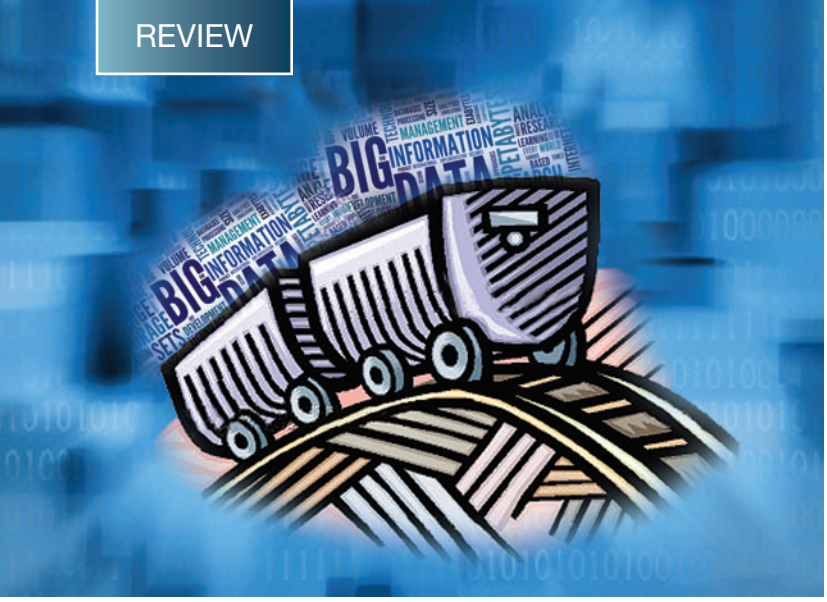
REVIEW



DATA MINING AND OFFICIAL STATISTICS:

The Past, the Present and the Future

Hossein Hassani,[1] Gilbert Saporta,[2] and Emmanuel Sirimal Silva[1]

Abstract

Along with the increasing availability of large databases under the purview of National Statistical Institutes, the application of data mining techniques to official statistics is now a hot topic that is far more important at present than it was ever before. Presented in this article is a thorough review of published work to date on the application of data mining in official statistics, and on identification of the techniques that have been explored. In addition, the importance of data mining to official statistics is flagged and a summary of the challenges that have hindered its development over the course of the last two decades is presented.

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.[1] Despite the fact that data mining is being utilized and introduced in many different fields ranging from astronomy to chemistry,[2,3] 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.[4] Half a decade since the new millennium, Sumathi and Sivanandam[5] state that there exist only a few if not any reported applications of data mining in official statistics. Furthermore, 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,[6] and as Letouzé[7] emphasizes, it is indeed opportune to use data mining to supplement official statistics in order to gain richer and deeper insights. 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 the common practice has been to outsource the analysis[5,6]; second, the objective of official statisticians is to answer precise questions and make forecasts as opposed to finding unexpected patterns or models.[4,6] Owing to these reasons, it is agreeable that, as Saporta[8] asserts, academic and official statisticians should interact more often. It is the responsibility of NSIs to open up and initiate a platform for such interaction as it is the NSIs that can gain immensely via the positive synergies created through such collaboration.

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, offers a workshop on big data and data mining. These two conferences can be construed as 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

[1]*Executive Business Centre, The Business School, Bournemouth University, Bournemouth, United Kingdom.*
[2]*Conservatoire National des Arts et Métiers, Paris, France.*

BIG DATA MARCH 2014 • DOI: 10.1089/big.2013.0038

amount of research work to be conducted in this field. This is perceptible following the review of the currently published work, which this article presents.

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.). In 2011, a poll conducted by kdnuggets.com ranked big data as the hottest data mining topic for 2012. 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 summarize the contributions that have been made to this field. Accordingly, this article categorizes, compares, and summarizes from 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 of Data Mining and Official Statistics takes a look at the definitions of data mining and its evolution. The section Why Should Data Mining Be Applied to Official Data? presents the need for applying data mining to official statistics, while the section Impediments for the Application of Data Mining to Official Data looks at the issues that must be addressed when applying data mining in official statistics. The section titled Applications of Data Mining to Official Data is a study of the successful applications over the last two decades, and the article wraps up in the Conclusion section.

