A FUZZY METRIC FOR CURRENCY IN THE CONTEXT OF BIG DATA

Completed Research

Heinrich, Bernd, University of Regensburg, Regensburg, Germany, Bernd.Heinrich@ur.de

Hristova, Diana, University of Regensburg, Regensburg, Germany, Diana.Hristova@ur.de

Abstract

Nowadays, companies rely more than ever on stored data to support decision making. However, outdated data may result in wrong decisions and economic losses. Thus, measuring data currency is extremely important. Existing metrics for currency either assume that their input parameters are given, or estimate them statistically, which may not always be possible in applications and especially in the context of big data. To address this issue, we propose a metric for currency based on expert estimations. The metric is modelled as a fuzzy inference system, which consists of a set of parallel IF-THEN rules with linguistic variables as inputs and output. It thus allows for a well-founded quantification of expert estimations and the consideration of both subjective and objective data. In addition to presenting our metric, we provide methods for estimating its input parameters (age of the considered attribute value and its decline rate). Furthermore, we demonstrate how the fuzzy inference system and thus the metric can be initialised and applied. The presented approach serves as a first step in modelling expert estimations as input to data quality metrics in a well-defined and structured way.

Keywords: Data quality, Currency, Metric, Fuzzy inference system, Expert estimations, Big data

1 Introduction

Due to the rapid technological development many companies around the world store and analyse large volumes of data from heterogeneous sources almost in real-time to support decision making. This type of data is often referred to as big data. Big data is usually characterised according to the three "V"s: volume, velocity and variety, where "volume" stands for the increasing number of (Tera)Bytes of stored data, "velocity" describes the increasing speed with which data is generated and analysed and the "variety" in big data is due to the integration of data from multiple data sources in heterogeneous formats (McAfee and Brynjolfsson, 2012). Recently, IBM Institute for Business Value (2012) has added to the characteristics a fourth "V" – veracity, which stands for the uncertainty in big data due to poor data quality (DQ). Poor DQ leads in many cases to wrong decisions and thus economic losses. According to a survey, conducted by Forbes (2010), poor DQ costs the majority of the participants more than $5 million annually. In addition, according to another survey (Experian QAS, 2013) 91% of the respondents "…admit that budgets have been lasted over the last 12 months as a result of poor data quality." (p. 7). This illustrates the growing importance of DQ nowadays.

DQ is defined in the literature as a multi-dimensional concept consisting of a number of dimensions such as currency, accuracy, completeness, etc. (Wang and Strong, 1996). Among them currency of an attribute value[1] is defined as the correspondence between a previously correctly stored attribute value and its real-world counterpart, which may have changed since the storage (Redman 1996; Pipino et al., 2002; Heinrich and Klier, 2011). Note that some authors refer to this DQ dimension as timeliness (Wand and Wang, 1996; Ballou et al., 1998), freshness (Cho and Garcia-Molina, 2000) or staleness (Chayka et al., 2012). Currency is considered to be one of the most important DQ dimensions (Redman, 1996), because most organisations suffer from a significant amount of outdated data (Experian QAS, 2013). This is true especially for the context of big data, where the data volume strongly increases the time for analyzing it, leading to possibly outdated data (velocity) at the time the analytical results are available. Moreover, the variety of big data exaggerates the problem as different sources with different currency are aggregated. Thus, it is important that the currency of big data is measured before making decisions based on analytical results.

A number of authors have developed metrics for currency (Ballou et al., 1998; Cho and Garcia-Molina, 2000; Even and Shankaranarayanan, 2007; Heinrich et al., 2009; Heinrich and Klier, 2011; Li et al., 2012; Wechsler and Even 2012). The general idea of these approaches is to assess the currency of an attribute value by considering its (storage) age and decline rate (e.g. decline rate is defined as the average percentage of attribute values that become outdated within a period of time (Heinrich and Klier, 2011)). Some of them assume that these input parameters are known, which is not always the case in reality. Exceptions are, for example the works by Cho and Garcia-Molina (2000), Heinrich et al. (2009), and Heinrich and Klier (2011), who statistically estimate the decline rate and thus require reliable historical data. However, such data is also not always given in reality, for example, due to short history - especially in the context of big data - or expensive data acquisition. Thus, expert estimations present a very reasonable alternative to historical data since they are easier to obtain and do not require such a large sample. However, in order to deliver precise results, expert estimations should be modelled in a well-founded mathematical manner, which is not straightforward due to their subjectivity.

In the current paper we develop a fuzzy metric for currency as a first step in modelling expert estimations. Thereby, the age and the decline rate of an attribute value are modelled as linguistic

[1] By an attribute value we mean a value of the domain of an attribute, for instance the attribute value "single" of the attribute marital status.

variables. Linguistic variables are variables, which values are words or sentences (Zadeh, 1975) and have a sound mathematical foundation in fuzzy set theory (Aliev, 2013). The metric is defined as a fuzzy inference system (FIS), which consists of parallel IF-THEN rules with linguistic variables in the rule-antecedent (age and decline rate) and in the rule-consequent (currency). The advantage of a FIS is that it allows for a precise modelling of complex non-linear relationships (Cingolani and Alcalá-Fdez, 2013) based on expert knowledge in a transparent and natural way (Mendel, 1995; Cherkassky, 1998).

The paper is structured as follows. Section 2 provides an overview of existing metrics for currency and briefly presents the theoretical foundations of fuzzy set theory. In Section 3 the fuzzy metric is defined and the derivation of its input parameters and output value is discussed. In Section 4 the metric is evaluated. In Section 5 main conclusions are drawn and limitations and future developments proposed.

2 Literature review and background

2.1 Existing metrics for currency

In this subsection we concentrate on well-known, formally defined metrics, which are based on the idea of considering the (storage) age and decline rate of an attribute value to measure its currency. One of the first metrics for currency in the literature is the approach by Ballou et al. (1998), which is defined as a metric for timeliness, but the definition corresponds to the above definition of currency. According to it, currency of an attribute value is determined by its age (i.e. time since its creation), shelf life and an additional attribute value specific parameter. The shelf life of an attribute value is defined as "…the length of time during which the data in question remain valid" (p. 468) and is directly connected to the decline rate defined above. Ballou et al. (1998) assume that both the age and the shelf life of the value are known and that both an increase in the age and a decrease in the shelf life reduce currency. Finally, the authors assess currency on a continuous scale between zero and one "for comparison purposes" (p. 468).

