HIGH-LEVEL GROUP FOR THE MODERNISATION OF STATISTICAL PRODUCTION AND SERVICES (HLG)

Modernisation Committee on Production and Methods

Machine Learning Documentation Initiative

Kenneth Chu and Claude Poirier, Statistics Canada

4 February 2015

ABSTRACT

1 The objective of this briefing is to present an overview of the machine learning techniques currently in use or in consideration at statistical agencies worldwide. Section I, outlines the main reason why statistical agencies should start exploring the use of machine learning techniques. Section II outlines what machine learning is, by comparing a well-known statistical technique (logistic regression) with a (non-statistical) machine learning counterpart (support vector machines). Sections III, IV, and V discuss current research or applications of machine learning techniques within the field of official statistics in the areas of automatic coding, editing and imputation, and record linkage, respectively. The material presented in this paper is the result of a literature review, of direct contacts with authors during conferences, and more importantly of an international call for input that was distributed on July 18, 2014 to participants from the 2014 MSIS Meeting, participants from the 2014 Work Session on Statistical Data Editing, and members of the Modernization Committee on Production and Methods. Section VI contains a list of machine learning applications in official statistics outside of the three areas mentioned above.

I. INTRODUCTION

A. What is Machine Learning?

2 In the statistical context, Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalisation in order to get a system that performs well on yet unseen data instances.

B. Why should statistical agencies consider machine learning?

3 Machine learning is a relatively new discipline within Computer Science that provides a collection of data analysis techniques. Some of these techniques are based on well established statistical methods (e.g. logistic regression and principal component analysis) while many others are not.

4 Most statistical techniques follow the paradigm of determining a particular probabilistic model that best describes observed data among a class of related models. Similarly, most machine learning techniques are designed to find models that best fit data (i.e. they solve certain optimization problems), except that these machine learning models are no longer restricted to probabilistic ones.

5 Therefore, an advantage of machine learning techniques over statistical ones is that the latter require underlying probabilistic models while the former do not. Even though some machine learning techniques use probabilistic models, the classical statistical techniques are most often too stringent for the oncoming Big Data era, because data sources are increasingly complex and multi-faceted. Prescribing probabilistic models relating variables from disparate data sources that are plausible and amenable to statistical analysis might be extremely difficult if not impossible.

6 Machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. It is imperative for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

II. CLASSES OF MACHINE LEARNING

7 There are two main classes of machine learning techniques: supervised machine learning and unsupervised machine learning.

A. Examples of supervised learning

Logistic regression (statistics) *vs* Support vector machines (machine learning)

8 Logistic regression, when used for prediction purposes, is an example of supervised machine learning. In logistic regression, the values of a binary response variable (with values 0 or 1, say) as well as a number of predictor variables (covariates) are observed for a number of observation units. These are called training data in machine learning terminology. The main hypotheses are that the response variable follows a Bernoulli distribution (a class of probabilistic models), and the link between the response and predictor variables is the relation that the logarithm of the posterior odds of the response is a linear function of the predictors. The response variables of the units are assumed to be independent of each other, and the method of

maximum likelihood is applied to their joint probability distribution to find the optimal values for the coefficients (these parameterise the aforementioned joint distribution) in this linear function. The particular model with these optimal coefficient values is called the "fitted model," and can be used to "predict" the value of the response variable for a new unit (or, "classify" the new unit as 0 or 1) for which only the predictor values are known. Support Vector Machines (SVM) are an example of a non-statistical supervised machine learning technique; it has the same goal as the logistic regression classifier just described: Given training data, find the best-fitting SVM model, and then use the fitted SVM model to classify new units. The difference is that the underlying models for SVM are the collection of hyperplanes in the space of the predictor variables. The optimization problem that needs to be solved is finding the hyperplane that best separates, in the predictor space, the units with response value 0 from those with response value 1. The logistic regression optimization problem comes from probability theory whereas that of SVM comes from geometry.

9 Other supervised machine learning techniques mentioned later in this briefing include decision trees, neural networks, and Bayesian networks.