Definitions of Data Mining and Official Statistics

Before addressing the concerns of this article, 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 they are 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, while the more modern definitions appreciate the emergence of supervised data mining. Accordingly, in the 1990s data mining was defined as follows:

> The search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data.[9]

> The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[10] (p. 40)

> The process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners.[3] (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.[11] (p. 443)

> Finding statistically reliable, previously unknown and actionable insights from data.[12]

> A set of automated techniques used to extract previously unknown pieces of information from large databases.[6] (p. 76)

"THE NEW MILLENNIUM HAS SEEN A SIMPLIFICATION OF THE DEFINITIONS OF DATA MINING AND THE EMERGENCE OF SUPERVISED DATA MINING."

Hassani et al.[6] also provide a more technical definition for data mining. This particular definition is closely related to that of Hand[3] but incorporates two new aims of data mining, in addition to discovering relationships, that is, the discovery of hidden trends and patterns. This is supervised data mining.

On the other hand, the objective of official data (which is used to produce "official statistics") is to accurately count a state's resources.[8,13] 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.[6] (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.[14] (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.[11] Furthermore, models do not come from a theory, but from data exploration. As such, data mining is not concerned with estimation and tests or prespecified models but with discovering models through an algorithmic search process exploring linear and nonlinear models, explicit or not.

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.[4] Official data that is collected by white-collar statisticians 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.[2–4,6]

Cheung[15] identifies two sets of statisticians, blue-collar and white-collar statisticians, where blue collar essentially refers to statisticians employed in producing official statistics for the benefit of policy makers and society as a whole, while white-collar statisticians are those employed by universities and research institutes as mathematical statisticians and academics. This distinction provides a second reason that warrants the application of data mining to official data as there exists an opposing view between blue-collar statisticians and white-collar statisticians because of the existence of problems in official data; for example, this could be because of problems relating to data accuracy and reliability.[8] The tools used for data mining can aid in understanding and overcoming these problems and help bridge the gap dividing academic and official statisticians while developing the quality of official statistics.

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.[3,4] 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.[6] 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,[4,8] 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.[4,11] This is exactly in line with the main responsibility of statisticians employed by NSIs, which is to build models. Brito and Malerba[13] 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.,[16] 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.

Impediments for the Application of Data Mining to Official Data

Hassani et al.[6] outline issues that must be addressed to achieve a successful application of data mining to official data. These issues are explained below in detail and expanded further using information obtained from other published articles.

Aggregated data

The law strictly prohibits NSIs from publishing or releasing individual responses because of privacy concerns.[6,13,17] As a consequence, NSIs are legally bound to aggregate data before releasing it to any external authority. Hassani et al.[6] state 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). Nonetheless, this is a challenge that should be welcomed, as the lucrative application of data mining should be to use techniques that integrate and appreciate privacy concerns.[6] "Symbolic data analysis" was introduced in order to overcome the challenges imposed by aggregated data.[13,18,19] 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. According to Brito and Malerba,[13] 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 Project 2001–2003, which developed the associated methodology and tools further; and the Spatial Mining for Data of Public Interest. Another project worthy of recognition is Knowledge Extraction for Statistical Offices, which was initiated in 1996 under Eurostat and DOSIS.

Data quality

McCarthy and Earp[20] state that data mining could be used to improve data quality. This is significant because, given that humans are likely to err, all datasets compiled by humans are

likely to contain errors. Over the years, data quality mining (DQM) has transformed to be an important concept because "real" data is noisy, inconsistent, and often incomplete.[21,22] The direct application of such data into data mining would result only in a garbage-in-garbage-out scenario. In addition, data mining techniques that concentrate on pattern detection are negatively affected by poor quality data.[23] DQM has been defined as the deliberate application of data mining techniques for data quality measurement and improvement.[6,24,25] The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases. According to the goal of DQM, it suggests that data cleaning[21] is similar to DQM. This is because data cleansing is defined as the process of identifying and correcting erroneous records.[26] Clustering approaches, dependency analysis, and data mining methods employed for deviation and outlier detection appear promising.[25] In addition, deficiencies in data can be recognized by the use of neural networks and artificial intelligence.[27,28] It is possible to classify the application of data mining to improve data quality as follows[6,29]:

1. Measuring and explaining data quality deficiencies
2. Correcting deficient data
3. Extension of Knowledge Discovery Database (KDD) process models to reflect the potentials of data quality management
4. Development of specialized process models for pure data quality management

"THIS IS SIGNIFICANT BECAUSE, GIVEN THAT HUMANS ARE LIKELY TO ERR, ALL DATASETS COMPILED BY HUMANS ARE LIKELY TO CONTAIN ERRORS."