A utility-based approach is proposed by Even and Shankaranarayanan (2007), who present two approaches for assessing currency. According to the first one, currency is determined by the age of an attribute value (defined as the time since the last update) and a decline factor (describing how utility declines with the increase of age), which should be estimated by experts. The second approach considers as input parameters the age of an attribute value, the age after which the attribute value is valueless (i.e. marginal age), and again a decline factor. In both cases, the age of the attribute value is assumed to be known and to have a negative effect on currency. Similar to Ballou et al. (1998), the metric provides results in the range between zero and one.

Heinrich et al. (2007, 2009) and Heinrich and Klier (2011) propose probability-based metrics for currency. The input parameters are the (storage) age and the decline rate, where the decline rate is statistically estimated from historical data. An increase in both parameters decreases currency. The metric results are interpreted as the probability that the stored value still corresponds to its real-world counterpart and are thus between zero and one. In contrast to the above methods, the metrics by Heinrich et al. (2007, 2009) and by Heinrich and Klier (2011) require the existence of reliable historical data, which, as mentioned in the introduction, is not always given in reality.

Li et al. (2012) present a metric for currency for pervasive applications, which considers the update dynamics of the data source in addition to the storage age (defined with respect to the last update) and the shelf life of an attribute value. The shelf life should be estimated by experts and the update interval is modelled as a random variable. Similar to Ballou et al. (1998), both a decrease in the shelf life and an increase in the storage age lead to lower currency. The metric takes values in the interval $[0,1]$.

Wechsler and Even (2012) propose a metric for accuracy, which however is defined as our definition of currency i.e. they address "…accuracies that are caused by failures to update data even when changes in the real-world entity require us to do so." (p. 1). Their metric is based on a Markov-Chain

model and currency is estimated with an exponential probability distribution leading to a very similar metric to that of Heinrich et al. (2007) and Heinrich and Klier (2011).

Probst and Görz (2013) apply the interpretation by Heinrich and Klier (2011) in the context of online social networks. They empirically demonstrate that lower age of the attribute value, longer shelf life of the attribute value (indicated by supplemental data), higher number of direct contacts of the user, and higher activity of the user all lead to an increase in currency of attribute values within a user's profile.

To sum up, existing approaches consider two main factors that negatively influence currency: the age and the decline rate of an attribute value, where the age is interpreted either as natural age (Ballou et al., 1998; Heinrich et al., 2009; Li et al., 2012; Probst and Görz, 2013) or as storage age (Even and Shankaranarayanan, 2007; Heinrich et al., 2007; Heinrich and Klier, 2011, Wechsler and Even, 2012). Moreover, the above metrics result in a currency between zero and one. However, many existing approaches either do not explicate how their input parameters are to be acquired/estimated or require a sound statistical basis, which is not always given in reality and especially in the context of big data. For example, statistically determining the decline rate of sensor data such as data from temperature sensors, smart meters, or car sensors is hardly possible, since very detailed historical data is required (i.e. the temperature of a device depends on its type, series, age, and operating history). The same holds for census data or social networks' data where precisely determining the decline rate requires considering additional factors (cf. Probst and Görz 2013). Experts, on the other hand, may estimate, based on their experience, the decline rate at a very detailed level without the need of historical data. This becomes even more obvious for unstructured data such as patients' clinical profiles, news, videos, etc., where human interpretation is required to extract the important information. Existing metrics for currency cannot be applied to these cases, even if historical data is available. Experts, on the other hand, can determine the decline rate of a piece of text, based on their experience. For example, a piece of news about the movements on the stock market on a given day will have much higher decline rate than one providing information about the currently elected president of the United States, which could change after four years and definitely would after eight years. In this paper we thus propose an alternative approach by developing a metric for currency based on fuzzy set theory which is initialised with expert estimations. Similar to the presented metrics, the input parameters for it are age and decline rate, where both have a negative effect on currency. Moreover, the metric delivers results between zero and one. In the next subsection we briefly present the theoretical basics of fuzzy set theory.

2.2 Fuzzy set theory

In classical (crisp) set theory an element x either belongs to a set N or it does not i.e. it has a degree of membership in $\{0,1\}$ assigned by its characteristic function. For example, if N represents the set of attribute values with an age of two years, then a two-year-old attribute value would have a membership of one and all other attribute values would have a membership of zero. Such sets are called "crisp" sets. However, in many cases, especially in expert estimations, it is not possible to define clear boundaries of a set (Wang, 1996) and a degree of membership in $[0,1]$ is required. For example, if A is the set of attribute values with an age of *approximately* five years, then a five-year-old attribute value would have a membership of one, but an attribute value with an age of five years and one month would also belong to A with positive membership lower than one (e.g. 0.9). Such "fuzzy" sets are modelled as follows:

*Definition 1 (Fuzzy Set): Given a collection of objects X with members $x \in X$, a **fuzzy set** A in X is a set of ordered pairs $A = \{(x, \mu_A(x)) | x \in X\}$, where $\mu_A: X \rightarrow [0,1]$ is called the **membership function** and X is called the **universe of discourse** of the fuzzy set.*

The membership function can be either a discrete set over the universe of discourse or defined by a parametric form. The most common parameterized membership functions are the triangular, the

trapezoidal, the L-function, the R-function (Straccia, 2014), the Gaussian and the bell-shaped ones (Jang et al., 1997). Similar to crisp sets, it is necessary to determine the union, intersection, and complement of fuzzy sets, which are often defined by a particular s-norm, t-norm, and negation, respectively. Definition 2 presents the so called *standard fuzzy set operations* (Straccia, 2014), which were initially proposed by Zadeh (1965):

Definition 2 (Standard fuzzy set operations)

Let $A = \{(x, \mu_A(x))|x \in X\}$ and $B = \{(x, \mu_B(x))|x \in X\}$ be two fuzzy sets. Then:

$(Union, A\ OR\ B)$: $A \cup B = \{(x, \mu_{A\cup B}(x))|x \in X, \mu_{A\cup B}(x) = max\{\mu_A(x), \mu_B(x)\}\}$

$(Intersection, A\ AND\ B)$: $A \cap B = \{(x, \mu_{A\cap B}(x))|x \in X, \mu_{A\cap B}(x) = min\{\mu_A(x), \mu_B(x)\}\}$

$(Complement, NOT\ A)$: $\neg A = \{(x, \mu_{\neg A}(x))|x \in X, \mu_{\neg A}(x) = \{1 - \mu_A(x)\}\}$

To facilitate the quantitative modelling of expert estimations, Zadeh (1975a) introduced the concept of a linguistic variable, which values are words or sentences rather than numbers (Aliev, 2013).

