Austrian Journal of Statistics
April 2014, Volume 43, 63–75.
http://www.ajs.or.at/

Linguistic Summaries Applied on Statistics - Case of Municipal Statistics

Miroslav Hudec
University of Economics in Bratislava, Slovakia

Abstract

Data collection in official statistics copes with missing values. In the municipal statistics we can recognize more or less similar municipalities and more or less dependent indicators. Therefore, an approach capable to process this uncertainty is desirable. Data produced in official statistics is a valuable source for users. Data dissemination which mimics human reasoning in searching and evaluating data could be a suitable solution. Hence, both processes could be improved by linguistic summaries which are based on the fuzzy logic. Finally, the paper discusses future research and development topics in these fields.

Keywords: missing values, data dissemination, data analysis, municipal statistics, fuzzy logic.

1. Introduction

Generally speaking, the mission of National Statistical Institutes (NSIs) is to collect data about various aspects of society, process them and disseminate to a variety of users. Policy decisions significantly depend on statistical data. This data is also a valuable source for businesses decisions.

However, data collection copes with the problem of missing values. Therefore, efforts focused on estimation of missing values should be continuously improved (De Leeuw, Hox, and Huisman 2003; Kľúčik 2012). In e.g. municipal statistics missing values are mainly due to the rare occurrence of a measured phenomenon, non-availability of instruments to measure values in all units, and late or no response. It implies that reminders and fines are not the solution as it is in e.g. enterprise and trade statistics. In municipal statistics we could recognise similarities between municipalities and dependencies between measured phenomena (indicators). It appears that, approaches which are able to process intensities of similarities and dependencies are promising. In hot deck imputation method each missing value is replaced with data from a more or less similar unit using the linear restriction rules (Coutinho and de Waal 2012). In this direction Linguistic Summaries (LSs) (Rasmussen and Yager 1997; Kacprzyk and Zadrozny 2009; Hudec 2013c), which operate on fuzzy logic, could also offer the solution (Hudec, Balbi, Juriová, Kľúčik, Marino, Scepi, Spano, Stawinoga, Tortora, and Triunfo 2012; Hudec 2013a). In addition, LSs are not limited to the linear rules.

In dissemination NSIs should provide tools which are able to process users' imprecise queries

expressed by linguistic terms and provide not only data but also summarized information about statistical data (Hudec 2013b). According to (Bavdaž, Biffignandi, Bolko, Giesen, Gravem, Haraldsen, Löfgren, Lorenc, Persson, Mohoric Peternelj *et al.* 2011) users of statistical data are interested either in raw data (large businesses) or aggregated information (small businesses). In the former, when users cannot unambiguously define boundary between relevant and not relevant data, fuzzy queries are suitable. In the later, linguistic summaries are able to offer summarized information which is in many cases sufficient for users.

LSs have been developed to express relational knowledge about the data (Rasmussen and Yager 1997) and its intensity in a useful and understandable way. A linguistic summary is a short sentence that describes relational knowledge in large data sets (Hudec 2013c). LSs are of structure *Q entities in database are (have) S* where *S* is a summarizer defined as linguistic term on the domain of examined attribute and *Q* is a fuzzy quantifier. An example of an elementary linguistic summary is *most municipalities have small pollution.* Linguistic summaries could be more complex e.g. *most highly situated (altitude above the sea level) municipalities have small migration.* The second structure is further examined in the paper. The truth value of a LS is called validity and gets values from the [0, 1] interval. Data summarization is one of basic capabilities needed to any "intelligent" system (Kacprzyk and Zadrozny 2009).

The main intent of the paper is to discuss opportunities of LSs in estimation of missing values, analysis and dissemination in municipal statistics. Section 2 shortly describes the municipal statistics and specific problems that LSs can help solve. Section 3 explains basic concepts of LSs which are required for this paper. Section 4 is devoted to problems in imputation and dissemination with touch on data analysis which could be solved by LSs. Section 5 concludes this paper and discusses challenges for the future research.

2. Some issues in the municipal statistics

Next two sub sections discuss some issues in imputation and dissemination of data and aggregated information in the municipal statistics.

2.1. Estimation of Missing Values

In the municipal statistics missing values are due to the fact that data is not available because of several reasons (rare occurrence of a measured phenomenon, non-availability of instruments to measure phenomena in all units, reluctance of administration units to cooperate in data collection, etc.). The Slovak municipal statistics currently collects 804 indicators for 2891 municipalities. The majority of indicators are collected on yearly basis except the indicators which contain stable values for a long period e.g. the altitude above the sea level and the year of the first written notice. In the collection phase missing values occur. However, no further imputation is realised for all indicators, presumably due to large number of small municipalities and not the same relevance of all indicators. Missing values could cause some issues in data analysis and dissemination. In case of database queries, we are not sure whether a non-selected municipality has indicator's value far from the query condition or because the value is missing. In case of classification, municipalities with missing values cannot be properly classified.

