KU LEUVEN
FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

A Contextual Data Quality Analysis for Credit Risk Management in Financial Institutions

Proefschrift voorgedragen tot het behalen van de graad van Doctor in de Toegepaste Economische Wetenschappen

door

Helen Tadesse Moges

Nummer 466 December, 2014

Dedicated to my family,
I love you very much!

Committee

Chairperson	Prof.dr. Erik Demeulemeester	K.U.Leuven
Promoter	Prof. dr. Wilfried Lemahieu	K.U.Leuven
Co-Promoter	Prof. dr. Bart Baesens	K.U.Leuven
	Prof. dr. Monique Snoeck	K.U.Leuven
	Prof. dr. Laure Berti-Eq́uille	AiX Marseille Universite
	Dr. Philip Woodall	University of Cambridge

Daar de proefschriften in de reeks van de Faculteit Economie en Bedrijfswetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

Contents

1 Introduction

1.1 Importance of data quality

The Space Shuttle Challenger broke apart 73 seconds into its flight on January 28, 1986 and killed seven astronauts on board. On July 3, 1986 an Iranian commercial passenger jet was shot down by the U.S Navy Cruiser USS Vincennes and killed all the 290 people on the jet. On February 1, 2003, 17 years later after the Challenger explosion, Space Shuttle Columbia exploded during re-entry to earth's atmosphere and killed all on board. Similar to the above disasters, on September 11, 2001, nineteen terrorists hijacked four commercial passenger jet by passing through airport securities unnoticed and killed more than 2000 people. On the high level scenario, these four disasters are similar that they are responsible for the tragic loss of life and they happened against the intent of their respective organizations. However, a deep analysis to these disasters revealed that information or data quality (DQ) is among the reasons responsible for the happening of the disasters [67, 11, 37, 16].

Incomplete and misleading information are found to be one of the reasons for the Challenger accident [16]. Fisher and Kingma [37] who conducted a thorough analysis of the Vincennes incident indicated that data quality or information quality was a major factor in the USS Vincennes accident. Similarly, the Columbia Accident Investigation Board [16] concluded that the

available data about the foam impact was enough to act upon, however they were considered as irrelevant. Finally, the 9/11 Commission [67] identified that relevant information from the National Security Agency and the CIA was not considered to be relevant to make their ways to criminal investigators. Although data quality or information quality is not the only responsible factor for these disasters, it is impossible to have perfect decisions with many examples of flawed data [37].

A more practical example is the death of a pediatric patient because of a misplaced decimal point in the medicine prescription [7] and the health care organization which overpaid $ 4 million per year in claims for patients who were no longer eligible [147]. Similarly, an eye-wear company has incurred one million dollars annually because of lens-grinding reworks which were caused by data errors [144]. Although losses from poor DQ vary, they are measured in the billions of dollars in addition to costs measured in lives lost, employee and customer dissatisfactions [37, 88, 105, 122]. This indicates corporations are losing millions of dollars due to poor DQ [37, 120, 123]. Davenport states, “no one can deny that decisions made based on useless information have cost companies billions of dollars” [23]. Moreover, the magnitude of DQ problems is continuously growing following the exponential increase in the size of databases [87, 99]. This certainly qualifies DQ management as one of the most important business challenges in today’s information based economy.

Unless specified otherwise, this PhD thesis uses data interchangeably with information. Hence, throughout the text, we use DQ (data quality) or IQ (information quality) and DP (data product) or IP (information product) synonymously.

In Section 1.2, we further describe the research context. Section 1.3 describes the research goal and questions that will be addressed in this PhD thesis. In Section 1.4, the research methodology that was used is presented. Section 1.5 indicates the outlines of the entire thesis. Finally, the chapter ends by listing the articles presented in the thesis (see, Section 1.6.)

1.2 Research Context

Concept of data quality

Data quality is sometimes considered as only incorrect or inaccurate data values [6]. For example, when the surname “Moges” is spelled in a telephone conversation, several misspellings can be made, such as “Mogges”, “Mogess” and “Mojes”, all are inaccurate. It is true that data are usually considered to be of poor quality if they are inaccurate. However, DQ is bigger than only data accuracy. There are many other important DQ dimensions such as completeness, consistency and timeliness which are necessary to fully indicate the quality of the data. In other words, DQ can be measured by many dimensions such as accuracy, completeness, timeliness, relevance, objectivity, believability and others [146, 33]. Some of these dimensions (e.g. accuracy and objectivity) lend themselves to objective measurement that is intrinsic to the data itself, independent from the context in which the data is used. There are however DQ dimensions that cannot be measured objectively. For example, the two recognized DQ dimensions, relevance and believability [152, 36], tend to vary with the usage context. Data relevance mostly depends on the task, since data that are highly relevant for one task may be irrelevant for another - for example, data on depreciation of stocks are required when making up the balance sheet, while being irrelevant for marketing tasks. Data believability is also difficult to assess objectively, since it often depends on the user’s experience and personal preferences - for example, certain data that seems to be believable to a beginner may be less believable to an expert [152, 36]. To understand the contextual effects of DQ, it is important to take factors pertaining to the use of data into account. Hence, DQ is often defined as ‘fitness for use’ which implies the relative nature of the concept [15, 75, 112]. Data has good quality if it satisfies the requirements of its intended use. It lacks quality to the extent that it does not satisfy these requirements. In other words, DQ depends as much on the intended use as it does on the data itself. For example, a database that has a 5% incorrect data element rate will probably be very troublesome to perform a credit risk assessment decision. Yet, the same database at a 5% incorrect rate would probably be very useful and considered high quality for performing an advertising task. Therefore, adapting DQ requirements of one task to others may not be productive as other tasks may have their own DQ requirements. Also, assessing the DQ level for one decision making task

requires using relevant DQ requirements or dimensions. Thus, choosing DQ dimensions to measure the level of quality of data is the starting point of any DQ-related activity.

This PhD thesis is positioned under the contextual DQ management which uses an iterative process towards improving DQ for the financial sector specifically in the credit risk assessment context. In this iterative process, the first step is defining and identifying the important DQ dimensions which are the basis for assessing the quality level of the credit risk databases. The second step is assessing the level of DQ using the identified DQ dimensions and identifying the causes of different DQ problems. Third in the process is to communicate the DQ level to respective users and identify whether the DQ level is in the acceptable range. The final step is implementing the improvement actions suggested.

1.3 Research goal and Questions

1.3.1 Credit Risk Assessment Task

DQ is of special interest and relevance in a credit risk setting because of the introduction of compliance guidelines such as Basel II and Basel III [74]. Since the latter have a direct impact on the capital buffers and hence safety of financial institutions, special regulatory attention is being paid to addressing DQ issues and concerns. Hence, given its immediate strategic impact, DQ in a credit setting is more closely monitored and/or scrutinized, than in most other settings and/or business units [45, 122].

The credit risk assessment task considered in this PhD thesis is subjected to Basel II regulation which demands complete transparency and traceability of data, and is primarily concerned with quantifying the risk of loss of principal or interest stemming from a borrower's failure to repay a loan or meet a contractual obligation. Thus, financial institutions are obliged to assess the credit risk that may arise from their investment. They may estimate this risk by taking into account information concerning the loan and the loan applicant.

The quality of the credit approval process from a risk perspective is determined by the best possible identification and evaluation of the credit risk resulting from a possible default on a loan. Credit risk can be decomposed

into four risk parameters as described in the Basel II documentation [45]. These are Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EaD) and Maturity (M). These parameters are used to calculate the regulatory capital. Regulatory capital, also referred to as a buffer capital, is the money set aside to anticipate future unexpected losses due to loan defaults.

$$Regulatory\ Capital = f(PD, LGD, EaD, M)$$

Incorrect parameters may result in a loss and even bankruptcy of the institution. Therefore, minimizing the errors when quantifying the credit risk parameters is a crucial process [4, 50]. Improving the quality of the data used for calculating these parameters is one way of improving the precision of the parameters.

In this PhD thesis, we define and measure DQ requirements in a credit risk management context from users perspectives. More specifically, in this thesis, we answer the following three broad research questions.

RQ1 What are the important DQ dimensions for a credit risk assessment task? How can we assess the DQ level in credit risk databases using the relevant DQ dimensions identified? (see Chapter 2)

RQ2 How important is incorporating data quality information (information about the quality of data) in databases for decision making? (see Chapter 3)

RQ3 How can we assess the maturity level of data and DQ management activities, and identify Key Process Areas for improvement in financial institutions? (see Chapter 4)

1.4 Research Methodology

The research in this PhD thesis is classified as empirical research. This type of research is fundamentally a problem solving paradigm, which addresses research through direct or indirect observation [49]. Such observation can be analyzed quantitatively or qualitatively. Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions, which should be clearly defined and answerable with the evidence collected usually called data.

1.4.1 Assumptions and Limitations

This thesis was based on the assumption that the data collected through the empirical method reflect the true knowledge and perceptions of the participants with respect to data quality in financial institutions. Similarly, although the data used in this thesis were collected from few participants, it is assumed that the data are enough to indicate the DQ requirements and the DQ level in the sector. This thesis focused on DQ in a credit risk context in financial institutions. As every application may have its own DQ requirements, the results of this thesis may not necessarily reflect the generalized views of organizations outside the financial institutions. However, the methods used in this PhD thesis can be repeated to analyze the data quality requirements of other sectors.

1.5 Outline

This dissertation is a collection of essays dealing with data quality in a credit risk management context in financial institutions. Chapter 2 identifies the relevant data quality dimensions for credit risk assessment and indicates the quality level of credit risk databases using a contextual assessment method. In order to answer RQ1, Chapter 2 used a Total Data Quality Management program (TDQM) [81] and a Methodology for Information Quality Assessment (AIMQ) framework [76]. The results of the analysis confirmed that accuracy is the most important DQ dimension. Also security, relevancy, actionability, accessibility, objectivity, timeliness, value-added and representational-consistency are found to be important DQ dimensions. Likewise, the scorecard index is used to assess the DQ level and to indicate the problem areas.

Chapter 3 extends the idea of continuously improving and mitigating the impact of poor DQ in financial institutions by analyzing the impact of keeping DQ measurement results in the form of metadata in databases as input to the decision making processes. As such, an exploratory study was conducted. The results indicated that the use of data quality metadata (DQM) can be affected by the characteristics of decision makers and task types in question.

As indicated in Chapter 2, data management processes were found to be responsible for DQ problems. Therefore, Chapter 4 further identifies the

data and DQ management processes which need to be improved to enhance the quality of data in financial institutions. As such, it assesses the maturity level of data and DQ management activities in financial institutions using the Information Quality Management Capability Maturity Model (IQM-CMM) [5]. Therefore, different key process areas for improvement are identified and a framework for DQ measuring activities is inferred from best practices in the organizations with high maturity level in the IQM-CMM.

Finally, we end this dissertation with a conclusion Chapter. This Chapter provides general conclusions and some ideas for future research.

1.6 Articles

As it is indicated in Section 1.5, this PhD thesis is a collection of articles either published in scientific journals and conference proceedings, or articles in the process of publication. As such, this section lists the articles with their respective chapters.

Chapter 2 is published in the following journals and proceedings

- Moges, H., Dejaeger, K., Lemahieu, W., Baesens, B. (2013). A multidimensional analysis of data quality for credit risk management: new insights and challenges. Information & Management, 50 (1), 43-58.
- Moges, H., Dejaeger, K., Lemahieu, W., Baesens, B. (2012). A total data quality management for credit risk: new insights and challenges. International Journal of Information Quality, 3 (1), 1-27.
- Moges, H., Dejaeger, K., Lemahieu, W., Baesens, B. (2011). Data quality for credit risk management: new insights and challenges. International Conference on Information Quality (ICIQ) 2011. University of South Australia, Adelaide (Australia), 18-20 November 2011.

Chapter 3 is published in the following journals and proceedings

- Moges, H., Lemahieu, W., Baesens, B. (2012). The use of data quality information (DQI) for decision-making: an exploratory study. International conference on business management and information systems (ICBMIS 2012). Singapore, 22-24 November 2012 (pp. 386-394). New Delhi 110 070: Bloomsbury Publishing India Pvt. Ltd.

- Moges, H., Van Vlasselaer, V., Lemahieu, W., Baesens, B., Determining the use of Data Quality Metadata (DQM) for decision making purposes and its impact on decision outcomes - An Exploratory Study, under second revision in Decision Support Systems Journal.

  Chapter 4 is in the process of submission

2

Data Quality Requirement Analysis For Credit Risk Management

2.1 abstract

Recent studies have indicated that companies are increasingly experiencing Data Quality (DQ) related problems as more and more complex data are being collected. In order to address such problems, literature suggests the implementation of a Total Data Quality Management Program (TDQM) that should consist of the following phases: *data quality definition*, *measurement*, *analysis* and *improvement*. DQ is often defined as 'fitness for use' and although this captures the essence of quality, it is difficult to measure DQ using this broad definition. Thus, it has long been acknowledged that the quality of data is best described or analyzed via multiple attributes or dimensions. Yet, despite broad discussion in the DQ literature, there is no single precisely defined set of DQ dimensions because DQ is context dependent, i.e. a data set with good quality for one task may not be appropriate for another task, even if it requires the same data. Moreover, the type and structure of the data often determine the applicability of a particular DQ dimension. Therefore, to achieve a suitable DQ level, DQ assessment using the most important DQ dimensions for a specific task is a recognized approach. As such, this paper performs an empirical study by means of a questionnaire distributed to financial institutions worldwide to identify and define the most important

DQ dimensions to a credit risk assessment context. This questionnaire is structured according to the framework of Wang and Strong, and incorporates three additional DQ dimensions which were found to be important to the context at hand (i.e. Actionable, Alignment and Traceable). In addition, this paper contributes by developing a scorecard index to assess the DQ level of credit risk databases using the DQ dimensions that were identified as most important. Finally, the paper explores the key DQ challenges and causes of DQ problems in financial institutions.

2.2 Introduction

The risk of poor Data Quality (DQ) increases as larger and more complex information resources are being collected and maintained [102, 80, 92]. Because most modern companies tend to collect increasing amounts of data, good data management is becoming ever more important. As a response, in the last two decades, the aspect of DQ has received a lot of attention, both by organizations worldwide and in academic literature. Several studies are exploring DQ challenges, focusing on DQ measurement and improvement [6, 15, 18, 19, 100, 24, 25, 33, 36, 61, 65, 76, 75, 80, 87, 94, 101, 102, 112, 119, 121, 122, 130, 128, 138, 140, 143, 146, 149, 148, 150, 151]. Fig. 2.1 illustrates this by plotting the increasing number of DQ related publications over the last ten years from ISI Web of Knowledge[1].

In practice, decision makers differentiate information from data intuitively, and describe information as data that has been processed. Unless specified otherwise, this paper uses data interchangeably with information.

DQ is often defined as 'fitness for use' which implies the relative nature of the concept [112, 75, 15]. Data with quality for one use may not be appropriate for other uses. For instance, the extent to which data is required to be complete for accounting tasks may not be required for sales prediction tasks. The first typically requires the availability of all cash balances, e.g. when making up a balance sheet. The latter task on the other hand will always be possible, irrespective of missing cash balances [112, 128]. In addition to task type, the contextuality of DQ can also be explained by the trade-offs between DQ dimensions where one dimension can be favored over others for a specific task. Data quality dimensions are not independent but

[1] http://apps.isiknowledge.com

are in fact correlated [76]. Moreover, if one dimension is considered more important than others for a specific application, then the choice of favoring this dimension may negatively affect other dimensions. For example, having accurate data may require checks which could negatively affect timeliness. Conversely, having timely data may cause lower accuracy, completeness or consistency. A typical situation in which timeliness can be preferred to accuracy, completeness, or consistency is given by most web applications: as the time constraints are often very stringent for web data, it is possible that such data are deficient with respect to other quality dimensions. For instance, a list of courses published on a university web site must be timely though there could be accuracy or consistency errors and some fields specifying additional course details could be missing. Conversely, when considering administrative applications, accuracy, consistency and completeness requirements are more essential than timeliness, and therefore delays are mostly admitted. Another example can be a trade-off between completeness and consistency. A statistical data analysis typically requires a significant and representative set of data and in this case, the approach will be to favor completeness, tolerating inconsistencies, or adopting techniques to solve them. Conversely, when publishing a list of student scores on an exam, it is crucial to check the list for consistency, possibly deferring the publication of the complete list [112, 15]. Accordingly, studying DQ in the context of a specific task is a recognized method [36, 94, 101, 102, 112, 150, 151].

2.2.1 Credit risk assessment task

DQ is of special interest and relevance in a credit risk setting because of the introduction of compliance guidelines such as Basel II and Basel III [4, 53]. Since the latter have a direct impact on the capital buffers and hence on the safety of financial institutions, special regulatory attention is being paid to addressing DQ issues and concerns in this context. Hence, given its immediate strategic impact, DQ in a credit risk setting is more closely monitored than in most other settings and/or business units [45, 122].

The credit risk assessment task is primarily concerned with quantifying the risk of loss of principal or interest stemming from a borrower's failure to repay a loan or meet a contractual obligation. Thus, financial institutions are obliged to assess the credit risk that may arise from their investment. They may estimate this risk by taking into account information concerning the loan

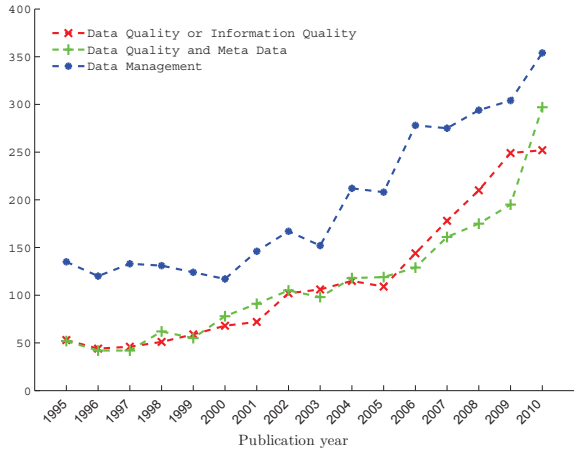


Figure 2.1: *Journal and conference proceedings from ISI Web of Knowledge searched by a query title and business economics domain using the key words information quality or data quality, data quality and metadata, and data management.*

and the loan applicant.

The quality of the credit approval process from a risk perspective is determined by the best possible identification and evaluation of the credit risk resulting from a possible default on a loan. Credit risk can be decomposed into four risk parameters as described in the Basel II documentation [45]. These are Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EaD) and Maturity (M). These parameters are used to calculate the Buffer Capital (BC), also referred to as regulatory capital, which is the money set aside to anticipate future unexpected losses due to loan defaults.

$$\mathrm{BC} = f(PD, LGD, EaD, M)$$

The correct estimation of these parameters and the appropriateness of the function or algorithm used to calculate the risk concentration are crucial since incorrect parameters or inappropriate algorithms may result in a loss and even bankruptcy of the institution. The Risk Concentration (RC) refers to an exposure with the potential to produce losses large enough to threaten a financial institution's health or ability to maintain its core operations [3].

Improving the quality of the data used for calculating these parameters is one way of improving the precision of the parameter estimates and consequently improving the correctness of credit approval decisions [4, 50].

2.2.2 Total Data Quality Management Program

Poor DQ impacts organizations in many ways. At the operational level, it has an impact on customer satisfaction, increases operational expenses and can lead to lowered employee job satisfaction. Similarly, at the strategic level, it affects the quality of the decision making process. An enterprise may experience various DQ problems [75, 122]. Yet, no improvement can be made without knowing and measuring the problems. It is argued in the literature that organizations should implement a Total Data Quality Management (TDQM) program which includes *DQ definition*, *measurement*, *analysis* and *improvement*. This enables them to achieve a suitable DQ level [79].

The *DQ definition* phase is the starting point for a TDQM program, identifying all the necessary DQ dimensions to be measured, evaluated and analyzed. Next, the *measurement* process is implemented. The results from the measurement process are *analyzed* and DQ issues are detected. These issues will be taken into account during the *improvement phase*. In this phase, the collection of poor quality data cases is thoroughly investigated and improvement actions are suggested. The four phases are iterated in this order over time as shown in Fig. 2.2. In fact, the primary goal of DQ assurance is the continuous control of data values and possibly, their improvement [146, 15].

The identification of DQ dimensions from a user perspective defines the list of important DQ dimensions for a specific task that have to be assessed, analyzed and improved [146, 15]. Therefore, the first aim of this paper is to identify the DQ dimensions considered relevant to assess the DQ in the context of credit risk assessment. Second, the paper investigates the impact of different factors such as the existence of DQ teams and the size of financial institutions on the importance of DQ dimensions. Thirdly, the DQ level of credit risk databases is assessed by incorporating the DQ dimensions categorized as relevant and finally, also frequent recurring DQ challenges and their causes in a credit risk assessment context are explored.

The remainder of the paper is structured as follows. The next section explores

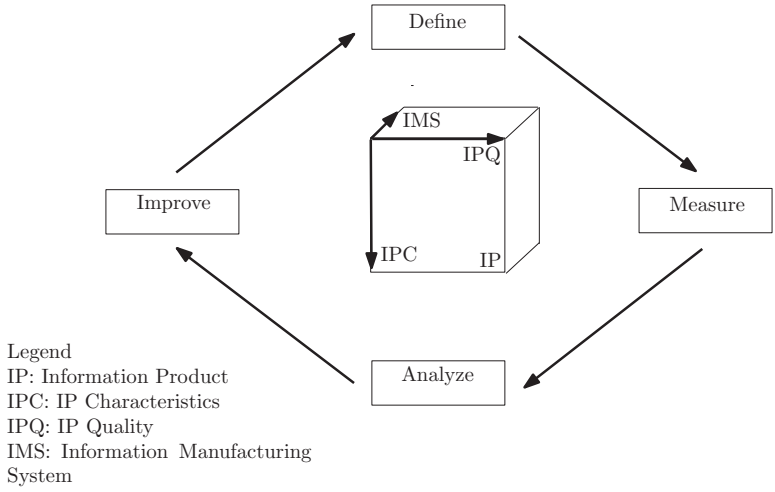


Figure 2.2: *A schematic overview of the TDQM methodology, adopted from Massachusetts Institute of Technology (MIT) [146]*

the related literature while the third section explains our research methodology. The fourth section elaborates on the key findings while the final section elucidates the conclusions and lists topics for further research.