Timeliness

The stated objectives of most NSIs require them to provide the public with timely statistics.[15] 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.[6] Therefore, timeliness becomes yet another important issue that needs to be addressed. Timeliness can be presented as a subcomponent of data quality and one that should complement DQM. This is because there is little or no use if DQM ends up delaying the availability of data for decision making. As stated by Hassani et al.,[6] what is required is a ready-to-use model that can guarantee smooth, efficient, and increased quality correction of data with time. For example, Miller et al.[30] show that the National Agriculture and Statistical Service in the United States is researching and developing data mining as a source of disseminating timely official statistics, while Klucik[31] shows that genetic programming as a data mining tool can be used for improving the timeliness of NSI 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 that are adopted must ensure that people or companies that provided the data are not recognized or publicized. Saporta[8] declares that confidentiality is one aspect of official statistics that stimulates the interest of academic statisticians. However, at the outset, data mining appears to be the complete opposite of protecting the confidentiality of official statistics. For that reason, it is imperative to emphasize that data mining is not looking for individuals and their details or their relationships; on the contrary, it is looking for patterns and statistical relationships in data.[6,32] At the same time, it is equally important to illustrate to the masses that data mining does in fact protect confidentiality. It is in this milieu that confidentiality issues and statistical disclosure methods have been developed to maximize the use of data while keeping with the confidentiality guarantee provided by the original data source.[32] In addition, it is suggested that as confidentiality is crucial, NSIs should carry out the data mining work on official statistics.[4] However, according to Sumathi and Sivanandam,[5] 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. While 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 in terms of not only counterterrorism,[33] 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 while safeguarding the data quality.[5] In addition, data perturbation and secure multiparty computation is also used to overcome privacy- and confidentiality-related issues.[34]

Metadata

Metadata refers to the descriptions of the meaning and context of the data.[3] More simply stated, it is data regarding the data.[5] Mining official data requires retrieving data from various surveys or administrative sources and correctly construing them as measures of observed phenomena.[6,14] Early into the millennium, Saporta[4] identified that text mining could be used to analyze metadata information, and through the work of Leckie and Yasinsac,[35] it is argued that metadata could be utilized for intrusion detection in an encrypted environment. However, over a decade since the millennium,

introducing metadata management practices into 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.[6,14] 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 that is able to provide the rough material for data integration by means of homogenously documenting data from different sources in a unique environment.[6,14]

Applications of Data Mining to Official Data

Over the years, NSIs, researchers, and academics have endeavored to apply data mining techniques to official data. The successful applications have been identified below. Interestingly, the oldest, published application of data mining to official data was in 1996. The successful applications have been classified below based on the application source, and then further categorized based on data mining techniques that were adopted. Accordingly, the applications can be grouped into three main categories as follows:

1. Third parties mining published official statistics

   The applications in this section relate to efforts by third parties to use data mining techniques on published official statistics that can aid in the successful exploration of previously undiscovered information and patterns.

   - Bayesian regression

     The Bayesian hierarchical model was used by Wheldon et al.[36] to model the population dynamics over the period of reconstruction of the female population of Burkina Faso from 1960 to 2000.

   - Decision trees

     Recently, Nithya and Sundaram[37] made use of Indian agricultural data and applied the C4.5 decision trees algorithm to discover classification rules for Indian rice diseases. Here they found that decision trees benefit from improved interpretability in comparison to methods such as neural networks.[37]

   - Neural networks

     Frutos et al.[19] adopted symbolic data analysis with a procedure based on neural networks to calculate economic indexes per household and censal section from official Spanish databases.

2. Third parties mining data collected for producing official statistics

   The following applications are concerned with third parties applying data mining techniques on data that is collected for the purpose of producing official statistics.