In the municipal statistics relations between municipalities and indicators are usually less complex. We could recognise some similarities between several municipalities (altitude above sea level, distance, population density, ...) and dependencies between measured indicators. If for example municipalities have similar population density and similar ratio of the built-up area and yard then we could expect that the waste production is also similar. The same holds for indicators describing the distance between municipalities and measured climatic indicators. Furthermore, high dependency (intensity of a relation) usually exists in parts of domains of considered indicators. Moreover, these parts of domains often do not have clear

boundaries. Figure 1, where entities are sorted from the smallest value (t_1) to the highest value (t_n) of examined attributes, depicts possible dependencies. Therefore, if a relation is very strong between small values of attribute A and high values of attribute B then we could roughly estimate missing value with the probability related to the intensity of dependency. If a relation is not sufficiently significant, we could divide domain into more parts and evaluate validity of a relation between e.g. very small values of attribute A and high values of attribute B. Roughly speaking, this is similar scenario as in hot deck imputation method that uses

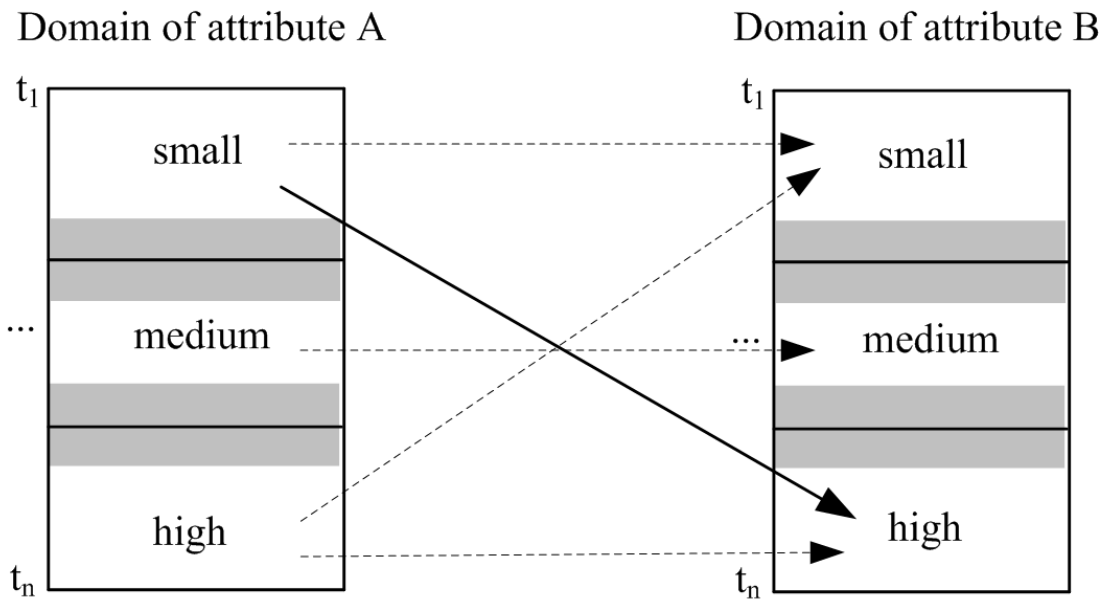

Figure 1: Relations between attribute A and attribute B

the data from other surveyed observations. In the hot deck method each missing value is replaced with data from a more or less similar unit using 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 other methods (Andridge and Little 2010). Contrary, theory of fuzzy logic and fuzzy sets is well developed but is rarely used in official statistics (Hudec *et al.* 2012). Furthermore, LSs are not limited to the linear constraints and therefore they could cope with variety of relations among data.

2.2. Data dissemination

NSIs offer their data to users by data portals or other services (either free or paid). The chief goal of the dissemination policy at NSIs should be satisfied users (Bavdaž *et al.* 2011). For example, the dissemination policy handbook of (Statistics Norway 2007) states that official statistics is a common good for society and should be available to everyone and main result should be presented in a way that makes them understandable for a broad variety of users including non-experts and lay audience. (Bavdaž *et al.* 2011) stated that smaller businesses seem to rather look for information (instead of data) and prefer simple presentation and short descriptions while larger businesses seem to favour raw data to analyse them on their own. Furthermore, users cannot know in advance which part of the data set is the most suitable for their purposes. A possible suitable solution for this problem is to reveal "abstracts" from specific parts of a large data set.

