DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 981–1010, http://dx.doi.org/10.2478/JOS-2018-0048

# Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language

*Miroslav Hudec*[1], *Erika Bednárová*[1], *and Andreas Holzinger*[2]

Data from National Statistical Institutes is generally considered an important source of credible evidence for a variety of users. Summarization and dissemination via traditional methods is a convenient approach for providing this evidence. However, this is usually comprehensible only for users with a considerable level of statistical literacy. A promising alternative lies in augmenting the summarization linguistically. Less statistically literate users (e.g., domain experts and the general public), as well as disabled people can benefit from such a summarization. This article studies the potential of summaries expressed in short quantified sentences. Summaries including, for example, "*most visits from remote countries are of a short duration*" can be immediately understood by diverse users. Linguistic summaries are not intended to replace existing dissemination approaches, but can augment them by providing alternatives for the benefit of diverse users of official statistics. Linguistic summarization can be achieved via mathematical formalization of linguistic terms and relative quantifiers by fuzzy sets. To avoid summaries based on outliers or data with low coverage, a quality criterion is applied. The concept based on linguistic summaries is demonstrated on test interfaces, interpreting summaries from real municipal statistical data. The article identifies a number of further research opportunities, and demonstrates ways to explore those.

*Key words:* Linguistic summaries; linguistic quantifiers; fuzzy sets; database queries; user interface.

## 1. Introduction

Businesses, public administrations, researchers, journalists, and the general public are increasingly interested in data and information that describe various aspects of our society. National Statistical Institutes (NSIs) are sources that are generally regarded as credible, due to their profound and reliable methodologies for data collection, production and dissemination, explained through the Generic Statistical Information Model (e.g., GSIM 2013; Scanu and Casagrande 2016). Various approaches to data dissemination have already been developed; applications such as Contestina (Zottoli et al. 2017) provide interfaces for creating questions, interpreting answers in tables, graphs and on maps, and storytelling based on specific parameters chosen by the users. However, all these approaches, although powerful, often require up-to-date information and communication technologies (web browser versions and fast internet connection running on up-to-date hardware) which are still not available to everybody. Consequently, the dissemination

[1] Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. Emails: miroslav.hudec@euba.sk and bednarovaa.erika@gmail.com
[2] Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Auenbruggerplatz 2, 8036 Graz, Austria. Email: andreas.holzinger@medunigraz.at

should also be accessible to those who rely on "low-tech" information and communication technologies, on a variety of platforms.

While larger businesses prefer raw data and analyze them by their own methods, smaller businesses rather look for information and prefer simple presentations and short descriptions (Bavdaž 2011). One reason might be that smaller businesses usually cannot afford specialists in data mining and statistics, or expensive consultancy, and their statistical and computational skills may not be sufficient to effectively interpret the data produced by NSIs. The same might hold true for some journalists searching for statistical information to support their articles (while data journalists would prefer access to the data). Disabled people frequently have to overcome obstacles when searching for data and information on websites: blind people need content that can be expressed by sound or voice instead of graphs and tables; people who are dyslexic or have cognitive impairments may benefit from the use of simpler language (Disability Rights Commission 2004; Heimgärtner et al. 2008).

Inspiration for an alternative approach emerged from the following five observations: (i) graphical interpretation is a valuable way of summarization; however, it is not always effective (Disability Rights Commission 2004; Lesot et al. 2016); (ii) users (e.g., small businesses) are often interested in summarized information rather than data (Bavdaž 2011); (iii) summaries should not be as terse as means (Yager et al. 1990), and should hold for any data type and distribution; (iv) a natural way for humans to communicate, compute and conclude is natural language (Zadeh 2001); and, (v) existing approaches in data dissemination are typically based on precise (crisp) conditions or questions, for example, "*find towns that accommodated more than 1,000 visitors*". The alternative is summaries of short quantified sentences of natural language, or Linguistic Summaries (LSs). For example, we can express: "*the mean value is 235.4 with a standard deviation of 123.3*", or linguistically: "*few observations are near the mean value*". The linguistic case clearly illustrates that the mean value is not a sufficiently representative characteristic in this example. The other option is interpreting the summary between attributes, for example, "*most visits from remote countries are of a short duration*". Such a summary, although neither based on traditional mathematical methods nor on visualisation, contains very valuable information for accommodation providers, marketers, journalists and local authorities. In addition, linguistic summaries can be interpreted by text-to-speech synthesis systems. They are especially useful whenever the users' visual attention is focused on something else (Arguelles and Triviño 2013), or for the aforementioned disabled and/or elderly people (Holzinger 2002). Kacprzyk and Zadrożny (2005, 282) recognized the benefits of linguistic summaries by emphasizing that "*Data summarization is one of [the] basic capabilities that is now needed by any 'intelligent system' that is meant to operate in real life*". People ask, evaluate and conclude by linguistic terms, which are vague, but on the other hand very effective. Here, "vague" means nonsharp boundaries of concepts (linguistic terms) expressed by fuzzy sets, whereas "effective" means that we distinguish elements by intensity of belonging to a set without adding further properties. This observation led Zadeh (2001) to formalize the concept known as *computing with words*.

In this article, we provide a more theoretical view on dissemination by linguistic summaries for the users of official statistics. The "test interfaces" have been developed

merely to illustrate our idea, to demonstrate applicability and to show procedures for calculating and interpreting linguistic summaries from real-world data.

The remainder of this article is organized as follows: Section 2 introduces linguistically quantified sentences and a theoretical basis consisting of related works and our observations, all of which is necessary for the subsequent sections. Section 3 is dedicated to dissemination through LSs, supported by illustrations and examples on data from the Municipal Statistics Database of the Slovak Republic. Section 4 is focused on discussing our findings, problems, challenges, potential obstacles and suggestions for future research topics, while Section 5 concludes the article. Moreover, Section 6 (Appendix A) addresses theoretical aspects of fuzzy logic and quality measures, whereas Section 7 (Appendix B) provides a list of symbols used.

## 2. Linguistic Summaries, Formalization and Quality

This section studies relevant theoretical aspects of flexible linguistic data summarization, which are used throughout the article.

### *2.1. Basic Types of Linguistic Summaries*

Linguistic summaries summarize information from data into a concise and easily understandable interpretation. Lesot et al. (2016) divided prototype forms (protoforms) of linguistic summaries into the following three main groups:

1. classic protoforms,
2. protoforms of time series, and
3. temporal protoforms.

The classic protoforms summarize attribute(s) on the whole data set, or relations among attributes (Kacprzyk and Zadrożny 2005; Rasmussen and Yager 1997; Yager 1982). These summaries are of the structure *Q entities are S* and *Q R entities are S*, respectively, where $Q$ is a flexible linguistic quantifier, $S$ is a summarizer and $R$ is a restriction. The former structure is illustrated by the sentence "*most houses have high gas consumption*". An illustrative example of the latter structure is: "*most old houses have high gas consumption*".

The protoforms of time series linguistically express behavior of attributes over time (Almeida et al. 2013; Kacprzyk et al. 2006). These summaries are divided into summaries describing a time series of the structure *Q Bs are A*, and summaries considering several time series together of the structure *Q Bs are A* $Q_T$ *time*, where $Q_T$ is a quantifier applied to the time attribute, $Q$ is a relative fuzzy quantifier, $A$ and $B$ are the examined concepts. Illustrative sentences are: "*most trends of topic B are of low variability*" and "*about half small businesses have small response rate most of the time*", respectively.

However, the temporal protoforms do *not* use linguistic quantifiers, but a mode of behavior for creating periodic summaries. This kind of summaries is of the structure *P, the data are A*, where $P$ is a temporal adjustment and $A$ is a fuzzy modality. An illustrative example would be: "*regularly in autumn, the participation is high*". Here, the term "*regularly*" describes the extent to which a summary holds in considering a particular temporal adjustment (Moyse et al. 2013).

While the other protoforms are also promising for data dissemination, and could be examined and applied in a similar manner, this work is focused on the classic ones in order to examine their applicability.

### 2.2. *Linguistic Variables and Quantifiers*

Linguistic summaries rely on the theories of fuzzy sets and fuzzy logic, where belonging to a set is a matter of degree. A fuzzy set $F$ over the universe of discourse $X$ is defined by the membership function $\mu_F$ that matches each element of $X$ with its degree of membership to the set $F$ (Zadeh 1965)

$$\mu_F(x)\colon X \rightarrow [0,\ 1] \tag{1}$$

where $\mu_F(x) = 0$ means that an element $x$ does not belong at all to $F$, while $\mu_F(x) = 1$ means that $x$ is a full member of $F$. A value of $\mu_F(x)$ between 0 and 1 indicates the intensity by which the element $x$ belongs to $F$. The concept of fuzzy sets is further discussed in Section 6, Appendix A.

The first major concept required for our work is Linguistic Variable (LV). An LV is a variable, whose values (often called labels) are words of natural language determined by a quintuple $(L, T(L), X, G, H)$ (Zadeh 1975), where

- $L$ is the name of the variable,
- $T(L)$ is a set of all linguistic labels related to variable $L$,
- $X$ is the universe of discourse,
- $G$ is the syntactic rule for generating $T(L)$ values, and
- $H$ is the semantic rule that relates each linguistic label of $T(L)$ to its meaning $H(L)$.

