Metodološki zvezki, Vol. 10, No. 2, 2013, 65-81

Perspectives of Data Mining in Improving Data Collection Processes in Official Statistics

Miroslav Hudec[1,2], Jana Juriová[2]

Abstract

Statistical offices are crucial institutions for collecting data about various aspects of society. Nevertheless, data collection copes with non-response in surveys and problem of missing values. Therefore, efforts focused on increasing response rates and the estimation of missing values are topics which need continual improvement. The paper examines advantages of soft computing techniques on small-scale case studies related to reminder letters, respondents' classification and estimation of missing values. Fuzzy sets have membership degree valued in the [0, 1] interval which implies that similar entities could be similarly treated in reminders and with some restriction in imputation. Neural networks are suitable when the borders of classes are not easily definable and databases contain incomplete records. In such a case the neural network can identify the most similar class for each entity and this enables the imputation of missing values. Finally, the paper discusses an efficient way for design and implementation of tools in the cooperation among statistical institutes.

1 Introduction

National Statistical Institutes (NSIs) collect data from various fields e.g. business and trade statistics, population censuses, municipal statistics. Policy decisions significantly depend on data collected by NSIs. The same holds for businesses that need data for different analyses and decisions.

Surveys cope with the problem of missing values which is a consequence of several reasons: reluctance of respondents to participate (e.g. business and trade statistics, Giesen, 2011) and non-availability of required data, e.g. in small area statistics due to rare occurrence of measured phenomenon, non-availability of instruments to measure values in all respondent units, etc. This ends up in unit non–response or item non-response which is far from negligible (Bavdaž, 2010).

[1] Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia; miroslav.hudec@euba.sk

[2] Infostat – Institute of Informatics and Statistics, Bratislava, Slovakia; juriova@infostat.sk

NSIs cope with these issues in two main ways: reminder letters to respondents in delay and the estimation of missing values. Both ways are demanding in skilled staff and tools.

Therefore, efforts focused on motivating and reminding respondents and estimation of missing values should be continuously improved (de Leeuw et al., 2003; Kľúčik, 2011; Torres van Grinsven et al., 2012). If we improve responding, less data will be missing. As a consequence, methods for estimation of missing values will be more efficient (fewer values need to be estimated and mining larger amount of available data leads to a better estimation).

However, usual tools are not able to capture and directly apply statisticians' knowledge, which cannot be always interpreted by crisp rules and numbers. Usual tools for data mining cannot envelop and evaluate all relevant relations and their intensities in the data. In this area the soft computing is a rational option which could offer the solution (Hudec et al., 2012; Kľúčik et al., 2012). Advantages of data mining and soft computing techniques in official statistics have been discussed in (Hasani et al., 2010) and some research has been already done (Kľúčik, 2012; Hudec and Juriová, 2013).

This research has also taken into account the European Plan of Research in Official Statistics (2007) which among others pointed out the inevitability of research and applicability of techniques such as neural networks/artificial intelligence and their comparison with classical statistical approaches.

The paper presents the discussion of applicability and advantages of soft computing for improving data collection in official statistics. Because data mining is the process of analysing large amount of data by various techniques and summarizing it into the useful information, we focus on soft computing techniques, more precisely fuzzy logic and neural networks. Section 2 shortly describes the data collection processes in official statistics and specific problems that these techniques can help solve. Section 3 is devoted to fuzzy logic and neural networks for solving problems of reminders, classification of respondents and estimation of missing values. Section 4 concludes this paper and discusses obstacles to use fuzzy logic and neural networks in official statistics as well as challenges for future research.

2 Specific problems in data collection processes in official statistics

The data collection of official statistics is a complex system. There are different types of collection processes used in official statistics: exhaustive surveys, sampling surveys, administrative data collection systems. Each field of collection has its own internal rules but general rules are more or less similar. Whether data are collected by questionnaires or interviews, some missing data will occur (de

Leeuw et al., 2003). NSIs are currently solving these issues by reminder letters to respondents in delay and by estimation of missing values.

Reminder letters should have different tone depending on respondents' past behaviour in regular surveys. An inappropriate tone of reminder letters could cause even higher reluctance. Another way to increase the response rate could be reward for active responding. In cases when respondents are in a significant delay or refuse to respond, NSIs have to estimate missing values e.g. flash estimates required by the government. The following sections describe the specific problems that examined data mining techniques could solve.

2.1 Reminder letters

In case of the Intra-EU trade statistics (database of trade statistics between EU countries) businesses whose trade value exceeds the exemption threshold value are becoming the respondents to data collection systems. NSI usually receives information when this happens, together with the trade value from the administrative sources. These businesses have a duty to send additional data to statistical offices. In practice, not all businesses satisfy this duty. Therefore, reminder letters with different level of tone should be created. Currently, two main approaches are used: sending the same intensity of warning to all businesses in delay regardless of their previous behaviour or importance for NSI and creating several levels of warnings using hand written rules without support of data mining (laborious work of analysing various data related to surveys).