On the other side of statistical data portals are people. People use their natural language in communication and searching for useful information. Human approximate reasoning, although without precise measurements, is a very powerful way for finding answers. "Computing, in its usual sense, is centred on manipulation of numbers and symbols" (Zadeh 1999). In contrast,

computing with words is inspired by the remarkable human capability to perform a wide variety of tasks without precise measurements and computations (Zadeh 1999).

LSs mimic human reasoning in looking for the information by expressions of linguistic terms. By LSs a user obtains the aggregated summarised information ("abstract" of a data set). This abstract in some cases contains the final information. In other cases, this abstract has similar meaning as an abstract of a paper (abstract inform potential readers if the paper could meet their interest and therefore should be ordered). Analogously, abstracts calculated by LSs inform users which group of data could be relevant for them.

For dissemination, several approaches have been suggested such as WEB 2.0 (Smith 2011) or visualizing selected indicators of territorial units on maps (Jern, Haldorson, and Thygesen 2011) for telling a geo visual analytics story about the regions. The eye tracking method (Wulff 2007) can monitor users' behaviour during revealing data from data portals and evaluate the level of difficulty of navigation to the relevant data. The eye tracking method conveys valuable information for improving the design of portals. Although all these approaches significantly improve the data dissemination capabilities of NSIs, the data dissemination capable to process human's approximate reasoning is still missing. Moreover, in dissemination we should provide aggregated data only if the sensitive data are safeguarded from the disclosure.

3. Linguistic summaries in brief

LSs have been developed to express relational knowledge about the data (Rasmussen and Yager 1997) that is concise and easily understandable for humans.

LSs are written in a general form:

$$Qx(Px) \tag{1}$$

where Q is a linguistic quantifier {few, about half, most, ... }, $X = x$ is a universe of disclosure (e.g. the set of all municipalities) and $P(x)$ is a predicate depicting summariser S {small, medium, high, ... }. An example is: *few municipalities have high pollution.*

The truth value of an elementary linguistic summary (Q *entities in database are* S) is computed by the following equation (Zadrozny and Kacprzyk 2009):

$$T(Qx(Px)) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_p(x_i)\right) \tag{2}$$

where n is the number of entities, $\frac{1}{n}\sum_{i=1}^{n}\mu_p(x_i)$ is the proportion of entities in a data set that satisfy $P(x)$ and μ_Q is the membership function of a quantifier. The truth value of summary is called validity and gets value from the $[0, 1]$ interval. High value of the validity means that the relation expressed by LS is essential. This kind of LS is easily applicable and therefore suitable for data dissemination.

A more complex type of a summary is of the form Q R *entities in database are (have)* S. One example of such a summary is the rule: *most low polluted municipalities have high altitude above sea level*. The procedure for calculating validity of this summary has the following form (Rasmussen and Yager 1997):

$$T(Qx(Px)) = \mu_Q\left(\frac{\sum_{i=1}^{n} t(\mu_p(x_i), \mu_R(x_i))}{\sum_{i=1}^{n}\mu_R(x_i)}\right) \tag{3}$$

where $\frac{\sum_{i=1}^{n} t(\mu_p(x_i),\mu_R(x_i))}{\sum_{i=1}^{n}\mu_R(x_i)}$ is the proportion of the R entities in a database that satisfy S, t is a minimum t-norm function (aggregation of the logical *and* operator from the two-valued logic, Q is the membership function of a quantifier. This kind of summarisation reflects intensity of a relation between particular parts of attributes' domains (Figure 1). The restriction R is more strict if contains several indicators connected by the *and* logical operator (e.g. *high*

pollution and small unemployment rate. In this case $\mu_R(x_i)$ is calculated in the following way:

$$\mu_R(x_i) = f(\mu_{Rj}(x_i)) \tag{4}$$

where R_j is the j-th atomic predicate and f is an aggregation function. In case of the fuzzy logical *and* operator an aggregation function is expressed by t-norm function (Klir and Yuan 1995). A commonly used t-norm function is the minimum t-norm:

$$\mu_R(x_i) = min(f(\mu_{Rj}(x_i))), j = 1, \ldots, n \tag{5}$$

3.1. Construction of membership functions for predicates

Domains of attributes in databases are designed to store all theoretically possible values. In practice, collected values are often situated in a part of a domain. Let D_{min} and D_{max} be the lowest and the highest domain values of attribute A i.e. $Dom(A) = [D_{min}, D_{max}]$ and L and H be the lowest and the highest values in the current content of a database respectively (Hudec and Sudzina 2012). It means that the following holds: $[L, H] \subseteq [D_{min}, D_{max}]$ (either $[D_{min}, L]$ or $[H, D_{max}]$ are empty or even both of them are empty). This fact should be considered in data summarisation by LSs. The aim of LSs is not to reveal intensity of relations from all theoretically possible values but only from the collected data.