An example of LV is any attribute whose domain can be divided into overlapping granules, for example *pollution* and *number of visits*. The LV "*pollution*" consisting of labels *low*, *medium* and *high* is plotted in Figure 1. For a finer granulation we can construct more labels, for example *very low*, *low*, *medium*, *high* and *very high*. The syntactic rule explains the required number of linguistic labels and their names, whereas the semantic rule assigns the context dependent meaning to each label by fuzzy sets. For instance, the

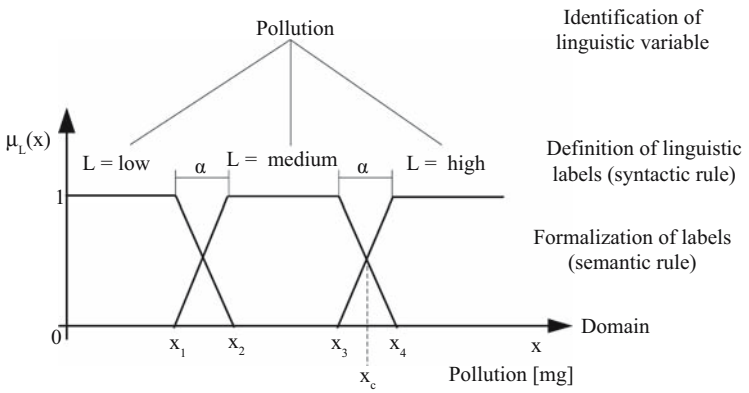


*Fig. 1. Linguistic variable "pollution" and its labels.*

fuzzy set *high* (using the $x$-values defined in Figure 1) is expressed as

$$\mu_{high}(x) = \begin{cases} 1 & x \geq x_4 \\ (x - x_3)/(x_4 - x_3) & x_3 < x < x_4 \\ 0 & x \leq x_3 \end{cases} \tag{2}$$

Value $x_c$ is the maximal uncertainty point. In a smooth transition from sets *medium* to *high*, $x_c$ belongs to both with 0.5 degree, that is, we cannot be sure whether this value is more *medium* than *high*. The intervals having the width $\alpha$ are uncertain areas. When $\alpha = 0$, these sets are *crisp* (an element is either a member of the set or not).

Generally, fuzzy sets can be formalized by non-linear functions. In this article, we adopted the linear ones due to their simplicity for the end users. In the case of non-linear functions, the users have to specify the shapes of fuzzy sets, which is not a simple task for the less mathematically literate users, as is, for example, the case in the medical domain (Holzinger et al. 2017).

The next element in LSs is the fuzzy quantifier. Fuzzy quantifiers are discussed in detail by, for example, Glöckner (2006). The formalization of fuzzy relative quantifiers can be realized by three approaches: sigma counts (Zadeh 1983), Ordered Weighted Averaging (OWA) operator (Yager 1988) and Competitive Type Aggregation (Yager 1984). For reasons of simplicity, the sigma count approach is chosen for this article. In this way, summarizer and restriction (explained later), as well as quantifier are modelled by the same approach, which is, in addition, more intuitive for diverse users. Within that approach, the quantifier *most of* is formalized by an increasing (usually linear) function where $\mu_Q(0) = 0$ and $\mu_Q(1) = 1$ as (Kacprzyk and Yager 2001; Kacprzyk and Zadrożny 2005)

$$\mu_Q(y) = \begin{cases} 1 & y \geq 0.8 \\ 2y - 0.6 & 0.3 < y < 0.8 \\ 0 & y \leq 0.3 \end{cases} \tag{3}$$

where $y$ is the proportion of units fully or partially satisfying a predicate in a summary expressed by fuzzy sets. In our application, we modified the parameters in (3) in such a way that the membership degree becomes higher than zero only for the proportions higher than 0.5 to meet the usual meaning of *most of* and *majority*, that is

$$\mu_Q(y) = \begin{cases} 1 & y \geq 0.8 \\ (y - 0.5)/0.3 & 0.5 < y < 0.8 \\ 0 & y \leq 0.5 \end{cases} \tag{4}$$

Analogously, the quantifier *about half* is a symmetric triangular or trapezoidal fuzzy set centered around the value of 0.5 ($\mu_Q(0.5) = 1$). The quantifier *few* is expressed by a decreasing function ($\mu_Q(0) = 1$, $\mu_Q(1) = 0$). Thus, a possible family of relative quantifiers plotted in Figure 2 is also a LV.

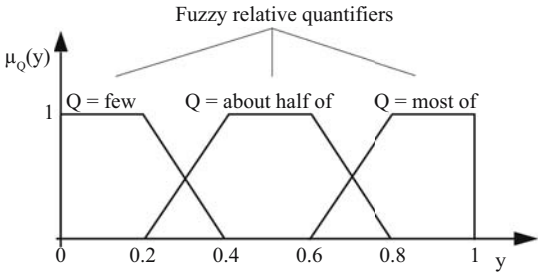


*Fig. 2. A possible family of relative quantifiers for a proportion y.*

### 2.3. *Formalization of Classic Protoforms and Their Quality Aspects*

A basic structure of LS for summarizing attributes is *Q entities in database are (have) S* (Yager 1982). Quantifier $Q$ and summarizer $S$ are usually both formalized by linguistic terms (fuzzy sets), for example "*most agricultural companies have a high turnover*". The proportion of records in a data set $\mathbf{X}$ that fully and partially satisfies the predicate $S$ are defined as

$$y_{LSb}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu_S(x_i) \tag{5}$$

where $n$ is the number of units in a data set and the membership function $\mu$ formalizes summarizer $S$ for the units. The validity (truth value) of the summary is calculated as

$$v_{LSb}(\mathbf{X}) = \mu_Q(y_{LSb}(\mathbf{X})) \tag{6}$$

where the function $\mu$ formalizes quantifier $Q$ for the summary. Both $y_{LSb}(\mathbf{X})$ and $v_{LSb}(\mathbf{X})$ assume values in the interval [0, 1].

Linguistic Summaries with restrictions take the form *Q R entities in database are (have) S*, where restriction $R$, also expressed in linguistic terms, focuses on a part of data set relevant for the summarization task (Rasmussen and Yager 1997), for example, "*most highly polluted municipalities have a high number of respiratory diseases*". The proportion of records in a data set $\mathbf{X}$ that fully or partially satisfies the restriction $R$ and also fully or partially satisfies the summarizer $S$, is defined as

$$y_{LSr}(\mathbf{X}) = \frac{\sum_{i=1}^{n}(\mu_S(x_i) \wedge \mu_R(x_i))}{\sum_{i=1}^{n}\mu_R(x_i)} \tag{7}$$

where $n$ is the number of units in $\mathbf{X}$ and the membership function $\mu$ formalizes, in term, both $S$ and $R$. The "and operator" in the numerator is expressed by a triangular norm (Section 6, Appendix A). The convention $0/0 = 0$ is used in order to avoid undefined proportions; this situation occurs when not a single record meets $R$ (and as a logical consequence, not a single record simultaneously meets $R$ and $S$). Analogously to (6), the validity of the summary is calculated as

$$v_{LSr}(\mathbf{X}) = \mu_Q(y_{LSr}(\mathbf{X})) \tag{8}$$

The concept of LSs was introduced by Yager (1982). Since then, the theory of LSs has been improved and applied in a variety of fields. Boran et al. (2016) provide an overview of recent developments. The linguistic terms used in $S$ and $R$ can be formalized by fuzzy sets having functions of different shape (as illustrated in Figure A.1 of Section 6, Appendix A), ensuring the smooth transition between belonging and nonbelonging to the set.

The basic quality criterion (validity or truth value as defined in (6) and (8)) does not cover all aspects of quality (Kacprzyk and Yager 2001). Due to the complexity of quality measures, problems with their aggregation and particularities of the considered data dissemination (see Section 6, Appendix A), we adopted a simplified quality measure that integrates two of the most important measures: validity and coverage introduced by Hudec (2017) for LSs with restriction

$$Q_c = \begin{cases} t(v, C) & C \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $C$ is data coverage (defined as a function (A.4) of the proportion of the whole data set affected by the summarizer and restriction) and $t$ is a nonidempotent t-norm, for example, a product t-norm (A.2). A discussion related to quality measures and the rationale for choosing the measure (9) is held in Section 6, Appendix A. This simplified measure, which is calculated from the data, contributes to the decreased complexity of interfaces, because users do not need to intervene. In (5) and (6), the whole data set is covered due to $n$ (the cardinality of the data set) being in the denominator. It means that the data coverage is implicitly calculated in $y_{LSb}(\mathbf{X})$.

### *2.4. A Case Study for Interpreting Data by Crisp and Fuzzy Logic*

Hypothetical values of pollution measured over all 30 days of a month in two districts are shown in Table 1. The authorities wish to disseminate information regarding the pollution dispersion. Let us have crisp set "*high pollution*" (*HP*) defined as $HP = \{x : x > 20\}$. When we apply this set in a query: *select districts where high pollution was recorded*, then district *D1* is selected, whereas *D2* is not. However, a quick glance at Table 1, applying common sense reasoning, leads to the conclusion that *D2* is more polluted than *D1*. Furthermore, it might happen that the recorded values for *D1* in days 10 and 14 are incorrect due to measurement errors. In that case, the disseminated information does not correspond with reality. Dissemination by proportion says that for *D1* pollution was high in 7% of the days, whereas high pollution was not recorded for *D2*.

Let us examine this problem from the fuzzy logic perspective. The concept "*high pollution*" can be expressed by a fuzzy set (2) as follows

$$\mu_{FHP}(x) = \begin{cases} 1 & x \geq 20 \\ (x - 15)/5 & 15 < y < 20 \\ 0 & x \leq 15 \end{cases} \tag{10}$$

where pollution above 20 units is still considered high without any doubt, but slightly lower values belong to the concept "*high pollution*" with membership degrees smaller than 1 (Table 1).

