concerns of this paper it is pertinent to understand key definitions relating to Data Mining and official statistics. It is important to note at this juncture that, the definitions of Data Mining differ according to the problem or industry in which it is being applied. This in turn has resulted in the definitions evolving and developing continuously over the years. Moreover, it will be clear from the definitions outlined below that the classical definitions of Data Mining stress upon exploratory (unsupervised) Data Mining whilst the more modern definitions appreciate the emergence of supervised Data Mining or Machine Learning.

According to Wikipedia, Machine learning and data mining are commonly confused, as they often employ the same methods and overlap significantly. A convenient distinction should be the following:

Machine learning focuses on prediction, based on known properties learned from the training data and corresponds to supervised Data Mining while Data mining focuses on the discovery of (previously) unknown properties in the data.

In the 1990's Data Mining was defined as:

"The search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data". (Siebes 1996)

"The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." (Fayyad et al. 1996, p. 40)

And,

"The process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners." (Hand 1998, p.112)

The new millennium has seen a simplification of the definitions of Data Mining and the emergence of supervised Data Mining. Accordingly the following definitions arise out of the literature whereby Data Mining is defined as;

"The process of unearthing unexpected, valuable, or interesting structures or patterns in large data sets." (Hand 2000, p. 443)

"Finding statistically reliable, previously unknown and actionable insights from data" (Elkan 2001).

On the other hand the objective of official data (which is used to produce 'official statistics') is to "accurately count a state's resources" (Saporta 1998; Brito and Malerba 2003). Thus, official data has been defined as:

"Data collected in censuses and statistical surveys by National Statistical Institutes (NSIs), as well as administrative and registration records collected by government departments and local authorities." (Hassani et al. 2010, p.75)

Given the above definitions, the application of Data Mining to official data can be defined as:

"Retrieving data from different surveys or administrative sources and properly interpreting them as measures of observed phenomena." (D'Angiolini 2002, p.1)

However, it should be noted that Data Mining is not concerned with efficient methods for collecting data such as surveys and experimental designs (Hand, 2000). Furthermore, models do not come from a theory, but from data exploration: they are "data driven". As such, Data Mining is not concerned with estimation and tests of pre-specified models, but with discovering models through an algorithmic search process exploring linear and non-linear models, explicit or not.

## 3 Why should Data Mining be applied to Official Data

Based on past literature, it is possible to identify many reasons that warrant the application of data mining to official data. First, as mentioned above, data mining employs specific tools to uncover hidden information in mountains of data, which is otherwise left invisible to the human eye. Official data relate to a variety of subjects and is utilized only for a specific purpose. Consequently, this leaves NSIs with large, untapped, and unexplored databases, and traditional techniques are not optimal for analyzing them. Data mining emerges as an essential tool as it has the potential to exploit such large databases by identifying relationships and discovering patterns that would otherwise remain unnoticed.

Third, the availability of large data sets (official data) is a resource for data mining, and the development of data mining itself is closely linked to the availability of such large databases. Therefore, it is evident that positive synergies would emerge benefiting both data mining and NSIs through the introduction and application of data mining to official statistics. Nevertheless, it is notable that in more recent work, Hassani et al. (2010) identify the availability of large data sets as a challenge for data mining as opposed to considering it a resource. It appears more accurate to label this increasing availability of large data sets as a ''resourceful challenge'' for data mining, because the availability of such large data sets provides positive benefits through the challenges it creates.

Additionally, through the work of Saporta, data mining is identified as an existing tool that is underused in official statistics, and it is emphasized that NSIs could profit by mining their large databases on agriculture, trade, population, and so on. Furthermore, the purpose of data mining is to find models, be it linear or nonlinear, and patterns in data. This is exactly in line with the main responsibility of statisticians employed by NSIs, which is to build models, but not exactly the same kind of models as we will see later on. Brito and Malerba (2003) add to this discussion by noting that public policy, which is the backbone of a democratic society, could benefit largely through the application of data mining to official data.