The key cause of non-response seems to lie in its character of being an 'irritation burden' (European Commission's High Level Group of Independent Stakeholders on Administrative Burdens, 2009), and not in the actual survey burden imposed on businesses that represents only around 0.5% of the total administrative burden. If NSIs create reminder letters that are not tailored to respondents' previous behaviour, it could presumably increase feeling of irritation.

2.2 Estimation of missing values

Missing values could appear due to different reasons and therefore should be differently treated. There are different statistical techniques which can be used to fill in the missing values with estimates (de Waal et al., 2010). However, in municipal statistics and foreign trade statistics also other approaches can be introduced.

In case of small area statistics (e.g. municipal statistics) missing values are due to the fact that data are not available because of several reasons (rare occurrence of measured phenomena, reluctance of a local administration unit to cooperate in surveys, non-availability of instruments to measure phenomena, etc.).

Municipal statistics collect variety of indicators (804 in case of the Slovak municipal statistics, for 2891 municipalities). In case of data selection, if a municipality is not selected is it because a value is far to meet the condition or because a value is missing? In case of classification, municipalities with missing values cannot be classified. However, the issue of missing values was usually neglected.

In this case, neither reminder letter nor motivation could bring the solution. In the municipal statistics we could recognise some similarities between municipalities and dependencies between measured phenomena (indicators). If for example the distance between municipalities is not high and they have similar altitude above the sea level, then some indicators could be more or less dependent. Fuzzy logic is able to detect not only dependency but also the intensity of dependency. If dependency is very strong then we could estimate missing value with the probability related to the intensity of dependency.

In case of foreign trade statistics, businesses whose trade value exceeds the exemption threshold value are becoming the respondents to data collection system. Parameters of businesses' trade could significantly vary from month to month. In addition, similar businesses could obtain different values of trade. Moreover, databases are incomplete and contain outliers due to measurement errors and item and/or unit non-response. Therefore, dependencies and relations are more complex to be properly evaluated in a reasonable time limit by fuzzy logic. Neural networks could offer the answer for this issue.

In our approach we use similar scenario as in the hot deck imputation method that uses the data from other surveyed observations. In the hot deck method each missing value is replaced with data from more or less similar unit using the linear restriction rules (Coutinho and de Waal, 2012). Hot deck is efficiently used in practice, even though theory is not as well developed as in the other methods (Andridge and Little, 2010). On the other hand, theory of neural networks is well developed for detecting similar units but is rarely used in official statistics (Juriová, 2012). Another approach, the fuzzy functional dependencies are able to detect not only dependency rules but also their intensities (Berzal et al., 2005; Vucetic et al., 2013). The last two methods are not limited to the linear constraints and therefore they could cope with nonlinear relations in the data.

2.3 Classification of respondents as a support for data collection

In area of respondents' management, we would like to demonstrate a framework for the flexible classification as a support for improving cooperation in surveys. Businesses often play role as respondents and users of statistical data. In many NSIs data are not free or at least businesses have to pay a service charge for the data preparation. Service charge and fee depend on the amount of ordered data.

The example is motivating respondents to participate in surveys by discounts of provided NSI's services. This is an initial idea which could be broadened into different directions. For any further communication or other purposes, businesses with good behaviour as respondents could obtain some services above the standard level. The intensity of above standard services depends on the intensity of belonging to the class of reliable respondents.

Traditional classification techniques in data mining uses sharp classes, which imply that similar respondents might be classified into different classes. If we want to avoid this issue and create a classification space described by linguistic terms rather than by numbers, then the answer might be the fuzzy classification.

3 Data mining approaches and illustrative examples of their use in official statistics

Data mining is the analysis of (often large) observational data sets to find unexpected relationships and to summarize the data in novel ways that are both understandable and useful for the owner of data (Hand et al., 2001). Therefore, data mining is the process of analysing large amount of data by various techniques and summarizing it into the useful information. Soft computing techniques are one of these techniques used in data mining. "In contrast to traditional hard computing, soft computing exploits the tolerance for imprecision, uncertainty, and partial truth to achieve tractability, robustness, low solution-cost, and better rapport with reality" (Zadeh, 1994). In this context, fuzzy logic, neural networks, probabilistic reasoning and genetic algorithms are considered as main components of computational intelligence (Feil and Abonyi, 2008). This Section discusses fuzzy logic and neural networks and their relations to the issues mentioned in the previous Section.

3.1 Fuzzy logic

The fuzzy set theory (Zadeh, 1965) provides a framework for systematically handling the vagueness (fuzziness). The fuzzy logic is capable to catch users' knowledge which contains subjectivity and uncertainty described by linguistic terms and quantifiers, and directly apply on a certain task. Fuzzy classes do not have sharp boundaries. In data retrieval selected respondents (entities) are ranked according to the matching degree to the query condition instead of just a list of entities which satisfy the condition (Branco et al., 2005). In classification, flexible classes' boundaries provide a resilient method of classifying by allowing same entity to reside in multiple classes with different membership degrees (Meier and Werro, 2007). In these ways similar entities are always similarly treated and the problem of falling into an inappropriate crisp class is significantly mitigated.