*Table 1. Measured pollution for two districts (illustrative data).*

| District D1 | | | | | | District D2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Day | Measured pollution [mg] | Matching degree to high pollution $\mu_{FHP}(x)$ | Day | Measured pollution [mg] | Matching degree to high pollution $\mu_{FHP}(x)$ | Day | Measured pollution [mg] | Matching degree to high pollution $\mu_{FHP}(x)$ | Day | Measured pollution [mg] | Matching degree to high pollution $\mu_{FHP}(x)$ |
| 1 | 2.950 | 0 | 16 | 8.375 | 0 | 1 | 16.577 | 0.315 | 16 | 18.925 | 0.785 |
| 2 | 6.740 | 0 | 17 | 8.079 | 0 | 2 | 16.923 | 0.385 | 17 | 17.223 | 0.445 |
| 3 | 1.669 | 0 | 18 | 9.183 | 0 | 3 | 15.102 | 0.020 | 18 | 14.465 | 0 |
| 4 | 5.887 | 0 | 19 | 7.104 | 0 | 4 | 19.383 | 0.877 | 19 | 18.465 | 0.693 |
| 5 | 2.621 | 0 | 20 | 16.005 | 0.2010 | 5 | 18.606 | 0.721 | 20 | 11.530 | 0 |
| 6 | 9.106 | 0 | 21 | 5.630 | 0 | 6 | 12.981 | 0 | 21 | 17.281 | 0.456 |
| 7 | 8.239 | 0 | 22 | 10.286 | 0 | 7 | 16.589 | 0.318 | 22 | 16.084 | 0.217 |
| 8 | 7.036 | 0 | 23 | 4.569 | 0 | 8 | 19.038 | 0.808 | 23 | 19.969 | 0.994 |
| 9 | 5.438 | 0 | 24 | 8.877 | 0 | 9 | 14.043 | 0 | 24 | 15.023 | 0.005 |
| 10 | 21.232 | 1 | 25 | 8.150 | 0 | 10 | 19.346 | 0.869 | 25 | 16.003 | 0.201 |
| 11 | 4.285 | 0 | 26 | 16.256 | 0.2512 | 11 | 18.443 | 0.689 | 26 | 17.226 | 0.445 |
| 12 | 7.494 | 0 | 27 | 4.456 | 0 | 12 | 19.889 | 0.978 | 27 | 18.099 | 0.620 |
| 13 | 2.831 | 0 | 28 | 3.187 | 0 | 13 | 19.886 | 0.977 | 28 | 18.402 | 0.680 |
| 14 | 20.006 | 1 | 29 | 2.041 | 0 | 14 | 18.359 | 0.672 | 29 | 12.049 | 0 |
| 15 | 1.810 | 0 | 30 | 2.950 | 0 | 15 | 19.039 | 0.808 | 30 | 18.077 | 0.615 |

Next, we calculate $y_{LSb}$ and adopt suitable quantifiers. The proportion for *D1* (i.e., the sum of matching degrees divided by 30 – i.e., the number of days) is obtained as 0.07 (as for the *crisp logic* example above), whereas the proportion is 0.51 for *D2*. Now, we are able to disseminate by proportions: "*for D1, in 7% of the days, pollution was high*, *for D2, in 51% of the days, pollution was high*". We may continue to elaborate a more sophisticated linguistic interpretation by the following two sentences: "*in D1, for a few days, pollution is high*; *in D2, for about half of the days, pollution is high*" (using the parameters shown in Figure 2, the validities of both sentences are obtained as $\mu_{few}(y_{D1}) = 1$ and $\mu_{about\ half}(y_{D2}) = 1$, respectively). The second sentence can be further summarized into "*in D2, for slightly above half of the days, pollution is high*", when we formalize the quantifier *slightly above half*.

## 3. Linguistically Summarizing Statistical Data

In this section the innovative potential of LSs for the official statistics data dissemination is demonstrated on the illustrative data, as well as on the real data from the Municipal Statistics Database managed by the Statistical Office of the Slovak Republic. This database consists of more than 800 attributes for 2,927 municipalities. The test interfaces were developed for the sole purpose of illustrating applicability and procedures for calculating linguistic summaries and have therefore not yet been tested on users. The interfaces were developed in Visual Studio 2013 and MS Access 2013, while the data was stored in an MS Access relational database.

### *3.1. An Option of Representing Data by a Set of High-Validity Sentences*

The extent to which observations are spread around their mean value is expressed by dispersion functions. However, these functions can be overlooked, especially by people with a lower level of statistical literacy, who may conclude that all essential information about a variable is encapsulated in its mean value. In the following example, we will illustrate how linguistic summaries could help remedy this.

*Example 1*

A fictive data set contains seven respondents with their respective ages {26, 28, 32, 40, 54, 56, 57} (Hudec 2016). Summarization by statistical methods reveals that the average age (arithmetic mean) is 41.9, the median age is 40, and the standard deviation is 13.7. The arithmetic mean and median lead us to the conclusion that the typical age of a respondent is around 40, but standard deviation shows that this is not the case.

The interpretation by linguistic summaries says the same, but differently. Three labels: *young*, *middle-aged* and *old* of the LV "*age*" required for summarizer *S* are formalized as follows

$$\mu_{young}(x) = \begin{cases} 1 & x \leq 30 \\ (35 - x)/5 & 30 < x < 35 \\ 0 & x \geq 35 \end{cases}$$

$$\mu_{middle\_aged}(x) = \begin{cases} (x-30)/5 & 30 < x < 35 \\ 1 & 35 \leq x \leq 50 \\ (55-x)/5 & 50 < x < 55 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_{old}(x) = \begin{cases} 0 & x \leq 50 \\ (x-50)/5 & 50 < x < 55 \\ 1 & x \geq 55 \end{cases}$$

The LV expressing the family of quantifiers: *few*, *about half of* and *most of* is depicted in Figure 2. With three labels and three quantifiers, $3{\cdot}3 = 9$ possible LSs exist. The high validity sentences and their respective validities are shown in Table 2.

From Table 2 it is clear that about half of respondents are young, and about half are old, whereas few are middle-aged (although the mean value is around 40). It is worth noting that a histogram would provide the same message visually; this is the corresponding verbal summary.

Linguistic summaries are able to generate all relevant sentences regarding the attributes under consideration and merge them to create a simple story. In our case, the story is*:* "*half of respondents are old, about half are young and few are middle-aged*". Moreover, such summaries might be supportive for automated or computational journalism, that is, technologically oriented journalism focused on the application of computational intelligence to the practices of information gathering and information presentation (Coddington 2015). Graefe (2016, 15) states that "Current solutions range from simple code that extracts numbers from a database, which are then used to fill in the blanks in prewritten template stories, to more sophisticated approaches that analyse data to gain additional insight and create more compelling narratives." The LSs concept presented in this article is situated between these two extremes.

This discussion naturally leads to the question of automated creation of relevant LSs, which is an important future topic in machine learning. This task is formalized by Liu (2011) as

$$\begin{aligned} &\text{find } Q, S, R \\ &\text{subject to} \\ &Q \in \bar{Q}, S \in \bar{S}, R \in \bar{R}, v(Q,S,R) \geq \theta \end{aligned} \tag{11}$$

*Table 2. Summaries of high validity, which express age of respondents.*

| Linguistic summary | Validity as defined in (6) |
|---|---|
| About half respondents are old | 1.0000 |
| Few respondents are middle-aged | 0.8575 |
| About half respondents are young | 0.8570 |

where $\bar{Q}$ is a set of quantifiers of interest, $\bar{R}$ and $\bar{S}$ are sets of relevant linguistic expressions for restriction and summarizer, respectively, and $\theta$ is a threshold value from the interval ]0, 1]. In this case, all feasible solutions ($Q*$ $R*$ *are* $S*$) create a story.

In Example 1, we have the following task:

find $Q, S$

subject to

$Q \in \{few,\ about\ half,\ most\ of\}, S \in \{young,\ middle-aged,\ old\}, v(Q,S,R) \geq 0.75$

To also take the quality aspect into account, the constraint related to the validity threshold in (11) could be replaced by

$$Q_c(Q,S,R) \geq \theta_k \tag{12}$$

where $\theta_k$ is a threshold value from the interval ]0, 1] related to quality expectations.

The following question naturally arises: how can we efficiently obtain LSs from large data sets? When the number of records and their attributes is relatively large, the computation might take much more time, hence might be costly. For instance, when having 2,927 records described by 800 attributes, it is necessary to compute 2927·800 membership degrees (Niewiadomski et al. 2006). We can avoid such an amount of computation by optimization procedures based on the calculated proportions using matching degrees and involving users to select sets of relevant attributes for $\bar{S}$, $\bar{R}$ and $\bar{Q}$ (11). Moreover, the processor power and memory size of modern computers ensure that the response time is not too high (the examples on municipal statistics were executed within a few seconds on an ordinary desktop computer).

### *3.2. Linguistically Expressing Data Distributions Around the Mean Value*

The well-known and often used SQL query language contains a function for computing arithmetic mean, abbreviated as AVG, as well as a function for calculating standard deviation, abbreviated as STDEV from databases or data warehouses.

We have extended this functionality for LSs. The procedure is as follows: In the first step, the SQL query retrieves the mean value $M$, standard deviation and number of records of a chosen attribute with the following SELECT statement

```
SELECT AVG(chosen_att) as M, STDEV(chosen_att) as st_dev,
COUNT(id_record) as n
```

where *chosen_att* stands for the attribute selected by the user. The retrieved mean value $M$ is a modal element of a triangular fuzzy set plotted in Figure 3. This fuzzy set is created by a widening factor *wf* of a membership function to get symmetric and convex fuzzy number "*around the mean value M*". The lowest and the highest values of support are calculated in the following way

$$a = M - wf \cdot M; \quad b = M + wf \cdot M \tag{13}$$

In the next step, all values of *chosen_att* that belong to the support of the fuzzy set "*around the mean value M*" (]*a*, *b*[ – Figure 3) are selected from the database by the following SELECT statement

SELECT chosen_att FROM municipalities WHERE chosen_att BETWEEN (a, b)

Consequently, matching degrees for these values to the fuzzy set "*around mean value M*" are calculated, summed and divided by *n* for $y_{LSb}$. In the final step, matching degrees to the respective quantifiers are calculated; the relevant linguistic interpretation is constructed and shown in the interface.

The benefit of triangular fuzzy numbers against the interval *around the mean value M* is in the intensity of belonging. The closer an element is to the boundary, the lower matching degree the element has, and accordingly, its influence on the proportions $y_{LSb}$ and $y_{LSr}$ is lowered. The next example, illustrating our procedure, is based on the Municipal Statistics Database.

*Example 2*

A historian wishes to examine the mean value of *the year of the first written notice* (an attribute in the aforementioned municipal database). In addition, the historian has divided municipalities into two sets: "*population less than 12,000*" and "*population greater than or equal to 12,000*". The interface for interpreting solutions is shown in Figure 4. In this interface, the user can choose the relevant attribute and relative dispersion *wf* around the mean value; *wf* is set to 10% by default. The user can add further conditions merged by the logical "and operator" for focusing on the more restrictive subset of municipalities, or merged by the logical "or operator" for the less restrictive subset.

Via this interface, the historian can discover that the mean value of *the year of first written notice* for the municipalities with low population is the year 1363, and also that about half of them have their year of first written notice in the vicinity of the mean value (Figure 4 – upper interface). Hence, the mean value is a suitable generalization. For the municipalities with high population, the situation is the opposite. The mean value is the year 1147, but few municipalities fully or partially belong to the neighbourhood of this mean value (Figure 4 – lower interface).

