Distr.
GENERAL

Working Paper
10 April 2013

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (iii): Innovation

Improvement of data collection and dissemination by fuzzy logic

Prepared by Miroslav Hudec, Infostat, Slovakia

I. Introduction

1. Data collection and data dissemination although on two different ends of the statistical data production influences each other. For respondents it might not be easy and straightforward to query and find relevant data in the statistical databases on the (National Statistical Institutes) NSIs' websites (Bavdaž et al, 2011). As a consequence motivation to cooperate in surveys might decreases and amplified by feeling of response burden causes higher frequency of non-response.

2. People rely on common sense and use linguistic terms when they solve the problems or search for data and information. Users on website want to make a selection on the basis of several criteria at the same time and prefer to see selected entities (e.g. municipalities) downwards from the best to the worst. We need a tool capable of giving answers to imprecise database questions. This kind of tool can solve more user demands and therefore improve image of NSIs and international statistical organizations as data providers.

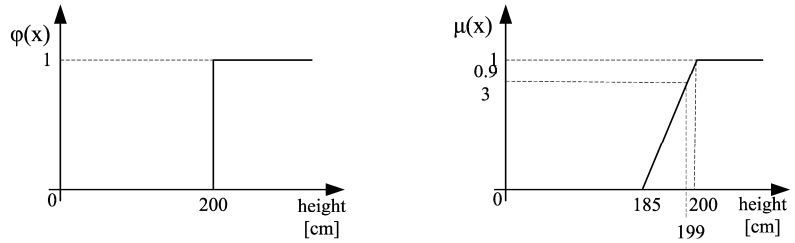
3. Data collection could be improved through evaluation of estimated values and creation of tailored rewards for key respondents/data users. If estimated and collected data share similar properties (distribution) then algorithm for estimation missing values works properly. If we express this task by linguistic terms and quantifiers then it is better understandable for users. Furthermore, identifying key customers (data users and respondents) and ensuring that similar respondents are always similarly treated (rewarded) when use statistical data could brings benefits for both sides.

4. In all this tasks fuzzy logic is an option which might offer the solution. Fuzzy logic is able to capture experts' knowledge which is often expressed by ambiguous terms and uncertainties and directly apply on databases.

II. Fuzzy logic in brief

5. The concept of fuzzy sets was initially introduced in (Zadeh, 1965) where it was observed that precisely defined criteria of belonging to a set often could not be defined. The fuzzy logic is an approach to computing based on degrees of truth rather than the usual true or false logic.

6. Let's say people having 200 cm and more are considered to be high. Query (select people where height >= 200) treats people with 199 cm in the same way as people having 170 cm. In fuzzy logic (select people where height is high) we can say that for example 198 cm significantly belong to this concept whereas height of 190 cm belongs with lower intensity. Membership degrees for sharp set and fuzzy set are depicted in Figure 1. The same holds for linguistic terms *small*, *medium* and *about*. Membership functions created on linguistic terms have overlapping boundaries that lead to smooth data analysis (similar objects are similarly treated).



a) crisp, sharp, usual set b) fuzzy set

Figure 1: Crisp (usual) and fuzzy set

III. Data dissemination

7. People prefer to use expressions of natural language in searching for useful data and information. Data required for these processes are mainly stored in relational databases. And there is a problem. People are familiar with linguistic terms e.g. *high unemployment rate*, *the majority of*, *about 20*, *low temperature*, etc. which describe particular objects e.g. territorial units. These terms include a certain vagueness or fuzziness that information systems based on two-valued logic {true, false} do not understand and therefore cannot use (Galindo et al, 2006). The presentation of results in a useful and understandable way is very important for all organisations dealing with data and information dissemination. For example, we would like to avoid a long list of retrieved entities, without any helpful ordering.

8. For the purpose of experiments data from the Urban and municipal statistics database of the Statistical Office of the Slovak Republic (SO SR) is used (2891 municipalities and more than 800 indicators).

A. Usual query improved with the fuzzy logic

9. In the following example the municipalities with *altitude about 700 meters above sea, small distance to the nearest train station, high number of beds in accommodation facilities and small population* are sought. Let's say tourists looking the destination for a calm holiday.