Example: We want to determine whether a respondent has the high turnover. We could say that each respondent with turnover above 500 000 EUR has high turnover (SQL: *where turnover > 500 000*). However, people do not really see a difference between 498 525, 500 000 and 500 035. Data are precise but people do not see the reason for sharp classes. The second aspect of fuzziness is that 498 525 belongs to high turnover but with a slightly lover degree than 500 000 and with higher degree than 497 700 (Fuzzy query: *where turnover is high*). Crisp (sharp) and fuzzy set are depicted in Figure 1.

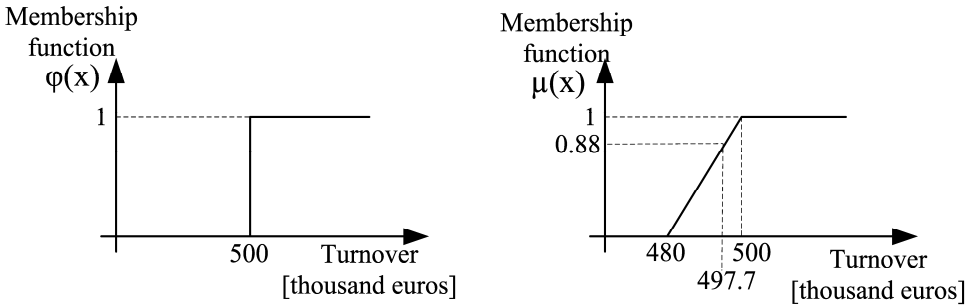

Figure 1: Comparison of traditional (crisp) and fuzzy set.

Let X be a universe of discourse. A fuzzy set A is characterized by a membership function $\mu_{A(x)}$ which associates with each element of X a real number in the interval [0, 1], i.e. $\mu_A(x): X \rightarrow [0,1]$. As the membership degree is closer to value of 1, an element more strongly belongs to the fuzzy set A.

Fuzzy logic operations are aggregations of classical operations of conjunction, disjunction and negation. In classical case one logical function exists for conjunction (*and* operator) because the elementary condition is satisfied (value of 1) or not (value of 0). In fuzzy logic there are many functions describing conjunction (these functions are called t-norms) (Zimmermann, 2001). The minimum t-norm which is widely used has the following form:

$$\mu(t) = \min(\mu_i(\mathrm{a}_\mathrm{i})), \qquad \mathrm{i} = 1,..., \mathrm{n} \tag{2.1}$$

where $\mu_i(a_i)$ denotes the membership degree of the attribute a_i to the i-th fuzzy set (i-th elementary condition) and $\mu(t)$ denotes matching degree to the fuzzy conjunction. Applicability of fuzzy logic in official statistics is examined in e.g. (Hudec, 2012).

3.1.1 Fuzzy logic in reminder letters

Reminder letters should take into account businesses' previous behaviour as respondents as well as their relevance for NSIs.

From the discussion in Slovak NSI the following rules for reminders appear. If a respondent is in a delay for e.g. one month or two it is considered as a small delay. A delay of three to four months is called a significant delay and a delay greater than five months is considered as a high delay. The relevant information is also the month of the first reporting duty. A delay could be caused by the fact that the duty to respond appears for the first time. If the task is totally new for the business, the letter should be less strict.

The value of trade could be also an interesting attribute. Non-responded unit of a high trade value could produce more significant bias than unit with small trade value. In the task we have recognized linguistic terms *small*, *significant* and *high*. It leads to the assumption that fuzzy logic could solve this issue.

Taking into account premises mentioned above, we could create several flexible queries. A query for the strong reminder letter has the following structure:

select businesses with high delay and long reporting duty and high trade value

A query for the selection of new businesses for a soft reminder letter with offering support is as follows:

select businesses with small delay and short reporting duty.

Fuzzy sets for attributes delay, reporting duty and trade value are depicted in Figure 2.