2.3 Related Research

2.3.1 Identification and definition of DQ dimensions

DQ problems cannot be addressed effectively without identifying the relevant DQ dimensions. Thus, a first objective of DQ research is to determine the characteristics of data that are important to, or suitable for data consumers [149]. While fitness for use captures the essence of DQ, it is difficult to measure DQ using this broad definition [6, 65]. Thus, it has long been acknowledged that data are best described or analyzed via multiple attributes or dimensions [87, 128, 140]. Yet, despite broad discussion in the DQ literature, there is no single precisely defined set of DQ dimensions because DQ is context dependent, see e.g. the studies presented in Table 1.

Different studies analyzed DQ from a task specific perspective. For example,

Ref	Intrinsic DQ	Contextual DQ	Representational DQ	Accessibility DQ
[100]	Accuracy, Completeness, Consistency, Validity	Timeliness	Uniqueness	
[149]	Accuracy, Believability, Reputation, Objectivity	Value-added, Relevancy, Completeness, Timeliness, Appropriate amount	Understandability, Interpretability, Concise and consistent representation	Accessibility, Ease of operations, Security
[143]	Correctness, Unambiguous	Completeness	Meaningfulness	
[25]	Accuracy, Precision, Reliability, Freedom from bias	Importance, Relevance, Usefulness, Informativeness, Content, Sufficiency, Completeness, Currency, Timeliness	Understandability, Readability, Clarity, Format, Appearance, Conciseness, Uniqueness, Comparability	Usableness, Quantitativeness

Table 2.1: *DQ dimensions from the literature ordered according to the framework of Wang and Strong [149]*

Zhu and Gauch [160] assessed the DQ of a web page in terms of a DQ framework comprising six DQ dimensions, namely currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness. They measured the dimensions through the properties of web pages. Similarly, Chien et al. [18] assessed different DQ dimensions in order to evaluate the quality of online product reviewing by customers. Of course, they adopted some sort of definitions of different DQ dimensions for the quality analysis of the online product reviews. For example, they defined objectivity as the extent to which an information item is biased, appropriate amount of data as the extent to which the volume of information in a review is sufficient for decision making, and completeness as the extent to which the information in a review is complete and covers various aspects of a product. Furthermore, they identified objectivity and appropriate amount of information as effective DQ dimensions in identifying product review quality but assessed completeness as a very ineffective DQ dimension to measure the quality of a product review by customers or other parties.

On the other hand, there are a number of studies which identify and define DQ dimensions regardless of the use of the data in order to facilitate the general applicability and comparability of their DQ dimensions. In this regard, Wand and Wang [143] based their definition of DQ on the internal view of information systems (data production and system design processes) because this view is context independent. This approach allows for a set of definitions of DQ dimensions that are comparable across applications. First, they identified different criteria for a real-world system to be properly represented by an information system. Based on these criteria, they defined four deficiencies namely ambiguous representation, incomplete representation, meaningless states, and operation deficiencies. Based on these deficiencies, they summarized different DQ aspects into complete, unambiguous, meaningful, and correct DQ dimensions. In addition, in the same study, they categorized different DQ dimensions from the literature as internal view (design or operation related) and external view (use or value related), whereby both views were further refined as either system or data related DQ dimensions. Within the internal view, accuracy or precision, timeliness or currency, reliability, completeness and consistency are defined as data related and reliability is defined as a system related DQ dimension. On the other hand, in the external view, timeliness, relevance, content, importance and sufficiency are defined as data related and timeliness, flexibility, format and efficiency are defined as system related DQ dimensions.

Similarly, Wang and Strong [149] analyzed the various DQ dimensions from end user's perspectives but regardless of the use of the data. They conducted a large scale survey to determine and categorize the DQ dimensions. Their analysis began by collecting information from users regarding various DQ descriptors that resulted in over 100 items that were grouped into 20 categories. These were further aggregated into four broad DQ categories: intrinsic (the extent to which data values are in conformance with the actual or true values), contextual (the extent to which data are applicable to the task of the data user), representational (the extent to which data are presented in an intelligible and clear manner), and accessibility (the extent to which data are available or obtainable). Table 2.1 illustrates the framework of Wang and Strong by classifying the DQ dimensions used in different studies according to this framework [149].

A waterfall[2] based literature survey was adopted to identify the most often recurring DQ dimensions and their definitions. As such, only dimensions adopted in three or more papers were retained, together with the DQ dimensions from our own pilot survey, see Section 2.4.2.1. In fact, we adopted the DQ framework of Wang and Strong to classify the DQ dimensions [149]. This framework is recognized as the only one that attempts to strike a balance between theoretical consistency and practicability. Furthermore, the framework has been found to be applicable to various domains [33]. The structure of the framework is hierarchical, and it organizes DQ aspects along fifteen DQ dimensions to comprehend the four broad DQ categories. Table 2.2 provides an overview of the DQ dimensions considered in this study. We believe that these DQ dimensions provide a comprehensive coverage of the multi-dimensional nature of DQ. Hence, in this paper, we used this summary to measure the applicability of the DQ dimensions for the credit risk assessment task.

2.3.2 Data quality: intrinsic and contextual

DQ can be measured by many dimensions such as accuracy, completeness, timeliness, relevance, objectivity, believability and others [146, 33]. Some of these dimensions (e.g. accuracy and objectivity) lend themselves to objective measurement that is intrinsic to the data itself, independent from the context

[2]A literature review which is based on the waterfall life cycle phases such as requirement specifications, selecting and analyzing studies based on the requirements [68]

Cat.	DQ dimensions	Definitions	References
Intrinsic	Accuracy (AC)	The extent to which data are certified, error-free, correct, flawless and reliable	[3, 6, 17, 24, 65, 76, 94, 102, 112, 119, 122, 128, 138, 140, 143, 146, 149, 151]
	Objectivity (OBJ)	The extent to which data are unbiased, unprejudiced, based on facts and impartial	[3, 17, 24, 65, 76, 94, 112, 138, 146, 149]
	Reputation (REP)	The extent to which data are highly regarded in terms of its sources or content	[6, 17, 65, 76, 94, 112, 138, 149]
Contextual	Completeness (COM)	The extent to which data are not missing and covers the needs of the tasks and is of sufficient breadth and depth of the task at hand	[3, 15, 17, 24, 65, 76, 94, 101, 102, 112, 122, 128, 138, 143, 146, 149, 151]
	Appropriate-amount (APM)	The extent to which the volume of information is appropriate for the task at hand	[17, 65, 76, 94, 102, 138, 146, 149]
	Value-added (VAD)	The extent to which data are beneficial and provides advantages from its use	[17, 65, 76, 94, 102, 138, 146, 149]
	Relevance (REL)	The extent to which data are applicable and helpful for the task at hand	[17, 65, 76, 94, 102, 138, 143, 146, 149]
	Timeliness (TIM)	The extent to which data are sufficiently up-to-date for the task at hand	[17, 24, 65, 76, 94, 112, 128, 138, 143, 146, 149, 151]
	Actionable (ACT)	The extent to which data is ready for use	pilot survey
Representation	Interpretable (INT)	The extent to which data are in appropriate languages, symbols, and the definitions are clear	[17, 65, 76, 94, 112, 146, 149]
	Easily-understandable (EU)	The extent to which data are easily comprehended	[17, 65, 76, 94, 112, 138, 146, 149]
	Representational-consistent (RC)	The extent to which data are continuously presented in same format	[17, 65, 76, 94, 112, 122, 138, 140, 143, 146, 149]
	Concisely-represented (CR)	The extent to which data is compactly represented, well-presented, well-organized, and well-formatted	[17, 65, 76, 94, 112, 122, 138, 140, 143, 146, 149]
	Alignment (AL)	The extent to which data is reconcilable (compatible)	pilot survey
Access	Accessibility (ACC)	The extent to which data is available, or easily and swiftly retrievable	[17, 65, 76, 94, 112, 138, 146, 149]
	Security (SEC)	The extent to which access to data is restricted appropriately to maintain its security	[17, 65, 76, 94, 112, 138, 146, 149]
	Traceability (TRA)	The extent to which data is traceable to the source	pilot survey & [81, 159]

Table 2.2: *Most cited DQ dimensions in literature (attributes) and their definitions*

in which the data is used. There are however DQ dimensions that cannot be measured objectively. For example, the two recognized DQ dimensions, relevance and believability [152, 36], tend to vary with the usage context. Data relevance mostly depends on the task, since data that are highly relevant for one task may be irrelevant for another - for example, data on depreciation of stocks are required when making up the balance sheet, while being irrelevant for marketing tasks. To understand the contextual effects of DQ, it is important to take factors pertaining to the use of data into account. In this regard, factors such as the relevance of the data to the task, the ability of the user to understand it, and the clarity of the task, all affect the usability of that data [151]. From this usage perspective, DQ assessment tends to be contextual. Furthermore, users that suppose data to be of poor quality are unlikely to weigh it heavily in their decision making tasks, even if it is objectively of high quality.

Many researchers such as Fisher et al. [36] and Shankaranarayanan et al. [129] identified the impact of experience, task type and time constraints on the possible use of DQ information. Their results indicated that when experience level increases and task complexity decreases, information about the specific quality of data is more often used in decision-making tasks. Likewise, other researchers such as Price et al. [116] investigated the impact of decision-making strategies on the use of DQ related information. In general, all these studies illustrate the existence of factors that can affect DQ assessment which encourages us to investigate whether other factors such as the existence of DQ teams and the size of organizations would impact DQ assessment.

2.3.3 Data quality: representation and access

The most frequently mentioned DQ dimensions in the representation and access DQ categories are representational-consistency, easily-understandable, accessibility and security. The representational-consistency and easily-understandable DQ dimensions assess the representation and understandability of data respectively. Typical issues such as using different currencies, different formats and different names for similar columns or rows are addressed by the representational-consistency DQ dimension. On the other hand, the latter two DQ dimensions assess, respectively, the easiness of accessing and the security of data. The accessibility DQ dimension, for example, deals with the

request and delivery time of output. For example, data can be classified as unaccessible if the gap between input and delivery time of output is too large [138].

2.3.4 DQ assessment

The level of DQ can be assessed using a questionnaire or metrics. Lee et al. [76] developed a methodology to assess DQ level using questionnaires. This methodology is called 'A Methodology for Information Quality Assessment' (AIMQ). The foundation of this methodology is a 2x2 table, called the PSP/IQ model, classifying DQ dimensions according to their importance from both user's and manager's perspectives. The axes of the table are conformity to specifications and conformity to user's expectations. Accordingly, four DQ categories are distinguished (sound, dependable, useful and usable) and DQ dimensions identified in Wang and Strong's framework [149] are classified along these categories. The *sound* DQ category relates to the intrinsic value of DQ and particulary deals with DQ dimensions such as free of error (accuracy), concise representation and completeness. The *useful* DQ category deals with the context dependent nature of DQ. This category includes aspects like appropriate amount, relevancy, interpretability and understandability of information. The *dependable* DQ category revolves around the timeliness and security of data while the *usable* DQ category is concerned with the accessibility, the reputation and believability of the data. The PSP/IQ model is used to aggregate scores of the DQ dimensions. Hence, two gap analysis techniques (IQ benchmark and role gap) are used to analyze the results from this model. IQ benchmark gap analysis is used to compare with the best performing organizations for the four DQ categories. The role gap technique is used to investigate differences of DQ level assessment among different roles in the organizations. For example, DQ assessments by information professionals and data users are compared.

On the other hand, DQ metrics are being developed to assess DQ level in organizations. Table 2.3 shows a number of exemplary DQ metrics proposed by different authors. Pipino et al. [112] developed DQ metrics based on three functional forms such as simple ratio, min/max operation, and weighted average. Simple ratio is used to build up DQ metrics for the accuracy, completeness and consistency DQ dimensions. The metric for the accuracy DQ dimension is given in Table 2.3. Metrics for the completeness and consistency

DQ Dimension	DQ Metric		Ref.
	Task independent	Task dependent	
Accuracy	$1 - \frac{number\ of\ data\ units\ in\ error}{Total\ number\ of\ data\ units}$ $f(accuracy\ percentage, a\ randomness\ measurement, probability\ distribution)$ *Bayesian Network Approach*		[112, 75] [38] [127]
Appropriate-amount	$\min[\frac{number\ of\ data\ units\ provided}{number\ of\ data\ units\ needed}, \frac{number\ of\ data\ units\ needed}{number\ of\ data\ units\ provided}]$		[112, 75]
Timeliness	$Q_{curr.} = e^{-decline(A) \cdot age(W.A)}$	$\max[(1 - \frac{Currency}{Volatility}), 0]^s$	[112, 75, 55]

Table 2.3: *DQ metrics from literature*
Conotations for task independent metrics *: f denotes 'function of', $Q_{curr.}$ is the currency level of the data,* A *is an attribute,* w *is an attribute value, age refers to the difference between the instant when DQ is assessed and the instant of data acquisition and $-decline$ refers to the average decline rate of the shelf life of attribute values of the attribute under consideration.*
Conotations for task dependent metrics*: $currency = (delivery\ time - input\ time) + age$, volatility refers to the length of time over which the data remains valid, delivery time refers to the time at which the data was delivered to the user, input time refers to the time at which the data was received by the system, age refers to the age of the data when it was first received by the system and the exponent s is task dependent and used to control the sensitivity of timeliness.*

DQ dimensions are defined analogously. Conversely, metrics for the believability, appropriate-amount, timeliness and accessibility DQ dimensions are developed using min/max operations. Metrics quantifying the appropriate-amount and timeliness DQ dimensions are given in Table 3 as examples. The appropriate-amount metric is based on its most recognized definition, namely that the amount of data should neither be too little nor too much [6]. On the other hand, a metric for the timeliness DQ dimension is defined as the maximum of one minus the ratio of currency and volatility, and 0 [75], see Table 2.3.

More recently, Fisher et al. [38] proposed an accuracy metric by changing the simple ratio scale to a vector approach which includes percentages, a randomness measure, and a probability distribution. The metric combines a simple ratio, number of cells in errors to total number of cells, with a randomness measure computed using the Lempel-Ziv complexity measure algorithm[3]. This algorithm is used to differentiate whether the errors in a database are random or systematic in nature. Once the randomness of the errors is determined, a probability distribution is used to help address various managerial questions. The metric is based on the assumption that the value corresponds to the possible validity range.

Similarly, Sessions et al. [127] suggested a measuring approach for accuracy using bayesian networks[4].

The metrics discussed earlier are task independent. Yet, a task dependent metric is formulated for currency, one aspect of the timeliness DQ dimension, by Heinrich and Klier [55] as shown in Table 2.3. The metric is based on quality of conformance, which is mainly related to data values and more independent of a particular user's demand in a specific business situation. The metric is defined as a probability that an attribute value stored in a database still corresponds to the current state of its real world counterpart at the moment when the DQ level is assessed. This metric depends on the context in which it will be applied. If the timeliness dimension is not critical to the task at hand, then a more relaxed sensitivity measure can be applied. Conversely, if the dimension is very critical, a conservative sensitivity measure

[3]Lempel Ziv is an algorithm for lossless data compression. In fact, it is not a single algorithm, but a whole family of algorithms, stemming from the two algorithms proposed by Jacob Ziv and Abraham Lempel in 1978 [158].

[4]A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) [127].

is suggested.

Many metrics are being developed by different DQ researchers and organizations to assess the DQ level in organizations using DQ dimensions. Many of the dimensions are multivariate in nature; which variables/components of the metrics are important to the organization must be clearly identified and defined. Choosing the specific variables or components to measure can be much more difficult than defining the general metric, which often reduces to the ratio form. Although falling within a specific dimensional category, the measure to assess a specific dimension will vary from organization to organization [75]. Thus, though a comparison of the objective and subjective assessment of the DQ level may indicate the actual DQ level and define the gap between objective and subjective DQ level assessment, the focus of this paper is only DQ assessment using a questionnaire.

2.3.5 DQ Challenges

As more data are collected and maintained, the risk of poor DQ increases. Multiple data sources, subjective judgment in data production, security/accessibility trade-off, and changing data needs are often mentioned challenges [75]. For example, multiple sources of the same data produce different values for that data. For instance, similar accounting data held in different files are very likely to differ to each other as updating or changing all the files at the same time is not always possible. This is also illustrated by system designers' tendency to avoid having similar data in different files or, in other cases, to enforce transactional consistency among replicated data. Similarly, using several different processes is also likely to produce different values for the same information [87]. Like multiple sources of data, subjective judgment of data is also a challenge for DQ. Information production using subjective judgment often produces biased information. Data stored in an organization's database is considered to be a set of facts. However, the process by which these 'facts' are collected may involve subjective judgments. For example, the expense codes assigned to indicate different allowances paid to employees by an accountant can be biased by the accountant's knowledge. The security/accessibility trade-off is also a challenge for DQ. Easy access to information may conflict with requirements for security, privacy, and confidentiality. For data consumers, high-quality data must be easily accessible. However, ensuring privacy, confidentiality, and security of infor-

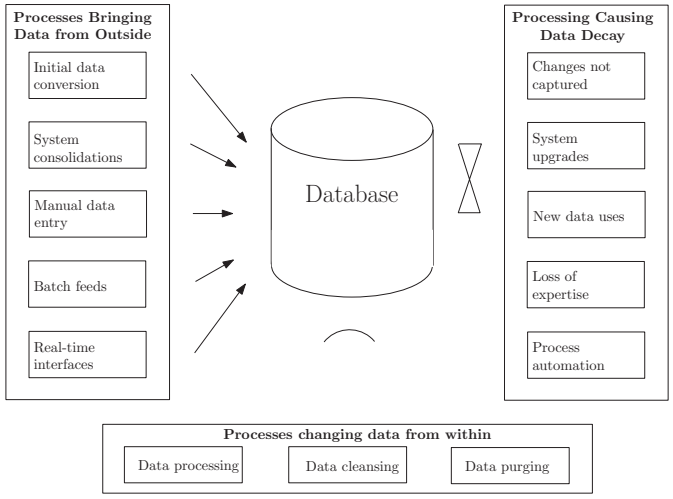


Figure 2.3: *Different data inputting and manipulating processes, adopted from Maydanchik [87]*

mation requires barriers to access. The other most recognized challenge is changing data needs. As information consumers' tasks and the organization environment change, the data that used to be relevant and useful may become obsolete [6].

In fact, DQ improvement actions require the identification of the causes of data errors and their permanent elimination through an observation of the whole process where data are involved [6, 15, 87]. Data are impacted by many processes, most of which affect their quality to a certain degree. Fig. 2.3 shows different data inputting and manipulation processes as identified by Maydanchik [87]. Measuring the impacts of data inputting and manipulating processes on DQ is necessary for proper DQ improving activities. In this paper, we identify different DQ challenges and their main causes in financial institutions.

2.4 Research Methodology

The research methodology is developed alongside four research aims. Fig. 2.4 shows the four aims of the study.

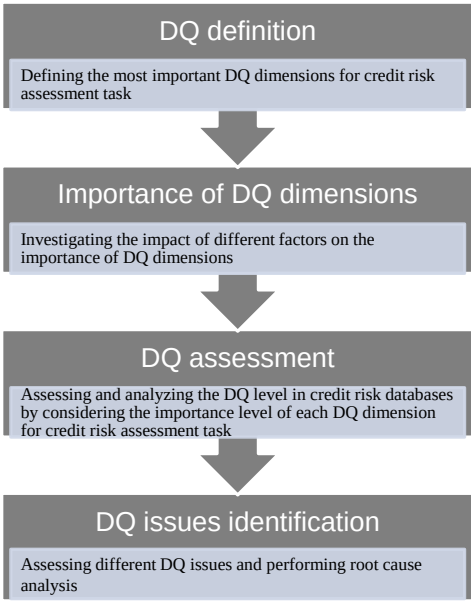


Figure 2.4: *The aims of the study*

2.4.1 Research aims

Data of sufficient quality considered appropriate for one task may not be of sufficient quality for another task [140]. Therefore, identifying and defining DQ dimensions which are relevant to assess the DQ of one specific task is a recognized approach [160, 18]. Thus, the first aim of this paper is to identify the most important and relevant DQ dimensions for the credit risk assessment task to assess the DQ level.

As we discussed in Sections 2.3.1 and 2.3.2, task type (simple or complex) and user experience (beginner or domain expert) are found to have an impact on DQ assessment [15]. Thus, these two factors are controlled in this study. The study subjects have similar experience on the credit risk assessment task. However, the impact of various DQ enhancing activities such as the implementation of DQ teams and the impact of size of financial institutions on DQ assessment by the decision makers remains, to the best of our knowledge, unexplored until now. Hence, taking these aspects into account, the second aim of this paper is to test whether the importance of the DQ dimensions in Table 2.2 differs depending on the existence of DQ teams, the size of financial

institutions and between financial institutions and other companies.

Assessing the DQ level is a crucial step for DQ improvement actions since it indicates the DQ problem areas [76]. Thus, the third aim of this paper is to assess the DQ level of credit risk databases by considering the degree of importance of each of the DQ dimensions identified in the first aim of the study. Assessing the DQ level using the most important DQ dimensions helps to identify the most critical DQ problem areas. Also, it indicates potential DQ improvement actions.

Finally, the frequent DQ challenges and their causes, the DQ improving activities and the motivation for DQ improvement activities in the context of credit risk assessment are investigated. Identifying frequent DQ challenges and sources of these challenges leads to sustainable DQ improvement actions because DQ problems can be mitigated from their source.

2.4.2 Empirical study

The data for this empirical study are collected in the form of a survey taken from financial institutions worldwide. The advantage of adopting an empirical approach is that it captures task specific user's requirements [149]. Furthermore, it may reveal characteristics that researchers have not defined as part of a general DQ definition.