Finally, as mentioned by Glasson et al. (2013), traditional statistical methods have trouble handling big samples and are unlikely to be fast enough when faced with the increasingly available big data found at NSIs. Accordingly, data mining techniques are mandatory in order to swiftly uncover information from big data.

Hassani et al. (2014) present a comprehensive review of applications of Data Mining to Official Data classified according to the algorithms used (bayesian regression, cluster analysis, decision trees, neural networks etc.) and three main categories: (i) third parties mining published official statistics, (ii) third parties mining data collected for producing official statistics and (iii) official statisticians using data mining techniques (as opposed to current statistical analyses) for producing official statistics.

It was clear from this grouped analysis that majority applications of Data Mining in Official Statistics has been from third parties mining data that has been collected for producing official statistics (mainly census data). Furthermore, third parties have explored a variety of Data Mining techniques for official data in comparison to the research by Government Statisticians which has mainly focussed on decision trees and cluster analysis alone. It is also evident that decision trees are the most popular data mining technique amongst NSIs at present. This increased application of decision trees by NSIs is not astonishing as polls by Kdnuggets indicate that decision trees are the most widely used data mining technique..

## 4 Some issues for successful applications

Hassani et al. (2010) outlines issues that must be addressed to achieve a successful application of Data Mining to Official Data.

**Aggregated Data**: The law strictly prohibits NSIs from publishing or releasing individual responses due to privacy concerns (Klosgen and May 2002; Brito and Malerba 2003; Hassani et al. 2010). As a consequence, NSIs are legally bound to aggregate data prior to releasing them to any external authority. Hassani et al. (2010) states that aggregated data presents a challenge for data analysts because the data would concern more or less homogenous classes or groups of individuals (macro data or second-order objects) as opposed to single individuals (micro data or first-order objects). “Symbolic data analysis” was introduced in order to overcome the challenges imposed by aggregated data (Diday and Esposito 2003; Brito and Malerba 2003; Frutos et al. 2003). Following its introduction, Eurostat pioneered and initiated various projects for developing symbolic data analysis further as it proved to be indispensable given the legal constraints. Three fine examples of such projects were the Symbolic Official Data Analysis System (SODAS) project which resulted in the SODAS software, the Analysis System of Symbolic Official data (ASSO) Project 2001-2003 which developed the associated methodology and tools further and the Spatial Mining for Data of Public Interest (SPIN). Another project worthy of recognition is Knowledge Extraction for Statistical Offices (KESO) which was initiated in 1996 under Eurostat and DOSIS.

**Timeliness**: The stated objectives of most NSIs require them to provide the public with timely statistics (Cheung 1998). However, we live in a world where public and private sector institutions are continuously urged to reduce the time lag between

data collection and decision making (Hassani et al. 2010). Therefore, timeliness becomes yet another important issue which needs to be addressed. For example, Miller et al. (2009) shows that the National Agriculture and Statistical Service in the USA is researching and developing Data Mining as a source of disseminating 'timely' official statistics, whilst Klucik (2011) shows that genetic programming as a Data Mining tool can be used for improving the timeliness of NSIs data collection and publication.

**Confidentiality**: Official data is collected following a guarantee of confidentiality for the informant which is also a legal requirement. Therefore, any Data Mining techniques which are adopted must ensure that people or companies which provided the data are not recognized or publicized. However, at the outset, Data Mining and Big Data Analytics appear to be the complete opposite of protecting the confidentiality of official statistics. In addition, it is suggested that as confidentiality is crucial, NSIs should carry out the Data Mining work on official statistics (Saporta 2000). However, according to Sumathi and Sivanandam (2006), exploratory Data Mining tools are able to expose sensitive and confidential facts about individuals; for example, link analysis is able to correlate phone and banking records to determine which customers have a fax machine at home. Whilst this is not good for the confidentiality of official data sources, the ability to narrow down possibilities has proven to be greatly helpful in criminal investigations not only in terms of counterterrorism (Fienberg 2005), but also cost reductions and efficient resource allocations. In order to overcome this situation, Data Mining now uses security control mechanisms known as query restriction or noise addition, to prevent the revelation of confidential individual information whilst safeguarding the data quality (Sumathi and Sivanandam 2006). In addition data perturbation and secure multiparty computation is also used to overcome privacy and confidentiality related issues (Vaidya and Clifton 2004).