   - Association analysis

     Malerba et al.[38] applied a relational approach to mining spatial association rules in census data in Stockport, United Kingdom. While the authors were successful in discovering certain new rules that can be interpreted as new knowledge, they acknowledge that the overall process is highly demanding of the data analyst. Klosgen and May[17] analyzed 1991 U.K. census data using association analysis to illustrate the association between spatial subgroup mining and GIS mapping to explain high mortality rates. Moreover, Appice et al.[39] applied spatial association rule mining to georeferenced U.K. census data of 1991 in order to address issues concerning accessibility of an urban area. This application enabled the identification of novel patterns that also happened to be new knowledge for urban planners. The following year, Brtio and Malerba[13] identified the growing demand for data analysis techniques that can link population data to their spatial distribution.

   - Bayesian regression

     Paaβ and Kindermann[40] used Bayesian regression not only to analyze complex relationships between georeferenced variables but also to allow estimating the intrinsic uncertainty of predictions. The Bayesian model was applied to predict the value of long-term illness in Stockport, United Kingdom, using statistics from the 1991 census.

   - Decision trees

     It is reported that researchers at the University of Ottawa applied the technique of decision trees to the Canadian census data of 1901 in order to uncover influences of bilingualism at the beginning of the last century.[41]

   - Genetic programming

     According to Smith,[42] genetic programming could be applied in official data for the purposes of regression, clustering, classification, or change and deviation detection. Klucik[31] complements the ideas presented by Smith[42] and further asserts how genetic programming could be successfully introduced to official statistics. However, a practical application of this technique is yet to be achieved.

   - Inductive learning algorithm

     The sole application of a machine learning technique for official statistics is found in Soares et al.,[43] where the authors combined inductive learning algorithms with outlier detection methods to detect errors in foreign trade forms of the Portuguese Institute Statistics.

   - Neural networks

     Nordbotten[44] applied neural network imputation to the Norwegian population census data of 1990 with

the aim of carrying out a population census by appending the administrative data with data amassed from sample surveys.

3. Blue-collar statisticians using data mining techniques (as opposed to current statistical analyses) for producing official statistics

This group includes exertions of data mining techniques by government statisticians for producing official statistics instead of relying on the established, current statistical analyses.

- Cluster analysis

Through the work of Earp et al.,[45] hierarchical clustering was used to show how individual states' questionnaires can be combined so as to reduce the number of versions required. Interestingly, however, this approach, which can save both time and money for the National Agricultural Statistics Service (NASS), is yet to be adopted and may be construed as evidence of traditions hindering the development of data mining in official statistics. McCarthy et al.[46] relied on cluster analysis to aid with the imputation of missing data by applying the technique for the 2007 census donor pool screening in order to enable creating and seeding a donor pool of records. In addition, McCarthy et al.[46] adopted clustering to identify subtypes of records missing from the census mailing list.

- Decision trees

The U.S. Public Health Service used classification tree models to perform postsurvey nonresponsive weighting adjustments in the 1996 Medical Expenditure Panel Survey Household Component.[47] As mentioned in McCarthy et al.,[46] the NASS of the United States applied decision tree models to numerous applications in the 2007 Census of Agriculture. In fact, reported in Cecere[48] is that the nonresponse adjustment methodology used by NASS was changed by dividing the 2007 census records into response propensity groups that represented the weighting adjustment cells with the aid of classification trees. Garber[49] records the use of classification trees to recognize records on the initial census mail list that were unlikely representations of farming operations. The U.S. Census Bureau made use of the decision trees technique during the 2010 Census Coverage Measurement study by applying a stepwise regression to the concept of CART modeling for recursive partitioning of racial classification cells.[50] In addition, McCarthy and Earp[20] adopted classification tree models for analyzing reporting errors in agricultural survey data and found them beneficial over methods such as logistic regressions as classification trees are not bound by assumptions relating to the characteristics of the data. Inspired by the success of the previous work, McCarthy et al.[51] used decision trees for the prediction of survey nonrespondents.

"INDEED, IF NSIs ARE WILLING TO EMBRACE AND INCORPORATE DATA MINING TECHNIQUES FOR OFFICIAL STATISTICS, THE POSSIBILITIES ARE ENDLESS."