B. Examples of unsupervised learning

Principal component analysis (statistics) *vs* Cluster analysis (machine learning)

10 The main example of an unsupervised machine learning technique that comes from classical statistics is principal component analysis, which seeks to "summarize" a set of data points in high-dimensional space by finding orthogonal one-dimensional subspaces along which most of the variation in the data points is captured. The term "unsupervised" simply refers to the fact that there is no longer a response variable in the current setting.

11 Cluster analysis and association analysis are examples of non-statistical unsupervised machine learning techniques. The former seeks to determine inherent grouping structure in given data, whereas the latter seeks to determine co-occurrence patterns of items.

III. AUTOMATIC CODING

A. Automatic coding via Bayesian classifier (Germany)

12 In a poster session at the Statistics Canada's 2014 International Methodology Symposium, Bethmann et al. (of Intitut für Arbeitsmarkt-und Berufsforschung) have reported on research on applying two types of probabilistic supervised machine learning algorithms -- Naïve Bayes (NB) and conjugate Bayesian analysis based on multinomial distributions (BMN) -- for automatic occupation coding for German panel surveys. The authors used a large volume (approximately 300,000) of manually coded occupation text strings from recent surveys as

training data. The rate of agreement between automatic coding and manual coding was used as a metric to evaluate the algorithms. Although both methods exhibited good agreement rates by common machine learning standards, the authors cautioned that they might not be sufficiently satisfactory given the considerably higher accuracy requirements of occupation coding in production settings (the authors suggested a minimum agreement rate of 95%). On the other hand, the authors pointed out that when the target variable was changed to "social-economic status" or "occupational prestige" (more precisely, ISEI-08 and SIOPS-08 scores, both derived from occupation codes), both methods yielded dramatically improved results. The authors concluded that the current versions of their methods may be sufficient for production of socio-economic status or occupational prestige predictions, but further improvements are required for production of reliable occupation coding. Possibilities for improvement include the addition of a preprocessing step (to "clean up" input text strings, thereby reduce noise in training data), incorporation of a certain distance measure in the existing models, as well as different machine learning methods altogether (such as random forests or support vector machines). The authors project that their methods will be ready for release as an open-source R package in several years.

B. Automatic occupation coding via CASCOT (United Kingdom)

13 **C**omputed-**a**ssisted **S**tructured **Co**ding **T**ool [1] is an automatic occupation coding software tool developed by the Institute for Employment Research at the University of Warwick, a partner in the EurOccupations project. The objective of the project is to construct a publicly available database of the most frequent occupations to facilitate multi-country data collection. Since 2009, CASCOT has been able to perform automated coding into the ISCO'08 classification of occupational texts in any of the seven languages of the eight EurOccupations partner countries[2]. CASCOT is available for online use for free and a desktop version is available for purchase should high-volume processing be required. However, CASCOT's underlying methodology has not been published.

C. Automatic coding via open-source indexing utility (Ireland)

14 The Central Statistics Office of Ireland has reported they are developing an automatic coding system for Classification of Individual Consumption by Purpose (COICOP) assignment for their Household Budget Survey, using previously coded records as training data. Their method is based on the open-source indexing and searching tool Apache Lucene (http://lucene.apache.org).

D. Automatic coding of census variables via Support Vector Machines (New Zealand)

15 Statistics New Zealand investigated the potential of using Support Vector Machines (SVM) to improve coding of item responses in their Census. They applied SVM to code the

variables Occupation and Post-school Qualification, using two disjoint sets of observations, each of size 10,000, from Census 2013 data for training and testing. They reported 50% correctness rate on testing data for both variables, and concluded that further investigations would be necessary to further evaluate SVM as an automatic coding methodology.

IV. EDITING AND IMPUTATION

A. Categorical data imputation via neural networks and Bayesian networks (Eurostat)

16 Eurostat [3] compared imputation results for missing categorical data (voting intensions) based on two machine learning methods (neural networks and Bayesian networks) against one of the current prevailing statistical imputation methods (multiple imputation using logistic regression). The data set used for this study was the subset of records from the 2008 Spanish general election poll containing no missing data. The method of evaluation was ten-fold cross-validation (a standard algorithm evaluation protocol in machine learning) and the evaluation metric used was proportion of correctly imputed records. The author reported that the current prevailing method, multiple imputation with logistic regression, yielded a proportion of correctly imputed records of 66.0%, whereas neural networks and Bayesian networks yielded 86.1% and 87.4%, respectively. The author pointed out the limited scope of the study, but emphasized the large margin of improvements of the two machine learning methods. Machine learning procedures, the author also pointed out, are generally less sensitive to outliers, and do not require model assumptions and variable selection (needed to avoid multicollinearity among predictor variables). The author recommended that more research on the application of machine learning procedures to imputation should be conducted.