10. In case of fuzzy queries, the main question from users' point of view is defining parameters of fuzzy sets *small*, *medium* and *about* (Figure 2). We have created way by mining database which help users to select these parameters. Parameters are calculated form the current database contend and offered to users. Therefore users can modify these parameters, if they are not satisfied with suggested ones, before running a query. If users are familiar with the context of query their can write parameters directly into the interface.

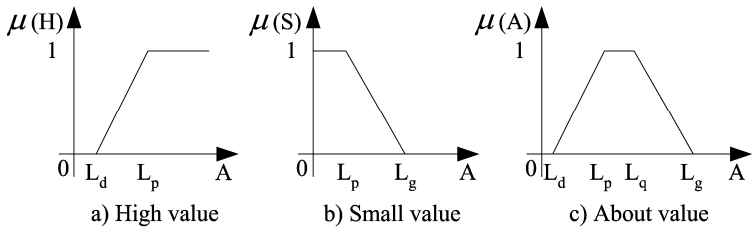


Figure 2: Fuzzy sets

11. The experimental interface has been adjusted to fit to the data retrieval process from statistical database dealing with the territorial units. The query selects 11 municipalities from the database. The Figure 3 shows two municipalities fully satisfying the query; one municipality is very close to satisfy the query and another nine municipalities partially satisfy the query condition. If SQL were used, this additional valuable information would remain hidden. The lower right part can be used for other types of presenting retrieved data and information, e.g. the answer could be presented on thematic map. Territorial units which fully satisfy the query criterion can be marked with one colour, territorial units which do not satisfy query are marked with second colour and territorial units which partially meet the query condition are marked with the third colour having a colour gradient from a faint hue to deep hue.

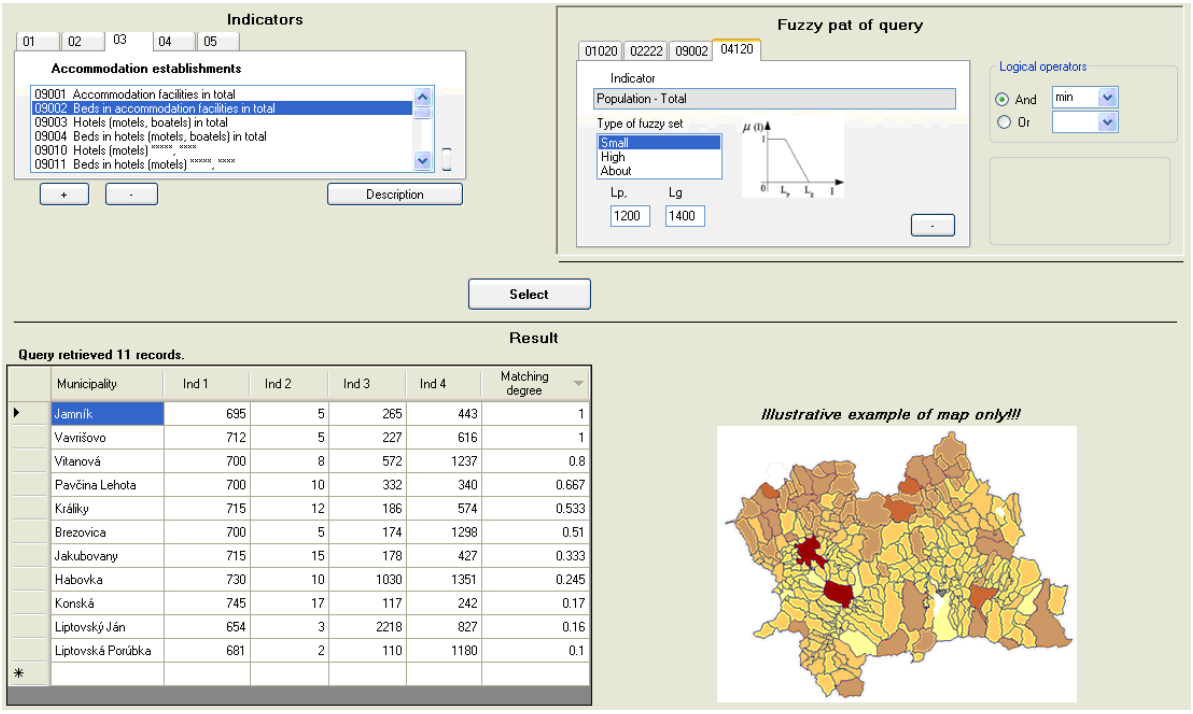


Figure 3: Flexible queries

12. Just imagine that the two municipalities with membership degrees equal to value of 1 do not exist in database. What is the result of the traditional SQL query? No data. In some cases it is informative enough but in other cases users would probably want to know why empty answer happened and how far are municipalities to meet the query condition. In that case user is forced to create another query.