Definition 3 (Linguistic Variable): A linguistic variable is characterised by a name n, a term set $T(n)$, which elements are called linguistic terms, and a universe of discourse X. The linguistic terms are represented by fuzzy sets in X with the corresponding membership functions.

Based on these definitions, in the next section we present our fuzzy metric for currency.

3 A fuzzy metric for currency

The main idea behind the fuzzy metric for currency, based on expert estimations, is that currency of a stored attribute value is influenced by both its age and decline rate in a negative way (cf. Subsection 2.1). In order to provide experts with a natural way of quantifying their estimations, we define age, decline rate and currency as linguistic variables. The relationship between the input parameters and output value currency is modelled by a FIS, consisting of a set of rules. The linguistic variables in the rule-antecedent are age and decline rate and the linguistic variable in the rule-consequent is currency. Age and decline rate are connected with the AND operation (cf. Definition 2). A simple example for a rule is:

Rule 1: $IF\ age\ is\ old\ AND\ decline\ rate\ is\ quick, THEN\ Currency\ is\ outdated$

We chose to model the metric as a FIS, because it is able to incorporate both subjective and objective estimations in a "unified mathematical manner" (Mendel, 1995, p. 1). This is important in our context, because, while the metric for currency is initialised with expert estimations, some of the inputs to the metric for a given attribute value (e.g. age in years) may be objective values. Moreover, as opposed to other methods such as Bayesian analysis with prior subjective information (Berger, 2010), this method gives experts the opportunity to make their estimation in a natural way represented by the linguistic terms. The same holds for modelling the relationship between the input variables (i.e. age and decline rate) and the output variable (i.e. currency), which does not take place in a formal fashion, but rather by stating rules such as *Rule* 1.

We begin with the definition of the linguistic variables. In the literature there are a number of recommendations, which the linguistic variables of a FIS should satisfy, presented in the following.

> **Recommendation 1 (Number of linguistic terms in the term set):** The number of linguistic terms is usually between three and seven (Zadeh, 1994), where most authors recommend at least two value signs (e.g. positive and negative) and a neutral value in the middle, which represents a normal situation (Lee, 1990; Driankov, 1996). Moreover, Adamy (2005) states that humans are able to differentiate at most nine levels of a certain subject.

> **Recommendation 2 (Cross point of two neighbour linguistic terms):** Every two neighbour linguistic terms should cross only once at the level of membership of 0.5 (Driankov, 1996). As a result all elements in the universe of discourse belong to at least one of the fuzzy sets with positive membership. Thus, discontinuities in the output are avoided (Driankov, 1996). Moreover, a value of 0.5 "provides for significantly less overshoot, faster rise-time and less undershoot." (Driankov, 1996, p. 120).

In addition, only the variables in the rule-antecedent should satisfy the following recommendation:

> **Recommendation 3 (Condition width):** For any two neighbour linguistic terms in the rule-antecedent, the left width of the right membership function should equal the right width of the left one and both should equal the distance between the peaks of the two membership functions (Driankov, 1996; Zimmermann, 2001). This condition should be satisfied for a smooth change in the system's output.

The linguistic variable "age" is interpreted as the time period in years between creating the attribute value in the real-world and assessing its currency. The universe of discourse of "age" is defined to be $[0, M]$ years, where M stands for the maximum number of years and thus allows for flexibility when constructing the FIS, based on the characteristics of the considered attribute value. The number of linguistic terms in the term set of "age" should satisfy **Recommendation 1**. Moreover, it influences the maximal possible number of rules of the FIS (Zimmermann, 2001) and thus the precision of the system. However, too many rules may lead to a computationally inefficient system. We thus define initially the term set of "age" as $\{young, not\ young\ and\ not\ old, old\}$ to keep the number of terms as low as possible[2] and to avoid too many rules (cf. Table 1). The linguistic term $\{not\ young\ and\ not\ old\}$ can be derived from the linguistic terms $\{young, old\}$ by applying Definition 2, which automatically fulfils **Recommendation 2** and **Recommendation 3**. Since with an increasing age attribute values become older, the membership function of $\{young\}$ should be decreasing with increasing age, while the membership function of $\{old\}$ should be increasing with increasing age. Moreover, since an attribute value with an age of zero years should be young with certainty, the membership function of $\{young\}$ should take a value of one at zero years. Similarly, the membership function of $\{old\}$ should take the value of one at M years. Finally, both membership functions must be defined on a bounded domain i.e. with the point above which a value is definitely not young anymore and the one below which it is definitely not old anymore, respectively.

Definition 3 (Age): The linguistic variable "age" is defined with the term set $T(age) = \{young, not\ young\ and\ not\ old, old\}$ *over the universe of discourse* $X = [0, M]$ *years and the linguistic terms:*

$$young = \{(x, \mu_{young}(x,b)) | \mu_{young}(0,b) = 1, \mu_{young}(b,b) = 0, \frac{\partial \mu_{young}(x,b)}{\partial x} \leq 0, x \in [0,b]\},$$

$$old = \{(x, \mu_{old}(x,c)) | \mu_{old}(M,c) = 1, \mu_{old}(c,c) = 0, \frac{\partial \mu_{old}(x,c)}{\partial x} \geq 0, x\ [c,M]\},$$

$$not\ young\ and\ not\ old = \{(x, \mu_{not\ young\ and\ not\ old}(x)) | \mu_{not\ young\ and\ not\ old}(x) = min(\big(1 - \mu_{old}(x,c)\big), \big(1 - \mu_{young}(x,b)\big), x \in [0,M]\big\}.$$

In order to define the linguistic variable "decline rate" we follow a similar approach as with "age". "Decline rate" is interpreted as the average percentage of attribute values that become outdated per year. We chose this definition, because experts feel more comfortable estimating percentages than pure numbers without any interpretation (Zimmermann, 2001). As opposed to "age", "decline rate"

[2] Note that the number of linguistic terms can be increased if necessary as long as Recommendations 1-3 are satisfied.

has a natural upper bound, thus the universe of discourse is $[0,100]\%$ per year. The term set is defined as $\{quick, not\ quick\ and\ not\ slow, slow\}$ (once again, the number of linguistic terms can be increased if necessary) and for the definition of the membership functions of $\{quick, slow\}$ we follow the same approach as with "age". Similar to "age", the membership functions must be defined on a bounded domain i.e. with the point above which the decline rate is not slow anymore and the one from which it is quick.