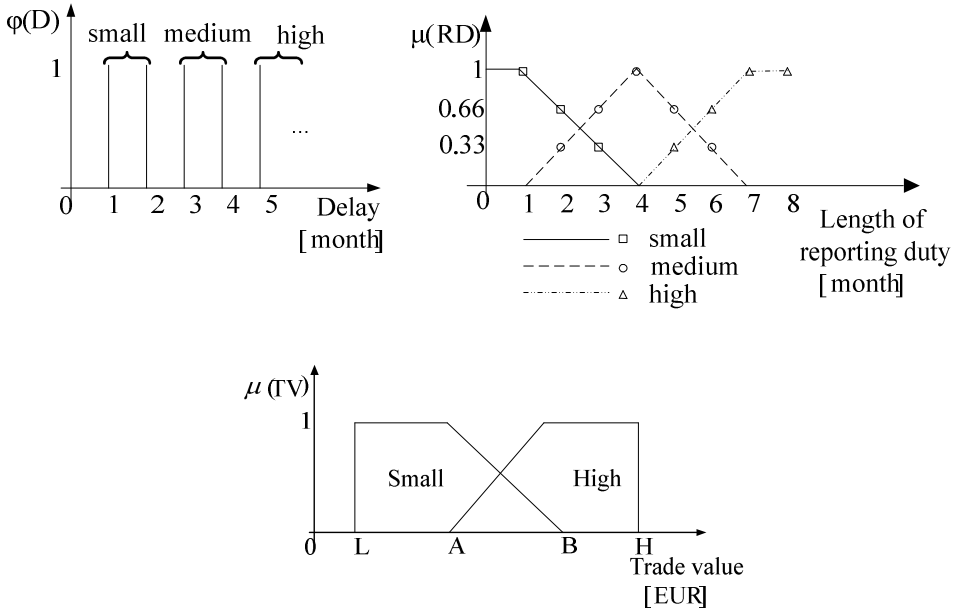

Figure 2: Indicators for reminder letters presented as fuzzy sets.

Fuzzy sets for the first and the second attribute are constructed according to the statistician's opinion (for the delay singleton fuzzy sets are used). The domain of the third attribute is theoretically $[0, +\infty]$. In practice the lowest L and the highest H values of current database content are far from the lowest and the highest domain values (Figure 2 – Trade value). In the construction of membership functions, we have to consider this fact. Ways for the construction of fuzzy sets by mining the current content of the database are examined in e.g. (Hudec and Sudzina 2012; Tudorie, 2008). For the logical *and* operator the min t-norm (2.1) is used. For selection of respondents the fuzzy query approach based on the fuzzy generalized logical condition (Hudec, 2009) has been used.

Anonymised data on foreign Intra-EU trade were provided by the Statistical Office of the Slovak Republic. Some of the selected businesses for the strong reminder letter are depicted in Table 1.

If a crisp selection were used, valuable information $\mu(t)$ would remain hidden. As a consequence businesses B4 and B5 would not be selected and therefore not reminded even though they are closer to meet the condition for the strong reminder than to meet the condition for the less strong reminder letter.

Table 1: Selected business for the strong reminder.

Business	μ (High delay)	μ (Long respondin g duty)	μ (High trade value)	Matching degree μ(t)
B1	1.00	1.00	1.00	1.00
B2	1.00	1.00	1.00	1.00
B3	1.00	1.00	1.00	1.00
B4	1.00	1.00	0.73	0.73
B5	1.00	0.66	0.90	0.66

3.1.2 Fuzzy logic for respondents' classification

In this Section we illustrate advantages of flexible classification on a small experiment. For other classification tasks, we need to select appropriate indicators, create the rule base and define fuzzy sets for each indicator but the main idea remains the same.

Let the domain, in our case study, for the attribute delay be limited by the [0, 20] interval. The domain for the attribute amount of ordered data is limited by the [0, 1000] interval. For the sake of simplicity both attributes are fuzzified into two fuzzy sets depicted in the Figure 3 together with the classification space (Meier et al., 2005).

The rule base has the following structure:

- if delay is high and amount is small then business belongs to C1;
- if delay is high and amount is high then business belongs to C2;

- if delay is small and amount is small then business belongs to C3;
- if delay is small and amount is high then business belongs to C4.

For the logical *and* operator the product t-norm (Zimmermann, 2001) is used. In case of classification the minimum t-norm (2.1) causes additional calculations: the normalisation in order to obtain sum of all membership degrees equals to 1.

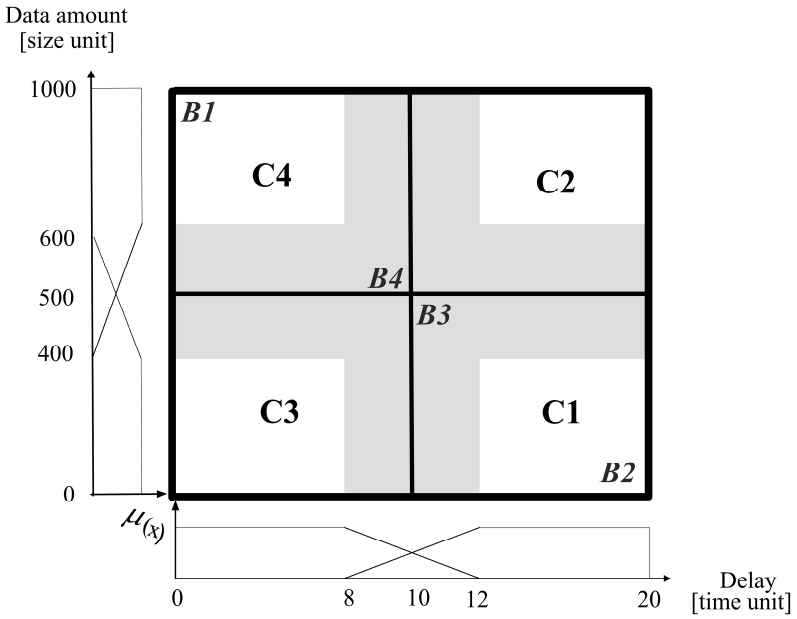

Figure 3: Classification space based on fuzzy logic (according to Meier et al., 2005).