The uniform domain covering method (Tudorie 2008) is an appropriate method for construction of membership functions for LSs (Hudec 2013c). Three fuzzy sets form the Figure 1, uniformly constructed on an attribute's domain are depicted in Figure 2 where L and H are the lowest collected value and the highest collected value of the examined indicator respectively. Slopes (α) and cores (β) of fuzzy sets are calculated using the following equations (Tudorie 2008):

$$\alpha = \frac{1}{8}(S - I) \tag{6}$$

$$\beta = \frac{1}{4}(S - I) \tag{7}$$

If higher number of fuzzy sets is required, e.g. five, then linguistic domain can be easily extended to five or even more fuzzy sets.

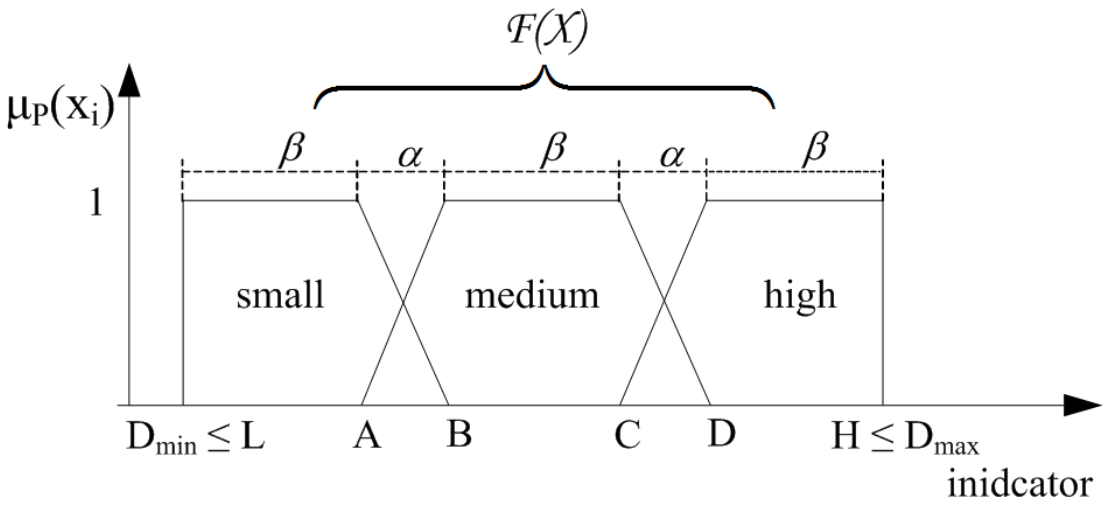

Figure 2: Linguistic and crisp domain of an attribute, $F(X)$ is the family of fuzzy sets (linguistic domain)

3.2. Construction of membership functions for quantifiers

The validity of summaries examined in the paper is computed by the relative quantifier *most*. The quantifier is generalization of the two-valued quantifier *all*. A relative quantifier is constructed by a fuzzy set on the $[0, 1]$ interval (Zadrozny and Kacprzyk 2009). Its membership function should meet the following property:

$$x \leq y \Rightarrow \mu_Q(x) \leq \mu_Q(y);\ \mu_Q = 0;\ \mu_Q(1) = 1 \tag{8}$$

Therefore, the quantifier *most* might be given as (Kacprzyk and Zadrozny 2009):

$$\mu_Q(y) = \begin{cases} 1, & \text{for } y > 0.8 \\ 2y - 0.6, & \text{for } 0.3 \leq y \leq 0.8 \\ 0, & \text{for } y < 0.3 \end{cases} \tag{9}$$

Two extreme situations might occur. If all records meet the predicate with value of 1 then μ_Q obtain a value of 1. It means that the statement is fully true. Opposite, if no record meets the predicate (even partially) then μ_Q obtains a value of 0. In all other cases, validity gets values from the $(0, 1)$ interval.

If numbers from (7) are replaced with parameters in the following way (Hudec 2013a):

$$\mu_Q(y) = \begin{cases} 1, & \text{for } y > n \\ (y - m)/(n - m), & \text{for } m \leq y \leq n \\ 0, & \text{for } y < m \end{cases} \tag{10}$$

then the strictness of the linguistic quantifier can be adjusted. The quantifier is stronger and closer to the crisp quantifier *all*, if $n \to 1$ and $m \to n$. Figure 3 shows three functions for a quantifier: function marked as high density dotted line is the least strict one (parameters m_1 and n_1). Function marked as low density dotted line represents extremely strict quantifier which is the crisp *all* quantifier. In this case, even only one entity (from very large number of entities) does not meet the predicate the truth value of quantifier is 0. The adjustment of m and n filters relations to select only those which significantly describe dependencies.