**Metadata**: Metadata refers to the descriptions of the meaning and context of the data (Hand 1998). More simply stated, it is data regarding the data itself (Sumathi and Sivanandam 2006). Mining official data requires retrieving data from various surveys or administrative sources and correctly construing them as measures of observed phenomena (D'Angiolini 2002; Hassani et al. 2010). Early into the millennium, Saporta (2000) identified that text mining could be used to analyse metadata information.

However, over a decade since the millennium, introducing metadata management practices in official data production continues to be a challenge regardless of the fact that ensuring the dissemination of such metadata to the end users is a primary task of NSIs (D'Angiolini 2002; Hassani et al. 2010). Moreover, the increasing need for integrating data from several sources obliges the NSIs to practice a policy of centralized metadata management. A centralized metadata system is one which is able to provide the rough material for data integration by means of homogenously documenting data from different sources in a unique environment (D'Angiolini 2002; Hassani et al. 2010).

## 5 The underuse of DM and ML at NSIs

There is substantial evidence that Government Statisticians are trapped in traditions which limit their exposure and willingness to exploit and explore lucrative and novel Data Mining techniques which can improve and enhance their efficiency and quality of information provided through official statistics. The answer to this challenge is to reiterate the call for increased engagement, co-operation and collaboration between data scientists and official statisticians. Such collaboration will undoubtedly encourage Government Statisticians to consider the usage of novel Data Mining techniques whilst creating synergies which will enhance the quality of official statistics published by NSIs and greatly improving the rate of methodological advances in Data Mining techniques.

Yet another interesting observation is in applications of Data Mining by Government Statisticians for producing official statistics. All applications of Data Mining that have been reported in this context emanate from US government agencies and we could not find a single application of this sort from any government organization in the UK or from INSEE (the French National Statistical Institute). We have scanned from 1991 to now the proceedings of the "Journées de Méthodologie Statistique" which is a major event gathering hundreds of methodologists and official statisticians from INSEE and there is not a single reference to neural networks, data mining, SVM, Lasso, decision trees etc.

If we consider that the Journal of Official Statistics correctly reflects the state of research in Official Statistics, one may be surprised to find only one paper (Grim et al. 2010) with "Data Mining" as a keyword since 1985. However the method used (mixture distribution) is a classical one in mathematical statistics. The same is true for the Statistical Journal of the IAOS.

However, it would be of course incorrect to assume that NSIs do not perform any exploratory data analysis and forecasting. The problem lies within the fact that NSIs rely on traditional methods as opposed to exploiting the novel Data Mining and Machine Learning techniques that are at their disposal today. NSIs seldom use emblematic Data Mining techniques such as association rules, neural networks or support vector machines. In fact, we could not find a single publication relating to the application of support vector machines in official data whilst only one application of machine learning could be found (by a third party).

We are of the opinion that statisticians at NSIs are reluctant to use these modern Data Mining techniques because they prefer models which can be written as simple equations in closed forms, and not like predictive black-boxes. This is likely to be a result of the economic background and training associated with official statisticians: eg in the syllabus of one of the most famous graduate studies in official statistics, the Msc of the University of Southampton, there is not a single hour devoted to DM or Machine Learning. The same is true for the master in "statistique publique" (ENSAI and Université Rennes 1).

Furthermore, NSIs are often ruled by economists who believe in their science, and Data Mining is not 'science' for such intellectuals, and researchers dislike

automatic processes. This cultural misconception must be changed sooner than later if we are to realise an increasing application of Data Mining to official data.

Karlberg and Skaliotis (2013) comment on the extremely cautious and conservative nature of official statisticians in terms of exploring novel types of data and thus concur with our analysis, which suggests that traditions are impeding the fruitful application of data mining in official statistics.