Definition 4 (Decline rate): The linguistic variable "decline rate" is defined with the term set $T(decline\ rate) = \{quick, not\ quick\ and\ not\ slow, slow\}$ *over the universe of discourse* $X = [0,100]\%$ *per year and the linguistic terms*

$$slow = \{(x, \mu_{slow}(x,d))|\mu_{slow}(0,d) = 1, \mu_{slow}(d,d) = 0, \frac{\partial \mu_{slow}(x,d)}{\partial x} \leq 0, x \in [0,d]\%\}$$

$$quick = \{(x, \mu_{quick}(x,e))|\mu_{quick}(100,e) = 1, \mu_{quick}(e,e) = 0, \frac{\partial \mu_{quick}(x,e)}{\partial x} \geq 0\ \forall x \in [e,100]\%\}$$

$$not\ quick\ and\ not\ slow = \{(x, \mu_{not\ quick\ and\ not\ slow}(x))|\mu_{not\ quick\ and\ not\ slow}(x) = min(\big(1 - \mu_{slow}(x,d)\big), \big(1 - \mu_{quick}(x,e)\big), x \in [0,100]\%\}.$$

Finally, the linguistic variable "currency" should be defined. However, since this is the output of the system, the definition of its term set depends on the structure of the rule base, which determines the number of necessary output linguistic terms. Thus, before we define "currency", we first discuss the derivation of the rule base.

In order to derive the rule base, we use a relational matrix (Mendel, 1995; Zimmermann, 2001; Adamy, 2005), which guarantees that all possible combinations of the input linguistic terms are considered and thus that for each combination of values for age and decline rate the currency of an attribute value can be determined. A relational matrix is a matrix, where the column and row names consist of the linguistic terms of the two input variables and the entries are the linguistic terms of the output variable. The two variables in the rule-antecedent are connected with an "AND" operation in our case (cf. Rule 1). For each combination of the input linguistic terms an integer value for the output linguistic term between -4 (outdated) and 4 (up-to-date) is written. The values should satisfy the assumption that both higher age and higher decline rate reduce currency. Table 1 presents an example for a rule base, where an increase in the age and an increase in the decline rate are assumed to have the same effect on currency. Rule 1, for example, can be derived from it.

Decline rate / Age	quick	not quick and not slow	slow
old	-4	-2	0
not young and not old	-2	0	2
young	0	2	4

Table 1. An example for a rule base

Let $Z = \{z_1, .. z_k\} \subseteq \{-4,-3,-2,-1,0,1,2,3,4\}, 3 \leq k \leq 7, z_i \neq z_j, i,j \in \{1,\dots,k\}$ be the set of different integers used in the matrix after applying the relational matrix method. Then $V = \{v_1, .. v_k\}, 3 \leq k \leq 7, v_i \neq v_j, i,j \in \{1,\dots,k\}$ represents the set of linguistic terms for currency derived with the relational matrix method, where $\forall i \in \{1,\dots,k\}\ z_i$ is mapped to v_i. For example, if $Z = \{-4,0,4\}$, then $V = \{outdated, neutral, up-to-date\}$. Note that the restriction on the number of linguistic terms (k) is due to **Recommendation 1**. Thus the definition of the term set of currency should satisfy the properties of the set Z (e.g. uniformly distributed terms, cf. Straccia 2014) and **Recommendation 2**.

The exact form of Table 1 and thus the definition of the sets $Z$and V depends strongly on the particular attribute and is determined by experts. For example, for some attributes a change in the decline rate from $not\ quick\ and\ not\ slow$ to $quick$ has a stronger influence on the currency than a change in the age from $not\ young\ and\ not\ old$ to old. Existing metrics for currency can only model this attribute-specific dependency if enough historical data is available, which is rarely the case for big data. Estimating the rule base by experts is thus another advantage of the fuzzy metric as opposed to existing metrics.

The linguistic variable "currency" is interpreted as the correspondence between previously correctly stored attribute value and its real-world counterpart, which may have changed since the instant of storage. In accordance with existing literature (cf. Subsection 2.1), the values of the metric (i.e. the universe of discourse) are normalised to $[0,1]$, where zero stands for a completely outdated attribute value and a currency of one means that the stored attribute value still corresponds to the real-world attribute value.

Definition 5 (Currency): Let the sets Z and V be as described above. The linguistic variable "currency" is defined with the term set $T(currency) = V$ over the universe of discourse $X = [0,1]$. The membership functions of the linguistic terms satisfy the characteristics of the set Z and ***Recommendation 2****.*

This completes the definition of the linguistic variables and the description of the derivation of the rule base for the FIS, which we call the *main FIS* from now on. In the next two subsections we discuss the estimation of the input parameters and the determination of the output value.

3.1 Derivation of the input parameters of the fuzzy metric

In order to determine the currency of attribute values, the *main* FIS requires as input parameters the age and decline rate of the attribute values. However, as discussed in Subsection 2.1, both parameters are in many cases not known. In the following we thus propose two possible solutions to this problem.

As discussed above, several authors (e.g. Even and Shankaranarayanan, 2007; Heinrich and Klier, 2011, Wechsler and Even, 2012) use the storage age, which is usually given as metadata, instead of the age of an attribute value to determine currency. Thus, we also adopt this idea by defining the variable of the rule-antecedent of the *main* FIS as "storage age" instead of "age" in Definition 3. The input parameter to the *main* FIS is then the storage age of the attribute value as an objective value. The assumption behind this approach is that the currency of an attribute value only depends on its history after storage and not on the development before that (i.e. "memorylessness"), which is not always satisfied in reality. An alternative is thus to estimate the age of an attribute value based on an *auxiliary* FIS, where "age" is the linguistic variable in the rule-consequent and the rule-antecedent consist of additional variables, which may include the storage age. The output of the system, which is a fuzzy set, will then be used as an input for the *main* FIS. For example, in professional networks such as Xing the time point when the user obtained the current professional position is not always given. Thus, it can be estimated based, for example, on the previous work experience, the education, the gender, and the age of the user, as well as the storage age of the attribute value (e.g. derived from the activity tab).

Similar to age, the decline rate of an attribute value is usually not known and is estimated in the literature either by experts or from statistical data (cf. Subsection 2.1). Thus, in the context of the fuzzy metric, the decline rate of an attribute value can be directly estimated as a fuzzy set by experts. A second possibility is, similar to age, to estimate the decline rate from additional information about the attribute value with an *auxiliary* FIS. For example, similar to age, in order to better estimate the decline rate of the current professional position in a Xing profile, experts may consider additional attributes such as the education or age of the user (cf. Probst and Görz, 2013). Similar to the age of an attribute value, the age of the user can also be modelled as a linguistic variable.