For the classification the fuzzy query classification approach (Hudec and Vujošević, 2012) has been used. A respondent can belong to more than one class with different membership degrees (if partially satisfying more than one fuzzy rule). The rank of a respondent is calculated by the aggregation of the class coefficient where the respondent belongs (P) and its membership degree to these classes using the following equation (Hudec and Vujošević, 2012):

$$R_O = \sum_{k=1}^{K} \mu_{Ock} P_k \tag{2.2}$$

where k is the number of classes, μ_{Ock} is the membership degree of entity O to class C_k and P_k is the coefficient describing class C_k.

The percentage of fee reduction can be associated with each output class. For instance the class C1 gets 0%, the class C2 gets 5%, C3 gets 10% and C4 gets 15%. In addition, if we replace class parameters of C2 and C3, we will prefer the use of data over the delay. This way motivates respondents to reach class C4.

Let us have four businesses with the following values of delay (in the time unit) and amounts of ordered data (in the size unit) (Hudec et al., 2012): B1 (2, 900); B2 (19, 50); B3 (11, 490); B4 (9, 502).

Business B1 fully belongs to the class C4. The notation is B1:C4(1). Other businesses belong to classes in the following way:

- B2:C1(1);
- B3:C1(0.4125), C2(0.3375), C3(0.1375), C4(0.1125);
- B4:C1(0.1), C2(0.15), C3(0.3), C4(0.45).

The fee reduction *R* is presented in the Table 2.

Table 2: Businesses ranked downward from the best.

Business	**R (%)**
B1	15.00
B4	10.50
B3	4.75
B2	0.00

This classification approach offers the fairly treatment of respondents. The motive for treating gradations, as is done in fuzzy logic, among others is reducing the complexity of the mathematical analysis of real problems by classical approaches (Radojević, 2008). For example, two elements of the analysed universe (businesses in the illustrative examples) can be discerned by crisp or two-valued logic only if one belongs and another does not belong to the particular class. As a consequence, the number of necessary properties (rules in classification or attributes in queries) increases. This complexity of the problem can be reduced by including intensity of examined property.

The same idea could be basis for other classification tasks. The fuzzy classification mines respondents and rank them downwards from the best to the worst. It could be used for example in providing services above standard for most perspective businesses.

3.1.3 Perspective of estimation of missing values in small area statistics

In case of municipal statistics of Slovakia the database consists of 2891 municipalities and more than 800 indicators. Most of them are collected on the yearly basis except the indicators which contain stable values e.g. altitude above the sea level. In this case respondents are from different administrative institutions: demography, hydrometeorology, districts, municipalities, ministries. The data collection copes with issues like rare occurrence of measured indicator, instruments for measurement are not installed in all units and also with late or no response.

In most of cases, neither reminder letter nor motivation is able to solve the problem. If we focus for example, on indicators describing climate conditions we recognise some similarities between municipalities and dependencies between

their indicators. If the distance between municipalities is not high and they have similar altitude above the sea level, then climate indicators (number of days with snow coverage, number of summer days) are more or less dependent.

Recent research has shown advantages of fuzzy functional dependencies (FFD) revealed by appropriate functions of proximity, similarity and implications (Vucetic et al., 2013). If FDD reveals the rule: the majority of database entities with value around *v* of attribute *A* have value around *m* of attribute *B*, then we could estimate some missing values of attribute *B* with the probability related to the intensity of dependency. On the other hand, if FDD reveals that values of attribute *B* are significantly dependent on attribute *A* then attribute *B* is redundant and should be dropped from database and surveys. In addition the approach developed in (Vucetic et al., 2013) could work with qualitative, quantitative, textual and imprecise values of indicators. The main problem lies in the need to compare each two municipalities and each two indicators, which could cause combinational explosion (Vucetic et al., 2013). However, in this case statisticians' knowledge could significantly reduce this problem by excluding indicators which are not relevant for a particular imputation task. This idea has been recently recognised and it needs further research, experiments and comparisons with the usual tools.

