Quantitative Approaches for Modeling Information Quality in Information Systems

DISSERTATION
zur Erlangung des Grades
eines Doktors der Wirtschaftswissenschaft (Dr. rer. pol.)
eingereicht an der
Fakultät für Wirtschaftswissenschaften
der Universität Regensburg

vorgelegt von
Diana Hristova, M.A.

Berichterstatter:
Prof. Dr. Bernd Heinrich
Prof. Dr. Günther Pernul

Regensburg, 26 Februar 2016

Tag der Disputation: 18.12.2015

To my parents, Ivanka and Rumen Hristovi.

Acknowledgements

I would like to express my deep gratitude to Prof. Dr. Bernd Heinrich for his great supervision during the last three years. He provided me with many interesting and innovative ideas, constructive feedback, and patiently guided me towards my aim. In addition, I would also like to thank Prof. Dr. Günther Pernul for his insightful comments and suggestions. Moreover, I would like to express my appreciation to my coauthor Prof. Dr. Guido Schryen for the many fruitful discussions and valuable advice. I would also like to offer my special thanks to the many students who worked on the topic during the last three years and thus helped me better clarify and understand the existing problems. In addition, the many anonymous reviewers who contributed to the improvement of the following papers should also be mentioned here. Last but not least, I would like to thank my family and friends for their ongoing support and encouragement.

Diana Hristova

Contents

Contents

List of Figures

List of Tables

1. Introduction

In this chapter, first a short motivation is provided followed by the addressed research questions. Then, the theoretical framework of the dissertation and the relevant methodologies for modeling uncertainty are described. Finally, the seven papers constituting the core of this dissertation together with the corresponding research questions and methodologies are presented.

1.1 Motivation

Nowadays and especially with the development of new technologies, companies increasingly use available information to support decisions. The types of decisions cover a wide range of problem domains. For example, information may be used to target customers in Customer Relationship Management campaigns (Kumar and Reinartz, 2012), for "predicting fraud or financial risks" (Henschen, 2013, p. 15), for revenue management (Wang *et al.*, 2014), or even for strategic decisions such as mergers and acquisitions (Witchalls, 2014). It may also be applied for sensor-based decisions such as activity recognition (Wu *et al.*, 2011) and augmented reality applications (Blum *et al.*, 2013), or for network-based decisions such as viral marketing (Hinz *et al.*, 2011) and crowdsourcing (Hobfeld *et al.*, 2014). Information can be used to support decisions by i) structuring the decision problem and ii) reducing the uncertainty faced by the decision maker (cf. Figure 1). This can happen 1) *directly* by incorporating raw information in the decisions or 2) *indirectly* by first analyzing raw information with knowledge discovery approaches. Thus, the *indirect* effects result from the application of knowledge discovery techniques from the field of Business Intelligence (BI) to derive new patterns from raw information. These patterns are then incorporated as extracted information (as opposed to raw information) in decision making (cf. Figure 1). In particular, the *indirect* effects differ from the *direct* effects in that for the former raw information is analyzed before being considered in decision making. For both effect types, information has a value[1] for the decision maker, expressed in the additional benefits due to better decisions.

To illustrate the idea, consider a Customer Relationship Management campaign where a company needs to decide which customers to target with a new product offer. First, in case i), the company must structure its decision problem, for example by determining the profit from accepting the offer and the probability of acceptance. Second, in case ii), it can use customers' attributes, such as age or gender, to refine the probability of acceptance for different customers. In both of these situations information can be used both 1) *directly* and 2) *indirectly*. For the *direct* effects, raw information would, for example, be the price of the corresponding product or a customer database with the stored personal attributes of each customer. For the *indirect* effects, based on BI techniques such as clustering, the company can categorize the customers in different groups according to the probability of accepting the offer. These categories can then be used as extracted information to reduce uncertainty in case ii). The value of information is expressed in the additional benefits of contacting particularly promising customers based on the (raw and extracted) information.

[1] The concept of value of information here is used to address both cases 1 i) and 1 ii). Later, the normative concept of the value of information will be used to address solely case 1 ii).

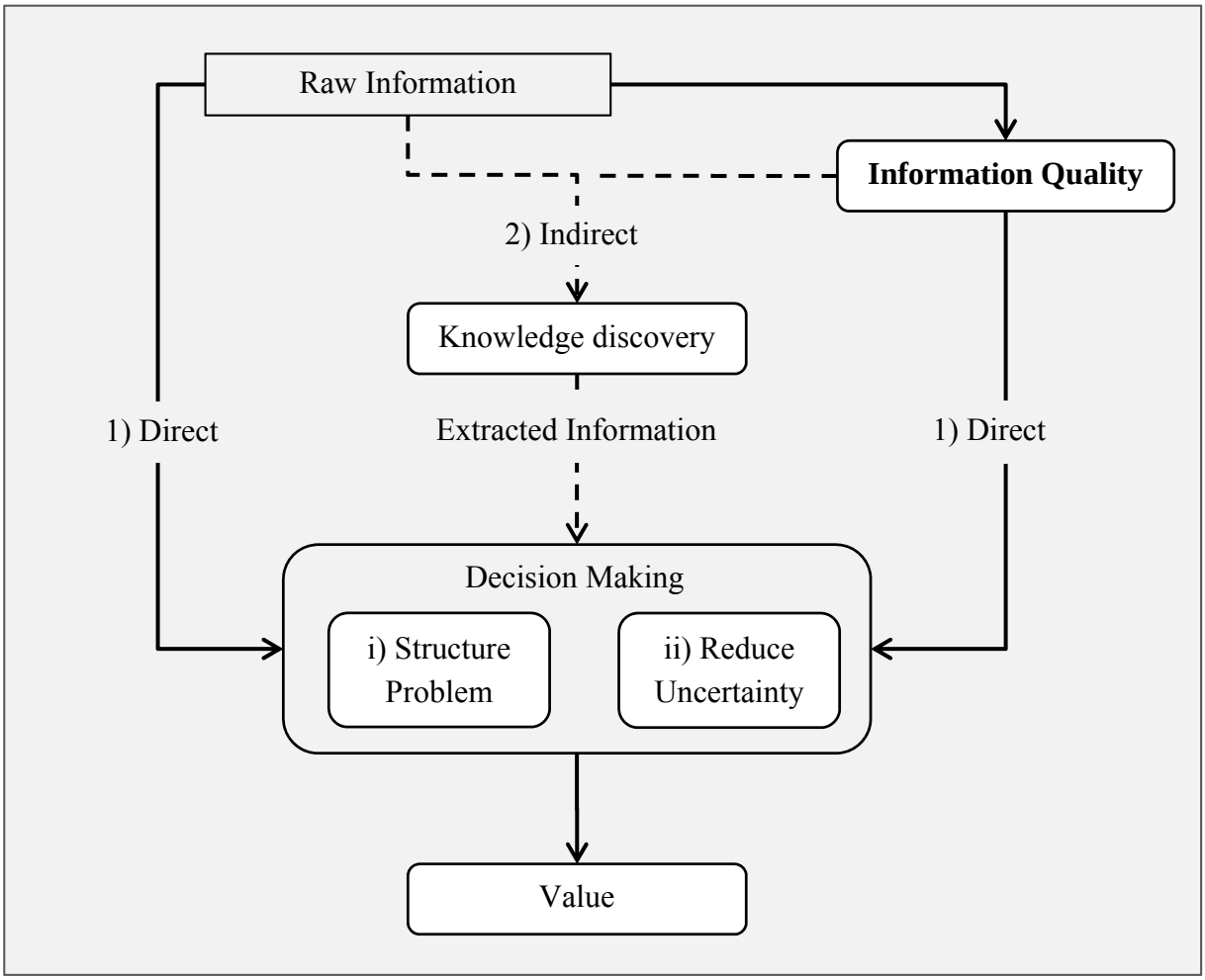


Figure 1: The Role of Information for Decision Making

However, there is one factor hurdling the value generation from information and this is low information quality. Information quality[2] is a multidimensional concept defined by different dimensions, such as accuracy, currency, and consistency (Wang and Strong, 1996), which represent different facets of the quality of data views (Orr, 1998; Heinrich and Klier, 2015). In a survey by Henschen (2013), 58% of the respondents name relevancy, timeliness, and reliability of their organization's information as the main information management problems and 59% blame information quality as "…the biggest barrier to successful analytics or BI initiatives." (p. 4). In another survey, conducted by The Information Difference (2013), only 63% of the participants stated to have "Good" (p. 4) information quality level in their organization. Finally, per company losses of more than $5 million annually are attributed to information quality problems (Forbes Insights, 2010).

Information quality is just as relevant in the context of big data, which is often characterized by the 3Vs: Volume, Velocity, and Variety (Minelli *et al.*, 2012). Volume stands for the large amounts of generated and stored information. Velocity describes the quick rate with which information is

[2] The difference between the terms "data" and "information" and "data quality" and "information quality" will be discussed later. For simplicity, for now these terms will be used interchangeably.

generated and thus changes, and Variety represents information diversity (i.e. with regard to structure, source and format). Information quality is seen as "…the main impediment…" (Witchalls, 2014, p. 27) for generating value from big data and this has been expressed in the addition of a fourth V (Veracity) (IBM Institute for Business Value, 2012), standing for the uncertainty due to low information quality.

As a result of these information quality issues, the number of initiatives for information quality improvement has doubled between 2009 and 2013 (The Information Difference, 2013). However, most companies still apply manual techniques to resolve these problems. In 2013, only 6% of the respondents in a survey had a fully automated system for information quality improvement (The Information Difference, 2013). One possible reason for this is that 39% of the respondents did not measure their information quality at all (The Information Difference, 2013). Measuring information quality is important to identify and analyze the consequences of low information quality and to improve it effectively. Moreover, doing this automatically decreases the required resources and therefore increases the chance of finding efficient improvement measures. The latter is very important in the context of big data with its Volume and Velocity characteristics.

The purpose of this dissertation is to develop quantitative approaches for measuring information quality and for considering the results from these approaches in decision making. Since the measured level is additional information for the decision maker, it can (and should) be considered both *directly* and *indirectly* in decision making (cf. Figure 1). Failing to do so may in both cases result in wrong decisions and economic losses. Moreover, it is easier to automatize quantitative approaches, which, as discussed above, is crucial for big data applications. The relevance of the topic is also demonstrated by recent surveys of the information quality literature (Xiao *et al.*, 2014; Sadiq *et al.*, 2011). In particular, according to Xiao *et al.* (2014) the areas of "data quality assessment" (p. 10) and "data quality for decision support" (p. 10) represent two of the most recent research tendencies in the field.

The focus of this dissertation is on the information quality dimensions accuracy, currency, and consistency as three of the most common and important ones (Experian QAS, 2013; Lee *et al.*, 2002; Eppler, 2006; LaValle *et al.*, 2013)[3]. In particular, already pioneering research works like the ones by Wang and Strong (1996), Redman (1996), and Pipino *et al.* (2002) consider these three dimensions to be among the most important and representative ones. This is also supported later by Eppler (2006) and Batini and Scannapieco (2006). In addition, in a review of the information quality literature for the last 20 years, Sadiq *et al.* (2011) found that "data consistency" is among the top 10 most frequent keywords in existing research. The importance of the three dimensions accuracy, currency, and consistency is evident not only among academics, but also among practitioners. For instance, according to a survey by Experian QAS (2013), low currency and low accuracy represent two of the most common information quality problems for organizations. Moreover, in another survey in the context of big data by LaValle *et al.* (2013), respondents rated consistency and currency of information among the top five priorities for their company. Thus, this dissertation also focuses on accuracy, currency, and consistency.

[3] Later, the relationship between the terms „currency" and "timeliness" as well as between the terms "accuracy" and "correctness" will be discussed. For simplicity, here "currency" and "accuracy" are used.

1.2 Research Questions

Based on the above motivation, the following four research questions are stated:

RQ 1 How can *currency* be adequately measured and how can the *direct* and *indirect* effects of the measured level on decision making be analyzed?

RQ 2 How can *accuracy* in subjective estimations be adequately measured and how can the *direct* effects of the measured level on decision making be analyzed?

RQ 3 How can *consistency* be adequately measured?

RQ 4 Which *requirements* should information quality metrics for data views satisfy to adequately, efficiently, and *directly* support decision making?

The research methods used to address these four research questions are shortly discussed here. Meredith *et al.* (1989) classifies research methods according to two dimensions: rational-existential and natural-artificial dimension. The first dimension concerns the process of generating knowledge in research, while the second one deals with "…the source and kind of information used in research…" (p. 305). The main focus of the research in this dissertation along the first dimension is the rational perspective which represents the formal, objective, and methodological derivation of knowledge. Along the second dimension, in this dissertation the concentration is on the artificial perspective, where the phenomenon of interest is reconstructed "…into another form that is more appropriate for testing and experimentation, such as analytical models, computer simulations, or information constructs." (p. 308).

1.3 Theoretical Framework

To address the above research questions, in this section, a theoretical framework is presented. It is based on the combination of three research areas: a) decision theory, b) knowledge discovery, and c) information quality management. Research area a) covers the approaches for considering both raw and extracted information in decision making (i.e. cases i) and ii) in Figure 1) and implicitly its *direct* effects. Research area b) deals with the methods for *indirectly* considering extracted information in decision making by first analyzing raw information with knowledge discovery techniques (i.e. case 2) in Figure 1). Finally, research area c) presents the methods for defining, measuring, analyzing, and improving information quality.

1.3.1 Decision Theory

Traditional decision theory[4] deals with the development of methodologies for optimal decision making under the assumption of rationality (Bell *et al.*, 1988; Peterson, 2009), where a decision maker is rational if she "…chooses to do what she has most reason to do at the point in time at which the decision is made." (Peterson, 2009, p. 5). There are many different types of decision models in decision theory (Clemen and Reilly, 2001; Marakas, 2003; Peterson, 2009), but all of them share the same basic idea: A rational decision maker has to choose the best (i.e. the optimal) alternative among a set of mutually exclusive alternatives, each of which has its corresponding

[4] In this dissertation the focus is on normative or prescriptive decision theory.

consequences in the form of payoffs, which can be monetary or of other nature (Peterson, 2009). The optimal alternative is the one that maximizes (or minimizes) the objective function. Thus, alternatives, payoffs, and an objective function are the main building elements of each decision model. Optionally, the decision maker may have to take into account other building elements such as budget constraints (Mas-Colell *et al.*, 1995; Hillier and Lieberman, 2005).

A decision model may be classified according to two sources of uncertainty: environment and time, both of which are considered in this dissertation. Environmental uncertainty[5] occurs when factors outside the control of the decision maker influence the outcome of her decisions. These factors are represented by states of nature and the occurring state of nature determines the payoff of the decision maker for the chosen alternative. Thus, states of nature are an additional building element of decision models with environmental uncertainty. In the absence of environmental uncertainty the decision maker knows exactly which state of nature will occur and thus the ex- post payoff for the chosen alternative is also known. The information about the occurrence of the different states of nature is often represented by probabilities (cf. Subsection 1.4.1). Decision models with environmental uncertainty for which these probabilities are known are often called decision models under risk, while such for which this is not the case are called decision models under uncertainty (Luce and Raiffa, 2012).

Time uncertainty[6] occurs when the structure of the decision problem changes over time. This change may be due to external factors (i.e. environmental uncertainty) or due to the chosen alternatives of the decision maker in former periods. Thus, as opposed to decision models with only environmental uncertainty, here the decision maker has to sequentially choose among multiple alternatives, knowing that a choice of an alternative at an earlier point in time has an effect on the possible choices at a later point in time (Clemen and Reilly, 2001; Marakas, 2003). As a result, decision makers may decide to "…give up short-run profit in order to increase long-run profit." (Edwards, 1962, p. 60). Sequential decisions are an additional building element of decision models with time uncertainty.

To sum up, a decision model may be classified in one of four categories depending on the uncertainty type. The first category is the one where there is neither environmental, nor time uncertainty. This decision model is rather trivial, because the decision maker simply chooses the alternative that maximizes the payoff. The second case is the one where only environmental uncertainty exists and the current state of nature is not known. Then, the decision maker takes a decision only once and it has no consequences for further time periods. In the third case only time uncertainty exists. Here, the decision maker knows the current state of nature with certainty, but the future structure of the decision problem is uncertain. As a result, the influence of current decisions on the feasibility of future ones is also unknown. Finally, in the presence of environmental and time uncertainty, both the current state of nature and the future structure of the decision problem are unknown. In this dissertation the focus is on decision models either with environmental uncertainty under risk (i.e. the second type), or with time uncertainty (i.e. the third type).

---

[5] Such decision models are often referred to as decision analysis (cf. Hillier and Lieberman (2005); Mladenic *et al.* (2003); Parnell (2013)).

[6] Such decision models are sometimes called dynamic decision models (cf. Edwards (1962); Mas-Colell *et al.* (1995)).

In both types of decision models, information plays an important role for decision support. In particular, (raw and extracted) information can support decisions by i) structuring the decision problem and/or by ii) reducing environmental uncertainty. In i), depending on the particular decision model, its elements such as alternatives, payoffs, or states of nature are determined from the given information (Hillier and Lieberman, 2005). This is especially relevant for decision models with time uncertainty. Here, approaches such as influence diagrams, decision trees, or the analysis of historical data and expert estimations can be applied (Clemen and Reilly, 2001; Marakas, 2003). In ii) and for decision models with environmental uncertainty, information is used to reduce the uncertainty faced by the decision maker. As a result, information may bring additional benefits due to better decisions and these benefits are referred to in the literature as the normative concept of the value of information (Clemen and Reilly, 2001; Demski, 1980; Goodwin and Wright, 2014; Hillier and Lieberman, 2005; Lawrence, 1999).

1.3.2 Knowledge Discovery

In the literature the terms “data” and “information” are distinguished from each other in that information is considered to be (organized) data which is meaningful for and relevant to the decision problem (Eppler, 2006; Marakas, 2003; Stvilia *et al.*, 2007). This means that, as opposed to data, information is considered in the particular decision context. The concept of knowledge is strongly related to those of data and information. Many researchers have engaged in defining knowledge, but its exact meaning is still rather vague (Marakas, 2003; Nissen, 2002; Dalkir, 2011; Löwstedt and Stjernberg, 2014). One of the most common definitions is the one by Davenport and Prusak (1998) stating that ”Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers.” (p. 5). In this dissertation, knowledge is defined as the processing of information by humans to support problem solving (Fayyad *et al.*, 1996; Liu *et al.*, 2010; Marakas, 2003). This processing is facilitated by different skills, experience and techniques. Once generated, knowledge can be stored and used as extracted information (Nissen, 2002) and in some cases even turned back to data (Davenport and Prusak, 1998).

The methodology of Knowledge Discovery in Databases (KDD) aims at identifying “valid, novel, potentially useful, and ultimately understandable patterns in data.” (Fayyad *et al.*, 1996, p. 83). It consists of five main steps. After understanding the decision problem, the first step is to select the relevant data by choosing a dataset or focusing on a subset of attributes. The result from this step is raw information, since it depends on the particular decision problem. In the second step, this information is preprocessed, for example by removing noise and imputing missing data. In the third step, the preprocessed information is transformed to a form which is more appropriate for analysis. For instance, in this step, feature reduction can be applied or attributes may be normalized. In the fourth step, well-known techniques and algorithms from the field of data mining (e.g. classification, clustering, or association, cf. Kantardzic, 2011) are applied to the generated information to derive useful patterns from it. In the last step, the results are evaluated and interpreted,

thereby discovering new knowledge which is then used as extracted information to support decision making (cf. Figure 1, Subsection 1.3.1). In this dissertation the focus lies on the data mining step of the KDD process.

1.3.3 Information Quality Management

After discussing the role of (data and) information for both the research areas of a) decision theory and b) knowledge discovery, now the research area of c) information quality management is presented, which is the main focus of this dissertation. In particular, less than perfect information quality is considered to generate uncertainty for the decision maker, which must be modeled to avoid wrong decisions. This uncertainty will be called quality uncertainty from now on and is a third type of uncertainty in addition to environmental and time uncertainty discussed above. In the following, the area of information quality management is discussed, dealing with the role of this uncertainty in decision making.

Wang (1998) presents an information quality management cycle consisting of the four phases of *Define*, *Measure*, *Analyze* and *Improve*. These phases are applied iteratively and aim at continuous information quality improvement. The framework presented by Wang (1998) is one of the most fundamental approaches in the information quality management literature and formed the basis of many of the further developments (Xiao *et al.*, 2014). The research in this dissertation is also based on these four phases and thus they will be described in detail in this subsection. Before doing that, first the difference between the terms "data quality" and "information quality" is discussed.

Many authors in the literature (Madnick *et al.*, 2009; Wang, 1998) use the terms "data quality" and "information quality" interchangeably. Based on the previous discussion, information quality is defined in this dissertation as data quality put into a decision context (cf. Eppler, 2006). This distinction is also evident in existing definitions in the literature. For instance, some authors define data/information quality as the "fitness for use" (Wang and Strong, 1996, p. 6, cf. also Olson, 2003). This implies that information must satisfy the user's needs, which depends on the decision context and is thus information quality here. Other authors consider data/information quality to be an objective characteristic of the data itself which does not depend on the decision context and thus address data quality as defined above. For instance, Orr (1998) defines data/information quality as "…the measure of the agreement between the data views presented by an information system and that same data in the real world." (p. 67). This division is also in line with the ideas presented by Wang and Strong (1996) who group data/information quality characteristics into four categories, two of which are "contextual" implying that "…quality must be considered within the context of the task at hand…" (p. 19) and "intrinsic" meaning that "…data have quality in their own right…" (p. 19) (cf. also Strong *et al.*, 1997). According to the previous definition, in the first case information quality is addressed, while in the second one data quality is regarded. However, considering the decision context may happen during each of the four phases of *Define*, *Measure*, *Analyze* and *Improve*. Thus, for now the term "information quality" will be used for simplicity and this discussion will be completed after describing these four phases.

1.3.3.1 *Define* Phase

During the *Define* phase, requirements regarding the different aspects of information quality are identified. These aspects are represented by information quality dimensions (Wang and Strong, 1996). Examples for information quality dimensions are accuracy, currency, completeness, and consistency. Among them, as mentioned above, accuracy, currency, and consistency stand for three of the most important dimensions (Eppler, 2006; Wang and Strong, 1996; Wang *et al.*, 2005) and are thus also the focus of this dissertation. In the following, they are discussed in detail.

The main factor influencing currency is the temporal change of information. Eppler (2006) defines currency with the question "Is the information up-to-date and not obsolete?" (p. 84). According to Nelson and Todd (2005) "Currency refers to the degree to which information is up to date, or the degree to which the information precisely reflects the current state of the world that it represents." (p. 203). Heinrich and Klier (2011) state that currency implies that "an attribute value, which was correct when it was stored in a database, still corresponds to the current value of its real world counterpart at the instant when data quality is assessed." (p. 3). Some authors define timeliness to be equivalent to currency (Ballou *et al.*, 1998), while others distinguish it from currency by considering the decision context (Batini and Scannapieco, 2006). In this dissertation, currency is defined based on Heinrich and Klier (2011) as the degree to which a correctly stored attribute value[7] still corresponds to the real-world attribute value at the time of the decision. This implies that the lower this degree is, the higher the quality uncertainty faced by the decision maker is. After defining currency, now the definition of accuracy is presented.

In the literature accuracy and correctness are seen as related issues. According to Wang *et al.* (2005) "Accuracy measures the degree of correctness of a given collection of data." (p. 24). Similarly, for Olson (2003) accuracy "…refers to whether the data values stored for an object are the correct values. To be correct, a data value must be the right value and must be represented in a consistent and unambiguous form." (p. 29). Eppler (2006) distinguishes between accuracy and correctness by defining accuracy with the question "Is the information precise enough and close enough to reality?" (p. 83) and correctness with the question "Is the information free of distortion, bias, or error?" (p. 84). In that sense, accuracy is the precision of stored information and correctness is defined as its correspondence to the true information. Thus, correctness is a binary characteristic (i.e. either correct or incorrect), while accuracy allows for different degrees of information quality level. If a value is correct, it will also be accurate and vice versa. Batini and Scannapieco (2006) state that "Accuracy is defined as the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v aims to represent." (p. 20). In this dissertation accuracy is defined based on this definition as the closeness of a given attribute value to the corresponding real-world attribute value. Again, the lower the closeness is, the higher the quality uncertainty for the decision maker is. This uncertainty may be due to objective (e.g. storage

[7] The terms "attribute value" and "data value" differ in that in the first case the particular attribute plays a role. For example, a data value would be "single" and an attribute value would be "single" for the attribute "marital status". Thus, they will be used interchangeably, unless the attribute must be mentioned, in which case "attribute value" will be used.

format restrictions) or subjective reasons (e.g. expert estimations). In this dissertation, the focus is on the latter. This completes the definition of accuracy. In the following consistency is defined.

Consistency has been addressed in different domains in the literature and as a result, there exist different, partly overlapping definitions for it. For instance, Blake and Mangiameli (2009) distinguish between integrity, representational consistency, and semantic consistency, where integrity is additionally divided into entity, referential, domain, column, and user-defined integrity. Most of the literature focuses on semantic consistency and so does this dissertation. Semantic consistency is violated "…when two or more data items are contradictory…(Blake and Mangiameli, 2009 based on Lee *et al.*, 2006) or when they are not "…logically compatible…" with each other (Liu and Chi, p. 298, cf. also Valle *et al.*, 2008; Mecella *et al.*, 2002). In this dissertation consistency is defined as the degree to which information "…is free of internal contradictions…" (Heinrich *et al.*, 2007, p. 4), which are determined based on a predefined set of rules (Batini and Scannapieco, 2006). Again, the lower the degree is, the higher the quality uncertainty is. This completes the description of the *Define* phase of the information quality management cycle. In the following the *Measure* phase is discussed.

1.3.3.2 *Measure* Phase

During the *Measure* phase, different information quality dimensions are measured. Sometimes the literature distinguishes between measurement and assessment approaches, in that assessment approaches include benchmarking of the measured values (Batini and Scannapieco, 2006). In this dissertation, a measurement approach is any methodology or technique that results in a numerical value for the level of an information quality dimension or of multiple dimensions. The measurement approaches in the literature can generally be divided into two groups: qualitative and quantitative ones. Qualitative approaches are often based on the methods of qualitative research such as surveys, interviews, and workshops, while quantitative approaches focus on the development of specific information quality metrics by using quantitative methods such as probability theory (cf. Section 1.4).

Table 1 presents a short overview[8] of existing measurement approaches together with the addressed dimensions and whether the approach is qualitative or quantitative. Note that it is difficult to draw a strict separation between the two types of approaches. For example, some methodologies define a metric based on a survey, but it has a quantitative form, while others define a metric in a quantitative way, but require expert estimations as input for it. To avoid confusion, existing approaches are classified as qualitative, only if the measured information quality level is directly determined in a qualitative way, for example by using survey responses and extracting it from verbal descriptions.

An important aspect of the *Measure* phase is the efficiency of a considered metric. If the instantiation and/or the application of a metric are/is too resource intensive, it may not be reasonable to use it. This is also a main drawback of qualitative approaches as opposed to quantitative ones. For example, conducting a survey every time information quality is assessed is less efficient than

[8] This overview does not claim to be complete. It only aims at providing the reader with a basic idea about the existing literature in the field of information quality measurement.

Source	Dimension(s)	Type
(Ballou *et al.*, 1998)	Timeliness	Quantitative
(Batini and Scannapieco, 2006)	Accuracy, Completeness, Currency (based on Ballou *et al.*, 1998), Consistency	Quantitative
(Batini *et al.*, 2011)	Accuracy, Currency	Quantitative
(Cai and Ziad, 2003)	Completeness	Quantitative
(Chayka *et al.*, 2012)	Staleness	Quantitative
(Even and Shankaranarayanan, 2007)	Accuracy, Completeness, Currency, Validity	Quantitative
(Fan *et al.*, 2011)	Currency, Consistency	Quantitative
(Fisher *et al.*, 2009)	Accuracy	Quantitative
(Görz and Kaiser, 2012)	Accuracy	Quantitative
(Heinrich *et al.*, 2009; Heinrich and Klier, 2009, 2011, 2015; Heinrich *et al.*, 2012)	Currency, Timeliness	Quantitative
(Hinrichs, 2002)	General (9 dimensions)	Quantitative
(Hipp *et al.*, 2001; Hipp *et al.*, 2007; Alpar and Winkelsträter, 2014)	Consistency	Quantitative
(Hüner *et al.*, 2011)	Accuracy, Change Frequency, Completeness, Consistency, Timeliness	Quantitative
(Lee *et al.*, 2002)	General (15 dimensions)	Qualitative
(Li *et al.*, 2012)	Availability, Currency, Validity	Quantitative
(Long and Seko, 2005)	Accuracy, Timeliness, Comparability, Usability, Relevance	Qualitative
(Olensky, 2014)	Accuracy	Quantitative
(Price *et al.*, 2008)	Completeness, Accessibility, Flexible content, Flexible layout, Security, Usefulness[9]	Qualitative
(Wang, 1998)	Accuracy, Currency, Completeness, Consistency	Qualitative

Table 1: Short Overview of Approaches for Measuring Information Quality

[9] This is the result after factor analysis and the combination of some of the criteria.

applying a mathematical formula, especially in today's big data era. Thus, in this dissertation the focus is on quantitative approaches. The aim is to quantitatively model the quality uncertainty generated by the less than perfect information quality and to use the results in decision making during the *Analyze* phase.

1.3.3.3 *Analyze* Phase

The *Analyze* phase examines the effects of the measured information quality level on decision making as well as the causes for low information quality (Huang *et al.*, 1999; Wang, 1998). As discussed above, the effects of information on decision making can be *direct* and *indirect* (cf. Figure 1 and Subsections 1.3.1, 1.3.2). Since the level of information quality is a type of information, it can also have these two effects on decision making. For the *direct* effect, the information quality level is incorporated as additional information by the decision maker. This is necessary, because both in the case of structuring the decision problem (i.e. i) above) and in the one of reducing environmental uncertainty (i.e. ii) above), considering quality uncertainty *directly* can result in the avoidance of wrong decisions and economic losses. In addition to that, the information quality level influences decision making also *indirectly* through the KDD process. If the raw information used to generate new knowledge is of low quality, then this may result in wrong extracted information which can cause wrong decisions when incorporated in decision making.

1.3.3.4 *Improve* Phase

Finally, based on the results from the *Analyze* phase, during the *Improve* phase, measures for information quality improvement are considered. Such measures generate costs other than those stemming from the instantiation and the application of the metric and thus must only be applied, if the total costs (including metric instantiation and application) outweigh the benefits (Heinrich *et al.*, 2007; Heinrich *et al.*, 2009). The benefits in this case are the economic improvements due to better decisions as a result of information with higher quality (i.e. lower quality uncertainty).

After describing the four phases, now the discussion about the terms "data quality" and "information quality" can be resumed. As stated above, information quality differs from data quality in that the decision context is taken into account. Generally, the context can be considered during each of the four phases of the information quality management cycle. However, during the *Define* and the *Measure* phases, the context may be considered, but does not have to be. For example, the definition of accuracy above is context-independent, while the definition of consistency may depend on the context, represented by a specific, predefined set of rules. As a result, a metric for consistency (i.e. *Measure* phase) based on this definition will also depend on the context. In some cases the definition may be context-independent, but the context may still be considered during the *Measure* phase. For instance, the above definition of currency is context-independent, but based on the same definition Heinrich and Klier (2009) measure currency by considering supplemental, context-specific data to make more precise estimations. Since the results from the *Measure* phase cannot be considered in decision making without taking the particular context into account, during the *Analyze* and *Improve* phases the context is always considered. Thus, for the *Define* and *Measure* phases, both terms "data quality" and "information quality" may be used, while during

the *Analyze* or *Improve* phase, always information quality is considered. As will be discussed later, all of the papers in this dissertation consider the decision context in at least one of the four phases and thus from now on the term "information quality" will be used. This completes the description of the three research areas which form the theoretical framework for this dissertation. The above ideas are graphically presented in Figure 2. For simplicity in this figure only information quality is illustrated.

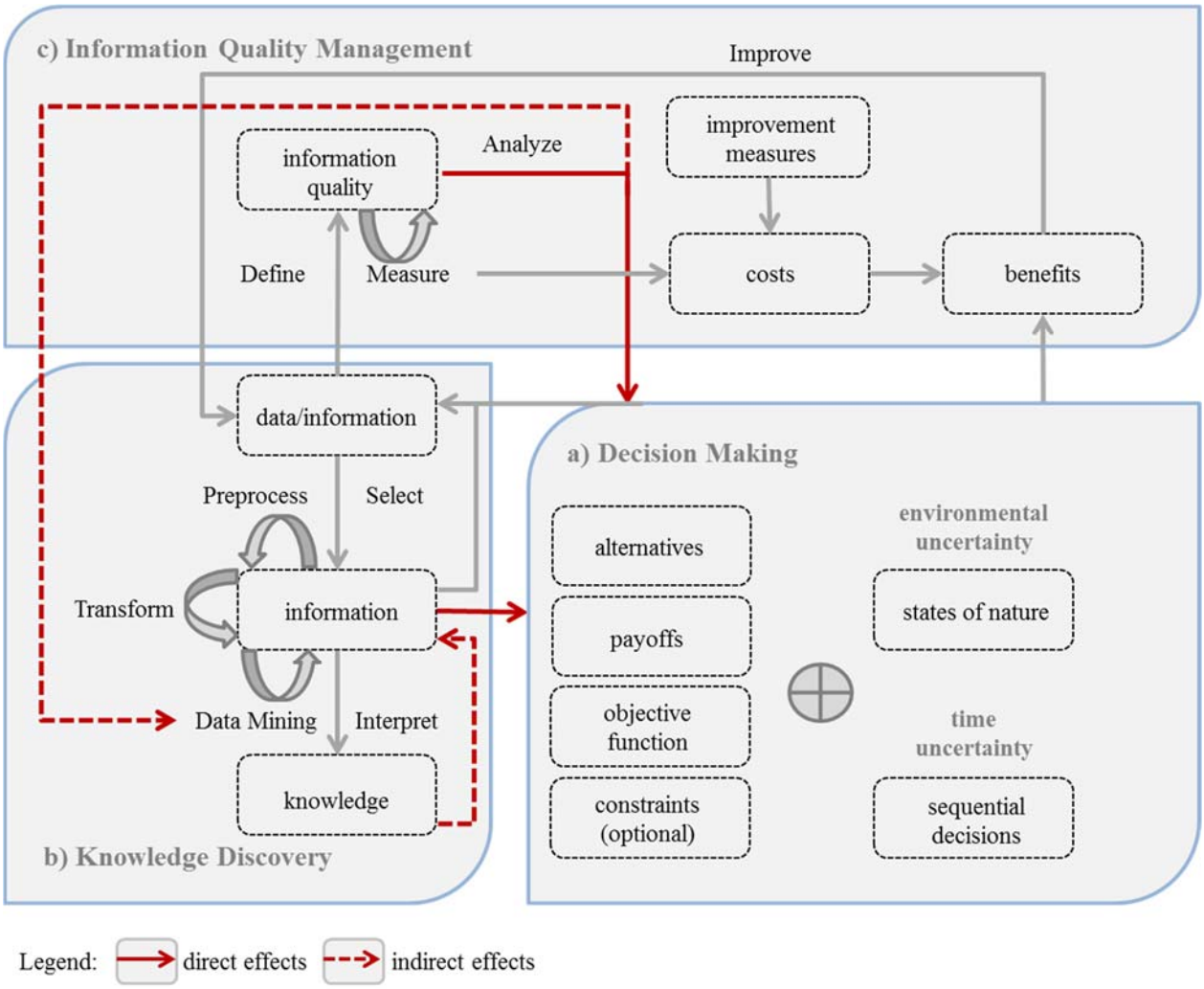


Figure 2: Theoretical Framework

1.4 Methodologies for Modeling Uncertainty

As mentioned above, less than perfect information quality generates quality uncertainty which must be taken into account by the decision maker to avoid wrong decisions and economic losses. Quantitatively modeling this uncertainty is the focus of this dissertation. For this aim, the two well-known methodologies of probability theory and fuzzy set theory are used.

1.4.1 Probability Theory

Probability theory is the standard approach for modeling uncertainty and has been studied for many centuries[10]. Nevertheless, there are still different interpretations of probability representing the different schools. On the one hand, there is the frequentist approach where probability is seen as the "relative frequency" (DeGroot and Schervish, 2012, p. 2) among a large number of repetitions and thus it is a fixed parameter. This approach requires a substantial amount of data from the same event. On the other hand, the Bayesian approach interprets probability as belief (Berger, 1985) which can be adjusted depending on the available data (Zyphur and Oswald, 2013). Although these two approaches represent two rival fields in probability theory, they do not have to be exclusive, but are rather complementary (Berger and Berry, 1988; Zyphur and Oswald, 2013) and also very often lead to the same results (Berger and Berry, 1988; DeGroot and Schervish, 2012). In this dissertation, probability is interpreted based on the frequentist perspective, but also some tools from the Bayesian perspective are used.

Conditional probability plays a very important role in the modeling of information quality, since it is very suitable to express words like "agreement" or "degree of correspondence" (cf. the definitions of information quality and currency above). Conditional probability represents the probability of occurrence of a certain event conditioned upon the already known occurrence of another event (cf. DeGroot and Schervish, 2012) which represents additional information. Thus, if the stored information is known, conditional probability can answer the question about the probability of a certain value of the real-world information conditioned on the stored one. A concept related to conditional probability is the one of a conditional distribution. Generally, a distribution is a collection of probabilities characterizing a given random variable[11] (DeGroot and Schervish, 2012, p. 94) and a conditional distribution is a distribution conditioned on additional information.

Another important approach which is used in this dissertation is the one of hypothesis testing. The idea behind it is to test a presumption about the statistical properties of a given random variable. This presumption may concern the distribution of the random variable or, if the distribution is known, the parameters of the distribution. Hypothesis testing consists of testing the null hypothesis (i.e. the presumption to be tested) against an alternative hypothesis (i.e. what happens if the presumption is not fulfilled). The null hypothesis is then rejected, if there is not enough support for it in the analyzed data. However, the null hypothesis can also be rejected by mistake, because of random effects in the data. The level of significance is the probability of rejecting the null hypothesis even though it is true. The p-value is then the "…lowest level of significance at which the null hypothesis could have been rejected." (Miller and Miller, 2004, p. 402). This idea is applied later when measuring consistency to determine the degree of contradiction with regard to a predefined set of rules (cf. Subsection 1.3.3).

Finally, in this dissertation probability theory is used to model currency for which, as mentioned above, the temporal change of information plays a very important role. This can be modeled by

[10] Some authors attribute the beginning of probability theory to Blaise Pascal and Pierre Fermat (cf. DeGroot and Schervish (2012)).

[11] A random variable is a "…real-valued function that is defined on…" some sample space for an experiment (cf. DeGroot and Schervish (2012), p.93).

using the concept of a stochastic process. In its simplified definition[12], a stochastic process is often described as a sequence of random variables each of which represents a certain point in time (cf. Billingsley, 2012; DeGroot and Schervish, 2012).

1.4.2 Fuzzy Set Theory

Fuzzy set theory presents an alternative to probability theory for modeling uncertainty. Zimmermann (2001) distinguishes probability theory from fuzzy set theory by stating that the first one describes well-defined "…events (elements of sets)…" (p. 3) and fuzzy set theory deals with the uncertainty regarding the "…semantic meaning of the events…" (p. 3). In that sense, fuzzy set theory is appropriate for modeling uncertainties which stem from the inability of humans to make precise estimations and not for uncertainties in objective estimations. It should however be noted that sometimes methods of probability theory are used in fuzzy set theory and vice versa (e.g. fuzzy probabilities, cf. Yager and Zadeh (1992)).

Generally, in the classical set theory, an element either belongs to a set or it does not (i.e. the membership of this element is binary). However, in many real-life situations, it may not be possible to exactly define the boundaries of a set. For example, how to define the boundaries of the set "tall person"? The answer to this question is both highly subjective and vague. In this case classical set theory cannot be applied and thus fuzzy set theory must be considered. In fuzzy set theory the membership of an element to a set is not simply one (equivalent to being a member) or zero (equivalent to being a non-member), but can also be any number in between[13] reflecting the above described semantic uncertainty. For example, a person with a height of 220 cm will certainly (and subjectively) be tall (a membership of one). However, a person with a height of 180 cm will (subjectively) be tall with a degree of 0.8.