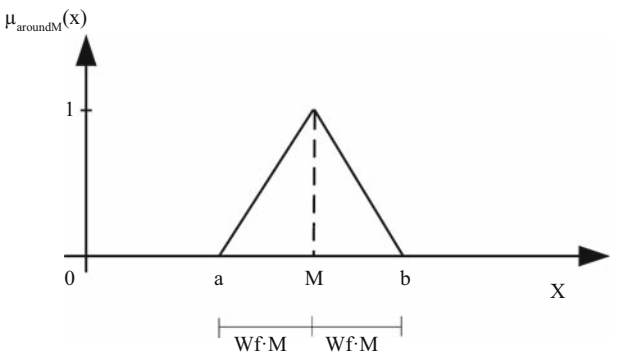


*Fig. 3. Triangular fuzzy set "around the mean value M".*

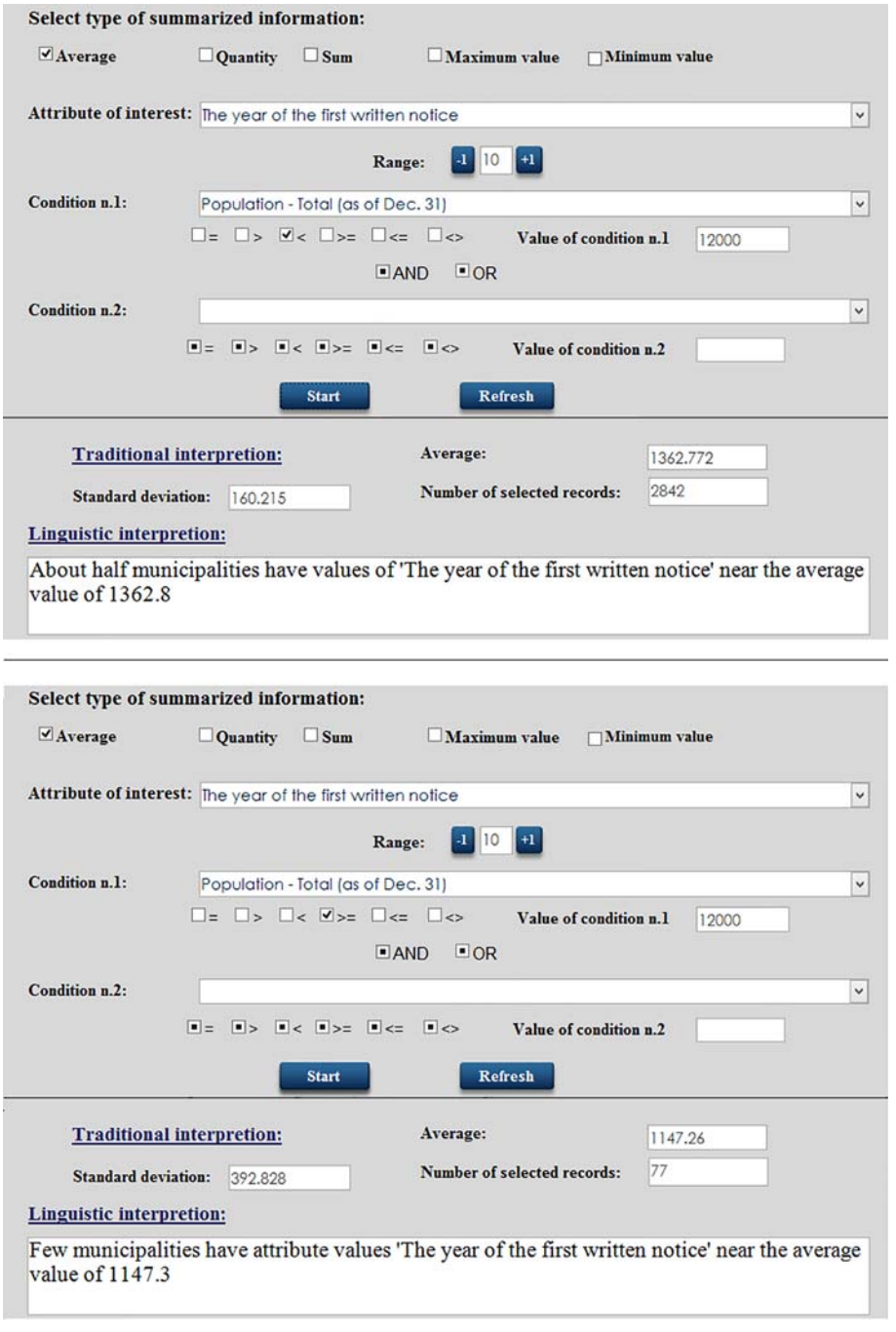


*Fig. 4. The interface for linguistically interpreting data distribution around the mean value (upper interface for population $<$ 12,000, lower interface for population $\geq$ 12,000).*

These interpretations are suitable for advanced, as well as for less advanced (in terms of statistical literacy and IT skills) users of official statistics, because the well-known statistical measures are disseminated together with their verbal interpretations. Additional functionalities can be added when required. For systems based on fuzzy logic, the following observation holds: "With any given system, it is easy to layer on more functionality without starting again from scratch" (Meyer and Zimmermann 2011, 432).

### 3.3. *Quantified Sentences as Nested Query Conditions*

This class of queries is suitable for the 1:N relationships in relational databases, or dimensions and facts in data warehouses, such as DISTRICT-RESPONDENT (one district contains multiple respondents, but each respondent is settled in one district).

An example of a quantified query condition is: "*find regions where most of the municipalities have a high amount of waste produced per inhabitant*". The algorithm is not complicated, but it might take more time depending on the number of entities on the "1" side of the considered relationship. The formula for calculating validities for each class $j$ on the "1" side is created as the extension of (5) and (6) in the following way (Hudec 2016)

$$v_{LSbj}(x) = \mu_Q\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\mu_S(x_{ij})\right), \quad j = 1 \ldots K \tag{14}$$

where $n_j$ is the number of entities in class $j$ (e.g., municipalities belonging to the region $j$), $K$ is the number of classes in a database (e.g., regions) and $v_{LSbj}$ is the validity of LS for $j$th class. Similarly, the nested query condition expressed by LS with restriction can be constructed by extending (7) and (8).

*Example 3*

A small enterprise is interested in extending its business activities related to agricultural equipment into the areas of low altitude and high ratio of arable land, but it is not sure which regions to favor. Hence, the SQL-like flexible query is: *SELECT regions WHERE most of the municipalities have low altitude and high ratio of arable land*. The decision maker considers an altitude of less than 200 m to perfectly match, between 200 m and 270 m to partially match and above 270 m to be out of the question. Thus, we formalize this user's linguistically expressed requirement by the fuzzy set *low*, where $m = 200$ and $b = 270$ (Figure A.1 (see Subsection 6.1)). The high ratio of arable land is formalized by the fuzzy set *high* plotted in Figure A.1, where $a = 40$ and $m = 60$. The quantifier *most of* is formalized by (4). For the "and operator" in the summarizer the minimum t-norm (A.1) was used. The result is presented in Table 3, where two regions (out of eight) partially meet the condition.

*Table 3. Selected regions by quantified query condition: "most of the municipalities have low altitude above sea level and high ratio of arable land".*

| Region | Validity as defined in (14) |
|---|---|
| Nitra | 0.930 |
| Trnava | 0.603 |

In the case of a classical database query, none of the regions meet the query condition and therefore the result is the empty set.

A further benefit for users may be to disseminate these results on thematic maps, for example, highlighting territorial units by hues, the intensity of which would be determined by the validity calculated by (14).

This type of summaries displays records on the higher hierarchy level, not data on lower levels. This is convenient when data on lower levels are sensitive. Hence, the risk of data disclosure is reduced, but care should be taken when summarizing from a small data set.

### *3.4. Summaries about Attributes*

The basic structure of LSs (6) provides a summary across the database for a particular subset of attributes. In order to practically illustrate this, we have developed a procedure and an illustrative interface (see Figure 5) for the aforementioned municipal database. The user selects the desired quantifier, chooses a relevant attribute from the database and selects the desired LV label (Subsection 2.2 and Figure 1). Consequently, the suggested parameters of a chosen label (fuzzy set) are shown under the picture of LV.

*Example 4*

A journalist examines distances to the nearest train stop for the purpose of writing an article regarding the train network coverage in municipalities. As shown in Figure 5, the interface requires the user to select a relevant quantifier, in our case *about half*, an attribute *Distance in km to the nearest passenger train station* and a label *low*. The value of 0 km means that the train station is situated within the municipality in question, whereas a

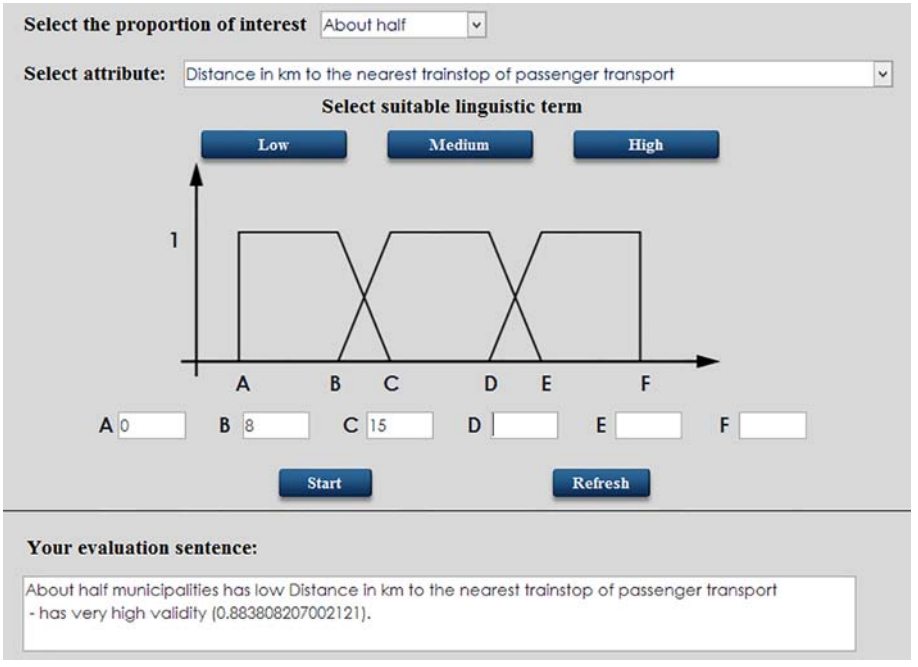


*Fig. 5. The illustrative interface for creating a LS and interpreting its validity.*

*Table 4. A possible mapping from validities (6) and (8) into a linguistic interpretation.*

| Validity | Linguistic explanation |
|---|---|
| 0 | Sentence is irrelevant |
| ]0, 0.15] | Sentence has very low validity |
| ]0.15, 0.4] | Sentence has low validity |
| ]0.4, 0.6] | Sentence has medium validity |
| ]0.6, 0.85] | Sentence has high validity |
| ]0.85, 1[ | Sentence has very high validity |
| 1 | Sentence excellently explains data |

distance greater than 0 indicates how far from the municipality the nearest train stop is. When the journalist chooses the label *low*, the lowest value from the database is shown (parameter $A$) and initial parameters for the fuzzy set *low* (parameters $B$ and $C$) are suggested. In the next step, the user can modify the parameters to values more suitable for a particular task. Finally, the linguistic interpretation is shown in the explanation box. In our case, the sentence "*about half of the municipalities have a low distance to the nearest passenger train station*" has a very high validity. The validity value in brackets is just shown for the purpose of illustration, and would be hidden by default. The rating of a linguistic explanation depends on the validity of quantified sentences. A possible mapping from $v_{LSb}$ or $v_{LSr}$ into a linguistic interpretation is shown in Table 4.