3.2 Derivation of the output value of the fuzzy metric

In order to determine the output value of the fuzzy metric (i.e. currency of an attribute value) from given input parameters, we use the approach illustrated in Figure 1. In this example the inputs of an individual attribute value to the fuzzy metric are a storage age of seven years (i.e. instead of age in Definition 3) and a decline rate, which was estimated by experts as a fuzzy set (methods for estimating fuzzy sets are presented in Section 4). The *main* FIS consists of two rules. First of all, for each rule, the membership of a storage age of seven years to the fuzzy set for the linguistic terms of storage age in the rule-antecedent is determined. In Figure 1 this corresponds to a membership of 0.25 to the fuzzy set "old" and a membership of 0.5 to the fuzzy set "not young and not old". Secondly, for each rule, the intersection of the fuzzy set describing the estimated decline rate with each of the fuzzy sets for the linguistic terms of decline rate in the rule-antecedent is determined and the maximum membership value is calculated. The result is a membership of 0.7 for the fuzzy set "quick" and a membership of 0.85 for the fuzzy set "not quick and not slow". Afterwards, for each rule-antecedent the AND operation (the minimum according to Definition 2) over the two membership values is applied resulting in a value of 0.25 for **Rule 1** and 0.5 for **Rule 2**, respectively. To derive the output of each rule, the fuzzy set in the consequent is clipped at the corresponding membership value (this is the commonly used Mamdani implication). Finally, the aggregated output of the system is obtained as the union (the maximum according to Definition 2) over the outputs of all rules. Since this is a fuzzy set, it needs for a better understanding to be defuzzified to a crisp value in a way that best represents the original fuzzy set (Jang et al., 1997). This is done most commonly by the application of the centroid method, which is analogous to the calculation of expected values in probability theory. For further defuzzification methods, see Driankov (1996) and Cingolani and Alcalá-Fdez (2013). The defuzzified set is the point z_D in Figure 1 and represents a currency of 0.49. This implies that all attribute values with a storage age of seven years and a decline rate estimated as the fuzzy set in Figure 1 will have a currency of 0.49 meaning that 49% from a set of stored attribute values correspondent to their real-world counterpart at the time of measurement. For example, if the attribute value for the current professional position in Xing is "postdoc", and the stored data consists of 100 postdocs, then 49 will be estimated to be still postdocs in the real-world at the time of measurement. This completes the definition of the fuzzy metric for currency.

3.3 Definition of the fuzzy metric from a fuzzy logic perspective

The definition of the fuzzy metric for currency as a FIS presented so far can be considered to be from an engineering perspective. In the literature (Hájek, 2001; Straccia, 2014), there exists another way to define it from a logic perspective. This definition is based on fuzzy logic, which can be defined as Łukasiewicz, Gödel, Product or Standard Fuzzy Logic (SFL) depending on the definitions of the fuzzy set operations for union, intersection, complement, and implication. The SFL is defined based on the operations in Definition 2 and an additional Kleene Dienes (K-D) implication operation defined as $a \Longrightarrow_S b \coloneqq \max(1-a, b)\,, a, b \in [0,1]$, where a, b are degrees of membership (Straccia, 2014).

In order to implement a FIS, the corresponding fuzzy logic is extended with fuzzy concrete domains. An example for a concrete domain is (*storage age* $\geq_{10}$), which represents the attribute values with storage age greater or equal to 10 years. Here *storage age* is a *feature name* and $\geq_{10}$ is a *hard constraint*. Analogously, a *fuzzy* concrete domain can be represented as (*storage age* $\widetilde{\mu_{slow}}(d)$) where $\widetilde{\mu_{slow}}(d)$ corresponds to the function $\mu_{slow}(x, d)$ in Definition 4 and is called a *soft constraint*.

To describe the FIS in Figure 1, we consider SFL, where $\wedge_S$ is interpreted as the AND-operation, $\vee_S$ as the OR-operation, both defined in Definition 2, and $\longrightarrow_S$ is interpreted as the K-D implication above. In addition, we replace $\longrightarrow_S$ with $\wedge_S$, which is called the Mamdani implication in the literature (cf. Subsection 3.2). Thus, the rules describing the FIS in Figure 1 are represented as *fuzzy statements* as follows:

$Rule\ 1 \leftrightarrow ((storage\ age\ \widetilde{\mu_{old}}) \wedge_S (decline\ rate\ \widetilde{\mu_{quick}}) \wedge_S (currency\ \widetilde{\mu_{outdated}})$

$Rule\ 2 \leftrightarrow ((storage\ age\ \widetilde{\mu_{not\ young\ and\ not\ old}}) \wedge_S (decline\ rate\ \widetilde{\mu_{not\ quick\ and\ not\ slow}}) \wedge_S (currency\ \widetilde{\mu_{neutral}})$

where *storage age*, *decline rate* and *currency* are feature names and the soft constraints are defined as in Figure 1. The input values in Figure 1 can also be described by (fuzzy) concrete domains. The input value for storage age is presented as ($storage\ age =_7$) and the input value for the decline rate as ($decline\ rate\ \widetilde{\mu_{decline}}$), where $\widetilde{\mu_{decline}}$ represents the corresponding trapezoidal membership function.

Next, the results from the rules are aggregated based on the maximum aggregation operator represented by $\vee_S$, which implies that the final output of the system is given by $Mamd = (Rule\ 1 \vee_S Rule\ 2)$ under the inputs described above (Hájek, 2001; Straccia, 2014). Finally, in order to defuzzify the output of the system, the maximum degree of satisfiability of $Mamd$ under the inputs described above can be determined (Straccia, 2014), which, however, will result in a different defuzzification method than the centroid method. The centroid method is not discussed in the fuzzy logic literature since it generally does not focus on the defuzzification part (Hájek, 2001).