B. Identification of error-containing records via classification trees (Portugal)

17 Statistics Portugal [4] described a method based on classification trees (a type of decision trees whose response variables are categorical) for error detection in foreign trade transaction data collected by the Portuguese Institute of Statistics. They reported that the method was able to select 47% of transaction records as potentially error-containing, and that subsequent human verification confirmed that the selected records turned out to contain 91% of error-containing records. The authors projected that, in a production setting, only the selected records would be manually reviewed by human experts to confirm or refute presence of errors, and corrected where appropriate. Without such a system, all records would have to be reviewed by human experts. The authors thus concluded that the system should be able to reduce manual examination of records by close to 50% while still successfully detecting about 90% of the error-containing records.

C. Imputation donor pool screening via cluster analysis (United States)

18 The United States Department of Agriculture [5] reported the use of cluster analysis in the 2007 Census of Agriculture for imputation donor pool screening. The 2002 Census of Agriculture questionnaires were scanned and data were captured via Optical Character Recognition (OCR) to be used as imputation donor pool for the 2007 Census of Agriculture. OCR however produced an unacceptably high proportion of errors. Cluster analysis was performed on the OCR-captured data to identify records containing OCR errors as a single outlying cluster. These erroneous records were removed and the remainder were used as imputation donor pool.

D. Imputation via Classification and Regression Trees (CART) (New Zealand)

19 CART is a common supervised machine learning method for classification and regression based on "recursive binary splitting". Statistics New Zealand investigated the use of CART to predict two binary variables based on Census 2013 data. The binary variables were i) the missingness of the income variable, and ii) the response to the question of whether the respondent has moved since the previous census. Results of this investigation are being evaluated.

E. Determination of imputation matching variables via Random Forests (New Zealand)

20 Statistics New Zealand is redesigning the editing and imputation methodology of their Household Economic Survey (HES). Their current proposed methodology will use the Canadian Census Edit and Imputation System (CANCEIS). The imputation module of CANCEIS is based on the Nearest Neighbour Imputation Methodology, which requires user specification of a distance measure of pairs of units based on a number of "matching variables" as well as weights which defines the relative importance of these matching variables. The weight of a matching variable should reflect its strength as a predictor for the variables to be imputed. Statistics New Zealand has reported promising results in using Random Forests to select the set of matching variables for CANCEIS, as well as their weights. Their investigations so far have used 2009/2010 HES data as training data and 2011/2012 HES data as testing data. Further investigation using 2012/2013 HES data has been planned. The Random Forests method is a tree-based classification and regression technique which, through averaging results over many de-correlated trees, is generally more stable (smaller variance) than single-tree methods (e.g. CART).

F. Creation of homogeneous imputation classes via CART (New Zealand)

21 Statistics New Zealand compared two methods for creating homogeneous imputation classes: the score method (a.k.a. predictive mean stratification or response propensity stratification; implemented as SAS macros by Statistics Canada) and CART (implemented in the **R** package **`rpart`**). Data from the New Zealand Income Survey, with missing values being stochastically introduced, were used for training and testing. Comparison was performed for both a continuous target variable as well as a discrete one. Statistics New Zealand reported that **`rpart`** performed equally well as the score method that Statistics Canada implemented in SAS macros, especially for the discrete variable.

G. Derivation of edit rules via association analysis (New Zealand)

22 Statistics New Zealand investigated the potential of using association analysis to derive additional edit rules to enhance the processing of census data. Association analysis is an unsupervised machine learning technique whose aim is to uncover "implication patterns" or "co-occurrence patterns" among a set of categorical variables (e.g. an American resident whose first language is English and who has completed college education in the U.S. also tends to have income above USD $50,000) based on given data. Statistics New Zealand performed the investigations based on Census 2013 data. Initial results, however, yielded no previously unknown edit rules. Statistics New Zealand plans to explore the possibility of using association analysis to derive an enhanced set of edit rules for their Household Economic Survey.