### 3.5. *Summaries for Subsets Expressed by Linguistic Summaries with Restrictions*

To demonstrate summaries based on (8), we have developed a procedure and an interface for summarizing from the aforementioned municipal database.

*Example 5*

An environmental agency is interested in learning whether "*the majority of municipalities with a high ratio of arable land have a low population density*". The interface is shown in Figure 6. On the upper left-hand side, the user chooses the relevant quantifier from a drop-down list (and modifies its parameters if needed). In the main part, the user selects

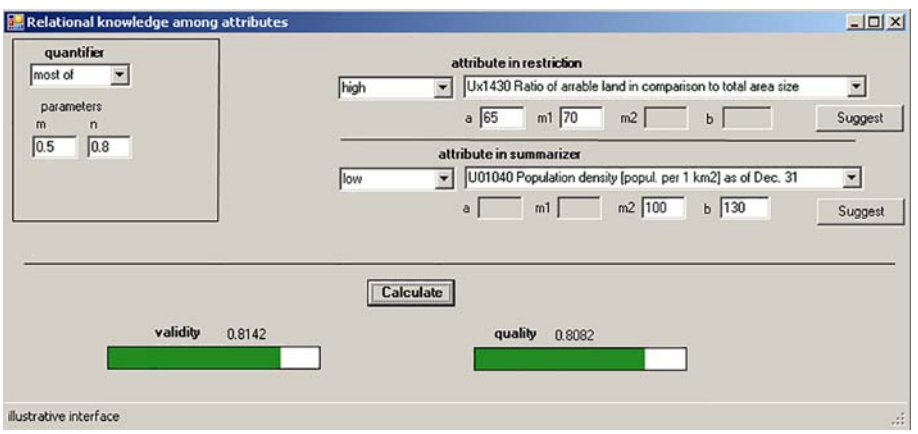


*Fig. 6. The illustrative interface for analysing LS with restriction.*

attributes and desired linguistic labels. The user can directly assign values to the respective parameters of labels, or ask for suggestions. In the latter case, parameters are mined from the database and presented to the user, who has the choice to either accept or modify them. Parameters $a$, $m_1$, $m_2$, $b$ correspond to the fuzzy sets parameters shown in Figure A.1. For the fuzzy set expressing the term *high* we have that $m_1 = m$, and for the fuzzy set formalizing the term *low* we have that $m_2 = m$. The chosen parameters are shown in Figure 6. The validity of this summary is 0.814 (defined by (8)) and its quality is 0.8082 (defined by (9)). These high values lead us to the conclusion that the summarized sentence is of a high quality.

If an agency is interested in investigating whether "*most of the municipalities with a high number of warm days (temperature above 25°C) have a small amount of waste produced per inhabitant*", then the validity (8) of this summary is equal to 1. However, the coverage (A.4) is equal to 0 and therefore the quality (9) is 0, for example, this summary is not representative. Focusing only on validity might lead us to draw inappropriate conclusions.

## 4. Discussion

This section provides the main features of the suggested approach and a reflection on its advantages and drawbacks, as well as some further research opportunities.

### *4.1. The Main Features of the Suggested Approach*

In this article, we adjusted well-known approaches for formalizing flexible predicates, quantifiers and quality measures, and provided the rationale for our choices. The suggested approach may be beneficial for NSIs due to the following features:

- It is less sensitive to the imprecise nature of some data and to inliers (i.e., erroneous values that lie in the normal range of a variable). When the measured value is not far from the real one, then this approach eliminates sharp jumps between belonging and not belonging to a set (Figures 1 and A.2 (see Subsection 6.1)).
- The suggested approach reveals summaries from the data, not the data itself. Generally speaking, the data disclosure would not be a problem; however, care should be taken when summarizing from small data sets. The decision regarding which data sources might be available for users to realize summaries should meet regulations and other relevant rules.
- The less complex interpretation of the data is especially welcome for less statistically literate users and disabled people, for whom the summarized sentences may be interpreted by voice.
- The *computing with words* concept can easily be applied to any human language. Adjectives such as *high* and quantifiers such as *most of* are always expressed by increasing functions, regardless of their translation to the other languages and examined concepts.
- LSs are able to offer an alternative answer when the initial sentence (summary) is of insufficient validity. For instance, if the proportion for the sentence "*most short visits are from countries with high GDP*" is 0.06, the answer is not only that the validity is

zero, but we can provide an alternative summary: "*few short visits are from countries with high GDP*".

- Statistical offices typically refrain from disseminating dispersion measures, although this information is valuable. Our suggested approach includes the linguistically interpreted deviation, which is suitable for all users, especially for the less statistically literate ones.

### *4.2. Further Research and Development Opportunities*

The following subsections identify opportunities for future research topics.

#### 4.2.1. Reflections on User Input

The quality measures for LSs are usually calculated from the data, excluding human intervention. While this might sometimes be convenient, it might sometimes be useful to develop an additional measure, for which the user would be able to assign relevance to each summary of interest.

The interfaces introduced in Section 3 were created for illustrative purposes. The interfaces in Figures 4 and 5 might be suitable for both types of users, since well-known statistical measures and linguistic interpretations are provided. The interface for summaries among attributes (Figure 6) may be difficult to use for less skilled users. On the other hand, experienced ones might welcome the possibilities of adjusting all relevant parameters of linguistic labels. The option provided to the less skilled users by the test interface is the automated support. Further options might be inspired by ReqFlex – a "fuzzy query engine for everyone", developed by Smits et al. (2013), where the users assign parameters by moving sliders rather than filling input boxes. Future research should include sophisticated usability testing and adjusting various designs according to the user feedback in order to meet the expectations of both advanced and less advanced users. While this approach is applicable for summarizing, for example at the European Union Member State level, the benefit of LSs is in general higher for larger data sets, such as on levels of Nomenclature of Territorial Units for Statistics (NUTS).

#### 4.2.2. Linguistic Quality

A possible obstacle might be the structure of short quantified sentences (indicated in italics throughout this article). Although such structures are widely used, the order of terms and the structure itself might not fully meet the usual terminology in official statistics, general public expectations and grammar rules. A mechanical construction of sentences may lead to grammatically incorrect expressions. Thus, there is room for experts from different fields, including linguists, to identify sound and practical solutions, but interactive machine learning could also be of help here. Moreover, verbal explanations are extremely important for the emerging field of "explainable artificial intelligence" (Goebel et al. 2018), which opens additional application fields.

#### 4.2.3. Applying SDMX to Summaries

The Statistical Data and Metadata eXchange (SDMX) standard was initially developed for the dissemination and exchange of data (SDMX 2012). The dimensional data structure is

solid, because it is based on a clear methodology and is therefore suitable for inclusion into business intelligence questions. This structure can be helpful for creating linguistic variables over a set of dimensions and measures. The possibility of managing fuzzy data by the SDMX standard is touched upon by Hudec and Praženka (2016).

#### 4.2.4. Applying Linguistic Summaries to New Data Sources

National Statistical Institutes (NSI) are also focusing their activities on alternative sources, including social networks (e.g., Torres van Grinsven and Snijkers 2015), web scraping (e.g., Barcaroli et al. 2015), mobile positioning data (e.g., Altin et al. 2015) and the like. Because the validities (6) and (8), as well as the quality measure (9) depend only on the intensities of belonging to fuzzy sets, it means that we can straightforwardly summarize from other data types. The only difference is in computing matching degrees of imprecise numbers (known as fuzzy numbers), (weighted) categorical data and sentence fragments to fuzzy sets in summarizer and restriction. For weighted categorical data (e.g., *negative (0.7) and neutral (0.3) opinion*) and fuzzy data (e.g,. *value is most likely 120 but for sure not lower than 100 and not higher than 150*) instead of calculating matching degrees of crisp numbers to the fuzzy sets, the possibility and necessity measures are applied (Galindo et al. 2006). For data expressed by short sentences or sentence fragments (e.g., *productivity is remarkably low*) matching degrees to the fuzzy concepts can be calculated by application of methods suggested by Duraj et al. (2015) and Niewiadomski (2002).

#### 4.2.5. Enhanced Dissemination as an Incentive for Data Providers

Although data collection and dissemination are at two opposite ends of the statistical data production process, they influence each other. Adolfsson et al. (2010) estimated that 30% of total data collection costs is allocated to data editing (imputation). Ross (2009) observed the paradox that users of official statistics are becoming more demanding with regard to data, but are less willing to provide their own data to NSIs. This problem results from the fact that respondents cooperate in many official surveys, but on the other hand, they often are not able to easily find and interpret relevant information on NSI data portals (Bavdaž 2011). One possible solution is in flexible and tailored data dissemination (Hudec and Torres van Grinsven 2013). As further motivation, we could offer sophisticated methods for linguistically interpreted summaries (means, deviations, time series, etc.) to businesses that cooperate timely in surveys. The practical feasibility of achieving this (while maintaining the principle of impartiality) is a topic for future research.