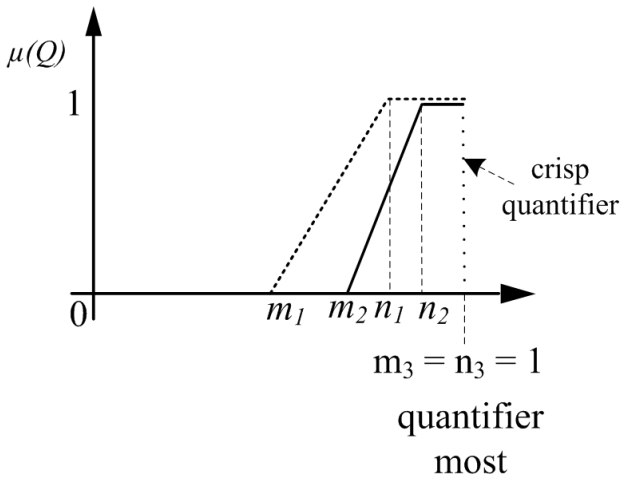

Figure 3: Adjustment of the quantifier *most*

4. Illustrative examples

For the purpose of illustrative examples data from the municipal statistics of the Slovak Republic have been used.

4.1. Estimation of missing values

For the example relations between attributes Population density and Production of waste were evaluated.

In the first step, both attributes were fuzzified into three fuzzy sets {small, medium, high} according to the uniform domain covering method (Figure 2). Experiments were realised in accordance with the definitions of the LSs (Section 3). For the purpose of estimation of missing values, quantifier *most* should be stronger than in (9). For imputation we should be focused on relations who cover majority of records. Consequently, we can say that value of 0.3 (9) hardly meet the membership to the set *most* even with a very low membership degree. Concerning full membership degree we are restricted to value of 1. Hence, the following values were used in experiment: $m = 0.5$ and $n = 1$, eq. (10). Regarding data analysis and dissemination, users can release or intensify strictness of quantifier according to their preferences.

Table 1 shows evaluated summaries. The table reveals very strong relation between small municipalities and small waste production. The first relation could be written in the form of the rule: *if population density is small then production of waste is small.* Therefore, rough estimates of missing values for waste production can be obtained from the waste production in municipalities of similar small population density. Regarding the second LS, we could say that there is no relation. Therefore, second LS cannot be transferred into the if-then rule. Concerning the third relation, the validity is insufficiently strong to be considered as a rule. The term sufficiently strong depends on expert's preferences and a particular task. Generally, the validity less than 0.5 are considered as insufficient. If we are interested for a

Table 1: Population density → Production of waste

Rule	**Validity**	**Fuzzy sets for population density** $[inhab./km^2]$	**Fuzzy sets for production of waste** $[t]$
Most of municipalities having small population density has small production of waste	0.984	$A = 214.50$ $B = 321.25$	$A = 2437.87$ $B = 3656.31$
Most of municipalities having medium population density has medium production of waste	0	$A = 214.50$ $B = 321.25$ $C = 534.75$ $D = 641.50$	$A = 2437.87$ $B = 3656.31$ $C = 6093.18$ $D = 7311.62$
Most of municipalities having high population density has high production of waste	0.362	$C = 534.75$ $D = 641.50$	$C = 6093.18$ $D = 7311.62$

rough estimation then the validity of 0.7 could be considered as a sufficient. But when we are interested in a fine estimation then we can continue with experiments to find more suitable relations having validity greater than for example value of 0.9. Anyway, this discussion is considered form the LSs point of view. From an imputation point of view further research and experiments are crucial.

When rule is not sufficiently strong, possible solution is an additional indicator in the R part of the summary to focus evaluation on a smaller part of the database. A convenient indicator in the example could be Non-agricultural land - built-up area and yard. Therefore, the third relation is modified into the following structure:

most of municipalities having high population density and high ratio of built-up area have high production of waste

The validity of this relation is equal to the value of 1. High validity of rule means that we can create the if-then rule and apply it in the imputation. The stronger restriction R might reveal parts of a database with a very significant relation.

Next step consists of the imputation of missing value. In the small scale test we have removed values of waste production in order to estimate them. Firstly, municipality which belong to the set *small population density* and has missing value of waste production is selected. Secondly, fuzzy query which is able to find similar municipalities (Hudec 2013b) according to population density selects candidates. The query is in the structure: *select municipalities where population density is about density M* (the value of M) is the population density of municipality which has missing value of the waste production). The fuzzy set about M is shown in Figure 4. Finally, the missing value is obtained as an average of population densities and respective membership degrees to fuzzy set about M of selected municipalities. The initial value of waste production was 30 tons. The experiment offered the value of 58.8 tons.