Key observations from reported applications of data mining in official statistics

It is clear from this grouped analysis that majority applications of data mining in official statistics have been from third parties mining data that have been collected for producing official statistics (mainly census data). Furthermore, we see that third parties have explored a variety of data mining techniques for official data in comparison to the research by government statisticians that has mainly focused on decision trees and cluster analysis alone. It is also evident that decision trees are the most popular data mining technique among NSIs at present. This increased application of decision trees by NSIs is not astonishing as polls by KDnuggets[52] indicate that decision trees are the most widely used data mining technique at present. However, it is alarming to notice that cluster analysis, which happens to be the third most popular data mining technique, is yet to be explored by third parties for official statistics.

On the basis of this, we are able to identify certain challenges that must be overcome for the development and increased application of data mining techniques in official statistics. First, there is substantial evidence that government statisticians are trapped in traditions that limit their exposure and willingness to exploit and explore lucrative and novel data mining techniques that can improve and enhance their efficiency and quality of information provided through official statistics. The answer to this challenge is the second point following from the analysis, which compels us to reiterate the call for increased engagement, cooperation, and collaboration between blue-collar and white-collar statisticians. Such collaboration will undoubtedly encourage government statisticians to consider the usage of novel data mining techniques while creating synergies that will enhance the quality of official statistics published by NSIs while greatly improving the rate of methodological advances in data mining techniques, as pursued by white-collar statisticians.

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 U.S. government

agencies, and we could not find a single application of this sort from government organizations in the United Kingdom—or any other country for that matter. It is indeed surprising that data mining techniques have thus far not been employed by the Office for National Statistics (ONS), which is also the recognized NSI for the United Kingdom. However, it should be noted that the ONS is on track to employing data mining techniques as was evident during the recently concluded ONS summer workshop, where the potential benefits of employing data mining techniques for obtaining added value out of the U.K. census were discussed.[53] Moreover, recent news reports indicate that the ONS is supposed to deliver substantial cost cuts of which majority are expected through efficiency gains,[54] and the increased application of appropriate data mining techniques could help the ONS achieve increased efficiency levels at a reduced cost.

Conclusions

Data mining is a process of secondary data analysis, and unlike the heavily model-driven modern statistics, data mining gives prominence to algorithms.[23] As a result, data mining can be considered a branch of exploratory statistics where the focus is on finding new and useful patterns through the extensive use of classic and new algorithms.

Buelens et al.[1] posit that the application of data mining for official statistics is still in its early life. We disagree with this notion and present as evidence the significant quantity of quality published work directly relating to the use of data mining in official statistics that has been reviewed in this article. However, we agree that data mining techniques are becoming increasingly vital in the production of official statistics,[1] and we present the notion that the cultural gap and conservative attitude seen in NSIs are hindering the minor or existent applications of data mining in official statistics.

Furthermore, it is 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 techniques that are at their disposal today. As evident from this review article, NSIs seldom use emblematic data mining techniques such as Bayesian regression, association rules, neural networks, and support vector machines. In fact, we could not find a single publication relating to the application of support vector machines in official data, while only one application of machine learning (the use of an inductive learning algorithm) could be found, and that too reported by a third party. The application of genetic programming is yet to be explored even though it has been identified as a potential technique.

As future work, the statisticians at ONS and other NSIs around the globe should consider the following research ideas. For example, Google's data mining efforts yield forecasts of unemployment trends long before the release of official government statistics by mining the timing and locations of search engine queries.[55] The ONS could research into adopting a similar approach for computing various official forecasts, which will not only improve the timeliness of official forecasts but also enable early policy alterations that have the potential to curb negative economic outcomes. Furthermore, as mentioned by Bollier,[55] historically, drops in claims for first-time unemployment benefits have signaled the end of recessions, and the ONS can predict such economic trends weeks in advance of the usual time scale by mining search queries for jobs and welfare. Moreover, the ONS can complement and enhance the richness of its economic data by mining tweets (among other social network status) related to inflation, housing, food, and fuel in the United Kingdom, a process that has, according to the UN Global Pulse,[56] revealed a close correlation between tweets relating to the price of rice in Indonesia and actual food price inflation. In addition, the ONS could consider mining UK's traffic records to supplement information for official traffic and transport statistics, and currently there exists ongoing research into the feasibility of mining mobile positioning data for producing tourism statistics.[16] Indeed, if NSIs are willing to embrace and incorporate data mining techniques for official statistics, the possibilities are endless.