## 5. Conclusions

One of the missions of NSIs is the dissemination of statistical data to a large variety of users, ranging from experts to the general public (including disabled people). Statistical agencies should offer flexibility in dissemination to avoid jeopardizing their mission (Bavdaž 2011). This may require rules for using natural human languages to describe key measures (Schield 2011) and to make statistics easily understandable and usable by the general public (Bier and Nymand-Andersen 2011). Thus, NSIs should apply different strategies in order to meet the expectations of diverse user categories. This article tackles innovative dissemination by short quantified sentences of natural language, which is

definitely a promising method to reach these goals. In particular, for some categories of users with disabilities, textual interpretation or interpretation by voice (rendered possible by LSs) would be more suitable than what is offered by current dissemination methods.

The potential of LSs is demonstrated on test interfaces on the real-world data. In order to reduce both complexity and interaction requirements on users, we have suggested approaches for constructing fuzzy sets and for measuring quality that minimally burden users. Further, our research has documented perspectives, obstacles and problems leading to future research directions. The important activity is in real-world testing with users to develop broadly accepted designs for full-featured and easy-to-use interfaces. These tasks should be solved in cooperation between NSI data dissemination units and scientists working in the aforementioned fields.

Finally, we emphasize that our approach based on LSs should not be considered as a rival to existing ones, but rather as a complementary dissemination practice to well-established ones.

## 6. Appendix A. Theoretical Concepts Related to Fuzzy Set Theory and Linguistic Summaries

This appendix provides an insight into fuzzy set theory, fuzzy "and operator" and quality measures of summaries.

### *6.1. Fuzzy Sets*

The linguistic terms *low*, *medium* and *high* (Figure 1) can be formalized by an L fuzzy set, a trapezoidal fuzzy set and a linear gamma fuzzy set, respectively, as illustrated in Figure A.1.

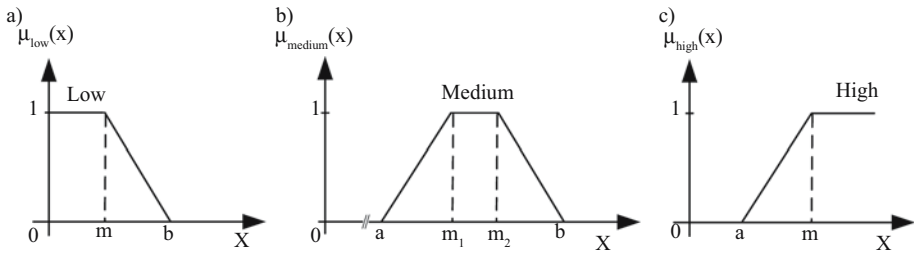


*Fig. A.1. Fuzzy sets: a) L fuzzy set, b) trapezoidal, c) linear gamma.*

Fuzzy sets are context dependent, for example, they may have different parameters for various given concepts. For instance, the set "*short distance*" has a different meaning – expressed by parameters $m$ and $b$ in Figure A.1 – for a small and densely populated country, and for a large but sparsely inhabited country. Two important concepts are the core and the support of fuzzy sets. The core of a fuzzy set contains all elements that fully belong to the set. The core of the fuzzy set *medium* contains all elements in interval $[m_1, m_2]$. The support of the fuzzy set contains all elements that belong to the set with degree greater than 0, that is, the support of fuzzy set *medium* is interval [a, b].

For instance, assume that someone wishes to know whether certain municipalities belong to the set "*high pollution*" (*HP*). The set *HP* is expressed as a fuzzy set shown in

Figure A.2a, and as a crisp set in Figure A.2b, where $\varphi$ is a characteristic function (a bi-valued function expressing membership of a crisp set). In Figure A.2a, the values 50 mg and 55 mg delimit the area where belonging to the set is a matter of degree. If we apply classical set theory, two similar values may be treated differently. For example: a municipality, in which a value of 54.73 mg was recorded, does not belong to the crisp set *HP*, whereas a municipality having a recorded value of 55 mg does belong to it. In the case of a fuzzy set, a municipality polluted with 54.73 mg participates in the set *HP* with a slightly lower degree than 1. The possible measurement error for values around 55 mg may cause assignment to the wrong crisp set.

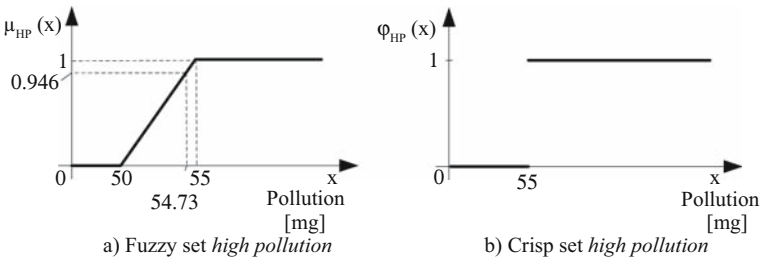


*Fig. A.2. Concept "high pollution" expressed as fuzzy set (a) and crisp set (b).*

On the other hand, when a categorization relies on precise or sharp rules, we should use crisp sets. For instance, the category Small and Medium-sized Enterprise (SME) is divided into three subsets (by number of employees): micro enterprises – fewer than 10 persons employed; small enterprises – from 10 to 49 persons employed; medium-sized enterprises – from 50 to 249 persons employed (e.g., EU Guide 2015). In this case, these sets have sharp boundaries (or $\alpha = 0$ in Figure 1).We can still use these sets in LSs, for example, to assess whether "*few micro enterprises in tourism have low turnover*".

### 6.2. *Triangular Norms*

The "and operator" is expressed by triangular norms, which were initially developed for statistical metric spaces and later modified and applied for the fuzzy "and operator" (Schweizer and Sklar 1983).

When a restriction $R$ and/or summarizer $S$ consisting of several atomic predicates aggregated by the "and operator", triangular norms (t-norms) should be used. Two well-known t-norms, both of which are discussed in Klement et al. (2005), are the minimum t-norm

$$\mu_P(x) = \underbrace{\min}_{i=1\ldots n} \mu_{Pi}(x) \tag{A.1}$$

and the product t-norm expressed as

$$\mu_P(x) = \prod_{i=1}^{n} \mu_{Pi}(x) \tag{A.2}$$

where $P$ stands for the compound predicate. All t-norms meet all axiomatic properties of "and operator", but differ in satisfying algebraic properties (Klement et al. 2005) to cover a variety of tasks.

### 6.3. Quality Measures of Summaries

The basic quality criterion (validity or truth value as defined in (6) and (8)) is the most important one, but it does not cover all aspects of quality (Kacprzyk and Yager 2001). Let us focus on LSs with restriction (7) and (8). It is possible that the validity equal to 1 explains the summary from the outliers (Hudec 2017). In order to avoid this problem, several quality measures have been suggested.

Hirota and Pedrycz (1999) have introduced five features for measuring quality of mined and aggregated information: validity, novelty, usefulness, simplicity and generality. Wu et al. (2010) have proposed equations for calculating these measures for linguistic summaries with restriction. In that approach, validity corresponds to (8). The generality measure is expressed by sufficient coverage that indicates whether a summary is supported by a sufficient subset of the data. First, the coverage ratio is calculated as (Wu et al. 2010)

$$i_c = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{A.3}$$

where $n$ is the number of records and $p_i = \begin{cases} 1 & \mu_S(x_i) > 0 \wedge \mu_R(x_i) > 0 \\ 0 & \text{otherwise} \end{cases}$

Because a summary of the structure (8) covers a subset of the whole database, $i_c$ is considerably smaller than 1. Thus, the following mapping $[0, 1] \rightarrow [0, 1]$ converts this ratio into the degree of sufficient coverage (Wu et al. 2010)

$$C = f(i_c) = \begin{cases} 0 & i_c \leq r_1 \\ 2((i_c - r_1)/(r_2 - r_1))^2 & r_1 \leq i_c < (r_1 + r_2)/2 \\ 1 - 2((r_2 - i_c)/(r_2 - r_1))^2 & (r_1 + r_2)/2 \leq i_c < r_2 \\ 1 & i_c \geq r_2 \end{cases} \tag{A.4}$$

where the suggested values for parameters $r_1$ and $r_2$ are 0.02 and 0.15, respectively.

The degree of usefulness is computed as a minimum of validity and coverage (i.e., $U = \min(v_{LSr}, C)$). The degree of outlyingness $O$, referring to novelty (unexpected summaries are very valuable for users if they cover the regular behavior in the data, not in outliers), is an aggregation of validity and coverage as: "the validity degree $v$ is very small or very high and the sufficient coverage C must be very small" (Wu et al. 2010, 14). To keep the best value of each measure equal to 1, instead of the outlier measure, we should use its negation $(1 - O)$. Finally, the simplicity measure expresses the length of a sentence as (Wu et al. 2010)

$$SL = 2^{2-|S\cup R|} \tag{A.5}$$

where $|S \cup R|$ is the cardinality of union between $R$ and $S$. When $R$ and $S$ contains one attribute each, the simplicity measure gets the value 1. All aforementioned measures get values from the unit interval, which makes their aggregation easier, but some measures are functionally dependent (Hudec 2017).

Kacprzyk and Strykowski (1999) have introduced the following quality measures: truth value or validity, degree of precision, degree of coverage, degree of appropriateness, and

length of summary. These measures are mainly focused on the basic structure of LSs, (5). The truth value ($T_1$) basically corresponds to validity (6). The degree of fuzziness is high for summaries based on very vague attributes in $S$. The wider the support of fuzzy set, the higher the value of fuzziness, that is

$$d_{fz}(S_j) = (|\{x \in A_j : \mu_{Sj}(x) > 0\}|)/(|A_j|) \tag{A.6}$$

where $S_j$ is predicate on attribute $A_j$ in summarizer $S$. This quality measure, the degree of precision, is defined as

$$T_2 = 1 - \sqrt{\prod_{j=1}^{s} d_{fz}(S_j)} \tag{A.7}$$

where $s$ is the number of atomic predicates in summariser $S$. Values close to 1 are associated with summaries of low fuzziness.