B. Similarities

13. Linguistic terms ‘more or less equal to’ are employed to find municipalities in a database with the same or similar values as indicators of selected municipality.

14. In this example we are interested to find whether municipalities with similar values of three indicators (altitude, total area size and population density) as the municipality Rohovce exist. The interface is shown in Figure 4. The procedure works in the following way: the user chooses one municipality from the list of municipalities and relevant indicators for calculation of similarities. Consequently, the software finds values of selected indicators for the chosen municipality and suggests triangular fuzzy "About value" set with the same value of L_p (Figure 2 with L_p=L_q) as for chosen municipality. If the users want to change these parameters their only needs to replace suggested values.

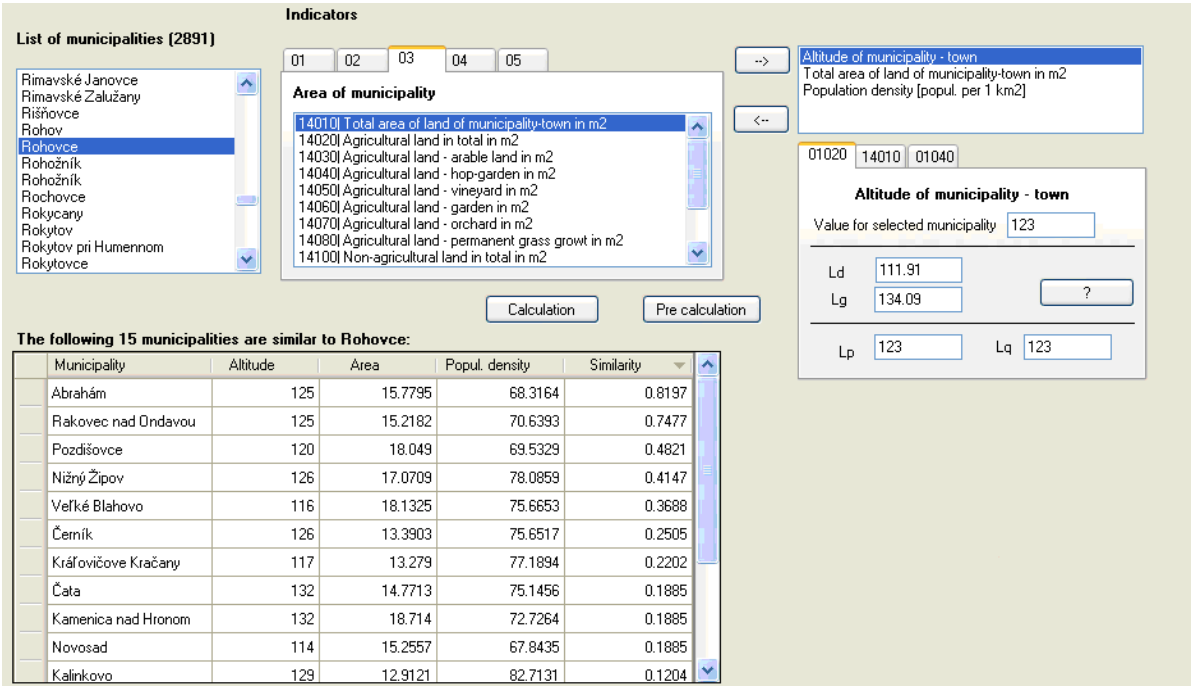


Figure 4: Similarity

C. Linguistic summaries as aggregators

15. In this kind of query data are not presented to users, only the mined information. A sharp (crisp) rule is either fully satisfied or fully rejected. If a rule is rejected, we are not sure whether the rule is about to be satisfied or whether the data are far away from the rule condition. Fuzzy rules can use linguistic terms like *small*, *medium*, *high* and quantifiers *most*, *about half* and *few* among others. Construction of fuzzy sets parameters is explained in (Hudec, 2012).