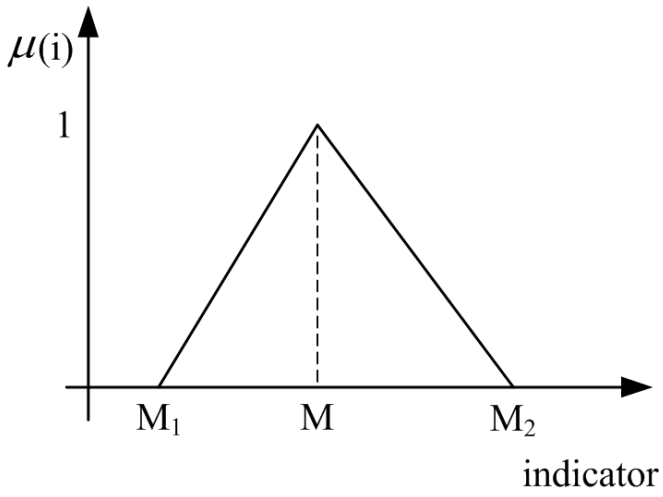

Figure 4: Fuzzy set about M

Discussion

The imputed value is at the first glance quite far from the real one, if we take into account absolute values of 58.8 (estimated) and 30 (real). However, concerning the whole domain of waste production; that is, $[1, 48021.3]$ then the result is not so bad. Anyway, results can be improved by adding additional attributes in the restriction part R of the summary or dividing fuzzy sets of restriction R and summarizer S (Figure 1) into more fuzzy sets e.g. *very small, small, medium, high and very high.*

Presumably, the important question is choosing appropriate indicators for LSs, especially when a database contains large number of indicators. In this way we avoid unnecessary computational burdens between less independent indicators. For example indicators: The year of the first written notice and Length of roads hardly have something in common. Statisticians should determine indicators of interest. The second task is reducing number of summaries between attributes (Figure 1) to relevant ones. Concerning this issue, we hardly expect that

small municipalities produce large amount of waste; that is, the rule: *most of municipalities having small population density has high production of waste* has very low validity and should not be evaluated. In this way, user can reduce number of relations.

On the other hand, evaluation of less expected relations could also reveal valuable information. For example, if the validity of such relation is significant then collected data might contain measurement errors or there is less expected development in some categories of municipalities. This is a topic which is suitable for the data analysis, discussed in the Section 4.2.

4.2. Data analysis

Retrieved relational knowledge from the data expressed in a concise and easily understandable way could support decision making, analysing trends, and behaviour in different areas of statistics. For the purpose of illustrative example we have evaluated relations between attributes Consumption of water for households per inhabitant and Number of summer days during year (temperature greater or equal than 25 degrees Celsius). Table 2 shows revealed intensities of relations. The aggregated information could be interpreted and explained in

Table 2: Number of warm days → Consumption of water (Hudec *et al.* in press)

Rule	**Validity**
Most of municipalities having high number of warm days has high consumption of water	0
Most of municipalities having medium number of warm days has medium consumption of water	0.695
Most of municipalities having small number of warm days has small consumption of water	0.905

the following way. The validity of the third rule is very strong. We can conclude that it is no special need for higher consumption of water in municipalities having small number of warm days. Concerning the second row, the validity is lower but still significant (> 0.5). In order to get more precise answer additional relations should be evaluated. Concerning the first relation and its zero validity, we could say that e.g. households' saving causes that the water consumption is not as high as expected. Admittedly, this is only one of possible interpretations used in the illustrative example.

Discussion

The relational knowledge expressed in a concise and easily understandable way could support executive decision making and analysing trends and behaviour in different groups of municipalities.

Let's examine the following rule *most of highly situated municipalities have small gas consumption*. High validity of the rule reveals that these municipalities use alternative source for heating in winter, for example. If we consider the same rule but in a traditional crisp way, the rule has to be expressed in a form*: all municipalities with altitude* $> A\,[m]$ *have gas consumption* $< B\,\left[m^3\right]$. At the first glance the meaning of the rule is not clear. Secondly, the rule is either satisfied or not. When the rule is not satisfied we are not sure whether municipalities are about to meet the summary or they are far to meet it. We can avoid this issue including intensities of examined summaries expressed by fuzzy sets.

Concerning choosing appropriate indicators and selecting relations of interest, the same as in discussion part of the Section 4.1 holds.

4.3. Data dissemination

NSIs offer data by portals or other services (free or paid). However, users cannot know in

advance which part of a data set is the most suitable for them.