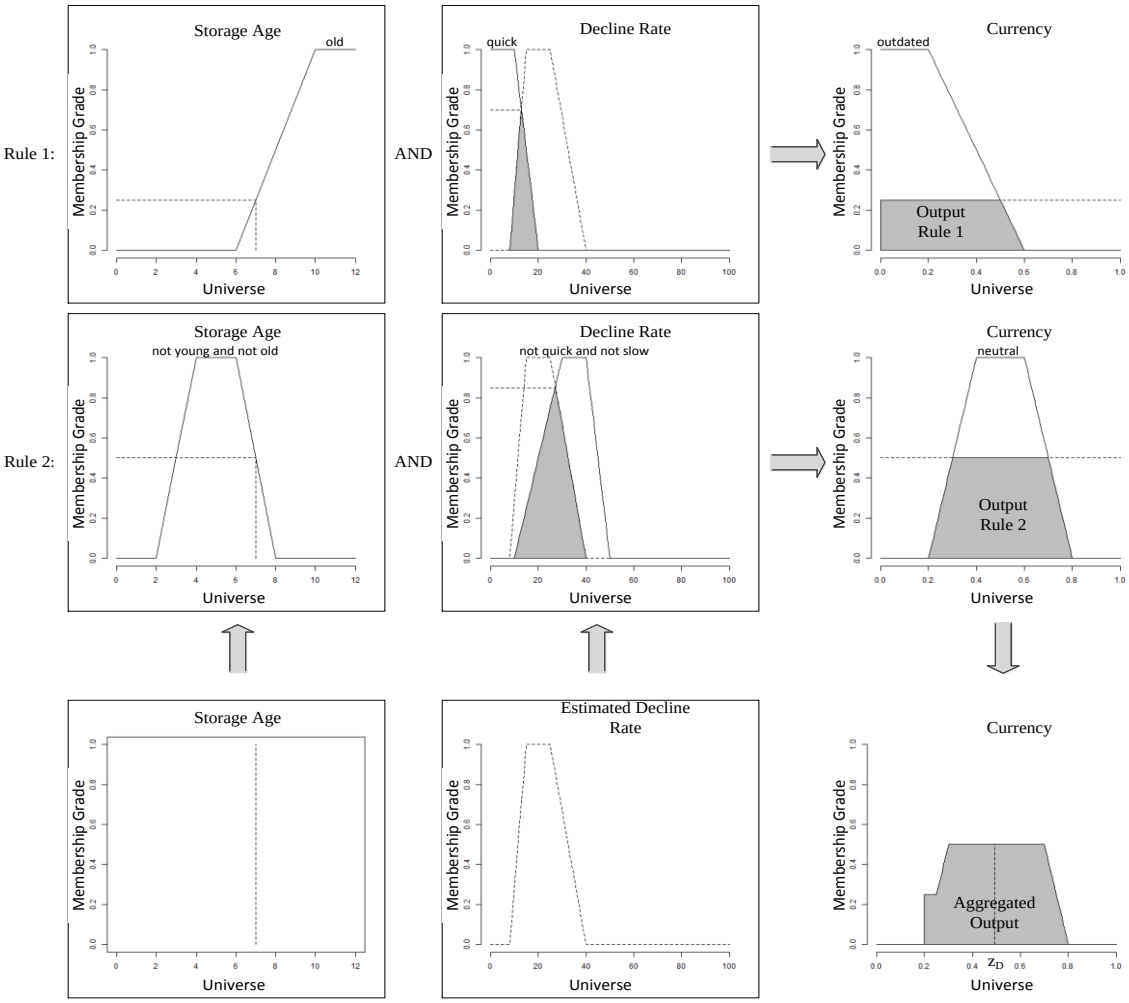

Figure 1. Example for estimating the output value of the fuzzy metric

As we see, the mathematically well-founded fuzzy logic perspective results in the same model as the engineering perspective without the centroid method in the defuzzification step. However, according

to the fuzzy logic perspective, the Mamdani implication does not satisfy the boundary condition axiom for an implication operation (Hájek, 2001; Straccia, 2014). It is an AND-operation in the above rules, which implies that the above rules are interpreted so that both the rule-antecedent and the rule-consequent should be satisfied. This does not change the validity of the approach, but its interpretation, which is extremely important for the acceptance of the metric in practice due to its simplicity. Alternative interpretations of the implication operation are the Łukasiewicz, Gödel, Product implications and the K-D implication defined above (Straccia, 2014), which are much more complex. In addition, the aggregation operator represented by $\vee_S$ is an upper bound for all other possible aggregation operators such as the weighted sum, where different inputs are given different weights. Thus, currency determined based on it represents an optimistic view of the correspondence between the stored attribute value and its real-world counterpart. Other aggregation operators (cf. Straccia 2014) would deliver a different interpretation. For a further discussion on the modelling of FIS from a fuzzy logic perspective refer to Hájek (2001), Aliev (2013) and Straccia (2014).

4 Evaluation

In this section we evaluate the presented metric by initialising and applying it to the attribute value "single" (marital status), which was chosen for a number of reasons. The first reason is that it can be found in many big datasets such as census data or social networks' data (e.g. in a CRM campaign based on the hobbies of Facebook users), where its currency cannot be precisely estimated from historical data. In particular, the decline rate of this attribute value is strongly dependent on other factors such as age, gender and nationality. Therefore, asking experts to estimate the decline rate with the help of the *auxiliary* FIS discussed in Subsection 3.1 (with the age of a person as additional information) is expected to provide better results than if additional information is not considered. The second reason is that it is enough for the initialization of the metric if experts in the field of DQ are interviewed and experts from other fields are not required as would be the case, for example for the temperature of a medical device measured by a sensor. This results in a smaller "expert" bias, which is the bias stemming from the choice of the experts (e.g. due to their experience in the field). The third reason for choosing this attribute value is that, due to its intuitive interpretation, the results from the expert estimations can be easily verified by non-experts (cf. Figure 2). Therefore, the metric developers can determine whether probably unexpected results (i.e. the output of the model) are due to the experts' estimations (i.e. the input to the model) or to the design of the metric (i.e. the model), which is very important for the sound development and evaluation of a new model. Finally, this attribute value is a standard example in the literature for measuring currency (Heinrich et al., 2009, Wechsler and Even, 2012) and thus allows for a good comparison between our metric and existing metrics in future research.

In order to initialise the metric, the linguistic variables for the *main* FIS and for the *auxiliary* FIS need to be derived as well as the input parameter for the decline rate to the *main* FIS as a fuzzy set. The input parameter of the age to the *main* FIS is represented by the storage age, which is a crisp value and is known. Moreover, the rule base in Table 1 is considered, based on which the linguistic variable currency is derived. In the literature there are a number of approaches for membership functions' elicitation, which can be grouped in the following categories: direct rating, pooling, reverse rating, membership exemplification, pairwise comparison and approaches based on a given set of data (Chameau and Santamarina, 1987; Santamarina and Salvendy, 1991; Turksen, 1991; Bilgic and Turksen, 2000). In the current paper we use a mixture between point estimation and interval estimation as they best fit the presented setting. In point estimation experts are asked for a value, while in interval estimation experts are asked for a range, which best represents a certain fuzzy set. Both methods are "easy to respond" (Santamarina and Salvendy, 1991, p. 30) and the derivation of the membership functions is relatively straightforward. The advantage of interval estimation as opposed to point estimation is that experts have more freedom in their estimation and thus the answers are more stable.