V. RECORD LINKAGE

23 Record linkage is different from automatic coding and editing and imputation in the sense that machine learning may not be as helpful for record linkage as it is for the other two types of applications.

24 In practice, record linkage consists of a preprocessing and standardization step, followed by a matching step. It is the matching step that can be regarded as a supervised machine learning problem, more precisely, classification of pairs of records as matching or non-matching.

25 However, many traditional matching techniques are not from machine learning [6], and the quality of record linkage is believed to be more sensitive to the quality of preprocessing and standardization than that of matching [7]. Consequently, innovations in record linkage are believed to come more probably from improvements in preprocessing and standardization, as well as scaling up to ever increasing file sizes, rather than from innovations in matching.

26 The traditional matching approach, known as Probabilistic Record Linkage (PRL), was introduced by [8]. It was pointed out in [7] that PRL and its variants can yield very high quality

record linkage results, as long as preprocessing and standardization was effective. On the other hand, there are many challenges in applying supervised machine learning to record linkage matching. Supervised machine learning methods require training data (i.e. known match and non-match status of given record pairs), which are usually unavailable in practice except via manual labeling by subject matter experts. On the other hand, PRL and its variants do not require training data. In addition, even if training data can be made available, a naïvely assembled training data set is likely to be highly imbalanced in the sense that it will probably contain far more non-matching record pairs than matching ones (proportion of matching pairs approaches zero as file size increases), which will render training and validation very difficult.

27 Several authors [9], [10] applied classification trees to record linkage matching, whereas other authors [11], [12], [13] employed support vector machines. More recently, it was reported that neural networks consistently yielded significantly enhanced results for linking genealogical records over PRL [14]. However, this study used a hand-labeled training data set of 80,000 genealogical record pairs; the author did not address how in a production setting such training data could be efficiently generated. No recent applications in official statistics of supervised machine learning techniques to matching in record linkage have been found.

VI. OTHER APPLICATIONS IN OFFICIAL STATISTICS

A. Questionnaire consolidation via cluster analysis (United-States)

28 The United States Department of Agriculture [15] reported research on the application of hierarchical clustering to reduce the number of USDA NASS Quarterly Agriculture Survey (QAS) questionnaire versions. The QAS collects data on 31 different crops and stocks from all 50 states in the US. As of 2007, each state had its own QAS questionnaire version (different states were surveyed on different items at different frequencies), as it was believed that this approach reduced respondent burden. However, maintaining state-specific questionnaire versions was complex and costly. The authors conducted research on the possibility of using hierarchical clustering to compare the state-specific questionnaire versions and produce a smaller set of regional versions (the output clusters of the clustering algorithm) as replacement. Regionalization of QAS questionnaires was believed to simplify survey administration and the production of national estimates.

B. Forming non-response weighting groups via classification trees (United States)

29 The United States Department of Agriculture [16] reported using classification trees for non-response adjustment for the 2007 Census of Agriculture. Farms within each state in the US were partitioned into groups of “homogeneous response propensity” using a classification tree model based on NASS sampling frame data (e.g. farm size, farm type, and operator

demographics). Non-response adjustments were performed within each such group based on the response rate within that group. The author concluded that the use of classification trees for formation of non-response weighting groups was both effective and efficient, and the ability to include an arbitrarily large number of (possibly correlated) sampling frame variables as predictors was a key advantage over previous methods. One disadvantage the authors mentioned was that tight production time frame allowed only limited extent of optimization over different tree models.