However a recent paper by Hal Varian (2014) who is Google's chief economist but also a renowned professor of economics, might be the sign of a deep change. H.Varian writes: "When confronted with a prediction problem an economist would think immediately of a linear or logistic regression. However, there may be better choices, particularly if a lot of data is available."

H.Varian presents regression and classification trees, boosting, random forests, and the Lasso among other techniques. He concludes by : "Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques." However it is clear that it necessitates deep changes in the curricula of graduate studies in economics and econometrics.

## 6 Big Data and Official Statistics

The irruption of the Big Data phenomenon is clearly changing the situation. The UNECE, in partnership with Eurostat and the OECD, organizes annual meetings on the management of statistical information systems (MSIS). The 2013 MSIS meeting decided that Big Data is a key issue for official statistics. One of the conclusions from a High-level Seminar on Modernization of Statistical Production and Services (St. Petersburg, Russian Federation, 3-5 October 2012) was: "Big data is an increasing challenge. The official statistical community needs to better understand the issues, and develop new methods, tools and ideas to make effective use of Big data sources. "

At the european level the DGINS Conference (DGINS means Director Generals of the National Statistical Institutes) took several initiatives to develop the use of Big Data in Official Statistics. The Scheveningen Memorandum[1] decided in 2013 to develop an "Official Statistics Big Data strategy" and "acknowledge that the use of Big Data in the context of official statistics requires new developments in methodology". An Action Plan and Roadmap has been adopted in 2014 by the ESS Taskforce on Big Data. Its aim is "to prepare the European Statistical System for

1 http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM Final version_0.pdf

integration of big data sources into the production of official statistics across the ESS". Combining official data and web data for a predictive purpose has a increasing interest, see Cheung (2012).

We consider that it implies the development of analytical capacities and the use of Data Mining and Machine Learning technologies since classical methods does not fit well with the size of Big Data. In the document "What does Big Data mean for Official Statistics" issued after after the St Petersburg meeting, one may notice item 39: "To use Big data, statisticians are needed with a different mind-set and new skills. The processing of more and more data for official statistics requires statistically aware people with an analytical mind-set, an affinity for IT (e.g. programming skills) and a determination to extract valuable 'knowledge' from data. These so-called "data scientists" can be derived from various scientific disciplines. "

## 7 Conclusions and perspectives

Data Mining techniques are of current use for exploratory purposes in NSIs but not popular for model building and prediction. This is due to a conservative attitude of many official statisticians due to their background and training. The use of predictive Machine Learning algorithms together with external sources could greatly improve predictions and timeliness objectives. The challenge of analyzing and processing Big Data will certainly deeply modify the actual situation.

## References

1. Brito, P., and Malerba, D., (2003). Mining Official Data. Intelligent Data Analysis, 7(6), pp.497-500.
2. Cheung, P. (1998). Developments in Official Statistics and Challenges for Statistical Education. In: Proceedings of the 5th International Conference on Teaching Statistics, 21-26 June 1998 Singapore, pp.1-3.
3. Cheung, P. (2012). Big Data, Official Statistics & Social Science Research: Emerging Data Challenges, Public lecture, Department of Statistics & Actuarial Science, The University of Hong Kong http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf
4. Daas P.J.H (2012). Big Data and Official Statistics. Software Sharing Newsletter, 7, http://www1.unece.org/stat/platform/display/msis/SAB+Newsletter
5. 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
6. Diday, E., and Esposito, F. (2003). An Introduction to Symbolic Data Analysis and the SODAS Software. Intelligent Data Analysis, 7(6), pp.583-601
7. Elkan, C. (2001). Magical Thinking in Data Mining: Lessons. In: The Proceedings of SIGKDD01 International Conference on Knowledge Discovery and Data Mining, August 2001 San Francisco, CA, pp.426-431.
8. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, AI Magazine, 17(3): Fall 1996, pp.37-54
9. Fienberg, S.E. (2005). Data Mining and the Hunt for Terrorists. Focus, 35(2), pp.1-7.
10. Friedman, J.H. (1997). Data Mining and Statistics: What's the Connection? 29th Symposium on the Interface: Computing Science and Statistics, 14-17 May 1997 Houston, TX, pp.3-9.