> "THIS CULTURAL MISCONCEPTION MUST BE CHANGED SOONER THAN LATER IF WE ARE TO REALIZE AN INCREASING APPLICATION OF DATA MINING TO OFFICIAL DATA."

We are of the opinion that statisticians at NSIs are reluctant to use these modern data mining techniques because they prefer models that 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 and, as mentioned by Glasson et al.,[16] NSIs are in need of employees with analytical mindsets and increased statistical awareness. Moreover, Karlberg and Skaliotis[57] 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.

In addition, it is important to note that while data mining techniques are computer-intensive, it poses no threat to the employability of humans as human expertise and intervention

in the process continues to be mandatory. This review article illustrates the benefits that official statistics could reap by exploiting data mining and outlines the application of various techniques to solve practical problems in the real world. However, the successful application of data mining in official statistics will depend greatly on how effectively data mining experts can address the obstacles highlighted in the section titled Impediments for the Application of Data Mining to Official Data.

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 realize an increasing application of data mining to official data. In addition to overcoming the said problems, NSIs will have to redefine their missions to welcome the lucrative application of data mining techniques. It is encouraging to see state-funded programs researching into the possible solutions for the issues highlighted through this review, and as identified in Karlberg and Skaliotis,[57] several NSIs are currently taking important and bold initiatives to overcome the issues hindering the use of data mining for official statistics.

Author Disclosure Statement

No conflicting financial interests exist.

References

1. Buelens B, Daas P, Van den Brakel J. Data mining for official statistics: challenges and opportunities. In: The Proceedings of the 12th IEEE International Conference of Data Mining Workshops, December 10, 2012, Brussels, Belgium, pp. 915.
2. Friedman JH. Data mining and statistics: what's the connection? Presented at the 29th Symposium on the Interface: Computing Science and Statistics, May 14–17, 1997, Houston, TX, pp. 3–9.
3. Hand DJ. Data mining: statistics and more? Am Stat 1998; 52:112–118.
4. Saporta G. 2000. Data mining and official statistics. Quinta Conferenza Nationale di Statistica, ISTAT, Roma. Available online at http://cedric.cnam.fr/PUBLIS/RC184.pdf (Last accessed on February 13, 2014).
5. Sumathi S, Sivanandam SN. Introduction to Data Mining and Its Applications. New York: Springer, 2006.
6. Hassani H, Gheitanchi S, Yeganegi MR. On the application of data mining to official data. J Data Sci 2010; 8:75–89.
7. Letouzé E. 2012. Big data for development: challenges & opportunities. UN Global Pulse. Available online at www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf (Last accessed on February 13, 2014).
8. Saporta G. The Unexploited Mines of Academic and Official Statistics. Academic and Official Statistics Cooperation, Eurostat, 1998, pp. 11–15.
9. Siebes A. Data mining: what it is and how it is done. In: Proceedings of SEDB96, July 1996, San Miniato, Italy, pp. 329–344.
10. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996; 17:37–54.
11. Hand DJ. Methodological issues in data mining. In: Bethlehem JG, van der Heijden PGM, eds. Compstat: Proceedings of Computational Statistics, 2000, Utrecht, Netherlands. Berlin: Physica-Verlag GMBH & Co., 2000, pp. 77–85.
12. Elkan C. Magical thinking in data mining: lessons. In: The Proceedings of SIGKDD01 International Conference on Knowledge Discovery and Data Mining, August 2001, San Francisco, CA, 2001, pp. 426–431.
13. Brito P, Malerba D. Mining official data. Intell Data Anal 2003; 7:497–500.
14. D'Angiolini G. 2002. Developing a metadata infrastructure for official data: the ISTAT experience. Available online at www.di.uniba.it/malerba/activities/mod02/pdfs/dangiolini.pdf (Last accessed on February 13, 2014).
15. Cheung P. Developments in official statistics and challenges for statistical education. In Pereira-Mendoza, L. et al (eds.). Statistical Education–Expanding the Network. In: Proceedings of the Fifth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute, 1998, Volumes 1–3.
16. Glasson M, Trepanier J, Patruno V, et al. What does "big data" mean for official statistics? In: UN Economic Commission for Europe: Conference on European Statisticians, September 25–27, 2013, Switzerland. Available online at www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 (Last accessed on February 13, 2014).
17. Klosgen W, May M. Census data mining—an application. In: Proceedings of the 5th European Conference on Principles of Data Mining, September 3–5, 2001, Germany. Berlin: Springer, 2002, pp. 65–79.
18. Diday E, Esposito F. An introduction to symbolic data analysis and the SODAS software. Intell Data Anal 2003; 7:583–601.
19. Frutos S, Menasalva E, Montes C, Segovia J. Calculating economic indexes per household and censal section from official Spanish databases. Intell Data Anal 2003; 7:603–613.
20. McCarthy JS, Earp MS. Who makes mistakes? Using data mining techniques to analyze reporting errors in total acres operated. In: RRD Research Report. Washington, DC, 2009, pp. 1–20.
21. Sund R. Utilisation of administrative registers using scientific knowledge discovery. Intell Data Anal 2003; 7:501–519.