C. Non-respondent prediction via classification trees (United States)

30 The United States Department of Agriculture [17] reported research results regarding application of classification trees to non-respondent prediction. The data used by the authors were the March, September and December NASS quarterly Crops/Stocks survey data from 2006 and 2007. The main response variable of interest was refusal to respond. The authors collected a large variety of auxiliary data for predictor variables. The resulting classification tree models showed that variables pertaining to response history could be used to effectively identify units likely to refuse to respond. The authors projected that data collection strategies for likely refusals could be modified in order to boost the response rate among them (e.g. by including them earlier in the data collection process, sending more experienced interviewers, etc.). Interestingly, the authors pointed out that successful non-response mitigation strategies based on this model may render this same model useless in the future. This is because this particular model deduced that response-history variables were the most relevant, and any so-inspired successful mitigation strategies will alter these very aspects of the sample units. The authors thus attempted to rebuild the classification tree model by excluding response-history variables, but unfortunately the resulting classification tree was far less effective. The authors interpreted this negative result as follows: on the one hand, there is little information in the auxiliary data they collected, apart from the response-history variables, that could be used to predict refusal (at least not via classification trees), while on the other hand, this may be taken as evidence that there was little systematic difference between refusals and non-refusals, at least in terms of the rather large variety of auxiliary data (apart from response-history) they collected for this study.

D. Analysis of reporting errors via classification trees (United States)

31 The United States Department of Agriculture [18] reported using classification trees to predict respondents likely to make reporting errors based on sampling frame data. The 2002 Census of Agriculture data were used for this analysis. Results of this analysis could suggest reasons for the reporting errors, types of respondents to be included in questionnaire testing, and editing strategies after data collection. Classification trees were believed to be superior to logistic regression (a classical statistical regression technique with a binary response variable) since, for the former, no modeling assumptions are required between response and predictor variables, and

missing data are naturally incorporated in the analysis. In addition, among the direct output of classification tree techniques is unambiguous identification of putatively error-prone respondents based on sampling frame data, which enables straightforward utilization of the results in production settings.

E. Substitutes for surveys via internet scraping and machine learning (Italy)

32 The Italian National Institute of Statistics [19] reported on research regarding the possibility of substituting (fully or partially) surveys by collecting data via internet scraping and extracting information therein using machine learning methods. The authors used as training data completed questionnaires from ISTAT's 2013 Community Survey on ICT Usage and e-Commerce in Enterprises (abbreviation: "ICT in Enterprises"), as well as data collected from respondent enterprises' web sites via internet scraping (using the open-source Apache suite Nutch-Solr-Lucene). The focus of this study was the questionnaire section pertaining to web sales activities of respondent enterprises. A variety of machine learning techniques were used to learn and predict this set of questionnaire responses, based on data scraped off web sites of respondent enterprises. The authors reported that the Naïve Bayes text mining algorithm exhibited the best performance with respect to a number of standard algorithm evaluation metrics but cautioned that sensitivity of all the examined methods were too low for production purposes ("low sensitivity" here meant that these methods incorrectly labelled too many enterprises as having no web sales activities, while in fact they do offer web sales according to their questionnaire responses). The authors reported that further and larger-scale investigations will be conducted.

F. Tax evader detection via k-nearest neighbours (Hungary)

33 The Hungarian Central Statistical Office [20] reported on research regarding possible use of the method of k-nearest neighbours (a supervised machine learning technique) to detect self-employed proprietors who are tax evaders. The authors used as training data enterprise survey microdata from Hungarian Central Statistical Office's databases, as well as Value Added Tax audit data they obtained externally from the Hungarian tax authorities. Seventy-one percent of units in the tax audit data were audited to be tax evaders while the rest (29%) law-abiding. The authors pointed out that the tax audit data could not be assumed to have been generated via probability sampling. The use of statistical techniques was thus precluded, and they decided to experiment with machine learning techniques instead. Unfortunately, even after much manual tuning, the performance of their chosen machine learning method, k-nearest neighbours, still failed to yield satisfactory results.

G. Crop yield estimation via image processing on satellite imaging data (Canada)

34 Statistics Canada has been experimenting with the use of satellite imaging data to assist with estimation of crop yields. Satellite imaging data of selected crop producing areas of the province of Prince Edward Island were used as predictor variables. Field surveyors were sent to corresponding actual locations to ascertain crop types and yields; these were used as response variables. Probabilistic image processing algorithms were used to learn and predict the field observations based on the satellite data. The eventual objective is to vastly reduce the number of field observations required to obtain the needed data. Preliminary results are promising though certain issues remain outstanding: Certain crop types are difficult to distinguish based on satellite images alone, even for human experts. Researchers have also observed that training data from one year currently cannot be used reliably for other years. Potential improvement from inclusion of auxiliary data (e.g. farm types) is being evaluated.