For each input variable of the *main* FIS two membership functions are required that satisfy the conditions in Definitions 3 and 4, respectively which means that they are monotonous with a fixed maximum or minimum value. This implies that here interval estimation cannot be applied, since it does not necessarily result in monotonous membership functions. Thus point estimation is used for the linguistic variables in the *main* FIS. In contrast, the estimation of the input parameter for the decline rate in the *main* FIS does not need to satisfy any conditions for the membership function, because it does not have the interpretation of a certain linguistic term (cf. Figure 1). Thus interval estimation is used. The input variable to the *auxiliary* FIS is "age of a person" with the linguistic terms $\{young, not\ young\ and\ not\ old, old\}$ satisfying the recommendations in Section 3 and naturally only people older than 18 are considered. In this case the membership functions of the fuzzy sets $\{young\}$ and $\{old\}$ are bounded from below and above, respectively and must be monotonous. Thus, similar to the *main* FIS, point estimation is used.

The results of both point and interval estimation are empirical, stepwise membership function. However, for a more stable and efficient system, continuous parametric membership functions are recommended. A number of authors have proposed ways for constructing parametric membership functions from stepwise ones (Chen and Otto, 1995; Chang et al., 2000; Medaglia et al., 2002). We decided to apply the approach by Chang et al. (2000), who derive continuous membership functions of a certain parametric form with the conjugate search method. The advantage of this approach is, as opposed to the others, that it allows generating membership functions of parametric form and its implementation is quite efficient. In order to derive the membership functions, we developed a questionnaire, which was then sent to experts (practitioners) from the field of DQ.

The process of questionnaire development began with a pre-test. The questions were ordered according to the guidelines in the literature (Marsden and Wright, 2010). In addition, the experts were also asked to provide some personal information, as well as their expertise to control for these factors. In order to keep the experts motivated, the initial questionnaire contained only two pages. To assure objectivity all of the guidelines and explanations were given in the questionnaire. The initial questionnaire was continually improved by asking each expert about the way s/he estimated the values and also for feedback regarding possible improvements. This addresses the content validity of the questionnaire (Litwin, 1995). The final questionnaire consisted of the following five *main* questions with total response time of about 10 minutes (translated from German).

1. Which value best represents the set of slow/quick decline rate? (point estimation, *main FIS*)
2. Which value best represents the set of young/old storage age? (point estimation, *main FIS*)
3. Please give the minimum and the maximum value for the number of people among 100 full aged, people, who were single one year ago and are married now. (interval estimation, *main* FIS)
4. Which value best represents the set of young/middle-aged/old people? (point estimation, *auxiliary FIS*)
5. Please give the minimum and the maximum value for the number of young/middle-aged/old people among 100 full aged, young/middle-aged/old people, who were single one year ago and are married now. (interval estimation, *auxiliary* FIS)

A full study was conducted with this questionnaire and 9 experts with an average experience of 12 years in the field of DQ. Based on the derived membership functions, currency for different storage age and decline rates (as linguistic terms) is determined. The results show that currency decreases with increasing storage age and decline rate, and that the decrease rate is relatively high at the beginning and reduces with the storage age. The reason for this is that the membership function for "old" storage age estimated by the experts begins relatively early.

Figure 2 presents the membership function of the decline rate without considering additional information and the membership functions of the output linguistic variable of the *auxiliary* FIS, where

the decline rate is estimated according to the age of the person. We can see that the decline rate of young people and that of old ones is more concentrated at a given point, as opposed to the group of not young and not old people. Moreover, the membership function describing the fuzzy set for the decline rate of old people is characterised by a very "fat tail", which illustrates the uncertainty of the experts regarding their estimation. The difference in the membership functions results in different currency for the corresponding age groups. This demonstrates the strength of our approach as opposed to existing metrics, because it allows considering additional information without requiring historical data, which, as discussed above, is a significant advantage in the context of big data.

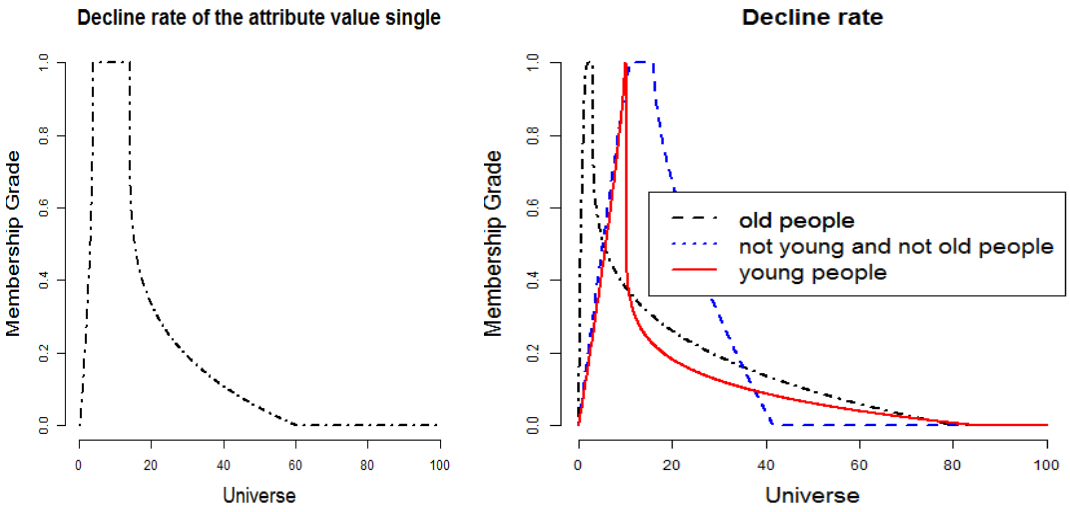

Figure 2. Decline rate of the attribute value single without and with considering additional data

5 Conclusion

In the current paper we present a fuzzy metric for currency, which is initialised with expert estimations in a mathematically well-founded way. The metric is modelled as a FIS with the linguistic variables "age" and "decline rate" as its input parameters and the linguistic variable "currency" as its output value, where currency decreases with both increasing age and decline rate and takes a value between zero and one. In addition, we present methods for deriving the input parameters of the FIS if they are not known. One such approach is the use of an *auxiliary* FIS to derive both the age and the decline rate. Moreover, the derivation of currency as an output value is discussed in detail. Finally, we evaluate our method by first discussing the derivation of the different membership functions according to the corresponding definitions of the linguistic variables and then describing the process of questionnaire development and application. The results show that the decline rate of the attribute value "single" changes depending on the age of the person and consequently that the attribute value has different currency for the same storage age but different ages of the people.