22. Kuonen D. Data mining and statistics: what is the connection? The Data Administration Newsletter, 2004, 30.0.
23. Hand DJ. Data mining: new challenges for statisticians. Soc Sci Comput Rev 2000; 18:442–449.
24. Hassani H, Anari M. Using data mining for data quality improvement. In: Proceedings of the 55th Session International Statistical Institute, April 5–12, 2005, Sydney, Australia. Netherlands: The International Statistical Institute, 2005.
25. Hassani H, Haeri Mehrizi A. Data mining and official statistics. J Stat Centre Iran 2006; 67:21–34.
26. National Statistical Commission. 2011. Report of the Committee on Data Management. New Delhi: Ministry of Statistics and Programme Implementation—Government of India. Available online at http://mospi.nic.in/mospi_new/upload/finalreportonData%20management01082011.pdf?status=1&menu_id=181 (Last accessed on February 13, 2014).
27. Hipp J, Guntzer U, Nakhaeizadeh G. 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), September 13–16, 2000, Lyon, France. New York: Springer, 2000, pp. 159–168.
28. Yeganegi MR, Hassani H, Haeri Mehrizi A. Artificial intelligence and its application to official statistics. In: Proceedings of the 8th Iranian International Statistics Conference, 2006, Shiraz, Iran, pp. 120–132.
29. Hipp J, Guntzer U, Grimmer U. Data quality mining—making a virtue of necessity. In: Proceedings of the 6th ACM SIGMOD Workshop in Research Issues in Data Mining and Knowledge Discovery (DMKD 2001), May 20, 2001, Santa Barbara, CA. New York: Cornell University, Department of Computer Science, 2001, pp. 52–57.
30. Miller D, McCarthy JS, Zakzeski A. A fresh approach to agricultural statistics: data mining and remote sensing. In: Proceedings of the Joint Statistical Meetings (2009), August 1–6, 2009, Washington, DC. Washington, DC: American Statistical Association, 2009, pp. 3144–3155.
31. Klucik M. Introducing new tool for official statistics: genetic programming. In: Eurostat International Conference (NTTS 2011), February 22–24, 2011, Brussels, Belgium. Belgium: Eurostat, 2011, pp. 22–24.
32. Nanopoulos Ph, King J. 2002. Important issues on statistical confidentiality. Available online at www.di.uniba.it/~malerba/activities/mod02/pdfs/nanopoulos.pdf (Last accessed on February 13, 2014).
33. Fienberg SE. Data mining and the hunt for terrorists. Focus 2005; 35:1–7.
34. Vaidya J, Clifton C. Privacy-preserving data mining: why, how and when. IEEE Security Privacy 2004; 2:19–27.
35. Leckie T, Yasinsac A. Metadata for anomaly-based security protocol attack deduction. IEEE Trans Knowledge Data Eng 2004; 16:1157–1168.
36. Wheldon MC, Raftery AE, Clark SJ, Gerland P. Estimating Demographic Parameters with Uncertainty from Fragmentary Data. Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington, 2011, Working Paper 108.
37. Nithya A, Sundaram V. Wheat disease identification using classification rules. Int J Sci Eng Res 2011; 2:244–248.
38. Malerba D, Lisi FA, Sblendorio F. Mining spatial association rules in census data: a relational approach. In: Proceedings of the ECML/PKDD'02 Workshop on Mining Official Data, August 19–23, 2002, Helsinki, Finland. Italy: University of Bari, 2002, pp. 1–14.
39. Appice A, Ceci M, Lanza A, et al. Discovery of spatial association rules in geo-referenced census data: a relational mining approach. Intell Data Anal 2003; 7:541–566.
40. Paaβ G, Kindermann J. Bayesian regression mixtures of experts for geo-referenced data. Intell Data Anal 2003; 7:567–582.
41. Drummond C, Matwin S, Gaffield C. Inferring and revising theories with confidence: data mining the 1901 Canadian census. In: Proceedings of the ECML/PKDD'02 Workshop on Mining Official Data (MOD'02), August 19–23, 2002, Helsinki, Finland. Italy: University of Bari, 2002.
42. Smith PWH. Genetic programming as a data-mining tool. In: Abbass HA, Sarker RA, Newton CS, eds. Data Mining: A Heuristic Approach. London: Idea Group Publishing, 2002, pp. 157–173.
43. Soares C, Brazdil P, Pinto C. Machine learning and statistics to detect errors in forms: competition or cooperation? In: Proceedings of the ECML/PKDD'02 Workshop on Mining Official Data, August 19–23, 2002, Helsinki, Finland. Italy: University of Bari, 2002.
44. Nordbotten S. Neural network imputation applied to the Norwegian 1990 population census data. J Off Stat 1996; 12:385–401.
45. Earp M, Cox S, McDaniel J, Crouse C. Exploring quarterly agricultural survey questionnaire version reduction scenarios. In: RRD Research Report. Washington, DC, 2009, pp. 1–36.
46. McCarthy JS, Jacob T, Atkinson D. Innovative uses of data mining techniques in the production of official statistics. In: UN Statistical Commission Session on Innovations in Official Statistics. New York: UN Statistical Commission, 2009.
47. Cohen SB, DiGaetano R, Goksel H. Estimation Procedures in the 1996 Medical Expenditure Panel Survey Household Component. MEPS Methodology Report No. 5, AHCPR Publication No. 99–0027. Rockville, MD: Agency for Health Care Policy and Research, 1999.
48. Cecere W. 2007 Census of Agriculture Non-Response Methodology. In: 2009 Joint Statistical Meetings, August 1–6, 2009, Washington, DC. Washington, DC: American Statistical Association, 2009, pp. 2762–2769.
49. Garber SC. Census mail list trimming using SAS data mining. In: RRD Research Report, May 9, 2009, Washington, DC.
50. Gilary A. Recursive partitioning for racial classification cells. In: Proceedings of the Survey Research Methods