The first ideas about fuzzy set theory were presented in 1965 by Zadeh (1965) and since then many additional elements such as types of fuzzy numbers[14], set operations, and fuzzy arithmetic have been developed. These concepts are defined later in the dissertation and thus their discussion is omitted here. However, a very important concept for this dissertation is the one of a linguistic variable. It is thus shortly introduced here and a more precise definition is provided later. A linguistic variable directly addresses the semantic uncertainty described above and as every variable takes values in a given range. However, the values of a linguistic variable are linguistic terms, described by fuzzy sets. For example, the linguistic variable "height of a person" would take the values "short" and "tall" which would be (subjectively) described by fuzzy sets. Linguistic variables have been very extensively applied in engineering, especially in the form of the so called fuzzy inference systems (FIS). They represent a set of parallel IF-THEN rules where both the rule antecedent and the rule consequent can consist of the values of linguistic variables (Wang, 1997).

---

[12] A more rigorous definition will be provided later.

[13] Here it is assumed that the membership function of a fuzzy set is the interval [0, 1]. Some authors define the range of the membership function of fuzzy sets to be the nonnegative real numbers (Zimmermann (2001)).

[14] Fuzzy numbers are defined as normalized, convex fuzzy sets of the real line with a piecewise continuous membership (cf. Zimmermann (2001), p. 59). A more rigorous definition will be provided later.

In addition to linguistic variables, fuzzy optimization represents just as important concept for this dissertation. It is thus also shortly described here. Naturally, fuzzy optimization deals with optimization (i.e. decision) problems where some or all of the elements are "fuzzified". This will happen in a decision situation where the uncertainty is modeled with fuzzy set theory. Thus, in fuzzy optimization there are fuzzy constraints, fuzzy payoffs, fuzzy objective functions, fuzzy alternatives and as a consequence fuzzy order operators for comparison and optimization. This requires the definition of new order operators and even more importantly of new solution approaches.

1.5 Structure and Content of the Dissertation

The dissertation consists of the following seven papers, which address the research questions from Section 1.2.

- Paper 1: A Quantitative Approach for Modeling the Influence of Currency of Information on Decision Making under Uncertainty (RQ 1)
- Paper 2: Considering Currency in Decision Trees in the Context of Big Data (RQ 1)
- Paper 3: A Fuzzy Metric for Currency in the Context of Big Data (RQ 1)
- Paper 4: Revenue Management for Cloud Computing Providers: Decision Models for Service Admission Control under Non-probabilistic Uncertainty (RQ 2)
- Paper 5: Duality in Fuzzy Linear Programming: a Survey (RQ 2)
- Paper 6: Assessing Data Quality – A Novel Probability-based Metric for Consistency (RQ 3)
- Paper 7: Requirements for Data Quality Metrics (RQ 4)

Figure 3 shows which phases of the information quality management cycle are addressed by each of the papers and also which methodology is applied to model quality uncertainty. For clarity, the *Define* phase is separated from the other three phases. As mentioned above, in this dissertation a particular focus is put on currency, accuracy, and consistency as three of the most important dimensions. Papers 1, 2, and 3 address the measurement and analysis of currency and thus RQ 1. Papers 4 and 5 focus on accuracy and RQ 2, while Paper 6 is concerned with the measurement of consistency and RQ 3. Finally, Paper 7 concentrates on the general measurement of information quality dimensions for data views and RQ 4. In the following, each of the papers is discussed shortly.

Paper 1 develops an extended metric referring to currency (*Measure* phase) which, as opposed to existing approaches, provides not only a measure of the correspondence between the stored and the real-world information[15] (cf. Subsection 1.3.3), but also an indication about the real-world information. This is especially important in cases of low information quality where the decision maker may not be able to use the stored information, but based on the indication, still be able to make an informed decision. The extended metric referring to currency is modeled based on a stochastic process. The measured currency level is then *directly* (cf. Figure 1) considered in decision making with environmental uncertainty (*Analyze* phase) by incorporating it in the normative concept of the value of information. The uncertainty (both quality and environmental) in this paper is modeled with probability theory.

[15] Here the focus is not on single attribute values, but on information (signals) which can also be a set of attribute values.

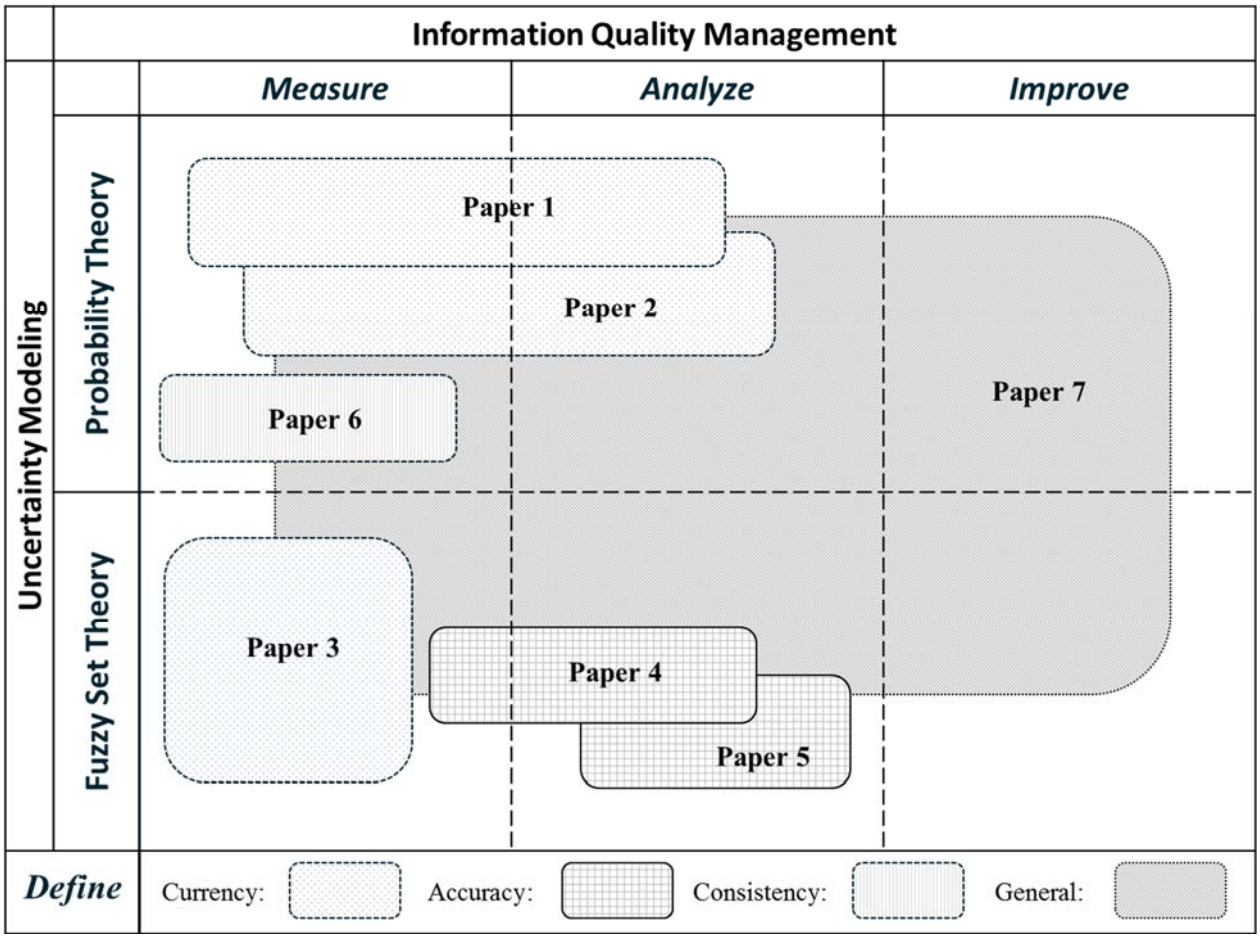


Figure 3: Research Areas and Methodologies for Modeling Uncertainty

Paper 2 also deals with the consideration of the information quality dimension currency (*Define* phase) in decision making, but *indirectly* through the data mining step of the KDD process (cf. Figure 1). In particular, this paper incorporates the ideas about the extended metric referring to currency from Paper 1 (*Measure* phase) in decision trees, which are one of the most common data mining methods. As a result, the negative effects of quality uncertainty on the discovered knowledge are taken into account and thus the likelihood for wrong decisions, based on this knowledge, can be reduced (*Analyze* phase). In addition, the method presented in Paper 2 is very efficient and thus suitable for big data applications.

Both Paper 1 and thus Paper 2 use probability theory to model quality uncertainty. However, applying probability theory may not always be possible because detailed historical data is required. This problem becomes especially evident in the case of big data due to its Volume and Velocity. To address it, Paper 3 develops a metric for currency (*Measure* phase), based on expert estimations by applying fuzzy set theory. The main idea is to use a FIS to model the dependency between the age, decline rate, and currency of attribute values which are all represented as linguistic variables. Paper 3 is the last paper which addresses the information quality dimension currency. Even though the definition of currency above is context-independent, the context is taken into account during the *Analyze* phase for Papers 1 and 2 and during the *Measure* phase for Paper 3. Thus, all of the papers in this group deal with information quality. In the following, the approaches considering the information quality dimension accuracy (*Define* phase) are discussed.

Paper 4 deals with the modeling of accuracy of resource requirements in the context of Cloud computing. Cloud service providers face a decision problem with time uncertainty (*Analyze* phase) regarding the resource requirements of arriving job requests. The reason for this is that jobs arrive sequentially, requiring real-time acceptance decisions. Moreover, according to the often applied pay-as-you-go policy, customers pay for the amount of used resources after the execution of a request. Before the execution, customers submit estimations of the required resources, but may eventually use more or less than that (i.e. quality uncertainty due to the missing/low accuracy of the estimations). Similar to Paper 3, estimating accuracy based on historical data in this case is hardly possible, because of the diversity of jobs and customers and also due to the subjective nature of the resource estimations. Thus, it is appropriate to apply fuzzy set theory for the modeling of accuracy in this scenario (*Measure* phase). The information about the estimated accuracy is then *directly* considered for structuring the decision problem (cf. Figure 1) of the Cloud provider (*Analyze* phase). As a result, the Cloud provider faces a fuzzy optimization problem for which there are no standard solution approaches in the literature. One way to solve this problem is to use the concept of duality in fuzzy optimization (i.e. fuzzy duality theory).

Traditional duality theory considers a pair of a primal and a dual optimization problem so that, under certain conditions, it is enough to solve only one of the two problems to determine the solution of the other one. The same idea can be applied to fuzzy linear optimization when a decision problem (*Analyze* phase) is very difficult to solve, but its dual is not. However, the literature on the topic is rather fragmented, because all of the elements in a decision problem as well as any combination of them can be "fuzzified", resulting in the lack of well-defined solution methods. The aim of Paper 5 is to review the existing literature on duality in fuzzy linear optimization, classify it, identify the research gaps, and propose directions for future work. The results may not only be used when *directly* considering accuracy in decision making as in Paper 4, but also in all other cases where quality uncertainty is modeled with fuzzy set theory such as Paper 3. This completes the presentation of the papers which address the information quality dimension accuracy. Since both papers in this group consider the decision context during the *Analyze* phase, they deal with information quality. In the following, Paper 6 which focuses on consistency (*Define* phase) is discussed.

Paper 6 develops a metric for consistency (*Measure* phase) based on probability theory. As mentioned above, consistency is defined with respect to a predefined set of rules. Many of the existing approaches in the literature treat the fulfillment of these rules as a zero-one decision. This is however unrealistic, because most real-world rules possess an implicit uncertainty with respect to their fulfillment. To address this issue, Paper 6 presents a metric for consistency which considers the probability of fulfillment of a given rule and is based on the p-value concept. In particular, the idea is that if the information of interest is consistent, then the probability (in terms of the relative frequency) of fulfillment of a given rule will correspond to the probability of fulfillment of the same rule for a consistent reference dataset. To measure the degree of this correspondence (and thus consistency), the p-value of the hypothesis test is used under the null hypothesis that the given rule is fulfilled with the same probability in both cases. Thus, the measured result has a clear interpretation and can, similar to Papers 1 and 2, be considered both *directly* and *indirectly* in decision making. Since consistency is defined in a context-specific way, Paper 6 considers information quality.

Papers 1, 3, 4, and 6 present measurement approaches for currency, accuracy, or consistency, the results from which are *directly* (cf. Papers 1, 4, 5) or *indirectly* (cf. Paper 2) considered in decision making. But how does one know whether a metric can adequately support decision making? Paper 7 addresses this question and thus RQ 4 by considering all four phases of the information quality management cycle. It presents a set of five requirements for information quality metrics for data views (*Measure* phase). These requirements are not restricted to any particular information quality dimension (*Define* phase). Metrics which satisfy them can adequately and *directly* be considered in decision making with environmental uncertainty (*Analyze* phase) and are also efficient from a cost-benefit perspective (*Improve* phase). The set of requirements from this paper can then be applied not only to determine the adequacy of existing metrics, but also as criteria for the development of new metrics. The paper covers both probability and fuzzy set theory approaches for modeling quality uncertainty. Since the decision context is considered for specifying the requirements, Paper 7 deals with information quality. The reason why this paper is the last one to be discussed is that, as opposed to the previous papers, it does not focus on a particular information quality dimension or application case. Thus, the examples for metrics for different dimensions provided in the previous chapters facilitate the understanding of the reader, before stating a general set of requirements. This completes the discussion of the seven papers in this dissertation. In Table 2 a short summary of their main components is provided.

Paper Nr.	Environmental uncertainty	Time uncertainty	Quality uncertainty	Direct effects	Indirect effects	Research method
1	✓		✓	✓		normative modeling
2			✓		✓	normative modeling
3			✓			normative modeling
4		✓	✓	✓		normative modeling
5			✓	✓		interpretive-historical analysis
6			✓			normative modeling
7	✓		✓	✓		reason/logic/theorems

Table 2: Uncertainty Types, Role in Decision Making, and Research Methods

First, this table presents the uncertainty types addressed in the different papers (i.e. environmental uncertainty due to the unknown state of nature; time uncertainty due to the sequential nature of decisions; and quality uncertainty). Second, for the papers addressing the *Analyze* phase, a classification with regard to the role of information quality in decision making is additionally provided (i.e. *direct* vs. *indirect*). Third, the classification of the papers, based on the framework by Meredith *et al.* (1989) (cf. Section 1.2) is also given. In particular, Papers 1, 2, 3, 4, and 6 apply a normative modeling approach for the explanation of the phenomenon. As opposed to that, Paper 5 takes more of an interpretive-historical analysis perspective (but still with a formal focus) and Paper 7 follows the reason/logic/theorems approach.

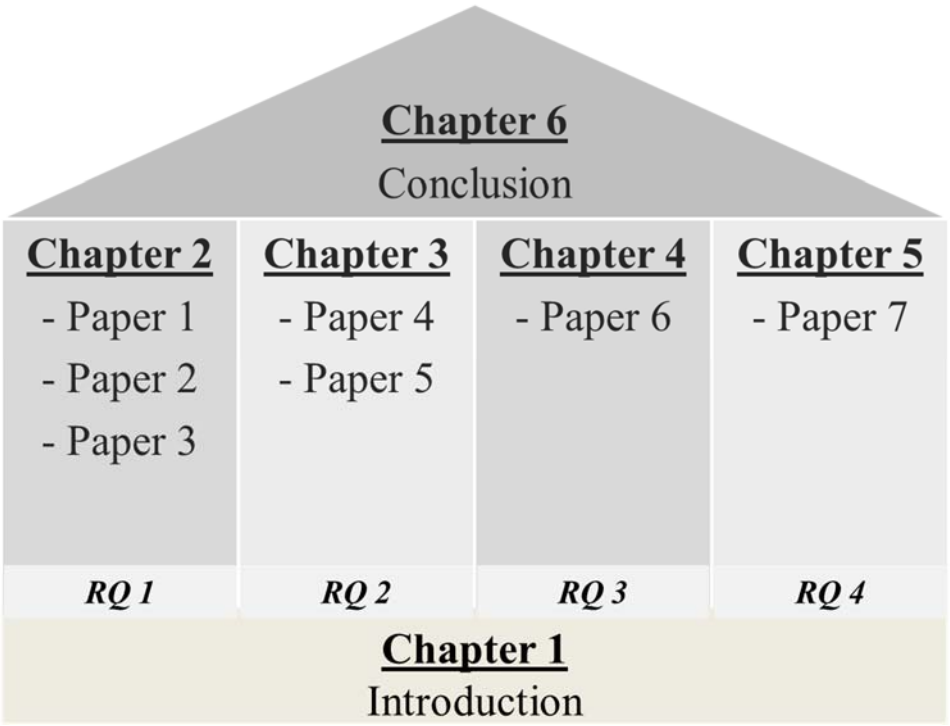


Figure 4: Structure of the Dissertation

This completes the short discussion of the content of the papers in this dissertation. In the following chapters, the seven papers are presented in detail. The chapters are organized according to the four research questions as illustrated in Figure 4. For each paper its highlights, a discussion about how the presented ideas were evaluated, and some limitations are provided. Finally, for each research question, the contribution to it is explicated and the remaining limitations and paths for future research are discussed.

2. Measuring and Analyzing Currency

In this chapter, RQ 1 is addressed by developing quantitative approaches for measuring currency and considering the measured level in decision making (*Measure* and *Analyze* phase). As mentioned above, a crucial factor for the measurement of currency is the temporal change of information. Thus, based on the above definition of currency and also on the existing literature, currency must generally decrease with the age of the stored information. However, the age is not the only factor influencing currency - it is just as important how quickly real-world information changes after storage (i.e. its decline rate). Thus, in this chapter, these two factors are applied to measure currency in two different ways. In the first approach (cf. Sections 2.1 and 2.2), based on conditional probabilities, and stochastic processes, the probability that the real-world information changes from one period to the other is modeled. In the second approach (cf. Section 2.3), based on a FIS, the age, decline rate and currency of attribute values are modeled as linguistic variables.

As discussed above, after measuring currency, the results are considered in decision making either *directly* or *indirectly* through the KDD process (cf. Figure 2). In this chapter, first (cf. Section 2.1) the *direct* effects of currency on decision models with environmental uncertainty are addressed, by extending the normative concept of the value of information. Then, also the *indirect* effects are modeled by considering the measured currency level during the classification of new instances in existing decision trees (cf. Section 2.2).

2.1 Paper 1: A Quantitative Approach for Modeling the Influence of Currency of Information on Decision Making under Uncertainty

Full citation: Heinrich, B. and Hristova, D. (2016), *A Quantitative Approach for Modeling the Influence of Currency of Information on Decision Making under Uncertainty*, Journal of Decision Systems, Vol. 25 No. 1, pp. 16-41, http://dx.doi.org/10.1080/12460125.2015.1080494

Status: accepted on 12.06.2015

Highlights: In this paper an extended metric referring to currency modeled with probability theory is developed. Based on the normative concept of the value of information, its results are then *directly* considered in decision making with environmental uncertainty. The contribution of the paper is twofold. First, the novel metric allows decision makers to consider the level of currency and the distribution of the real-world information in their decisions. This is particularly important in the case of low currency of the stored information. Second, a tool for incorporating the results from the extended metric in decision making is provided by modifying the normative concept of the value of information. The proposed approach presents a dependency between currency and the value of information which is not covered by existing approaches.

Evaluation: The approach is evaluated with two datasets from the SOEP[16] panel data, representing two scenarios from the field of Customer Relationship Management of insurance companies. The results from the evaluation show that low currency exists in real-world applications and that measuring and analyzing it is crucial for avoiding wrong decisions. Moreover, not considering currency in the normative concept of the value of information can lead to wrong estimates of this value and thus wrong decisions in the *Improve* phase.

Limitations: The presented approach also has some limitations. First, the instantiation of the extended metric referring to currency requires a substantial amount of data which is not always given in reality and even less so in the context of big data. Thus, other, less data-intensive approaches for modeling quality uncertainty must be considered. This is addressed in Section 2.3. Second, the approach is restricted to currency, but also other information quality dimensions need to be measured and considered in decision making. This is addressed in Chapters 3 and 4. Finally, considering more than one information quality dimension at a time will be the task of future research.

[16] Socio-Economic Panel Study (SOEP), Data for the years 1984-2011, Version 28: doi:10.5684/soep.v28.

2.2 Paper 2: Considering Currency in Decision Trees in the Context of Big Data

Full citation: Hristova, D. (2014), *Considering Currency in Decision Trees in the Context of Big Data*, in Proceedings of the 2014 International Conference on Information Systems (ICIS), 14-17 December, Auckland, New Zealand.

Status: accepted on 26.09.2014.

Highlights: In this paper an approach is developed for *indirectly* considering currency in decision making through the data mining step of the KDD process. In particular, currency is measured based on the extended metric from Paper 1 and considered during the classification of new instances in existing decision trees. The contributions of the paper are as follows: First, it presents an approach for efficiently considering currency in decision trees, which is very suitable to address the Volume characteristic of big data. Second, by only adjusting the classification of new instances in existing decision trees, this method can be applied to any decision tree method. Third, by considering the structure of the tree and supplemental data, this approach is not only context-specific (resulting in more accurate classification), but also requires less detailed historical data than existing methods for measuring currency. This makes it again very suitable for big data applications due to their Velocity characteristic.

Evaluation: The presented approach is evaluated with three datasets representing two scenarios: a Customer Relationship Management scenario and an activity recognition scenario. The first dataset stands for the first scenario and is again the panel data from the SOEP. The second and the third datasets address the second scenario and are representative for big data. They consist of sensor measurements for the different physical activities performed by a given person. The results from the evaluation show that this approach not only leads to more accurate classification than not considering currency, but also that it is more efficient than standard methods from the literature.

Limitations: Since currency is measured based on the extended metric from Paper 1, the first limitation here is again the necessity of historical data. This limitation can be addressed with alternative methods for measuring currency such as the one presented in Section 2.3, but would require the corresponding adjustment of the decision tree (e.g. using fuzzy decision trees). Second, the focus of the paper is the decision tree method, but the approach can be easily transferred to other data mining methods. Third, this approach does not consider modern big data techniques for increasing efficiency and also it is evaluated only on structured data. Addressing these issues will be the task of future research.

Considering Currency in Decision Trees in the Context of Big Data

Completed Research Paper

Diana Hristova
Department of Management Information Systems
University of Regensburg
Universitätsstraße 31
93053 Regensburg, Germany
Diana.Hristova@wiwi.uni-regensburg.de

Abstract

In the current age of big data, decision trees are one of the most commonly applied data mining methods. However, for reliable results they require up-to-date input data, which is not always given in reality. We present a two-phase approach based on probability theory for considering currency of stored data in decision trees. Our approach is efficient and thus suitable for big data applications. Moreover, it is independent of the particular decision tree classifier. Finally, it is context-specific since the decision tree structure and supplemental data are taken into account. We demonstrate the benefits of the novel approach by applying it to three datasets. The results show a substantial increase in the classification success rate as opposed to not considering currency. Thus, applying our approach prevents wrong classification and consequently wrong decisions.

Keywords: Decision trees, Currency, Information[i] quality, Big data mining

Decision Analytics, Big Data, and Visualization

Introduction

In the current information age companies around the world store huge volumes of quickly changing, distributed, heterogeneous data (IBM Institute for Business Value 2012) to support decision making in areas such as marketing, investment, risk management, production, health care, etc. (Economist Intelligence Unit 2011; Giudici and Figini 2009; Hems et al. 2013; Ngai et al. 2009; Yue 2007). Such data is often called big data and is characterized by the three Vs i.e. Volume, Velocity and Variety[ii]. Data streams are a typical example for big data as they are characterized by both high Volume and high Velocity. Stored (big) data has business value only, if it is analyzed to discover new patterns. Thus, in a recent survey by the IBM Institute for Business Value (2012) more than 75% of the participants "with active big data efforts" (p.12) stated that they apply data analytics techniques to derive valuable insights from it. However, not only data quantity, but also its quality matters.

Stored data may be outdated due to improper update frequency (e.g. address data), attribute values may be missing due to malfunctioning sensors or be inaccurate, because of privacy protection reasons (Aggarwal et al. 2013; Liang et al. 2010) or data integration problems. As a result, if such data is used for analysis without taking its quality into account, false patterns may be identified and thus wrong decisions may be made ("Garbage in, garbage out."). According to a survey by Forbes (2010), most of the participants estimate the cost of poor information quality to more than $5 million annually. The problem is especially evident in the context of big data (Fan 2013; Li et al. 2012; Yang et al. 2012) and to demonstrate its importance IBM has added a fourth V to the characteristics of big data, which stands for Veracity (IBM Institute for Business Value 2012) and represents the uncertainty due to low information quality.

Information quality can be defined as "the measure of the agreement between the data views presented by an information system and that same data in the real world" (Orr 1998, p. 67). Information quality is a multi-dimensional concept consisting of dimensions such as currency, accuracy, completeness, etc. (Wang and Strong 1996). In this paper we focus on currency, as one of the most important among them (Experian QAS 2013; Redman 1996). We define currency as *the degree to which a previously correctly stored attribute value still corresponds to its real-world*[iii] *counterpart at the time of analysis*. This implies that low currency causes uncertainty regarding the correspondence between the stored and the real-world attribute value which should be taken into account in data mining.

The process of Knowledge Discovery in Databases consists of five main steps: selection, pre-processing, transformation, data mining and interpretation/evaluation (Fayyad et al. 1996). Among them, (big) data mining is the application of analytic methods to search for new patterns in the pre-processed and/or transformed data. The aim of classification is to assign a stored instance characterized by the values for a set of independent attributes to one of a predefined set of classes of a given dependent attribute. Decision trees are one of the most commonly applied classification methods (Tsang et al. 2011) due to their simple interpretation (i.e. "white box") and efficiency. They consist of a set of non-leaf and leaf nodes, where each non-leaf node is characterized by a splitting independent attribute condition and each leaf node is characterized by a class of the dependent attribute (Vazirigianniset al. 2003; Witten and Frank 2005). Common applications of decision trees are credit scoring (Koh et al. 2006), fraud detection (Pathak et al. 2011), medical diagnosis (Azar and El-Metwally 2013), and sensor networks (Yang et al. 2012). In all of these applications low currency causes uncertainty in the analyzed data, which should be taken into account in the classification process. Blake and Mangiameli (2011) confirm this point by empirically showing that currency has an effect on the accuracy of classification methods.

In the literature, incorporating data uncertainty in (big) data mining is called uncertain (big) data mining (Aggarwal and Yu 2009; Chau et al. 2006; Leung and Hayduk 2013; Tsang et al. 2011: Yang and Fong 2011). For example, due to the emergence of the Internet of Things, data streams are becoming one of the most common examples of big data in the practice and as a result a number of approaches have been developed for the mining of data streams (Aggarwal 2013) and in particular of uncertain data streams. Typical for the uncertain big data mining literature is that it does not focus on the source of uncertainty, but rather takes it as given without discussing how it can be derived. However, modelling this uncertainty properly (e.g. deriving a probability distribution) with the corresponding interpretation is just as important for real-world applications as incorporating it in the classification process.

In this paper we develop a probability-theory based approach for considering currency of the attribute values of stored instances when classifying them in existing decision trees. The advantage of applying probability theory is that it allows for a mathematically-sound modelling of uncertainty. Our approach addresses the Volume, Velocity and Veracity characteristics of big data. It addresses the Volume by considering currency in an efficient way; Velocity by requiring less detailed historical data than existing approaches for currency; and Veracity by dealing with data uncertainty. Moreover, our approach is universal, because it is independent of the particular decision tree classifier, and adaptable to the given context of application and to the structure of the decision tree.

Thus, our approach aims at extending both the information quality and the uncertain big data mining literature and at closing[iv] the gap between them. The first stream of research is extended by proposing an efficient way for measuring currency in decision tress which is applicable in the context of big data. The second stream of research is extended by demonstrating how the uncertainty, which is assumed to be given in existing works, can be measured in an interpretable, efficient and context-specific way.

The practical relevance of our approach can be illustrated by an example from the field of Customer Relationship Management (CRM)[v]. Consider a financial service provider who would like to conduct a campaign for winning new customers based on their annual income. The annual income of a customer depends strongly on other personal characteristics such as education, age, and employment status and this relationship can be modeled as a decision tree. If the personal characteristics are outdated (e.g. because they were stored long before the time of the campaign), the customer may be classified in the wrong income class and thus offered the wrong product resulting in losses for the company.

Another example, which is a typical big data case, comes from the field of sensor measurements. Consider a dataset for handicapped individuals, whose movements, location, speed, etc. are measured by sensors integrated in their clothing. Based on the measurements and a decision tree classifier, their current activity is determined (Yang et al. 2012). As a result, emergency situations (e.g. an accident) can be detected and thus corresponding measures derived (e.g. sending an ambulance). If the data measured by the sensors is outdated, for example due to a delayed transmission[vi] and if this is not considered in the classification, an emergency case can be detected either with a strong delay or not at all, causing serious consequences for the patient.

The paper is structured as follows. In the next section we give a literature review and provide the required background on the topic. In the third section our approach for incorporating currency in decision trees in the context of big data is presented. In the fourth section it is evaluated based on three different datasets representing the two applications above. In the final section main conclusions are drawn and limitations and paths for future research are discussed.

Related Work and Background

Since we extend both the literature on modelling currency in decision trees and the one on uncertain big data mining with decision trees, in the following we first present the main findings in these two streams of research as well as our contribution to them.

Modelling Currency in Decision Trees

As mentioned above, information quality is a multi-dimensional concept including characteristics such as currency, accuracy, completeness (Wang and Strong 1996). In the context of data mining, lower information quality can be seen as causing two types of uncertainty: existential and attribute-level uncertainty (Aggarwal and Yu 2009). Existential uncertainty represents the uncertainty whether an attribute value does or does not exist (i.e. completeness), while attribute-level uncertainty stands for the uncertainty regarding the existing attribute value which is possibly of poor quality (i.e. concerning accuracy, currency, etc.).

Existential uncertainty is well-studied by researchers (Dasu and Johnson 2003; Hawarah et al. 2009; Quinlan 1986; Witten and Frank 2005; Yang et al. 2012). The reason for this is that it is rather straightforward to measure and thus to identify it. To reduce existential uncertainty, missing attribute values are either replaced (i.e. imputed) based on point estimation (e.g. by the mean of the non-missing attribute values) or represented as a probability distribution over the domain of the attribute (Quinlan

1986). Already Quinlan (1986) proposes a probability-theory based approach for considering data completeness in classifying data instances in decision trees.

Attribute-level uncertainty is a relatively new topic of research. A few authors have developed metrics for measuring accuracy (Fisher et al. 2009), currency (Ballou et al. 1998; Blake and Mangiameli 2011; Even and Shankaranarayanan 2007; Heinrich and Klier 2009; Heinrich and Klier 2011; Li et al. 2012; Wechsler and Even 2012) and other attribute-level information quality dimensions (Alpar and Winkelsträter 2014; Mezzanzanica et al. 2012). Since the focus of this paper is on currency, we concentrate in the further discussion on it.

Based on our definition of currency above, it is very natural to interpret currency as probability. This is also in accordance with the literature. For example, Heinrich and Klier (2011, p. 6) interpret currency as "...the probability that an attribute value stored in a database still corresponds to the current state of its real world counterpart..." i.e. as the probability that the stored attribute value is up-to-date. Thus, we can estimate the currency of a stored attribute value with the help of historical data, which can be either publicly available (e.g. Federal Bureau of Statistics 2013) or company-internal data. To demonstrate the idea, consider the attribute $Marital\ status$ which can take the values $\{single, married, divorced, widowed\}$. If the stored attribute value is $single$, then the currency of this attribute value will be the probability that a person, who was single at the time of storage, is still single at the time of classification. Thus, currency can be defined as the conditional probability $P(single\ in\ t_1 |\ single\ in\ t_0)$, where t_1 is the time of classification and t_0 is the time of storage. Based on the Bayes' theorem (Berger 2010), the currency for the attribute value $single$ can be derived as:

$$P(single\ in\ t_1 |\ single\ in\ t_0) = \frac{P(single\ in\ t_1\ AND\ single\ in\ t_0)}{P(single\ in\ t_0)} \quad (1)$$

where the operator AND stands for the intersection of the two events. To calculate these probabilities, we follow the frequentist perspective (Berger 2010) according to which a probability "...is the proportion of the time that events of the same kind will occur in the long run." (Miller et al. 2004, p. 24). This implies that the probability of a certain event is represented by the percentage of instances in the population characterizing this event. Thus,

$$P(single\ in\ t_1 |\ single\ in\ t_0) = \frac{|single\ in\ t_1\ AND\ single\ in\ t_0||instances|}{|single\ in\ t_0||instances|} = \frac{|single\ in\ t_1\ AND\ single\ in\ t_0|}{|single\ in\ t_0|} \quad (2)$$

where $|single\ in\ t_0|$ stands for the number of instances in the population who were single at the time of storage. Analogously, $|single\ in\ t_1\ AND\ single\ in\ t_0|$ represents the number of instances in the population who were single both at the time of storage and at the time of classification. Finally, $|instances|$ stands for the total number of instances in the population.

Calculating currency in this manner provides context-free results, which are independent of the other attribute values of a stored instance. This has its advantages and is in accordance with existing metrics for currency based on probability theory (Heinrich and Klier 2009, 2011; Li et al. 2012; Wechsler and Even 2012). However, as the literature has shown, the context of application plays a role for the measurement of currency. Even and Shankaranarayanan (2007) develop context-specific metrics for different information quality dimensions including currency. They use a so called data utility measure, which reflects the business value in the specific context and give a contextual interpretation of currency as "The extent to which outdated data damages utility."(p. 83). Their metric for currency is thus represented as the relative decrease in utility due to outdated data for a particular *context*. Similarly, Heinrich and Klier (2009) propose the use of supplemental data for a more precise estimation of currency (defined as timeliness by Heinrich and Klier (2009)) adapted for the particular context of application. Supplemental data is different from metadata such as storage age in that it is attribute-value specific. Based on this approach, we consider the context of application by determining the conditional probabilities, so that the probability that an attribute value is up-to-date depends on the values of the supplemental data.

To demonstrate this idea, consider again the attribute $Marital\ status$ from above. The currency of the attribute value $single$ can be derived much more precisely, if additional information about the personal characteristics of the individual is considered. For example, the age and the gender of the person would have an influence on the changes in his/her marital status. Given a single, 20-year-old (at the time of storage) female, we can compute the currency of her marital status as the conditional probability $P(single\ in\ t_1 | single\ in\ t_0, female, 20\ in\ t_0)$, which after applying the Bayes' theorem can be rewritten as:

$$P(single\ in\ t_1 |\ single\ in\ \ t_0, female, 20\ in\ t_0) = \frac{P(single\ in\ t_1 AND\ single\ in\ \ t_0, female, 20\ in\ t_0)}{P(single\ in\ \ t_0, female, 20\ in\ t_0)} \quad (3)$$

Again, based on the frequentist perspective, (3) can be rewritten as:

$$P(single\ in\ t_1 |\ single\ in\ \ t_0, female, 20\ in\ t_0) = \frac{|single\ in\ t_1 AND\ single\ in\ \ t_0, female, 20\ in\ t_0|}{|single\ in\ \ t_0, female, 20\ in\ t_0|} \quad (4)$$

The result in (4) will deliver a much more precise indication about the currency of the stored attribute value than the one in (2) as it considers additional information about the stored instance. However, such a detailed historical data is not always available in reality (Velocity) and determining the conditional probability for all values can be computationally rather intensive (Volume). These challenges are addressed by our approach.

In this paper we aim to draw on the findings in the literature on modelling currency by 1) measuring currency based on the interpretation by Heinrich and Klier (2011) and 2) considering supplemental data in the measurement process. We extend the literature by a) proposing measurement methods, which are more efficient and do not require such a detailed historical data, and by b) demonstrating how the measured currency can be considered during the classification of stored instances in existing decision trees. In particular a) is very relevant in the context of big data. b) is partly based on the ideas from the field of uncertain big data mining. In the next subsection we discuss existing approaches in this field by focusing on decision trees.

Uncertain Big Data Mining for Decision Trees

In order to analyze uncertain big data, traditional data mining approaches need to be modified to consider both the characteristics of data uncertainty and these of big data. On the one hand, data uncertainty is considered in the field of uncertain data mining. However, most of these approaches are not suitable for big data because the Volume and Velocity of big data require more efficient methods. On the other hand, the field of big data mining concentrates on the efficiency of the algorithms (especially for data streams as one of the most common applications), but the methods there do not consider data uncertainty. The approaches in the field of uncertain big data mining can thus emerge either i) from the field of uncertain data mining by improving the efficiency of the methods, or ii) from the field of big data mining by increasing the robustness to uncertainty, or iii) by combining both of them. Figure 1 presents this idea. Since our approach is part of the first group, in the following we first describe the ideas in the field of uncertain data mining for decision trees.

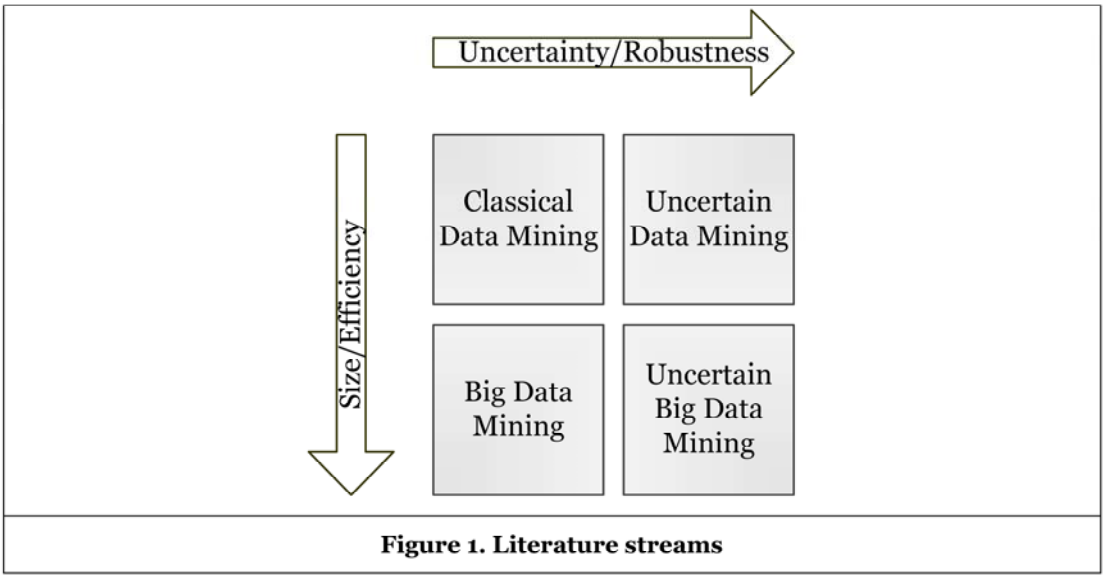


Figure 1. Literature streams

Stream of research i)

A number of authors have developed approaches that modify decision trees to work with attribute-level uncertainty, usually based on probability theory. One major group is represented by the works of Qin et al. (2009) and Tsang et al. (2011). In these approaches the independent attribute values of the data instances are assumed to be uncertain and described by a probability distribution (Qin et al. 2009). For each instance, these distributions are propagated down the tree by multiplying the probabilities characterizing a certain path of the tree. Finally, the resulting path probabilities are multiplied with the probability for a certain class at the leaf of each path and the probabilities for the same class are summed over the paths forming the class probabilities. The data instance is classified in the class with the highest class probability (majority vote). New decision trees are built by defining suitable splitting functions such as an information gain measure based on probabilistic cardinality (Qin et al. 2009). The result is a non-probabilistic tree. Similarly, Magnani and Montesi (2010) generate a table for each possible world alternative (data sources) that can occur and assign to each of these tables the probability of the corresponding alternative. This corresponds to the probability distributions in Qin et al. (2009) and Tsang et al. (2011). Then, from each of the tables (i.e. based on certain data) a decision tree is built, which would occur with the probability of the corresponding possible world alternative. In order to classify a data instance, for each class and each tree, the probability that the data instance belongs to this class is multiplied with the probability that the tree occurs. The final probability for a given class is the sum over the alternative trees, similar to the class probability in Qin et al. (2009) and Tsang et al. (2011).