VII. CONCLUDING REMARK

35 The international inquiry initiated in July 2014 was very useful for determining current work on machine learning applications. Gaps were noticed though, especially in the area of record linkage, most likely due to the nature of the process which deals with score functions and search algorithms rather than artificial intelligence. Also, it is felt that more attention could be given to broader applications of machine learning, i.e. those outside the current statistical paradigms. Applying Machine Learning on Big Data still need to be done, for instance to recognize patterns of related units.

36 The applications reported in section VI show promising avenues, especially the web scraping proposed by Italy. Although the current scraping is modest, it opens a new world for predicting industrial activities. More research on the area is advised, for instance for identifying industrial coding based on text mining on official web sites.

37 Given the current portrait of Machine Learning for statistical production, the Modernisation Committee on Production and Methods recommends a second report documenting a development road map, this means future directions, development challenges and partnership opportunities related to Machine Learning applications.

VIII. ACKNOWLEDGEMENTS

38 The authors thank the HLG Executive Board for having supported the inquiry which led to this portrait, and more importantly the contributors who responded positively in providing details on their respective applications.

Bibliography

[1] [Online]. Available: http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/.

[2] P. Elias, R. Ellison and R. Jones, "Development of associated software for EurOccupations," EurOccupations, 2009.

[3] P. Rey del Castillo, "Use of Machine Learning Methods to Impute Categorical Data," in *UNECE Work Session on Statistical Data Editing*, Oslo, Norway, 2012.

[4] C. Soares, P. Brazdil and C. Pinto, "Machine Learning and Statistics to Detect Errors in Forms: Competition or Cooperation?," in *ECML/PKDD'02 Workshop on Mining Official Data*, Italy, 2002.

[5] J. McCarthy, T. Jacob and D. Atkinson, "Innovative Uses of Data Mining Techniques in the Production of Official Statistics," in *UN Statistical Commission Session on Innovations in Official Statistics*, New York, 2009.

[6] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, New York: Springer, 2012.

[7] W. Winkler, "Matching and record linkage," *WIREs Comput Stat,* vol. 6, pp. 313-325, 2014.

[8] Fellegi and Sunter, "A theory for record linkage," *Journal of the American Statistical Association,* vol. 64, no. 328, pp. 1183-1210, 1969.

[9] W. Cohen, "The WHIRL approach to data integration," *IEEE Intelligent Systems,* vol. 13, no. 3, pp. 20-24, 1998.

[10] M. Elfeky, V. Verykios and A. Elmagarmid, "TAILOR: A record linkage toolbox," *IEEE ICDE,* pp. 17-28, 2002.

[11] M. Bilenko and R. Mooney, "Adaptive duplicate detection using learnable string similarity," *ACM SIGKDD,* pp. 39-48, 2003.

[12] P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," *ACM SIGKDD,* pp. 151-159, 2008a.

[13] P. Christen, "Automatic training example selection for scalable unsupervised record linkage," *PAKDD, Springer LNAI,* vol. 5012, pp. 511-518, 2008b.

[14] D. R. Wilson, "Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage," in *Proceedings of International Joint Conference on*

Neural Networks, San Jose, California, USA, 2011.

[15] M. Earp, S. Cox, J. McDaniel and C. Chadd, "Exploring Quarterly Agricultural Survey Questionnaire Version Reduction Scenarios," United States Department of Agriculture, Washington, DC, 2009.

[16] W. Cecere, "2007 Census of Agriculture Non-Response Methodology," in *2009 Joint Statistical Meetings*, Washington, D, 2009.

[17] J. S. McCarthy, T. Jacob and A. McCraken, "Modeling Non-response in National Agricultural Statistics Service Surveys Using Classification Trees," United States Department of Agriculture, Washington, DC, 2010.

[18] J. S. McCarthy and M. S. Earp, "Who Makes Mistaktes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated," United States Department of Agriculture, Washington, DC, 2009.

[19] G. Barcaroli, A. Nurra, M. Scarno and D. Summa, "Use of Web Scraping and Text Mining Techniques in the ISTAT survey on "Information and Communication Technology in Enterprises"," 2014.

[20] I. R. Kazimir, G. Horvath and J. Giczi, "Modeling Tax Evasion of the Self-employed in Hungary".