16. For example we want to know to which extent is the following rule (query) satisfied *most of municipalities has small attitude above sea level*. Answer could be obtained for each region or whole country. The query and the result for all eight regions of the Slovak Republic are presented in Figure 5. It is obvious from Figure 5 that regions 1, 2 and 4 are flat whereas regions 5 and 7 are hilly. Region 3 is more flat than hilly. The same holds for region 8 but it is a slightly hillier than region 3.

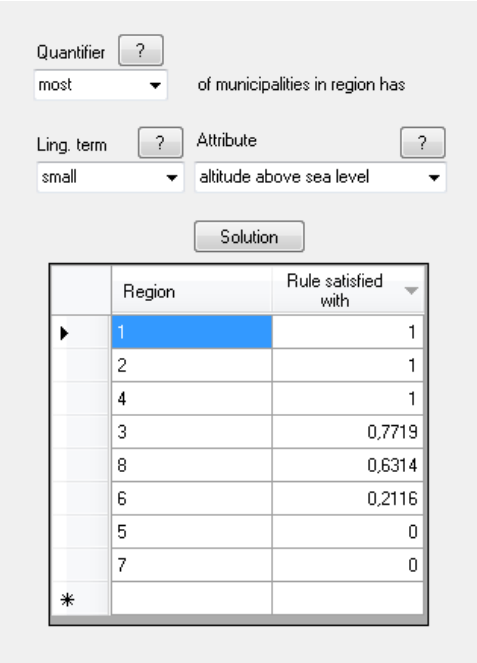


Figure 5: Fuzzy (flexible) rules

IV. Data collection

17. Statistical offices are crucial institutions for collecting data about various aspects of society. However, data collection copes with the non-response to surveys and therefore missing values. Efforts focused on increasing response rates and the estimation of missing values are topics which need continual improvement. This section presents advantages of fuzzy logic for evaluation of imputed values and in classification of respondents.

A. Evaluation of imputed values

18. After the estimation of missing values, evaluation of these values and their comparison with surveyed could reveal useful information for statisticians. If collected and imputed data share similar properties then algorithms for estimation missing values work properly. Fuzzy logic and fuzzy rules are able to evaluate similarity between these two kinds of data.

19. For the experiment anonymised data on the Intrastat trade were provided by the SO SR. Database contains an attribute which indicates whether the row is collected or estimated. It helps easier evaluating, because the structure of database is the same for collected and estimated values. The SO SR receives information from the administrative source about realised trades. Respondents are obliged to send additional data about realised trade to the SO SR. If a respondent do not respond then required parameters of its trade like country of dispatch, number of items (goods) in dispatch, amount of goods are estimated. The current algorithm finds trades which have some common features as with the missing one, and therefore, these values are filled in.

20. Fuzzy rules of same structure as in Section III. C can be used to reveal properties of collected and estimated values. For the purpose of our research we could create pair of rules (Kľúčik et al, 2012):

most (about half, few) of non-responded exports have small (medium, high) number of items (goods) in report
and
most (about half, few) of responded exports has small (medium, high) number of items (goods) in report

and run the rules on two parts of database: estimated data and collected data respectively. Results are truth values of rules. If truth values of both rules gravitate to each other, then both parts of database have similar

properties (data distribution), which mean that the current algorithm works properly. The opposite result suggests that algorithms should be improved.

21. For this purpose we have suggested an interface and procedures inside the tool. The interface organises rules as it is depicted in Figure 6. Our idea was to offer the users a possibility to create their own rules.

22. The first kind of rules evaluates single truth value of distribution of selected attribute (e.g. number of items in report) for imputed values (non-response) and compare result of the same rule for responded data (the upper left part of the interface). Using respective combo boxes we are able to create a particular rule using quantifiers *most*, *about half* and *few*, linguistic terms *small*, *medium* and *high* and selecting one of available database attributes. The second rule (for collected data) is a read only rule. It ensures that the same rule structure is applied to both parts of the database.