2.4.2.1 Pilot study

To verify the setup and clarity of the questions/items in the survey, a pilot study was conducted. The pilot and final study questionnaires are based on Lee et al. [75] and Wang et al. [149]. The pilot study, which has 21 questions (see appendix), was organized using an online survey tool. The link to the pilot study questionnaire was sent to three respondents (two males and a female), who can be considered experts to this research. The respondents' minimum educational level is a master degree, and all previously participated in data governance activities. Overall, they have 6-10 years of managerial experience in the risk department of financial institutions. They took an average answer time of 30 minutes to finish the questionnaire.

The pilot study also helped to identify DQ dimensions important to the credit risk assessment task but not shown in Wang and Strong's IQ framework [149].

Subjects were asked to list as many DQ dimensions as they found relevant for their task in addition to the given framework (see Question 1, 2 and 3 of part I). As a result, 'actionable', 'alignment' and 'traceability' DQ dimensions were identified. The categories and definitions of these three dimensions are also determined by the subjects (see Table 2.2). These three DQ dimensions are included in the full study.

Actionable means that data do not require additional manipulation and are readily usable. This dimension is related to the 'relevant' dimension in the sense that the former is dependent on the latter; the actionability of the data should be investigated once its relevancy is established.

The *alignment* DQ dimension is defined as 'the extent to which data are well-matched or compatible'. If two similar data elements from different data sources contain similar attributes, it may very well be that both data elements are still very differently structured and/or formatted. If, despite these differences, it is possible to integrate both data elements, they are said to be reconcilable or compatible.

Traceability refers to the extent to which data are traceable to their sources.

The pilot study finally provided feedback as to the usability and clarity of the study instruments and the clarity and consistency of the procedures.

2.4.2.2 Final study

Participants

Among a set of 500 financial institutions worldwide determined by multiple business experts, a random subset of 150 financial institutions was taken. The study subjects are managers of the credit risk department who are responsible for developing or assessing credit risk models. The majority of subjects (48%) have more than 10 years of working experience and most of them (64%) worked in the company where the survey is conducted for already 1 to 5 years. Using a Friedman test at $\alpha = 5\%$, it was verified that no statistically significant differences exist across experience groups.

Subjects of the other sectors are also selected in a similar way.

Study design and procedures

The survey was organized into two different sessions, and consisted of three separate parts. The first session, which consisted of the first and second part of the survey, has been sent to a sample group of 150 financial institutions. In the first part, the importance of the DQ dimensions defined in Table 2.2 is measured. Subjects are provided with Table 2.2 and are asked to rate the importance of the DQ dimensions listed on a scale from 0 to 10 for their task (credit risk assessment), where 0 was not important at all, 5 was somehow important and 10 indicated high importance. In the second part of the survey, subjects are asked to assess the DQ level of their own data using the same scale (0-10). They are provided with different controlling questions/items of the same DQ dimension. For example, they are asked to indicate the extent to which two analogue statements apply to their data, e.g. 'the data are error free' and 'the data are accurate'. Two or three questions for each DQ dimension, totaling 31 questions, are asked. Of the 150 questionnaires mailed to financial institutions, 64 (an effective response rate of 42.67%) were returned.

Similarly, of the 150 questionnaires mailed to organizations in other sectors, 30 (an effective response rate of 20%) were returned.

Next, during a follow up session, the third part of the survey was sent to those who replied in the first session. Note that this second session was also organized in the same time frame as the first session and was only conducted in financial institutions. As a result, among the 64 respondents in the first session, only 37 also participated in the second session. In the third part of the survey, the respondents are asked 6 questions to identify recurring DQ problems and their magnitude, the motivation for DQ initiatives in their department and if there are any DQ improving activities in place.

2.4.3 Statistical analysis

In order to test the significance of the obtained results, a number of statistical tests are applied in accordance with the literature. Each of the different tests is assessed at a significance level of 5% unless stated otherwise. The following notation is adopted throughout the remainder of this paper. Financial institutions are denoted by $i = 1 \ldots N$, while DQ dimensions are denoted as $j = 1 \ldots P$. N and P indicate the total number of financial

institutions and DQ dimensions respectively. d_j is used to indicate the j^{th} DQ dimension.

When items are used to form a scale they need to have internal consistency. Items measuring the same aspects should be correlated with one another. A useful coefficient for assessing internal consistency is Cronbach's alpha, α_c [63]. It is defined as:

$$\alpha_c = \frac{K}{K-1}\left(1 - \frac{\sum s_l^2}{s_T^2}\right)$$

where K is the number of items/questions under one DQ dimension, s_l^2 is the variance of the l-th item and s_T^2 is the variance of the total score formed by summing all the items. The reliability of the study instruments is confirmed as α_c was found to be 0.82 or higher for each dimension.

Before adopting specific statistical tests, the underlying assumptions made by these tests should be fulfilled. Parametric tests such as Analysis Of Variance (ANOVA) and t-tests both assume the data are normally and IID (Independently and Identically Distributed) [54]. A Jarque-Bera test was adopted to verify the normality of the data. The Jarque-Bera test is a two-sided goodness-of-fit test used to verify the null hypothesis that the data comes from a normal distribution with unknown variance and mean. It has an asymptotic χ^2 distribution with 2 degrees of freedom. The test statistic takes on the following form:

$$JB = \frac{N}{6}\left(s^2 + \frac{(k-3)^2}{4}\right)$$

where N represents the sample size, s the sample skewness and k the sample kurtosis.

As the null hypothesis of normality was rejected for ten out of seventeen DQ dimensions at α = 5%, we used non-parametric tests in the remainder of the analysis.

To compare the survey results across DQ dimensions, a Friedman test was adopted which is a non-parametric equivalent to the well known ANOVA test [41]. This test detects differences across all DQ dimensions and is defined as:

$$\chi_F^2 = \frac{12N}{P(P+1)}\left[\sum_{j=1}^{P} AR_j^2 - \frac{P(P+1)^2}{4}\right]$$

with AR_j the average rank of j-th DQ dimension for N financial institutions. Under the null hypothesis, the Friedman test statistic is χ_F^2 distributed with $P-1$ degrees of freedom, at least when N and P are big enough ($N > 10$ and $P > 5$). In this survey, $N = 64$ and $P = 17$.

Next, since the assumption of equality between all DQ dimensions is rejected, we proceed with a post-hoc Bonferroni-Dunn test. The Bonferroni-Dunn test is a non-parametric alternative to the Tukey test and compares the DQ dimensions with the dimension associated with the highest average rank (AR). The difference between two dimensions is found to be significant if the corresponding average ranks (ARs) differ by at least the critical difference:

$$CD = q_\alpha \sqrt{\frac{P(P+1)}{6N}}$$

where q_α is drawn from a Studentized range statistic divided by $\sqrt{2}$. This test also incorporates an additional Bonferroni correction by dividing the confidence level α by the number of comparisons made, $P-1$, to control for family wise testing, thus resulting in a stronger test.

In case of comparing the sample median between two groups, a (non-parametric) Wilcoxon ranked sum test is used. This test hypothesizes that the data comes from two unknown distributions with equal median [153]. All statistical tests were implemented in Matlab[5].

2.5 Results and Discussion

In this section, we present and discuss the key findings of the study. In Section 2.5.1, we present the results of the statistical analysis which define and identify the most important DQ dimensions for the credit risk assessment task. In Section 2.5.2, we discuss the DQ level assessment by using the outputs of Section 2.5.1. Finally, the results of Section 2.5.3 explain the key DQ challenges, the key causes of DQ problems and the motivations of DQ enhancing activities in financial institutions.

[5] www.mathworks.com

2.5.1 Aim 1: Importance of DQ dimensions

Our first research aim is addressed by analyzing the first part of the survey where the respondents were asked to rate the importance of each of the DQ dimensions given in Table 2.2. The overall results are presented in Table 2.4. All seventeen DQ dimensions in Table 2.2 are attributed a score higher than 7/10, indicating the importance of each dimension for credit risk assessment. The results in Table 2.4 are further analyzed by first performing a Friedman test, which detects if there are statistically significant differences between the scores of all DQ dimensions. The null hypothesis is strongly rejected (p value < 0.001) indicating significant differences exist in the results of the survey. Thus, we proceed with a Bonferroni-Dunn test. The results of the Bonferroni-Dunn test are depicted in Fig. 2.5. The x-axis in this figure corresponds to the average rank (AR) for each of the DQ dimensions. The DQ dimensions are represented by a horizontal line; the more this line is situated to the right, the higher the scores on that DQ dimension. The right end of this line depicts the average ranking while the length of the line corresponds to the critical distance. If the difference in average ranking between a DQ dimension and the 'best' DQ dimension is more than this critical distance, the difference is significant at the 99% confidence level. The 'best' DQ dimension is a DQ dimension which has the highest average ranking. The dotted, dashed and full vertical lines in the figure indicate the critical difference at respectively the 90%, 95% and 99% confidence level. The scores on a DQ dimension are significantly lower than those of the 'best' dimension if it is located at the left hand side of the vertical line.

Accuracy clearly was attributed the highest score as it is the most right-positioned DQ dimension as shown in the results of the Bonferroni-Dunn test in Fig. 2.5 and consequently is confirmed to be the most important DQ dimension. Since accuracy is found to be the best scoring dimension, it is used to compare the average scores of each of the other sixteen DQ dimensions. The scores for security, relevancy, actionability, accessibility, objectivity, timeliness, value-added and representational-consistency are found to be not significantly different at the 99% confidence level. Based on these results, we can conclude that accuracy and those dimensions with no significant lower AR are the most important DQ dimensions to assess the DQ level for the credit risk assessment task. On the other hand, the completeness, interpretability, reputability, traceability, easily-understandable, appropriate-amount, alignment and concise-representation DQ dimensions are found to be significantly

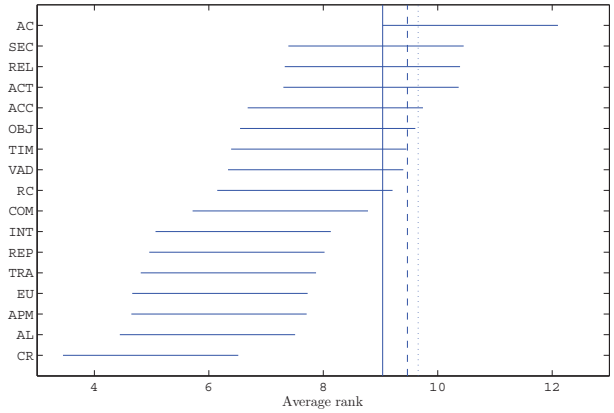


Figure 2.5: *Bonferroni-Dunn plot of the relative importance of the DQ dimensions in financial institutions*

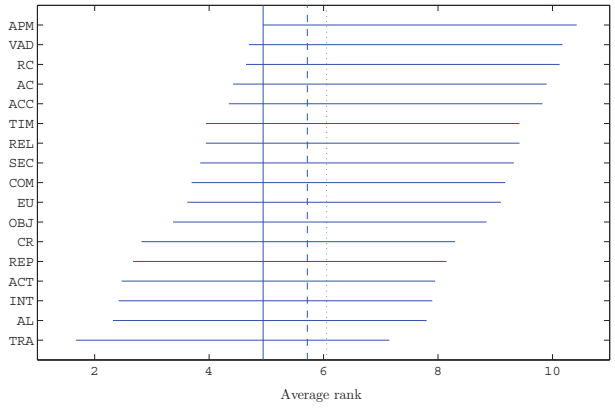


Figure 2.6: *Bonferroni-Dunn plot of the relative importance of the DQ dimensions as assessed by other sectors*

DQ Dimension	Mean	SD	95% C.I.
Accuracy(AC)	9.08	1.54	8.69–9.46
Actionable(ACT)	8.53	1.63	8.12–8.94
Relevancy(REL)	8.52	1.53	8.13–8.9
Security(SEC)	8.47	2.08	7.95–8.99
Accessibility(ACC)	8.41	1.61	8.00–8.81
Timeliness(TIM)	8.28	1.79	7.83–8.73
Value-added(VAD)	8.27	1.94	7.78–8.75
Objectivity(OBJ)	8.19	2.20	7.64–8.74
Representational-consistent(RC)	8.13	2.22	7.57–8.68
Completeness(COM)	8.02	2.31	7.44–8.59
Reputability(REP)	7.89	1.88	7.42–8.36
Interpretability(INT)	7.86	2.05	7.35–8.37
Appropriate-amount(APM)	7.84	1.86	7.38–8.31
Easily-understandable(EU)	7.81	1.93	7.33–8.30
Alignment(AL)	7.75	2.05	7.24–8.26
Traceability(TRA)	7.73	2.23	7.18–8.29
Concisely-Represented(CR)	7.36	2.21	6.81–7.91

Table 2.4: *Basic statistical description of DQ dimensions (mean, standard deviation (SD) and confidence interval (C.I.))*

less important, see Fig. 2.5.

A Bonferroni-Dunn test is also performed to the other sectors' data of which the results are depicted in Fig. 2.6. These sectors include telecommunication, retail, food, pharmaceutical, chemical and health care industries. From the results in Fig. 2.6, we can see that all the DQ dimensions are suggested as very important unlike the financial sector. Also, the relative importance of the DQ dimensions is very different compared to Fig. 2.5. The results suggest that DQ requirement analysis should be done in each sector individually instead of combining different sectors together, because DQ is context dependent. Nevertheless, the same method can be used. Although we cannot compare the Bonferroni-Dunn result for the financial sector against each other sector's Bonferroni-Dunn results because of the small sample size in the other sectors' data, we believe that the Bonferroni-Dunn result for another individual sector would be much different than the one shown in Figure 2.6 (which includes all the sectors together) because DQ depends much on the context of the task as it does on the data itself [111, 18].

Finally, a non-parametric Wilcoxon signed-rank test was performed in order to test if there is a difference in the relative importance of the DQ dimensions between financial institutions with and without DQ teams, and between large and small and medium (SME) financial institutions of which the results are shown in Fig. 2.7 and Fig. 2.8. Both figures (Fig. 2.7 and 2.8) show there is no significant difference in the relative importance of DQ dimensions between

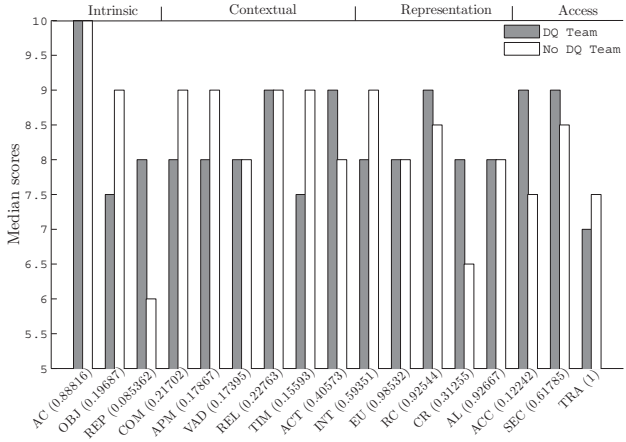


Figure 2.7: *The results of the Wilcoxon ranked sum test, comparing the medians of the DQ dimensions for financial institutions with and without DQ teams; p values are indicated between brackets*

financial institutions with and without DQ teams, and between large and SME financial institutions respectively. This result confirms that DQ depends much on the characteristics of the task [159]. Therefore, we can conclude that the results in Fig. 2.5 are applicable to all financial institutions irrespective of the presence or absence of DQ teams and the size of financial institutions.

For financial institutions' data, the correlation between DQ dimensions is also investigated using the Spearman's rank correlation, ρ. This is a non-parametric correlation measure which investigates the monotonic relationship between any two DQ dimensions. ρ is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} r_i^2}{N(N^2 - 1)}$$

with N the sample size and r_i the difference between the ordinal ranks assigned to each of the observations. The significance of the Spearman's rank correlation measure is given in Table 2.6 and 2.6. These results show that most of the DQ dimensions are correlated with each other. The black and grey cells show the significance of the correlation among the DQ dimensions at the 99% and 95% confidence level respectively while a white cell indicates no correlation between two DQ dimensions.

As the results in Table 2.6 indicate, the importance level of the majority of the

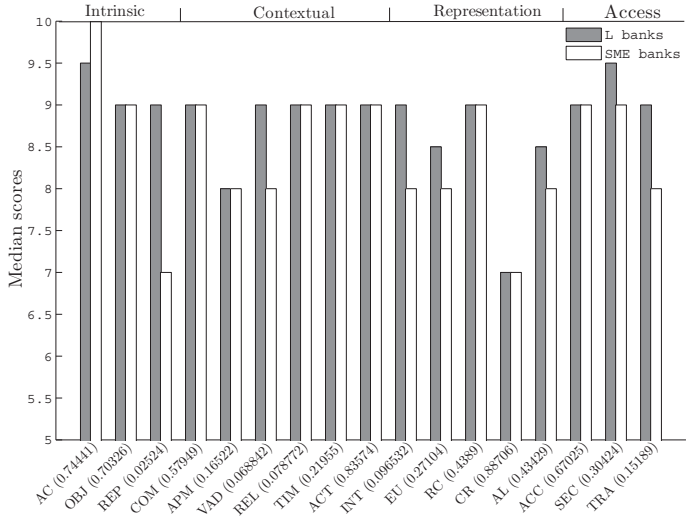


Figure 2.8: *The results of the Wilcoxon ranked sum test, comparing the medians of the DQ dimensions for large and SME financial institutions; p values are indicated between brackets*

DQ dimensions are positively correlated to each other. Accuracy is correlated with the majority of other DQ dimensions, clearly illustrating the business analyst's tendency of equalizing it with the total DQ requirements. In fact, the problem of inaccuracy can be related to many of the DQ dimensions. For example, a null value for the age of a customer can be both associated to completeness and accuracy DQ dimensions. Accuracy can also relate to the representational-consistency DQ dimension. For example, a birthdate value of a person represented in DDMMYY and MMDDYY format can indicate both inaccuracy and inconsistency problems. Easily-understandable, interpretability and actionability DQ dimensions are also highly correlated to each other. As a rule of thumb, a person that understands the data is able to interpret the data. Likewise, if a decision maker understands the data, it is more likely that he/she will use it, hereby enhancing its actionability. The strong positive correlation observed in the results of Table 2.6 are also supported by the literature. Lee et al. also found high correlation between the importance level of a number of DQ dimensions. They reported a significant correlation at the 95% confidence level between the importance level of accessibility DQ dimension and the appropriate-amount,

believability, completeness, concise-representation, consistent-representation, free-of-error, interpretability, relevance, reputation, security, timeliness and easily-understandable DQ dimensions [76].

Table 2.5: *Correlation between the importance level of the DQ dimensions*

	AC	OBJ	REP	COM	APM	VAD	REL	TIM	ACT	INT	EU	RC	CR	AL	ACC	SEC	TRA
AC	1																
	0																
OBJ	0.252446	1															
	0.044169	0															
REP	0.2605	0.346904	1														
	0.03762	0.00498	0														
COM	0.228264	0.003598	0.072626	1													
	0.069655	0.977491	0.568466	0													
APM	0.137191	0.311675	0.480231	0.368127	1												
	0.279689	0.012175	5.92E-05	0.002764	0												
VAD	0.491512	0.220929	0.427274	0.096249	0.247091	1											
	3.72E-05	0.079371	0.000429	0.449305	0.049022	0											
REL	0.198114	0.224751	0.449875	0.241511	0.395312	0.467646	1										
	0.116584	0.074182	0.000192	0.054534	0.001225	9.77E-05	0										
TIM	0.228709	0.174024	0.117943	0.587841	0.372854	0.376575	0.470993	1									
	0.069097	0.169046	0.35331	3.27E-07	0.002411	0.002162	8.57E-05	0									
ACT	0.432112	0.107858	0.383201	0.254649	0.353605	0.636142	0.647521	0.410077	1								
	0.000363	0.396251	0.001775	0.042291	0.004153	1.62E-08	7.36E-09	0.000765	0								
INT	0.244413	0.230851	0.479009	0.439484	0.4379	0.353463	0.555655	0.441587	0.438121	1							
	0.051607	0.066463	6.22E-05	0.00028	0.000296	0.004169	1.88E-06	0.000259	0.000294	0							
EU	0.194107	0.136513	0.377356	0.300995	0.430025	0.362254	0.500757	0.371792	0.557832	0.754115	1						
	0.124316	0.282095	0.002113	0.015656	0.000391	0.003266	2.51E-05	0.002487	1.68E-06	6.31E-13	0						
RC	0.452627	0.191273	0.196586	0.463533	0.334	0.414019	0.26989	0.3565	0.426922	0.614232	0.53583	1					
	0.000173	0.130014	0.119488	0.000115	0.006991	0.000672	0.031024	0.003835	0.000435	6.73E-08	5.05E-06	0					
CR	0.325934	0.334841	0.48522	0.274208	0.404152	0.506259	0.380893	0.316399	0.465907	0.633552	0.56711	0.594237	1				
	0.008582	0.0068441	4.83E-050	.028333	0.000926	1.97E-05	0.001903	0.10862	0.000105	1.92E-08	1.03E-06	2.26E-07	0				
AL	0.406922	0.324045	0.339646	0.430396	0.331717	0.485235	0.232156	0.271109	0.332455	0.388638	0.287846	0.520336	0.570255	1			
	0.000847	0.008997	0.006037	0.000385	0.007413	4.83E-05	0.064899	0.030244	0.007274	0.001506	0.021087	1.05E-05	8.7E-07	0			
ACC	0.199658	0.155104	0.353577	0.355718	0.253552	0.354263	0.556064	0.360933	0.62701	0.466148	0.464598	0.377482	0.419347	0.472153	1		
	0.113705	0.221029	0.004156	0.003918	0.043217	0.004079	1.84E-06	0.003389	2.97E08	0.000104	0.00011	0.002105	0.000562	8.19E-05	0		
SEC	0.217996	0.39783	0.454412	0.191851	0.367704	0.167309	0.410671	0.209936	0.206063	0.530728	0.361428	0.257688	0.456047	0.259578	0.298892	1	
	0.083546	0.001132	0.000162	0.128836	0.002798	0.186358	0.00075	0.095913	0.102344	6.45E-06	0.003343	0.039809	0.000152	0.038326	0.016434	0	
TRA	0.1956	0.475829	0.40956	0.405492	0.326457	0.498089	0.513802	0.433849	0.391099	0.531129	0.415167	0.454955	0.701534	0.517412	0.4821	0.436286	1
	0.12139	7.07E-05	0.000778	0.000887	0.00847	2.81E-05	1.41E-05	0.000342	0.001396	6.32E-06	0.000647	0.000159	1.07E-10	1.2E-05	5.49E-05	0.000314	0

ρ and p-values are indicated in the first and second row of each DQ dimension respectively.