The main advantage of the presented metric is that it does not require historical data as opposed to existing approaches in the literature, which makes it very appropriate for the field of big data. Thus, future research should concentrate on comparing the performance of the fuzzy metric with other metrics for currency, especially as the amount of historical data decreases. Moreover, in the current paper we have considered SFL with Mamdani implication and the maximum defuzzification operator. Future research may test the sensitivity of the metric by defining it with other types of fuzzy logic. In addition, the sensitivity of the metric to the number of linguistic terms should be evaluated. Finally, the application to a different attribute is of special interest, especially if this is a big data case with unstructured data, where the subjective information extraction plays such an important role.

References

Adamy, J. (2005). Fuzzy Logic, Neurol Networks and Evolutionary Algorithms (in German). Shaker Verlag.

Aliev, R. A. (2013). Fundamentals of the fuzzy logic-based generalized theory of decisions. Springer.

Ballou, D., Wang, R., Pazer, H. and Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. Management Science, 44(4), 462-484.

Berger, J. O. (2010). Statistical decision theory and Bayesian analysis, 2nd ed., Springer.

Bilgic, T. and Turksen, I. (2000). Measurement of membership functions: Theoretical and empirical work. In Fundamentals of fuzzy sets, 95-227, Springer.

Chameau, J.-L. and Santamarina, J. C. (1987). Membership functions I: Comparing methods of measurement. International Journal of Approximate Reasoning, 1(3), 287-301.

Chang, P.-T., Huang, L.-C. and Lin, H.-J. (2000). The fuzzy Delphi method via fuzzy statistics and membership function fitting and an application to the human resources. Fuzzy Sets and Systems, 112(3), 511-520.

Chayka, O., Palpanas, T. and Bouquet, P. (2012). Defining and Measuring Data-Driven Quality Dimension of Staleness. Trento : University of Trento, Technical Report # DISI-12-016.

Chen, J. E. and Otto, K. N. (1995). Constructing membership functions using interpolation and measurement theory. Fuzzy Sets and Systems, 73(3), 313-327.

Cherkassky, V. (1998). Fuzzy inference systems: a critical review. In Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications, 177-197, Springer.

Cho, J. and Garcia-Molina, H. (2000). Synchronizing a database to improve freshness. SIGMOD Rec., 29(2), 117-128.

Cingolani, P. and Alcalá-Fdez, J. (2013). jFuzzyLogic: a Java Library to Design Fuzzy Logic Controllers According to the Standard for Fuzzy Control Programming. International Journal of Computational Intelligence Systems, 6(sup1), 61-75.

Driankov, D. H. (1996). An Introduction to Fuzzy Control. Springer.

Even, A. and Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. ACM SIGMIS Database, 38(2), 75-93.

Experian QAS. (2013). The Data Advantage: How accuracy creates opportunity.

Forbes Insights. (2010). Managing Information in the Enterprise: Perspectives for Business Leaders.

Hájek, P. (2001). Metamathematics of fuzzy logic. Kluwer.

Heinrich, B., Kaiser, M. and Klier, M. (2007). How to measure data quality? – a metric based approach. Proceedings of the 28th ICIS 2007, Paper 108.

Heinrich, B., Kaiser, M. and Klier, M. (2009). A Procedure to Develop Metrics for Currency and its Application in CRM. Journal of Data and Information Quality (JDIQ), 1(1), 5.

Heinrich, B. and Klier, M. (2011). Assessing data currency-a probabilistic approach. Journal of Information Science, 37(1), 86-100.

IBM Institute for Business Value. (2012). Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data.

Jang, J.-S. R., Sun, C.-T. and Mizutani, E. (1997). Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Inc.

Lee, C.-C. (1990). Fuzzy logic in control systems: fuzzy logic controller. I. IEEE Transactions on Systems, Man and Cybernetics, 20(2), 404-418.

Litwin, M. S. (1995). How to measure survey reliability and validity. Sage.

Marsden, P. V. and Wright, J. D. (2010). Handbook of survey research. Emerald Group Publishing.

McAfee, A. and Brynjolfsson, B. (2012). Big data: The management revolution. Harvard Business Review.

Medaglia, A. L., Fang, S.-C., Nuttle, H. L. and Wilson, J. R. (2002). An efficient and flexible mechanism for constructing membership functions. European Journal of Operational Research, 139(1), 84-95.

Mendel, J. M. (1995). Fuzzy logic systems for engineering: a tutorial. In Proceedings of the IEEE, 83(3), 345-377.

Li, F., Nastic, S. and Dustdar, S. (2012). Data Quality Observation in Pervasive Environments. In 2012 IEEE 15th International Conference on Computational Science and Engineering (CSE), 602-609, Nicosia.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.

Probst, F. and Görz, Q. (2013). Data Quality Goes Social: What Drives Data Currency In Online Social Networks? In: Proceedings of the 21st European Conference on Information Systems, (ECIS), Utrecht, The Netherlands.

Redman, T. C. (1996). Data Quality for the Information Age. Boston, MA: Artech House.

Santamarina, C. and Salvendy, G. (1991). Fuzzy sets based knowledge systems and knowledge elicitation. Behaviour and Information Technology, 10(1), 23-40.

Straccia, Umberto (2014). Foundations of fuzzy logic and semantic Web languages. Boca Raton: CRC Press (Chapman & Hall/CRC studies in informatics series).

Turksen, I. (1991). Measurement of membership functions and their acquisition. Fuzzy sets and systems, 40(1), 5-38.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundation. Communications of the ACM, 39(11), 86-95.

Wang, L.-X. (1996). A Course in Fuzzy Systems. Prentice-Hall press, USA.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(2), 5-33.

Wechsler, A. and Even, A. (2012). Using a Markov-Chain Model for Assessing Accuracy Degradation and Developing Data Maintenance Policies. In Proceedings of the AMCIS 2012, Paper 3.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

Zadeh, L. (1975a). The concept of a linguistic variable and its application to approximate reasoning-I. Information Sciences, 8(3), 199-249.

Zadeh, L. A. (1994). Soft computing and fuzzy logic. Software, IEEE, 11(6), 48-56.

Zimmermann, H. J. (2001). Fuzzy set theory-and its applications. Springer.