The main aim of the above approaches is to consider data uncertainty in decision trees and not to provide efficient methods[1]. Thus, they need to be extended with regard to their efficiency to be applied to big data. To our knowledge, there is no such approach for decision trees in the literature (i.e. stemming *only* from the uncertain data mining literature). However, Leung and Hayduk (2013) extend the well-known UF-growth algorithm for mining uncertain frequent patterns (Leung et al. 2008) for application to big data. Their point is that the presence of uncertainty increases the search space and thus the runtime of the algorithm. Thus, Leung and Hayduk (2013) apply the MapReduce framework (Aggarwal 2013) which efficiently mines large volumes of distributed data to identify frequent patterns in the case of big data. Our approach is also based on such an idea. To better justify it, in the following we discuss the papers for uncertain big data mining, stemming from the big data mining literature. We focus on data streams, which are one of the most common applications.

Stream of research ii)

Attribute-level uncertainty in data streams can be divided into noise and concept drift uncertainty. Noise is seen as data, which "do not typically reflect the main trends but makes the identification of these trends more difficult" (Yang 2013, p. 322[vii]) and can be interpreted as representing incorrect, inaccurate or outdated attribute values (Zhu and Wu 2004). Concept drift describes the change in the data distribution over time and appears when the distribution is not stationary, but "is drifting" (Hulten and Domingos 2001). For example, the relationship between annual income and employment status may change over time and as a result, the decision tree describing this relationship will change. The idea of a concept drift is related to currency in that it considers the development of data over time, but also differs from it, as it is concerned with the change in the underlying distribution over time, while currency is determined by the change of the corresponding attribute values in the real-world. A change in currency in the above example will be, if the employment status of the stored instance changes in the real world resulting in a different annual income class than the stored one.

Many of the papers that deal with classifying *certain* data streams with decision trees are based on the Hoeffding bound for tree growing upon the arrival of new data. The idea is that a new split takes place only when there is a statistically significant support for it in the data. The most famous such approach is the Very Fast Decision Tree (VFDT) (Domingos and Hulten 2000), which classifies data streams efficiently and incrementally. Since it has been shown that the presence of noise can reduce the classification accuracy and increase the tree size of decision trees for data streams (Yang and Fong 2011), existing approaches for classifying uncertain data streams are designed to be more robust against noise as compared to the ones for certain data. Examples for such approaches are the extensions of the VFDT

[1] Note, however that the authors still discuss some efficiency improvements (Tsang et al. 2011).

method (Yang and Fong 2011) and the FlexDT method (Hashemi and Yang 2009)[viii]. VFDT has been extended to account for noise by introducing a dynamic tie threshold based on the changing mean of the Hoeffding bound (Yang and Fong 2011). Hashemi and Yang (2009) apply fuzzy sets in the FlexDT method to mitigate the effect of noise on the classification accuracy. The approach is similar to the approaches by Qin et al. (2009) and Tsang et al. (2011), but is based on fuzzy set theory.

In order to consider the occurrence of a concept drift, approaches for classifying data streams use new instances to test the suitability of the existing decision trees. For example, Yang et al. (2012) and Hulten and Domingos (2001) update the sufficient statistics of the tree based on new instances and as a result decide if the tree structure should be changed. Hulten and Domingos (2001) grow gradually a new subtree which eventually replaces the old one. Hashemi and Yang (2009) backpropagate new instances up the tree and update the fuzzy parameters for potential changes in the data. If necessary, they grow a new subtree based on Hulten and Domingos (2001). Finally, Wang et al. (2003) apply ensemble classifiers to classify data streams and adjust their weights based on the values of new instances. As mentioned above, our focus is on the currency of the attribute values and not on[ix] the stationarity[x] of its distribution over time.

To sum up, the approaches for uncertain big data mining stemming from the big data mining literature are designed to be robust against noise and thus currency. However, they do not examine the source of the noise and do not model it explicitly[2]. This is the research gap we aim to close.

Stream of research iii)

An approach that has emerged from both the big data mining and the uncertain data mining literature is the one by Liang et al. (2010). They apply the idea for probabilistic cardinality (Qin et al. 2009) to model the uncertainty in data streams and update the sufficient statistics for new samples similar to Yang et al. (2012) and Hulten and Domingos (2001). However, they do not model the source of uncertainty.

Typical for all the presented approaches in the uncertain big data mining literature is that uncertainty is considered to exist, but without providing an interpretation for it. The approaches in the uncertain data mining literature assume that it is given in the form of a probability distribution without discussing the derivation of this distribution. The works dealing with the classification of uncertain data streams do not model uncertainty explicitly[2], but either modify their methods to increase their robustness against noise or update the decision trees in the case of a concept drift. Our approach contributes to the uncertain big data mining literature in *Stream of research i)*. We model currency as a probability distribution and classify stored instances by propagating their probabilities down the tree similar to Qin et al. (2009) and Tsang et al. (2011). We extend the literature by a) giving an interpretation of the probabilities in the form of currency, b) showing how these probabilities can be efficiently determined for an application in big data, and c) deriving them in a context-specific way with the use of supplemental data. We have chosen probability theory, because: 1) it allows for a mathematically-sound modelling of uncertainty, 2) it is a common well-founded approach for modelling currency in the literature and 3) the literature on uncertain data mining is mainly based on it. In the next section we present our approach.

A Method for Considering Currency in Decision Trees

In this section we present our two-phase approach, which considers the currency of stored instances when classifying them in *existing* decision trees. As a result, we do not concentrate on the process of building the tree, but only on the classification of stored instances in existing decision trees. Therefore, our approach is independent of the particular decision tree classifier (e.g. CHAID, C4.5[xi]). The considered independent attributes can be numerical or categorical, while categorical dependent attributes are assumed, which is a standard assumption in the uncertain data mining literature[3] (e.g. Liang et al. 2010; Qin et al. 2009).

---

[2] An exception here is the paper by Hashemi and Yang (2009), who apply fuzzy set theory, but we focus on probability theory.

[3] Otherwise, a leaf will not be characterized by a given class of the dependent attribute, but by a probability distribution and thus final classification will not be possible.

Notation

In order to describe the approach we first introduce some necessary notation. As mentioned above, decision trees aim at classifying instances according to the values of a set of independent attributes in one of the classes of the dependent attribute. A path in the tree begins at the root and ends in one of its leaves. It is thus represented by a sequence of the splitting conditions of the independent attributes and the corresponding class of the dependent attribute. Let $A_i^I \in D_i^I, i \in \{1, \dots, n\}$ be the set of independent attributes with their corresponding domains D_i^I. Let, in addition, $d_i^{j(k)}, j(k) \in \{1(1), \dots, m(k)\}, k \in \{1, \dots, t\}$ represent the disjoint splitting independent attribute conditions for the attribute A_i^I at depth k of the tree (i.e. $\bigcup_{j=1}^{m} d_i^{j(k)} = D_i^I$) where t is the maximal depth of the tree. We call $d_i^{j(k)}, j(k) \in \{1(1), \dots, m(k)\}, k \in \{1, \dots, t\}$ *splitting subsets*. Let $A^D \in D^D$ analogously be the dependent attribute with its corresponding domain D^D and disjoint classes $c_l, l \in \{1, \dots, p\}$, $\bigcup_{l=1}^{p} c_l = D^D$. A path leading to a leaf node with the class c_l is then given by a sequence $path_{lu} = \left\{d_{i1}^{j(1)}, d_{i2}^{j(2)}, \dots, d_{it}^{j(t)}, c_l\right\}, \{i1, \dots, it\} \subseteq \{1, \dots, n\}$[xii], where u represents the different paths leading to the same class[4][xiii]. A stored data instance is given by $G = \{s_1, \dots, s_n\}$, where $s_i \in D_i^I, \forall i$ represent the stored values of the corresponding independent attributes.

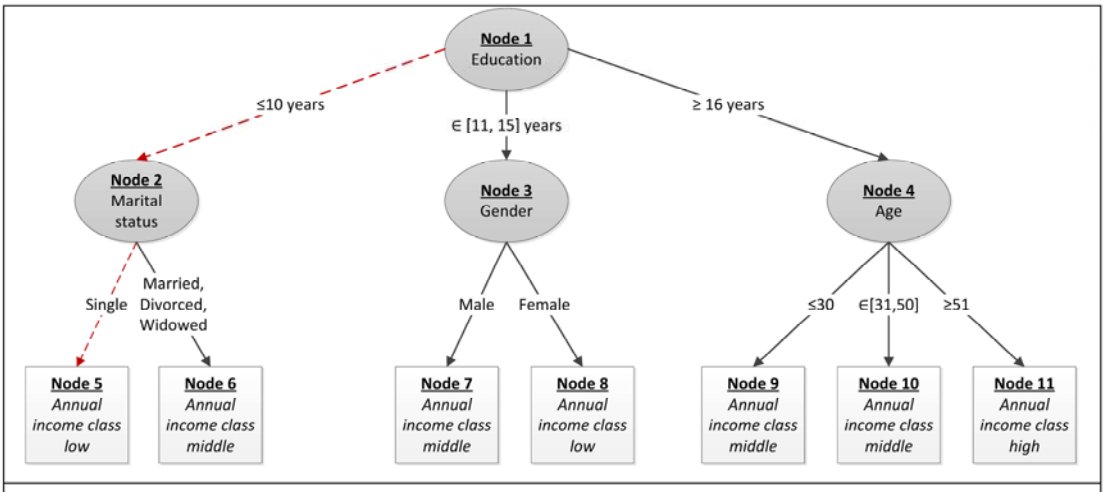


Figure 2. Decision Tree Example

To illustrate the idea, consider the tree in Figure 2, where the instances need to be classified according to the income of the person. The independent attributes are *Education* (*in years*), *Marital status*, *Gender*, and *Age* (*in years*), the dependent attribute is *Income* with $c_1 = low, c_2 = middle, c_3 = high$. An example for a path is $path_{11} = \{\leq 10\ years, single, low\}$ (marked in Figure 2 with a dashed line) and an example for a stored instance that would follow this path (i.e. it would be classified in the class $c_1 = low$) is $G = \{10\ years, single, male, 18\ year\ old\}$.

Since the instance which must be classified, was stored some time ago, it may be the case that some or all of its values are outdated at the time of classification. Not considering the currency of these values may result in wrong classification and thus wrong decisions. For example, for the instance $G = \{10\ years, single, male, 18\ year\ old\}$, the individual may have got married or completed his university education (at least 16 years of education) in the meantime and thus be in the class of medium or high earners. If this is not considered, wrong decisions and thus economic losses may occur.

[4] Note that the values $\{i1, \dots, it\}$ represent single numbers from the set $\{1, \dots, n\}$.

First Phase

In order to consider currency in decision trees, we first need to measure it and this is the *first* phase of our approach. For this aim, we follow the interpretation from above, where currency is seen as the probability that the stored attribute value is up-to-date. Thus, as discussed in the second section, currency can be determined independently of the given decision tree based on the percentage of instances in the population which had the stored attribute value (e.g. single) at the time of storage and still have it at the time of classification (e.g. are still single).

This way of determining currency derives the probability for every possible value in the domain of the independent attribute separately. However, it is rather impractical, especially in the context of big data, since it requires reliable and detailed historical data and is computationally very intensive. Moreover, for decision trees, this high precision of the estimations is very rarely needed. The reason is that the splitting subsets in decision trees are often sets or intervals rather than single values. This is especially true for numerical attributes such as $Education$ or Age, but may also occur with categorical attributes such as $Marital\ status$ or $Industry$. For example, in Figure 2 the splitting subsets for the attribute $Marital\ status$ are given by the sets $d_2^{1(2)} = \{single\}$ and $d_2^{2(2)} = \{married, divorced, widowed\}$, and for $Education$ by the intervals $d_1^{1(1)} = [0, 10]$, $d_1^{2(1)} = [11, 15]$, $d_1^{3(1)} = [16, \infty)$. As a result, for a correct classification, it is not anymore crucial that the stored attribute value is up-to-date. Rather, it is enough that even if the value changed in reality, it still belongs to the same splitting subset. For example, in Figure 2, if the stored attribute value is *married* and the person is *divorced* at the time of classification, the correct classification of the stored instance will not be affected.

Thus, we propose an approach for considering currency in decision trees which is based on the splitting subsets of the particular tree. To describe the idea, consider for some $i \in \{1, .., n\}$ the attribute value s_i from the stored instance $G = \{s_1, \dots, s_n\}$ with the splitting subset(s)[5] $d_i^{j*(k)}$ such that $s_i \in d_i^{j*(k)}$ i.e. the attribute A_i^l is at depth k of the tree. We call $d_i^{j*(k)}$ *storage splitting subset(s)*. Then the currency of s_i is the probability that the instances which attribute values were in $d_i^{j*(k)}$ upon storage are still in it upon classification. We denote this probability by $p_i^{j*(k)}$. If the stored attribute value is up-to-date, then $p_i^{j*(k)} = 1$. This probability can be derived from historical data based on the Bayes' theorem. For example, in Figure 2 and $s_1 = 9\ years$, $p_1^{1(1)}$ will be the probability that during the time span between storing and classifying the instance, a person with less or equal to **ten** years of education at the time of storage did not study **more** than ten years until the time of classification. This implies that a stored instance with $s_1 = 9\ years$ will follow the path to $d_1^{1(1)} = [0, 10]$ with a probability $p_1^{1(1)}$. As a result, it will follow the path to $d_1^{2(1)} = [11, 15]$ or to $d_1^{3(1)} = [16, \infty)$ with a total probability of $1 - p_1^{1(1)}$. The exact probability with which the stored instance will follow each of these two paths is then the probability that the stored attribute value changed between storage and classification to a value, which belongs to one of the two splitting subsets. For example, for $d_1^{2(1)}$ this is the probability that a person who had less or equal to ten years of education at the time of storage, has received between eleven and fifteen years of education until the time of classification. It can be derived analogously to $p_1^{1(1)}$ from historical data.

Note that, in order to determine all these probabilities, we only need to know the corresponding storage splitting subset(s) $d_i^{j*(k)}$ and the structure of the tree. Thus, they can be derived independently of the stored instances (for all splitting subsets $d_i^{j*(k)}$) resulting in higher efficiency. Let, for each possible storage splitting subset $d_i^{j*(k)}$ (i.e. each splitting subset of the tree), $p_i^{j(k)}$ represent the so derived probabilities for each splitting subset $d_i^{j(k)}$ (including $d_i^{j*(k)}$) at the time of classification. We call $p_i^{j(k)}$ the *node* probability of $d_i^{j(k)}$ for the storage splitting subset $d_i^{j*(k)}$. Then, for a given stored instance $G = \{s_1, \dots, s_n\}$ and for each of its stored attribute values $s_i, i \in \{1, \dots, n\}$, only the corresponding storage

[5] Note that an independent attribute may happen to be more than once in a tree a splitting independent attribute.

splitting subset(s) $d_i^{j*(k)}$ with $s_i \in d_i^{j*(k)}$ need(s) to be identified to derive the node probabilities for $G = \{s_1, \ldots, s_n\}$. Figure 3 provides the node probabilities for the tree in Figure 2 and the stored instance $G = \{10\ years, single, male, 18\ year\ old\}$.

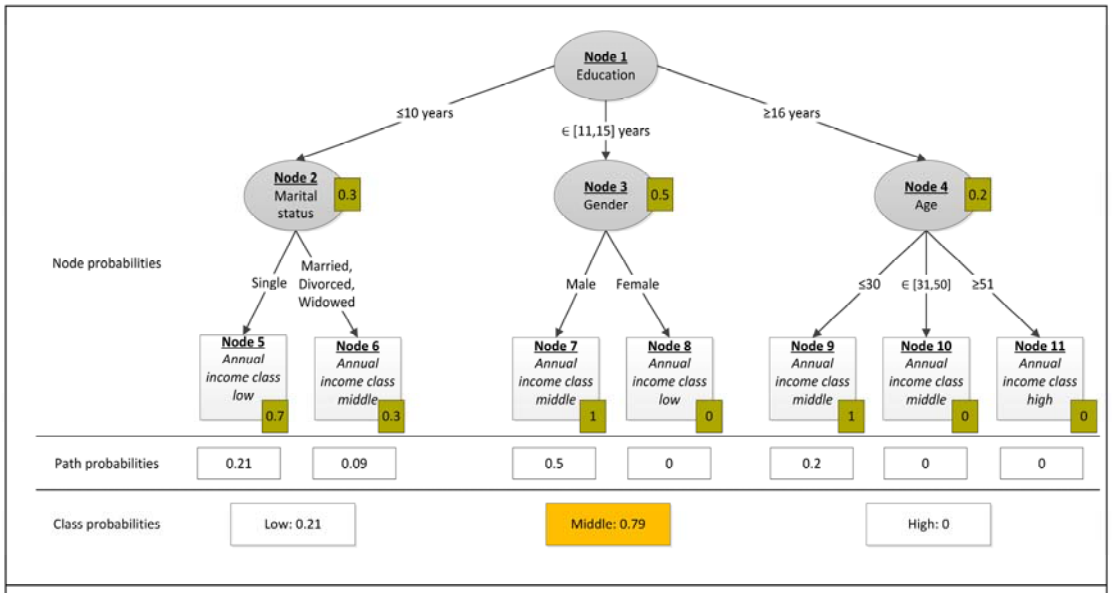


Figure 3. An Example for Considering Currency in a Decision Tree

This completes the description of the derivation of the node probabilities and thus the *first* phase of the algorithm. By deriving the node probabilities according to the splitting subsets of the decision tree, we model them more efficiently than the existing approaches for measuring currency. In addition, such a derivation considers the structure of the decision tree and thus the context of application. Finally, as opposed to the approaches in the uncertain big data mining literature, these probabilities have an interpretation based on the currency of the stored attribute values.

Second Phase

The *second* phase consists of classifying the stored data instance, which follows a certain path to reach a leaf with a class for the independent attribute. Based on the results from the *first* phase, for a given stored instance, we can assign to each path $path_{lu} = \left\{d_{i1}^{j(1)}, d_{i2}^{j(2)}, \ldots, d_{it}^{j(t)}, c_l\right\}, \{i1, \ldots it\} \subseteq \{1, .. n\}$ a sequence of the corresponding node probabilities $\left\{p_{i1}^{j(1)}, p_{i2}^{j(2)}, \ldots, p_{it}^{j(t)}\right\}, \{i1, \ldots it\} \subseteq \{1, .. n\}$. For example, in Figure 3, the node probabilities for the path $path_{11} = \{\leq 10\ years, single, low\}$ are given by $\{0.3, 0.7\}$[xiv]. This implies that the first splitting subset in the path will be followed with a probability of 0.3 and the second one will be followed with a probability of 0.7. In the next step, we need to determine the total probability that the stored instance follows a particular path, which we call *path* probability. We derive it, based on the uncertain data mining literature, by multiplying the node probabilities of the splitting subsets of the path. In the example above, the path probability that the instance $G = \{10\ years, single, male, 18\ year\ old\}$ follows $path_{11}$ is 0.21. Figure 3 provides the path probabilities for this instance and all the possible paths in the tree.

In order to classify the stored instance in a particular class, we consider the path probabilities for each of the classes. In Figure 3 the stored instance will be classified in the income class *low* either if it follows $path_{11}$ (i.e. a probability of 0.21) or if it follows $path_{12} = \{\in [11,15]\ years, female, low\}$ (i.e. a probability

of 0). To determine the probability with which a given instance belongs to a class, we sum the probabilities of all the paths leading to this class (e.g. in Figure 3 for the class "low" 0.21+0) and call this *class* probability. Note that summation is possible because the splitting subsets are disjoint sets. Finally, the stored data instance is classified in the class with the highest class probability. In Figure 3 the instance $G = \{10\ years, single, male, 18\ year\ old\}$ is assigned to the class *middle* with a class probability of 0.79.

Classifying the instance in the class with the highest class probability corresponds to the majority vote from the literature concentrating on the combination of multiple classifiers (Fred 2001; Kittler et al. 1998; Kuncheva 2004; Seewald et al. 2001). The idea is that the different paths an instance could follow[xv] represent the different classifiers and the class of a path and the path probability stand for the result and the probability of each of the classifiers, respectively. According to the majority vote, the instance is assigned to the class with the highest class probability among the classifiers. In such cases ties are either resolved arbitrarily (Kuncheva 2004; Street and Kim 2001) or based on the prior probability of the particular class, which is the number of instances in the training set that belong to this class (Seewald et al. 2001). In case the prior probabilities are also equal, then the class is chosen arbitrarily. We resolve ties arbitrarily, as we believe that it is important that our method remains independent of the particular training set, not least due to efficiency reasons. This completes the description of the *second* phase of our approach, which is based on the ideas in the uncertain data mining literature. In Figure 4 the whole approach is summarized.

Phase I: Derivation of the node probabilities

1. For each splitting subset $d_i^{j*(k)}, i \in \{1, \dots, n\}, k \in \{1, \dots, t\}, j* \in \{1, \dots, m\}$ of the tree
   i. Based on historical data, determine the node probability $p_i^{j*(k)}$ that an attribute value which belonged to $d_i^{j*(k)}$ at the time of storage still belongs to it at the time of classification
   ii. For all $d_i^{j(k)}$ with $j \neq j*$ determine the node probabilities $p_i^{j(k)}$ that an attribute value which belonged to $d_i^{j*(k)}$ at the time of storage belongs to $d_i^{j(k)}$ at the time of classification
   iii. Store the node probabilities $p_i^{j(k)}, j \in \{1, \dots, m\}$ from 1.i or 1.ii[xvi] with $d_i^{j*(k)}$
2. For each stored instance $G = \{s_1, \dots, s_n\}$
   i. For each stored attribute value $s_i, i \in \{1, \dots, n\}$
      a. Identify the storage splitting subset(s) $d_i^{j*(k)}$ s.t. $s_i \in d_i^{j*(k)}$
      b. For $d_i^{j*(k)}$ from 2.i.a.[xvii] assign to the splitting subsets $d_i^{j(k)}, j \in \{1, \dots, m\}$ the probabilities $p_i^{j(k)}, j \in \{1, \dots, m\}$ stored with $d_i^{j*(k)}$ in 1.iii

Phase II: Classification of the stored instance

1. For each path $path_{lu} = \left\{d_{i1}^{j(1)}, d_{i2}^{j(2)}, \dots, d_{it}^{j(t)}, c_l\right\}, \{i1, \dots it\} \subseteq \{1, .., n\}, l \in \{1, \dots, p\}$[xviii] in the tree with the corresponding probabilities $\left\{p_{i1}^{j(1)}, p_{i2}^{j(2)}, \dots, p_{it}^{j(t)}\right\}, \{i1, \dots it\} \subseteq \{1, .. n\}$ derived in Step I.2.i.b[xix] determine the path probability $prob_{lu} = \prod_{k=1}^{t} p_{lk}^{j(k)}$
2. For each class of the dependent attribute $c_l, l \in \{1, \dots, p\}$, sum the probabilities for the paths of the type $path_{lu}$ to determine the class probability
3. Classify the data instance in the class $c_l, l \in \{1, \dots, p\}$ with the highest class probability (in case of a tie choose the class randomly)

Figure 4. A Two-Phase Method for Incorporating Currency in Decision Trees

In order to demonstrate the efficiency of our approach, we compare its complexity to the complexity of the approach presented in the second section. It is enough to compare only the complexities of the calculation

of the node probabilities (i.e. Step 1 of Phase I) as we assume that in the other steps the two approaches are identical[6]. We calculate the complexity for our approach based on Figure 4. The complexity for Steps 1.i-1.iii depends on the available historical data. If the historical dataset consists of N instances, for each $d_i^{j*(k)}$ the algorithm will take $O(N)$ for each splitting subset. Thus, Step 1 requires $O(N\ n\ MAXk^2)$[xx], where $MAXk$ represents the maximal number of splitting subsets for a given attribute and n stands for the number of attributes. In the approach from the second section, for each attribute the probabilities for all possible stored attribute values need to be determined. This results in a complexity in $O(N\ n\ MAXv\ MAXk)$, where $MAXv$ represents the maximal number of different stored independent attribute values for a given attribute. Already here the complexity of our approach is lower or equal (in the worst case) to that[xxi] in the second section, because $MAXk \leq MAXv$, especially for numeric attributes. This proves the efficiency of our approach. In the next subsection we show how the node probabilities can be more precisely and context-specifically modeled based on supplemental data.

Supplemental Data

In the approach presented above, the node probabilities are determined for each splitting subset without considering the values of the other attributes. This implies, for example in Figure 3, for the splitting subset *single* and for the instance $G = \{10\ years, single, male, 18\ year\ old\}$ that the probability for a person staying single in the time span between storage and classification is independent of supplemental data such as the years of education, gender, and age of this person. However, as mentioned in the second section, the literature has shown that considering supplemental data can lead to a more precise, context-specific estimation. Thus, it is reasonable to modify the derivation of the node probabilities for a given instance so that not only $d_i^{j*(k)}$ is considered, but also the other stored values. Based on the discussion above, for a given stored instance $G = \{s_1, \dots, s_n\}$ we would derive $p_i^{j(k)}$ as the probability that s_i belongs to $d_i^{j(k)}$ at the time of classification provided that the values of the other attributes were $\{s_1, \dots s_{i-1}, s_{i+1}, \dots s_n\}$ at the time of storage.

This approach considers high amount of additional information in the derivation of the probabilities, but it is again very impractical in the context of big data since very detailed historical data is required to determine all the probabilities. For example, for the instance $G = \{10\ years, single, male, 18\ year\ old\}$ and the attribute value *single*, we need historical data on the males, who were 18-year-old, single, and had ten years of education at the time of storage, and who got married until the time of classification. To avoid this problem and make the approach better applicable in the context of big data, we consider as supplemental data only the stored attribute values which *precede* the stored value in the path. Moreover, we do not calculate the probabilities based on single values such as *18 year old*, but rather use the corresponding splitting subset of the tree (e.g. ≤ 30 year old). Thus, we still determine the node probabilities $p_i^{j(k)}$ based on supplemental data, but with fewer and less granular conditions. For example, for $G = \{10\ years, single, male, 18\ year\ old\}$ and the stored attribute value *single*, the node probability $p_2^{1(2)}$ for the path $path_{11} = \{\leq 10\ years, single, low\}$ will be derived by considering all singles who had less or equal to 10 years of education at the time of storage without posing any additional restrictions on their gender or age.

This approach is very reasonable in the context of decision trees, since the order of the independent attributes represents their importance with respect to the classification of the dependent attribute. The root is the most important attribute and with increasing depth the importance decreases. Moreover, a path in the tree is also a sequence of conditions, where each consecutive condition is considered, only if the preceding one is fulfilled and regardless of the stored attribute values, which are not part of the path. For example, for $G = \{10\ years, single, male, 18\ year\ old\}$ and the path $path_{11} = \{\leq 10\ years, single, low\}$, the fact that the person is an 18-year-old male is not relevant for the classification and would thus not influence the result when incorporating currency.

---

[6] The approach in the second section was not developed for the application to decision trees. This is an assumption we make to be able to compare the two approaches.

In some cases, due to missing historical data, it may happen that a given node probability is zero. This is especially the case when supplemental data is used, as detailed historical data is needed, which is not always available in reality. The problem is then that these zero probabilities are propagated down the tree and no instance is classified in the corresponding class. To solve this "zero-frequency-problem" we apply a smoothing technique by transforming the zero probabilities with a Laplace transformation. This approach is often used in data mining for such situations (Witten and Frank 2005).

To sum up, in this section we[xxii] presented our novel approach for considering currency in decision trees. The presented two-phase algorithm is interpretable, universal, context-specific, and efficient. It is interpretable, because currency is interpreted as probability. In addition, it is universal, because it is independent of the particular decision tree classifier. Moreover, it is context-specific, because both determining the node probabilities and considering supplemental data strongly depend on the stored instance and on the structure of the tree. Finally and most importantly in the context of big data, it is efficient because 1) it only modifies the process of classifying stored instances and not the decision tree algorithm itself, 2) it determines the node probabilities based on the set of splitting subsets and not on a single value and 3) it incorporates supplemental data based on the tree structure and not on all the stored values. In the next section our approach is evaluated with three real-world datasets.

Evaluation

In this section we evaluate our approach by applying it to three different datasets for the two applications presented in the introduction. The first dataset consists of the publicly available panel data from the SOEP (2012)[7] where for each year the personal characteristics (age, gender, marital status, education, industry, employment status, annual income, etc.) of the same individuals are stored. This dataset represents the CRM[xxiii] scenario, in which a financial service provider conducts a campaign to attract new customers based on their annual income. It derives the relationship between the annual income of the customers and their personal characteristics, based on an up-to-date database of existing customers, in the form of a decision tree. Since the CRM campaign took place some time ago, some of the stored personal characteristics of the people who took part in it may be outdated resulting in a wrong income class. Since the company targets the customers according to their income, this may lead to the wrong product being offered and thus economic losses.

The second dataset is from the UCI Machine Learning Repository (Bache and Lichman 2014), which is a source commonly used in the big data mining literature. The dataset is called "Localization Data for Posture Reconstruction" and contains the results from sensors attached to four different parts of the body of individuals who were performing different activities in a number of sessions. For each sensor, its x, y and z coordinates are given. This dataset stands for the scenario mentioned in the introduction for handicapped individuals, where the aim is to determine the current activity of the patient based on the measurements coming from the sensors. The data is used to build the relationship between the sensors and their values on the one side and the current activity on the other side in the form of the decision tree. Based on this tree, instances stored some time ago (this time span can also be as small as a couple of minutes) are classified to determine the current activity of the individuals. If the currency of the sensor values is not taken into account, wrong activities may result causing the wrong responses, which may have devastating consequences.

Finally, the third dataset was generated by a mobile application, developed for the purpose of this paper and can be provided upon request. This dataset also represents the scenario for handicapped individuals, but as opposed to the second one, it contains the values of all the sensors of a mobile device, thus extending the scenario by considering the development in mobile technologies. Nowadays, many companies develop mobile applications that are applied for activity recognition[8]. In the case of elderly and handicapped care, such applications can increase the probability of an appropriate response and also result in higher acceptance as it is enough to have a mobile device and not many different sensors

[7] The data used in this publication was made available to us by the German Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin.

[8] E.g. https://actitracker.com/

attached to the body. Thus, in this scenario, the relationship between the current activity of a person and the values from the sensors delivered from the mobile application is examined in the form of a decision tree. As for the second dataset, the stored sensor values may be outdated (e.g. due to wrong transmission, sensor error, etc.) resulting in the wrong activity and thus a wrong response. In the next subsection we provide the evaluation procedure we follow to demonstrate the advantages of our approach and describe the three datasets in detail.

Data and Evaluation Procedure

In order to evaluate our approach we define for each of the three datasets a point in time in which the data was stored (t_o) and another, later point in time in which the classification takes place (t_1). Then each of the three datasets is randomly divided into three *mutually exclusive* subsets, which are then used to derive the *training and test* set (approx. 50% of the instances), the *currency* set (approx. 25% of the instances) and the *validation* set (approx. 25% of the instances).

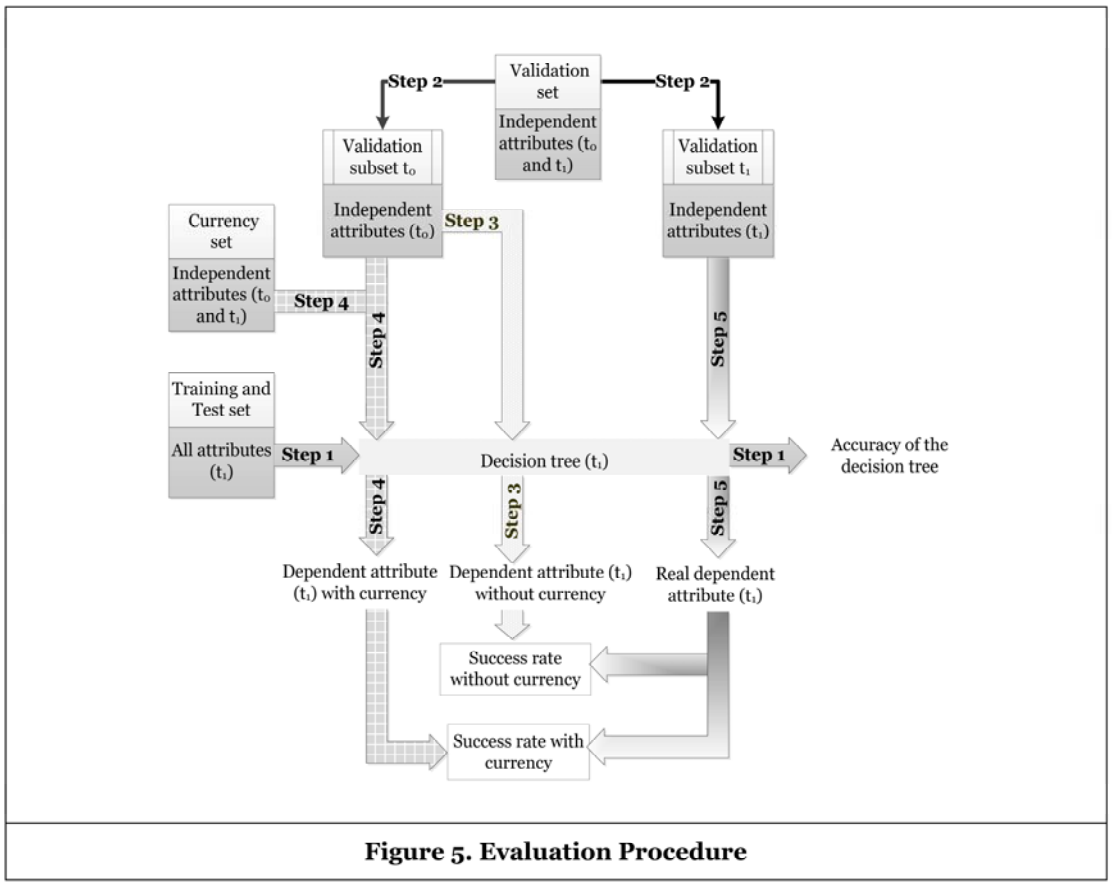


Figure 5. Evaluation Procedure

The evaluation procedure is presented in Figure 5. In **Step 1** the decision tree is built and its accuracy is tested based on the *training and test* set. This set contains the values of the independent attributes and of the dependent attribute of the instances for t_1 and represents the real-world values at the time of classification. In **Step 2** the *validation* set, which contains the values for the independent attributes of the individuals for both t_o and t_1, is divided into two *validation* subsets according to the point of acquisition. In **Step 3** the decision tree is applied to the *validation* subset for t_o and the result is the class of the dependent attribute based on the values of the independent attributes for t_o and the decision tree for t_1. In **Step 4** the *currency* set is used to determine the node probabilities for the independent attribute values

in t_1, given the stored independent attribute values in t_0. This set contains the values for the independent attributes for both t_0 and t_1. The node probabilities are used in this step to classify the stored instances from the *validation* subset for t_0 in the decision tree for t_1. The result is the class of the dependent attribute based on the decision tree for t_1 and the values of the independent attributes in t_0 after considering the currency of the stored attributes. To determine the success rate of our approach, in **Step 5** the decision tree is applied to the *validation* subset for t_1 and the resulting classes (corresponding to the real-world classes of the independent attribute in t_1) are compared with the classes from **Step 3** and **Step 4**. We additionally restrict the validation and the currency sets to instances, for which the attribute value corresponding to the root of the tree changed,[xxiv] to demonstrate our approach. Note that in the case when currency is considered, we additionally examine the results for incorporating supplemental data. In the following we describe the three datasets in detail.

The first dataset is the panel data SOEP (2012) from which we consider t_0=2001 and t_1=2011, as 2011 is the last year in[xxv] the dataset. We filter only the individuals that are contained in both datasets resulting in 9522 instances. The dependent attribute[xxvi] for this dataset is the annual income of the individuals and the independent attributes are: *In Training,* which shows if the individual is currently in training; *Employment*, which gives the employment status of the individual; *Education*, which stands for the number of years of education; *Marital Status*; *Year of Birth*; and *Gender*. Since the dependent attribute for the decision tree is numerical[xxvii], we categorize the annual income in classes with the software tool IBM Corp. Released (2011) so that each class contains an approximately equal part of the individuals in the panel data. Note that if this is not done, as mentioned above, the result will not be a distribution over the classes, but rather a distribution over the distributions of the income in the different leaves. Thus determining the final class will not be possible, which is crucial in real-world applications.

The second dataset is the "Localization Data for Posture Reconstruction" consisting of 164860 instances, where each instance consists of the person who was measured (5 people), the session (5 sessions per person), the type of sensor (4 sensors), the time of measurement, a unique timestamp, the sensor's three coordinates, and the current activity of the individual (4 activities). The independent attributes are *Person*, *Sensor*, *xCoordinate*, *yCoordinate* and *zCoordinate*. The dependent attribute is *Activity*. t_0 was chosen to represent the first two sessions, while t_1 was chosen to represent the second two sessions.

The third dataset was gathered with a mobile device with an Android 4.4.2 operating system and a Quad-Core Qualcomm Snapdragon 600 processor. The dataset consists of 1839062 instances in two sessions (for t_0 and t_1). The dependent attribute is again *Activity* (4 classes based on Wu et al. 2011) and the independent attributes are the values of the sensors of the device (18 sensors). Most sensors had three values and thus for a better comparison to the second dataset, we omitted the sensors with less or more than three values. In the next subsection we present the decision trees[xxviii].

Decision Trees

We applied to the training and test set of each of the three datasets above both the CHAID and the C4.5 algorithm (Step 1 in Figure 5). The CHAID is based on the chi-squared test of independence and results in non-binary trees (Kass 1980). The C4.5 is based on an information gain measure and also results in non-binary trees (Quinlan 1993). We used IBM Corp. Released (2011) with default options and an adjustable depth (depending on the dataset) for the CHAID algorithm, and WEKA (Hall et al. 2009) with reduced-error pruning and a minimum number of instances (depending on the dataset) for C4.5. For CHAID all the p-values were significant at the 5%-level. We tested the models with both a randomly chosen test set (33 %) and a cross validation (10-fold). The accuracy of the decision trees is presented in Table 1. We can see that depending on the dataset, the method, and the validation, the results may differ, but generally the accuracies are reliable enough for the models to be considered for further analysis.

Table 1. Accuracy of the Decision Trees

Dataset/Method	CHAID a	CHAID b	C4.5 a	C4.5 b
1	71%	74%	73%	74%
2	65%	67%	71%	73%
3	75%	77%	77%	78%

Legend: *a = Test set, b = Cross validation*

Decision Analytics, Big Data, and Visualization

Performance of the New Approach

As presented in Figure 5, to evaluate our approach we conduct Steps 3-5 based on the models from Table 1. The results are presented in Table 2. As we can see, considering currency leads always to higher success rate than not doing so and also considering supplemental data can additionally increase the success rate. The success rate of supplemental data would be even higher, if the trees contained attributes that strongly changed according to the additional information, which is not the case in the models in Table 1.

Table 2. Comparison of the Success Rates of the Approaches

Dataset/Method	CHAID i	CHAID ii	CHAID iii	C4.5 i	C4.5 ii	C4.5 iii
1 a	6%	33%	33%	5%	31%	31%
1 b	7%	29%	29%	7%	29%	29%
2 a	33%	59%	60%	30%	54%	60%
2 b	32%	62%	62%	30%	44%	48%
3 a	45%	63%	63%	51%	73%	73%
3 b	46%	64%	65%	37%	53%	53%

Legend: *a = Test set, b = Cross validation,*
i =without currency, ii =with currency, iii =with currency and supplemental data

Table 3. Runtimes of the Approaches in milliseconds

Method/Dataset	1a(M1 /M2)	1b(M1 /M2)	2a(M1 /M2)	2b(M1 /M2)	3a(M1 /M2)	3b(M1 /M2)
CHAID I ii	47/ 78	47/ 94	281/ 280	297/ 577	5546/ 10358	6156/ 11840
CHAID II ii	47/ 78	47/ 78	218/ 250	203/ 296	5437/ 10093	6062/ 11606
C4.5 I ii	47/ 109	62/ 109	187/ 218	203/ 296	1484/ 2153	1797/ 3183
C4.5 II ii	47/ 78	63/ 109	172/ 203	172/ 296	1312/ 2121	1907/ 3135
CHAID I iii	62/ 94	78/ 171	875/ 1689	972/ 1873	100159/ 443448	102324/ 459578
CHAID II iii	62/ 94	62/ 109	844/ 1436	953/ 1592	99273/ 441532	101944/ 456993
C4.5 I iii	63/ 109	156/ 250	33797/ 68267	61817/ 159574	677299/ 2484836	748587/ 3192598
C4.5 II iii	46/ 78	156/ 128	33344/ 66785	61750/ 128992	602771/ 2479723	717455/ 3141878

Legend: *a = Test set, b = Cross validation, ii =with currency, iii =with currency and supplemental data, I = approach from the second section, II= our approach*

[xxix]Finally, to demonstrate the efficiency of our approach, we measured the runtimes of computing the node probabilities with our approach and with the approach presented in the second section. Note that, as mentioned above, to conduct this comparison certain assumptions need to be made and the approach of

the second section needs to be additionally modified. We measured the runtimes on two different machines (M1 with Intel(R) Core™ i5-2520M CPU @ 2.50GHz, 4.00 GB RAM and M2 with Intel(R) Core™2 Duo CPU T6570 @ 2.10GHz, 4.00 GB RAM) to test the reliability of the results. As we can see and as shown in the third section, our approach is almost[9xxx] always faster than the one presented in the second section which proves its efficiency.