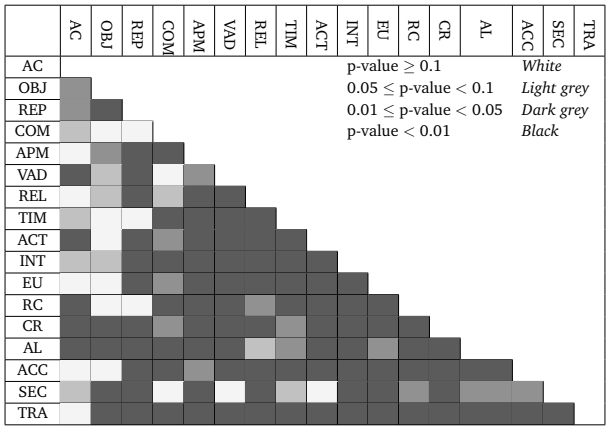


Table 2.6: *The results of Spearman's rank correlation test which show the significance of correlation between the importance level of DQ dimensions*

Although there is a trade-off between the improvement of the DQ dimensions, their importance level is positively correlated.

2.5.2 Aim 2: Scorecard index

The second research aim of the study is investigated by analyzing the second part of the survey where the DQ level of credit risk databases is assessed using the DQ dimensions in Table 2.2. For this analysis, we aggregated the DQ dimensions into Wang and Strong's DQ categories [149]. The value of each DQ category is computed as the weighted average of the values of its constituing DQ dimensions using their degrees of importance. The degree of importance of each DQ dimension is assessed in the first phase of the study. The weights (w_j) which indicate the degrees of importance are computed using the average ranks (AR_j) in Fig. 2.5. A simple-average model was also investigated. However, the equal weight of 0.25 for each of the four DQ categories is different from the range of weights (0.183-0.366) computed. Hence, we used only the weighted average model in our analysis. The weighted average distribution of the DQ level for each DQ category for each financial institution is given in Fig. 2.9. In line with the earlier introduced notation, let d_{ij} be the score attributed by the i-th financial institutions to the j-th DQ dimension. Then in the rest of the analysis, we use $\bar{x}_i$, $\bar{x}$ and s to indicate the weighted average for individual financial institution i, the

sector weighted average and the sector standard deviation respectively. We calculate $\bar{x}_i$, $\bar{x}$ and *s* for each of the four DQ categories. The weight for each DQ dimension is computed as:

$$w_j = \frac{AR_j}{\sum_{j=1}^{P} AR_j}$$

where w_j is the weight of each DQ dimension as per its degree of importance for assessing the DQ level. AR_j is the average rank of each DQ dimension as shown in the Bonferroni-Dunn results in Fig. 2.5. The weighted average is computed as:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\sum_{j=1}^{P} w_j d_{ij}}{\sum_{j=1}^{P} w_j} \right]$$

where $\bar{x}$ is the sector weighted average of the DQ level for each DQ category and d_{ij} is the DQ level score of each DQ dimension from the second part of the survey under each DQ category for each financial institution. The summation across the DQ dimensions is particular for each DQ category. Hence, the summation includes only DQ dimensions under a specific DQ category.

The DQ level distribution (Fig. 2.9) is used to indicate the performance of the sector for the four DQ categories. Therefore, best practices and areas for improvement can easily be identified. Consequently, it helps an individual financial institution to focus improvement activities. The four categories can be compared to detect common patterns or to focus on the category that most needs to be improved.

A common concern in organizations is how well they are doing relatively to others in the sector. The scorecard index addresses this concern. It is defined as a managerial system that can motivate breakthrough improvements by indicating critical areas such as product, process, customer, and market development [45]. Also, it is a measurement of products, services, or practices against tough competitors, industry leaders, or other sources or best practices. These best practices form the benchmark against which performance is measured. The scorecard index is used to benchmark the DQ level of an individual financial institution.

The distribution in Fig. 2.9 provides a method to establish the state of DQ benchmarks. Hence, financial institutions can assess their DQ level using the best practice institutions in the sector. To identify the best practice institutions, we defined four limits in the distributions. These are above

(a) $\bar{x}$ distribution for *intrinsic*

(b) $\bar{x}$ distribution for *contextual*

(c) $\bar{x}$ distribution for *representation*

(d) $\bar{x}$ distribution for *accessibility*

Figure 2.9: *DQ levels in financial institutions for the four DQ categories*

	DQ Dimension	Mean	AR_j	w_j	d_{fj}
Intrinsic	Accuracy(AC)	9.08	12.10	0.079	*3*
	Objectivity(OBJ)	8.19	9.61	0.063	*2*
	Reputability(REP)	7.89	8.02	0.052	*9*
Contextual	Completeness(COM)	8.02	8.78	0.057	*8*
	Appropriate-amount(APM)	7.84	7.71	0.050	*7*
	Value-added(VAD)	8.27	9.40	0.061	*4*
	Relevancy(REL)	8.52	10.39	0.068	*7*
	Timeliness(TIM)	8.28	9.45	0.062	*3*
	Actionable(ACT)	8.53	10.37	0.068	*1*
Representation	Interpretability(INT)	7.86	8.13	0.053	*8*
	Easily-understandable(EU)	7.81	7.73	0.051	*8*
	Representational-consistent(RC)	8.13	9.21	0.060	*3*
	Concisely-Represented(CR)	7.36	6.52	0.043	*3*
	Alignment(AL)	7.75	7.51	0.049	*3*
Access	Security(SEC)	8.47	10.45	0.068	*3*
	Accessibility(ACC)	8.41	9.74	0.064	*3*
	Traceability(TRA)	7.73	7.88	0.051	*4*

Table 2.7: *The columns indicate the mean, the average rank (AR_j) and the weight (w_j) from the first part of the study and the DQ level assessment scores (d_{fj}) of one fictitious financial institution (f) from the second part of the study for each DQ dimension.*

upper limit ($\bar{x}_i > (\bar{x} + s)$), between the upper limit and the sector weighted average ($\bar{x} < \bar{x}_i < (\bar{x} + s)$), between the sector weighted average and the lower limit ($(\bar{x} - s) < \bar{x}_i < \bar{x}$) and less than the lower limit ($\bar{x}_i < (\bar{x} - s)$). If the weighted average for an individual financial institution ($\bar{x}_i$) falls above the upper limit, between the upper limit and the sector weighted average, between the sector weighted average and the lower limit, and below the lower limit for the specific DQ category, then the DQ level is assessed to be very good, good, below average and worst respectively. In general, if the DQ level is assessed to be below average, the institution is in a poor DQ state. Therefore, an improvement action should be taken. This is illustrated by an example in the following section.

Scorecard index illustration

The scorecard index is illustrated using a fictitious financial institution. The mean, the average rank (AR_j) and the weight (w_j) from the first part of the study, and the DQ level assessment scores from the second part of the study for each of the DQ dimensions are given in Table 2.7. The weighted average for each DQ category is computed using the (w_j) and the score, of which the results are given in Table 2.8. The colors of the cells indicate the DQ level of each DQ category. The black, dark grey, light grey and white cells indicate very good, good, below average and worst DQ levels respectively. The DQ level for the three DQ categories (intrinsic, contextual and access)

are worst and the DQ level for representation DQ category is below average for the financial institution. This scorecard allows financial institutions to directly locate potential areas for improvement and guide their improvement actions.

Limits	Color index	DQ scorecard for XYZ	
$\bar{x}_i > (\bar{x} + s)$		Intrinsic	Contextual
$\bar{x} < \bar{x}_i < (\bar{x} + s)$		$\bar{x}_i$ = 4.28	$\bar{x}_i$ = 4.86
$(\bar{x} - s) < \bar{x}_i < \bar{x}$		Representation	Access
$\bar{x}_i < (\bar{x} - s)$		$\bar{x}_i$ = 5.03	$\bar{x}_i$ = 3.28

Table 2.8: *Scorecard index for one fictitious financial institution's DQ level for each DQ category, where $\bar{x}_i$ is the weighted average for the DQ level of the institution for each DQ category, $\bar{x}$ is the sector weighted average and* s *is the sector standard deviation.*

2.5.3 Aims 3 & 4: DQ issues for credit risk management

In this section, the third and fourth research aims of the study are discussed based on the third part of the survey. Note that among 64 financial institutions, only 37 participated in this third part of the survey. The 37 institutions are in fact a subset of the 64 institutions that participated in the first session of the survey. We verified using a Friedman test if there were significant institutional (size and geographical area) and background (education level and experience) differences between the 37 and 27 institutions and subjects respectively at α = 5%. The results indicated no significant statistical differences and therefore, the results of this section can also be considered as equally valid as sections 2.5.1 & 2.5.2.

2.5.3.1 Different DQ problems and their causes

In this third part, the respondents were asked to indicate the major DQ challenges or problems that they encounter on a daily basis. The results are depicted in Fig. 2.10a. 63% of the respondents indicated that inconsistency (value and format) and diversity of data sources are main recurring DQ challenges. This indicates that there are many similar data which are kept in different files. Since these data may not be updated or changed at the same

time, it is very likely that the data can differ to each other. As a result, decision makers either must rely on their own DQ assessment in order to choose the data source most suited for their decision tasks or must reconcile the different data sources to get one reliable data source. However, we can infer from the results that both processes are not easy. In line with the results from Fig. 2.10a, Cappiello et al. [15] indicated that mismatches among sources of the same data are a common cause of intrinsic DQ concerns. They identified in their study that mismatches among sources of the same data encourage a subjective DQ assessment by decision makers which gradually affects the intrinsic or objective DQ dimensions. Initially, data consumers do not know the source to which DQ problems should be attributed; they only know that data is conflicting. These concerns initially appear as *believability* problems. Over time, data users assess the *accuracy* of the data for the sources based on experience and personal preferences, which leads to a poor *reputation* for sources considered inaccurate. Hence, less reputable sources are viewed as having little *added value* for the task, resulting in reduced use [15, 138]. However, these less reputable data sources may be of high quality.

	Process bringing data from outside	Processes changing data within	Processes causing data decay
Inconsistent Data representation	Large	Medium	Small
Inconsistent copies of data	Large	Medium	Small
Data collection and its costs	Large	Medium	Small
Diversity of data sources	Large	Medium	Small
Making use of the available data	Small	Medium	Large

Table 2.9: *Cause-effect relationship between DQ problems and different data processes [76].*

In addition to the inconsistency and diversity of data sources, the results in Fig. 2.10a show that data collection problems and the high costs associated with them are recurring DQ challenges. Data are often produced or maintained by different departments and by different data producers. However, these data are typically also needed by other departments which are not responsible for the production and maintenance of it. Although cross departmental data access is typically facilitated by enterprise wide information systems, collecting all the necessary data is still found to be a common challenge that consumes an important share of the decision makers' time. Another reported

DQ related problem in the results of Fig. 2.10a are difficulties when making use of the available data. This is related to the relevancy and timeliness DQ dimensions. Decision makers will discard irrelevant data as it has no added value in a particular context; they might also opt to not use outdated data. Unfortunately, in many cases, assessing whether data is relevant and/or timely will again consume a fair amount of the decision makers' valuable time.

Data processes as causes of DQ problems

The impact of different data related processes on DQ has been assessed earlier [87, 75]. The third part of the survey further investigates these processes and quantifies their impact on DQ in financial institutions. The results hereof are depicted in Fig. 2.11. These results indicate that though with different degree, all data related processes have caused DQ problems.

Predominantly, manual data entry processes are confirmed to be a major DQ problem cause. This indicates that despite high automation, a lot of data is manually entered into databases inducing a heightened risk of faults. One example could be mixing up the age of two customers or not entering any data at all, resulting in inconsistent data. This can create a DQ problem which can not easily be identified or explained. These different human manual data entry process problems however can be mitigated by well-designed data entry processes and accompanying instructions [87].

System consolidation and initial data conversion are also confirmed to cause database impurity. The main common problem in system consolidation is data duplication. Previous research also acknowledges that the data in the consolidated systems often overlap [6]. Similarly, when data are transferred from previous/old systems or paper documents to a new system, data may be lost in the process. This is exacerbated by the fact that there is typically no well recorded metadata [6, 87, 75]. In addition to the above identified causes of DQ problems, data mutations taking place internally without being captured by the system and loss of expertise are also indicated as common DQ problem causes as shown in the results in Fig. 2.11. The changes are known only by those who made the changes and whenever those employees leave, these changes may get lost. This clearly indicates that much information that is essential for appropriate use of the data exists as tacit knowledge, rather than in metadata format. Though very rarely, the respondents also admitted

that processes meant to clean impure data in fact caused DQ problems. Wang and Strong [149] reported that every database has impurity, thus trying to fix one problem may create another one. This finding warns that in order to ensure DQ, the effects of all data related processes need to be taken into account as well.

The data processes in Fig. 2.11 can be catalogued as processes bringing data from the outside, those changing data within the company, and processes causing data decay, see also Fig. 2.3 [87]. In Table 2.9, we have summarized the cause-effect relationship by mapping the major DQ problems reported in Fig. 2.10a with the different data processes reported as the major causes of the problems in Fig. 2.11. Initial data conversion, system consolidation, manual data entry, batch feeds and real-time interfaces are processes bringing data from the outside into databases and are confirmed to be major causes for most of the DQ problems reported in a context of credit risk management. Consequently, many DQ problems such as inaccuracy, incompleteness and inconsistency can be traced back to these processes. For example, during initial data conversion, data may not enter into new databases simply because the new databases are not prepared to accept those data. Similarly, the person who manually enters data can make different mistakes such as entering the wrong data and leaving the cell/column empty where there is supposed to be a value. Likewise, DQ can be impacted by data processing, data cleansing and data purging which are processes changing data inside databases. If there is a bug in the program responsible for data processing, it can create different DQ problems such as inaccuracy, inconsistency and incompleteness. Similarly, as there are always DQ problems in databases, data cleansing and purging may impact DQ. For instance, the correct data which exist in databases can accidentally be cleaned or purged because they fit the cleansing or purging criteria [75]. Changes not captured, system upgrades, new data uses, loss of expertise and process automation are processes which can cause data decay. Sometimes physical changes which happened in organizations may not be recorded into the systems. For example, a married employee may be known as a single. In organizations, there are lots of daily changes. New production methods can be created or new ways of sales are proposed but the data previously collected may not be useful for these new tasks. Similarly, computer programs take the data literally and cannot make a proper judgment about the likelihood of it been correct. Some validation screens may be implemented in the automated processes, but these will often fail to see all data peculiarities, or are turned off in the interest of

performance. As a result, automation can cause data decay.

2.5.3.2 Magnitude of DQ problems

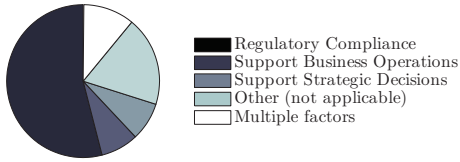
In order to properly manage DQ, one should know the challenges and the causes of DQ problems. However, an important step is the ability to measure the DQ of the data stores [76]. In the third part of the survey, the respondents were also asked to indicate the magnitude of DQ problems. The results in Fig. 2.12 depict the observed magnitude of poor DQ. More than 10% of the data in credit risk management databases are estimated to be of poor quality. The majority of the institutions estimated that between 10-20% of the data is subject to errors. However, 19% of the questioned institutions are unaware of the magnitude of their DQ problems. This result indicates that most financial institutions are still unable to develop comprehensive measures and are unable to assess the magnitude of DQ problems. As a consequence, the impact of the existing poor DQ on the decision tasks is hard to assess as well. Yet, it is clear that addressing the reported 10-20% of DQ problems may take more than 50% of an employee's time [87]

In addition, the inherent difficulty of accurately measuring DQ might discourage any initiative to improve it. This is confirmed by the results shown in Fig. 2.10b which indicate that regulatory requirements (e.g. Basel I and II) are cited as the main reason of many DQ enhancing projects. The Basel Accord requires the calculation of detailed loss modeling factors to determine the capital requirement as explained earlier. Accurate quantitative modeling of PD, LGD, EaD and M is not only required by this regulation but can become a competitive advantage leading to superior credit performance [4]. However, a competitive advantage is considered as less important to initiate DQ enhancing activities. Because of these regulatory compliance requirements, financial institutions are organizing DQ teams to improve DQ and cross-functional efforts to improve the comparability and applicability of data sources across different business units. However, such efforts are reported to be not mature enough yet.

Generally, the above explained key findings show that although poor DQ appears to be the norm, rather than the exception, DQ is not given much attention in financial institutions.

Every data related process has an impact on the quality of data. However, the DQ problem(s) that can be caused by one data related process may be very

(a) Major data quality issues in financial institutions



(b) Major Data quality initiative motivations in financial institutions

Figure 2.10: *The major DQ problems and reasons for improvement actions*

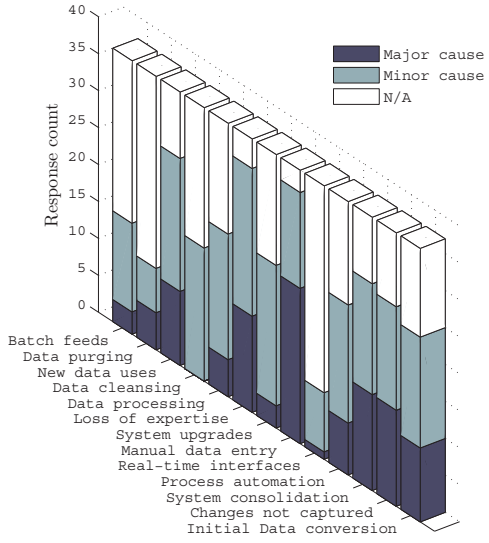


Figure 2.11: *Different causes of DQ problems in financial institutions*

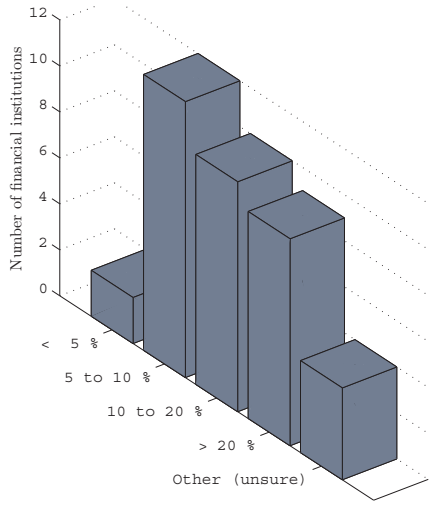


Figure 2.12: *The % of the data estimated to be of poor quality in credit risk databases as assessed by data users in financial institutions*

different to the ones induced by other processes. This statement also holds for processes meant to improve data quality.

2.6 Limitations

As every application may have its own DQ requirements, the results of this paper may not necessarily reflect the generalized views of organizations outside the scope of financial institutions. In addition, if a larger sample can be collected, the results of this study may be different to some extent. Furthermore, as organizations are trying to improve their DQ level every day, the results in this paper only reflect the DQ requirements of financial institutions in the study period. However, the methods used by this paper can be repeated to analyze the data quality requirements of financial institutions and other sectors in different time periods.

2.7 Conclusion and Future Research

This paper explored the important DQ dimensions and assessed the self reported DQ level using a scorecard index. Also, the paper identified different DQ challenges and their possible causes. In general, the paper demonstrated a TDQM effort in a financial setting. In the definition phase, the identification of various DQ dimensions relevant to credit risk assessment is considered. Similarly, in the measurement phase, the DQ level in credit risk databases is assessed and DQ issues are analyzed. The results of the analysis help to identify the problem areas and to focus improvement actions, completing the TDQM cycle.

We started with a literature overview of the different DQ dimensions, focussing on the framework of Wang and Strong [149]. Based on the results of the pilot survey, this framework was extended with three additional DQ dimensions (i.e. 'alignment', 'actionability' and 'traceability'), totalling seventeen DQ dimensions. The importance of this extended framework has been assessed by credit risk managers. These decision makers rated the DQ dimensions on a scale from 0-10. The results were analyzed using a Friedman test which indicated a significant difference among the scores of the DQ dimensions. The results of the post-hoc Bonferroni-Dunn test confirmed that accuracy is the most important DQ dimension. Also security, relevancy, actionability, accessibility, objectivity, timeliness, value-added and representational-consistency are found to be important DQ dimensions. The Wilcoxon ranked sum tests confirmed that the most important DQ dimensions identified are valid, irrespective of the size of financial institutions and the presence of DQ teams.

A Bonferroni-Dunn test was also performed to other sectors' data. The results indicated that there is a difference between financial and other sectors in assessing the importance of DQ dimensions. This result also confirmed the contextual behavior of DQ.

The correlation between the importance level of DQ dimensions has also been assessed and it was found that the majority of DQ dimensions are correlated, implying that DQ, although intrinsically a multidimensional concept, is often perceived from a single perspective.

Second, the DQ levels in credit risk databases are assessed using the weighted average model. The distributions of the weighted average of each DQ category

have been used to benchmark the DQ level as very good, good, below average and worst. The scorecard index is used to assess the DQ level and to indicate the problem areas.