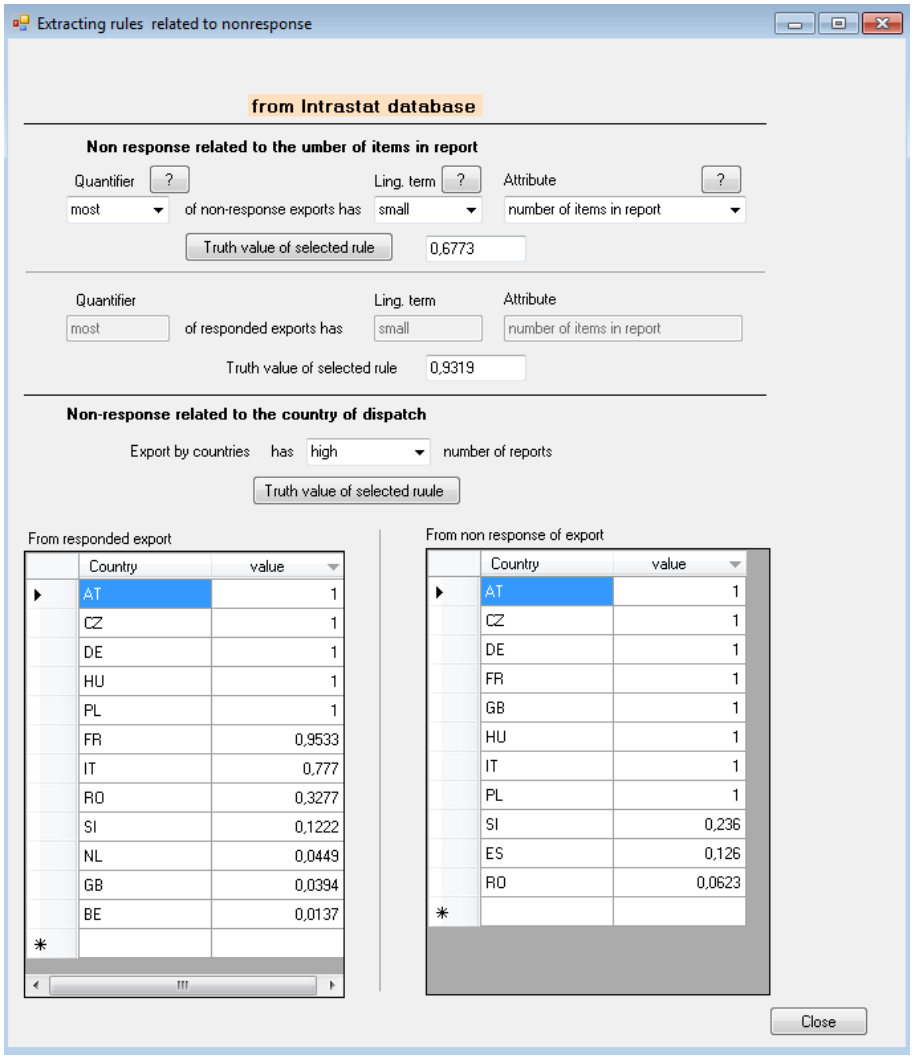


Figure 6: Interface

An example of created rule is the following:
most of non responded exports has small number of items in report.
The tool calculated the truth value of rule is 0.6773.

We applied the same rule on responded data:
most of responded exports has small number of items in report
in this case the rule is satisfied with the value of 0.9313.

23. If we compare this truth value with the truth value of rule on non-responded data, we reveal significant difference in rule satisfaction. We could conclude that the distribution is quite different which implies the current algorithm for estimation should be improved.

24. The second kind of rule examines distribution of all countries of dispatch (the lower part of the same interface - Figure 6). The Intrastat database covers trade among EU countries. It means that currently 26 countries appear in Slovak Intrastat database as destinations of goods.

Designed rule for this purpose is as follows:
export by countries has high number of reports.

25. This rule shows us how data for countries to which export is most frequent are distributed. Table 1 depicts most frequent export countries from declarations obtained from respondents. We have evaluated the same rule for the part of the database that consists of estimated data and obtained the result depicted in Table 2.

Table 1: Countries with high number of reports – surveyed data

Country	**High number of reported trade**
AT	1
CZ	1
DE	1
HU	1
PL	1
FR	*0,9533*
IT	0,777
RO	0,3277
SI	0,1222
NL	0,0449
GB	0,0394
BE	0,0137

Table 2: Countries with high number of reports – estimated data

Country	**High number of estimated trade**
AT	1
CZ	1
DE	1
FR	*1*
GB	1
HU	1
IT	1
PL	1
SI	0,236
ES	0,126
RO	0,0623

26. Table 1 reveals that the rule is fully satisfied for five countries and partially for another seven. If crisp rule were used, countries from FR downward would not be selected even FR is about to meet the condition. This relevant information would remain hidden. At the first glance it is obvious that both parts of database have similar properties. Strength of fuzzy rule is obvious in case of FR. We see that FR has a value of 1 (Table 2) and almost 1 (Table 1). In addition, crisp case will select FR only for non-response which might lead to the conclusion that the used algorithm for estimation values could be improved. Tables 1 and 2 bring us useful information for the conclusion that the current algorithm properly evaluates missing values for countries of dispatch.