The degree of coverage ($T_3$) basically corresponds to (A.3) and (A.4). The degree of appropriateness is a measure functionally dependent on $T_3$

$$T_4 = \left| \prod_{j=1}^{s} k_j - T_3 \right| \tag{A.8}$$

where $k_j = \left(\sum_{i=1}^{n} h_i\right)/n$, and $h_i$ is defined to be equal to 1 when the $i$th record satisfies membership function $\mu$ for $S_j$, and 0 otherwise. The role of this measure is to exclude trivial summaries of high validity.

The length of the summary corresponds to (A.5), but it is adjusted to the basic structure of LSs by

$$T_5 = 2 \cdot 0.5^{|S|} \tag{A.9}$$

This measure gets value 1 when the cardinality of $S$ is equal to 1, that is, $S$ consists of one atomic predicate.

The problem of applying (A.6) to summaries on the Municipal Statistics Database is that for many attributes, the data distribution is unbalanced and therefore a low value of $T_2$ does not necessarily imply low quality. In addition, users may have particular reasons to express requirements by "wide" fuzzy sets. Regarding the summary length, we should use (A.5) for LSs with restriction and (A.9) for the basic structure.

Another problem is the aggregation of quality measures. Kacprzyk and Yager (2001) suggest the weighted average

$$T = \sum_{i=1}^{5} w_i T_i \tag{A.10}$$

where $\sum_{i=1}^{5} w_i = 1$.

For example, this way is suitable for decision support (e.g., in the medical domain), where decision makers assign values to $w_i$ either individually or by consensus. On the other hand, this way is not applicable for disseminating statistical data to the general public, because assigning weights imposes a burden on users.

George and Srikanth (1996) have developed a genetic algorithm for fitness function to compute the best summary. Having the simplicity and robust solution for statistical dissemination in mind, this way is not elaborated further.

### *6.4. A Brief Review of Using Fuzzy Sets in Queries*

The first practical implementations of flexible queries were FQUERY introduced by Kacprzyk and Zadrożny (1995) and SQLf introduced by Bosc and Pivert (1995). These approaches faced the problems of covering complex aggregation operators. Quantified query conditions, that is, selecting entities that meet the majority of atomic conditions, were introduced by Kacprzyk and Ziółkowski (1986). An illustrative example is to find municipalities where most of the conditions "*altitude above sea level is around 700 m and population density is small and municipality size is medium and pollution is low and opinion about municipality is positive*" are satisfied. The empty answer problem is an issue when a higher number of atomic conditions is merged by the "and operator". Quantified query conditions based on LSs mitigate this problem by retrieving not only entities that meet all atomic conditions, but also entities that meet the majority of these conditions.

The first querying tool for summarizing the data was SummarySQL (Rasmussen and Yager 1997) followed by SAINTETIQ (Raschia and Mouaddib 2002) and the extension of FQUERY (Kacprzyk and Zadrożny 2005). Achievements related to the official statistics data dissemination community were mainly focused on the fuzzy queries (Hudec 2013).

## 7. Appendix B. Overview of Symbols Used

| | |
|---|---|
| a | Left border of fuzzy set support |
| A | Attribute, topic |
| $\alpha$ | Length of the uncertain area in fuzzy set |
| b | Right border of fuzzy set support |
| C | Coverage |
| $d_{fz}$ | Degree of fuzziness |
| f | Function |
| F | Fuzzy set |
| $\varphi$ | Characteristic function of crisp set |
| G | Syntactic rule for *LV* |
| h | Parameters used to calculate $T_4$ |
| H | Semantic rule for *LV* |
| $i_c$ | Coverage ratio |
| k | Parameter used to calculate $T_4$ |
| K | Number of classes |
| L | Name of linguistic variable |
| LS | Linguistic Summary |
| LV | Linguistic Variable |
| m | Modal value of fuzzy set |
| $m_1$ | Left border of fuzzy set core |

| | |
|---|---|
| $m_2$ | Right border of fuzzy set core |
| M | Mean value (average) |
| $\mu$ | Membership function of fuzzy set |
| $\mu_F(x)$ | Membership degree of element $x$ to fuzzy set $F$ |
| n | Number of records |
| O | Outlier degree |
| p | Parameters used to calculate $i_c$ |
| P | Predicate |
| Q | Fuzzy quantifier |
| $Q_c$ | Quality measure aggregating validity and coverage |
| r | Parameters used to calculate coverage $C$ from its ratio $i_c$ |
| R | Restriction |
| s | Number of anomic predicates in summarizer $S$ |
| S | Summarizer |
| SL | Simplicity measure |
| t | t-norm |
| T | Set of labels (linguistic terms) |
| $T_1$ | Truth value |
| $T_2$ | Degree of precision |
| $T_3$ | Degree of covering |
| $T_4$ | Degree of appropriateness |
| $T_5$ | Length of summary |
| $\theta$ | Threshold value |
| $Q_T$ | Fuzzy quantifier applied to time attribute |
| U | Usefulness measure |
| v | Validity or truth value of summary |
| w | Weight |
| wf | Widening factor |
| X | Universe of disclosure |
| x | Element of universal set |
| y | Proportion |
| $y_{LSb}$ | Proportion in basic structure of a summary |
| $y_{LSr}$ | Proportion in summary with restriction |

## 8. References

Adolfsson, C., G. Arvidson, P. Gidlund, A. Norberg, and L. Nordberg. 2010. "Development and Implementation of Selective Data Editing at Statistics Sweden." In Proceedings of the European Conference on Quality in Official Statistics, May 4, 2010. Helsinki Available at: https://q2010.stat.fi/media//presentations/Norberg_et_all__Statistics_Sweden_slutversion.pdf (accessed April 2017).

Almeida, R.J., M-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse. 2013. "Linguistic Summaries of Categorical Time Series Septic Shock Patient Data." In Proceedings of the 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2013), July 7–10, 2013. 1–8. Hyderabad.

Altin, L., M. Tiru, E. Saluveer, and A. Puura. 2015. "Using Passive Mobile Positioning Data in Tourism and Population Statistics." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2015), March 10–12, 2015. Brussels. Available at: https://ec.europa.eu/eurostat/cros/system/files/Altin-etal_abstract_ntts_2301LA_0.pdf (accessed January 2017).

Arguelles, L. and G. Triviño. 2013. "I-struve: Automatic Linguistic Descriptions of Visual Double Stars." *Engineering Applications of Artificial Intelligence* 26: 2083–2092. Doi: http://dx.doi.org/10.1016/j.engappai.2013.05.005.

Barcaroli, G., M. Scannapieco, D. Summa, and M. Scarnò. 2015. "Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2015), March 10–12, 2015. Brussels. Available at: https://ec.europa.eu/eurostat/cros/system/files/Barcaroli-etal_WebScraping_Final_unblinded.pdf (accessed February 2017).

Bavdaž, M. (editor). 2011. *Final Report Integrating Findings on Business Perspectives Related to NSIs Statistics*. Brussels: European Commission. (Deliverable 3.2 from FP7 project BLUE-Enterprise and Trade Statistics). Blue-Ets Project: SSH-CT-2010-244767.

Bier, V. and P. Nymand-Andersen. 2011. "Communicating Statistics to Frequent Users – One Size Fits All?" In Proceedings of the Committee for the Coordination of Statistical Activities (CCSA Special Session), September 8, 2011. Luxembourg.

Boran, F.E., D. Akay, and R.R. Yager. 2016. "An Overview of Methods for Linguistic Summarization with Fuzzy Sets." *Expert Systems with Applications* 61: 356–377. Doi: http://dx.doi.org/10.1016/j.eswa.2016.05.044.

Bosc, P. and O. Pivert. 1995. "SQLf: a Relational Database Language for Fuzzy Querying." *IEEE Transactions on Fuzzy Systems* 3: 1–17. Doi: http://dx.doi.org/10.1109/91.366566.

Coddington, M. 2015. "Clarifying Journalism's Quantitative Turn." *Digital Journalism* 3: 331–348. Doi: http://dx.doi.org/10.1080/21670811.2014.976400.

Disability Rights Commission. 2004. *The Web Access and Inclusion for Disabled People – A Formal Investigation conducted by the Disability Rights Commission*. London: TSO. Available at: https://www.city.ac.uk/__data/assets/pdf_file/0004/72670/DRC_Report.pdf (accessed, May 2018).

Duraj, A., P.S. Szczepaniak, and J. Ochelska-Mierzejewska. 2015. "Detection of Outlier Information Using Linguistic Summarization." In Proceedings of the 11th International Conference Flexible Query Answering Systems (FQAS 2015), October 26–28, 2015. 101–113. Cracow.

EU Guide. 2015. *User guide to the SME Definition*. Luxembourg: Publications Office of the European Union. Available at: http://ec.europa.eu/growth/tools-databases/newsroom/cf/itemdetail.cfm?item_id=8274&lang=en (accessed November, 2016).

Galindo, J., A. Urrutia, and M. Piattini. 2006. *Fuzzy Databases—Modeling. Design and Implementation*. Hershey: Idea Group Publishing.

George, R. and R. Srikanth. 1996. "Data Summarization Using Genetic Algorithms and Fuzzy Logic." In *Genetic Algorithms and Soft Computing*, edited by F. Herrera and J.L. Verdegay, 599–611. Heidelberg: Physica–Verlag.

Glöckner, I. 2006. *Fuzzy Quantifiers – A Computational Theory*. Berlin Heidelberg: Springer-Verlag.

GSIM. 2013. *Generic Statistical Information Model (GSIM): Specification*. Geneva: United Nations Economic Commission for Europe (UNECE). Available at: http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification (accessed February 2017).

Goebel, R., A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger. 2018. "Explainable AI: The New 42?" In *Machine Learning and Knowledge Extraction, Springer Lecture Notes in Computer Science LNCS 11015*, edited by A. Holzinger, P. Kieseberg, A. Tjoa, and E. Weippl, 295–303. Cham: Springer.

Graefe, A. 2016. *Guide to Automated Journalism*. New York: Tow Center for Digital Journalism. Available at: https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php (accessed April 2018).

Heimgärtner, R., A. Holzinger, and R. Adams. 2008. "From Cultural to Individual Adaptive End-User Interfaces: Helping People with Special Needs." In Proceedings of the 11th International Conference on Computers Helping People with Special Needs (ICCHP 2008), July 9–11, 2008. 82–89. Linz.