Finally, the paper identified different DQ challenges and their causes in financial institutions. The results indicated that inconsistency and diversity of data sources are among the most recurring challenges. Likewise, manual data entry processes are found to cause the majority of the DQ problems. Although DQ problems are endangering the effectiveness of the task, only little DQ enhancement activities are currently in place. Moreover, these activities are mostly instigated by regulatory authorities, rather than by internal considerations. Surprisingly, creating a competitive advantage was not found to be an important stimulus in any DQ improving activity.

It is confirmed in this paper that the majority of financial institutions are unaware of the magnitude of their DQ problems, this refraining them from taking holistic measures to tackle these issues. This is a clear indication of the need for comprehensive DQ metrics.

Although DQ is contextual and should be addressed with respect to the task at hand, it has also intrinsic characteristics which can be valuable to other tasks. As such, since credit risk assessment involves mostly analytical tasks, it is believed that the DQ requirements and findings of this study can be extended towards different tasks and organizations of similar nature. The empirical validation of this conjecture is considered to be an interesting topic for future research.

Although this study confirms our assumption that DQ is contextual by indicating only nine out of seventeen DQ dimensions as very important to fulfill the DQ requirements of credit risk management task, the sensitivity analysis on the parameters (PD, LGD, EaD and M) in order to understand the possible impact of DQ on risk concentration as well as the relative importance of individual DQ dimensions on these parameters are both considered to be interesting topics for future research.

3

Factors Determining The Use of Data Quality Metadata (DQM) for Decision Making Processes

3.1 abstract

Decision making processes and their outcomes can be affected by a number of factors, among them, the quality of the data is critical. Poor quality data causes poor decisions. Although this fact is widely known, data quality (DQ) is still a critical issue in organizations because of the huge data volumes available in their systems. Therefore, literature suggests that communicating the DQ level of a specific data set to decision makers in the form of DQ metadata (DQM) is essential. However, the presence of DQM may overload or demand cognitive resources beyond decision makers' capacities, which can adversely impact the decision outcomes. To address this issue, we have conducted an experiment to explore the impact of DQM on decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. Findings of a statistical analysis suggest that the use of DQM can be enhanced by data quality training or education. Decision makers who have a prior data quality knowledge used DQM more than those who do not have a prior DQ knowledge to solve the decision task and achieved a

higher decision accuracy. However, the efficiency of decision makers suffers when DQM is used. Our suggestion would be that DQM can have a positive impact on decision outcomes if it is associated with some characteristics of decision makers, such as a high data quality knowledge. However, the results do not confirm that DQM should be included in data warehouses as a general business practice, instead organizations should first investigate the use and impact of DQM in their setting before maintaining DQM in data warehouses.

3.2 Introduction

Although the importance of data quality has been recognized for more than decades, different DQ problems continue to exist even in simple traditional systems because of huge data volumes and their complexity [19]. The problem is exacerbated by the fact that decision support systems are becoming vital to support decision making processes. The DQ level in decision support systems may not be good for different reasons. One reason is that DQ problems can be aggravated when data are merged or integrated from different sources which is typically the case in decision support systems or data warehouses. The other reason can be that soft data analysis is needed for strategic planning. Soft data is a subjective assessment or a future trend forecast which can be used for decision making [36]. For example, decision makers need to utilize soft data, such as the marketing strategies of competitors in order to change or adapt the marketing strategy of the company accordingly. Most of the time, managers make decisions without considering the DQ level of the data. Decision makers who are familiar with the data have an intuitive knowledge about the data. However, this intuitive knowledge can be missed when data are used by different decision makers for purposes other than the original purpose for which the data were created, which is becoming more and more the case with the increasing use of data warehouses. Decision makers who do not have prior experience with the data may avoid using the data because they can not verify the quality of the data. Because of such and other reasons, DQ is very important for decision making processes. However, organizational data warehouses are still facing different DQ problems.

As one of the different ideas to reduce the impact of poor DQ on decision outcomes, the literature suggests the inclusion of metadata about the quality of data (DQM) [19, 36, 115, 132]. DQ is context-dependent, meaning that

data with quality for one use may not be appropriate for other uses. For instance, the extent to which data is required to be complete for accounting tasks may not be required for sales prediction tasks. Therefore, DQM can help decision makers to determine the appropriateness of the DQ level in the context of the task at hand [149, 36, 128, 92]. Additionally, DQ practitioners have acknowledged the importance of providing DQM to facilitate the decision making process [86, 32]. However, the presence of DQM may overload or demand cognitive resources beyond decision makers' capacities, which can adversely impact the decision outcomes.

Maintaining DQM in databases means maintaining the level of DQ measured along DQ dimensions such as accuracy, completeness and timeliness. Capturing and maintaining DQM in databases may induce equal or greater costs to any data capturing and maintaining process. Capturing DQM requirements is a difficult process and requires investment in software tools and/or human resources [128]. Likewise, the process of maintaining DQM requires additional storage space. Yet, as there are no standard models for evaluating the benefits of metadata in general [35], it is difficult for organizations to evaluate the benefits of DQM and justify its maintenance costs [128]. Therefore, prior DQ researchers assessed the benefits of DQM in terms of its impact on decision making and outcomes. Hence, a set of decision outcome measures have been proposed such as decision accuracy, consistency, consensus, complacency, confidence and efficiency [19, 36, 129, 116]. However, there is no full agreement on the results of prior studies. Some researchers have found that DQM is used in certain situations [36], and others did not find any statistical evidence that DQM is actually used, even when it is available [116]. In addition, the impact of DQM on the effectiveness of decision outcomes has not been studied conclusively. To fill this gap, this paper assesses the benefits of DQM for decision making purposes by comparing the decision complacency, accuracy, confidence and efficiency of decision outcomes with and without DQM using a different decision task and analysis techniques from previous researches. We have developed a critical decision task (bankruptcy prediction) based on an Altman-Z model [2] to understand the impact of DQM on decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. When deciding on actual DQ initiatives, the costs of acquiring and maintaining DQM, which can be assumed by project costs and resource requirements, can be compared to DQM's assessed benefits. The latter is, however, beyond the scope of this paper, because the benefits and

costs of DQM can be different among organizations depending on the size, business directions and the degree of necessity of decision support systems [116].

The paper is structured as follows. The next section reviews previous research in DQM. The third section discusses the research design and the fourth section explains the results. Finally, the paper ends by giving concluding remarks and indicating future research ideas.

3.3 Literature Review

3.3.1 Data Quality

Recently, data quality (DQ) is becoming a concern to organizations where plenty of data are available. Similarly, DQ is constantly growing as a crucial research topic in academic world. DQ research can be categorized into two broad types, *intrinsic* and *contextual* DQ studies. The *intrinsic* DQ research concerns about the intrinsic value of the data. It depends on the data themselves without considering the context in which the data is used. Plenty of research has been conducted to measure the correctness, the completeness and consistency of data independent of the context. Therefore, the main deliverables are different techniques to improve DQ, such as data cleansing [56], statistical process control [122], data source calculus [103], data stewardship [32], and dimensional gap analysis [65]. These techniques treat the data intrinsically and usually do not consider *contextual factors* such as the purposes for which the data is used and the characteristics of the data users. However, prior research has indicated that these contextual factors can strongly affect the way DQ is assessed for daily use. This has led to the well recognized DQ definition of "fitness for use". In addition, Wang and Strong [149] have indicated the importance of recognizing the multi-dimensionality nature of DQ and measure data items accordingly using users' perceptions. They identified important DQ dimensions by considering users' perceptions such as accuracy, relevancy, representation and accessibility. Accuracy indicates the extent to which data are certified, reliable and error-free. This DQ dimension is also a dimension which intrinsic DQ assessment usually considers. Relevancy reflects the extent to which data are applicable and appropriate for the task at hand. Representation describes the extent to which data is

compactly represented, well-presented, well-organized and well-formatted, and accessibility indicates the extent to which data is available, or easily and swiftly retrievable. Reliability, believability, currency, and completeness are also identified as important in the literature [149, 38, 92]. Generally, the importance of considering contextual DQ assessment increases the complexity level of DQ management. In intrinsic assessment, DQ measurement focuses solely on the data item, therefore, the aim is easy to define and measure. For example, the output of an intrinsic DQ measurement can be that the data is 90.05% error-free. However, as discussed before, intrinsic measurements do not show the full picture of the relevancy and appropriateness of the data because they do not consider the context in which the data are used or the characteristics of data users. For example, consider a production company sales sheet which shows "item codes", "quantities", "cost" and "selling prices" where some values for the "cost" column are missing. For decisions regarding production efficiency, the sheet with missing "cost" data would be considered incomplete. However, the same sales sheet can be considered as complete for making inventory decisions (reconciling the amount of quantities on the sheet and the physical quantities in a store) because all the values for the "quantities" column are present. This indicates the importance of considering the contextual nature of DQ in order to improve DQ management in databases. Therefore, it is important that decision makers can determine the level of DQ for the task at hand. This is also one of the reasons why recent DQ research has suggested the integration of DQM along with the data in decision support systems [116].

3.3.2 Data quality metadata (DQM)

As data are often created, managed and maintained by sophisticated information systems, decision makers often use data which they do not produce. As a result, knowledge that would be needed to assess the relevance and appropriateness of the data has been lost [36]. Consequently, decision makers are left to take the DQ level of the available data for granted. In turn, the quality of decision outcomes is negatively affected. Including data quality metadata (DQM) in databases is believed to provide this missing information [19]. Data quality metadata (DQM) is information about the quality level of stored data in organization databases, and is measured along different dimensions such as accuracy, currency, and completeness. Also, DQM is considered to be intrinsic to the data because the metadata is usually produced

objectively. DQ tagging is the process by which DQM is created [131]. There are different types of metadata in information systems which are maintained and managed, such as data dictionary metadata, administrative metadata, and metadata about the system infrastructure (see Table 3.1). Although many of the metadatas in Table 3.1 are maintained in databases, DQM capturing and maintaining processes are not considered as business routines as DQM's benefits are not vivid yet.

Types of metadata	Description
Data quality metadata	indicates the quality level of specific data in databases. For example, it can be indicated that sales data are 90% complete for the month of January 2014.
Descriptive metadata	indicates the purpose for which and by whom the data are created. It shows the author and title of the data.
Terms and conditions metadata	indicates the intellectual property rights.
Administrative metadata	indicates when and how the data are created, and who can access them.
Data dictionary metadata	indicates meaning and relationships of data.
Structural metadata	indicates how a system or metadata works. It indicates the hardware and software records.

Table 3.1: *Different types of metadata as discussed in literature [51, 14, 47].*

There are different issues in DQ tagging. First, there are no established rules, to the best of our knowledge, at which level DQM should be maintained in databases. It is possible to have DQM at the level of the individual data item, at the record level, at an attribute/column level and at the level of a relational table [36, 19]. However, the merits and demerits of these levels of DQM representations are not fully discussed in the literature. The most common level of DQM representation used by previous researchers is at the data item level [36, 19, 131, 115].

Second, determining the quality measure of which DQ dimension should be stored as DQM is context-dependent. In other words, the measurement of which DQ dimension is important to be stored as DQM, such as the accuracy or completeness level, should be determined by the importance level of these dimensions for the task at hand. However, the most commonly used DQ dimension in the literature is the accuracy dimension [36, 19, 131, 115]. This may emphasize the enormous interdependency between accuracy and most of the other DQ dimensions. Also, the inclusion of the accuracy dimension as DQM acknowledges the importance of this dimension for different tasks [92].

The third important consideration is the format of DQM, in particular how

DQM is created, maintained and represented to the end users. The format in which DQM is represented can affect the decision making process and should be designed to facilitate the process [115, 137]. Indeed, DQM needs to be presented in databases in the appropriate format to enhance effective decision making. There are different DQM representations used in previous DQM research. Chengalur-Smith et al. [19] considered two approaches: 0-100 interval and n-level ordinal representation. A 0-100 interval representation assumes that, for example, data with a score of 70 is better than data with a score of 60. The n-level ordinal representation categorizes the DQ level as excellent, good, average, and so forth. The n-level ordinal representation could also be mapped into a two-point scale with a value of "above average" and "below average." Fisher et al. [36] only used the interval DQM representation. Shankaranarayanan et al. [129] used a percentage DQM representation where the quality level of the data represented with an 80% accuracy or completeness level is better than the quality of data represented by a 70% level. However, the percentage representation is similar to that of an interval representation of DQM. Moges et al. [91] conducted a pilot study to evaluate DQM representations by using two different types of DQM formats. These are DQM with lower and upper value limits (range representation) and probability representation. The range DQM representation shows the minimum and maximum possible values for specific data where, for example, a specific data item can be in a range between 50-70. The probability DQM representation presents the likelihood that the value of a specific data item represents its real value. Their pilot survey indicated the understandability of the probability DQM format. However, the probability and the interval DQM representation can be considered as similar with a minor distinction. On the other hand, Even et al. [34] used a graphical representation of process metadata as information product map (IPMAP) which uses colors to describe the quality level of data, for example, the color red used to indicate the poor quality level of the data given. Although there is no standard for DQM representations, many of prior DQM researchers agreed on the understandability of an interval, a percentage and a probability DQM representation [19, 36, 128, 91]. As can be inferred from the text above, these three representations are similar. Table 3.2 summarizes the different DQM representations.

The use of DQM for decision making purposes and the impact of DQM on the outcomes of decisions are the functions of the three discussed issues in DQ tagging: the level at which DQM can be maintained, the DQ dimension

DQM formats	Description
Ordinal	It indicates whether the data quality level is above average or below average, or it categorizes the DQ level as excellent, good, average and so forth.
Interval	It represents the DQ level using a 0-100 interval scale, where a higher DQ level indicates higher accuracy.
Probability	It represents the DQ level using a 0-1 probability scale which indicates the chances that the data are correct.
Range	It gives the lower and the upper limit where a specific data set can be.
Graphical	It uses colors to indicate the DQ level of a specific data set.

Table 3.2: *Different DQM formats explored in literature*

and the DQM representation format. Therefore, the use of DQM should be investigated in consideration of those three elements. Although providing DQ metadata (DQM) along with the actual data set is considered to regain or complete the intuitive knowledge that is lost, it is important to identify whether decision makers are not complacent to the DQM (see Table 3.3). In addition, it is important to identify the impact of DQM on the outcomes of decision making. Including DQM in databases together with the actual data would be beneficial only if DQM is used and improves the decision outcome. As a result, many information systems researchers have responded to this need.

Chengalur-Smith et al. [19] investigated the use of DQM by using two DQM formats (ordinal and interval) and two decision strategies (conjunctive and weighted additive) (see Section 3.3.3.1). Their results indicated that when an ordinal DQM format was implemented, complacency [1] was accepted for the conjunctive decision strategy but rejected for the weighted additive strategy for both simple and complex tasks. However, when the interval DQM format was implemented, complacency was rejected for both decision strategies for the simple decision task. Yet, complacency couldn't be rejected for both decision strategies for the complex task scenario which was further explained by the interaction effect of task complexity and DQM formats. The interval DQM representation gives detailed information about DQM unlike the ordinal DQM representation. Therefore, it is reasonable to assume that the interval DQM format has a potential to increase the level of task complexity. This phenomena is also known as *information overload*. Information overload happens when the information given for solving a specific task is too much to

[1] Complacency means the extent to which decision makers ignore DQM [19]

be used by decision makers [12]. Maintaining too detailed DQM in databases may have a negative impact because it adds one level of complexity to the task at hand. Finally, their results showed that there is an interaction between DQM format, decision strategy and task complexity.

Fisher et al. [36], on the other hand, investigated how the experience of the decision maker, the available decision time, the time pressure and task complexity influence the use of DQM in decision making using an interval scale DQM representation. Their results indicated that complacency was rejected for experienced decision makers but couldn't be rejected for novices. However, these researchers haven not taken different types of decision making strategies into account (see Table 3.4).

Price et al. [115], investigated the use of DQM for two kinds of decision making strategies, namely Weighted Additive (WA) and Elimination By Attributes (EBA) (see Section 3.3.3.1 and Table 3.4). They found that complacency couldn't be rejected for both decision strategies. However, decision time has been significantly extended.

In fact, Chengalur-Smith et al. [19] reported DQM use for the WA decision strategy when an ordinal DQM is implemented for both simple and complex task scenarios. Yet, they reported DQM use for both decision strategies (weighted additive and conjunctive) when an interval DQM is implemented for only the simple decision task. On the other hand, Shanks and Tansley [132] reported DQM use for an EBA decision strategy when an interval DQM is implemented. However, Price et al. [116] couldn't reject complacency for both EBA and WA decision strategies.

Although the above studies agreed on the use of DQM in some circumstances, the circumstances are not similar. In addition, though many researchers have investigated the complacency of decision makers about DQM associated with many variables such as decision strategy, decision makers' experience and task complexity, to the best of our knowledge, with the exception of the study by Shankaranarayanan et al. [129], research on the impact of DQM on the effectiveness of decision outcomes lags behind. To fill this gap, this study analyzes the impact of DQM on the effectiveness of decision outcomes along with different variables such as the level of education, experience, DQ awareness, different decision making strategies and task complexity.

Shankaranarayanan et al. [129] investigated the impact of DQM on the accuracy of the decision outcomes. They investigated the impact of DQM on

decision outcomes for two variables namely the level of work experience and decision task complexity using undergraduate and graduate MBA (Master of Business Administrations) students. They measured the impact of DQM on decision outcomes using two measures such as decision accuracy and time. Their results indicated that DQM increases task complexity and decreases decision performance when it is provided along with the complex decision task. In addition, the results indicated that DQM increased the decision accuracy of experienced users yet decreased their decision efficiency.

This study is different from the Shankaranarayanan et al. [129] study with respect to the variables, the decision task, the decision outcome measures and the analysis techniques. In addition to task complexity and experience levels, out study investigates the impact of DQM in relation with decision strategy, data quality awareness, domain experience and educational level. Likewise, the study uses a bankruptcy prediction task unlike the Shankaranarayanan et al. [129] study which used an MP3 player and digital camera purchasing tasks in order to understand the influence of the application nature on the degree to which DQM is used and benefits the decision outcomes. The nature of the application domain is indicated as one of the factors which may influence the use and benefits of DQM [116]. For example, subjects of the experiment may be more concerned to avoid the serious consequences of using poor quality data to solve the bankruptcy prediction task than an MP3 player or digital camera purchasing task. Moreover, this study incorporates other decision outcome measures such as decision confidence and complacency in addition to decision accuracy and efficiency to assess the benefits of DQM (see Section 2.2, paragraph 11). Finally, the manuscript used a tree based algorithm to analyze the benefits of DQM for decision making processes unlike prior studies [19, 129, 36, 116, 93].

Mostly, prior DQM studies used an attribute level of DQ tagging, an interval DQM representation, two types of task complexity (simple and complex) and two types of decision making strategies in order to investigate the use of DQM for decision making purposes [36, 19, 132, 115, 90]. Additionally, these researchers define DQM usage in terms of the change in the preferred decision choice or the inclusion of a specific attribute in the decision processes.

This paper implements two levels of task complexity and an attribute DQ tagging as prior studies in order to enhance comparison with the results of the above studies. Furthermore, the paper contributes by proposing a different method which is used to identify whether decision makers incorporated DQM

Decision outcome assessment	Description
Complacency	Measures whether or not decision makers used the new variable, in this case the DQM, in their decision making process. If the decision outcome is similar for decision processes with and without DQM, then the decision maker is complacent to DQM.
Efficiency	The time used by decision makers to accomplish a specific decision task.
Accuracy	Measures the accuracy of the decision outcomes.
Confidence	Measures the extent to which the decision confidence of decision makers is affected with and without DQM. It can be defined as the confidence level that decision makers have for the correctness of decision choices they made.
Consensus	Measures the extent to which decision makers agreed on their decision making outcomes with and without DQM.
Consistency	Refers to the rankings of all alternatives from the most preferred to the least preferred.

Table 3.3: *Measurements of decision outcomes discussed in DQM literature*

given along with the data. In other words, the complacency of decision makers towards the DQM is clearly known in the decision making strategies which subjects used to perform the decision task instead of a vague association of DQM usage in the change of preferred decision choices like previous studies [36, 19, 132, 115, 90]. Additionally, the paper measures the impact of DQM on the effectiveness of decision outcomes.

3.3.3 Relevant variables for the use of DQM

In this section, we will discuss the important variables considered to evaluate the impact of DQM on decision outcomes. Figure 2.4 summarizes them in a structured way.

3.3.3.1 Decision making strategy

Many researchers have investigated different decision strategies in the past decades [134]. Decision making is a process which involves solving a specific decision problem by considering all the relevant information available [70]. Simon [134] stated decision making as the process where decision makers can consider all attributes of all alternatives before selecting one. On the other hand, Nut [97] identified four different decision making processes which aid decision makers to investigate different alternatives in order to choose the

optimal one. These are analytical, judgmental, subjective and bargaining. As their name indicates, the judgmental, subjective and bargaining decision making processes depend mainly on the subjective analysis of the decision makers and their characteristics. Yet, the analytical decision making strategy allows the objective assessment of alternatives to reach to an optimal decision outcome [142]. In a structured task where a decision maker collects all relevant information, designs a decision scenario with a set of alternatives and chooses the optimal one, the analytical decision strategy is preferred because of its objectivity.

Payne et al. [109] explored three types of decision making strategies under the analytical decision making processes which are Weighted Additive (WA), Conjunctive (CON) and Elimination By Attribute (EBA). Table 3.4 describes these decision strategies in more detail. However, Figure 2.4 displayed only the EBA and WA decision strategies for the reasons that the subjects of this study have implemented only these two decision strategies to solve the decision tasks in the final experiment.

3.3.3.2 Experience

It can be reasonable to assume that experience is an important variable in decision making because it aids the decision process by incorporating life-time knowledge. Experienced decision makers can easily identify errors. They are also able to identify important aspects of a decision problem which may lead to a better decision outcome than inexperienced decision makers [107, 71].