Availability of an Up-to-Date Training and Test Set

In our approach and thus in the evaluation above, we implicitly assume the existence of an up-to-date Training and Test set for building a decision tree that describes the current relationships. If this is not the case and there is a concept drift in the data, then the relationships presented in the tree may be outdated. Such a situation may occur when in the CRM-scenario the company does not possess the up-to-date information of its current customers, as a customer is not obliged to inform the insurer regarding all changes in his/her personal characteristics. It may also occur if there is no up-to-date available data about the sensor measurements and the corresponding activities in the handicapped-people-scenario. In such a situation, the approaches from the uncertain big data mining literature presented above (e.g. Tsang et al. 2011) can be applied to modify the decision tree. This will reduce the accuracy and increase the runtime of the approach with increasing negative effects for more outdated data. However, the results will still be better than not considering currency as the decision tree in both cases is the same and only the classification of stored instances is determined by the two approaches.

Conclusion

In this paper we present an approach for considering currency in the decision tree classification method in the context of big data. Our idea is based both on the literature for modelling currency and on the one for uncertain big data mining in decision trees. Based on the first stream of research, we demonstrate how currency can be derived for the consideration in decision trees in an efficient and context-specific way. In particular, we show how the probability, which commonly represents currency in the literature, can be derived with little historical data and depending on the structure of the decision tree. This makes our approach suitable for the context of big data. In addition, we demonstrate how this probability can be refined through the use of supplemental data, but again in an efficient and decision-tree-specific way. Based on these considerations, we determine the node probabilities for a given stored instance and a decision tree and apply the ideas from the uncertain data mining literature to classify these stored instances. We extend this stream of research by showing how the uncertainty, which is assumed to be given by the authors there, can be derived in an interpretable, efficient and context-specific way, where the interpretation is given by the currency of stored data. The applicability and the contribution of our approach are demonstrated based on three different datasets, two of which stemming from the context of big data. The results demonstrate that our approach leads to a substantial improvement in the classification accuracy as opposed to not considering currency and also that it is more efficient than the approach presented in the second section.

Our method also has some limitations. First of all, it is developed for decision trees. However, the problem of input data of low quality is just as relevant for other data mining methods such as clustering, for example. The idea to develop an approach based on the information quality and the uncertain data mining literature can be applied there in a similar fashion. In addition, in some cases (especially with supplemental data) historical data may not be provided. An alternative idea would be to consider expert estimations instead. These can be modeled with fuzzy set theory based on fuzzy decision trees (Yuan and Shaw 1995) and fuzzy metrics for currency (Heinrich and Hristova 2014). Moreover, in the future our approach can be additionally optimized by using new methods such as the MapReduce framework (Aggarwal 2013) for even more efficient application in the context of big data. Finally, applying the approach to other types of big data such as data streams or unstructured data, for example, will be an additional challenge for future research.

---

[9] The runtime measurement in Java is not always stable. Thus, in the future this analysis must be conducted again and the average runtime must be taken.

References

Aggarwal, C. C., and Yu, P. S. 2009. "A survey of uncertain data algorithms and applications," *IEEE Transactions on Knowledge and Data Engineering* (21:5), pp. 609-623.

Aggarwal, C. C., Ashish, N., and Sheth, A. 2013. "The internet of things: A survey from the data-centric perspective," in *Managing and mining sensor data*: Springer, pp. 383–428.

Aggarwal, C. C. 2013. *Managing and mining sensor data*, New York: Springer

Alpar, P., and Winkelsträter, S. 2014. "Assessment of data quality in accounting data with association rules," *Expert Systems with Applications* (41:5), pp. 2259–2268.

Azar, A. T., and El-Metwally, S. M. 2013. "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications* (23:7-8), pp. 2387-2403.

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

Bache, K. and Lichman, M. 2014. "*UCI Machine Learning Repository*," http://archive.ics.uci.edu/ml, Irvine, CA: University of California, School of Information and Computer Science.

Berger, J. O. 2010. *Statistical Decision Theory and Bayesian Analysis*, New York, NY: Springer.

Blake, R., and Mangiameli, P. 2011. "The Effects and Interactions of Data Quality and Problem Complexity on Classification," *Journal of Data and Information Quality* (2:2), pp. 1–28.

Chau M., Cheng R., Kao B., and Ng, J. 2006. „Uncertain Data Mining: An Example in Clustering Location Data," in *Advances in Knowledge Discovery and Data Mining*, Ng W., Kitsuregawa M., Li J. and Chang K. (eds), vol 3918. Springer Berlin Heidelberg, pp 199-204.

Dasu, T., and Johnson, T. 2003. *Exploratory data mining and data cleaning*, Wiley Online Library.

Domingos, P., and Hulten G. 2000. "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*, ACM, New York, NY, USA, pp. 71-80.[xxxi]

Economist Intelligence Unit 2011. *Levelling the Playing Field: How Companies Use Data for Competitive Advantage, http://fm.sap.com/images/kern/assets/sap_EIU_Levelling_The_Playing_Field.pdf, accessed on 01.06.2015.*[xxxii]

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

Experian QAS 2013. The Data Advantage: How accuracy creates opportunity http://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf, *accessed on 01.06.2015.*[xxxiii]

Fan, W. 2013. "Querying Big Social Data," in *Big Data*, G. Gottlob, G. Grasso, D. Olteanu, and C. Schallhart (eds.): Springer Berlin Heidelberg, pp. 14-28.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From data mining to knowledge discovery in databases," *AI magazine* (17:3), pp. 37-54[xxxiv].

Federal Bureau of Statistics 2013. *Annual Abstract of Statistics (in German).*

Fisher, C. W., Lauria, E. J., and Matheus, C. C. 2009. "An accuracy metric: Percentages, randomness and probabilities," *Journal of Data and Information Quality (JDIQ)* (1:3), No[xxxv]. 16.

Forbes, I. 2010. *Managing Information in the Enterprise: Perspectives for Business Leaders, http://images.forbes.com/forbesinsights/StudyPDFs/SAP_InformationManagement_04_2010.pdf, accessed on 01.06.2015.*[xxxvi]

Fred, A. 2001. "Finding consistent clusters in data partitions," in *Multiple classifier systems, Kittler, J. and Roli, F. (Eds.)*: Springer, pp. 309–318.[xxxvii]

Giudici, P., and Figini, S. 2009. *Applied Data Mining for Business and Industry*. John Wiley and Sons.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. 2009. „The WEKA Data Mining Software: An Update," *SIGKDD Explorations* (11:1), pp. 10-18.[xxxviii]

Hashemi, S., and Yang, Y. 2009. "Flexible decision tree for data stream classification in the presence of concept change, noise and missing values," *Data Mining and Knowledge Discovery* (19:1), pp. 95-131.

Hawarah L., Simonet A., Simonet M. 2009. "Dealing with Missing Values in a Probabilistic Decision Tree during Classification," in *Mining Complex Data*, Zighed D., Tsumoto S., Ras Z., Hacid H. (eds), vol. 165. Springer Berlin Heidelberg, pp. 55-74.

Heinrich, B., and Klier, M. 2009. „A Novel Data Quality Metric for Timeliness considering Supplemental Data," in *Proceedings of the 17th European Conference on Information Systems*, University of Verona, pp. 2701-2713.

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

Heinrich, B. and Hristova, D. 2014. „A Fuzzy Metric for Currency in the Context of Big Data," in *Proceedings of the 22nd European Conference on Information Systems*, Tel Aviv, Israel.

Hems, A., Soofi, A., and Perez, E. 2013. *How innovative oil and gas companies are using big data to outmaneuver the competition*. A Microsoft White Paper.

Hulten, G., Spencer, L. and Domingos, P. 2001. "Mining time-changing data streams," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 97-107[xxxix].

IBM Corp. Released 2011. *IBM SPSS Statistics for Windows. Version 20.0*. Armonk, NY: IBM Corp.

IBM Institute for Business Value 2012. *Analytics: The real-world use of big data, How innovative enterprises extract value from uncertain data, http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf, accessed on 01.06.2015.*[xl]

Kass, G. V. 1980. "An exploratory technique for investigating large quantities of categorical data," *Applied statistics*, pp. 119-127.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. 1998. "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (20:3), pp. 226–239.

Kuncheva, L. I. 2004. *Combining pattern classifiers: methods and algorithms*: John Wiley & Sons.

Koh, H. C., Tan, W. C., and Goh, C. P. 2006. "A two-step method to construct credit scoring models with data mining techniques," *International Journal of Business and Information* (1:1), pp. 96-118.

Leung, C. K.-S., Mateo, M. A. F., and Brajczuk, D. A. 2008. "A tree-based approach for frequent pattern mining from uncertain data," in *Advances in Knowledge Discovery and Data Mining*: Springer, pp. 653–661.

Leung, C.-S., and Hayduk, Y. 2013. "Mining Frequent Patterns from Uncertain Data with MapReduce for Big Data Analytics," in *Database Systems for Advanced Applications*, W. Meng, L. Feng, S. Bressan, W. Winiwarter, and W. Song (eds.): Springer Berlin Heidelberg, pp. 440-455.

Li, F., Nastic, S., and Dustdar, S. 2012. "*Data Quality Observation in Pervasive Environments,*" in *Proceedings of the IEEE 15th International Conference on Computational Science and Engineering (CSE)*, 5-7 Dec. 2012, Nicosia, pp. 602-609.

Liang, C., Zhang, Y., and Song, Q. 2010. "Decision Tree for Dynamic and Uncertain Data Streams," *Journal of Machine Learning Research-Proceedings Track* (13), pp. 209-224[xli].

Magnani, M., and Montesi, D. 2010. "Uncertainty in Decision Tree Classifiers," in *Scalable Uncertainty Management*, Lecture Notes in Computer Science, Vol. 6379, pp. 250-263.

Mezzanzanica, M. Boselli, R., Cesarini, M., and Mercorio, F. 2012. "Towards the use of Model Checking for performing Data Consistency Evaluation and Cleansing," in *Proceedings of the 17th International Conference on Information Quality (ICIQ 2012)*.

Miller, I., Miller, M., and Freund, J. E. 2004. *John E. Freund's mathematical statistics with applications*, Upper Saddle River, NJ: Prentice Hall.

Ngai, E. W., Xiu, L., and Chau, D. C. 2009. "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications* (36:2), pp. 2592-2602.

Orr, K. 1998. "Data quality and systems theory," *Communications of the ACM* (41:2), pp. 66-71.

Pathak, A. N., Sehgal, M., and Christopher, D. 2011. "A Study on Fraud Detection Based on Data Mining Using Decision Tree," *International Journal of Computer Science* (8:3), pp. 258-261.

Qin, B., Xia, Y., and Li, F. 2009. "DTU: a decision tree for uncertain data," in *Advances in Knowledge Discovery and Data Mining*, Theeramunkong, T., Kijsirikul, B., Cercone, N. and Ho, T. (eds), Springer, pp. 4-15.

Quinlan, J. R. 1993. *C4. 5: programs for machine learning*: Morgan Kaufmann.

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

Seewald, A. K., and Fürnkranz, J. 2001. "An evaluation of grading classifiers," in *Advances in Intelligent Data Analysis*: Springer, pp. 115–124.

SOEP 2012. *Socio-Economic Panel Study (SOEP), Data for the years 1984-2011, Version 28*. doi:10.5684/soep.v28.

Street, W. N., and Kim, Y. 2001. "A streaming ensemble algorithm (SEA) for large-scale classification," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 377–382.

Tsang, S., Kao, B., Yip, K. Y., Ho, W.-S. and Lee, S. D. 2011. "Decision trees for uncertain data," *IEEE Transactions on Knowledge and Data Engineering* (23:1), pp. 64-78.

Vazirigiannis, M., Halkidi, M., and Gunopulos, D. 2003. *Uncertainty handling and quality assessment in data mining*, Springer.

Wang, H., Fan, W., Yu, P. S., and Han, J. 2003. "Mining concept-drifting data streams using ensemble classifiers," in *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, ACM, New York, NY, USA, pp. 226-235.[xlii]

Wang, R. Y., and Strong, D. M. 1996. "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems* (12:4), pp. 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.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Wu, J., Jiang, C., Houston, D., Baker, D., and Delfino, R. 2011. "Automated time activity classification based on global positioning system (GPS) tracking data," *Environ Health* (10), p. 1-13[xliii].

Yang, H., and Fong, S. 2011. "Moderated VFDT in Stream Mining Using Adaptive Tie Threshold and Incremental Pruning," in *Data Warehousing and Knowledge Discovery*, A. Cuzzocrea, and U. Dayal (eds.): Springer Berlin Heidelberg, pp. 471-483.

Yang, H., Fong, S., Sun, G., and Wong, R. 2012. "A Very Fast Decision Tree Algorithm for Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network," *International Journal of Distributed Sensor Networks* (2012), pp. 1–16.

Yang, H. 2013. "Solving Problems of Imperfect Data Streams by Incremental Decision Trees," *Journal of Emerging Technologies in Web Intelligence* (5:3), pp. 322-331[xliv].

Yuan, Y., and Shaw, M. J. 1995. "Induction of fuzzy decision trees," *Fuzzy Sets and Systems* (69:2), pp. 125-139.

Yue, D., Wu, X., Y, W., Li, Y., and Chu, C-H. 2007. "A review of data mining-based financial fraud detection research," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, WiCom 2007, pp. 5519-5522.

Zhu, X., and Wu, X. 2004. "Class Noise vs. Attribute Noise: A Quantitative Study," *Artificial Intelligence Review* (22:3), pp. 177-210.

Appendix A: Post Publication Changes

In this section the changes in the paper which were made after its publication are listed. The reasons for these changes are either to achieve consistency with the rest of the dissertation or to correct typos or other mistakes in the paper.

---

[i] In the whole paper "data quality" was changed to "information quality" for consistency reasons.

[ii] In the whole paper "volume", "velocity", and "variety" were changed to "Volume", "Velocity", and "Variety" for consistency reasons.

[iii] In the whole paper "real world" when used as adjective was changed to "real-world" for consistency reasons.

[iv] "to closc" was replaced by "at closing".

[v] (CRM) was added to be used later.

[vi] A comma was removed.

[vii] A space was added here.

[viii] The reference was added here.

[ix] "on" was added here.

[x] Changed from "stationary" to "stationarity" due to a typo.

[xi] A comma was removed here.

[xii] A comma was added here.

[xiii] The footnote was added for clarification.

[xiv] A space was added here.

[xv] The commas on the two sides of the expressions were removed.

[xvi] Changed from "1.a or 1.b" to "1.i or 1.ii" due to inconsistencies in the numbering.

[xvii] Changed from "a." to "2.i.a" for consistency. Also "each" was removed.

[xviii] A comma was added here.

[xix] Changed from "I.1.d" to "I.2.i.b" due to inconsistencies in the numbering.

[xx] Spaces were added here for better readability.

[xxi] "that" was added as it was missing from the text.

[xxii] "have" was removed as it was unnecessary.

[xxiii] Changed from "Customer Relationship Management" to "CRM".

[xxiv] A comma was inserted.

[xxv] "in" was added as it was missing.

[xxvi] In the whole paper "(in)dependent variable" was replaced by "(in)dependent attribute" for consistency reasons.

[xxvii] "categorical" was replaced by "numerical" as it was incorrect.

[xxviii] "tree" was changed to "trees" as it was a typo.

[xxix] The text was moved here for better visualisation.

[xxx] The word "almost" was added here for clarification. In addition, a footnote was added.

[xxxi] The reference was changed for consistency reasons.

[xxxii] A link was added here.

[xxxiii] A link was added here.

[xxxiv] A page range was added.

[xxxv] „pp" was replaced by „No".

[xxxvi] A link was added here.

[xxxvii] Editors' names were added here.

[xxxviii] A space was removed.

[xxxix] The page range was corrected.

[xl] A link was added here.

[xli] The page range was corrected.

[xlii] The reference was changed for consistency reasons.

[xliii] The page range was corrected.

[xliv] The page range was corrected.

2.3 Paper 3: A Fuzzy Metric for Currency in the Context of Big Data

Full citation: Heinrich, B. and Hristova, D. (2014), *A Fuzzy Metric for Currency in the Context of Big Data*, in Proceedings of the 22nd European Conference on Information Systems (ECIS), 9-11 June, Tel Aviv, Israel.

Status: accepted on 09.04.2014

Highlights: In this paper a metric for currency is developed which is initialized with expert estimations rather than with historical data. The metric is modeled based on fuzzy set theory in the form of a FIS, following the guidelines from the literature. The contribution of the approach is twofold. On the one hand, it facilitates the measurement of currency without requiring detailed historical data, which, as mentioned above, is especially important in the context of big data. On the other hand, as opposed to existing qualitative approaches (cf. Subsection 1.3.3), this metric quantifies expert estimations in a well-founded and natural way by using linguistic variables. As a result, it is not only expected to be more efficient than them (i.e. due to a calculation of the metric values in an automated way), but also less biased.

Evaluation: The approach is evaluated by initializing the metric for the attribute value "single" of the attribute marital status and by using the age of a person as supplemental data. A questionnaire was developed following the guidelines from the literature, which was then answered by 9 experts from the field of information quality with an average experience of 12 years. The results show that the age of a person plays an important role for the currency of the attribute value "single" and also that some fuzzy sets may be easier to estimate than others.

Limitations: First, the metric has been developed by making particular assumptions regarding the elements of the FIS. Thus, it would be the task of future research to examine the necessity and also the restrictiveness of these assumptions. Second, similar to Paper 2, it would also be interesting to transfer the approach to other types of data addressing the Variety characteristic of big data. This is expected to bring great benefits, especially for unstructured data, where an expert estimation may be much better than the one based on historical data. Finally, it would also be the task of future research to compare the developed metric with the extended metric from Paper 1 (both in terms of efficiency and precision) and also with other metrics from the literature.

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 modeled 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 initialized[i] and applied. The presented approach serves as a first step in modeling expert estimations as input to information quality[ii] metrics in a well-defined and structured way.

Keywords: Information 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 analyze 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 characterized according to the three "V"s: Volume, Velocity and Variety[iii], where Volume[iv] stands for the increasing number of (Tera)Bytes of stored data, Velocity describes the increasing speed with which data is generated and analyzed 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 information quality (IQ^{v}). Poor IQ leads in many cases to wrong decisions and thus economic losses. According to a survey, conducted by Forbes (2010), poor IQ 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 IQ nowadays.

IQ 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 IQ 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 IQ dimensions (Redman, 1996), because most organizations 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 modeled 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 modeling expert estimations. Thereby, the age and the decline rate of an attribute value are modeled 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 modeling 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[vi] 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 modeled 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[vii], such as data from temperature sensors, smart meters, or car sensors[viii], 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[ix], 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 initialized 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 modeled 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 parametric[x] 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 modeling of expert estimations, Zadeh[xi] (1975) introduced the concept of a linguistic variable, which values are words or sentences rather than numbers (Aliev, 2013).

Definition 3 (Linguistic variable[xii]): A linguistic variable is characterized 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 modeled 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 initialized 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 modeling 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 neighbor linguistic terms):** Every two neighbor 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 neighbor 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 the[xiii] 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 4[xiv] (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:*

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

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

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

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 5[xvi] *(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:*[xvii]

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

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

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

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. $Rule\ 1$[xix]). For each combination of the input linguistic terms[xx], 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. $Rule\ 1$[xxi], for example, can be derived from it.

Age \ Decline rate	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 a[xxii] 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 normalized 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 6[xxiii] (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 4[xxiv]. 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 consists[xxv] 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 modeled 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 4[xxvi]) 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**[xxvii]. 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 0.43[xxviii]. 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 0.43 meaning that 43% 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 43 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 $a \Rightarrow_S b := \max(1-a, b), a, b \in [0,1]$, 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 5[xxix] 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 $\rightarrow_S$ is interpreted as the K-D implication above.

In addition, we replace $\rightarrow_S$ 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:

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

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

$(currency\ \mu_{\widetilde{neutral}})$

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 $(decline\ rate\ \mu_{\widetilde{decline}})$, where $\mu_{\widetilde{decline}}$ 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).

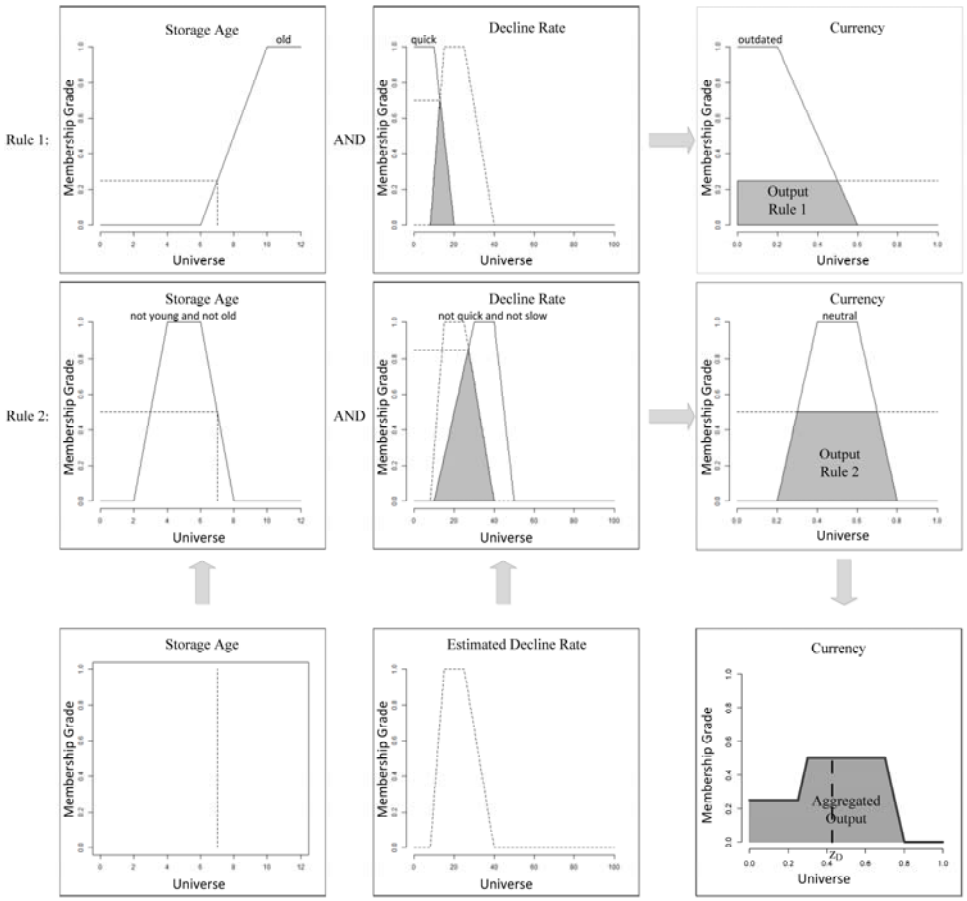


Figure 1. Example for estimating the output value of the fuzzy metric[xxx]

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 modeling 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 initializing 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 IQ 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 initialize 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 4 and 5[xxxi], 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 functions[xxxii]. 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 gradient search method[xxxiii]. 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 IQ.

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[xxxiv], 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 IQ. 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 characterized 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.

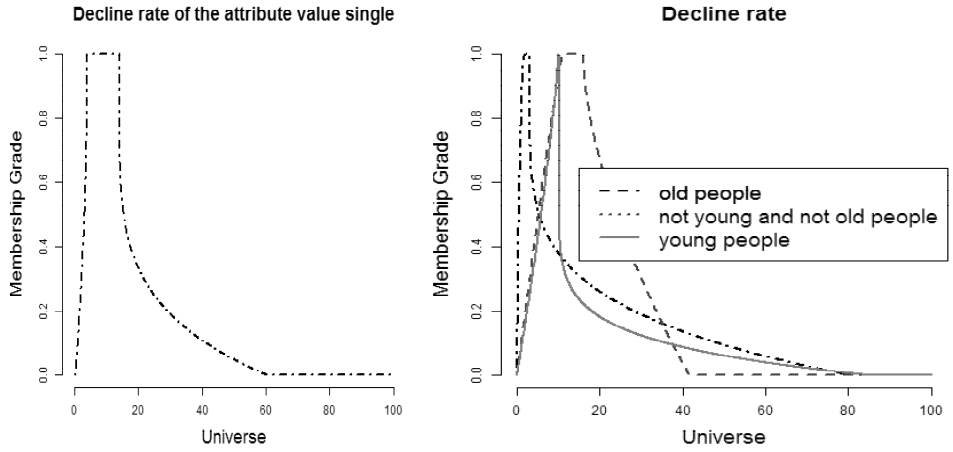


Figure 2. Decline rate of the attribute value single without and with considering additional data

5 Conclusion

In the current paper we present a fuzzy metric for currency, which is initialized with expert estimations in a mathematically well-founded way. The metric is modeled 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 In Dubois, D. and Prade, P. (Eds.): Fundamentals of Fuzzy Sets, vol. 7, 95-227, Springer.[xxxv]

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 Kaynak, O., Zadeh, L. Türkşen, B. and Rudas, I. (Eds.): Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications, vol. 162, 177-197, Springer. [xxxvi]

Cho, J. and Garcia-Molina, H. (2000). Synchronizing a database to improve freshness. In Proceedings of the ACM SIGMOD '00, 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, http://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf, accessed on 02.06.2015.[xxxvii]

Forbes Insights. (2010). Managing Information in the Enterprise: Perspectives for Business Leaders, http://images.forbes.com/forbesinsights/StudyPDFs/SAP_InformationManagement_04_2010.pdf, accessed on 02.06.2015.[xxxviii]

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, http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf, accessed on 02.06.2015.[xxxix]

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.

Li[xl], 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.

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, 60-68.[xli]

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.

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 ECIS 2013[xlii].

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)[xliii].

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. IEEE Software, 11(6), 48-56.[xliv]

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

Appendix A: Post Publication Changes

In this section the changes in the paper which were made after its publication are listed. The reasons for these changes are either to achieve consistency with the rest of the dissertation or to correct typos or other mistakes in the paper.

---

[i] In the whole paper British English words were changed to American English words for consistency reasons.
[ii] In the whole paper "data quality" was replaced by "information quality" for consistency reasons.
[iii] In the whole paper "volume", "velocity", and "variety" were changed to "Volume", "Velocity", and "Variety" for consistency reasons.
[iv] The quotation marks were removed here and also later for consistency reasons.
[v] "DQ" was changed in the whole paper to "IQ" for consistency reasons.
[vi] Hyphens were added here.
[vii] A comma was inserted here.
[viii] A comma was inserted here.
[ix] A comma was inserted here.
[x] "parameterized" was replaced by "parametric" here.
[xi] The reference has been updated in the reference list.
[xii] "Variable" was changed to "variable" for consistency reasons.
[xiii] "a" was replaced by "the" for consistency reasons.
[xiv] Changed from "3" to "4" as it was a typo.
[xv] An "∈" operator was added here as it was missing.
[xvi] Changed from "4" to "5" as it was a typo.
[xvii] ":" was added for consistency reasons.
[xviii] "∀" was replaced by comma for consistency reasons.
[xix] The font was changed for consistency reasons.
[xx] A comma was added here.
[xxi] The font was changed for consistency reasons.
[xxii] "a" was added here.
[xxiii] Changed from "5" to "6" as it was a typo.
[xxiv] Updated to "4".
[xxv] „consist" was replaced by "consists" due to a typo.
[xxvi] Updated to "4".
[xxvii] "respectively" was removed as it was unnecessary.
[xxviii] The value was adjusted from 0.49 to 0.43 due to necessary changes in Figure 1. The adjustment was then applied to the other relevant places in the paper.
[xxix] Updated to "5".
[xxx] The aggregated output was improved.
[xxxi] Updated to "4" and "5".
[xxxii] "function" was replaced by "functions".
[xxxiii] „conjugate search method" was changed to "conjugate gradient search method".
[xxxiv] A hyphen was added here and later.
[xxxv] The reference was updated.
[xxxvi] The reference was updated.
[xxxvii] A link was added here.
[xxxviii] A link was added here.
[xxxix] A link was added here.

[xl] The reference was moved according to the alphabetical order.
[xli] The page range was added here.
[xlii] The reference was updated.
[xliii] The reference was updated.
[xliv] The reference was updated.

2.4 Contribution to RQ 1

The research question addressed in this chapter was:

RQ 1 How can *currency* be adequately measured and how can the *direct* and *indirect* effects of the measured level on decision making be analyzed?

This research question consists of two parts, concerning the *Measure* and the *Analyze* phases discussed above. To address the *Measure* phase, two approaches for measuring currency were proposed: the extended metric modeled with probability theory and the fuzzy metric defined with fuzzy set theory. The extended metric is based on the idea that the temporal change of real-world information can be represented as a stochastic process. Thus, if the history of the real-world information before the time of the decision is known, then the distribution of the real-world information at the time of the decision can be determined. This is important, especially if the stored information is of low currency. In case not all of the history is known, an alternative metric was provided for which it is enough to know the real-world information at the time of storage. The fuzzy metric for currency is modeled in the form of a FIS where the input linguistic variables to the system are the age and the decline rate of a stored attribute value and the output variable is currency. The FIS is built with a set of rules, determined by experts and its application results in a currency in [0,1]. If the age and/or the decline rate of the stored attribute value is/are not known, then other methods such as an auxiliary FIS can be applied to make more precise estimations.

The results from these two metrics must then be adequately considered in decision making to avoid wrong decisions (*Analyze* phases). In this chapter, it was demonstrated how this can be done both *directly* and *indirectly* for the extended metric. To *directly* incorporate the results from the extended metric in decision making with environmental uncertainty, an approach based on the normative concept of the value of information was developed. The idea behind it is that just as information delivers an indication about the state of nature that will occur, so does the extended metric referring to currency deliver an indication about the real-world information based on the stored information. To *indirectly* consider the results from the extended metric, it was shown how the above indication can be incorporated in the classification of new instances in existing decision trees. The point is that if these instances are characterized by quality uncertainty, then they would follow multiple paths of the tree. Thus, it is not possible to uniquely classify each of them in one of the classes of the variable of interest. As a result, a new instance must be classified in multiple classes, from which at the end one is chosen with majority voting. Based on the presented approaches in this chapter, both the *direct* and *indirect* effects of the extended metric can be incorporated in decision making. In Section 3.2, it will be shown how quality uncertainty modeled with fuzzy set theory can be considered *directly* in decision making. This completes the analysis of the information quality dimension currency. In the next chapter, the information quality dimension accuracy is discussed.

3. Measuring and Analyzing Accuracy

In this chapter, RQ 2 is addressed by developing a quantitative approach for measuring low accuracy due to subjective estimations and for *directly* incorporating the results of the measurement in decision making (*Measure* and *Analyze* phases). In particular, in Section 3.1 the scenario of a Cloud service provider is considered who, due to the sequential arrival of the job requests and the often applied pay-as-you-go policy, faces a decision problem with time uncertainty regarding incoming job requests. The aim of the Cloud service provider is to maximize revenue by considering the resource constraints for a defined time slot. However, since a job can require resources for more than one time slot, accepting a request at a certain point of time affects the feasibility of later decisions. To structure this decision problem (i.e. case i) in Figure 1), the decision maker needs to know the amount of resources required by each incoming request. However, this amount is only vaguely estimated by customers. Thus, the decision maker faces quality uncertainty in terms of low accuracy due to subjective estimations, which is modeled with fuzzy numbers (*Measure* phase). This uncertainty is then *directly* considered in decision making (*Analyze* phase), resulting in a fuzzy optimization problem, for which there are no standard solution approaches in the literature.

One way to solve such problems is to use the concept of duality theory. In traditional duality theory, for a primal-dual pair of two optimization problems, if certain conditions are satisfied, it is enough to solve one of them to obtain the solution of the other one. This is particularly useful when one of the two problems is very difficult or inefficient to solve, but the other one is not. The idea has been analogously translated to fuzzy linear optimization resulting in fuzzy duality theory. However, the literature on the topic is very fragmented, inconsistent, and partly incomplete making its application difficult. Thus, in Section 3.2 a state-of-the-art analysis of this stream of research is provided, by systematically analyzing and classifying existing approaches. Moreover, directions for future research are proposed. The results can be applied to all cases where quality uncertainty is modeled with fuzzy set theory such as Papers 3 and 4.

3.1 Paper 4: Revenue Management for Cloud Computing Providers: Decision Models for Service Admission Control under Non-probabilistic Uncertainty

Full citation: Püschel, T., Schryen, G., Hristova, D. and Neumann, D. (2015), *Revenue management for Cloud computing providers: Decision models for service admission control under non-probabilistic uncertainty*, European Journal of Operational Research, Vol. 244 No. 2, pp. 637-647, http://dx.doi.org/10.1016/j.ejor.2015.01.027.

Status: accepted on 15.01.2015

Highlights: In this paper the problem of a Cloud service provider is considered, who faces time and quality uncertainty with respect to the resource requirements of incoming requests. The time uncertainty is due to the sequential nature of the job requests and the often applied pay-as-you-go

policy, implying that future resource requirements are characterized by imperfect information. The quality uncertainty (in the form of lower accuracy) is due to the subjective estimations of the customers. As a result, Cloud service providers need to make real-time decisions regarding the acceptance of incoming requests, while *directly* considering accuracy in their decision. In this paper, three different admission control policies based on the literature are presented (first-come first-served, dynamic pricing, and client classification). In all three cases, considering accuracy of the resource requirements in the decision problem results in the "fuzzification" of the constraints of the problem. The contribution of the paper is threefold. First, it extends existing literature by providing an approach for modeling (subjective) accuracy of resource requirements in admission control for Cloud service providers. Second, the use of the three admission control policies allows examining the effect of quality uncertainty on both revenue maximization and quality of service for Cloud service providers. Finally, the proposed approach is very efficient, making it suitable for real-time decisions.

Evaluation: The approach is evaluated both analytically and with a simulation using real-world data. The analytical evaluation shows that under certain conditions, the dynamic pricing policy leads to higher revenue than the first-come first-served policy both with and without quality uncertainty. Moreover, the client classification policy leads to higher acceptance rate of important customers than the first-come first-served policy also both with and without quality uncertainty. Finally, it is shown that the approach runs in polynomial time. In addition to the analytical evaluation, the methodology is tested by using real-world workloads and a simulation experiment. The results support the findings from the analytical evaluation and also show that increasing quality uncertainty of the resource estimations reduces revenue in the first-come first-served policy. As opposed to that, the exact effect for the dynamic pricing policy cannot be clearly determined. Finally, in the presence of quality uncertainty, the client classification policy leads to lower revenue than the first-come first-served and the dynamic pricing policies in most of the cases.

Limitations: The presented approach also has some limitations. First, in future research the exact effect of quality uncertainty on the revenue level especially for the dynamic pricing policy, needs to be further investigated with other datasets. Second, low accuracy stemming from the vague estimation of the available capacity from the Cloud service provider (i.e. supply side) can also influence revenue and must be analogously considered in decision making. This will result in the need for alternative fuzzy optimization solution approaches, which is addressed in Section 3.2.

3.2 Paper 5: Duality in Fuzzy Linear Programming: a Survey

Full citation: Schryen, G. and Hristova, D. (2015), *Duality in fuzzy linear programming: a survey*, OR Spectrum, Vol. 37 No. 1, pp. 1-48.

The final publication is available at Springer via http://dx.doi.org/10.1007/s00291-013-0355-2.

Status: accepted on 09.10.2013

Highlights: The paper presents a systematic literature review of the existing approaches in the field of duality theory for fuzzy linear programming. The approaches are classified according to the components of the decision problem that are "fuzzified". These include the objective function, the order operator, and the constraints. Since quality uncertainty due to subjective estimations, can appear in any of these components during the structuring of the decision problem (case i) in Figure 1), this is a reasonable approach. The classified approaches are additionally analyzed with respect to the used fuzzy numbers and fuzzy order operators, as well as the presented duality theorems (i.e. weak duality, strong duality, complementary slackness, and the fundamental theorem of duality). The contribution of the paper is twofold. First, it identifies, classifies, presents, and compares existing approaches, thus supporting both academics and practitioners with regard to the choice of appropriate methods. For example, if there are inaccurate (cf. Paper 4) or outdated (cf. Paper 3) estimations in the constraints, then the decision maker can concentrate on the presented approaches in the corresponding class. Second, based on the results from the literature analysis, the paper provides multiple paths for future research that address the current issues in the field.

Evaluation: As mentioned above, as opposed to Papers 1, 2, 3, 4, and 6 this paper does not follow the normative modeling approach, but rather presents an interpretive-historical analysis approach. In that sense, no evaluation as in the other papers takes place. It is replaced by the analysis of the existing literature with respect to the proven duality theorems. This analysis is conducted both on a class basis (i.e. according to the "fuzzified" components), and on a fuzzy operator basis. The results show that the weak and strong duality theorems are (directly or indirectly) proven for all possible classes, which however is not the case for the complementary slackness and the fundamental theorem of duality. Moreover, according to the analysis from a fuzzy operator point of view, there is a big difference with regard to the homogeneity, completeness, consistency, and complexity of the approaches in the different operator groups.

Limitations: The paper also has some limitations. First, the focus is on linear optimization, as one of the most commonly applied approaches. However, the advantages of duality theory exist also in non-linear optimization and the approaches there can be used to address the current issues identified in the paper. Second, the presented overview can be applied to solve different decision problems where quality uncertainty is modeled with fuzzy set theory, such as the one presented in

Section 3.1. In particular, based on duality theory, more efficient solution approaches can be identified, which is especially relevant in the context of big data. Both will be the task of future research.

3.3 Contribution to RQ 2

The research question addressed in this chapter was:

RQ 2 How can *accuracy* in subjective estimations be adequately measured and how can the *direct* effects of the measured level on decision making be analyzed?

Similar to RQ 1, this research question also consist of two parts, addressing the *Measure* and *Analyze* phase of the information quality management cycle. To begin with, in Section 3.1, an approach for modeling (subjective) estimations and their accuracy in the field of Cloud computing was developed (*Measure* phase). The idea is to model the quality uncertainty in the resource estimations of the customers with fuzzy numbers, thus accounting for the possibility of using more or less resources after execution. This approach is not specific to Cloud computing, but can be applied to other scenarios with accuracy problems due to subjective estimations.

The results from the *Measure* phase are then *directly* considered in decision making with time uncertainty (*Analyze* phase) by using three different admission control policies. In all three decision problems, quality uncertainty results in a fuzzy number on the left-hand-side and a non-fuzzy number on the right-hand-side of the constraints. There are no standard solution approaches for such decision problems in the literature and in this paper they are solved by employing symmetric, triangular fuzzy numbers and a simple order operator. However, this is not universally applicable. Thus, there is a need for more advanced solution approaches.

One such approach from the field of fuzzy linear programming is duality theory. It allows transforming a difficult fuzzy optimization problem into another (fuzzy) optimization problem which is not so difficult (or time-consuming) to solve. However, as the analysis in Section 3.2 shows, there are several problems with this field of research. First, the existing approaches are not complete with respect to the different duality theorems. Second, since there are different ways to define fuzzy order operators, there are also different treatments of fuzzy constraints and also different definitions of maximization and minimization operators. This leads to different approaches in fuzzy duality (even for the same operator and/or the same decision problem) resulting in a very fragmented and incomplete field of research. Moreover, some of the approaches are very complex and also do not follow the traditional interpretation of duality. To address these issues, in Section 3.2 multiple solutions are proposed, such as the development of a unifying fuzzy duality theory. This completes the discussion of the information quality dimension accuracy. In the next chapter, the information quality dimension consistency is considered.