B. Classification of respondents

27. Fuzzy classification can help identifying key customers (data users and respondents) and reveal their potential and weaknesses so as to ensure that similar customers/data providers are always similarly treated e.g. rewarded. Flexible classes provide a resilient method of classifying by allowing to same entity to reside in multiple classes with different membership degrees.

28. Businesses often play role as respondents to surveys and users of statistical data. In NSIs often data are not free or at least businesses have to pay a service charge for the data preparation to the required structure. Service charge and fee depend on amount of ordered data. This information could be a relevant input for tailored reward: Motivation of respondents to participate in surveys by discounts of provided services. To

simplify we use only the delay in response and the amount of ordered data to create a tailored reward. In experiment data are fictional but NSIs can store and use these records.

29. Traditional classification approaches imply two shortcomings which conflict with the human reasoning: on the one hand, businesses with similar characteristics can be classified in different classes; on the other hand, businesses with different behaviour can belong to the same class (Werro et al, 2005). Shortcomings could be solved by other approaches but fuzzy logic approach is easier to use.

30. In fuzzy classification rules are created by linguistic terms which makes them understandable and easy to maintain. Secondly, a small number of rules are capable to deal with smooth distinctions between businesses. Let the domain 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 6 together with the classification space (Meier et al, 2005). The rule base has the following structure:

(a) If delay is high and amount is small then business belongs to C1;
(b) If delay is high and amount is high then business belongs to C2;
(c) If delay is small and amount is small then business belongs to C3;
(d) If delay is small and amount is high then business belongs to C4.

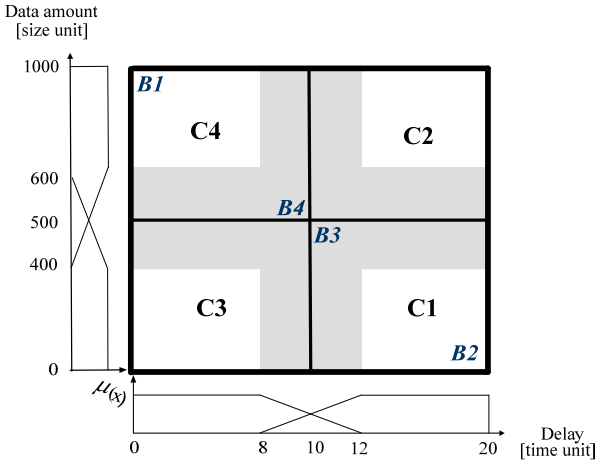


Figure 6: Fuzzy classification space

31. 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. The rank of a business is calculated by the aggregation of the classes' coefficients where the business belongs to and the business's membership degree to these classes. The obtained solution for selected businesses is presented in Table 3

Table 3. Classified businesses

Business	Delay	Data amount	Fee reduction
B1	2	900	15
B4	9	502	10.5
B3	11	490	4.75
B2	19	50	0

32. The flexible classification leads to the transparent and fair judgment (Meier et al, 2005). B1 gets the maximum reduction of the fee and at same time the higher fee reduction than B4, although both belong to the class C4, but with different degrees. (In a classical classification both obtain the same reduction). B3 and B4 obtain smaller difference of reduction although they belong to two different classes. In classical classification

they obtain 0% and 15% respectively. B2 which is in the same class as B3 (in classical case) does not have a benefit from the fee reduction.

33. For classification we can also use categorical data e.g. business opinions about NSIs that could be collected by interviews (Giesen, 2013). Fuzzy logic is capable to capture categorical and numeric data in the same way because fuzzy sets could be created on both kinds of domains.

V. Applications in NSIs

34. The fuzzy Generalized Logical Condition (GLC) which converts fuzzy queries to crisp ones and therefore select relevant data from databases in usual SQL way has been created in (Hudec, 2009). For the classification, we have extended the GLC (Hudec and Vujošević, 2012). In this way, no modification inside databases has to be undertaken.