An important consideration is the cognitive capacity limit of decision makers which might be positively affected by life-time knowledge [69]. The interaction between working-memory (WM) and long-term memory (LM) creates a cognitive capacity limit. WM stores data for a short period of time while LM stores data associated with life-time experiences. WM stores data for a while to merge with the data in the LM so that the specific decision task is conceptually represented. If the conceptual representation of the task is not enough to solve the decision problem, then WM draws data from the LM and applies logical rules to explore other conceptual representations of the decision task. This process will continue until the optimal solution for the task is found or the capacity of the WM is exceeded [107]. Prior research reported that experienced decision makers can create a well-organized con-

Decision strategies	Description
Weighted Additive (WA)	It selects the highest sum value of the products of all criteria by their corresponding values. Let $W1_{a1}$ & $V1_{a1}$ and $W2_{a1}$ & $V2_{a1}$, and $W1_{a2}$ & $V1_{a2}$ and $W2_{a2}$ & $V2_{a2}$ represent the weight and value of criteria 1 (C1) and criteria 2 (C2) for alternative 1 (A1) and alternative 2 (A2) respectively. The WA strategy would compare the value of ($W1_{a1}$ x $V1_{a1}$) + ($W2_{a1}$ x $V2_{a1}$) for A1 and ($W1_{a2}$ x $V1_{a2}$) + ($W2_{a2}$ x $V2_{a2}$) for A2 and selects the alternative with the highest value as an optimal solution of the decision task.
Elimination By Attribute (EBA)	It compares all alternatives based on the value of an attribute, the so-called deciding attribute, and then eliminates the alternatives which have lower values for the deciding attribute. The deciding attribute is usually selected based on its highest weight or decisive power for the decision task. If for example C1 is a decisive criterion because of its highest weight for a specific decision task, the EBA decision strategy would compare the values of ($W1_{a1}$ x $V1_{a1}$) and ($W1_{a2}$ x $V1_{a2}$) for A1 and A2 respectively, and selects the alternative with the highest value as an optimal decision outcome.
Conjunctive (CON)	It assigns a cut-off value for each criterion and selects an alternative with the value of each criterion above the cut-off value as an optimal decision outcome. The cut-off value is determined by business specialists where the value of it varies from task to task. For example, for some tasks a cut-off value of 50 for each criterion would be more appropriate but for other tasks it would be not sufficient. Mostly, an alternative with values above the cut-off for all or most of the criteria is selected as an optimal solution. Let the cut-off value for a specific decision task be 50 for each criterion. If $V1_{a1}$ is 50 and $V2_{a1}$ is 45 for A1 and $V1_{a2}$ is 65 and $V2_{a2}$ is 50, the CON strategy would select the alternative with the values of all or most of the criteria are above 50. In this case, the optimal decision outcome would be A2.

Table 3.4: *Summaries of the three kinds of decision strategies in literature [108]*

ceptual representation of decision tasks which can lead to an optimal solution compared to inexperienced decision makers [126].

However, it is not entirely correct to assume that experience is always good to arrive at the optimal solution for a decision task. Prior experience with the data may affect the feelings towards a specific data set. As a result, decision makers may rely more on their prior knowledge instead of using all the available data objectively. Therefore, decision makers may end the decision making processes early which, in turn, may negatively impact the outcomes of decisions [29]. Mao and Benbast [85] stated that experienced decision makers may consider their life-time knowledge more than the given information. Yet, they suggested that specialization may improve performance.

Prior research indicated that when the level of experience increases, so will the use of DQM for decision making purposes. However, the level and type of experience are also found to have a different impact on the use of DQM. Fisher et al. [36] indicated that decision makers who have more managerial experience used DQM more but managers who have domain-specific experience used DQM to a lesser extent.

3.3.3.3 Time

Decision time is a scarce resource for decision makers. Therefore, it needs to be utilized in a very efficient way. If providing DQM increases the decision time but doesn't increase the effectiveness of the decision output, then providing DQM can be assumed to have a negative impact on decision making.

Some researchers studied decision making with time pressure. Time pressure was measured differently by different researchers. Some researchers measured time pressure only by determining a specific time duration for a task [95]. However, other researchers differentiated between the time constraint and time pressure. They defined a time constraint as the specific time allowed and time pressure was defined as a subjective reaction about the time allowed for performing a decision task [36]. Time pressure can happen whenever decision makers perceive the allowed time as not sufficient to complete the decision task [139].

Surprisingly, Fisher et al. [36] indicated that some decision makers may feel greater time pressure when they are given a longer time limit. In the

same study, it is indicated that the time constraint did not affect the use of DQM, while the time pressure was found to have a positive impact on the use of DQM. Decision makers who felt time pressure integrated DQM more than those who did not feel time pressure in the same time constraint group. Conversely, a study by Price et al. [115] showed that providing DQM can significantly extend decision time.

In this study, the experiment did not impose any time constraint on subjects. However, they were asked to register the time at which they started and finished working on the experiment.

3.3.3.4 Data Quality Awareness (DQA)

One of the major goals of marketing is to make and maintain brand awareness. This is particularly important in an era when consumers actively search for information to assess their brand choices. Brand awareness is believed to have an impact on consumers' decision making, whereby the probability of brands being considered and selected can increase with the brand awareness level of the customers [78]. For example, customers usually say "I chose the brand I know," and, "I have heard of the brand so many times, I think it must be good."

Likewise, decision makers without DQ awareness may not fully use the DQM available in decision support systems. Similarly, DQ practitioners indicated the importance of creating DQA to bring DQ problems into consideration [109].

Although DQ is a problem that organizations are facing currently, creating DQA about the problem is not considered thoroughly. One reason can be that the impact of poor DQ on organizations' performance is not clearly known. Although there is an intuitive feeling that DQ awareness improves the use of DQM by making decision makers alert, the impact of DQ awareness on the use of DQM and its effect on decision performance are not studied.

Therefore, in the experiment of this paper, the impact of DQ awareness on the use and effect of DQM on decision performance is investigated.

3.3.3.5 Task complexity

Task complexity can be determined by different factors such as, the amount of relevant information (the number of decision alternatives and attributes) and decision time available [109]. Task complexity increases with the amount of data which need to be processed for a specific decision task [154]. Prior research has defined task complexity using number of cells in the matrix of decision alternatives and decision criteria. A task which has 20 or less cells is categorized as a simple task, while a decision task with more than 20 cells is categorized as a complex task [109]. This study used this threshold to classify the task as simple and complex.

3.3.3.6 Demographics

Education, age and gender are the three most important demographical variables. These variables are considered important in studying the impact of DQM on decision making purposes [36]. The subjects in this study belong to one age category (20-30 years old) and most of them are males, therefore there are no significant variances between the groups in respect to these two variables. Therefore, only the education variable is included in the experimental setting.

3.3.3.7 Dependent variables

Among the dependent variables in Table 3.3, this paper considers decision complacency, confidence, accuracy and efficiency to evaluate the impact of DQM on decision outcomes. Because we believe that these decision outcome measures are comprehensive enough to indicate the impact of DQM on decision outcomes. Thus, Figure 3.1 displays only the four decision outcome measures.

In summary, all the variables except data quality awareness (DQA) discussed from Section 3.3.3.1 to 3.3.3.7 were considered relevant for studying the use and impact of DQM for decision making purposes in prior studies [19, 130, 36, 129, 116]. In order to compare the results of this study with prior studies, all these variables are considered. In addition, DQA is the novel variable for this study as discussed in Section 3.3.3.4. Figure 3.1 shows the variables considered in a structured way.

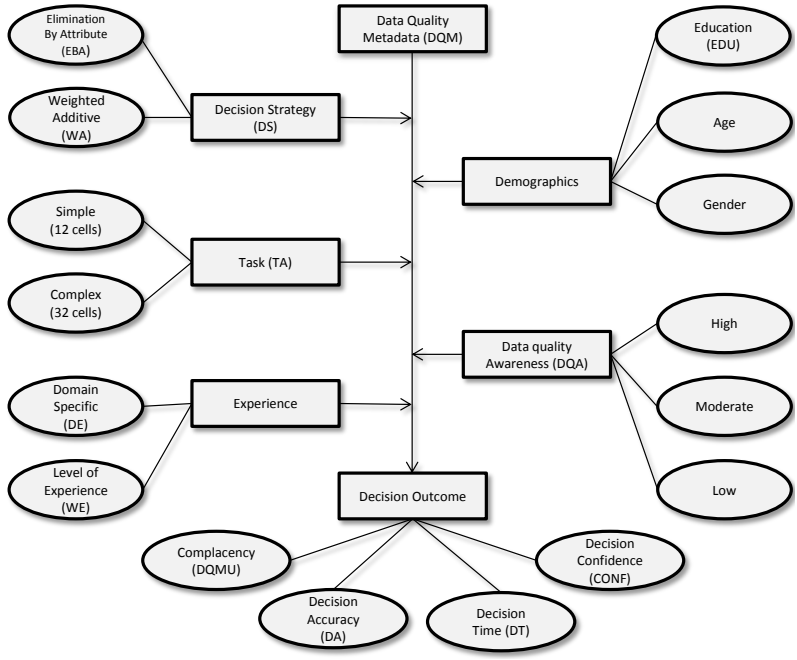


Figure 3.1: *The research setup which shows the use and impact of DQM on decision outcomes, and its interaction with other variables.*

3.4 Research Methodology

The research methodology is developed by considering different factors which would have an impact on the use of DQM as depicted in Figure 3.1.

3.4.1 Research aim

Prior research indicated the importance of providing DQM along with the actual data so that decision makers can gauge the appropriateness of the DQ level for the task at hand [19]. However, there are two important questions which should be answered before deciding to include DQM into databases because of the associated cost of creating, maintaining and manipulating it. First, the question of whether DQM positively impacts the effectiveness (the accuracy/quality) of decision outcomes should be answered. Second, it should be clearly known whether decision makers are not complacent to

the new information provided, in this case DQM. The latter question is adequately investigated in the DQ literature though there is no consensus about the results [19, 36, 130, 131, 116]. However, with the exception of a study by Shankaranarayanan et al. [129], the former has not sufficiently been investigated, to the best of our knowledge. Although both questions are important, the first one is more critical as the impact of DQM on the effectiveness of the decision outcomes can be either positive or negative.

In order to investigate the first research question, we employed three decision outcome measures: decision accuracy (effectiveness), decision confidence and decision time (efficiency). In other words, the impact of DQM on decision outcomes is measured in terms of those three dependent variables. Also, the interaction that DQM will have with other important variables such as experience level of decision makers and decision strategy is addressed under this research question.

Although DQM may have a positive impact on decision outcomes, maintaining DQM in databases would be more beneficial if the decision makers are not complacent to it. To investigate the second research question, we proposed the following hypotheses in order to identify different groups of decision makers who are not complacent to the DQM provided.

H1 Decision makers who are less educated equally incorporate DQM into their decisions in comparison to decision makers who are more educated.

H1a Educated decision makers **include DQM more** into their decisions compared to less educated decision makers.

When the education level increases, the complacency to the new information decreases [128].

H2 Decision makers who are less experienced equally incorporate DQM into their decisions in comparison to decision makers who are more experienced.

H2a Experienced decision makers **include DQM more** into their decisions compared to inexperienced decision makers.

Experienced decision makers are less complacent to new information than novices [36].

H3 Decision makers who have no domain-specific experience equally incor-

porate DQM into their decisions in comparison to decision makers who have domain specific experience.

H3a Decision makers without domain-specific experience (DE) **include DQM more** into their decisions than those with domain specific experience.

Specialization may prevent the use of all information objectively [48].

H4 Decision makers who have prior DQ knowledge equally include DQM into their decisions in comparison to decision makers who have no prior DQ knowledge.

H4a Decision makers with DQ knowledge **include DQM more** into their decisions than those with no DQ knowledge.

Generally, decision makers are inclined to base their decisions on the known variables by ignoring the unknown variables [119]. Therefore, decision makers who have no prior DQ knowledge are less likely to include DQM into their decisions for the reason that DQM is an unknown variable for them. In other words, the knowledge about DQ or DQ awareness triggers the use of DQM for decision making purposes.

H5 Decision makers who use a relatively simple decision making strategy to solve a decision task equally integrate DQM into their decisions in comparison to decision makers who use a relatively complex decision making strategy to solve a decision task.

H5a Decision makers who use a relatively simple decision strategy (DS) **include DQM more** into their decisions than those who use a relatively complex decision strategy.

As reported above, if the decision process taxes cognitive capacity, decision makers tend to simplify the process by being complacent to new variables [69]. Prior research has indicated that the simplicity of a decision strategy can depend on the type of task [97]. A previous study by Moges et al. [93] identified an EBA decision strategy as relatively complex and WA as relatively simple because of its compensatory nature. Therefore, we expect that decision makers who use a WA decision strategy include DQM more into their decision process than those who use an EBA strategy.

The hypotheses from H1(H1a) to H5(H5a) are similarly assumed in both

the simple and complex decision task scenarios. Therefore, in the results in Section 3.5, we use those hypotheses to present the results for both simple and complex decision environments.

H6 Decision makers facing either a simple or a complex task include DQM equally into their decisions.

H6a Decision makers facing a simple task include DQM into their decisions **more than** those who are assigned to solve a complex task.

More data and more choices may complicate the decision task. Decision makers can avoid cognitive overload by being complacent to additional variables such as DQM [114]. This can be explained by two well-known concepts in literature; information overload and cognitive capacity limit. Information overload can happen when a decision maker is asked to process more information than he/she is capable of. The second concept, cognitive capacity limit, is the result of an interaction between a working and long-term memory where decision making is processed [12]. Processing more data and many alternatives may load and demand a high cognitive capacity level where decision makers can hold data to process. Therefore, solving a complex decision scenario may load and demand a high cognitive capacity unlike the simple decision task, whereby it leads to omitting some information from decision processes. However, the feeling of information overload and cognitive capacity limit varies with the characteristics of decision makers. Prior research found that for the same amount of information novices may feel an information overload while experienced decision makers may not [31].

3.4.2 Experimental setting

3.4.2.1 Pilot study

To determine the appropriate DQM representation, in a previous study [93], we have conducted a pilot experiment using two different types of DQM formats. These are DQM with lower and upper value limits (interval representation), and probability representation. The interval DQM format shows the minimum and maximum possible values for specific data. The probability DQM format presents the likelihood that the value of a specific data item

represents its real value. Three groups were formed for the experiment, one that received interval DQM, one that received probability DQM and one without DQM. For experimental control, the decision strategies were limited to additive and EBA, and the experience level was limited to PhD students in Applied Economics.

The PhD students were randomly assigned to one of the three groups, each consisting of 10 students. A χ^2 statistical test indicated that there is no significant difference in the use of DQM between the interval and probability DQM format groups at a 95% confidence level. We conducted an exit interview with the PhD students in the two groups (interval and probability) to find out (1) how they incorporated DQM in their decision-making, (2) how they understood the meaning of DQM, and (3) what kind of DQM format they would prefer. The interview analysis indicated that the PhD students in the interval DQM group did not find an easy and uniform way of including the DQM in their decision-making process, which was also confirmed by a slightly higher time usage to finish the decision task compared to the other two groups. In contrast, the groups with the probability DQM format could easily understand the meaning of the DQM and used it all in a similar manner at a 95% confidence level.

Therefore, in the final experiment of this study, we employed a probability DQM format. In addition, we incorporated the usability study results that Price et al. [115] found. For example, we used the term accuracy for tag nomenclature and gave a detailed explanation of the meaning of the DQM with an example in the instruction section of the experiment.

Similarly, a pilot experiment with PhD students was conducted in order to investigate the clarity and understandability of the experiment. We identified three decision strategies EBA, WA, CON which subjects implemented to solve the decision tasks. Thus, the final experiment was designed to have the same decision solution using these three decision strategies so that the decision accuracy of each subjects can be evaluated using a similar decision solution. Finally, the pilot study confirmed that there are no ambiguities in the experiment.

3.4.2.2 Final task

It has been indicated that the nature of the application domain used in DQM experiments may influence the extent to which DQM is used [115]. This is

explained by the fact that participants may be less concerned that basing decisions on poor quality data negatively affects the decision outcome for a particular domain. Thus, previous research has suggested that the use of DQM for decision making should be investigated in different decision making environments, particularly in critical environments [115]. We therefore developed a new decision making environment which is a bankruptcy prediction task[2]. For example, subjects of the experiment may be more concerned to avoid the serious consequences of using poor quality data to solve the bankruptcy prediction task than an MP3 player or digital camera purchasing tasks. The bankruptcy prediction task is based on the Altman-Z model of bankruptcy prediction for non-manufacturing companies [2]. We employed all four criteria ($\frac{Retained\ earnings}{total\ assets}$, $\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$, $\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$ and $\frac{working\ capital}{Total\ assets}$) to determine the financial health of a firm. The relative importance of each criterion is directly adopted from the Altman-Z model. The decision task was developed similarly to previous studies to boost comparison of the results [19, 38, 132]. Therefore, the decision making process in this study is described as the process of ranking firms according to their financial health from best to worst, based on the given criteria. According to the results of the pilot study in Section 3.4.2.1, the experiment was designed to have the same ranking result using any of the three decision mechanisms (EBA, WA and CON) so that the ranking of the experiment can be evaluated using one correct answer. However, it was found that subjects in the final experiment used only the two decision strategies (EBA and WA) in their decision processes.

The task was categorized into two types, simple and complex. The simple task asked subjects to rank the financial health of four banks based on the first three criteria ($\frac{Retained\ earnings}{total\ assets}$, $\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$, and $\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$), which has a total of 12 cells. Meanwhile, the complex task asked subjects to rank the financial health of eight banks based on the four criteria ($\frac{Retained\ earnings}{total\ assets}$, $\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$, $\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$ and $\frac{working\ capital}{Total\ assets}$), which has a total of 32 cells. Both the simple and complex tasks were further grouped into two types where some subjects receive DQM upfront and other subjects receive DQM later in the decision processes. Subjects who did not get DQM upfront with the experiment were provided with DQM later and asked if they would change their decisions because of the DQM provided. This is done particu-

[2]Appendix

larly to increase the sample size for the complacency test. The four types of the experiment (simple task with upfront DQM, simple task with later DQM, complex task with upfront DQM, complex task with later DQM) were distributed randomly to subjects. Using a χ^2 test at $\alpha = 5\%$, it was verified that no statistically significant differences exist in the DQM usage between subjects who were provided with DQM upfront and who were provided with DQM later in the decision processes. Therefore, both groups of subjects are merged to determine the complacency level about DQM.

A clear description which explains the contents of the experiment, the meaning of each attribute and the expectations from subjects was also included in the experiment. The experiment was conducted in a controlled environment. An exit survey which consists of 28 questions was also conducted to gather demographic information after the experiment was finished. Finally, the subjects of the experiment were asked to register the time when they started and finished working on the experiment.

3.4.2.3 Participants

A total of 106 (80 business information system and 26 applied economics students) subjects participated in the experiment. The participants have been further segmented based on the exit survey provided with the experiment. 60 students solved a decision task with DQM upfront and 46 students solved a decision task where they received DQM later in the decision process. 42 and 64 participants solved the complex and simple decision task respectively. 30 of 106 participants have work experience. 35 have domain experience. 77 of the 106 participants have medium or high prior data quality knowledge and 29 participants have no prior data quality knowledge.

3.4.2.4 Variables

All the dependent and independent variables, and their descriptions and acronyms are included in Table 3.5.

Variable	Description
Decision Accuracy (DA)	it is a discrete variable which is measured using the ranking of the firms for both complex and simple tasks. If a subject gets the ranking entirely correct, the accuracy is said to be 10/10. Otherwise, if a person gets one ranking wrong, the accuracy level decreases by 1.25 or 2.5 marks for the complex and simple task respectively. Although in a real life scenario, it is not always possible to a priori determine the optimal decision with uncertainties such as poor quality data, we designed the experiment to have a relatively correct ranking order by considering the DQ level given using the three commonly used decision strategies[a] in decision making processes. Therefore, the decision accuracy of each subject can be evaluated using a similar decision solution. As such, the definition of a 100% decision accuracy in this experiment's context is that one has achieved the predefined relatively correct ranking order without making any ranking error using one of the three decision strategies. In short, accuracy measures the number of errors that a decision maker makes in solving the decision task. The errors include calculation and inconsistencies in the decision making process.
Decision Confidence (CONF)	is an ordinal variable which is measured with a 5-Likert scale which ranges from -2 to 2. -2 and -1 represent very low or low confidence about the decision outcome respectively, while 0 represents a neutral feeling about the decision task. Finally, 1 and 2 represent high and very high confidence about the decision outcomes.
Decision Time (DT)	is a continuous variable which is measured by subtracting the start time from the finish time of the experiment. It is measured using minutes.
Data Quality Metadata Used (DQMU)	is a categorical variable with a "yes" or "no" label. It is measured based on three questions. First, subjects are asked to write the formula or method which they used to solve the task, if the method contains DQM and the decision solution belongs to the solution with DQM category, then the DQMU variable will get a "yes" label. Also, the response to the question "which variables were important to solve the decision task" is checked for its consistency to the solution provided. In almost all cases, the subjects' responses to the three questions were consistent. Therefore, the use of DQM is correctly determined.
Data Quality Metadata (DQM)	is a categorical variable with a "yes" or "no" label. It is measured based on the type of the experiment. For the experiment types which include DQM upfront, the DQM variable has a "yes" label, but otherwise, it has a "no" label.
Decision Strategy (DS)	it is categorized into two broad types, Weighted Additive (WA) and Elimination by Attribute (EBA). Subjects are asked to explain which kind of decision strategy they used to solve the decision task. Based on their explanations, the decision strategy is classified into either EBA or WA decision strategies.
Task Clear (TAclear)	is a variable which is measured by subjects' responses to the question "whether the experiment was fully clear". It has a "yes" or "no" label.
Task Type (TA)	is a categorical variable with "Complex" or "Simple" label which is determined from the type of the experiment.
Data Quality Awareness (DQA)	is measured by subjects' responses to the six basic data quality related questions asked in the exit survey. If a subject answers the four important questions[b c d e] correctly, a subject is said to have full data quality awareness. If a subject answers two or three of the four important questions correctly, then the subject is said to have medium data quality awareness. Subjects who answered only one question or did not answer any of the questions correctly, are said to have no data quality awareness.
Work Experience (WE)	is a variable which shows the work experience level of subjects. It is determined by the response to the question "how many years of work experience do you have" in the exit survey.
Domain Experience (DE)	is a variable which shows the level of previous similar experience with that of the experiment that subjects have. It is also determined by the response to the question "how many times did you solve this kind of exercise previously".
Education (EDU)	is a variable which shows the level of education that a subject has achieved.