4. Measuring Consistency

In this chapter, RQ 3 is addressed by proposing a quantitative approach for measuring consistency based on probability theory (*Measure* phase). As mentioned above, consistency is defined as the degree to which the assessed information agrees with a predefined set of rules. Thus, similar to currency, it is appropriate to model this (objective) degree as probability, where the lower the probability is, the higher the quality uncertainty is. This will also facilitate the *direct* consideration of the metric results in decision making (*Analyze* phase) and an effective information quality management (*Improve* phase). Moreover, an adequate metric for consistency must take into account a wide range of possible rule sets. In particular, it should not only be applicable to rules which are always either true or false, but also to such that are characterized by implicit uncertainty. For example, a rule stating that all 40-year-olds are married is true for many people, but obviously not for all of them. Thus, it is not "true by definition", but possesses implicit uncertainty. This rule uncertainty must also be modeled for an adequate metric for consistency.

4.1 Paper 6: Assessing Data Quality – A Novel Probability-based Metric for Consistency

Full citation: Heinrich, B., Hristova, D., Klier, M., Schiller, A., Wagner, G. (2015), *Assessing Data Quality – A Novel Probability-based Metric for Consistency*, Working Paper, University of Regensburg.

Current Status: Submitted to the International Conference on Information Systems on 05.05.2015, under review

Highlights: In this paper a probability-based metric for consistency is developed, which is defined for a predefined set of rules. The idea of the metric is to compare the probability with which a rule is fulfilled in a consistent reference dataset with the relative frequency (i.e. empirical probability) with which the same rule is fulfilled in the assessed dataset. If the two coincide, then the attribute value is assigned to be consistent. In order to measure the degree of consistency, hypothesis testing is applied. In particular, consistency is measured as the two-sided p-value of the hypothesis test under the null hypothesis that the two input probabilities coincide. In that sense, the measured consistency level can be interpreted as probability. The contribution of the paper is twofold. First, as opposed to existing approaches, the metric allows for the consideration of rules which are not "true by definition". This is crucial, because most rules in real-world applications possess implicit uncertainty in their consequent and this should be taken into account for adequate consistency measurement. Second, as opposed to existing works, the presented metric has a clear interpretation in terms of probability yielding the advantages mentioned above.

Evaluation: The metric for consistency was implemented and evaluated based on a real-world dataset consisting of the GPS sensor measurements of six mobile devices. These measurements were experimentally generated for the purpose of the paper and compared with a consistent reference dataset of coordinates (i.e. longitude and latitude) extracted from GoogleMaps. Moreover, a

survey among 27 IS professionals was conducted, the results from which support the validity of the metric.

Limitations: The proposed metric for consistency has few limitations. First, it addresses the *Measure* phase of the information quality management cycle. However, since it is interpreted as probability, its results can, similar to the metric in Section 2.1, be *directly* and *indirectly* considered in decision making (i.e. *Analyze* phase). Thus, the task of future research would be to develop approaches for supporting decisions based on the metric results. Second, the metric is defined for structured data, but similar to the papers in Chapter 2, it is important to extend it to other types of data, thus addressing the Variety characteristic of big data.

Assessing Data Quality – A Novel Probability-based Metric for Consistency

Completed Research Paper

In this paper we present an information quality metric for consistency which is based on probability theory and a predefined set of rules. Our metric for consistency addresses the existing research gap, by both 1) considering rule uncertainty and 2) providing a clear interpretation of the metric results. We measure consistency by statistically comparing the probability with which a rule is considered to be fulfilled with the relative frequency with which it is observed to be fulfilled in the dataset to be assessed. The metric results can be interpreted as the probability that the assessed dataset is free of internal contradictions with regard to the predefined set of rules. We demonstrate the applicability and validity of our approach based on experimentally generated datasets containing the GPS measurements of six different mobile devices and a survey among IS professionals. Our metric can be applied to efficiently and adequately support decision making.

Introduction

Nowadays, companies around the world use large amounts of internal and external data to support decision making. The volume of this data is constantly growing due to the rapid technological development (e.g., sensor measurements in the context of the Internet of Things) and the increasing number of open data initiatives. In 2013 the number of open government datasets worldwide exceeded one million (Manyika et al. 2013). However, despite the large amount of available data, companies still struggle with generating the best value from it due to information quality (IQ) problems (Witchalls 2014). In a survey conducted by Henschen (2013), 59% of the respondents consider IQ to be "the biggest barrier to successful analytics or BI initiatives." Moreover, according to another survey by Moges et al. (2011, p. 639), "63% of the respondents indicated that inconsistency (value and format) and diversity of data sources are main recurring challenges of DQ." Finally, in a survey by Forbes Insights (2010), on average respondents attributed more than $5 million annual losses in their company to IQ problems.

IQ can be defined as the "agreement between the data views presented by an IS and the same data in the real world" (Orr 1998, p. 67; for a similar definition cf. also Parssian et al. 2004). IQ is a multidimensional construct comprising different dimensions such as correctness, consistency, completeness, and currency (Batini et al. 2009; Eppler 2006; Heinrich and Hristova 2014; Lee et al. 2002; Redman 1996). We focus on consistency as one of the most important dimensions (Blake and Mangiameli 2009; Helfert 2002; Shankaranarayanan et al. 2012; Wand and Wang 1996; Wei-Liang et al. 2009). Thereby, as opposed to correctness, assessing consistency does not require a so-called real-world test (Heinrich and Klier 2015). This is a huge advantage of the dimension consistency, since comparing data values to their real-world counterparts is often far too time-consuming and cost-intensive. In the case of large volumes of analyzed data such a real-world test is not practicable at all. We define consistency as the degree to which the assessed data "is free of internal contradictions" (Heinrich et al. 2007, cf. also Liu and Chi 2002; Mecella et al. 2002; Redman 1996). Contradictions are determined based on a predefined set of rules (Batini and Scannapieco 2006; Heinrich et al. 2007; Mezzanzanica et al. 2012). Thereby, a rule represents a proposition consisting of two logical statements, where the first statement (antecedent) implies the second one (consequent). For example, in a database with weather data for Munich, Germany, such a rule may be: $month = \text{August} \rightarrow night\ temperature\ in\ [℃] > 0$. A temperature of minus four degrees Celsius measured in Munich on August 23rd would, for example, contradict this rule.

The vast majority of existing IQ metrics for consistency (e.g., Batini et al. 2009; Cordts 2008; Heinrich et al. 2007; Pipino et al. 2002) are based on the fulfillment of such rules and consider rules as "true by definition", which means that they have to be true for all the assessed data. A typical example refers to address data. Here, the values for the attributes *city*, *street* and *street number* (antecedent) imply a unique value for the attribute *zip code* (consequent). However, as the example for the weather database shows, there also exist rules which are not "true by definition" but only a subset of the data fulfilling the antecedent also fulfills the consequent. Indeed, not all measurements for the night temperature in Munich on August 23rd will be higher than zero degrees Celsius, even if they are correctly assessed and stored. Therefore, a contradiction to this rule does not necessarily imply that such data is inconsistent and of low IQ. Rather, we have to distinguish. On the one hand, contradictions may stem from the fact that the rule is not "true by definition" but has a probabilistic character (e.g., the low value for the night temperature may also result from a cyclone). On the other hand, they may also occur due to random (e.g., a single sensor value is transmitted incorrectly) or systematic (e.g., a sensor device is damaged and needs to be repaired) data errors (cf. Alkharboush and Yuefeng Li 2010; Fisher et al. 2009). Moreover, in some cases, rules are determined by experts and thus their fulfillment cannot be treated as "true by definition" either. In all of these situations, we have rules with uncertain consequent, to which we refer to as rule uncertainty from now on. This uncertainty must be considered in the assessment of consistency as it exists in most of the cases. However, literature only provides very few approaches aiming at addressing this important issue when measuring consistency. Based on association rule mining (Agrawal et al. 1993), some metrics (Alpar and Winkelsträter 2014; Hipp et al. 2001; Hipp et al. 2007) take the rule confidence (i.e., the percentage of relevant transactions that satisfy a rule) into account. However, these approaches suffer from a lack of clear interpretation which is important to adequately support business decisions, for benchmarking, for repeated analysis over time, and also for determining the effectiveness of IQ improvement measures (cf. Heinrich and Klier 2015).

Thus, (1) to consider rule uncertainty in a well-founded way and (2) to ensure a clear interpretation of the metric results, we propose a novel IQ metric for consistency which is based on probability theory. To address uncertainty, a metric for consistency shall deliver an indication or estimation rather than a "true by definition" statement. We argue that the well-founded methods of probability theory are adequate and valuable to describe and analyze uncertain rules. The design of our metric is based on a statistical comparison of the probability with which a rule is considered to be fulfilled in a reference dataset with the relative frequency with which it is observed to be fulfilled in the assessed dataset. By this means, we consider (1) rule uncertainty in a well-founded way when assessing consistency. Moreover, we ensure (2) a clear interpretation of the metric results by measuring consistency as the probability that the assessed data is free of internal contradictions with regard to the predefined set of rules. These contradictions may be due to both random and systematic errors.

The remainder of the paper is structured as follows. In the next section, we discuss related literature and identify the research gap. Thereafter, we present a novel metric for consistency and provide different possible procedures to design metric instantiations. In the fourth section, the metric is applied to a real-world dataset comprising GPS data generated by different mobile devices. Finally, we briefly summarize the findings and conclude with a discussion of limitations and directions for further research.

Related Work

Consistency is seen "as a multi-faceted dimension" (Blake and Mangiameli 2009) which can be defined in terms of representational consistency, integrity, and semantic consistency (Blake and Mangiameli 2009). Since these three aspects stem from different domains, they overlap in some cases. Representational consistency requires that data are "presented in the same format and are compatible with previous data" (Blake and Mangiameli 2009; p. 32; cf. also Wang and Strong 1996) (e.g., consistent use of abbreviations). Integrity is often defined as entity, referential, domain, column, and user-defined integrity (Blake and Mangiameli 2009; Lee et al. 2004). Entity integrity requires that attribute values considered as primary keys are unique and different from NULL. Referential integrity states that, given two relations, if an attribute is a primary key in one of them and is contained as a foreign key in the other one, then the non-NULL attribute values from the second relation must be contained in the first one (Lee et al. 2004). Domain and column integrity require that the attribute values are part of a predefined domain (e.g., $income \geq 0$) and user-defined integrity requires that a predefined set of general rules is satisfied. Finally, semantic consistency refers to the absence of contradictions between different attribute values based on a set of rules (Blake and Mangiameli 2009; Della Valle et al. 2008; Heinrich et al. 2007; Helfert 2002; Hinrichs 2002; Lee et al. 2006; Liu and Chi 2002; Mecella et al. 2002; Mezzanzanica et al. 2012; Redman 1996; Scannapieco et al. 2005). Generally, semantic consistency is equivalent to user-defined integrity.

In this paper, in accordance with our definition of consistency, we focus on semantic consistency due to three major reasons. First, while both integrity and representational consistency have already been extensively studied in the literature (Blake and Mangiameli 2009), semantic consistency is a relatively new research field which gains more and more importance in the course of growing data volumes. Second, while integrity and representational consistency can be assured during the data integration process, semantic consistency must be assured already during the acquisition of the data which may be resource intensive or even impossible. Thus, it is important to assess semantic consistency to identify potential conflicts. Finally, assuring semantic consistency is crucial for decision support, as decision making is typically based on attribute values which are also the ones assessed for semantic consistency. In the following, for simplicity, we will use the term consistency instead of semantic consistency.

To provide an overview of existing works on IQ metrics for consistency, we concentrate on metrics that are (i) formally defined (e.g., by a closed-form mathematical function) and (ii) result in a specific metric value representing the consistency of the data to be assessed. In that sense, we do not consider approaches that aim at identifying groups of potentially (in)consistent attribute values without providing specific metric values for (in)consistency (e.g., Fan et al. 2013; Mezzanzanica et al. 2012). Table 1 presents existing metrics for consistency satisfying (i) and (ii). They generally follow the idea that consistency of an attribute value or a data record can be determined based on the number of fulfilled rules, with a higher number of fulfilled rules implying a higher consistency value and vice versa (Alpar and Winkelsträter 2014; Batini et al. 2009; Cordts 2008; Heinrich et al. 2007; Hinrichs 2002; Hipp et al. 2001; Hipp et al. 2007; Kübart et al. 2005; Pipino et al. 2002).

In the following we discuss these metrics with regard to (1) the consideration of rule uncertainty and (2) the clear interpretation of the metric results. The metrics that consider (1) rule uncertainty model this uncertainty in terms of weights assigned to the fulfillment (and violation) of different rules (Alpar and Winkelsträter 2014; Hinrichs 2002; Hipp et al. 2001; Hipp et al. 2007; Kübart et al. 2005), which are then summed up to determine consistency. For example, Alpar and Winkelsträter (2014), Hipp et al. (2001) and Hipp et al. (2007) use the confidence $conf(r_n)$ of a rule r_n to determine these weights. In particular, they assign a weight of $conf(r_n)^\tau$ to the fulfillment of the rule and a weight of $-conf(r_n)^\tau$ to its violation, where τ represents a calibration parameter. Although the above authors present first, promising steps in considering rule uncertainty, due to the summation of the weights, they suffer from a lack of (2) clear interpretation, which is also related to the fact that the metric results are not normalized.

There are approaches in the literature which have a (2) clear interpretation in terms of the percentage of consistent attribute values (Batini et al. 2009; Cordts 2008; Heinrich et al. 2007; Pipino et al. 2002). However, they treat the fulfillment of a rule as "true by definition" and thus do not take (1) rule uncertainty into account. The metric results are either summed (Batini et al. 2009; Cordts 2008; Pipino et al. 2002) or multiplied (Heinrich et al. 2007) to determine consistency with respect to the number of rules. To sum up, the existing metrics for consistency do not aim at 1) considering rule uncertainty and/or at 2) ensuring a clear interpretation of the metric results. In the next section we address this research gap.

Source	Metric
Batini et al. 2009;Cordts 2008; Pipino et al. 2002	g: attribute value, N: number of relevant rules[1] for g $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$, $cons(g) = 1 - \frac{\sum_{n=1}^{N} r_n(g)}{N}$
Heinrich et al. 2007	g: attribute value, N: number of relevant rules for g $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$, $cons(g) = \prod_{n=1}^{N}(1 - r_n(g))$
Hinrichs 2002	g: attribute value, N: number of relevant rules for g, w_n: weights $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$, $cons(g) = \frac{1}{\sum_{n=1}^{N} w_n r_n(g)+1}$
Alpar and Winkelsträter 2014; Hipp et al. 2001; Hipp et al. 2007	t:record, N:number of relevant rules for t, L: number of irrelevant rules for t, ${w^-}_n$, ${w^+}_n$, ${w^0}_l$: weights $r_n(t) = \begin{cases} 0, \text{if } t \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$, $h_n(t) = \begin{cases} 0, \text{if rule } r_n \text{ applies for } t \\ 1 \text{ else} \end{cases}$ $cons(t) = (\sum_{n=1}^{N}({w^-}_n r_n(t) + {w^+}_n(1 - r_n(t))))(1 - h_n(t)) + \sum_{l=1}^{L} h_l(t){w^0}_l$
Kübart et al. 2005	t:record, N:number of relevant rules for t, ${w^-}_n \geq 0$: weights $r_n(t) = \begin{cases} 0, \text{if } t \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$, $incons(t) = \sum_{n=1}^{N} {w^-}_n r_n(t)$

Table 1. Existing Metrics for Consistency

Development of the Probability-based Consistency Metric

In this section, we present our novel metric for consistency. First, we outline the general setting and the basic idea. Then, we describe the methodological foundations, based on which we define the metric in the subsequent subsection. Finally, we outline possible ways to instantiate the metric.

General Setting and Basic Idea

We consider the common relational database model. In particular, let D_B be a given database and let a relation in it consist of a set of attributes $\{a_1, \ldots, a_m\}$ and a set of records $T = \{t_1, t_2, \ldots, t_n\}$. We denote the

[1] The "consistency checks" mentioned in these works can be interpreted as checks for fulfillment of rules.

attribute value of record t_j regarding attribute a_i with $\phi(t_j, a_i)$. In line with existing literature (cf. section "Related Work"), in order to assess consistency, we use a set of predefined rules which can be estimated by experts or derived from a reference dataset. The rules are contained in a rule set R and are propositions of the form $r: A \rightarrow C$, where A (antecedent) and C (consequent) are logical statements addressing either single attributes in D_B or the relations between them. However, as mentioned above, contrary to existing approaches, we do not treat rules as "true by definition". Rather, we incorporate rule uncertainty in measuring consistency. In particular, based on probability theory, we compare the relative frequency with which a rule r is fulfilled in D_B with the probability with which this rule is expected to be fulfilled in a reference dataset D_R (e.g., a quality assured database). As a result, uncertain rules are treated differently than certain rules, which, as discussed above, is crucial in many cases.

For an attribute value $\phi(t_j, a_i)$ in D_B and a given rule r, we interpret consistency as the probability that $\phi(t_j, a_i)$ does not contradict D_R with regard to r but rather corresponds to the attribute values in D_R in this respect. This guarantees that the metric values are normalized and have a clear interpretation. Moreover, based on this interpretation, the metric results can be integrated in decision support, for instance, into the calculation of expected values.

To illustrate the idea of our metric and its advantages, consider the following running example: A health insurance company conducts a product campaign targeting exclusively married customers, who are younger than 20. The campaign is based on a customer database, where for each customer his/her year of birth, marital status, and education are stored. If the data stored in this database is incorrect, wrong decisions and economic losses may result. For instance, if a customer is stored as "married" in the database and in reality this is not the case, contacting him/her with a product offer will not bring any benefits, but will only generate costs and may lead to lower customer satisfaction. Measuring consistency in that context is more appropriate than measuring correctness, because to assess correctness, the customers need to be contacted personally (real-world test), which is time-consuming and cost-intensive. Consider the rule $r_1: year\ of\ birth > 1995 \rightarrow marital\ status = single$, where $year\ of\ birth > 1995$ is the rule antecedent and $marital\ status = single$ is the rule consequent. This rule is fulfilled by most people who are younger than 20. Existing metrics for consistency which do not consider rule uncertainty will classify all married customers who are younger than 20 as inconsistent. In that case, no customer will be contacted and thus the campaign will fail. We address this problem by considering the frequency of a given rule and in addition to r_1, we also consider $r_2: year\ of\ birth > 1995 \rightarrow marital\ status = married$. In particular, our approach compares the correspondence between the probability that a person who is younger than 20 is married in a consistent reference dataset (e.g., census data) and the relative frequency of such people in the customer database. Since the number of married people below 20 is generally low, if this is also the case in the customer database, they will not be considered inconsistent. Moreover, by measuring consistency as a probability, the company may decide to contact only customers with consistency values above a certain threshold (e.g., probability of 70%). In the following subsections, we present the methodological foundations and formally define our metric.

Methodological Foundations

Rules and their Application to a Dataset

As mentioned above, a rule $r: A \rightarrow C$ consists of the logical statements A and C, where A and C describe the relations between different attributes in D_B. The simplest form of a logical statement S is defined (Chiang and Miller 2008; Fan et al. 2013) as:

<attribute><operator><attribute>
or
<attribute><operator><constant>

where <attribute> is one of the attributes a_i and <operator> is a binary operator such as $=, \geq, >, \neq$ or $substring\_of$. Simple logical statements can be linked by conjunction (AND, $\wedge$), disjunction (OR, $\vee$) or negation (NOT, $\neg$) to form more complex logical statements. For instance, in the running example, we may have a rule of the form

$r_3: year\ of\ birth > 1995 \wedge school\ education = high\ school \rightarrow marital\ status = single,$

where $year\ of\ birth > 1995$, $school\ education = high\ school$, $marital\ status = single$ and $year\ of\ birth > 1995 \wedge school\ education = high\ school$ are logical statements. To determine whether a logical statement S is true or false for a record t of D_B, it can be applied to t by replacing each attribute a_i contained in S by $\phi(t, a_i)$ (i.e., by the corresponding attribute value of the record t).

We define the set of records in D_B rendering S true as $fulfilling\ records(D_B, S) := \{t \in T |\ S(t)\ is\ true\}$. As an example, we can apply the antecedent $year\ of\ birth > 1995$ and the consequent $marital\ status = married$ of the rule r_2 to a record t of the customer database with $\phi(t, year\ of\ birth) = 1997$ and $\phi(t, marital\ status) = married$ and, since $1997 > 1995$ and $married = married$, it follows that $A(t) = true$ and $C(t) = true$. Thus, $t \in fulfilling\ records(D_B, A)$ and $t \in fulfilling\ records(D_B, C)$.

We call a rule $r: A \rightarrow C$ *relevant* for a record $t \in T$ if $t \in fulfilling\ records(D_B, A)$. If r is relevant for t and $t \in fulfilling\ records(D_B, A \wedge C)$, we say that t *fulfills* r, otherwise t *violates* r. As mentioned above, contrary to existing approaches for measuring consistency, we examine the relative frequency with which a rule is fulfilled in D_B in comparison with the probability with which it is fulfilled in the reference dataset D_R. To be more precise, we assign to each rule $r \in R$ a number $p(r) \in [0,1]$ representing the probability with which r is fulfilled in the reference D_R. Later, when discussing the metric instantiation, we will outline how the rule set R and the corresponding values $p(r)$ can be obtained in practice.

Statistical Background

Let D_R be a reference dataset and $r: A \rightarrow C$ be a rule in R. Consider the records $t \in T$ in D_R for which r is relevant (i.e., $t \in fulfilling\ records(D_R, A)$). Such a record fulfills r with probability $p(r)$ (i.e., $t \in fulfilling\ records(D_R, A \wedge C)$ with probability $p(r)$). Hence, the application of r to t can be seen as a Bernoulli trial with success probability $p(r)$ (resp. failure probability $1 - p(r)$), where success is defined as $t \in fulfilling\ records(D_R, A \wedge C)$. The Bernoulli trial can be represented by a random variable $r(t)$ with

$$r(t) = \begin{cases} 1 & \text{if } t \in fulfilling\ records(D_R, A \wedge C) \\ 0 & \text{if } t \notin fulfilling\ records(D_R, A \wedge C), t \in fulfilling\ records(D_R, A) \end{cases} \quad (1)$$

such that $r(t) \sim Bern(p(r))$.

In the same manner, we can apply r to all records $t \in T$ in D_R with $t \in fulfilling\ records(D_R, A)$ and sum up the results. This is defined by the random variable $X(r) = \sum_{t \in fulfilling\ records(D_R, A)} r(t)$ which represents the number of records t with $t \in fulfilling\ records(D_R, A \wedge C)$. As a sum of independent Bernoulli distributed random variables, $X(r)$ follows a binomial distribution with parameters $|fulfilling\ records(D_R, A)|$ and $p(r)$ (i.e., $X(r) \sim B(|fulfilling\ records(D_R, A)|, p(r))$). An illustration for such a distribution with parameters 100 and 0.5 is presented in Figure 1. The expected value of $X(r)$ represents the expected number of records $t \in fulfilling\ records(D_R, A \wedge C)$ for D_R.

Consider now a dataset D_B the consistency of which we aim to assess and again the rule $r: A \rightarrow C$. Then, $|fulfilling\ records(D_B, A \wedge C)|$ is the cardinality of the set of records in D_B which fulfill r. Since $X(r)$ represents $|fulfilling\ records(D_R, A \wedge C)|$, if D_B was a consistent dataset with $|fulfilling\ records(D_B, A)| = |fulfilling\ records(D_R, A)|$, then $|fulfilling\ records(D_B, A \wedge C)|$ has to be distributed as $X(r)$. Thus, to determine the consistency of the records in D_B, we compare $|fulfilling\ records(D_B, A \wedge C)|$ with the distribution of $X(r)$. Generally, the higher the absolute distance between $|fulfilling\ records(D_B, A \wedge C)|$ and the expected value of $X(r)$ is, the less likely it is that D_B is consistent with respect to r. In Figure 1, $|fulfilling\ records(D_B, A \wedge C)| = 60$ and the expected value of $X(r)$ is 50, thus there is an indication for potential inconsistency.

Based on this idea, we develop a probability-based metric for consistency which is founded on the well-known concept of the (two-sided) $p - Value$ in hypothesis testing. Let $p'(r)$ be the probability with which the rule r is fulfilled by a record in the dataset D_B to be assessed. Since in a consistent dataset, each record fulfills the rule r with probability $p(r)$, if D_B was consistent, then $p'(r)$ should be equal to $p(r)$ (e.g., 0.5 in Figure 1). Thus, in statistical terms, measuring consistency implies testing the null hypothesis $H_0: p'(r) = p(r)$ against the alternative hypothesis $H_1: p'(r) \neq p(r)$ under a binomial distribution (i.e., $X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))$). We use a two-sided alternative because both too many and too few fulfillments of r indicate inconsistency. Generally, in statistical terms, if the null hypothesis is true,

the two-sided $p - Value$ represents the probability that a value occurs which is equal to or more extreme (i.e., further away from the expected result) than the observed value. For example, in Figure 1, the two-sided $p - Value$ is calculated by summing up the probabilities $p(X(r) \geq 60)$ and $p(X(r) \leq 40)$, represented by the grey bars.

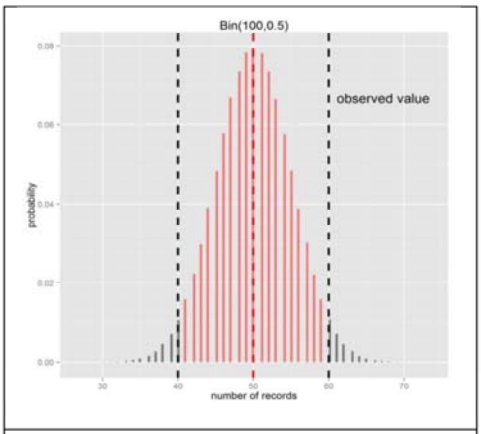


Figure 1. Binomial Distribution

In our case, the observed value is $|fulfilling\ records(D_B, A \wedge C)|$ and the expected value is determined from the distribution of $X(r)$. Thus, the $p - Value$ represents the probability that, under the null hypothesis and for the binomially distributed random variable $X(r)$, a value occurs which is equal to or more extreme than $|fulfilling\ records(D_B, A \wedge C)|$. Hence, it represents the probability that the observed attribute value in D_B corresponds to the considered dataset D_R with respect to the rule r. In the following, we denote the two-sided $p - Value$ of the random variable $X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))$ with respect to the observed value $|fulfilling\ records(D_B, A \wedge C)|$ as $p - Value((X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))), |fulfilling\ records(D_B, A \wedge C)|)$. We would like to note that we are aware of the discussion regarding the interpretation of the $p - Value$ (cf., e.g., Goodman 2008) and since this is not the main focus of our paper, we follow the above, standard interpretation. The outlined statistical background allows for a mathematically sound definition of our metric in the next subsection and ensures a clear interpretation of the consistency values determined by the metric.

Definition of the Consistency Metric

In the following, based on the methodological foundations above, we define our probability-based metric for consistency. Let D_B be a database, $t_j \in T$ be a record in D_B, a_i be an attribute in D_B, and $r: A \rightarrow C$ be a rule[2] such that a_i is part of r and $t_j \in fulfilling\ records(D_B, A \wedge C)$. We define the consistency of the attribute value $\phi(t_j, a_i)$ with respect to r as:

$$Q_{Cons}(\phi(t_j, a_i), r: A \rightarrow C) \coloneqq p - Value((X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))), |fulfilling\ records(D_B, A \wedge C)|) \quad (2)$$

This definition ensures that only attributes are considered which are contained in the rule and only records are considered which fulfill the rule. Our metric identifies inconsistent attribute values due to both random and systematic data errors which possibly warrant a manual examination. Within our running example, for instance, if the attribute value *married* of a customer has a low consistency, it might be better not to contact the corresponding customer in order to prevent unnecessary costs.

[2] We consider a single rule for simplicity here and will address later the case with multiple rules.

The metric in Equation (2) measures consistency on the level of attribute values. Based on this, we derive aggregated consistency metrics for records, attributes, relations, and the whole database. On the level of records, the metric for a record $t_j \in T$ is defined as the weighted arithmetic mean of the consistency values of the corresponding attribute values:

$$Q_{Cons}\left(t_j, R_{t_j}\right) \coloneqq \sum_i w_{a_i} \cdot Q_{Cons}(\phi(t_j, a_i), r_{a_i}) \quad (3)$$

Similar to above, for a given record t_j, we define the rule set R_{t_j} such that for each $\phi(t_j, a_i)$ there is exactly one rule $r_{a_i}: A \rightarrow C$ in R_{t_j} where a_i is part of r_{a_i} and $t_j \in fulfilling\ records(D_B, A \wedge C)$. The weights $w_{a_i} \in [0,1]$ are specific for each attribute a_i and satisfy $\sum_i w_{a_i} = 1$. They can be adjusted to adapt the metric according to the relative relevance of certain attributes for the consistency of the record t_j. This definition can be used to detect records with a low consistency value, for example in order to examine them manually in a later step. For instance, within our running example, customers whose entry in the database has a low consistency value could be contacted in order to verify the stored attribute values of the attributes *school education*, *marital status* and *year of birth*.

Similarly, on the level of attributes, the metric for an attribute a_i is defined as the weighted arithmetic mean of the consistency values of the corresponding attribute values:

$$Q_{Cons}(a_i, R_{a_i}) \coloneqq \sum_{t_j \in T} w_{t_j} \cdot Q_{Cons}(\phi(t_j, a_i), r_{t_j}) \quad (4)$$

The weights $w_{t_j} \in [0,1]$ are specific for each record $t_j \in T$ and satisfy $\sum_{t_j \in T} w_{t_j} = 1$. They reflect the relative relevance of certain records (e.g., certain customers) for the consistency of the attribute a_i. Again, the rule set R_{a_i} is defined so that for each $\phi(t_j, a_i)$ there is exactly one rule $r_{t_j}: A \rightarrow C$ in R_{a_i} where a_i is part of r_{t_j} and $t_j \in fulfilling\ records(D_B, A \wedge C)$. Using this definition allows the recognition of attributes in D_B with low consistency value, which can subsequently be treated as less trustworthy or indicate the need for an IQ improvement measure.

The metric definition on the level of records or, equivalently, on the level of attributes, is used to derive a consistency value for the whole database D_B (with requirements regarding the rule sets as above):

$$Q_{Cons}(D_B, R_{D_B} = \bigcup_{t_j \in T} R_{t_j}) = \sum_{t_j \in T} w_{t_j} \cdot Q_{Cons}\left(t_j, R_{t_j}\right) \quad (5)$$

So far, for reasons of simplicity, it was assumed that D_B consists of only one relation. For a database D_B consisting of a set of disjoint relations M_k, $k = 1, \ldots, K$, the consistency of D_B is defined as the weighted arithmetic mean of the consistency values of the relations M_k, where R_k is the corresponding rule set:

$$Q_{Cons}(D_B, R = \bigcup_k R_k) \coloneqq \sum_k w_k \cdot Q_{Cons}(M_k, R_k) \quad (6)$$

Again, the relation specific weights $w_k \in [0,1]$ for each relation M_k satisfy $\sum_k w_k = 1$. Both Equations (5) and (6) allow assessing the consistency of the database as a whole.

Metric Instantiation

In this subsection, we describe how to instantiate our metric. In particular, we describe how a rule set R can be obtained and how the calculation of the metric values can be performed.

Obtaining the Rule Set

For the application of the metric, it is crucial to determine an appropriate rule set R and the corresponding values $p(r)$ for each $r \in R$. Generally, there are different possibilities to determine this rule set. Similar to Heinrich and Klier (2015), we briefly describe the following three possibilities: (1) Analysis of a reference dataset (e.g., open data), (2) Conducting a study, and (3) Surveying experts.

Ad (1): A promising option is to use a reference dataset D_R (in case such exists) which is representative for the data of interest in D_B and is known or considered to contain consistent data. Such a reference dataset may, for example, be quality-assured historical data owned by the organization itself. With more and more external data being provided by recent open data initiatives, reliable publicly available data from

public or scientific institutions (e.g., census data, government data, data from federal statistical offices and institutes) can be analyzed as well. The German Federal Statistical Office, for instance, offers detailed data about the population of Germany (e.g., regarding the marital status according to the person's age, which could be used in the running example). Further common examples are traffic data as well as healthcare databases, providing detailed data about diseases and patients. From such a reference dataset D_R, it is possible to determine the rule set for D_B directly and with a high degree of automation. In the following we exemplarily present three possible ways for the determination of a rule set from a reference database.

First, an association rule mining algorithm (Agrawal et al. 1993; Kotsiantis and Kanellopoulos 2006) can be applied to D_R (similar to, e.g., Alpar and Winkelsträter 2014). The resulting association rules can subsequently be used as input for the metric with $p(r)$ equal to the confidence of the rule r. Using an association rule mining algorithm, however, does not ensure a rule set based on which the consistency of each attribute value can be assessed. For instance, it may happen that for certain attribute values, no rule can be applied. We thus propose the following additional alternatives:

Second, we can create rules with tautological antecedent A in an automated way. This means that the logical statement A is always true. Such a tautology will in the following be denoted by $\top$. The rules consist of a tautological antecedent $\top$ and $a_l = \phi(t_m, a_l)$ as a consequent for all records t_m in D_R and attributes a_l of D_R. This leads to the rule set of the form $R_1 = \{(r\colon \top \rightarrow a_l = \phi(t_m, a_l))\}$. We call these rules *column rules*.

Third, we can generate rules with tautological antecedent and $\bigwedge_{a_l} a_l = \phi(t_m, a_l)$ as consequent for all t_m in D_R in an automated way. This results in the rule set of the form $R_2 = \{(r\colon \top \rightarrow \bigwedge_{a_l} a_l = \phi(t_m, a_l))\}$. We call these rules *row rules*. In both of the last two cases, the values $p(r)$ can be computed by using the relative frequency with which r is fulfilled in D_R. For example, for a record t in D_R with $\phi(t, year\ of\ birth) = 1997$ and $\phi(t, marital\ status) = single$ (and no other attributes in D_R), the rules added to R_1 would be $r_1\colon \top \rightarrow year\ of\ birth = 1997$ and $r_2\colon \top \rightarrow marital\ status = single$ and the rule added to R_2 would be $r_3\colon \top \rightarrow year\ of\ birth = 1997 \wedge marital\ status = single$.

Using column rules to assess the consistency of D_B leads to disregarding the interdependencies between different attributes, as only the frequency with which certain attribute values occur is considered. In other words, the better the distributions of the attribute values of D_R and D_B match, the higher the consistency of D_B will be. Row rules, on the other hand, are very strict with regard to their fulfillment, as all of the attribute values of a record need to match. If D_R and D_B are equally structured, these two ways of proceeding guarantee a rule set such that for each attribute value of each record t_j in D_B exactly one rule is relevant. In other words, a unique consistency value can be determined for each $\phi(t_j, a_i)$ and the consistency of D_B as a whole can be assessed.

These three ways for creating a rule set based on a reference dataset D_R were chosen because of their general applicability. A large variety of further rule sets can be determined, for example by considering fixed attributes in the antecedent or by using different operators. Depending on the dataset and the specific application case, usage of any of these possibilities (or a combination of them) can be favorable as the relations between attribute values may vary. For example, using column rules is obviously promising when interdependencies of attributes do not have to be examined at all.

It needs to be mentioned that when using a reference dataset D_R for obtaining the rules, if the number $|fulfilling\ records(D_R, A \wedge C)|$ of records in D_R fulfilling a rule is too small, the results of the consistency metric with respect to this rule will not be reliable. To be more precise, the statistical significance of $p(r)$ needs to be assured. If an association rule mining algorithm is used to create the rule set, the analyst can fix a suitable minimum support to exclude rules based on a non-significant probability. In any case, a statistical (one-sided) t-test can be applied in order to determine the minimal number of records required such that a rule has a statistically significant explanatory power. However, to provide a statistically reliable basis and to circumvent the aforementioned issue, rules can be aggregated (e.g., by using a disjunction).

Ad (2): If neither external nor internal data is available, conducting a study is a further possibility. For example, if a customer database is to be assessed, a random sample of the customers can be drawn and

surveyed. The survey results can be used to determine an appropriate rule set (by analyzing their statements) and to obtain the corresponding values of $p(r)$ for each rule r (by analyzing how many of the surveyed customers fulfill the rule), thus providing the input parameters for the metric. Hence, as a result of the survey, not only does one obtain correct data of the surveyed customers, but also one can assess the consistency of the data of customers which were not part of the survey.

Ad (3): Another possibility is to use an expert-based approach (similar to Mezzanzanica et al. 2012, cf. Baker and Olaleye 2013; Meyer and Booker 2001). Here, the idea is to survey qualified individuals. For example, for rules in a company database taking into the account the attributes *year of foundation*, *profit* and *turnover*, business experts could be surveyed. Another example concerns very rare events such as predicting earthquakes, for which there is not enough available data. The experts can rate which rules are suitable to describe the expected structure of the considered attribute values and can specify the respective values of $p(r)$ for each rule.

It is worth mentioning here that, although for (1) we always have a reference *dataset*, this is not the case for (2) and (3). However, in such cases our probability-based approach is still valid, since we can use the results from the study and the estimations of the experts to derive the rule set R and the corresponding values $p(r)$. Thus, all input parameters for the metric are provided. In that sense we implicitly assume that a consistent reference dataset exists.

Calculating the Metric Values

Once a suitable rule set R (with values of $p(r)$ for each $r \in R$) has been created, the metric values can be determined in an automated way. The values $|fulfilling\ records(D_B, A)|$ and $|fulfilling\ records(D_B, A \wedge C)|$ can be efficiently determined. In addition, based on the value of $p(r)$, the corresponding binomial distribution can be instantiated. The (two-sided) $p - Value$ with regard to $|fulfilling\ records(D_B, A \wedge C)|$ can be computed in order to calculate the metric values. In the literature, several different ways to calculate the two-sided $p - Value$ have been proposed. These include doubling the one-sided $p - Value$ and clipping to one; summing up the probabilities less than or equal to the probability of the observed result; and more elaborate ways (Dunne et al. 1996). In practical applications, for non-symmetric distributions, these definitions lead to slightly different results. However, the larger the sample size (in our case $|fulfilling\ records(D_B, A)|$), the smaller the differences between the results from the different approaches are, as in this case (for $p(r) \in (0,1)$) the binomial distribution converges to the (symmetric) normal distribution (de Moivre-Laplace theorem).

In the previous subsection, the consistency metric for an attribute value of a record has been defined with respect to a single rule. This must not always be the case, because we can have multiple rules as well as no rule for a certain attribute value. On the one hand, if multiple rules $r_1, \dots r_l$ are available that can be used to determine a consistency value (e.g., this situation may arise when an association rule mining algorithm is applied to obtain the rule set R), the respective consistency values can be aggregated by, for example, using rule-specific weights $w_{r_k} \in [0,1]$ with $\sum_k w_{r_k} = 1$ such that

$$Q_{Cons}(\phi(t_j, a_i), r_1, \dots, r_l) = \sum_k w_{r_k} \cdot Q_{Cons}(\phi(t_j, a_i), r_k). \quad (7)$$

The rule-specific weights can be chosen to reflect the relative importance of the rules. Similarly, the consistency definitions on the level of records and on the level of attributes can be adjusted to the use of multiple rules. On the other hand, if a reference dataset D_R was used to create the rules and for a certain attribute value of a record no rules are applicable because the attribute values were not represented within D_R, then this attribute value can be considered as an outlier and hence receives a consistency value of zero.

Evaluation

In this section, we evaluate (1) the applicability and practical efficacy as well as (2) the validity of our approach. To address (1) we apply our metric in an experimental setting to datasets generated by the GPS measurements of six different mobile devices. In this context, publicly available data from map services can serve as reference dataset. To address (2), we conduct a survey among IS professionals. In particular, we measure the perceived consistency ranking (Lee et al. 2002; Litwin 1995) of the six datasets by visualizing the GPS coordinates measured by the respective mobile device together with the GPS

coordinates in the reference dataset. To evaluate the validity of our metric, we compare the perceived consistency ranking to the consistency ranking determined by the metric results.

Assessing consistency in this context is very important, because GPS coordinates are used as input for many location-based applications such as fleet management, routing, or even entertainment (D'Roza and Bilchev 2003). On the one hand, low consistency on a single record level results in the wrong position being identified and thus in making wrong decisions (e.g., "turn left"). On the other hand, low consistency on the dataset level may be an indication that a device is not working properly und should thus be exchanged or repaired. More precisely, this application context is particularly suitable for illustrating the applicability and practical efficacy of our metric due to the following reasons: First of all, since in most cases GPS coordinates are used as input for real-time decisions, it is important that their consistency can be determined in an automated way. Second, based on the open data from existing mapping services (e.g., GoogleMaps, OpenStreetMaps), it is possible to determine an objective and consistent (i.e., big enough) reference dataset (Blum et al. 2013), which can also be validated by other researchers. Third, in this context the fact that our metric has a clear interpretation in terms of a probability is particularly important, since it facilitates comparisons of the consistency between different records, devices, and points in time. In that case, the consistency values can be applied, for example, to identify and repair specific inconsistent locational measurements, to assess the effectiveness of IQ improvement measures, or in case of conflicts between redundant devices to identify the most consistent one. To illustrate our approach we determine the consistency of the GPS measurements of the different devices both on a single record level (i.e., one location) and on a dataset level (i.e., one device).