3.2 Neural networks

Neural networks (NNs) can be used for an effective analysis of large databases. Their big advantage is the ability to generalize from abstract and this function can be in general used also for data classification in official statistics when the borders of classes are not exactly defined (Juriová, 2012). However, performance of the NNs is mostly dependent on the success of the training process (Kulluk et al., 2012). The process of training a NN is generally interested in adjusting the individual weights between each pair of the individual neurons. At the beginning of the learning process a dataset, which is named as a training set, is presented to the inputs to determine the correct outputs. When the learning process is finished, a testing dataset is used to evaluate the generalization capability of the classifier. Feed-forward NNs, which are also known as Multi-Layer Perceptrons (MLP), are one of the most popular and most widely used NNs models in many practical applications due to their high capability to classification and forecasting (e.g. in Aminian et al., 2006). Neural network classifies each entity into the class for which output neuron is activated (value of 1). Therefore, neural network can easily detect entities which are not correctly classified having difference between the result from neural network and the actual value in the training dataset. In the next training step network tries to fix this difference and classify the entity to the right class. Percentage of correctly classified entities in the training process depends on several attributes: size of a data set, number of neurons in hidden layers and

selected learning algorithm. A simplified feed-forward neural network is depicted in Figure 4 (Hudec and Juriová, 2013).

Neural networks are especially appropriate for recognizing patterns in the presence of noise and incomplete data sets (missing values) or making decisions for current problems based on the prior experience.

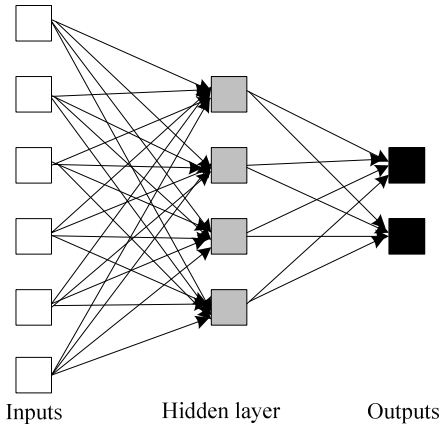

Figure 4: Example of feed-forward neural network.

3.2.1 Neural networks in foreign trade statistics

An example of using the neural network approach for the estimation of missing data is described on data surveyed in foreign trade statistics like in (Hudec and Juriová, 2013). Anonymised data on foreign trade were provided by the Statistical Office of the Slovak Republic. After reaching the exemption threshold value the company has to fulfil the declaration. However, not all respondents fulfil their duty. Individual business reports contain several items characterising their activity. The characteristics considered useful were the following 8 items: time period (month), code of goods (simplified, i.e. three-digit level), invoiced value, region of dispatch, state of destination, delivery terms, nature of transaction and mode of transport.

The usefulness of imputation can be seen e.g. for the item Nature of transaction which has nine classes altogether. For experimental purposes we have tested the classification into two classes of transaction (Kľúčik, et al., 2012): DOA4 and DOA5 where it can be distinguished between Operations with a view to processing under contract (DOA4) and Operations following processing under contract (DOA5).

In our research a feed-forward neural network with three layers is used. Our proposed and tested neural network has seven inputs, logistic activation function in the hidden layer and two outputs. Seven inputs correspond to seven different items surveyed in Intra EU trade (Intrastat) system that are available to

characterise each statistical report in the survey. The logistic activation function was chosen as this function is often successfully used in the case of classification into two classification groups. Further, two outputs stand for two classification groups. The number of hidden neurons can be selected according to the success of the training process; that is, increasing the number of neurons above a certain level does not increase the classification ability of the network. As a searching algorithm a scaled conjugate gradients algorithm was used to find a local minimum of the function.

The neural network approach for the purpose of Intrastat data classification is proposed in the following steps: dividing data into training and validating parts, allocation of training dataset into 2 classes – 1 means that a unit belongs to the class, 0 means that a unit does not belong to the class, creating the neural network, training the neural network with optimisation algorithm, classification of validating dataset by means of the trained neural network.

The output of the neural network is thus the computed probability, with which a certain data belongs to classes. The neural network is trained well if the probability for one of the classes gets close to 1, which is the maximal credibility. The result of the classification is then interpreted as data belongs with computed probability into the particular classes.

The results from the training process are included in the Table 3 (Hudec and Juriová, 2013). The best results were obtained with the neural network with 15 neurons and 1000 training cycles. The highest probability of inclusion into classes was with this neural network gained above 70% for both classes when the number of training cycles was 1000. Increasing the number of neurons in the hidden layer or increasing the number of training cycles did not result in gaining a better trained network. After the network has been trained the best one was used for the classification of the original data to verify the proposed classifier. The validating set consists of 2000 units coming from the class DOA5. The probability of inclusion into the class DOA5 proved to be 76.8%. This confirmed the ability to use the trained network for suggesting the missing values.