The municipal statistics is organized in a hierarchical way: municipalities belong to respective districts, districts to respective regions, etc. In case of Slovak municipal statistics (and probably in other countries as well) the majority of municipal data are not free of charge. The fee depends on the amount of ordered data. LSs provide an "abstract" of each unit on higher hierarchy level by summarising data on lower hierarchical level. An example is the following: *rank regions where most of municipalities have small unemployment and high migration.* In the first step, validity of a summary is calculated for each region. In the second step regions are ranked downwards starting with region having the highest value of the summary validity. Therefore, users can order data only for regions which significantly meet their requirements.

Let's for the illustrative purpose, user wants to deeply analyse regions where most of municipalities have high ratio of agricultural land (arable land). This question is expressed by the linguistic summary *most of municipalities have high ratio of agricultural land.* The result for all eight regions of the Slovak Republic is presented in Table 3 using eq. (9). Table 3 shows regions ranked from the most suitable ones to regions out of interest. Depending of user's capacities for data analysis, interest etc. user can decide to order data for municipalities belonging to regions which fully meet the LS only or to order data for municipalities from regions which significantly meet the LS as well. The query based on LSs keeps data that are

Table 3: Validity of the linguistic summary for each region

Region	**Validity of the summary**
Nitra	0.9469
Trnava	0.8255
Bratislava	0.2015
Košice	0.0603
Prešov	0
Bánska Bystrica	0
Žilina	0
Trenčín	0

not free of charge or sensitive (in our case the ratio of agricultural land for municipalities) hidden. The ratio of agricultural land is used in the query condition of the summary without any modifications and restrictions. This indicator is not presented to user in the Table 3. A user obtains only the summarised information for each region. Therefore, LSs could meet disclosure control requirements. In this promising approach further research is required e.g. the critical size of data and the structure of dependencies on lower hierarchical level to avoid any risk of disclosure. Furthermore, even if all data are open users presumably prefer to obtain summarized overview in order to decide to which region focus their interest.

Finally, obtained summaries can be visualised by other means. In area of municipal statistics a suitable mean are thematic maps. Districts having validity of rule equal to value of 1 can be marked with one colour, districts which do not meet the summary (validity is equal to value of 0) can be marked with the second colour and district having the validity of the summary in the $(0, 1)$ interval can be marked with the third colour having a colour gradient from a faint hue to a deep hue following the value of validity.

(Hudec 2011) argued that the essence of fuzzy queries is reducing or eliminating the communication barrier between the human and the computer during querying process. The goal of many websites is to target broad audience. Queries by linguistic terms offer the natural way for querying databases and, therefore, websites could become more user friendly oriented in retrieving relevant data and relational knowledge. The web application realised in this way could be an effective way to motivate users to use data portals (Hudec and Torres van Grinsven 2014) and to provide their data in mandatory or voluntary surveys. We could expect

that if users can find relevant data and information easily, they will be more willing to provide their own data.

5. Conclusion

The paper discusses issues in a municipal statistics related to treatment missing values, data analysis and data dissemination. For all these tasks LSs have potential which was proven on illustrative examples. The main objective of this paper was introducing LSs to statistics and illustrating their potential by small scale case tests. Definitely, this approach needs further research and experiments. The second intent was sharing this idea to researchers and developers in official statisticians to initiate further research activities. It especially holds for estimation of missing values because it is a very sensitive task. In the further research comparison between LSs, traditional tools (e.g. hot deck imputation mentioned above) and approaches based on soft computing: neural networks (Juriová 2012) and genetic programming (Kľúčik 2012) could be very valuable for NSIs. For data dissemination, LSs could be applied as a standalone tool allowing quickly mining relational knowledge from the data. LSs could be also a complementing tool for existing approaches in data dissemination.

Furthermore, variability among the indicators could be expected as random. However, this is not always true and some relation might exist (lower degree of similarity between entities and dependency between measured phenomena). LSs could be used as easy to apply tool for an initial mining of these relations to limelight the path for a deeper analysis. LSs could be also used to check hypotheses about dependency and relations between attributes. In order to solve problems which worry practitioners, LSs should be used with other approaches in a complementing rather than competitive way.

The main advantage of development of software tool based on LSs is that the core of the software remains the same only modules for imputation, data analysis and dissemination should be adapted to particular needs. Standalone application in NSI should be more complex to cover all needs whereas web application should be less complex to offer comfortable work for non-experts.

Finally, the development of a full functional software tool is a demanding task. There are some tools like SummarySQL (Rasmussen and Yager 1997) or FQUERY (Kacprzyk and Zadrozny 2009). For official statistics we should create a tailored tool to meet specific needs. 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 efforts inside NSIs. The term limited means that the tool is open only for the NSIs community due to the specific requirements. One group of NSIs could be focused on development of several tools and other institutes will use these tools and will be at the same time able to use their resources to participate in development of other software tools.

References

Andridge RR, Little RJA (2010). "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review*, **78**(1), 40–64.