Section, American Statistical Association–Session 628: Survey Analysis and Issues with Data Quality. Miami Beach, 2011. Washington, DC: American Statistical Association, 2011, pp. 2706–2720.
51. McCarthy J, Jacob T, McCracken A. Modeling NASS survey non-response using classification trees. In: RDD Research Report, November 2010, Washington, DC, pp. 1–28.
52. KDnuggets. 2011. Polls: algorithms for data mining (Nov 2011). Available online at www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html (Last accessed on February 13, 2014).
53. Leventhal B. 2013. Views of commercial users. In: Presentation at the ONS Summer Workshop, July 24, 2013, Manchester. Available online at www.ons.gov.uk/ons/guide-method/census/analysis/summer-workshops/index.html (Last accessed on February 13, 2014).
54. Office for National Statistics. 2013. ONS opens up consultation on statistical products. Available online at www.statisticsviews.com/details/news/5267271/ONS-opens-up-consultation-on-statistical-products.html (Last accessed on February 13, 2014).
55. Bollier D. The Promise and Peril of Big Data. Washington, DC: The Aspen Institute, 2010.
56. UN Global Pulse. 2013. Twitter and perceptions of crisis-related stress. Available online at www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress (Last accessed on February 13, 2014).
57. Karlberg M, Skaliotis M. 2013. Big data for official statistics—strategies and some initial European applications. In: UN Economic Commission for Europe: Conference on European Statisticians, September 25–27, 2013, Switzerland. Available online at www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP30.pdf (Last accessed on February 13, 2014).

Address correspondence to:

Hossein Hassani
Statistical Research Centre
Executive Business Centre
The Business School
Bournemouth University
Bournemouth BH8 8EB
United Kingdom

E-mail: hhassani@bournemouth.ac.uk