Table 3.5: *Summaries of the different variables in the experiment*

[a] Table 3.4
[b] How do you define accuracy in the data quality context?
[c] Can you please give one example of inaccurate data?
[d] Can you please explain what is meant by data quality dimensions/attributes?
[e] Can you please mention some data quality dimensions/attributes?

3.4.3 Statistical Analysis

In order to test the significance of the obtained results, a number of statistical tests are applied in accordance with the literature. Each of the different tests is assessed at a significance level of 5% unless stated otherwise.

3.4.3.1 Chi-square

The χ^2 test basically investigates the null hypothesis, whether the frequency distribution of observed events in a sample is equal to an expected frequency distribution of the same events derived from a particular theoretical distribution or from the control groups in the observed samples. A χ^2 analysis can also be used to determine whether paired observations on two variables are independent from each other (e.g. the education level of subjects and their DQM usage). In this paper, a χ^2 test is used in this latter way to investigate the complacency level of the decision outcomes between different groups, to determine if there is a relationship between subjects characteristics and DQM usage for decision making purposes [84].

3.4.3.2 Regression trees - Leave-one-out-cross validation

Regression trees

Tree-building algorithms generally define a set of logical environments by which different cases can be predicted or classified with some degree of accuracy. Regression trees are non-linear and non-parametric algorithms which predict continuous dependent variables using one or more continuous or categorical independent variables [54]. Regression trees are non-parametric and avoid the assumptions where tests such as Analysis Of Variance (ANOVA) and t-tests rely on, namely that data are normally, and independently and identically distributed [iid] [54] . In addition, in most cases, interpreting the results from the trees is very straightforward. Moreover, the trees do not assume any relationship (linear, non-linear or monotonic) between the predictor and the dependent variables. For example, decision accuracy can be negatively related to the use of DQM, but can also be positively related to the use of DQM if subjects have a high data quality awareness or high experience level whereby the tree can reveal such a non-monotonic relationship between

the variables. Thus, linear regression trees are good methods for cases where there is little or no prior knowledge about the relationship between the dependent and independent variables. Consequently, they are particularly suited to analyze this experiment's data where there is no prior assumption about the three dependent variables i.e., decision accuracy (DA), decision confidence (DC) and decision time (DT), and their predictors [60].

Leave-one-out-cross validation

Cross-validation is used to assess the performance of a regression or classification model on previously unseen data. Assessing the performance of the model is mainly the case in predictive analytic, where one predicts a model and determines its performance in practice. In general, model performance is measured by splitting the data in a training and test set. The model is estimated using the training set; the test set assesses its performance. However, in cross-validation, the original data set is split into several subsets, each of which is once used for testing purposes to assess the model's performance after the training phase. This technique helps to minimize overfitting and gives an insight on how the model will work on an independent real-life data set. In addition, it is very useful when a small sample size makes it difficult to split the data into separate training, validation and test sets. Leave-one-out-cross validation is one type of cross validation where each observation in the sample is used once as the validation data and the remaining observations as training data [44].

In this paper, we implemented a regression tree with leave-one-out-cross validation in order to predict the values of the three dependent variables (decision accuracy (DA), decision confidence (DC) and decision time (DT)) from all the available independent variables.

3.4.3.3 Stepwise regression

A stepwise linear regression was also implemented in order to choose predictive variables with their interaction effect for each dependent variable, Decision Accuracy (DA), Confidence (CONF) and Time (DT). However, the results from the leave-one-out-cross validation and the stepwise regression are found to be similar with some minor distinctions. In addition, the results of leave-one-out-cross-validation outperformed that of the stepwise

regressions when the models were compared by the mean squared error (MSE). Therefore, in this paper, we will only present the results from the leave-one-out-cross validation tests [60].

3.5 Results and discussions

3.5.1 The use of DQM in decision making processes

3.5.1.1 Education, Experience, Data Quality Awareness and Decision Strategy - Simple decision task

As defined in Table 3.3, complacency measures the degree to which DQM is used in decision making processes. A χ^2 test is conducted on the variable DQMU (see Table 3.5) in order to investigate the complacency level of different decision makers towards the DQM. The χ^2 test results are depicted in Table 3.6 and indicate that for the simple decision task, the complacency level about DQM of decision makers who are more or less educated is not significantly different at a 95% confidence level. Similarly, the complacency level between decision makers who have and do not have work experience; and who have and do not have domain experience is not significantly different. In addition, there is no relationship between the complacency level and the decision strategies implemented. Therefore, H1-H3 and H5 are accepted, and H1a-H3a and H5a are rejected at a 95% confidence level for the simple decision task. Yet, there is a significant relationship between the complacency level of the decision makers and their DQA level at a 95% confidence level. In other words, subjects who have a high DQA integrated DQM more into their decision processes than subjects who have little or no DQA. Among 24 subjects who have little or no DQA, 17 of them did not integrate DQM into their decision processes, yet, from 40 subjects who have a high DQA, only 15 of them did not integrate DQM to solve the decision task. Therefore, H4 is rejected in favor of H4a at a 95% confidence level.

Simple Task				
Variables		DQMU	Obs.	Complacency
EDU (H1)	Under graduates	Yes	20	$\chi^2 = 0.2591$
		No	18	
	Post graduates	Yes	12	$p = 0.6107$
		No	14	
WE (H2)	No experience	Yes	23	$\chi^2 = 1.6967$
		No	18	
	With experience	Yes	9	$p = 0.1927$
		No	14	
DE (H3)	Without DE	Yes	22	$\chi^2 = 0.2771$
		No	20	
	With DE	Yes	10	$p = 0.5986$
		No	12	
DQA (H4)	Without DQA	Yes	7	$\chi^2 = 6.6667$
		No	17	
	With DQA	Yes	25	$p = 0.00098^{**}$
		No	15	
DS (H5)	WA	Yes	18	$\chi^2 = 1.0667$
		No	22	
	EBA	Yes	14	$p = 0.3017$
		No	10	

Table 3.6: *The complacency level of different groups of subjects on their decision outcomes when Data Quality Metadata (DQM) is given and the decision task is relatively simple.* $^{**} = p < 0.05$.

3.5.1.2 Education, Experience, Data Quality Awareness and Decision strategy - Complex decision task

The χ^2 test results in Table 3.7 indicate that for the complex decision task, the complacency level about DQM of the decision makers who are more or less educated is not significantly different at a 95% confidence level. Similarly, the complacency level between decision makers who have and do not have work experience is not significantly different. In addition, there is no relationship between the complacency level and the decision strategies. Therefore, H1, H2 and H5 are accepted and H1a, H2a and H5a are rejected at a 95% confidence level for the complex decision task. Yet, the complacency level between subjects who have DE and who have no domain experience is significantly different at a 95% confidence level. Subjects who have prior DE on the decision task are more complacent towards the given DQM to solve the decision task than those subjects who have no domain experience. In other words, among 13 subjects who have prior DE, only 3 of them integrated

DQM. Yet, among 29 subjects who have no prior DE, 23 of them integrated DQM into their decision making processes. This may suggest the fact that domain relevant knowledge is used more often when the task is complex [95]. In addition, decision makers may fully rely on their domain experience to reduce their cognitive effort by ignoring the less relevant variables when the decision task is difficult [46]. There is also a significant relationship between the complacency level of decision makers towards DQM and the DQA level of the decision makers at a 95% confidence level. In other words, subjects who have more DQA integrated DQM more than subjects who have no DQA, similar to the simple decision task. Among 5 subjects who have little or no DQA, 4 of them did not integrate DQM to solve the decision task, yet, from 37 subjects who have a high DQA, only 12 of them did not integrate DQM to solve the decision task, though both groups (subjects with DQA and subjects without DQA) were provided with DQM. Therefore, H3 and H5 are rejected and instead H3a and H5a are accepted at a 95% confidence level.

To summarize the results from Table 3.6 and 3.7, for both simple and complex decision tasks, the DQA level of decision makers has a high impact on the degree to which decision makers are complacent towards DQM. The higher the DQA level is, the more the decision makers integrate DQM into their decision making processes. This reinforces the suggestion from an earlier study by Fisher et al. [36] that organizations should conduct a seminar and DQM education prior to maintaining DQM in databases in order to fully benefit from it. Similarly, the results in Table 3.7 reveal that decision makers with prior knowledge about the decision task used DQM less than those without prior knowledge for the complex decision task. On the other hand, the results in Table 3.6 and 3.7 indicate that there is no relationship between the level of education, work experience and the use of different decision strategies, and the complacency level of decision makers towards DQM. These results agree with the earlier findings that the education level and the type of decision strategy did not affect the use of DQM for decision making processes [36, 116].

3.5.1.3 Task type

As the results of the χ^2 test in Table 3.8 show that there is no significant difference found between the complacency level of decision makers who solved the simple and complex decision task at the 95% confidence level. Therefore, H6 is accepted and H6a is rejected. This result may be explained by the fact that

Complex Task				
Variables		DQMU	Obs.	Complacency
EDU (H1)	Under graduates	Yes	21	$\chi^2 = 0.1958$
		No	12	
	Post graduates	Yes	5	$p = 0.6581$
		No	4	
WE (H2)	No experience	Yes	22	$\chi^2 = 0.0808$
		No	13	
	With experience	Yes	4	$p = 0.7763$
		No	3	
DE (H3)	Without DE	Yes	23	$\chi^2 = 12.0361$
		No	6	
	With DE	Yes	3	$p = 0.0005^{**}$
		No	10	
DQA (H4)	Without DQA	Yes	1	$\chi^2 = 4.2262$
		No	4	
	With DQA	Yes	25	$p = 0.0398^{**}$
		No	12	
DS (H5)	WA	Yes	19	$\chi^2 = 1.2620$
		No	9	
	EBA	Yes	7	$p = 0.2613$
		No	7	

Table 3.7: *The complacency level of different groups of subjects on their decision outcomes when Data Quality Metadata (DQM) is given and the decision task is relatively complex.* $^{**} = p < 0.05$.

both decision tasks (simple and complex) entailed a similar problem, with the task complexity determined only by the number of alternatives.

3.5.2 Data quality metadata and its impact on decision outcomes

3.5.2.1 Decision accuracy

The regression tree in Figure 3.2 indicates that the dependent variable, decision accuracy (DA), can be predicted by the independent variables DQM, DQA, DS and TA with a low mean squared error of 0.5715. The DA variable is measured with a scale 0 to 10, with the lowest value 0 and highest value 10. If decision makers have a high DQA, the probability of having a good level of decision accuracy is high. Yet, if decision makers have little or no DQA, decision accuracy depends on the presence of DQM, DS and TA variables. In

Complacency for the decision task when DQM is provided				
Variables		DQMU	Obs.	Complacency
TA (H6)	Simple	Yes	32	$\chi^2 = 1.4505$
		No	32	
	Complex	Yes	26	$p = 0.2285$
		No	16	

Table 3.8: *The complacency level of subjects on their decision outcomes when Data Quality Metadata (DQM) is given in combination with the complexity of the decision task.*

general, if decision makers who have no DQA integrate DQM into their decision processes, the decision accuracy will be very low. However, if decision makers who have moderate DQA integrate DQM into their decision processes and use a weighted additive decision strategy, their decision accuracy will be high. Conversely, if decision makers do not integrate DQM into their decision processes, the decision accuracy depends on the the level of difficulty of the decision task, whereby a complex decision task leads to a lower decision accuracy. Summarizing the results, decision accuracy mainly depends on the level of DQA that decision makers have. Similarly, decision makers who have a high DQA use DQM more than those decision makers who have little or no DQA (see Table 3.6 and 3.7) and reach a high consensus on their results. The consensus level is indicated by the high DA that those decision makers with high DQA have. In general, the results indicate that decision makers who have a high DQA can have high decision accuracy regardless of the decision strategy or the complexity of the task they dealt with. This may be explained by the fact that decision makers who have a high DQA are more educated as DQA knowledge is mostly acquired from the extra training in addition to the formal education. Although the results in Figure 2 and 3 indicate that DQA increases decision accuracy regardless of the use of DQM and increases decision confidence for decision tasks with DQM respectively, the results in Figure 4 indicate that DQA considerably increases the decision time. Therefore the results in Figure 2 do not necessarily suggest that a high DQA always leads to a good decision performance as decision performance needs to be evaluated based on all the decision measures (decision accuracy, time and confidence).

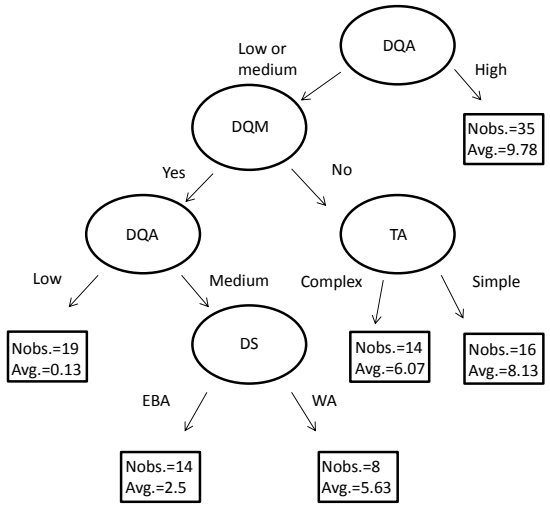


Figure 3.2: *A regression tree for the decision accuracy (DA) with MSE=0.5715. The minimum score is zero and the maximum is 10.*

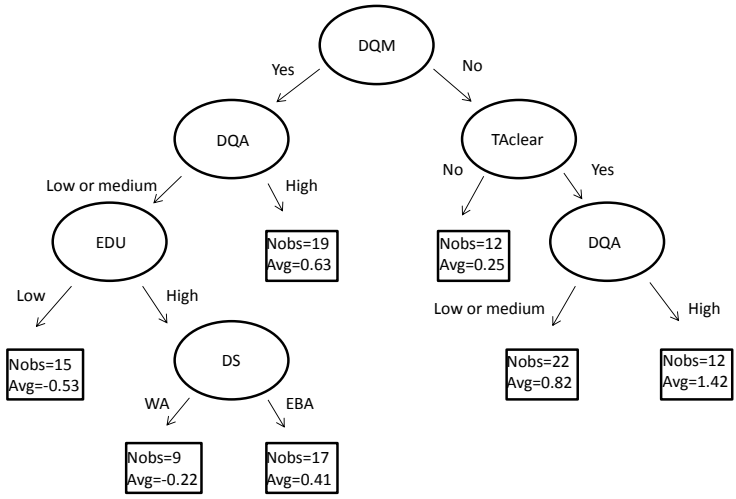


Figure 3.3: *A regression tree for the confidence level of decision makers on their decision outcomes with MSE=0.089.*

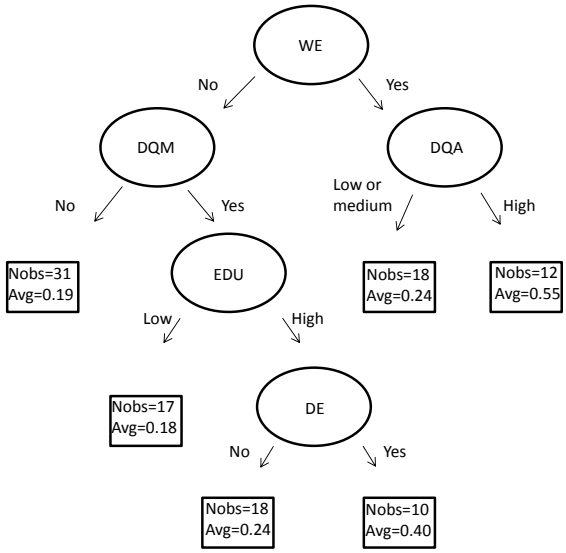


Figure 3.4: *A regression tree for the decision time measured in minutes with MSE=0.0118.*

3.5.2.2 Decision confidence

The regression tree in Figure 3.3 indicates that the dependent variable, decision confidence, can be predicted by the independent variables DQM, DQA, TAclear, EDU and DS with a MSE of 0.089. If decision makers do not integrate DQM into their decision processes and the decision task is not clear, the decision confidence is very low. Yet, if decision makers do not integrate DQM into their decision processes and the decision task is clear, then the decision confidence depends on the DQA level, as such, a high DQA leads to a better decision confidence. Similarly, if decision makers who have a high DQA integrate DQM into their decision processes, the decision confidence will be high. However, if decision makers who have little or no DQA integrate DQM into their decision processes, the decision confidence depends on the education level and decision strategy, whereby a high education level and an EBA decision strategy lead to a relatively high decision confidence. Summarizing the results, DQM usage, a high DQA level, the clarity of the decision task, a high educational level and a complex decision strategy have a positive impact on the confidence of decision makers on their decision outcomes. In other words, decision makers with a good prior data quality

knowledge and a high educational background, and who understand the decision task clearly will have a high decision confidence when they integrate DQM into their decision processes and use a more complex decision strategy such as EBA. Although Figure 3.3 depicts the interaction effects of different variables with DQM on the decision confidence and should be compared with other studies with interaction effect, the results can partially be compared with the findings of Moges et al. [93] where the confidence level of decision makers was slightly higher for decision makers who used an EBA decision strategy to integrate DQM into their decision processes.

3.5.2.3 Decision time

The regression tree in Figure 3.4 indicates the prediction of the decision time in terms of the different independent variables. The results indicate that, in general, decision makers who have work experience will take more time to solve a similar task than those who do not have many years of work experience. Yet, for decision makers who have no work experience the decision time will depend on their integration of DQM into their decision processes. Decision makers without experience and who do not integrate DQM into decision making processes take less time to solve the decision task compared to decision makers with experience and a high DQA. However, decision makers who have no experience, low educational background and who did integrate DQM into decision making processes were more quicker to solve the decision task. Those without experience who do integrate DQM will vary their decision time based the education and the domain experience level of the decision makers. More precisely, a high DQA, a high educational background and experience level extend the decision time. One possible explanation would be that decision makers with plenty of experience and high DQA level may consider different information from their past experiences and encounters instead of limiting themselves only to the information provided or may rely on their past experiences instead of considering all the given information to solve the decision task. This may have resulted in an increased decision time. In the same way, decision makers who integrate DQM can have extended decision time compared with those who did not integrate DQM into their decision processes. This result confirmed previous findings by Price et al. [116] that the use of DQM can increase decision time. Yet, the results in Figure 3.4 indicate that if DQM is provided to decision makers with lower educational background, the decision time considerably decreases. Whereas

a shorter decision time in itself may be desirable, the latter should always be considered in the context of other variables such as decision accuracy. Therefore, for concrete tasks, the tradeoff between time pressure and required quality of the decision outcome will determine the desirability of including DQM, given other factors such as task type and experience level.

3.6 When and to whom is DQM beneficial for decision making purposes?

Decision makers who have DQ knowledge used DQM more than those decision makers who do not have DQ knowledge to solve both the simple and complex decision tasks. Similarly, decision makers who have a general experience regarding the decision task used DQM more than those decision makers who are specialists with respect to the complex task. Although there are no significant differences in the decision outcomes in terms of decision accuracy, time and confidence between decision makers who did and did not use DQM at a 95% confidence level, the results indicate that DQM affects decision outcome if it is associated with certain characteristics of decision makers. More precisely, decision makers who have a moderate level of DQA and used a compensatory decision making strategy have a high decision accuracy. Likewise, if decision makers who have a high DQA level use DQM, their confidence level increases. Similarly, decision makers with low educational background who integrated DQM into their decisions use less time to solve the decision tasks. Therefore, these results advise organizations to include DQM into decision support systems with compensatory decision making strategies and intended for general managers. In addition, raising DQ awareness and providing training on how to use DQM in decision support systems is beneficial for improving decision accuracy and confidence when DQM is provided.

Although ideally, a decision outcome is optimal if all the corresponding decision outcome measures such as decision accuracy, efficiency and confidence are high, in reality decision making processes consider the tradeoffs between the decision outcome measures. For example, some medical diagnosis tasks may need accuracy more than efficiency. On the other hand, stock trading tasks may need more efficiency than accuracy. Therefore, system designers are advised to consider the benefits of DQM with respect to the intended use

of the decision support systems.

3.7 Limitations

Although students are used as study subjects in this paper's experiment, so as to be able to compare the results of this study with the results of prior studies of the same kind which have used students as subjects, we would like to recommend an extension of this study using other subject groups such as business people who are very familiar with real-life business decision making processes. In that way, the impact of DQM on decision making processes can properly be better revealed.

3.8 Conclusion

In this study, a new experiment was conducted in order to investigate the use of DQM for decision making purposes and its impact on decision outcomes. The experiment was motivated by prior research results where there was no agreement on the use of DQM for decision making processes [36, 131, 116].

This study addressed different notions either suggested by previous studies or inferred from missing factors in their experimental designs. One of the suggestions addressed is that we created an equal or a similar understanding of DQM among the participants by using a clear description and an example of what DQM represents which, consequently, helped to clear ambiguities. Also, the study incorporated all the variables studied in previous DQM research in addition to novel variables such as DQA which makes the study inclusive. This, in turn, helped to measure the effect of the variables on the use of DQM in a similar environment where similar subjects are used, removing the impact of an experimental design.