Hirota, K. and W. Pedrycz. 1999. "Fuzzy Computing for Data Mining." *Proceedings of IEEE* 87: 1575–1600. Doi: http://dx.doi.org/10.1109/5.784240.

Holzinger, A. 2002. "User-Centered Interface Design for Disabled and Elderly People: First Experiences with Designing a Patient Communication System (PACOSY)." In Proceedings of the 8th International Conference on Computer Helping People with Special Needs (ICCHP 2002), July 15–20, 2002. 33–40. Linz.

Holzinger, A., B. Malle, P. Kieseberg, P.M. Roth, H. Müller, R. Reihs, and K. Zatloukal. 2017. "Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach." In *Towards Integrative Machine Learning and Knowledge Extraction*, edited by A. Holzinger, R. Goebel, M. Ferri, and V. Palade, 13–50. Cham: Springer.

Hudec, M. 2013. "Improvement of Data Collection and Dissemination by Fuzzy Logic." In Proceedings of the Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), April 22–24, 2013. Paris and Bangkok. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_3_Slovakia.pdf (accessed January 2017).

Hudec, M. 2016. *Fuzziness in Information Systems – How to Deal with Crisp and Fuzzy Data in Selection, Classification, and Summarization*. Cham: Springer.

Hudec, M. 2017. "Merging Validity and Coverage for Measuring Quality of Data Summaries." In *Information Technology and Computational Physics*, edited by P. Kulczycki, L.T. Kóczy, R. Mesiar, and J. Kacprzyk, 71–85. Cham: Springer.

Hudec, M. and D. Praženka. 2016. "Collecting and Managing Fuzzy Data in Statistical Relational Databases." *Statistical Journal of the IAOS* 32: 245–255. Doi: http://dx.doi.org/10.3233/SJI-160956.

Hudec, M. and V. Torres Van Grinsven. 2013. "Business' Participants Motivation in Official Surveys by Fuzzy Logic." In Proceedings of the 1st Eurasian Multidisciplinary Forum (EMF 2013), October 24–26, 2013. 42–52. Tbilisi.

Kacprzyk, J. and P. Strykowski. 1999. "Linguistic Data Summaries for Intelligent Decision Support." In Proceedings of the fourth European Workshop on Fuzzy Decision

Analysis and Recognition Technology for Management, Planning and Optimization (EFDAN 1999), June 14–15, 1999. 3–12. Dortmund.

Kacprzyk, J., A. Wilbik, and S. Zadrożny. 2006. "Linguistic Summarization of Trends: A Fuzzy Logic Based Approach." In Proceedings of the 11th Information Processing and Management of Uncertainty in Knowledge Based Systems (IPMU 2006), July 2–7, 2006. 2166–2172. Paris.

Kacprzyk, J. and R.R. Yager. 2001. "Linguistic Summaries of Data Using Fuzzy Logic." *International Journal of General Systems* 30: 133–154. Doi: http://dx.doi.org/10.1080/03081070108960702.

Kacprzyk, J. and S. Zadrożny. 1995. "FQUERY for Access: Fuzzy Querying for Windows-Based DBMS." In *Fuzziness in Database Management Systems*, edited by P. Bosc and J. Kacprzyk, 415–433. Heidelberg: Physica-Verlag.

Kacprzyk, J. and S. Zadrożny. 2005. "Linguistic Database Summaries and Their Protoforms: Towards Natural Language Based Knowledge Discovery Tools." *Information Sciences* 173: 281–304. Doi: http://dx.doi.org/10.1016/j.ins.2005.03.002.

Kacprzyk, J. and A. Ziółkowski. 1986. "Database Queries with Fuzzy Linguistic Quantifiers." *IEEE Transactions Systems, Man and Cybernetics SMC-16* 3: 474–479. Doi: http://dx.doi.org/10.1109/tsmc.1986.4308982.

Klement, E.P., R. Mesiar, and E. Pap. 2005. "Triangular Norms: Basic Notions and Properties." In *Logical, Algebraic, Analytic, and Probabilistic Aspects of triangular Norms*, edited by E.P. Klement and R. Mesiar, 17–60. Amsterdam: Elsevier.

Lesot, M-J., G. Moyse, and B. Bouchon-Meunier. 2016. "Interpretability of Fuzzy Linguistic Summaries." *Fuzzy Sets and Systems* 292: 307–317. Doi: http://dx.doi.org/10.1016/j.fss.2014.10.019.

Liu, B. 2011. "Uncertain Logic for Modeling Human Language." *Journal of Uncertain Systems* 5: 3–20. Available at: www.jus.org.uk (accessed September 2012).

Meyer, A. and H.J. Zimmermann. 2011. "Applications of Fuzzy Technology in Business Intelligence." *International Journal of Computers, Communications & Control* VI(3): 428–441. Doi: http://dx.doi.org/10.15837/ijccc.2011.3.2128.

Moyse, G., M-J. Lesot, and B. Bouchon-Meunier. 2013. "Mathematical Morphology Tools to Evaluate Periodic Linguistic Summaries." In *Flexible Query Answering Systems*, edited by H.L. Larsen, 257–268. Berlin Heidelberg: Springer-Verlag.

Niewiadomski, A. 2002. "Appliance of Fuzzy Relations for Text Documents Comparing." In Proceedings of the 6th Conference on Neural Networks and Soft Computing (ICNNSC' 2002), June 11–15, 2002. Zakopane.

Niewiadomski, A., J. Ochelska, and P.S. Szczepaniak. 2006. "Interval-Valued Linguistic Summaries of Databases." *Control and Cybernetics* 35: 415–443. Available at: http://matwbn.icm.edu.pl/ksiazki/cc/cc35/cc35212.pdf (accessed June 2016).

Raschia, G. and N. Mouaddib. 2002. "SAINTETIQ: A Fuzzy Set-Based Approach to Database Summarization." *Fuzzy Sets and Systems* 129: 137–162. Doi: https://doi.org/10.1016/S0165-0114(01)00197-X.

Rasmussen, D. and R.R. Yager. 1997. "Summary SQL – A Fuzzy Tool for Data Mining." *Intelligent Data Analysis* 1: 49–58. Doi: http://dx.doi.org/10.1016/S1088-467X(98)00009-2.

Ross, M.P. 2009. "Official Statistics in Malta – Implications of Membership of the European Statistical System for a Small Country/NSI." In Proceedings of the 95th DGINS Conference, October 1, 2009. Malta. Available at: https://ec.europa.eu/eurostat/documents/1001617/4339944/MPR-opening-address-00909.pdf/7c298770-0869-415c-9833-d702e8b3ce9e (accessed October, 2016).

Scanu, M. and C. Casagrande. 2016. "The Generic Statistical Information Model (GSIM): State of Application of the Standard." In Workshop on Implementing Standards for Statistical Modernisation, 21–23 September 2016. Geneva. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2016/mtg4/Paper_17_Italy_-_The_Generic_Statistical_Information_Model__GSIM__and_the_Sistema_Unitario.pdf (accessed March 2017).

SDMX. 2012. *SDMX 2.1 User Guide, SDMX 2.1 Documentation*. SDMX Consortium. Available at: https://sdmx.org/?page_id=1119 (Accessed January 2017).

Schweizer, B. and A. Sklar. 1983. *Probabilistic Metric Spaces*. Amsterdam: North-Holland.

Schield, M. 2011. "Statistical Literacy: A New Mission for Data Producers." *Statistical Journal of the IAOS* 27: 173–183. Doi: http://dx.doi.org/10.3233/SJI-2011-0732.

Smits, G., O. Pivert, and T. Girault. 2013. "ReqFlex: Fuzzy Queries for Everyone." In Proceedings of the 39th International Conference on Very Large Data Bases, 26–30 August, Trento.

Torres van Grinsven, V. and G. Snijkers. 2015. "Sentiments and Perceptions of Business Respondents on Social Media: An Exploratory Analysis." *Journal of Official Statistics* 31: 283–304. Doi: http://dx.doi.org/10.1515/jos-2015-0018.

Wu, D., J.M. Mendel, and J. Joo. 2010. "Linguistic Summarization Using If-Then Rules." In Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, July 18–23, 2010. 1–8. Barcelona.

Yager, R.R. 1982. "A New Approach to the Summarization of Data." *Information Sciences* 28: 69–86. Doi: http://dx.doi.org/10.1016/0020-0255(82)90033-0.

Yager, R.R. 1984. "General Multiple-Objective Decision Functions and Linguistically Quantified Statements." *International Journal of Man-Machine Studies* 21: 389–400. Doi: http://dx.doi.org/10.1016/S0020-7373(84)80066-8.

Yager, R.R. 1988. "On Ordered Weighted Averaging Operators in Multicritera Decision Making." *IEEE Transactions on Systems, Man and Cybernetics,* SMC-18: 183–190. Doi: http://dx.doi.org/10.1080/03081070108960702.

Yager, R.R., M. Ford, and A.J. Canas. 1990. "An Approach to the Linguistic Summarization of Data." In Proceedings of the 3rd International Conference of Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 1990), July 2–6, 1990. 456–468. Paris.

Zadeh, L.A. 1965. "Fuzzy Sets." *Information and Control* 8: 338–353. Doi: http://dx.doi.org/10.1016/S0019-9958(65)90241-X.

Zadeh, L.A. 1975. "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning: Part I." *Information Sciences* 8: 199–249. Doi: http://dx.doi.org/10.1016/0020-0255(75)90036-5.

Zadeh, L.A. 1983. "A Computational Approach to Fuzzy Quantifiers in Natural Languages." *Computers & Mathematics with Applications* 9: 149–184. Doi: http://dx.doi.org/10.1016/0898-1221(83)90013-5.

Zadeh, L.A. 2001. "From Computing With Numbers to Computing With Words—From Manipulation of Measurements to Manipulation of Perceptions." In *Computing with Words*, edited by P. Wang, 35–68. New York: Wiley.

Zottoli, M., S. Laurita, and F. Monteleone. 2017. "Contestina: A Visibly Understandable Path toward More Effective Data Dissemination." In Proceedings of the New Techniques and Technologies in Statistics (NTTS 2017), March 14–16, 2017. Brussels. Available at: https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_151.html (accessed May 2017).

Received June 2017
Revised September 2018
Accepted October 2018