Table 3: Evaluation of the training process

Characteristics of the training process	Type of transaction	Probability of inclusion into the class (%)	Root Mean Square Error
10 hidden neurons, 300 training cycles	DOA4	59	0.41
	DOA5	46	0.54
10 hidden neurons, 400 training cycles	DOA4	64	0.37
	DOA5	57	0.43
15 hidden neurons, 1000 training cycles	DOA4	**71**	**0.30**
	DOA5	**77**	**0.23**

To supplement, any introduction of new methods for the purpose of missing values imputation at the NSIs needs further research of variance estimation of proposed values. When the proposed method will be improved, the variance estimation of values suggested by NNs should be computed and verified as well.

4. Concluding remarks

This paper relates some issues in data collection from the soft computing (fuzzy logic and neural networks) perspective. Advantages of soft computing have been explained and initial research has been done. Efficient approaches for reminders, motivation and estimation of missing values could improve the collection and the quality of data produced by NSIs. Data users and society in general could benefit by using NSIs' data of higher quality and earlier available. This could improve the image of NSIs and therefore motivate respondents to respond timely and accurately.

Definitely, approaches examined in this paper need further research and experiments. It especially holds for the estimation of missing values because it is a very sensitive task. In the further research comparison between traditional tools e.g. the hot deck imputation and approaches based on the NNs and FDD could be very valuable for NSIs. The main advantage of neural networks and FDD lies in the fact that they are able to autonomously or semi autonomously mine data sets and detect patterns and rules. Patterns and rules could be hidden for users especially in large sets of observations, administrative and auxiliary data.

The main obstacles are in higher computational demand for neural networks and fuzzy logic tools. Concerning the former, they are usually expensive and require trained users. Regarding the later, we need to develop full functional software tool, which is also a demanding task. The second problem lies in the fact that these approaches are not broadly used in official statistics (learned from discussion with several NSIs) and especially in the data collection. Finally, missing values are imputed with a certain degree of probability. In the traditional databases the degree of probability cannot be efficiently stored and further processed. The answer could provide the fuzzy databases (Galindo et al., 2006) which contains variety of possibilities to manage and store uncertainties on unit or item levels.

If analysed approaches offer significant improvement in some of discussed issues, the next step should be the development of full functional software tools for NSIs. In general, NSIs share common mission and obligations and share common standards for data collection. Having this fact in mind soft computing approaches without significant modifications could be applied in other NSIs.

The development of a full functional software tool is a demanding task. One possible answer is software sharing among NSIs (Lehtinen and Gløersen, 2009). Sharing of software tools, through the limited open source approaches could

reduce the development effort inside NSIs. One group of NSIs could be focused on development of some tools and other institutes will use these tools and will be able to use their resources for development of some other tools. The Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) provides standardized environment for this purpose (Seo, 2011; Vale, 2009).

Acknowledgement

The research reported herein was funded by the European Commission via the Seventh Framework Programme for Research (FP7/2007-2013) under Grant agreement n°244767. This work was supported by the Slovak Research and Development Agency under the contract No. DO7RP-0024-10.

References

[1] Aminian, F., Suarey, E.D., Aminian, M. and Walz, D.T. (2006): Forecasting Economic Data with Neural Networks. *Computational Economics*, **28**, 71-88.

[2] Andridge, R. and Little, R. (2010): A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, **78**(1), 40–64.

[3] Bavdaž, M. (2010): Sources of Measurement Errors in Business Surveys. *Journal of Official Statistics,* **26**(1), 25-42.

[4] Branco, A., Evsukoff, A.., and Ebecken, N. (2005): Generating Fuzzy Queries from Weighted Fuzzy Classifier Rules. Proceedings of the ICDM workshop on Computational Intelligence in Data Mining, Houston, 27 November, 21-28.

[5] Berzal, F., Blanco, I., Sánchez, D., Serrano, J.M., and Vila, M.A. (2005): A definition for fuzzy approximate dependencies. *Fuzzy Sets and Systems*, **149**(1), 105-129.

[6] Coutinho W. and de Waal T. (2012): Hot deck imputation of numerical data under edit restrictions. Discussion paper, Statistics Netherlands.

[7] De Leeuw, E., Hox, J. and Huisman, M. (2003): Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics*, **19**(2), 153-176.

[8] European Plan of Research in Official Statistics (EPROS). (2007): Main conclusions from the activities in the 5th Framework Programme. Office for Official Publications of the European Communities, Luxembourg, Luxembourg.

[9] Feil, B. and Abonyi, J. (2008): Introduction to fuzzy data mining methods. In J. Galindo (Ed.), *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 55-96). London: IGI Global.

[10] Galindo, J., Urrutia, A. and Piattini, M. (2006): *Fuzzy databases: Modeling, Design and Implementation*. Hershey: Idea Group Publishing Inc.

[11] Giesen, D. (ed.) (2011): Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes. Blue-ETS Project Report.

[12] Hand, D., Mannila, H. and Smyth, P. (2001): *Principles of Data Mining*. Cambridge: The MIT Press.