11. Frutos, S., Menasalva, E., Montes, C., and Segovia, J. (2003). Calculating Economic Indexes per Household and Censal Section from Official Spanish Databases. Intelligent Data Analysis, 7(6), pp.603-613.
12. Glasson M, Trepanier J, Patruno V, et al. (2013). What does ‘‘big data’’ mean for official statistics? In: UN Economic Commission for Europe: Conference on European Statisticians, September 25–27, 2013, Switzerland. www1.unece.org/stat/platform/pages/viewpage.action?pageId = 77170614 .
13. Grim J., Hora J., Boček P., Somol P., Pudil P.(2010). Statistical Model of the 2001 Czech Census for Interactive Presentation, Journal of Official Statistics, Vol.26, No.4, pp. 673–694
14. Hand, D.J. (1998). Data Mining: Statistics and More? The American Statistician, 52(2), pp.112-118.
15. Hand, D.J. (2000). Methodological Issues in Data Mining. In: Bethlehem, J.G., and van der Heijden, P.G.M., ed. Compstat: Proceedings in Computational Statistics, 2000 Utrecht, Netherlands. Berlin: Physica-Verlag GMBH & Co., pp.77-85.
16. Hand, D.J. (2000a). Data Mining: New Challenges for Statisticians. Social Science Computer Review, 18(4), pp.442-449.
17. Hassani, H., Gheitanchi, S., and Yeganegi, M.R. (2010). On the Application of Data Mining to Official Data. Journal of Data Science, 8(1), pp.75-89.
18. Hassani, H., Saporta, G., Silva, E.M. (2014). Data Mining and Official Statistics : the Past, the Present and the Future. Big Data, 2 (1), pp.1-10.
19. Karlberg M, Skaliotis M. (2013). Big data for official statistics strategies and some initial European applications. In: UNECE Conference on European Statisticians, September 25–27,Switzerland. www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP30.pdf
20. Klosgen, W., and May, M. (2002). Census Data Mining – An Application. In: Proceedings of the 5th European Conference on Principles of Data Mining, 3-5 September 2001 Germany. Springer, pp.65-79.
21. Klucik, M. (2011). Introducing new tool for Official Statistics: Genetic Programming. In: Eurostat International Conference (NTTS 2011), 22-24 February 2011 Brussels, Belgium, pp.22-24.
22. Letouzé, E. (2012). Big Data for Development: Challenges & Opportunities. UN Global Pulse, http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf
23. McCarthy, J.S., Jacob, T., and Atkinson, D. (2009). Innovative Uses of Data Mining Techniques in the Production of Official Statistics. In: UN Statistical Commission Session on Innovations in Official Statistics, 20 February 2009 New York.
24. Miller, D., McCarthy, J.S., and Zakzeski, A. (2009). A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing. In: Proceedings of the Joint Statistical Meetings (2009), 1-6 August 2009 Washington DC, pp.3144-3155.
25. Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. Journal of Official Statistics, 12(4), pp.385-401.
26. Saporta, G. (1998). The Unexploited Mines of Academic and Official Statistics. In Academic and Official Statistics Co-operation, Eurostat, pp.11-15.
27. Saporta, G. (2000). Data Mining and Official Statistics. Quinta Conferenza Nationale di Statistica, ISTAT, Roma. http://cedric.cnam.fr/PUBLIS/RC184.pdf
28. Siebes, A. (1996). Data Mining: What it is and How it is Done. In: Proceedings of SEDB96, July 1996 San Miniato, Italy, pp.329-344.
29. Sumathi, S., and Sivanandam, S.N. (2006). Introduction to Data Mining and its Applications. New York: Springer.
30. Vaidya, J., and Clifton, C. (2004). Privacy-Preserving Data Mining: Why, How and When. IEEE Security & Privacy, 2(6), pp.19-27.
31. Varian, H. (2014). Big Data: New Tricks for Econometrics, Journal of Economic Perspectives, 28 (2), pp. 3–28