Bavdaž M, Biffignandi S, Bolko I, Giesen D, Gravem D, Haraldsen G, Löfgren T, Lorenc B, Persson A, Mohoric Peternelj P, *et al.* (2011). "Final Report Integrating Findings on Business Perspectives Related to NSIs' Statistics. Deliverable 3.2., Blue-Ets Project."

Coutinho W, de Waal T (2012). "Hot Deck Imputation of Numerical Data Under Edit Restrictions." *Technical report*, Discussion paper, Statistics Netherlands.

De Leeuw ED, Hox JJ, Huisman M (2003). "Prevention and Treatment of Item Nonresponse." *Journal of Official Statistics*, **19**(2), 153–176.

Hudec M (2011). "What Could Fuzzy Logic Bring to Statistical Information Systems?" *Statistika*, **48**(1), 58–70.

Hudec M (2013a). "Applicability of Linguistic summaries." In *Proceedings of the 11th Balkan conference on operational research*. Belgrad.

Hudec M (2013b). "Fuzzy Database Queries in Official Statistics: Perspective of Using Linguistic Terms in Query Conditions." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **29**(4), 315–323.

Hudec M (2013c). "Issues in Construction of Linguistic Summaries." In *R. Mesiar and T. Bacigál, editors, Proceedings of Uncertainty Modelling 2013*. STU, Bratislava.

Hudec M, Balbi S, Juriová J, Kl'účik M, Marino M, Scepi G, Spano M, Stawinoga A, Tortora CTN, 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*.

Hudec M, Sudzina F (2012). "Construction of Fuzzy Sets and Applying Aggregation Operators for Fuzzy Queries." In *14th International Conference on Enterprise Information Systems(ICEIS 2012)*. Wroclaw.

Hudec M, Torres van Grinsven V (2014). "Business Participants Motivation in Official Surveys by Fuzzy Logic." *European Scientific Journal*, **SPECIAL(3)**, 42–52.

Hudec M, Vučetić M, Vujošević M (in press). "Comparison of Linguistic Summaries and Fuzzy Functional Dependencies Related to Data Mining." In *S. Alam, Y. S. Koh and G. Dobbie, editors, Biologically-Inspired Techniques for Knowledge Discovery and Data Mining*. IGI Global.

Jern M, Haldorson M, Thygesen L (2011). "Storytelling-How to Visualize Statistics." In *Proceedings of the New Techniques and Technologies for Statistics (NTTS) conference*. Brussels.

Juriová J (2012). "Neural Network Approach Applied for Classification in Business and Trade Statistics." In *Proceedings of the 46th scientific meeting of the Italian statistical society (SIS 2012)*. Rome.

Kacprzyk J, Zadrozny S (2009). "Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining." *International Journal of Software Science and Computational Intelligence*, **1**, 1–11.

Klir GJ, Yuan B (1995). *Fuzzy Sets and Fuzzy Logic*, volume 4. Prentice Hall New Jersey.

Kl'účik M (2012). "Estimates of Foreign Trade Using Genetic Programming." In *Proceedings of the 46th Scientific Meeting of the Italian Statistical Society (SIS 2012), Rome*.

Lehtinen H, Gløersen R (2009). "Cooperation in Development of Open Source Software." In *Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information (MSIS2009)*. Oslo.

Rasmussen D, Yager RR (1997). "Summary SQL–a Fuzzy Tool for Data Mining." *Intelligent Data Analysis*, **1**(1), 49–58.

Smith A (2011). "Web 2.0 and Official Statistics: The UK Perspective." In *Proceedings of the New Techniques and Technologies for Statistics (NTTS) conference. Brussels*.

Statistics Norway (2007). "Statistics Norway's Dissemination Policy." In *Documents 2007/10*. Statistics Norway, Oslo.

Tudorie C (2008). "Qualifying Objects in Classical Relational Database Querying." In *J. Galindo, editor, Handbook of research on fuzzy information processing in databases*. Information Science Reference Hershey.

Wulff A (2007). "Experiences Using the Eye-Tracking Method to Test Website Usability." In *Proceedings of the UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS 2007)*. Geneva.

Zadeh LA (1999). "From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions." *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, **46**(1), 105–119.

Zadrozny S, Kacprzyk J (2009). "Issues in the practical use of the OWA operators in fuzzy querying." *Journal of intelligent information systems*, **33**(3), 307–325.

Affiliation:

Miroslav Hudec
Department of Applied Informatics, Faculty of Economic Informatics
University of Economics in Bratislava
852 35 Bratislava, Slovakia
E-mail: miroslav.hudec@euba.sk
URL: http://www.euba.sk

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43 *Submitted:* 2013-12-03
April 2014 *Accepted:* 2014-03-20