[13] Hassani, H., Gheitanchi, S. and Yeganegi, M.R. (2010): On the Application of Data Mining to Official Data. *Journal of Data Science*, **8**, 75-89.

[14] Hudec, M. and Juriová, J. (2013): Evaluation and checking non-response data by soft computing approaches - case of business and trade statistics. Proceedings of the New Techniques and Technologies for Statistics (NTTS 2013), Brussels, 5-7 March.

[15] Hudec, M. (2013): Fuzzy database queries in official statistics: Perspective of using linguistic terms in query conditions. *Statistical Journal of the IAOS* (forthcoming).

[16] Hudec, M. (2009): An Approach to Fuzzy Database Querying, Analysis and Realisation. *Computer Science and Information Systems*, **6**(2), 127-140.

[17] Hudec, M., and Sudzina, F. (2012): Construction of fuzzy sets and applying aggregation operators for fuzzy queries. Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS 2012), Wroclaw, 28 June - 1 July, Proceedings volume 1, 253-257.

[18] Hudec, M. and Vujošević, M. (2012): Integration of data selection and classification by fuzzy logic. *Expert Systems with Applications*, **39**(10), 8817–8823.

[19] Hudec, M., Balbi, S., Juriová, J., Kľúčik, M. Marino, M., Scepi, G., Spano, M., Stawinoga, A., Tortora, C. and Triunfo, N. (2012): Report on Principles of Fuzzy Methodology and Tools Developed for Use in Data Collection (Soft Computing and Text Mining tools for Official Statistics). Blue-ETS Project Report.

[20] Juriová, J. (2012): Neural Network Approach Applied for Classification in Business and Trade Statistics. Proceedings of the 46th scientific meeting of the Italian statistical society, Rome, 20-22 June.

[21] Kľúčik, M. (2011): Introducing New Tool for Official Statistics: Genetic Programming. Proceedings of the New Techniques and Technologies for Statistics (NTTS 2011), Brussels, 22-24 February.

[22] Kľúčik, M. (2012): Estimates of Foreign Trade Using Genetic Programming. Proceedings of the 46th scientific meeting of the Italian statistical society, Rome, 20-22 June.

[23] Kľúčik M., Hudec, M. and Juriová, J. (2012): Final Report on the Case Study Results on Usage of IT Tools and Procedures Developed for Data Collection (Soft Computing tools for Official Statistics). Blue-Ets Project Report.

[24] Kulluk, S., Ozbakir, L. and Bayakasoglu, A. (2012): Training neural networks with harmony search algorithms for classification problems. *Engineering Applications of Artificial Intelligence*, **25**, 11-19.

[25] Lehtinen, H. and Gløersen, R. (2009): Cooperation in development of open source software. Proceedings of the Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), Oslo, 18-20 May.

[26] Meier, A. and Werro, N. (2007): A Fuzzy Classification Model for Online Customers. *Informatica*, **31**, 175–182.

[27] Meier A., Werro, N., Albrecht, M. and Sarakinos, M. (2005): Using a Fuzzy Classification Query Language for Customer Relationship Management. Proceedings of the Conference on Very Large Data Bases, Trondheim, 30 August – 2 September, 1089-1096.

[28]Radojević, D. (2008): Real sets as consistent Boolean generalization of classical sets. In L.A. Zadeh, D. Tufis, F. Filip, FG. Diztac (Eds), *From natural language to soft computing: New paradigms in artificial intelligence* (pp. 150-171). Bucharest: Editing House of Romanian Academy.

[29] Seo, C. (2011): Development of the Generic Statistical Information System. Proceedings of the Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), Luxembourg, 7-9 April.

[30] Tudorie, C. (2008): Qualifying Objects in Classical Relational Database Querying. In J. Galindo (Ed): *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 218-245). London: IGI Global.

[31] Torres van Grinsven, V., Bolko, I. and Bavdaž, M. (2012): Sources of Motivation in Business Surveys. Proceedings of the Fourth International Conference on Establishment Surveys (ICES IV), Montréal, 11-14 June.

[32] de Waal, T., Pannekoek, J. Scholtus S. (2011): *Handbook of Statistical Data Editing and Imputation*. New Jersey: John Wiley & Sons.

[33] Vale, S. (2009): Developing a Generic Statistical Business Process Model. Proceedings of the Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), Oslo, 18-20 May.

[34] Vucetic M., Hudec, M. and Vujošević, M. (2013): A new method for computing fuzzy functional dependencies in relational database systems. Expert Systems with Applications, **40**(7), 2738–2745.

[35] Zadeh, L.A. (1994). Soft Computing and Fuzzy Logic. IEEE Software, **11**(6), 48-56.

[36] Zadeh, L.A. (1965): Fuzzy sets. *Information and Control*, **8**, 338-353.

[37] Zimmermann, H.-J. (2001): *Fuzzy Set Theory – and Its Applications*. London: Kluwer Academic Publishers.