35. Although all three kinds of approaches are based on fuzzy logic, there are some differences in ways of their applicability. For data dissemination interface should be created as web application connected to particular database in order to allow choosing existing indicators.

36. For the evaluation of estimated values data and rule extraction the tool could be applied as a standalone tool. This tool could be activated after the imputation process in order to evaluate results of data imputation. In addition, this tool could be very useful in NSIs for data analysis.

37. For classification the construction of full software is more demanding and should include experts working with respondents/data users in order to create the classification space and a full functional tool. This tool should also work with categorical data.

VI. Conclusion

38. Respondents have to respond to many surveys. If they can find relevant data and information on NSIs' websites in a user friendly way they will be more willing to cooperate in surveys. Providing the same functionality for general public could improve the image of NSIs and international organizations as an interesting source of data and information.

39. Fuzzy approach introduces an additional computation due to the substantial amount of calculations (membership degree to fuzzy sets, quantifiers and matching degrees). We need to emphasize that this additional amount of calculation is balanced with additional valuable information obtained from the database in a way that is more suitable for users. If task requires sharp query conditions, then the SQL is better solution. In data classification, contrary to data selection, fuzzy approach offers faster processing due to a significantly lower number of rules (in fuzzy classification not only rules but also intensities of matching them are included). This is the main reason for an expansion of fuzzy rule based systems in many areas from controlling of technical systems to support decision making.

40. Relevant equations, models and experimental tools have been created in order to evaluate pros and cons. Preliminary results are also presented on small-scale case studies on official statistics data (Hudec et al, 2012; Kľúčik et al, 2012). The next step is creation of a framework for the further development of full functional tools which could be applied in several parts of the Generic Statistical Business Process Model.

41. Modernization of the first and the last stage of data collection could create a chain reaction of improvements in data quality. Better data dissemination could motivate respondents to provide their own data timely and accurately and reduce the frequency of missing values implying more efficient imputation (less missing values and powerful imputation tools). Finally, better and earlier data will be available for dissemination (websites) or exchange among institutes (e.g. by SDMX).

Acknowledgements

This work was partially inspired and performed as part of the Blue-ETS project (funded by the European Commission via the Seventh Framework Programme for Research (FP7/2007-2013) under Grant agreement n°244767 and supported by the Slovak Research and Development Agency under the contract No. DO7RP-0024-10).
In particular I would like to thank Jorgen Mortensen at CEPS for providing feedback and suggestions for improvement of the research and dissemination of results.

VII. References

Bavdaž, M., Biffignandi, S., Bolko, I., Giesen, D., Gravem, D.F., Haraldsen, G., Löfgren, T., Lorenc, B., Persson, A., Mohoric Peternelj, P., Seljak, R., Torres van Grinsven, V. (2011). Final report integrating findings on business perspectives related to NSIs' statistics. Deliverable 3.2., Blue-Ets Project, http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable3.2.pdf

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

Giesen D., Bavdaž M., Bolko I. (Eds.) (2013). Comparative report on integration of case study results related to reduction of response burden and motivation of businesses for accurate reporting. Deliverable 8.1., Blue-Ets Project, waiting for the final approval.

Hudec M., Balbi S., Juriová J., Kľúčik M., Marino M., Scepi G., Spano M., Stawinoga A., Tortora C. 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). Deliverable 5.1., Blue-Ets Project, waiting for the final approval.

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

Hudec M. (2012). Dynamically modelling of fuzzy sets for flexible data retrieval. 46th scientific meeting of the Italian statistical society (SIS 2012), Rome.

Hudec M. (2009). An approach to fuzzy database querying, analysis and realisation, Computer Science and Information Systems 6(2):127-140.

Kľúčik M., Hudec M., 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). Deliverable 8.3., Blue-Ets Project, waiting for the final approval.

Werro, N., Meier, A., Mezger, C., Schindler, G., 2005. Concept and Implementation of a Fuzzy Classification Query Language. International Conference on Data Mining, Houston.

Meier A., Werro N., Albrecht M., Sarakinos M. (2005). Using a Fuzzy Classification Query Language for Customer Relationship Management. Conference on Very Large Data Bases, Trondheim.

Zadeh L. A. (1965) Fuzzy sets. *Information and Control*, 8, 338–353.