GPS technology is based on the idea that the time required for a signal to travel from a satellite to a mobile device can be used to determine the distance between the satellite and the device and thus the position of the device. GPS signal quality is lower if there is no clear sky view (which happens in or between buildings), or in the presence of other obstacles (D'Roza and Bilchev 2003). In such cases the consistency of the values is rather low. There are different coordinate systems to represent a location. In this paper we use decimal degrees, where each position is represented by a pair of latitude and longitude values.

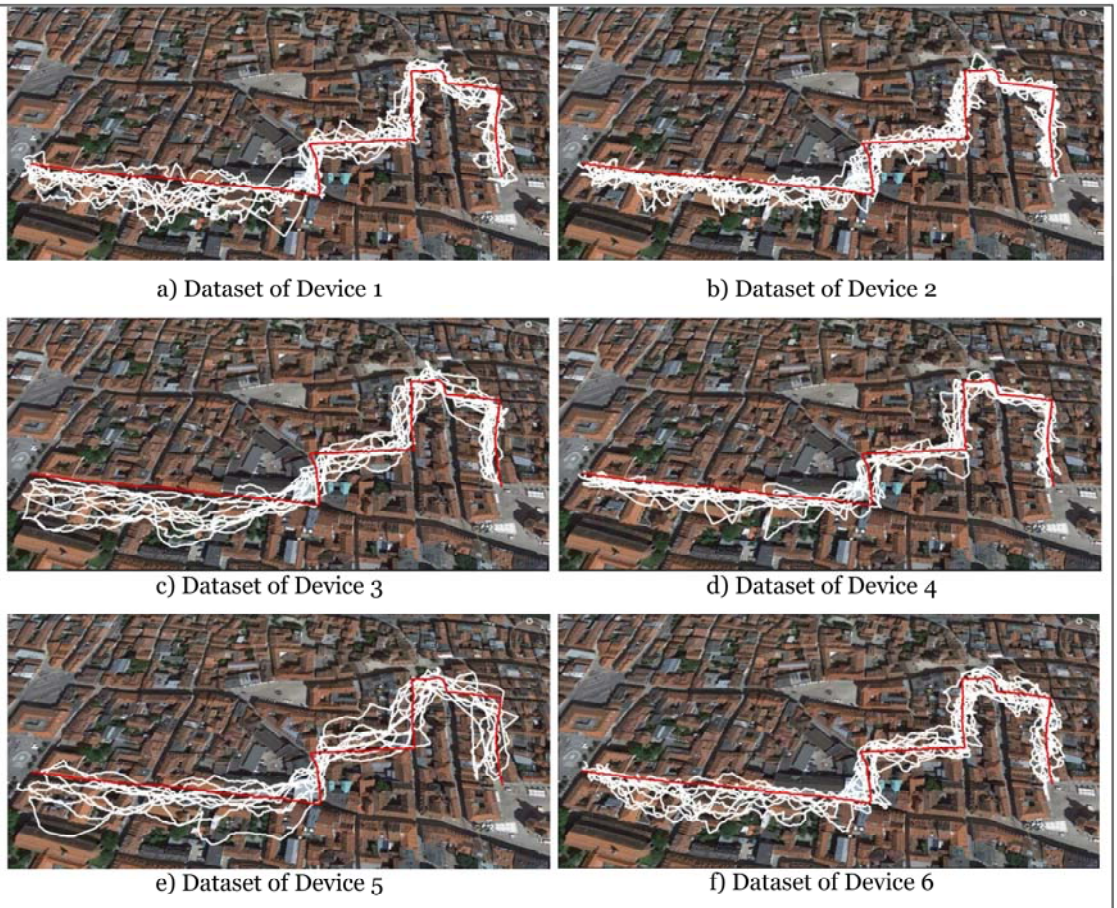
The values of the six datasets were generated on a route with a length of approximately 800 meters, which is presented in Figure 2. It was chosen because of the relatively narrow streets and the many houses along the route, which both influence the quality of GPS signals, as mentioned above. The reference dataset was extracted from Google Maps (https://www.google.de/maps) and consists of 764 two-dimensional data points covering each meter of the route. Each record is represented by a latitude and longitude value, where the latitude values are in the range [49.01848°, 49.02003°] and the longitude values in the range [12.08989°, 12.09628°]. The values were stored with five decimal points corresponding to approximately one meter precision.

Figure 2. Visualization of the Reference Dataset

The datasets to be assessed were generated by six different mobile devices with an Android operating system and without access to the mobile network. The devices were chosen to cover a wide range of price categories and release dates (Blum et al. 2013) such as Vodafone Smart Tab II, Samsung Galaxy S4, Nexus 7 Tablet, and HTC One. The GPS values were extracted by a mobile application which was developed for

the purpose of this paper. The devices were placed in a vertical position and were taken with a constant walking speed along the route. In addition, it was assured that the start and end time stamp (accurate up to 4 milliseconds) of the acquired data for all devices were the same. In order to receive a higher number of records in each dataset and to assure the convergence of the resulting consistency values, we walked the above route multiple times. To determine the number of walks required, we analyzed the average consistency per device and determined the minimum number of times where the absolute error was lower or equal to 0.01. Then, we aggregated the records from the multiple walks (about 550 records per walk). As a result, each of the six device datasets contains between 5,332 and 5,893 records.

A visualization of the values of the six datasets with respect to the reference dataset (red line) is presented in Figure 3. As we can see, although all six datasets contain records which correspond to the red line, this is not the case for all records and intuitively we will expect that such records have lower consistency. Moreover, Figure 3 illustrates that for some devices the data tend to be closer to the route than for others and in such cases we will expect them to have a higher consistency.



a) Dataset of Device 1 b) Dataset of Device 2

c) Dataset of Device 3 d) Dataset of Device 4

e) Dataset of Device 5 f) Dataset of Device 6

Figure 3. Visualization of the Values of the Six Datasets

For the six datasets, both the consistency of the records and of the whole dataset were determined. To do this, the proposed consistency metric was implemented in Java for both column and row rules as discussed in the previous section. In the following, for simplicity and due to length limitations, we will illustrate the application of the metric only for row rules (the results from the application of the column rules are analogous). Row rules are also more suitable for the given context since both the latitude and the

longitude coordinates are needed to identify a position. In order to determine the consistency values we used Equation (2). The two-sided $p-Value$ was calculated by doubling the one-sided $p-Value$ and clipping the value to one (Dunne et al. 1996). As described above, to determine the minimal number of required records for each rule, we applied a one-sided t-test based on the binomial distribution. To achieve this minimum requirement at a level of significance of 5%, we aggregated the values by building intervals based on values with four decimal places. An example for a row rule is $r_1: \top \rightarrow latitude = 49.0185 \wedge longitude = 12.0936$ ($p(r_1) = 0.014$), which means that 1.4% of the records in the reference dataset have a latitude in the interval represented by 49.0185° (i.e. [49.01845°, 49.01854°]) and a longitude in the interval represented by 12.0936° (i.e. [12.09355°, 12.09364°]).

To better illustrate the idea behind our evaluation, in Table 2 we consider two different intervals from the reference dataset. For each interval, we provide its probability based on the reference dataset, its relative frequency from each of the six device datasets, and its consistency. As we can see, the same interval can have different consistency values depending on the dataset. The reason for this is that when measuring consistency, we compare the relative frequency of records in the interval in the particular device dataset to the corresponding probability derived from the reference dataset. If these two values are far from each other, then consistency is low. To illustrate the idea, consider Interval II) in Table 2 and the dataset for device 1. The probability for Interval II) in the reference dataset is 0.0105. Thus, it is much higher than the respective frequency of records in this interval in dataset 1 (i.e., 0.0030). This implies that the observed number of records in dataset 1 that fall into Interval II) is much lower than the expected number of records based on the reference dataset. As a result, consistency is low. Generally, the closer the two probabilities (or numbers of records) are to each other, the higher the metric value is. For instance, the probability in dataset 5 is much closer to 0.0079 than the probability in dataset 1 and as a result its consistency is higher. Figure 4 illustrates this idea for datasets 1 and 5. We chose these datasets due to the fact that they have a similar number of records and thus instead of comparing the probabilities, we can directly compare the frequencies. As we can see, in dataset 1 (16 white squares), there are less points which lie in the Interval II) (orange rectangle) as opposed to dataset 5 (41 red squares). This is also the reason for the lower consistency. A metric value of one would require that 56 and 58 records of dataset 1 and 5, respectively, lie in Interval II).

Nr.	Interval (latitude, longitude)	Reference Probability	Dataset Frequency (Probability)	Device	Consistency Value
I)	latitude in [49.01845-49.01854] longitude in [12.09305-12.09314]	0.0092	49/5332 (0.0092)	1	1
			32/5883 (0.0054)	2	0.0017
			45/5845 (0.0077)	3	0.2662
			45/5839 (0.0077)	4	0.2694
			31/5478 (0.0057)	5	0.0048
			40/5893(0.0069)	6	0.0566
II)	latitude in [49.01995-49.02004] longitude in [12.09525-12.09534]	0.0079	16/5332 (0.0030)	1	8.3E-06
			69/5883 (0.0117)	2	0.0020
			12/5845 (0.0021)	3	5.1E-09
			45/5839 (0.0077)	4	0.9773
			41/5478 (0.0075)	5	0.8346
			23/5893 (0.0039)	6	0.0002

Table 2. Consistency Values of Single Records of the Six Datasets

Figure 4. Visualization of the Results from Interval II for Dataset 1 (white) and 5 (red)

Besides analyzing the consistency of records, we also investigate whether there is a significant difference regarding the consistency values on the dataset level for the six device datasets. We used the parametric ANOVA method at a 5% level of significance (Mason et al. 2003) and, to consider the case where the assumption for a normal distribution may not be satisfied, the non-parametric Kruskal-Wallis test at a 5% level of significance. Both showed that the consistency values significantly differ among the six datasets. In order to identify the ranking of the different datasets with regard to their consistency, we conducted post-hoc analysis by applying the Tukey HSD test (Mason et al. 2003). This test allows determining whether the difference between two datasets is significant and whether it is positive or negative (i.e., which dataset has higher consistency). The results for each pair of datasets are presented in Table 3, where in case of significance we added an indication about the direction of the relationship. For example, the value 5.10E-06*(-) in the third row indicates that on average the consistency values in dataset 3 were lower than the consistency values in dataset 6 with ≈99.999% confidence. Based on Table 3, we receive the following ranking for the consistency of the different datasets: dataset 4 > dataset 1 > dataset 2 > datasets 3, 5, 6 and dataset 6 > dataset 3. To demonstrate the aggregation possibilities of the metric, based on Equation (5) we obtain the following ranking: dataset 4 > dataset 1 > dataset 2 > dataset 6 > dataset 5 > dataset 3, which supports the results from the test. The reason why datasets 3 and 5 are ranked together based on the Tukey HSD test and separately based on Equation (5) is that, even though the average consistencies of the two datasets differ, this difference is not significant.

	Dataset 2	**Dataset 3**	**Dataset 4**	**Dataset 5**	**Dataset 6**
Dataset 1	6.28E-06*(+)	0*(+)	1.15E-07*(-)	0*(+)	6.01E-14*(+)
Dataset 2		5.72E-14*(+)	2.80E-14*(-)	6.52E-14*(+)	3.07E-07*(+)
Dataset 3			0*(-)	0.2483	5.10E-06*(-)
Dataset 4				0*(+)	0*(+)
Dataset 5					0.0523
Legend: * 5% significance, (+) on average the consistency of the row dataset is higher, (-) on average the consistency of the row dataset is lower					

Table 3. P-values for the Differences in the Consistency of the Datasets

So far, we discussed (1) the applicability and efficacy by determining and analyzing the consistency values of the six datasets on a record and dataset level. Next, we will address (2), which means the validity of our approach. To determine the perceived ranking of the consistency of the devices based on Figure 3, we

conducted a survey among 27 IS professionals. The results were then aggregated based on Borda's method (Dwork et al. 2001). This method works as follows: to each dataset in a given response, we assign a weight corresponding to the number of datasets that were ranked below the dataset in question. For example, if dataset 1 was ranked first, it will be assigned a weight of five. Then the weights are summed over the responses providing a number for each dataset. Based on these numbers, the datasets are ranked in decreasing order. The application of Borda's method delivered the following ranking: dataset 2 > dataset 4 > dataset 6 > dataset 1 > dataset 3 > dataset 5 where we should note that the difference between dataset 2 and dataset 4 was not that high (115 vs. 110 points). This supports the results discussed above except for dataset 1. The reason is, as mentioned above, that the corresponding relative frequencies of the records in dataset 1 are closer to the probability based on the reference dataset than the relative frequencies of the records in datasets 2 and 6.

To sum up, in this section we evaluated our approach by applying it to the GPS measurements of six different mobile devices. We first illustrated the application of the metric by presenting and discussing the results from the single records. Then, we demonstrated the strength of our approach by comparing the ranking of the six datasets with regard to their consistency to the ranking which we determined through a survey among IS professionals.

Conclusion and Future Work

In this paper we provide a metric for (semantic) consistency based on probability theory and a predefined set of rules. We measure consistency by statistically comparing the probability of fulfillment of a rule in a reference dataset with the relative frequency with which this rule is fulfilled in the dataset to be assessed. Thus, as opposed to existing approaches in the literature, we consider (1) rule uncertainty in the assessment of consistency. As a result, our metric is much more generally applicable than metrics that only consider rules to be "true by definition". In particular, in our approach a rule is satisfied according to a Bernoulli distribution and we define our consistency metric as a two-sided p-value. Thus, the metric results can be interpreted as the probability that the assessed attribute values do not contradict the predefined set of rules and thus correspond to the attribute values in the reference data. Therefore, our metric has (2) a clear interpretation. We provide a metric for the level of attribute values as well as aggregation functions on the level of records, relations, and databases. Further, we present different possibilities for metric instantiation and in particular for the determination of a suitable rule set. We evaluate our approach by measuring the consistency of experimentally generated GPS measurements of mobile devices and using data from GoogleMaps as a reference dataset. Moreover, to confirm the validity of our metric, we conduct a survey among IS professionals.

Since the results of our metric can be interpreted as probabilities, future research should concentrate on considering them in decision support in a well-founded way. For example, if the consistency of a given attribute value (or tuple) is low, the decision maker may consider this in the calculation of the expected utility. Another possible path for future research is to develop other aggregation procedures which also take the statistical properties of the consistency value into account. For instance, the aggregation can be defined based on the sum of random variables following a Bernoulli distribution and thus also be interpreted as p-value. Furthermore, we evaluated our metric by using the proposed column and row rules, but future research should apply other types of rules such as association rule mining and rules derived by experts. Finally, our metric is defined for structured data. However, it can be extended to semi- and unstructured data by applying text mining methods such as inverted term frequency.

References

Agrawal, R., Imieliński, T., and Swami, A. 1993. "Mining association rules between sets of items in large databases," in *ACM SIGMOD Record*, pp. 207–216.

Alkharboush, N., and Yuefeng Li 2010. "A Decision Rule Method for Assessing the Completeness and Consistency of a Data Warehouse," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Alpar, P., and Winkelsträter, S. 2014. "Assessment of data quality in accounting data with association rules," *Expert Systems with Applications* (41:5), pp. 2259–2268.

Baker, E., and Olaleye, O. 2013. "Combining Experts: Decomposition and Aggregation Order," *Risk Analysis* (33:6), pp. 1116–1127.
Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. "Methodologies for data quality assessment and improvement," *ACM Computing Surveys* (41:3), pp. 1–52.
Batini, C., and Scannapieco, M. 2006. *Data-Centric Systems and Applications*: *Concepts, Methodologies and Techniques:* Springer.
Blake, R. H., and Mangiameli, P. 2009. "Evaluating the Semantic and Representational Consistency of Interconnected Structured and Unstructured Data," in *Proceedings of the Americas Conference on Information Systems (AMCIS 2009)*.
Blum, J., Greencorn, D., and Cooperstock, J. 2013. "Smartphone Sensor Reliability for Augmented Reality Applications," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, K. Zheng, M. Li and H. Jiang (eds.): Springer Berlin Heidelberg, pp. 127–138.
Chiang, F., and Miller, R. J. 2008. "Discovering data quality rules," *Proceedings of the VLDB Endowment* (1:1), pp. 1166–1177.
Cordts, S. 2008. *Implementierung eines Datenqualitätsdienstes zur evolutionären Datenqualitätsverbesserung in relationalen Datenbankmanagementsystemen (in German)*. Dissertation, University of Hamburg.
Della Valle, E., Celino, I., Dell'Aglio, D., Kim, K., Huang, Z., Tresp, V., Hauptmann, W., Huang, Y., and Grothmann, R. 2008. "Urban Computing: a challenging problem for Semantic Technologies," in *2nd International Workshop on New Forms of Reasoning for the Semantic Web (NEFORS 2008) co-located with the 3rd Asian Semantic Web Conference (ASWC 2008)*.
D'Roza, T., and Bilchev, G. 2003. "An Overview of Location-Based Services," *BT Technology Journal* (21:1), pp. 20–27.
Dunne, A., Pawitan, Y., and Doody, L. 1996. "Two-sided P-values from discrete asymmetric distributions based on uniformly most powerful unbiased tests," *The Statistician* , pp. 397–405.
Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. 2001. "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web (WWW '01)*.
Eppler, M. J. 2006. *Managing information quality: Increasing the value of information in knowledge-intensive products and processes:* Springer Science & Business Media.
Fan, W., Geerts, F., Tang, N., and Yu, W. 2013. "Inferring data currency and consistency for conflict resolution," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pp. 470–481.
Fisher, C. W., Lauria, Eitel J. M., and Matheus, C. C. 2009. "An Accuracy Metric: Percentages, randomness, and probabilities," *Journal of Data and Information Quality* (1:3), pp. 1–21.
Forbes Insights 2010. "Managing Information in the Enterprise: Perspectives for Business Leaders," Forbes.
Goodman, S. 2008. "A dirty dozen: twelve p-value misconceptions," in *Seminars in hematology*, pp. 135–140.
Heinrich, B., and Hristova, D. 2014. "A Fuzzy Metric for Currency in the Context of Big Data," in *22nd European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014,* Tel Aviv, Israel.
Heinrich, B., Kaiser, M., and Klier, M. 2007. "Metrics for Measuring Data Quality - Foundations for an Economic Oriented Management of Data Quality," in *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT)*: INSTICC/Polytechnic Institut of Setúbal Barcelona, Spain.
Heinrich, B., and Klier, M. 2015. "Metric-based data quality assessment—Developing and evaluating a probability-based currency metric," *Decision Support Systems* (72), pp. 82–96.
Helfert, M. 2002. *Planung und Messung der Datenqualität in Data-Warehouse-Systemen (in German)*. Dissertation.
Henschen, D. 2013. *2014 Analytics, BI, and Information Management Survey*. http://reports.informationweek.com/abstract/81/11715/Business-Intelligence-and-Information-Management/Research:-2014-Analytics,-BI,-and-Information-Management-Survey.html. Accessed 19 February 2015.
Hinrichs, H. 2002. *Datenqualitätsmanagement in Data Warehouse-Systemen (in German)*. Dissertation.
Hipp, J., Güntzer, U., and Grimmer, U. 2001. "Data Quality Mining-Making a Virtue of Necessity," in *DMKD*.

Hipp, J., Müller, M., Hohendorff, J., and Naumann, F. 2007. "Rule-Based Measurement Of Data Quality In Nominal Data," in *12th International Conference on Information Quality*, pp. 364–378.
Kotsiantis, S., and Kanellopoulos, D. 2006. "Association rules mining: A recent overview," *GESTS International Transactions on Computer Science and Engineering* (32:1), pp. 71–82.
Kübart, J., Grimmer, U., and Hipp, J. 2005. "Regelbasierte Ausreißersuche zur Datenqualitätsanalyse (in German)," (14), pp. 22–28.
Lee, Y. W., Pipino, L., Strong, D. M., and Wang, R. Y. 2004. "Process-embedded data integrity," *Journal of Database Management (JDM)* (15:1), pp. 87–103.
Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y. 2006. *Journey to Data Quality:* The MIT Press.
Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. "AIMQ: a methodology for information quality assessment," *Information & management* (40:2), pp. 133–146.
Litwin, M. S. 1995. *How to Measure Survey Reliability and Validity:* SAGE Publications.
Liu, L., and Chi, L. N. 2002. "Evolutional data quality: a theory-specific view," in *Proceedings of the Seventh International Conference on Information Quality (ICIQ-02)*, pp. 292–304.
Manyika, J., Chui, M., Groves, P., Farrell, D., van Kuiken, S., and Almasi Doshi, E. 2013. "Open data: Unlocking innovation and performance with liquid information," McKinsey Global Institute.
Mason, R. L., Gunst, R. F., and Hess, J. L. 2003. *Statistical design and analysis of experiments: with applications to engineering and science:* John Wiley & Sons.
Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T., and Batini, C. 2002. "Managing Data Quality in Cooperative Information Systems," in *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, R. Meersman and Z. Tari (eds.): Springer Berlin Heidelberg, pp. 486–502.
Meyer, M. A., and Booker, J. M. 2001. *Eliciting and analyzing expert judgment: a practical guide:* SIAM.
Mezzanzanica, M., Cesarini, M., Mercorio, F., and Boselli, R. 2012. "Towards the Use of Model Checking for Performing Data Consistency Evaluation and Cleansing," in *Proceedings of the 17th International Conference on Information Quality, IQ 2012, Paris, France, November 16-17, 2012*, Laure Berti-Equille, Isabelle Comyn-Wattiau and Monica Scannapieco (eds.), MIT, pp. 163–177.
Moges, H.-T., Dejaeger, K., Lemahieu, W., and Baesens, B. 2011. "Data quality for credit risk management: new insights and challenges," in *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)*, Adelaide, Australia, pp. 632–646.
Orr, K. 1998. "Data quality and systems theory," *Communications of the ACM* (41), pp. 66–71.
Parssian, A., Sarkar, S., and Jacob, V. S. 2004. "Assessing data quality for information products: impact of selection, projection, and Cartesian product," *Management Science.* (50:7), pp. 967–982.
Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data quality assessment," *Communications of the ACM* (45:4), pp. 211–218.
Redman, T. C. 1996. *Data Quality for the Information Age:* Artech House.
Scannapieco, M., Missier, P., and Batini, C. 2005. "Data Quality at a Glance," *Datenbank-Spektrum* (14), pp. 6–14.
Shankaranarayanan, G., Iyer, B., and Stoddard, D. 2012. "Quality of Social Media Data and Implications of Social Media for Data Quality," in *17th International Conference on Information Quality, IQ 2012, Paris, France, November 16-17, 2012*, Laure Berti-Equille, Isabelle Comyn-Wattiau and Monica Scannapieco (eds.), MIT, pp. 311–325.
Wand, Y., and Wang, R. Y. 1996. "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM* (39:11), pp. 86–95.
Wang, R. Y., and Strong, D. M. 1996. "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.
Wei-Liang, C., Shi-Dong, Z., and Xiang, G. 2009. "Anchoring the Consistency Dimension of Data Quality Using Ontology in Data Integration," in *Web Information Systems and Applications Conference (WISA), 2009*, pp. 201–205.
Witchalls, C. 2014. *Gut & gigabytes*: *Capitalising on the art & science in decision making*. http://www.economistinsights.com/business-strategy/analysis/gut-gigabytes. Accessed 19 February 2015.

4.2 Contribution to RQ 3

In this chapter, the following research question was addressed:

RQ 3 How can *consistency* be adequately measured?

RQ 3 was addressed by proposing a metric for consistency based on probability theory (*Measure* phase). The metric uses a predefined set of rules and identifies contradictions to these rules by comparing their probability of fulfillment in a consistent reference dataset with the one in the assessed dataset. Thus, it models consistency adequately, by considering rule uncertainty which exists in a wide range of cases. The metric results have a clear interpretation in terms of probability, which facilitates their consideration in decision making (*Analyze* phase) and in information quality management (*Improve* phase). This completes the discussion of the information quality dimensions currency, accuracy, and consistency. In the next section, necessary requirements for information quality metrics for data views are presented.

5. Requirements for Information Quality Metrics

In the previous chapters, approaches for measuring and analyzing particular information quality dimensions were presented. Their main idea was to model the quality uncertainty stemming from less than perfect information quality by applying probability theory or fuzzy set theory. However, even though these measurement approaches were evaluated with regard to different criteria, such as validity and accuracy, an important question in this context remains: How can the adequacy of an information quality metric for decision support (*Analyze* phase) as well as its efficiency from a cost-benefit perspective (*Improve* phase) be examined? To address this question, in this chapter, five requirements for information quality metrics for data views are developed and formally proven. They are applicable to metrics for different information quality dimensions of data views (*Define* phase) and also cover different methodologies for modeling quality uncertainty.

5.1 Paper 7: Requirements for Data Quality Metrics

Full citation: Heinrich, B., Hristova, D. and Klier, M. (2015), *Requirements for Data Quality Metrics*, Working Paper, University of Regensburg.

Current status: Submitted to the journal Business & Information Systems Engineering, submission of the 1st revised manuscript after a review with major revision on 04.02.2015, under review.

Highlights: In this paper, a set of five requirements for information quality metrics for data views is developed and proven. They have an axiomatic nature and are justified based on a sound theoretical foundation. This foundation is in accordance with Figure 2 and considers both decision making with environmental uncertainty and information quality management. The requirements contribute to the literature in two ways. First, as opposed to existing works, they have a clear interpretation, which is very important for practical applications. Second, the necessity of the requirements is formally proven and as a result metrics which do not satisfy them cannot adequately (*directly*) support decision making with environmental uncertainty and economically oriented information quality management. The requirements can be applied for both the verification and improvement of existing metrics and for the adequate design of new metrics.

Evaluation: Similar to Paper 5, this paper differs from Papers 1, 2, 3, 4, and 6 in that it does not follow a normative modeling approach. As opposed to them, it defines a set of necessary axioms (or theorems) following the reason/logic/theorems approach. In that sense, the evaluation focuses on the applicability and efficacy of the requirements by providing a detailed analysis of two well-known metrics from the literature. The results show that the requirements are easy to verify, but neither trivial nor impossible to fulfill.

Limitations: The requirements presented in this paper have two main limitations. First, they are designed to consider the information quality of data views and in particular of structured data. As a result, future research has to investigate whether it is possible to transfer them to metrics measuring other types of data (cf. Variety in big data). Second, although the requirements represent a

set of necessary conditions, it is not possible to prove that these conditions are also sufficient. Therefore, future work should iteratively evaluate and enhance them. Finally, the requirements focus on the *direct* consideration of information quality in decision making with environmental uncertainty. Thus, it would be also the task of future research to extend them to the *indirect* consideration through the KDD process, as demonstrated in Section 2.2.

Requirements for Data Quality Metrics

Teaser

Many information quality metrics are developed to support decision making under uncertainty and an economically oriented management of information quality. However, if not adequately defined, these metrics can lead to wrong decisions and economic losses. Thus, based on a decision-oriented framework we propose a set of five requirements for information quality metrics. These requirements are widely applicable and not restricted to a certain information quality dimension. Furthermore, they are easy to verify, but neither trivial, nor impossible to fulfill. The requirements can be considered in the development of information quality metrics and in the assurance of their adequacy.

Abstract

Information quality and especially the assessment of information quality have been intensively discussed in research and practice alike. To adequately support an economically oriented management of information quality and decision making under uncertainty, it is essential to assess the information quality level by means of well-founded metrics. Therefore, based on a decision-oriented framework, we present a set of five requirements for information quality metrics. If these requirements are met, the respective metric and its values are not capable of supporting an economically oriented management of information quality and decision making under uncertainty. Finally, we demonstrate the applicability and efficacy of these requirements by evaluating two well-known information quality metrics.

Keywords: *Information quality, Information quality assessment, Information quality metrics, Requirements for metrics*

1 Introduction

Due to the rapid technological development, companies increasingly rely on data to support decision making and to gain competitive advantage. To make informed and effective decisions, it is crucial to assess and assure the quality of the underlying data. 91% of the respondents of a survey conducted by Experian QAS (2013) state that failing to do so has resulted in wrong decisions and wasted budgets. In addition, 82% of the participants in a Forbes Insights (2010) survey reveal that poor IQ leads to costly mistakes for businesses and 51% of them estimate their annual losses due to IQ problems to be at least 5 million US-$. In the light of the current proliferation of big data, with large amounts of heterogeneous,

quickly-changing data from distributed sources being analyzed to support decision making, assessing and assuring IQ become even more relevant (IBM Global Business Services 2012; Buhl et al. 2013). Indeed, the three characteristics Volume, Velocity and Variety, which are often called the three Vs of big data, make the assurance of IQ even more challenging (e.g. due to the integration of various data sources) and the consequences of wrong decisions even more costly (Economist Intelligence Unit 2011; SAS Institute 2013). This has resulted in the addition of a fourth V (=Veracity) reflecting the importance of IQ in the context of big data (IBM Global Business Services 2012; Lukoianova and Rubin 2014).

IQ can be defined as "the measure of the agreement between the data views presented by an information system and that same data in the real world" (Orr 1998, p. 67; cf. also Parssian et al. 2004; Heinrich et al. 2009). IQ is a multi-dimensional construct (Redman 1996; Lee et al. 2002; Eppler 2003) comprising different IQ dimensions like currency, completeness, consistency, and accuracy (Wang et al. 1995). Each IQ dimension provides a particular perspective on the quality of data views. As a result, researchers have developed corresponding metrics for the quantitative assessment of these dimensions for data views (Ballou et al. 1998; Hinrichs 2002; Even and Shankaranarayanan 2007; Heinrich et al. 2007; Fisher et al. 2009; Heinrich et al. 2009; Blake and Mangiameli 2011; Heinrich et al. 2012; Wechsler and Even 2012), and such metrics are in the focus of this paper.

IQ metrics are needed for two main reasons. First, the metric values are used to support data-based decision making under uncertainty. Here, well-founded IQ metrics are required to indicate to what extent decision makers should rely on the underlying data values. Second, the metric values are used to support an economically oriented management of IQ (cf. e.g. Wang 1998; Heinrich et al. 2009). In this context, IQ improvement measures should be applied, if and only if the benefits (due to higher IQ) outweigh the associated costs. To be able to analyze which IQ improvement measures are efficient from an economic perspective, well-founded IQ metrics are necessary to assess (the changes in) the IQ level.

While both research and practice have realized the high relevance of well-founded IQ metrics, many IQ metrics still lack an appropriate methodical foundation, as they are either developed on an ad hoc basis to solve specific problems (cf. e.g. Pipino et al. 2002), or are highly subjective (cf. e.g. Cappiello and Comuzzi 2009). Thus, both researchers and practitioners set out to propose requirements for IQ metrics (e.g. Pipino et al. 2002; Even and Shankaranarayanan 2007; Heinrich et al. 2007; Mosley et al. 2009; Loshin 2010; Hüner 2011). Most of them, however, did not aim at justifying the requirements based on a

methodical foundation. As a result, the literature on this topic is fragmented, and it is not clear which requirements are indeed necessary to support decision making in a well-founded way. Moreover, as some of the requirements leave room for interpretation, their verification is difficult and subjective. This results in a research gap, which we aim to address by answering the following research question:

Which clearly defined requirements must an IQ metric satisfy to adequately support both decision making under uncertainty and an economically oriented management of IQ?

To address this research question, we propose a set of five clearly defined requirements for IQ metrics (e.g. the metric values should be normalized and also interval-scaled). We take the existing literature into account and justify our set of requirements based on a decision-oriented framework. As a result, our requirements are necessary to adequately support both decision making under uncertainty and an economically oriented management of IQ. Thus, IQ metrics which do not meet them either lead to wrong decisions or to economic losses (e.g. because the efficiency of the metric's application is not ensured).

We follow the axiomatic-artificial perspective according to Meredith et al. (1989) and apply the reason/logic/theorems methodology. The requirements for IQ metrics have an axiomatic nature (cf. "work place ideology") and are defined and justified by means of "formal procedures" in the form of "logico-mathematical theorems" (Meredith et al. 1989, p. 305 and p. 314). To demonstrate the applicability and efficacy of our requirements for the considered "object reality" (Meredith et al. 1989, p. 308), we apply them to two renowned IQ metrics from literature. Our evaluation demonstrates that the requirements are straightforward to verify and that they are neither trivial nor impossible to fulfill, which is important for the application and the acceptance of the requirements in practice.

The remainder of the paper is structured as follows. In the next section, we provide an overview of the related work and identify the research gap. Section 3 comprises the decision-oriented framework which is the theoretical foundation for our work. In Section 4, we propose a set of five requirements for IQ metrics which are defined and justified based on this framework. In Section 5, we demonstrate the applicability and the efficacy of these requirements with the help of two renowned IQ metrics. The last section provides conclusions, limitations and directions for future research.

2 Related work

In this section, we analyze existing works which propose requirements for IQ metrics. Following the guidelines of standard approaches to prepare the related work (e.g. Webster and Watson 2002; Fettke 2006; Levy and Ellis 2006; Vom Brocke et al. 2009), we searched the databases Science direct, ACM Digital Library, EBSCO Host, IEEE Xplore, and the AIS Library as well as the Proceedings of the International Conference on Information Quality (ICIQ) for the following search term and without posing a restriction on the time period: *("data quality" and metric* and requirement*) or ("data quality" and metric* and standard*) or ("information quality" and metric* and requirement*) or ("information quality" and metric* and standard*)*. This search amounted to a total of 136 papers which were manually screened based on their title, abstract, and keywords. The remaining 43 papers were analyzed in detail and could be divided into three main groups: A) requirements for IQ metrics and metric values, B) requirements for the IQ assessment process in general, and C) requirements and (practical) recommendations for the integration of IQ metrics within organizations (e.g. within business processes). We focused on group A) consisting of five relevant papers on which we performed an additional forward and backward search, resulting in a total of eight relevant papers which are presented in the following.

Pipino et al. (2002) propose the functional forms *simple ratio*, *min or max operation*, and *weighted average* to develop IQ metrics. *Simple ratio* measures the ratio of the number of desired outcomes (e.g. number of accurate data units) to the total number of outcomes (e.g. total number of data units). *Min or max operation* can be used to define IQ metrics requiring the aggregation of multiple assessments, for instance, on the level of data values, tuples, or relations. Thereby, the minimum (or maximum) value among the normalized values of the single assessments is calculated. *Weighted average* is an alternative to the *min or max operation* and represents the weighted average of the single assessments. The major goal of Pipino et al. (2002) is to present feasible and useful functional forms which can be seen as a first important step towards requirements for IQ metrics. They ensure the range [0; 1] for the metric values and address the aggregation of multiple assessments.

Even and Shankaranarayanan (2007) aim at an economically oriented management of IQ. They propose four consistency principles for IQ metrics. *Interpretation consistency* states that the metric values on different data view levels (data values, tuples, relations, and the whole database) must have a consistent semantic interpretation. *Representation consistency* requires that the metric values are interpretable for business users (typically in the range [0; 1] with

respect to the utility resulting from the assessed data). *Aggregation consistency* states that the assessment of IQ on a higher data view level has to result from the aggregation of the assessments on the respective lower level. The aggregated result should take values which are not higher than the highest or lower than the lowest metric value on the respective lower level. *Impartial-contextual consistency* means that IQ metric values should reflect whether the assessment is context-dependent or context-free.

Heinrich et al. (2007; 2009; 2012) analyze how IQ can be assessed by means of metrics in a goal-oriented and economic manner. To evaluate IQ metrics, they define six requirements. *Normalization* requires that the metric values fall into a bounded range (e.g. [0; 1]). *Interval scale* states that the difference between any two metric values can be determined and is meaningful. *Interpretability* means that the metric values have to be interpretable, while *aggregation* states that it must be possible to aggregate metric values on different data view levels. *Adaptivity* requires that it is possible to adapt the metric to the context of a particular application. *Feasibility* claims that the parameters of a metric have to be determinable and that this determination must not be too cost-intensive. Moreover, this requirement states that it should be possible to calculate the metric values in an automated way.

Mosley et al. (2009) and Loshin (2010) discuss requirements for IQ metrics from a practitioners' point of view. Both contributions comprise the requirements *measurability* and *business relevance* claiming that IQ metrics have to take values in a discrete range and that these values need to be connected to the company's performance. Loshin (2010) adds that it is important to clearly define the metric's goal and to provide a value range and an interpretation of the parts of this range (*clarity of definition*). In addition, Mosley et al. (2009) require *acceptability*, which implies that a metric is assigned a threshold at which the IQ level meets business expectations. If the metric value is below this threshold, it has to be clear who is accountable and in charge to take improvement actions. The corresponding requirements *accountability/stewardship* and *controllability*, however, refer to the integration of an IQ metric within organizations (cf. Group C) and are thus not within the focus of this paper. The same holds for the requirements *representation* and *reportability* found in both works and also *drill-down capability* by Loshin (2010). *Representation* claims that the metric values should be associated with a visual representation, *reportability* points out that they should provide enough information to be included in aggregated management reports, and *drill-down capability* states that it should be possible to identify an IQ metric's impact factors within the

organization. Finally, *trackability* which requires a metric to be repeatedly applicable at several points of time (cf. Group B) is also beyond the focus of this paper.

Hüner (2011) proposes a method for the specification of business oriented IQ metrics to support both the identification of business critical data defects and the repeated assessment of IQ. Based on a survey among experts, he specifies 21 requirements for IQ assessment methods (cf. Appendix B). However, only some of them constitute requirements for IQ metrics and metric values (cf. Group A) and are thus considered further. These are *cost/benefit*, *definition of scale*, *validity range*, *comparability*, and *comprehensibility*. The other requirements refer to Group B (e.g. *repeatability*, *definition of measurement frequency*, *definition of measurement point, definition of measurement procedure*) or Group C (e.g. *responsibility*, *escalation process*, *use in SLAs*) and are not within the focus of this paper.

To sum up, prior works provide valuable contributions by stating a number of possible requirements for IQ metrics and their respective values. While some of them overlap, the existing literature is still very fragmented. In addition, many of the requirements are not clearly defined which makes their application and verification very difficult. To address these issues, we organize the existing requirements in six groups with each group being characterized by a clear, unique characteristic (cf. Table 1). Note that some of the requirements which leave room for interpretation (cf. brackets in Table 1) are classified in more than one group.

Group 1 comprises requirements stating that IQ metrics have to take values within a given range. *Simple ratio* and *representation consistency* aim at metric values in the range [0; 1]. *Measurability* results in a bounded range defined by the lowest and the highest discrete value. Hence, these requirements as well as *clarity of definition* (with respect to the range), *normalization* and *validity range* are assigned to this group. Group 2 contains requirements regarding the scale of measurement of the metric values. Since *definition of scale* may not only concern the interpretation of the metric values but also their scale, this requirement is included as well. Group 3 covers requirements claiming an interpretation of the metric values. Here, *clarity of definition* is interpreted as *interpretability*. In addition, metric values satisfying the *simple ratio* requirement can be interpreted as a percentage, and *interpretation consistency* requires a consistent semantic interpretation of the metric values regardless of the hierarchical level. While *comparability*, *comprehensibility* and *definition of scale* require some kind of interpretation of the metric values (e.g. as a percentage), *representation consistency* directly implies a clear interpretation with respect to the utility of the data under

consideration. The requirements in Group 4 state that IQ metrics should be able to adequately consider the particular context of application, for example by means of weights that decrease or increase the influence of contextual characteristics. Group 5 focusses on the application of an IQ metric from a cost-benefit perspective. *Feasibility* is part of this group, because it requires that the costs for determining a metric's parameters are taken into account and that it should be possible to calculate the metric values in a widely automated way – a fact that results in lower application costs. *Business relevance* implies that a metric goes along with some benefit for the company, whereas *acceptability* is part of this group because business expectations are defined considering a cost-benefit perspective. Finally, Group 6 concerns the (consistent) aggregation of the metric values on different data view levels. *Min or max operation* and *weighted average* state how this aggregation has to be done and *interpretation consistency* requires the same interpretation of the metric values on all data view levels.

Group	Requirements
1	*normalization, validity range, clarity of definition (range), simple ratio (bounded in [0;1]), representation consistency (range), measurability*
2	*interval scale, definition of scale (scale)*
3	*interpretability, clarity of definition (interpretation), simple ratio (interpretation), interpretation consistency (interpretation), comparability, comprehensibility, definition of scale (interpretation), representation consistency (interpretation)*
4	*weighted average (context), impartial-contextual consistency, adaptivity*
5	*cost/benefit, feasibility, acceptability, business relevance*
6	*aggregation consistency, aggregation, min or max operation, weighted average (aggregation), interpretation consistency (aggregation)*

Table 1: Groups of requirements

Table 1 provides an overview of the existing requirements for IQ metrics, which are partly fragmented and vaguely defined. Prior work does in fact lack a formal methodical framework and does not aim at stating and justifying which requirements for IQ metrics are necessary to adequately support decision making under uncertainty and an economically oriented management of IQ. To address this research gap, in the next section we present a decision-oriented framework, which serves as a theoretical foundation for stating and justifying a set of clearly defined requirements for IQ metrics in Section 4.

3 Theoretical Foundation

The decision-oriented framework which serves as a theoretical foundation for our work is based on the following fields: i) decision making under uncertainty by considering the influence of assessed IQ metric values and ii) economically oriented management of IQ by considering the costs and benefits of applying IQ metrics[1].

The literature on decision making under uncertainty (and in particular under risk) uses the well-known concept of decision matrices to represent the situation decision makers are faced with (Adam 1996; Nitzsch 2006; Laux 2007; Peterson 2009). Decision makers can choose among a number of alternatives, while the corresponding payoff depends on the state of nature. Each possible state of nature occurs with a certain probability. Hence, in case of a risk-neutral decision maker (if this is not the case, the payoffs can be replaced by risk-adjusted utilities), the one alternative is chosen which results in the highest expected payoff when considering the probability distribution over all possible states of nature. Table 2 illustrates a decision matrix for a simple situation with two alternatives a_i (i = 1, 2), two possible states of nature s_j (j = 1, 2), and the respective payoffs p_{ij} for each pair (a_i, s_j). The probabilities of occurrence of the possible states of nature are represented by $w(s_j)$. To select the alternative with the highest expected payoff, the decision maker has to compare the expected payoffs for choosing alternative a_1 (i.e. $p_{11}w(s_1) + p_{12}w(s_2)$) and alternative a_2 (i.e. $p_{21}w(s_1) + p_{22}w(s_2)$). The two-by-two matrix serves for illustration purposes only. Generally, we represent the possible states of nature s_j (j = 1, …, n) by the vector $S = (s_1, s_2, \dots, s_n)$, the respective probabilities of occurrence by $w(s_j)$, the alternatives a_i (i = 1, …, m) by the vector $A = (a_1, a_2, \dots, a_m)$[2], and the payoffs for alternative a_i by the vector $P_i = (p_{i1}, p_{i2}, \dots, p_{in})$. The *expected* payoff for choosing alternative a_i is denoted by $E(a_i, P_i, S) = \sum_{j=1}^{n} p_{ij} w(s_j)$; the maximum expected payoff is given by $max_{a_i} E(a_i, P_i, S)$. An overview of the notation is provided in Appendix A.

[1] Note that i) may also be seen as an important tool for ii). However, due to the high relevance of i) in the context of DQ metrics, we decided to distinguish these two cases.

[2] In case of a continuous decision space, this will be a vector of infinitely many alternatives. If not all alternatives are known, the concept of bounded rationality is applied (Simon 1956; 1969; Jones 1999).

	Probability $w(s_1)$	Probability $w(s_2)$
	State s_1	State s_2
Alternative a_1	Payoff p_{11}	Payoff p_{12}
Alternative a_2	Payoff p_{21}	Payoff p_{22}

Table 2: Decision matrix

Requirements for IQ metrics must guarantee that the metric values can adequately support i) decision making under uncertainty. To address i) it is necessary to examine the influence of IQ and thus of the IQ metric values on the components of the decision matrix (i.e. the *probabilities of occurrence,* the *payoffs,* and the *alternatives*). In this respect, literature provides useful insights. Heinrich et al. (2012), for example, propose a metric for the IQ dimension currency. The metric values represent probabilities that the data values under consideration still correspond to their real-world value at the instant of assessing IQ. They apply the metric to determine the *probabilities of occurrence* (represented by the metric values) in a decision situation. The influence of IQ on the *payoffs* is considered, for example, by Ballou et al. (1998), Even and Shankaranarayanan (2007), and Cappiello and Comuzzi (2009). All of them argue that less than perfect IQ (represented by the IQ metric values) may affect and reduce the payoffs. Other works such as Fisher et al. (2003), Heinrich et al. (2007), and Jiang et al. (2007) examine the influence of IQ on the choice of the *alternative*.

To formally specify the influence of IQ on the decision matrix, let IQ represent the value of the IQ metric and $E(a_i, IQ, P_i, S)$ the *expected* payoff for choosing alternative a_i when considering IQ as well as the payoff vector P_i and the vector of states of nature S. Let further $max_{a_i} E(a_i, IQ, P_i, S)$ be the maximum *expected* payoff when considering IQ. It is obvious and in line with prior works (cf. above) that considering IQ may result in choosing a different optimal alternative as compared to not considering IQ (i.e. $a_1 = argmax_{a_i} E(a_i, IQ, P_i, S)$ and $a_2 = argmax_{a_i} E(a_i, P_i, S)$ with $a_1 \neq a_2$). Hence, it is indispensable to consider IQ by means of well-founded metrics in decision making under uncertainty.

To avoid inefficient or impractical metrics, it is further necessary to take the field of ii) economically oriented management of IQ into account when developing requirements for IQ metrics. Existing literature already addressed the question of whether or not to apply IQ improvement measures from a cost-benefit perspective (Campanella 1999; Feigenbaum 2004; Heinrich et al. 2007; 2012). Indeed, applying IQ improvement measures may increase the IQ level and thus bring benefits. At the same time, the associated costs have to be taken into account and the improvement measures should only be applied, if the benefits outweigh these

costs. In decision making, the benefits result from being enabled to choose a better alternative (i.e. with an additional expected payoff) due to the improved IQ. The costs include the ones for conducting the improvement measures as well as the ones for assessing IQ by means of IQ metrics. The latter have rarely been considered in the literature, even though they play an important role and must not be neglected. Indeed, if applying an IQ metric is too resource-intensive, it may not be reasonable to do so from a cost-benefit perspective. Thus, requirements for IQ metrics have to explicitly take this aspect into account.

Based on the literature on i) and ii) and the above discussion, Figure 1 presents the decision-oriented framework which serves as a theoretical foundation to justify our requirements (for a similar illustration cf. Heinrich et al. 2007; 2009). IQ metrics are applied to data views to assess the IQ level (cf. I-III). The assessed IQ level (represented by the metric values) influences i) decision making under uncertainty and in particular the chosen alternative, and the expected payoff of the decision maker (cf. IV-VI). Thus, the decision maker may apply improvement measures to increase the IQ level represented by the metric values (cf. IX). However, applying IQ improvement measures creates costs (cf. VII). This also holds for the application of the metric including the determination of its parameters (cf. II). Hence, the optimal IQ level (cf. VIII) has to be determined based on an economical perspective.

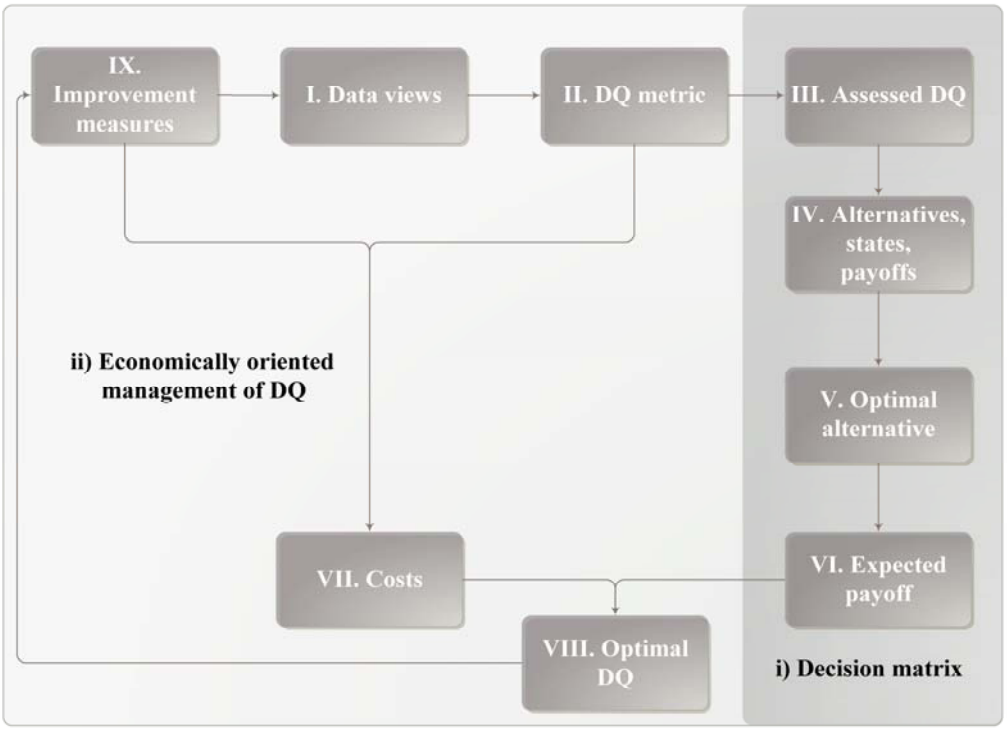


Figure 1: Decision-oriented framework

4 Requirements for information quality metrics

In this section, we present a set of five clearly defined requirements for IQ metrics. They combine, specify, and enhance existing approaches covering the six groups of requirements identified in Section 2. Moreover, based on the decision-oriented framework, we justify and formally prove that our requirements are *necessary* conditions for IQ metrics to adequately support both i) decision making under uncertainty and ii) an economically oriented management of IQ.

Group 1 states that IQ metrics have to take values within a given range. Most of the requirements in this group (e.g. *validity range* and *clarity of definition*) are vaguely defined and thus difficult to verify. Moreover, none of them has been formally proven yet. Hence, both the indispensability of these requirements and the possible consequences of them not being fulfilled remain unclear (e.g. *measurability* claims that the range should be discrete). To address these issues, we propose and prove the following requirement:

Requirement 1 (R1): Normalized metric values

The metric values have to be bounded from below and from above and must attain their minimum (representing perfectly poor IQ) and maximum (representing perfectly good IQ).

Proof: To justify (R1) and evaluate its necessity, we provide a proof by contradiction. Let us assume that a given IQ metric does not fulfill (R1), but that its values are still unambiguous and meaningful in order to support decision making under uncertainty and an economically oriented management of IQ in a well-founded way (cf. Section 3). The fact that (R1) is not fulfilled implies that the values of the metric

(i) are not bounded from below and from above and/or
(ii) do not attain their minimum and/or maximum.

Let ω be a stored data value (e.g. a stored customer address) of perfectly good IQ. As a consequence, the data value ω perfectly represents the corresponding real-world value ω_m. The metric value for ω is given by $IQ(\omega, \omega_m)$.

Re (i): If there exists no upper bound for the metric values, another data value ω' exists which – compared to ω – results in a higher metric value (i.e. $IQ(\omega', \omega_m) > IQ(\omega, \omega_m)$). As higher metric values represent better IQ, this implies that ω' is of better IQ than ω. This is, however, a contradiction to the fact that ω is – by definition – of perfectly good IQ. Hence, the metric values must be bounded from above. The existence of a lower bound can be shown

analogously by using a data value of perfectly poor IQ (e.g. the value 'NULL' stored for an unknown customer address which, however, does exist in the real world).

Re (ii): The metric values are bounded from below and from above (cf. re (i)). Hence, there must exist a supremum M (lowest upper bound). If the metric values do not attain a maximum, it follows that $IQ(\omega, \omega_m) < M$ for a data value ω which is of perfectly good IQ. As M is by definition the lowest upper bound, it follows that there exists another data value ω'' with $IQ(\omega, \omega_m) < IQ(\omega'', \omega_m) < M$ (otherwise, $IQ(\omega, \omega_m)$ would be an upper bound and the maximum of the values of the metric). This is, however, again a contradiction to the fact that ω is – by definition – of perfectly good IQ. Hence, the metric values must attain a maximum. The fact that they have to attain a minimum as well can be shown analogously by using a data value of perfectly poor IQ.

To sum up, it follows that a metric has to be bounded from below and from above and must attain both a minimum and a maximum (cf. I-III in Figure 1). ■

The proof shows that (R1) is necessary. However, some existing metrics (cf. e.g. Hipp et al. 2001; Hinrichs 2002; Hipp et al. 2007; Alpar and Winkelsträter 2014) do not attain a minimum/maximum and may thus lead to wrong decisions when evaluating decision alternatives (cf. III-VI in Figure 1). In these cases it is, for example, not possible to decide whether the assessed IQ level can or should be increased to allow for better decision making (cf. VI-IX in Figure 1). This may result in unnecessary improvement measures for data values of already perfectly good IQ (since the metrics do not indicate that the maximum IQ has already been reached). In addition, with respect to an economically oriented management of IQ, if (R1) is not satisfied, neither the comparability nor the validation (e.g. against a benchmark) of the metric values in repeated assessments over time are guaranteed.

The requirements in Group 2 focus on the scale of measurement of the metric values. The requirements in this group have not been formally proven, and some of them do not specify a precise scale (e.g. *definition of scale* is not defined, but only illustrated by a very wide range of examples). To address this gap, we state and justify the following requirement:

Requirement 2 (R2): Interval-scaled metric values

Based on the classification of scales of measurement (cf. Stevens 1946), the values of an IQ metric have to be at least interval-scaled.

Proof: To justify (R2) and evaluate its necessity, we provide a proof by contradiction. Assume that a metric does not provide interval-scaled values (cf. I-III in Figure 1), but that its values are still unambiguous and meaningful and thus support decision-making under

uncertainty and an economically oriented management of IQ in a well-founded way (cf. Section 3). Let us further consider the decision matrix in Table 1 where the payoff vectors $P_1 = (p_{11}, p_{12})$ and $P_2 = (p_{21}, p_{22})$ for the alternatives a_1 and a_2 are estimated based on the metric values IQ_1 and IQ_2, respectively. By assumption, these metric values are not interval-scaled. Hence, they are at most ordinal-scaled. Let the expected payoffs for choosing alternative a_1 and alternative a_2 be the same (i.e. $E(a_1, IQ_1, P_1, S) = E(a_2, IQ_2, P_2, S)$) while $p_{11} > p_{21}$, $p_{12} = p_{22}$, and $IQ_1 < IQ_2$ holds. Hence, the decision maker faces a situation in which in state s_1 choosing alternative a_1 goes along with a higher payoff than choosing a_2 ($p_{11} > p_{21}$), but due to the lower metric value IQ_1, as compared to IQ_2, the expected payoff for both alternatives is the same (cf. III-VI in Figure 1). In this situation, the decision maker is indifferent between the two alternatives[3]. Thus, a reduction of the payoff is accepted if its estimation is based on data of higher IQ. This means that an increase in the payoff of $(p_{11} - p_{21})$ has the same value to the decision maker as an increase in the metric value of $(IQ_2 - IQ_1)$. As the payoffs are interval-scaled, the differences between the payoffs are meaningful. However, by assumption, the difference $(IQ_2 - IQ_1)$ is calculated from metric values which are at most ordinal-scaled and therefore not meaningful. This interpretability of the differences between the metric values compared to the respective differences in the payoffs is therefore a contradiction to the assumption. In particular, it is contradictory that a decision maker is indifferent between the meaningful difference $(p_{11} - p_{21})$ and the not meaningful difference $(IQ_2 - IQ_1)$. Therefore, the assumption that at most ordinal-scaled IQ metric values are sufficient to support both decision-making under uncertainty and an economically oriented management of IQ in a well-founded way has to be wrong. Instead, the metric values have to be interval-scaled to guarantee a meaningful difference $(IQ_2 - IQ_1)$. ■

(R2) has a significant practical impact. Indeed, many existing IQ metrics (cf. e.g. Ballou et al. 1998; Hinrichs 2002) which do not ensure interval-scaled values, may lead to wrong decisions when evaluating different decision alternatives (cf. III-VI in Figure 1). Moreover, interval-scaled metric values are necessary for evaluating, interpreting, and comparing the effects of IQ improvement measures to enable an economically oriented management of IQ. In particular, it is not sufficient to state which measure results in the greatest relative

[3] If such a situation does not exist, the decision is trivial: Assume that $E(a_1, IQ_1, P_1, S) > E(a_2, IQ_2, P_2, S)$ holds for $p_{11} > p_{21}$, $p_{12} = p_{22}$, and all possible values for IQ_1 and IQ_2. Then, the decision maker will always choose a_1 regardless of the metric values. In this case IQ, does not matter, which means that assessing IQ is not necessary at all. The same argumentation applies analogously for $E(a_1, IQ_1, P_1, S) < E(a_2, IQ_2, P_2, S)$.

improvement of the IQ level (in case of an ordinal scale). In fact, to ensure the selection of the efficient IQ improvement measure(s), it is necessary to exactly determine its/their benefits (i.e. the additional expected payoff) resulting from the increase in the IQ level and compare them to its/their costs (cf. VI-IX in Figure 1).

The requirements in Group 3 state that the metric values must have a given interpretation. However, existing requirements (e.g. *comprehensibility*, *comparability*, *interpretability*, *definition of scale*, *interpretation consistency*, and *clarity of definition*) were neither formally proven nor do they specify what exactly is meant by interpretation, making the verification of IQ metrics with regard to this requirement very difficult. In the following, we show that we do not need to define a separate requirement for Group 3, because a clear interpretation is already ensured by the combination of (R1) and (R2). Indeed, a metric which meets both (R1) and (R2) is *interpretable* in terms of the measurement unit *one* (Bureau International des Poids et Mesures 2006). To show this, let m be the minimum (representing perfectly poor IQ) and M be the maximum (representing perfectly good IQ) of the metric values (cf. (R1)). Since equal differences result in equidistant numbers on an interval scale (cf. (R2)), each value of the metric IQ can be interpreted as the $\frac{(IQ-m)}{(M-m)}$ fraction of the maximum difference $(M-m)$. Thus, an IQ metric that meets both (R1) and (R2) is inherently interpretable in terms of the measurement unit *one* (i.e. as percentage).

A simple interpretation of the metric values is helpful to understand the actual meaning of the IQ level and is thus important in practical applications, such as the communication to business users. This is the case if the metric values are ratio-scaled. Ratio-scaled metric values support statements like ‘a metric value of 0.6 is twice as high as a metric value of 0.3’. Ratio-scale can be achieved by a simple transformation of each interval-scaled IQ metric by transforming the minimum m of the metric values to 0 so that each value of the metric can be interpreted as a fraction with respect to the maximum IQ value.

Group 4 contains requirements stating that it must be possible to adjust an IQ metric to adequately reflect the particular context of application which, however addresses only one relevant aspect. There are general quality criteria (i.e. objectivity, reliability, and validity) that must be satisfied by IQ metrics but have not been considered in the literature so far. In addition, not only the metric values, but also the configuration parameters of an IQ metric should satisfy these quality criteria to avoid inadequate results (cf. II-III in Figure 1). To address these drawbacks, we propose and justify the following requirement:

Requirement 3 (R3): Quality of the configuration parameters and the determination of the metric values

It must be possible to determine the configuration parameters of an IQ metric according to the quality criteria objectivity, reliability, and validity (cf. Allen and Yen 2002; Cozby and Bates 2012; Zikmund et al. 2012). The same holds for the determination of the metric values.

There exists a large body of literature dealing with the quality criteria *objectivity*, *reliability,* and *validity* of measurements in general (cf. e.g. Allen and Yen 2002; Cozby and Bates 2012; Zikmund et al. 2012). In the following, we first briefly discuss these criteria for the context of IQ metrics. Afterwards, we prove their necessity based on our decision-oriented framework.

Objectivity of both the configuration parameters and the IQ metric values denotes the degree to which the respective parameters and values as well as the procedures for determining them (e.g. SQL queries) are independent of external influences (e.g. interviewers). This criterion is especially important for IQ metrics requiring expert estimations to determine the configuration parameters or the metric values (cf. e.g. Ballou et al. 1998; Hinrichs 2002; Cai and Ziad 2003; Even and Shankaranarayanan 2007; Hüner et al. 2011; Heinrich and Hristova 2014). Here, *objectivity* is violated, if the estimations are provided by too few experts or if external influences such as the particular behavior of the interviewers are not minimized. In general, *objectivity* becomes an issue, if the metric lacks a precise specification of (sound) procedures for the determination of the respective parameters and values. To avoid highly subjective results and ensure *objectivity*, the IQ metric and its configuration parameters have to be unambiguously (e.g. formally) defined and be determined with objective procedures (e.g. statistical methods, cf. e.g. Heinrich et al. 2012).

Reliability of measurement refers to the accuracy with which a parameter is determined. Reliability means the replicability of the results of the methods used for the determination of the configuration parameters or the metric values. In particular, methods will not be reliable, amongst others, if expert estimations are involved which change over time or among different groups of experts. In this case, *reliability* can be analyzed based on the correlation of the results obtained from the different measurements. Thus, IQ metrics which rely on expert estimations (cf. e.g. Ballou et al. 1998; Even and Shankaranarayanan 2007; Hüner et al. 2011; Heinrich and Hristova 2014) have to define a reliable procedure to determine the configuration parameters and the metric values. Generally, to ensure *reliability* of the configuration parameters and the metric values, correct database queries or statistical methods

may be used. In this case, the result of the respective procedure remains the same, if the procedure is applied to the same data at different points in time.

The *validity* of a method for determining the configuration parameters or the metric values refers to the degree of accuracy with which a proposed method actually measures what it should measure. Typically, the *validity* of the determination of a configuration parameter or a metric value is violated, if the determination contradicts the aim. There are several examples which illustrate the practical relevance of validity in the context of IQ metrics. The metric for timeliness by Batini and Scannapieco (2006, p. 29), for example, involves the configuration parameter $Currency$ which is intended to represent "how promptly data are updated". Its mathematical specification $Currency = Age + (DeliveryTime - InputTime)$, however, seems to contradict this aim. Similarly, Hüner et al. (2011, p. 150) state that a metric value of zero indicates that "each data object validated contains at least one critical defect". However, the mathematical definition of the metric shows that for a value of zero, each data object must actually contain *all* critical defects. *Validity* can be achieved by consistent definitions, database queries, or statistical estimations constructed to determine the corresponding parameter or value according to its definition. Additionally, restricting the application domain of a metric (cf. e.g. Ballou et al. 1998; Heinrich et al. 2007) also contributes to ensuring *validity*.

Proof: To prove the necessity of (R3) based on the decision-oriented framework, let an IQ metric violate *objectivity, reliability* and/or *validity* and let its values be used to support decision making under uncertainty (cf. Figure 1). For example, *objectivity* and/or *reliability* may be violated due to different expert estimations of the configuration parameters of the metric; *validity* may be violated due to an inaccurate definition of the metric or its configuration parameters (cf. above). Consider a decision situation as illustrated by the decision matrix in Table 1. In case *objectivity* and/or *reliability* are violated, two applications of the IQ metric result in two different IQ levels, IQ_1 and IQ_2 with $IQ_1 \neq IQ_2$ (e.g. depending on different expert estimations). In case *validity* is violated, the IQ level IQ_1 estimated by the metric does not accurately represent the actual IQ level IQ_2 in the real world. In both cases, considering each of IQ_1 and IQ_2 separately in decision making may result in choosing a different alternative, which means, $a_1 = argmax_{a_i} E(a_i, IQ_1, P_i, S)$ and $a_2 = argmax_{a_i} E(a_i, IQ_2, P_i, S)$ with $a_1 \neq a_2$ (cf. III-VI in Figure 1). If *objectivity* and/or *reliability* are violated, it is not clear to the decision maker if IQ_1 or IQ_2 or none of them correctly reflects the actual IQ level and thus if a_1 or a_2 or none of them is the accurate

decision. Similarly, if *validity* is violated, then the decision maker will choose a_1 instead of the right choice a_2. Thus, in case *objectivity*, *reliability* and/or *validity* are violated, decision makers will make wrong decisions. ■

The proof shows that IQ metrics which do not fulfill (R3) can lead to wrong decisions when evaluating different alternatives (cf. III-VI in Figure 1). In addition, such metrics result in serious problems when evaluating IQ improvement measures (cf. VII-IX in Figure 1). Indeed, IQ metrics not fulfilling (R3) provide inaccurate metric values (cf. above) when evaluating the IQ level before and after conducting an IQ improvement measure. Hence, it is not possible to determine the increase in the IQ level in a well-founded way. For example, it may occur that IQ improvement measures evaluated as efficient before their application do not even result in an increase of the IQ level afterwards. Thus, to enable a sound evaluation of IQ measures in an economically oriented management of IQ, it is necessary to ensure (R3).

Group 5 comprises requirements addressing the cost-benefit perspective when applying IQ metrics. Existing requirements in this group are not justified based on a theoretical framework. Moreover, for some of them, their definition, specification, and interpretation remain unclear (e.g. *business relevance* and how to determine the threshold for *acceptability*), making them difficult to verify. We address these issues by proposing and justifying the following requirement (for a similar guideline in the context of business process modelling cf. Becker et al. 2000):

Requirement 4 (R4): Economic efficiency of the metric

The configuration and application of an IQ metric have to be efficient from an economic perspective. In particular, the additional expected payoffs from the application of a metric have to outweigh the costs for determining both the configuration parameters and the metric values.

Proof: To justify (R4), consider a decision situation as shown in the decision matrix in Table 1. Let alternative a_1 be chosen by a decision maker who does not apply the metric and thus does not consider IQ in decision making (i.e. $a_1 = argmax_{a_i} E(a_i, P_i, S)$), and let alternative a_2 be chosen if IQ is considered (i.e. $a_2 = argmax_{a_i} E(a_i, IQ, P_i, S)$). It is thus possible to analyze the difference between the expected payoffs of both choices (cf. III-VI in Figure 1). Here, the application of the IQ metric is economically efficient and therefore justifiable with respect to the decision-oriented framework (cf. Figure 1), if and only if the difference between the expected payoffs (i.e. $max_{a_i} E(a_i, IQ, P_i, S) - max_{a_i} E(a_i, P_i, S) =$

$E(a_2, IQ, P_2, S) - E(a_1, P_1, S)$) outweighs the costs for applying the IQ metric. Otherwise, the IQ metric contradicts an economically oriented management of IQ. ■

From a practical perspective, (R4) is particularly relevant. Especially those metrics have to be carefully analyzed which require configuration parameters that are not directly available to the user or only at relatively high costs. For example, the metric for correctness by Hinrichs (2002) is based on the comparison of the stored data value and the corresponding real-world value. However, determining the real-world value as input for an IQ metric is usually very resource-intensive. Actually, simply updating the stored data value with the corresponding real-world value would result in a perfectly good IQ and the calculated value of the metric would no longer be needed (as this metric value should represent perfectly good IQ). If the determination of the configuration parameters is too costly or if the procedure for a repeated determination of the metric values is too resource-intensive as compared to the additional expected payoffs (cf. I-IX in Figure 1), approximations and automated estimations (especially for configuration parameters) may be used to reduce the effort. Overall, it has to be stated that a metric which meets the other requirements, but does not fulfill (R4), may be theoretically meaningful, but of no practical value.

Finally, we consider Group 6, addressing the consistent aggregation of the metric values on different data view levels. Again, the requirements in this group are not based on a sound theoretical framework. In addition, applying the *min or max* and the *weighted average operations* – as proposed by existing works – does not always assure a consistent aggregation. We address these issues by the following requirement:

Requirement 5 (R5): Sound aggregation of the metric values

An IQ metric has to be applicable to single data values as well as to sets of data values (e.g. tuples, relations, and a whole database). The consistent interpretation of the resulting values of the metric has to be assured on all levels.

Proof: To justify the necessity of (R5), consider a situation (cf. Figure 1) in which data applied in decision making are not restricted to the level of single data values, but also cover sets of data values (e.g. tuples, relations, and the whole database). This implies that for sound decision making under uncertainty and an economically oriented management of IQ, it must be possible to determine IQ at several data view levels. Let an IQ metric be defined at a lower data view level l (e.g. relations) and a higher data view level $l + 1$ (e.g. database). In the following, we provide a proof by contradiction that the metric values must have a consistent interpretation on both levels: Assume that an aggregation function f for determining the

metric value at the level $l+1$ based on the metric values at level l does not assure a consistent interpretation of the metric values at l and $l+1$, but that the metric values nevertheless support decision-making under uncertainty and an economically oriented management of IQ in a well-founded way (cf. Section 3). In this case, the aggregation of the metric values at l to the metric value at $l+1$ does not adequately reflect the characteristics of the underlying datasets at l (e.g. size, importance). The unweighted arithmetic mean, which is for example used by Hinrichs (2002) to determine the metric value on the database level based on its values on the level of relations, may serve as an example for such an aggregation function. Consider a disjoint decomposition of a dataset D_{l+1} at $l+1$ into the subsets R_l^h $(h=1,\dots,H)$ at l (e.g. a database D_{l+1} which is decomposed into non-overlapping relations R_l^h): $D_{l+1}=R_l^1\cup R_l^2\cup\dots\cup R_l^H$ and $R_l^i\cap R_l^j=\emptyset\ \forall i\neq j$. The metric values for the subsets R_l^h are denoted by $IQ(R_l^h)$. Then, the metric value for D_{l+1} is determined by means of the aggregation function f: $IQ(D_{l+1})=f(IQ(R_l^1)),\dots,IQ(R_l^H))$. Let us now consider that without loss of generality the subset R_l^1 is further divided into two disjoint subsets $R_l^{1\prime}$ and $R_l^{1\prime\prime}$ at l (i.e. $R_l^1=R_l^{1\prime}\cup R_l^{1\prime\prime}$, $R_l^{1\prime}\cap R_l^{1\prime\prime}=\emptyset$). Based on the decomposition $R_l^{1\prime}$, $R_l^{1\prime\prime}$, R_l^2, …, R_l^H at l, the metric value for D_{l+1} is determined as $IQ'(D_{l+1})=f(IQ(R_l^{1\prime}),IQ(R_l^{1\prime\prime}),IQ(R_l^2),\dots,IQ(R_l^H))$. However, as the aggregation function f does – by assumption – not assure a consistent interpretation of the metric values at l and $l+1$, it follows that $IQ'(D_{l+1})\neq IQ(D_{l+1})$ (e.g. in case of an unweighted arithmetic mean, the same subsets R_l^2, …, R_l^H of D_{l+1} are weighted with $1/H$ or $1/(H+1)$, respectively depending on the particular decomposition used). Indeed, the resulting metric value for D_{l+1} depends on the decomposition of the dataset and can hence be manipulated accordingly (i.e. there are two or more possible metric values for the same dataset). Thus, we face the same situation as in the proof of (R3) where it is also not known which metric value actually represents the real IQ of the dataset D_{l+1} at level $l+1$. It analogously follows that this fact results in wrong decisions (cf. III-VI in Figure 1). Hence, the assumption that the metric values must not have a consistent interpretation on the different data view levels to be able to support decision making under uncertainty and an economically oriented management of IQ in a well-founded way (cf. Section 3) has to be wrong. ■

This requirement is of particular practical relevance as decision situations often rely on the IQ of (large) sets of data values. For example, the IQ level of a whole customer database may be considered for the decision of whether or not to conduct a marketing campaign. However,

many IQ metrics in the literature do not provide (consistent) aggregation rules for different data view levels (cf. e.g. Alpar and Winkelsträter 2014; Hipp et al. 2001; 2007; Li et al. 2012). In addition, a consistent interpretation of the metric values on all aggregation levels is important to support an economically oriented management of IQ. If this is not the case, (repeated) measurements of IQ will result in inconsistent and/or wrong results, and it will not be possible to decide whether improvement measures should be applied from a cost-benefit perspective (cf. VI-IX in Figure 1).

5 Application of the requirements

We demonstrate the applicability and efficacy of our requirements by evaluating the metrics by Ballou et al. (1998) and Blake and Mangiameli (2011), published in the high-quality journals Management Science and ACM Journal of Data and Information Quality. To make the evaluation of the metrics more transparent and comprehensible, we refer to the following context of application (cf. Even et al. 2010; Heinrich and Klier 2011): Based on the stored data of existing customers (e.g. corporate customers), a company has to decide which customers to contact with a new product offer in a CRM mailing campaign. The two decision alternatives for the company with respect to each customer in the database are a_1: to select a customer for the campaign or a_2: not to do so. The possible states of nature (occurring depending on a certain probability of acceptance) are s_1: the customer accepts or s_2: the customer rejects the offer. The benefits of applying an IQ metric in this context are generally non-negligible. Indeed, considering the quality of the customer data (as discussed by Even et al. 2010 and Heinrich and Klier 2011) will lead to better decisions (e.g. if an offer is sent to an outdated or incomplete address, this will only cause mailing costs).

5.1 Metric for timeliness by Ballou et al. (1998)

The IQ metric for timeliness proposed by Ballou et al. (1998) is defined as follows:

$$Timeliness := \max\left[1 - \frac{age\ of\ the\ data\ value}{shelf\ life}, 0\right]^s \qquad (1)$$

The parameter *age of the data value* represents the time difference between the occurrence of the real-world event (i.e. when the data value was created in the real-world) and the assessment of timeliness of the data value. The parameter *shelf life* is defined as the maximum length of time the values of the considered attribute remain up-to-date. Thus, a higher value of the parameter *shelf life*, ceteris paribus, implies a higher value of the metric for timeliness,

and vice versa. The exponent $s > 0$, which has to be determined based on expert estimations, influences the sensitivity of the metric to the ratio $\frac{age\ of\ the\ data\ value}{shelf\ life}$. In the following, we present the evaluation of the metric based on the requirements.

R1: Normalized metric values (Fulfilled)

For all values of the parameter $s > 0$, the metric values are within the bounded interval [0; 1]. The minimum of zero (which represents perfectly poor IQ) is attained if the parameter *age of the data value* is greater than or equal to the parameter *shelf life*. The maximum of one (which represents perfectly good IQ) is attained if the parameter *age of the data value* equals zero (e.g. a stored customer address is certainly up-to-date). It follows that (R1) is fulfilled.

R2: Interval-scaled metric values (Not fulfilled)

For $s = 1$ the metric values can be interpreted as the percentage of the data value's remaining shelf life (e.g. a stored customer address is up-to-date with 50%). As a consequence, for $s = 1$ we observe a ratio scale which implies that the values are interval-scaled as well. Apart from this particular case (i.e. for $s \neq 1$), however, the metric values are not interval-scaled. This is due to the fact that for any two interval scales it is always possible to transform one of them to the other by applying a positive linear transformation of the form $x \mapsto ax + b$ (with $a > 0$) (Allen and Yen 2002). Obviously, such a transformation does not exist for $s \neq 1$, as the mapping $x \mapsto x^s$ is not linear for $s \neq 1$. That is why the metric values are generally not interval-scaled and (R2) is not fulfilled.

R3: Quality of the configuration parameters and the determination of the metric values (Not fulfilled)

In the context of customer data, the values of the attribute "address" do not have a known and fixed maximum shelf life. Indeed, company addresses are not characterized by a maximum length of time during which they remain up-to-date (e.g. some companies have been located at the same address for hundreds of years). In this case, it is not possible to determine a fixed value for the configuration parameter *shelf life* of the metric. That is why (R3) is not fulfilled.

R4: Economic efficiency of the metric (Not fulfilled)

Ballou et al. (1998) define the parameter *age of the data value* based on the point of time when the data value was created in the real world. Therefore, to determine the parameter *age of the data value* for a customer's address, it has to be known when the customer moved to this address. This point of time, however, is usually neither stored nor easily accessible for companies (e.g. due to privacy protection laws), making the costs of configuration parameter determination very high. Indeed, for the above context it would not be efficient to determine

the configuration parameter *age of the data value* to be able to calculate the metric values for the company's customers. Actually, it would even be easier and less resource-intensive – independent of the benefits of the campaign – to directly evaluate whether the address values are still up-to-date (e.g. by contacting the customers). As a consequence, (R4) is not fulfilled.

R5: Sound aggregation of the metric values (Fulfilled)

The authors propose to use the weighted arithmetic mean to aggregate the metric values from single data values to a set of data values. (R5) is fulfilled, as this aggregation rule ensures a consistent interpretation of the metric values on all levels.

Overall, while the metric for timeliness proposed by Ballou et al. (1998) fulfills (R1) and (R5), it does not fulfill (R2), (R3), and (R4).

5.2 Metric for completeness by Blake and Mangiameli (2011)

The metric for completeness by Blake and Mangiameli (2011) is defined as follows. On the level of data values, a data value is incomplete (i.e. the metric value is zero) if and only if it is 'NULL', otherwise it is complete (i.e. the metric value is one). A tuple in a relation is defined as complete if and only if all data values are complete (i.e. none of its data values is 'NULL'). For a relation R, let T_R be the number of tuples in R which have at least one 'NULL'-value and let N_R be the total number of tuples in R. Then, the completeness of R is defined as

$$Completeness := 1 - \frac{T_R}{N_R} = \frac{N_R - T_R}{N_R} \quad (2)$$

The evaluation of the metric with respect to the requirements is presented below:

R1: Normalized metric values (Fulfilled)

The values of the metric are within the bounded interval [0; 1]. This holds for all aggregation levels. The minimum of zero (which represents perfectly poor IQ) on the level of a data value, a tuple, and a relation is attained, if a data value equals 'NULL' (e.g. the street of a single customer address is not stored), if a tuple contains at least one data value which equals 'NULL', and if each tuple of a relation contains at least one data value which equals 'NULL', respectively. The maximum of one (which represents perfectly good IQ) on the level of a data value, a tuple, and a relation is attained if a data value does not equal 'NULL', if a tuple does not contain any data value which equals 'NULL', and if a relation does not contain any tuple with data values which equals 'NULL', respectively. It directly follows that (R1) is fulfilled.

R2: Interval-scaled metric values (Fulfilled)

On the levels of data values and tuples, the values of the metric are interval-scaled (i.e. the difference between the only two possible metric values zero and one is meaningful). On the level of relations, the values of the metric are defined as the percentage of tuples which do not contain any data values which equal 'NULL' (e.g. 50% of all tuples storing customer data are complete). That implies a ratio scale, and thus the values are also interval-scaled. Therefore (R2) is fulfilled.

R3: Quality of the configuration parameters and the determination of the metric values (Fulfilled)

All configuration parameters of the metric (i.e. whether a data value equals 'NULL'; whether a tuple contains a data value, which equals 'NULL'; and the number of tuples in a relation and how many of them contain at least one data value, which equals 'NULL') can be determined by means of simple database queries. Hence, the quality criteria objectivity, reliability, and validity are fulfilled. The values of the metric can be determined by means of mathematical formulae in an objective and reliable way. As the metric quantifies the IQ dimension completeness at different levels according to the corresponding definition, the determination of the metric values is valid. To sum up, (R3) is fulfilled.

R4: Economic efficiency of the metric (Fulfilled)

The parameters of the metric can be determined by means of database queries and the values of the metric can be determined by means of mathematical formulae, both of them in an automated and effective way and at negligible costs. In case the benefits from applying the metric are non-negligible (cf. the given context of application), the application of the metric is efficient and thus fulfills (R4).

R5: Sound aggregation of the metric values (Fulfilled)

The metric is applicable to single data values as well as to sets of data values (tuples and relations). The determination of the values of the metric on the different aggregation levels follows well-defined rules allowing for a consistent interpretation. Therefore, (R5) is fulfilled.

Overall, the metric by Blake and Mangiameli (2011) satisfies all requirements (R1) to (R5). To sum up, the evaluation of the two IQ metrics shows that our requirements are neither trivial nor impossible to fulfill.

6 Conclusion, Limitations and Future Research

In this paper, we propose a set of five requirements for IQ metrics to support both decision making under uncertainty and an economically oriented management of IQ. Our requirements contribute to the existing literature in two ways. First, as opposed to existing approaches, which are fragmented and leave room for interpretation, we present a set of clearly defined requirements, thus making it possible to easily and transparently verify them. This is very important for practical applications. Second, in contrast to existing works, we justify our requirements based on a sound theoretical foundation. If this foundation is missing, it is neither possible to prove the necessity of the requirements nor is it clear what happens if a requirement is not met. As a result, our requirements are essential for the evaluation of existing metrics as well as for the design of new metrics (e.g. in the context of Design Science Research). Based on our requirements, inadequate metrics, which may lead to wrong decisions and economic losses, can be identified and improved. The applicability and efficacy of the proposed requirements are demonstrated by means of two well-known IQ metrics. The application to the metric for completeness by Blake and Mangiameli (2011) has revealed the existence of metrics which satisfy all requirements. The application to the metric for timeliness by Ballou et al. (1998), however, shows that the requirements are not trivial to fulfill. Both of these results are crucial from a methodical and practical point of view.

The proposed requirements constitute a first but essential step to support both decision making under uncertainty and an economically oriented management of IQ. Nevertheless, they also have some limitations. To begin with, they are designed for IQ metrics concerning data views and therefore do not directly consider IQ metrics addressing the quality of data schemes, for example. However, in future research, the ideas underlying the derivation of the requirements can be transferred analogously to other types of IQ metrics. Moreover, as already discussed for many other sets of requirements (e.g. in the context of software engineering), it is impossible to prove the completeness and sufficiency of a set of requirements. Indeed, extending a set of requirements is an iterative process, which should consider both theoretical and practical aspects. Thus, future research should extend the proposed set of requirements in a well-founded manner.

References

Adam D (1996) Planung und Entscheidung: Modelle-Ziele-Methoden. Mit Fallstudien und Lösungen. Springer Gabler, Wiesbaden

Allen MJ, Yen WM (2002) Introduction to measurement theory. Waveland Press, Long Grove Ill

Alpar P, Winkelsträter S (2014) Assessment of data quality in accounting data with association rules. Expert Syst Appl 41(5):2259–2268

Ballou D, Wang R, Pazer H, Tayi GK (1998) Modeling information manufacturing systems to determine information product quality. Manag Sci 44(4):462–484

Batini C, Scannapieco M (2006) Data quality: concepts, methodologies and techniques. Springer, New York

Becker J, Rosemann M, Uthmann C (2000) Guidelines of business process modeling. In: Business Process Management. Springer, London, pp 30-49

Blake R, Mangiameli P (2011) The effects and interactions of data quality and problem complexity on classification. J Dat Infor Qual 2(2):No. 8

Buhl HU, Röglinger M, Moser F, Heidemann J (2013) Big data. A fashionable topic with(out) sustainable relevance for research and practice? Bus Inform Syst Eng 5(2):65–69

Bureau International des Poids et Mesures (2006) The international system of units (SI). National Institute of Standards and Technology, Paris

Cai Y, Ziad M (2003) Evaluating completeness of an information product. In: AMCIS (2003), pp 2273–2281

Campanella J (1999) Principles of Quality Costs: Principles, Implementation and Use. ASQ Quality Press, Milwaukee

Cappiello C, Comuzzi M (2009) A utility-based model to define the optimal data quality level in IT service offerings. In: ECIS (2009), Paper 76

Cozby P, Bates S (2012) Methods in behavioral research, 11th edn. McGraw-Hill Higher Education, New York

Economist Intelligence Unit (2011) Big data: harnessing a game-changing asset. The Economist. http://www.sas.com/resources/asset/SAS_BigData_final.pdf. Accessed 04 Jun 2014

Eppler MJ (2003) Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. Springer, Berlin

Even A, Shankaranarayanan G (2007) Utility-driven assessment of data quality. Database Adv Inform Syst 38(2):75–93

Even A, Shankaranarayanan G, Berger PD (2010) Evaluating a model for cost-effective data quality management in a real-world CRM setting. Decis Support Syst 50(1):152–163

Experian QAS (2013) The Data Advantage: How accuracy creates opportunity. http://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf. Accessed 04 Jun 2014

Feigenbaum AV (2004) Total quality control. McGraw-Hill Professional New York

Fettke P (2006) State-of-the-Art des State-of-the-Art. Bus Inform Syst Eng 48(4):257-266

Fisher CW, Chengalur-Smith I, Ballou DP (2003) The impact of experience and time on the use of data quality information in decision making. Inform Syst Res 14(2):170–188

Fisher CW, Lauria EJM, Matheus CC (2009) An accuracy metric: Percentages, randomness, and probabilities. J Dat Infor Qual 1(3):No. 16

Forbes Insights (2010) Managing information in the enterprise: perspectives for business leaders. Survey by Forbes Insights, New York

Heinrich B, Hristova D (2014) A Fuzzy Metric for Currency in the Context of Big Data. In: ECIS (2014)

Heinrich B, Kaiser M, Klier M (2007) How to measure data quality? A metric-based approach. In: ICIS (2007), Paper 108

Heinrich B, Klier M, Kaiser M (2009) A procedure to develop metrics for currency and its application in CRM. J Dat Infor Qual 1(1):No. 1

Heinrich B, Klier M (2011) Assessing data currency-a probabilistic approach. J Inform Sci 37(1):86–100

Heinrich B, Klier M, Görz Q (2012) Ein metrikbasierter Ansatz zur Messung der Aktualität von Daten in Informationssystemen. Z Betriebswirtsch 82(11):1193–1228

Hinrichs H (2002) Datenqualitätsmanagement in Data-Warehouse-Systemen. Dissertation. Universität Oldenburg

Hipp J, Güntzer U, Grimmer U (2001) Data Quality Mining-Making a Virute of Necessity. In: 6TH ACM SIGMOD DMKD (2001), pp 52–57

Hipp J, Müller M, Hohendorff J, Naumann F (2007) Rule-Based Measurement Of Data Quality In Nominal Data. In: ICIQ (2007), pp 364–378

Hüner KM (2011) Führungssysteme und ausgewählte Maßnahmen zur Steuerung von Konzerndatenqualität. Dissertation. Universität St. Gallen

Hüner KM, Schierning A, Otto B, Österle H (2011) Product data quality in supply chains: the case of Beiersdorf. Electron Mark 21(2):141–154

IBM Global Business Services (2012) Analytics: Big Data in der Praxis. IBM Global Business Services, Armonk

Jiang Z, Sarkar S, De P, Dey D (2007) A framework for reconciling attribute values from multiple data sources. Manag Sci 53(12):1946–1963

Jones BD (1999) Bounded rationality. Annu Rev Polit Sci 2(1):297–321

Laux H (2007) Entscheidungstheorie, 7th edn. Springer Gabler, Wiesbaden

Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inform Manag 40(2):133–146

Levy Y, Ellis TJ (2006) A systems approach to conduct an effective literature review in support of information systems research. Inf Sci 9(1):181–212

Li F, Nastic S, Dustdar S (2012) Data quality Observation in Pervasive Environments. In: CSE (2012), pp 602–609

Loshin D (2010) The practitioner's guide to data quality improvement. Morgan Kaufmann

Lukoianova T, Rubin VL (2014) Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Adv Class Res Online 24(1):4–15

Meredith JR, Raturi A, Amoako-Gyampah K, Kaplan B (1989) Alternative research paradigms in operations. J Oper Manag 8(4):297–326

Mosley M, Brackett M, Earley S (eds) (2009) The DAMA guide to the data management body of knowledge enterprise server version. Technics Publications, LLC, Westfield

Nitzsch R von (2006) Entscheidungslehre. Verlag Mainz, Mainz

Orr K (1998) Data quality and systems theory. Comm ACM 41(2):66–71

Parssian A, Sarkar S, Jacob VS (2004) Assessing data quality for information products: impact of selection, projection, and Cartesian product. Manag Sci 50(7):967–982

Peterson M (2009) An introduction to decision theory. Cambridge University Press, Cambridge

Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Comm ACM 45(4):211–218

Redman TC (1996) Data quality for the information age. Artech House, Boston

SAS Institute (2013) 2013 Big data survey research brief. SAS Institute, Cary, N.C.

Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63(2):129–138

Simon HA (1969) The sciences of the artificial. MIT press, Cambridge

Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680

Vom Brocke J, Simons A, Niehaves B, Niehaves B, Reimer K, Plattfaut R, Cleven A (2009) Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: ECIS (2009), Paper 372

Wang RY, Storey VC, Firth CP (1995) A framework for analysis of data quality research. IEEE Trans Knowl Data Eng 7(4):623–640

Wang RY (1998) A product perspective on total data quality management. Comm ACM 41(2):58–65

Webster J, Watson RT (2002) Analyzing the past to prepare for the future: Writing a literature review. Manag Inform Syst Quartely 26(2):13–23

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

Zikmund W, Babin B, Carr J, Griffin M (2012) Business research methods, 8th edn. Cengage Learning, Mason

Appendix A: Notation

Notation	Definition
s_j	State of nature, $j \in \{1, \dots, n\}$
$S = (s_1, s_2, \dots, s_n)$	Vector of all considered states of nature
$w(s_j)$	Probability of occurrence for a state of nature s_j
a_i	Decision alternative, $i \in \{1, \dots, m\}$
$A = (a_1, a_2, \dots, a_m)$	Vector of all considered decision alternatives
p_{ij}	Payoff if alternative a_i is chosen and state of nature s_j occurs
$P_i = (p_{i1}, p_{i2}, \dots, p_{in})$	Vector of the payoffs for alternative a_i and all considered states of nature
$E(a_i, P_i, S)$	Expected payoff without considering IQ for alternative a_i, given a vector S of states of nature and a vector P_i of payoffs for alternative a_i
DQ	IQ metric value of the considered data value/set
$E(a_i, IQ, P_i, S)$,	Expected payoff when considering IQ for alternative a_i, given a vector S of states of nature, a vector P_i of payoffs for alternative a_i, and a value of the IQ metric IQ
$IQ(\omega, \omega_m)$	A metric value for a stored data value ω and a real-world value ω_m
M	Supremum/maximum of the considered metric values
m	Minimum of the considered metric values
l	Data view level with $l \in \{1, \dots, L\}$
D_l	A dataset at data view level l
R_l^h	A subset of the dataset D_l, $h \in \{1, \dots, H\}$
f	Aggregation function
$IQ(R_l^h)$	IQ level of the dataset R_l^h

Table 3: Notation

Appendix B: Requirements for IQ metrics proposed by Hüner (2011)

Requirement	Description of the proposed requirement
Cost/benefit	The costs for the definition and the calculation of the IQ metric values ought to be in a positive ratio (< 1) to the benefits (controlled error potential).
Definition of measurement frequency	The instants of time at which the values of an IQ metric are calculated should be defined.
Definition of measurement point	The measurement point (e.g. data repository, process, department) of an IQ metric should be defined.
Definition of measurement procedure	The instrument (e.g. survey, software) to determine the IQ metric value should be defined.
Definition of scale	A scale (e.g. percentage, school grades, time) should be defined for an IQ metric value.
Limitation of the application data	For an IQ metric, the data to be applied to (e.g. material master, European customers) should be defined.
Escalation process	For an IQ metric appropriate measures should be defined depending on certain threshold values (i.e. metric values to initiate IQ measures).
Validity range	A range should be defined for an IQ metric in which its values are valid.
SMART criteria	An IQ metric should fulfill the SMART criteria (specific, measurable, attainable, relevant and time-bounded).
Disturbance variables	The metadata of an IQ metric should contain information about possible disturbance variables (i.e. it should describe possible events or impacts which may distort the values of the IQ metric).
Responsibility	For an IQ metric clear responsibilities should be defined such as to whom and which values of the IQ metric are reported, who is responsible for the maintenance of the metric (e.g. up-to-date/meaningful definition, implementation of the measurement procedure).
Comparability	An IQ metric should be defined so that its values can be compared to those of other metrics (IQ metrics or process metrics).
Comprehensibility	For an IQ metric metadata should be available, which describes its purpose and the correct interpretation of its values.
Use in SLAs	It should be possible to use IQ metric values in Service Level Agreements.
Visualization	It should be possible to visualize the values of an IQ metric (e.g. time series, diagrams).
Repeatability	It should be possible to determine the values of an IQ metric not only once, but multiple times.
Target value	For an IQ metric a target value should be defined.
Assignment to an IQ dimension	It should be possible to assign an IQ metric to one or more IQ dimensions.
Assignment to a business problem	It should be possible to assign an IQ metric to a specific (company-specific) business problem.
Assignment to a process figure	It should be possible to assign an IQ metric to one or more process figures.
Assignment to the company strategy	It should be possible to assign an IQ metric to one or more strategic goals of the company.

Table 4: Requirements for IQ metrics proposed by Hüner (2011, p. 302)

5.2 Contribution to RQ 4

In this chapter, the following research question was addressed:

RQ 4 Which *requirements* should information quality metrics for data views satisfy to adequately, efficiently and *directly* support decision making?

It was shown that an information quality metric must result in normalized and interval-scaled metric values; the determination of its configuration parameters and of the metric values must be objective, reliable, and valid; its application must be efficient (after considering the costs for the application itself); and the sound aggregation of the metric values must be guaranteed. These five requirements were defined and proven by both considering the existing literature and the framework in Figure 2. Thus, a metric for data views which does not satisfy them and the results of which are *directly* considered in decision making with environmental uncertainty (*Analyze* phase), will lead to wrong decisions and economic losses (*Improve* phase). Moreover, by evaluating these requirements on two well-known metrics from the literature, it was shown that they are easy to apply, but neither trivial, nor impossible to fulfil. This together with their clear definition makes them suitable for practical applications. The presented paper is the last one in this dissertation. In the next chapter, main conclusions are drawn and paths for future research are proposed.

6. Conclusion

In this chapter, the major findings are summarized and the limitations and paths for future research are discussed.

6.1 Major Findings

- *Currency can be adequately measured with both probability theory and fuzzy set theory*

Two metrics for currency were developed: the extended metric in Subsection 2.1 and the fuzzy metric in Subsection 2.3. The first metric is based on probability theory and models the temporal change of real-world information as a stochastic process. It contributes to the literature by not only delivering a statement regarding the currency of stored information, but also by providing an indication about the current real-world information. In Section 2.1, two different metric forms were provided, depending on the information the decision maker possesses. The first one assumes that the history of the real-world information after storage is known (general form), while for the second one (Markov form) this is not necessary. A drawback of the extended metric is that it requires detailed historical data which may not always be given in reality.

The second metric, presented in Section 2.3, addresses this drawback as it is based on expert estimations and fuzzy set theory. It measures currency by a FIS where the input linguistic variables to the system are age and decline rate of the stored attribute value and the output linguistic variable is currency. In case the input parameters to the system are not known, a method was provided to estimate them with the help of an auxiliary FIS. The output value of the system, after aggregation and defuzzification, is a currency level in [0,1]. The metric was evaluated by developing a questionnaire according to the guidelines in the literature and by specifying the corresponding membership functions from the responses. The evaluation showed that additional information plays a role for the quality of expert estimations and that some fuzzy sets may be more difficult to estimate than others.

- *Currency must be directly considered in decision making with environmental uncertainty to avoid wrong decisions*

In Section 2.1, a method was developed for *directly* considering currency in decision making with environmental uncertainty, based on the normative concept of the value of information. The idea is that outdated information leads to quality uncertainty which is comparable to environmental uncertainty. In particular, just as real-world information delivers an indication about the current state of nature, so does stored information deliver an indication about the current real-world information. Based on this idea, the normative concept of the value of information was modified by incorporating the extended metric referring to currency. As a result, the influence of the additional information about the level of currency on the decision was modeled and it was also shown under which conditions the decision will change based on this information. Moreover, it was demonstrated that the level of currency may influence the (normative concept of the) value of information in a way that has not been addressed by existing approaches. This is very important for the *Improve* phase of the information quality

management cycle, where the improvement measures are evaluated from a cost-benefit perspective. Finally, the evaluation showed that quality uncertainty does influence decisions and the value of information in real-world applications and that not taking it into account may result in wrong ex-post decisions and economic losses.

- *Currency must be indirectly considered in decision making to decrease the likelihood for the generation of wrong knowledge and thus wrong decisions*

A two-phase approach for *indirectly* considering quality uncertainty in decision making in the context of decision trees was developed. In particular, failing to take into account the currency of stored instances in the application of the method may lead to wrong classification and thus wrong knowledge. To address this issue, the results from the extended metric referring to currency were incorporated during the classification of new instances in existing decision trees. This approach regards the structure of the tree, resulting in a more efficient, context-specific and less data-intensive estimation. Moreover, based on supplemental data, a method for increasing the precision of the estimations was provided, which is also efficient and context-specific. The evaluation supported the merits of the approach, by demonstrating that it leads to higher success rates than not taking currency into account and that it is very efficient.

- *Accuracy due to subjective estimations can be adequately modeled with fuzzy set theory*

An approach was developed for modeling accuracy due to subjective estimations in the context of resource requirements for Cloud service providers. In particular, due to their subjective nature, customer's resource estimations are rather vague. This results in subjective quality uncertainty, which was modeled with fuzzy numbers. The idea to use fuzzy numbers to model subjective quality uncertainty is not restricted to the context of Cloud computing, but can analogously be applied to other contexts.

- *Accuracy due to subjective estimations must be directly considered in decision making with time uncertainty to decrease the likelihood for wrong decisions*

The *direct* role of accuracy in decision making with time uncertainty was modeled for the problem faced by Cloud service providers (Section 3.1). In particular, Cloud service providers need to decide whether to accept an incoming job request by taking into account resource restrictions, revenue maximization, and quality of service requirements. It was shown how low accuracy, modeled with fuzzy numbers, can be incorporated in three different admission control policies: first-come first-served, dynamic pricing and client classification. The first policy simply accepts requests in their incoming order (if there are enough resources). The second policy aims at revenue maximization, while the third one focuses on the quality of service requirements. Incorporating accuracy in each of these three policies leads to "fuzzification" of the constraints of the decision problem and as a result, appropriate solution approaches are required. The evaluation showed that, similar to Section 2.1, quality uncertainty influences revenue and if not taken into account, may cause wrong decisions and economic losses.

- *Fuzzy duality theory approaches can be applied when analyzing quality uncertainty, but the literature on the topic suffers from some drawbacks*

In many cases subjective quality uncertainty is modeled with fuzzy set theory. Thus, considering this uncertainty in decision making results in fuzzy optimization problems, which may not be easy to solve. One solution approach is fuzzy duality theory. However, the analysis of the literature in the field of fuzzy linear optimization (cf. Section 3.2) showed that this stream of research is rather fragmented, incomplete, and inconsistent. In particular, not all duality theories are covered for a given fuzzy optimization problem and depending on the used order operator the approaches differ substantially with regard to their homogeneity, interpretation, complexity, and completeness. As a result, applying these approaches to real-world problems such as modeling quality uncertainty is difficult and even dangerous. To solve these issues, a number of possible directions for future research were proposed including a unifying fuzzy duality theory.

- *Consistency can be adequately measured based on probability theory*

A metric for consistency based on hypothesis testing was developed. The idea behind the metric is to model the fulfillment of a rule as a binomial random variable with the corresponding probability of fulfillment. In particular, for a given rule and under the assumption of a binomial distribution, the probability of fulfillment in a consistent reference dataset is compared with the relative frequency (empirical probability) with which the same rule is fulfilled in the assessed dataset. Thus, consistency is measured as the two-sided p-value of the hypothesis test with the null hypothesis that the two probabilities coincide. As a result, the measured consistency level has a clear interpretation in terms of probability. The evaluation supports the applicability and the validity of the approach.

- *Adequate and efficient information quality metrics for data views must satisfy a set of five clearly defined requirements*

In Section 5.1, five requirements were developed and proven. These requirements must be satisfied by information quality metrics for data views to *directly* and adequately support decision making with environmental uncertainty. In particular, it is required that the metric values are normalized and interval-scaled; that their configuration parameters and their values are estimated in an objective, reliable, and valid way, that their application is efficient; and that their values are aggregated in a sound way. If any of these requirements is not satisfied, then wrong decisions and/or economic losses will result. It was shown that the requirements are easy to verify, but neither trivial nor impossible to fulfill, which additionally supports their appropriateness.

6.2 Limitations and Future Research

Naturally, this work also leaves some open questions, which should be addressed by future research. To begin with, here quality uncertainty was modeled with both probability and fuzzy set theory. Generally, the latter would be preferred in the absence of historical data and in the presence of subjective estimations. These two methodologies were addressed to a different extend for the different information quality dimensions. In particular, two metrics for currency were presented

covering both probability theory and fuzzy set theory. However, for accuracy only fuzzy set theory and for consistency only probability theory was considered. Thus, the task of future research would be to develop approaches for addressing this gap. In particular, (objective) accuracy can be measured as the probability that a given attribute value stems from the distribution of the real-world attribute value, thus having a clear interpretation. Moreover, (subjective) consistency can be measured by modeling the fulfillment of a rule in consistent reference dataset as a fuzzy set and by comparing its values with the (crisp) relative frequency with which the same rule is fulfilled in the assessed dataset.

In addition, it would certainly be very interesting to compare the two ways for modeling uncertainty for different dimensions, especially with regard to the ease of obtaining the required data, the efficiency of application, and the precision of estimations. Probability theory requires enough historical data, while fuzzy set theory is based on expert estimations and both may be difficult to obtain depending on the application case. Moreover, human beings cannot deliver such precise estimations as historical data can and also expert estimations are characterized by different psychological biases. However, too little historical data for probability-theory based approaches also leads to unreliable and biased estimations. Thus, methods for addressing these issues must additionally be applied in both cases. A possible solution may be to combine the two ways for modeling uncertainty as is done in the field of financial engineering (Huang, 2007).

In future research, also further information quality dimensions must be considered. Examples for such dimensions are completeness, believability, and relevancy. Similar to the distinction between data and information quality, these dimensions can be context-independent or context-specific. For instance completeness can be measured as the percentage of attribute values for a given attribute that are different from NULL or as the percentage of *relevant* attribute values for a given attribute that are different from NULL. In the second case and also when measuring relevancy, it is impossible to model quality uncertainty without incorporating the decision context. However, doing this may require the involvement of experts and also be very resource-consuming. The same holds for believability. There are some works in the literature that deal with this dimension, such as Lukoianova and Rubin (2014), which however only present an initial idea. Future research may use such approaches as a starting point for more sophisticated methodologies.

Developing further the approaches for *directly* considering the metric results in decision making is also an important research direction. In this dissertation it was demonstrated how to *directly* incorporate currency (measured with probability theory) and accuracy (modeled with fuzzy set theory). However, this does not cover all metrics presented in this dissertation (e.g. consistency) and also, as mentioned above, there are many other information quality dimensions that need to be measured and analyzed. One possibility to do this is, based for example on historical data, to analyze, the dependency between the information quality level of the corresponding dimension and the different building elements of decision models. For instance, in the case of consistency, the relationship between the consistency level and the distance between the assessed and the real-world (consistent) payoffs can be analyzed and the results can be applied for structuring decision models with known consistency levels and (inconsistent) payoff values. Another possibility for metrics with clear interpretation is to use expert estimations. Based on their experience, experts

can define the influence of the corresponding information quality level, for instance by using linguistic variables.

A further interesting path for future research is the *indirect* effect of quality uncertainty on decision making. This was examined here with one data mining method, one information quality dimension, and one way of modeling quality uncertainty. Thus, it would be important to apply similar ideas to other data mining methods (e.g. clustering), other information quality dimensions (e.g. reliability), and quality uncertainty modeled with fuzzy set theory. Here the field of uncertain data mining and other related fields may be a useful starting point. As discussed above, these approaches assume that quality uncertainty is given and do not examine its sources. However, modeling this uncertainty in a valid way is an important step. In particular, the modeling methodology must adequately consider the interpretation of the corresponding dimension and also its subjective or objective nature. Thus, it is not possible to use one approach for all information quality dimensions, but rather each dimension should be examined separately depending on its interpretation.

In the context of big data and its Veracity, it would certainly be interesting to apply the presented approaches to unstructured data, which importance is significantly growing, especially with the emergence of social networks. There are generally two ways to do this: 1) by structuring the data with text mining approaches or 2) by developing new approaches trimmed for unstructured data. Both directions seem very promising. Again, in the context of big data and with respect to its Volume, the efficiency of all of the presented approaches can be additionally improved by the use of state-of-the-art techniques such as MapReduce or Hadoop.

Future research should also apply the presented requirements for information quality metrics to the metrics discussed in this dissertation. Generally, the extended metric referring to currency, the fuzzy metric, and the metric for consistency are all normalized and efficient in their application. However, they do not necessary deliver interval-scaled results (e.g. fuzzy metric) or sound aggregation (e.g. extended metric referring to currency). Moreover, it may not always be possible to determine their configuration parameters and the metric values in an objective, reliable, and valid way. As a result, these metrics may be improved to meet the requirements. Another possible direction in this respect would be to introduce a fulfillment index for the requirements. For example, if a metric fulfills only two out of five requirements, then the value of the index would be 40%. This index can then *directly* be incorporated in decision making as a fourth type of uncertainty stemming from the adequacy of information quality metrics.

Additionally, future research should apply the provided requirements to further metrics from the literature in order to investigate whether or not the set of proposed requirements has to be extended. One interesting path in that respect is the development of requirements for *indirectly* regarding the metric results in decision making. In particular, it is not clear, whether the presented requirements are also suitable for this case. For instance, for the consideration of the metric results in data mining, as presented in Section 2.2, no aggregation is necessary, but efficiency has a huge importance.

Finally, in this dissertation relative little attention was paid to the *Improve* phase of the information quality management cycle, although it is just as important as the other three phases. The reason for this is that both the effectiveness and the efficiency of information quality improvement measures depend very much on the particular application case and thus it is difficult to propose a

general approach. Therefore, future research should concentrate on the development of approaches for valuing quality improvement measures, based on the particular decision context. Here again different methodologies can be used, for instance from the field of risk management, where risk mitigation measures are applied with the same aim.

7. References

Alpar, P. and Winkelsträter, S. (2014), "Assessment of data quality in accounting data with association rules", *Expert Systems with Applications*, Vol. 41 No. 5, pp. 2259–2268.

Ballou, D., Wang, R., Pazer, H. and Tayi, G.K. (1998), "Modeling information manufacturing systems to determine information product quality", *Management Science*, Vol. 44 No. 4, pp. 462–484.

Batini, C., Daniele, B., Federico, C. and Simone, G. (2011), "A data quality methodology for heterogeneous data", *International Journal of Database Management Systems*, Vol. 3 No. 1, pp. 60–79.

Batini, C. and Scannapieco, M. (2006), *Data Quality: Concepts, Methodologies and Techniques*, Springer-Verlag Berlin Heidelberg.

Bell, D.E., Raiffa, H. and Tversky, A. (1988), *Decision making: Descriptive, normative, and prescriptive interactions*, Cambridge University Press.

Berger, J.O. (1985), *Statistical decision theory and Bayesian analysis,* 2nd ed., Springer-Verlag.

Berger, J.O. and Berry, D.A. (1988), "Statistical analysis and the illusion of objectivity", *American Scientist*, Vol. 76 No. 2, pp. 159–165.

Billingsley, P. (2012), *Probability and Measure, Anniversary Edition, Wiley Series in Probability and Statistics*, John Wiley & Sons.

Blake, R.H. and Mangiameli, P. (2009), "Evaluating the Semantic and Representational Consistency of Interconnected Structured and Unstructured Data", paper presented at Americas Conference on Information Systems (AMCIS 2009).

Blum, J., Greencorn, D. and Cooperstock, J. (2013), "Smartphone Sensor Reliability for Augmented Reality Applications", in Zheng, K., Li, M. and Jiang, H. (Eds.), *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Springer, pp. 127–138.

Cai, Y. and Ziad, M. (2003), "Evaluating completeness of an information product", paper presented at Americas Conference on Information Systems (AMCIS 2003).

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

Clemen, R.T. and Reilly, T. (2001), *Making hard decisions with DecisionTools*, Brooks/Cole.

Dalkir, K. (2011), *Knowledge Management in Theory and Practice*, MIT Press.

Davenport, T.H. and Prusak, L. (1998), *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, Mass.

DeGroot, M.H. and Schervish, M.J. (2012), *Probability and statistics,* 4th ed., Pearson Education, Inc.

Demski, J.S. (1980), *Information analysis,* 2nd ed., Addison-Wesley Publishing Company.

Edwards, W. (1962), "Dynamic decision theory and probabilistic information processings", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 4 No. 2, pp. 59–74.

Eppler, M.J. (2006), *Managing information quality: Increasing the value of information in knowledge-intensive products and processes,* 2nd ed., Springer Science & Business Media.

Even, A. and Shankaranarayanan, G. (2007), "Utility-driven assessment of data quality", *ACM SIGMIS Database*, Vol. 38 No. 2, pp. 75–93.

Experian QAS (2013), "The Data Advantage: How accuracy creates opportunity", available at: http://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf (accessed 7 March 2015).

Fan, W., Geerts, F. and Wijsen, J. (2011), "Determining the currency of data", paper presented at 13th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996), "Knowledge Discovery and Data Mining: Towards a Unifying Framework", paper presented at Knowledge Discovery and Data Mining 1996.

Fisher, C.W., Lauria, E. J. M. and Matheus, C.C. (2009), "An Accuracy Metric: Percentages, Randomness, and Probabilities", *Journal of Data and Information Quality (JDIQ)*, Vol. 1 No. 3, pp. 1–21.

Forbes Insights (2010), "Managing Information in the Enterprise: Perspectives for Business Leaders", available at: http://images.forbes.com/forbesinsights/StudyPDFs/SAP_InformationManagement_04_2010.pdf (accessed 19 February 2015).

Goodwin, P. and Wright, G. (2014), *Decision analysis for management judgment,* 5th ed., John Wiley & Sons Ltd.

Görz, Q. and Kaiser, M. (2012), "An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy", in Rahman, H., Mesquita, A., Ramos, I. and Pernici, B. (Eds.), *Knowledge and Technologies in Innovative Information Systems*, *Lecture Notes in Business Information Processing*, Vol. 129, Springer Berlin Heidelberg, pp. 169-184.

Heinrich, B., Kaiser, M. and Klier, M. (2007), "How to measure data quality? A metric-based approach", paper presented at 28th International Conference of Information Systems (ICIS).

Heinrich, B. and Klier, M. (2009), "A novel data quality metric for timeliness considering supplemental data", paper presented at 17th European Conference on Information Systems (ECIS).

Heinrich, B. and Klier, M. (2011), "Assessing data currency-a probabilistic approach", *Journal of Information Science*, Vol. 37 No. 1, pp. 86–100.

Heinrich, B. and Klier, M. (2015), "Metric-based data quality assessment — Developing and evaluating a probability-based currency metric", *Decision Support Systems*, Vol. 72, pp. 82–96.

Heinrich, B., Klier, M. and Görz, Q. (2012), "Ein metrikbasierter Ansatz zur Messung der Aktualität von Daten in Informationssystemen", *Zeitschrift für Betriebswirtschaft*, Vol. 82 No. 11, pp. 1193-1228.

Heinrich, B., Klier, M. and Kaiser, M. (2009), "A Procedure to Develop Metrics for Currency and its Application in CRM", *Journal of Data and Information Quality (JDIQ)*, Vol. 1 No. 1, pp. 1–28.

Henschen, D. (2013), "2014 Analytics, BI, and Information Management Survey", available at: http://reports.informationweek.com/abstract/81/11715/Business-Intelligence-and-Information-Management/Research:-2014-Analytics,-BI,-and-Information-Management-Survey.html (accessed 19 February 2015).

Hillier, F.S. and Lieberman, G.J. (2005), *Introduction to Operations Research,* 8th ed., McGraw-Hill Education (Asia).

Hinrichs, H. (2002), *Datenqualitätsmanagement in Data Warehouse-Systemen: (in German)*, Dissertation, Universität Oldenburg.

Hinz, O., Skiera, B., Barrot, C. and Becker, J.U. (2011), "Seeding strategies for viral marketing: An empirical comparison", *Journal of Marketing*, Vol. 75 No. 6, pp. 55–71.

Hipp, J., Güntzer, U. and Grimmer, U. (2001), "Data Quality Mining-Making a Virute of Necessity", paper presented at 6TH ACM SIGMOD DMKD.

Hipp, J., Müller, M., Hohendorff, J. and Naumann, F. (2007), "Rule-Based Measurement Of Data Quality In Nominal Data", paper presented at International Conference on Information Quality 2007.

Hobfeld, T., Hirth, M., Korshunov, P., Hanhart, P., Gardlo, B., Keimel, C. and Timmerer, C. (2014), "Survey of web-based crowdsourcing frameworks for subjective quality assessment", paper presented at 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP).

Huang, K.-T., Lee, Y.W. and Wang, R.Y. (1999), *Quality information and knowledge*, Prentice Hall PTR.

Huang, X. (2007), "Two new models for portfolio selection with stochastic returns taking fuzzy information", *European Journal of Operational Research*, Vol. 180 No. 1, pp. 396–405.

Hüner, K.M., Schierning, A., Otto, B. and Österle, H. (2011), "Product data quality in supply chains: the case of Beiersdorf", *Electronic Markets*, Vol. 21 No. 2, pp. 141-154.

IBM Institute for Business Value (2012), "Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data", available at: http://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf (accessed 17 June 2015).

Kantardzic, M. (2011), *Data Mining:Concepts, Models, Methods, and Algorithms*, Wiley-IEEE Press.

Kumar, V. and Reinartz, W. (2012), *Customer Relationship Management: Concept, Strategy, and Tools*, Springer.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. and Kruschwitz, N. (2013), "Big data, analytics and the path from insights to value", *MIT Sloan Management Review*, Vol. 21.

Lawrence, D.B. (1999), *The Economic Value of Information*, Springer New York.

Lee, Y.W., Pipino, L.L., Funk, J.D. and Wang, R.Y. (2006), *Journey to Data Quality*, The MIT Press.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002), "AIMQ: a methodology for information quality assessment", *Information & management*, Vol. 40 No. 2, pp. 133–146.

Li, F., Nastic, S. and Dustdar, S. (2012), "Data Quality Observation in Pervasive Environments", paper presented at 2012 IEEE 15th International Conference on Computational Science and Engineering (CSE).

Liu, L. and Chi, L.N., "Evolutional Data Quality: a Theory-specific View", in *7th International Conference on Information 2002*.

Liu, S., Duffy, Alex H. B., Whitfield, R.I. and Boyle, I.M. (2010), "Integration of decision support systems to improve decision support performance", *Knowledge and Information Systems*, Vol. 22 No. 3, pp. 261-286.

Long, J. and Seko, C. (2005), "A cyclic-hierarchical method for database data-quality evaluation and improvement", in Wang, Richard Y., Pierce, Elizabeth M., Madnick, Stuart E. (Ed.), *Information Quality*, M.E. Sharpe, Inc.

Löwstedt, J. and Stjernberg, T. (2014), *Producing management knowledge: research as practice*, Routledge.

Luce, R.D. and Raiffa, H. (2012), *Games and decisions: Introduction and critical survey*, Courier Corporation.

Lukoianova, T. and Rubin, V. (2014), "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?", *Advances In Classification Research Online*, Vol. 24 No. 1, pp. 4–15.

Madnick, S.E., Wang, R.Y., Lee, Y.W. and Zhu, H. (2009), "Overview and framework for data and information quality research", *Journal of Data and Information Quality (JDIQ)*, Vol. 1 No. 1, pp. 2:1-2:22.

Marakas, G.M. (2003), *Decision support systems in the 21st century,* 2nd ed., Prentice Hall Upper Saddle River.

Mas-Colell, A., Whinston, M.D. and Green, J.R. (1995), *Microeconomic theory*, Oxford university press.

Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T. and Batini, C. (2002), "Managing Data Quality in Cooperative Information Systems", in Meersman, R. and Tari, Z. (Eds.), *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, *Lecture Notes in Computer Science*, Vol. 2519, Springer Berlin Heidelberg, pp. 486-502.

Meredith, J.R., Raturi, A., Amoako-Gyampah, K. and Kaplan, B. (1989), "Alternative research paradigms in operations", *Journal of Operations Management*, Vol. 8 No. 4, pp. 297–326.

Miller, I. and Miller, M. (2004), *John E. Freund's Mathematical Statistics: With Applications,* 7th ed., Pearson Education.

Minelli, M., Chambers, M. and Dhiraj, A. (2012), *Big data, big analytics: emerging business intelligence and analytic trends for today's businesses*, John Wiley & Sons.

Mladenic, D., Lavrač, N., Bohanec, M. and Moyle, S. (2003), *Data mining and decision support: integration and collaboration*, Springer Science & Business Media.

Nelson, R.R. and Todd, P.A. (2005), "Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing", *Journal of Management Information Systems*, Vol. 21 No. 4, pp. 199–235.

Nissen, M.E. (2002), "An extended model of knowledge-flow dynamics", *Communications of the Association for Information Systems*, Vol. 8 No. 1, pp. 251–266.

Olensky, M. (2014), "Testing an Automated Accuracy Assessment Method on Bibliographic Data", *Journal of Library and Information Studies*, Vol. 12 No. 2, pp. 19–38.

Olson, J.E. (2003), *Data quality: the accuracy dimension*, Morgan Kaufmann.

Orr, K. (1998), "Data quality and systems theory", *Communications of the ACM*, Vol. 41 No. 2, pp. 66–71.

Parnell, G. (2013), *Handbook of Decision Analysis, Wiley Handbooks in Operations Research and Management Science*, John Wiley & Sons.

Peterson, M. (2009), *An introduction to decision theory*, Cambridge University Press.

Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002), "Data quality assessment", *Communications of the ACM*, Vol. 45 No. 4, pp. 211–218.

Price, R., Neiger, D. and Shanks, G. (2008), "Developing a measurement instrument for subjective aspects of information quality", *Communications of the Association for Information Systems*, Vol. 22 No. 1, pp. 49–74.

Redman, T.C. (1996), *Data quality for the information age*, Artech House, Boston.

Sadiq, S., Yeganeh, N.K. and Indulska, M. (Eds.) (2011), *20 years of data quality research: themes, trends and synergies*, Australian Computer Society, Inc.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997), "Data quality in context", *Communications of the ACM*, Vol. 40 No. 5, pp. 103–110.

Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C. (2007), "A framework for information quality assessment", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 12, pp. 1720–1733.

The Information Difference (2013), "The State of Data Quality Revisited", available at: http://www.sap.com/bin/sapcom/en_us/downloadasset.2013-05-may-29-16.the-state-of-data-quality-revisted--research-study-pdf.html (accessed 19 February 2015).

Valle, E.D., Celino, I., Dell'Aglio, D., Kim, K., Huang, Z., Tresp, V., Hauptmann, W., Huang, Y. and Grothmann, R. (2008), "Urban Computing: a challenging problem for Semantic Technologies", paper presented at Workshop on New forms of Reasoning for the Semantic Web: scalable, tolerant and dynamic (NEFORS 2008).

Wang, L.-X. (1997), *A course in fuzzy systems*, Prentice-Hall PTR.

Wang, R.Y. (1998), "A product perspective on total data quality management", *Communications of the ACM*, Vol. 41 No. 2, pp. 58–65.

Wang, R.Y., Pierce, E.M. and Madnick, S.E. (2005), *Information Quality*, M.E. Sharpe, Inc.

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5–33.

Wang, Z., Deng, S. and Ye, Y. (2014), "Close the gaps: A learning-while-doing algorithm for single-product revenue management problems", *Operations Research*, Vol. 62 No. 2, pp. 318–331.

Witchalls, C. (2014), "Gut & gigabytes. Capitalising on the art & science in decision making", available at: http://www.economistinsights.com/business-strategy/analysis/gut-gigabytes (accessed 19 February 2015).

Wu, J., Jiang, C., H., D., Baker, D. and Delfino, R. (2011), "Automated time activity classification based on global positioning system (GPS) tracking data", *Environmental Health*, Vol. 10 No. 1, pp. 1–13.

Xiao, Y., Lu, Louis Y. Y., Liu, J.S. and Zhou, Z. (2014), "Knowledge diffusion path analysis of data quality literature: A main path analysis", *Journal of Informetrics*, Vol. 8 No. 3, pp. 594–605.

Yager, R.R. and Zadeh, L.A. (Eds.) (1992), *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, Kluwer Academic Publishers.

Zadeh, L.A. (1965), "Fuzzy sets", *Information and control*, Vol. 8 No. 3, pp. 338–353.

Zimmermann, H.-J. (2001), *Fuzzy set theory—and its applications,* 4th ed., Kluwer Academic Publishers.

Zyphur, M.J. and Oswald, F.L. (2013), "Bayesian Probability and Statistics in Management Research A New Horizon", *Journal of Management*, Vol. 39 No. 1, pp. 5–13.