The main contribution of this study is the way the decision outcome measures were defined and analyzed. First, the complacency of decision makers towards DQM, one of the most widely recognized decision outcome measures in the DQM literature, is inferred from three standards instead of associating complacency indirectly from the change of the first decision choice or the

usage of one attribute as in prior studies [36, 116, 19]. Complacency is determined by participants' decision strategy where they explicitly indicated the formula they used to reach the decision solution. Second, it is derived from the category of the decision solution where the decision solution can either be categorized as a solution with DQM or without DQM. Finally, complacency is measured by the responses to the question "which variables were important in the decision processes". In nearly all cases, the three indicators of complacency were consistent for each subject. Second, the decision accuracy of each subjects was evaluated using one possible decision solution. This could be possible because the experiment was designed to have one possible decision outcome for the three decision strategies (WA, EBA, and CON).

Another key contribution is that the way the three decision outcome measures, decision accuracy, decision time and decision confidence were analyzed. We used a tree based algorithm to identify the impact of DQM and its interaction effect with other independent variables on these three measures. The results gave new insights on the impact of DQM on decision outcomes.

The use of DQM for decision making purposes was investigated using a χ^2 test. The results in Table 3.6, 3.7 and 3.8 indicated that the use of DQM is enhanced by prior data quality knowledge of decision makers where decision makers with prior DQ knowledge integrated DQM into their decision processes significantly more than those decision makers with no prior DQ knowledge. One possible explanation would be that prior DQ awareness could increase the understanding of potential consequences of making decisions using incorrect data. Understanding the consequence of using flawed data encourages the incorporation of DQM into decision processes. Another explanation can be, as prior market research indicated, that brand awareness increases the chance of the brand being purchased by users. Similarly, the more decision makers are familiar with the variable, the more they use it into their decision making processes [78]. Hereby, we can suggest that organizations conduct a seminar or DQM training prior to maintaining DQM in data warehouses. The results in Table 3.7 also indicate that the use of DQM decreases with domain experience level, whereby, a high domain experience or specialization level is associated with a lower usage of DQM to solve the complex decision task. This result is consistent with the findings by Fisher et al. [36] that more specialization may avert the use of all available information. Additionally, decision makers who have more experience about the decision task are may be more influenced by their prior experience than other information given, such as DQM. On the contrary, education level, work

experience level and decision strategy are found to have no effect on the use of DQM. Based on these results, we can suggest that a special effort should be paid to encourage users with domain experience to still make use of DQM.

The regression trees in Figure 3.2 suggest that in order to reach a high consensus or a high decision accuracy, DQM should be used by those who have prior DQ knowledge. Yet, in general, the tree indicated that those who have not used DQM reached a high decision accuracy though they solved a complex decision task. Put differently, although decision makers who have prior DQ knowledge benefited from using DQM, decision makers who did not use DQM could still reach to a high decision accuracy. This can be explained by the preference of decision makers to use a complex decision strategy, such as EBA, to integrate DQM into their decision making processes. To clarify, among 38 subjects who used an EBA decision strategy, 26 integrated DQM into their decision processes.

The regression tree in Figure 3.3 indicates that, in general, decision confidence decreases when decision makers used DQM. However, decision confidence is found to be very high when decision makers have a high DQA. Similarly, when decision makers who have a high DQA solved the decision task using DQM, the confidence level is said to be high. In the same way, the regression tree in Figure 3.4 indicates that a high DQA and DQM use increased the decision time.

Although the decision accuracy and the decision confidence can be improved when decision makers who have a high DQA integrate DQM, the decision time increases. As DQ is contextual, for tasks which are critical and where the consequence of flawed data is high, providing DQM seems advantageous. However, for tasks which need a high efficiency, DQM seems rather counterproductive.

The general conclusion we can draw from this analysis is that DQM can be used and positively impacts the decision outcome when it is associated with certain characteristics of decision makers and some decision strategies. However, the benefits and costs of DQM can be different among organizations depending on the size, business directions and the degree of necessity of decision support systems [116]. Therefore, management should evaluate and decide upon the cost and benefits of integrating DQM in detail as the the benefits of DQM are discovered to be application dependent.

Although the results of this study identify different characteristics of decision makers who can possibly integrate DQM into their decision making processes, the possible impact of DQM on decision making outcomes in association with different groups of decision makers and the results are validated by comparing them with the results of prior studies of the same kind, the results of this study can further be validated using different experiment groups in a real-life context instead of using a simulation as this paper did. We believe that validating the results of this paper using a real-life experiment can be an interesting future research idea.

4

Maturity Assessment of Data Quality (DQ) Management Activities in Financial Institutions

4.1 abstract

Good quality data are necessary for healthy business operations and strategic decisions in every organization. However, maintaining good quality data is difficult because of the vast amount of data being collected every day and the difficulty of executing successful data quality (DQ) management activities. DQ is broadly defined as fitness for use and measured along different dimensions such as accuracy, completeness and consistency. Mature DQ management practices are relevant to enhance the quality level of the data. Thus, assessing the maturity level of DQ management activities is useful to understand the best practices of the DQ activities in mature organizations and to identify different process areas for improvement. In line with this, this paper assesses the maturity level of DQ management activities in five financial institutions using the Information Quality Management Capability Maturity Model (IQM-CMM) in a case study methodology. The results indicate that among the five financial institutions only one has a relatively high DQ management maturity level; (Level 4 in the IQM-CMM). Furthermore, seven key process areas for improvement are identified: Information Quality Needs

Analysis, Information Quality Assessment, Information Quality Management (IQM) Roles and Responsibilities, IQM Governance, Enterprise Information Architecture Management, and Continuous DQ Improvement. In addition, a framework is suggested on how to organize DQ measuring practices in financial institutions because DQ measuring activities are one of the most important DQ management activities.

4.2 Introduction

The introduction of the Basel II and III accords is often the reason for data quality (DQ) improving activities in financial institutions [91]. These accords impose different requirements, among which solid risk data aggregation and risk reporting practices[1] are the major ones. These requirements are directly and strongly related to the quality of the data used for risk assessment. As a result, financial institutions are currently engaged in different activities to attain the required level of DQ. Although the negative consequences of poor DQ on operational and strategic decisions seem clear, in fact, DQ activities are often solely motivated by Basel II and III requirements [92].

Some of the important DQ dimensions indicated by the accords are accuracy, completeness, timeliness, clarity and comprehensiveness [4]. Accuracy is defined as the extent to which a record and its true value are close. Completeness refers to the availability of all relevant or required data to perform risk assessment. Clarity indicates the extent to which risk reporting is understandable. Finally, comprehensiveness is defined as the extent to which risk reports include all the relevant risks to the financial institution.

It has been indicated that many operational and decision making activities in financial institutions are impacted by poor quality data. Moreover, it was argued that organizations which do not consider DQ management as part of their business activities will have difficulty in maintaining their status in business environments because inaccurate and incomplete data may adversely affect the competitiveness of an organization where operational and strategic activities are mainly based on the analysis and interpretation of data [96, 89]. Although many financial institutions are conducting different DQ activities to

[1] "Risk data aggregation means defining, gathering and processing risk data according to the bank's risk reporting requirements to enable the bank to measure its performance against its risk tolerance"

mitigate the impact of poor quality data, the activities are not well-organized instead they are ad hoc [9]. However, it has been indicated that without mature DQ management activities, it's difficult to create good quality data [135]. Therefore, the major aim of this paper is to identify the general and key process areas where enhancement is often due, and provide optimization recommendations by conducting a comprehensive maturity assessment of the data and DQ management processes and approaches in the five financial institutions.

The paper is organized as follows. The next section discusses different studies related to DQ maturity, DQ measuring framework, and accuracy and completeness metrics. Section three clarifies the case study methodology used in order to assess the maturity level of the DQ management activities in five financial institutions. Section four discusses important findings. Finally, the paper gives concluding remarks and indicates future research ideas.

4.3 Literature review

In this section, different DQ maturity assessment models and metrics will be discussed.

4.3.1 Maturity assessment

The main aim of maturity assessment is to determine the maturity level or existence of necessary activities to achieve the intended goals [5]. There are many models proposed to assess the maturity level of organizational activities [22, 39, 64, 13, 5].

Assessing the maturity level of quality management was first proposed by Crosby [22]. The author developed a Quality Management Maturity Grid (QMMG) which has five levels (Uncertainty, Awakening, Enlightenment, Wisdom and Certainty). The levels are used to assess the maturity of different activities such as company quality posture, quality improvement actions, cost of quality as % of sales, problem handling, management understanding and attitudes of quality. Later, this grid has been adapted by many to assess the maturity level of activities in different areas such as software engineering and information quality management [118, 32].

Maturity stages	Definitions & Benefits
Define	*Identify and define the key attributes of the product or process*. "Identify the business needs met by the process, scope the process, and identify the Critical to Quality (CTQ) characteristics of the process output and tolerance limits".
Measure	*Determine how (e.g., deciding the measurement device)the key attributes will be measured for their quality*. "Obtain quality process data and begin analysis. Measure quality based on user requirements".
Analyze	*Analyze and identify source of variation or weak processes to improve*. "Identify the root causes of process problems and key factors in a process".
Improve	*Clean the source of variation and improve weak process areas.* "develop appropriate process and/or product improvements while considering business needs".
Control	*Put controlling mechanisms to keep the best processes as defined.* "If implemented incorrectly, could result in having to repeat the entire process".

Table 4.1: *Six Sigma framework to create or enhance the stability of processes in organizations. Table 4.1 is directly adopted from [124].*

Similarly, a Total Quality Management framework was proposed to enhance the quality of products and services in organizations [26]. The principal focus of the framework is customer satisfaction which means quality should be valued by customers and should always be directed towards their needs and expectations. Since then, the framework has been adapted by many organizations and extended by many researchers. As a result, many critical success factors (CSFs), different activities which are key to assess the maturity of organizations' performance, were proposed [83, 98]. In the same way, the Total Data Quality Management (TDQM) framework at Massachusetts Institute of Technology (MIT) has been developed based on the TQM theories to assess information quality management by assuming a similarity between manufacturing of tangible products and data products (DPs) [146].

Following the Total Quality Management framework, the Six Sigma methodology was developed to enhance or create stable processes in organizations [113]. Table 4.1 shows the Six Sigma levels and descriptions as one example of quality maturity models.

Based on the QMMG, TQM and Six Sigma frameworks, many maturity models have been developed to assess the maturity level of different activities and

processes in different sectors.

The Data Quality Management (DQM) Maturity Model (DQMMM) was developed to improve data structure and management quality [64]. The model has four maturity levels to assess the maturity of data structures and it indicates the requirements for each level. The model was based on the assumption that poor data structure causes poor data values and service quality. It also gives guidelines for reaching the highest maturity levels. The model was constructed based on the results of case studies.

Likewise, the Information Quality Management Maturity (IQM3) model with five maturity levels (Initial, Defined, Integrated, Quantitatively Managed and Optimized) has been suggested to assess information quality management activities in general [13]. The model is developed based on the well-known Capability Maturity Model Integration (CMMMI) model in software engineering [106].

Similarly, the Information Quality Management Maturity (IQMM) model which is based on Crosby's maturity framework has also five levels to assess DQ [22]. The model aims to understand the necessary resources required to develop a DQ management tool.

In the same way, the Information Quality Management Capability Maturity Model (IQM-CMM) with five maturity levels (chaotic, reactive, measuring, managed and optimizing) was proposed to assess the maturity level of DQ management related activities in asset management organizations [5].

All the models discussed above have similarities in the way that they are organized into different levels with distinct CSFs to assess the maturity levels of different business activities and processes in organizations [5].

There are many models which can be used to assess the maturity of DQ management activities in different organizations as discussed in Section 4.3.1. We adopted the IQM-CMM model to assess the maturity level of DQ management activities in financial institutions because the model is developed inductively and it is based on empirical data gathered from many DQ experts and data users. Therefore, the model includes the perception and knowledge of different data stakeholders in the practical world. In addition, the model allows to assess the maturity level of both data and DQ management activities. Furthermore, it is generic to be applicable to different sectors [5].

The high level view of the IQM-CMM model is depicted in Figure 4.1. The

model has five evolutionary levels. Organizations on level one, Chaotic, do not consider DQ problems as an issue. Although such organizations may have basic DQ management processes, they are not clearly documented and consistently implemented. There are no plans to enhance the DQ levels. Organizations that do not satisfy the maturity indicators on level 2 are classified under this level [5].

Organizations can be assigned to level 2, Reactive, if they are aware of any DQ problems that exist and have been practicing basic DQ management activities. Such organizations may have identified data stakeholders, information needs, and developed conceptual, logical and physical data models. In addition, there may be appropriate storage management policies which indicate how information should be backed-up, archived and destroyed. Proper access control may have been practiced, i.e., permitting only authorized personnel to access the information system. All organizations that do not fulfill the appraisal criteria of level 3 are classified under this level [5].

Organizations on Level 3 are becoming more aware of their data resources and have started managing them as products. Such organizations have configuration management processes which ensure the recording and rolling back of any changes. Relevant DQ dimensions may have been elicited from all the major stakeholders. Therefore, these organizations may have developed qualitative or quantitative DQ metric and thus, a regular DQ assessment may be possible. In addition, a DQ team or manager may exist. However, the DQ team and its function may not be well developed. Organizations that do not qualify the assessment criteria of level 4 may be classified under this level [5].

Organizations on level four have governance of DQ management which ensures the assignment of roles and responsibilities, ensuring accountability, and providing rewards and incentives. Such organizations may have benchmarked their DQ management activities within or outside the organization. Therefore, DQ is properly managed and aligned with strategic and business goals. Such organizations may have implemented processes which ensure the root-cause analysis of any DQ problems. Moreover, Managed organizations may have developed and documented the entire data architecture [5]. Organizations that do not qualify the appraisal criteria of level 5 are classified under this level.

Organizations on level five continuously enhance DQ management efforts [5]. These organizations can be best examples of how DQ can be managed

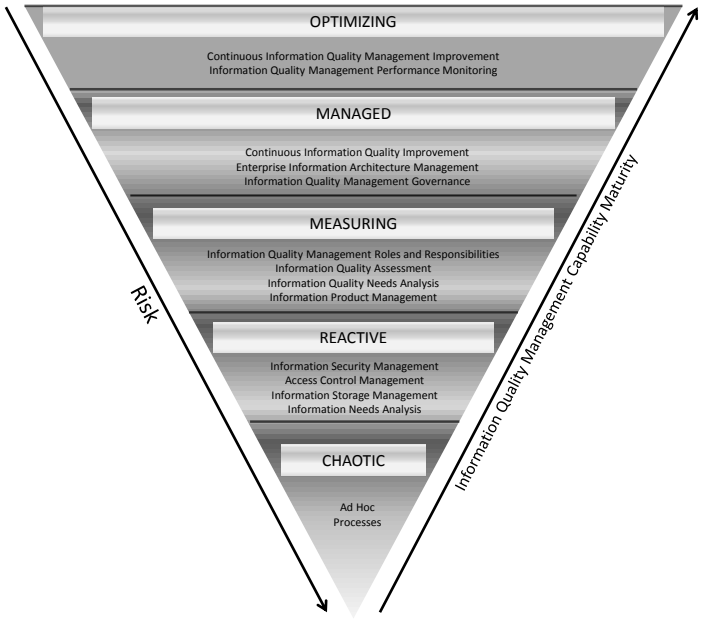


Figure 4.1: *IQM-CMM model [5].*

and optimized.

4.3.2 Accuracy and completeness metrics

Accuracy and completeness DQ dimensions are indicated as relevant dimensions to fulfill the DQ requirements for different tasks [92]. Therefore, in this section, we will identify different metrics to measure the accuracy and completeness level of the data in databases. This section mainly serves by indicating different DQ metrics to quantify the DQ level of the accuracy and completeness DQ dimensions so that financial institutions which are below Level 4 in the IQM-CMM model can benefit by implementing one of the metrics to assess the quality level of their data.

4.3.2.1 DQ metrics

A metric is generally defined as a unit of measurement for specific need [38]. DQ metrics can broadly be classified into objective and subjective measurement. Objective DQ measurement uses simple or complex maths to quantify the DQ level [75, 112]. On the other hand, subjective DQ metrics measure the DQ level based on the personal judgement of individuals at different stages of the data production processes in the organization [76, 92].

Among the objective DQ metrics, we can have application-dependent and application-independent metrics. The application dependent metric only serves for its intended use, in other words, the metric cannot be generally applicable to many purposes [72]. Further, the application-dependent metric can be refined into task-dependent and task-independent. Some examples of task-dependent and independent metrics are provided in Table 4.2. For example, a metric by Heinrich and Klier [55] to quantify the currency level, one aspect of the timeliness dimension which is more dependent on a particular user's demand in a specific business situation. The metric is defined as a probability that an attribute value stored in a database still corresponds to the current state of its real world counterpart at the moment when the DQ level is assessed. This metric depends on the context in which it will be applied. If the timeliness dimension is not critical to the task at hand, then a more relaxed sensitivity measure can be applied. Conversely, if the dimension is very critical, a conservative sensitivity measure is suggested.

On the other hand, application-independent metrics can be generally applicable. For example, the metric which counts the records which violate the entity integrity constraints in the relational model can be applicable to any database. For example, the ratio of the total number of null values for the key fields divided by the total number of rows in the table can indicate the extent to which the integrity constraint rules are violated in any database [21].

Therefore, the type of DQ metric (an objective (an application-dependent or application-independent, a task-dependent or task-independent) or subjective) is an important aspect which should be determined based on the purposes for which the metric is used.

The next aspect of any DQ metric is to determine the scale of measurement. Lack of the proper measurement scale can lead to improper interpretation

and application of DQ measurement results. In particular, this may happen when different DQ dimensions are combined for developing a DQ level indication index in an organization. For example, measuring a DQ error whose measurement scale is a ratio together with a DQ error whose measurement scale is ordinal gives a report which cannot be clearly interpreted. Therefore, determining the appropriate DQ measurement scale is important [125]. Moreover, consistency in the scales of measurement facilitates the comparison of the DQ level across databases.

However, sometimes choosing the specific variables or components to measure can be much more difficult than defining the general metric, which often reduces to the ratio form. Moreover, measuring a similar DQ dimension may require a different approach from organization to organization [75].

Although there are many classifications discussed above, all the DQ metrics can be categorized into two classes: first-generation and second-generation metrics. First-generation DQ metrics deal with data already present in databases. It serves to find erroneous data and correct them within the database [66]. In contrast, second-generation techniques serve to prevent erroneous data from entering into databases.

In this paper, we will discuss different metrics to quantify inaccuracies and incompleteness in the data. The accuracy and completeness dimensions are selected for the reason that they are relevant to many data purposes [76, 149, 92].

4.3.2.2 Accuracy metrics

Data users or customers can have different requirements about the data. Some may require timely data, others may need data to be detailed, and some may require data to be presented in an-easy-to-understand format. However, almost all data users need the data to be correct and provided in an appropriate amount [149, 92]. In other words, many data users need the data they use to be accurate and complete. The main reason is that the two dimensions (and in particular accuracy) are the most basic ones and are considered to be essential to fulfil the DQ requirements of many operational and decision making activities. Accuracy literally means the extent to which the data values agree with their real-world values and completeness indicates the extent to which data are appropriate in amount for the task at hand [149].

DQ Dimension	DQ Metric		Ref.
	Task independent	Task dependent	
Accuracy	$1 - \frac{number\ of\ data\ units\ in\ error}{Total\ number\ of\ data\ units}$ $f(accuracy\ percentage, a\ randomness\ measurement, probability\ distribution)$ $Bayesian\ Network\ Approach$		[112, 75] [38] [127]
Appropriate-amount	$\min[\frac{number\ of\ data\ units\ provided}{number\ of\ data\ units\ needed}, \frac{number\ of\ data\ units\ needed}{number\ of\ data\ units\ provided}]$		[112, 75]
Timeliness	$Q_{curr.} = e^{-decline(A) \cdot age(W.A)}$	$\max[(1 - \frac{Currency}{Volatility}), 0]^s$	[112, 75, 55]

Table 4.2: *DQ metrics from literature*

Conotations for task independent metrics *: f denotes 'function of', $Q_{curr.}$ is the currency level of the data,* A *is an attribute,* w *is an attribute value, age refers to the difference between the instant when DQ is assessed and the instant of data acquisition and $-decline$ refers to the average decline rate of the shelf life of attribute values of the attribute under consideration.*

Conotations for task dependent metrics*: $currency = (delivery\ time - input\ time) + age$, volatility refers to the length of time over which the data remains valid, delivery time refers to the time at which the data was delivered to the user, input time refers to the time at which the data was received by the system, age refers to the age of the data when it was first received by the system and the exponent s is task dependent and used to control the sensitivity of timeliness.*

Measuring the accuracy level of a set of data elements can be a simple or complex process depending on the clarity of the domain values, the existence of real values to compare to and the business situation. For some attributes the domain values can be clearly defined or the real values exist and can easily be attained. In a similar way, it may be easy to determine the costs of inaccurate data for a particular business function.

For example, measuring the accuracy of the GENDER attribute is easy because the two possible values are male and female. In addition, a value to this attribute can only be either correct or not. Similarly, the BIRTHDATE attribute is easy to measure considering the persons exist and know their birthdates. In this case, the persons can inform us the real values for their BIRTHDATE with which the recorded values can be reconciled. This process is usually applicable in e.g. an employer-employee situation.

However, there are situations where measuring data accuracy is difficult because of the need for a standard or real value to measure the accuracy of a recorded value. Mostly, the standard values are unknown or not clearly defined and the real values may not exist. This is frequently the case with data collected in the past, with no supporting evidence still existing. For example, determining the accuracy level for an estimated RISK-AVERSIVE behaviour of a customer is difficult because it is an estimate and doesn't have a real value to compare with. Similarly, there may be a situation where a customer's name has two or more alternative spellings in different documents [125].

Considering different aspects of the accuracy dimension, many metrics have been proposed. One of the simplest metrics to measure data accuracy is the simple ratio for either error rate or accuracy percentage by Pipino et al. [112]. The error rate is defined as the number of incorrect records divided by the total number of records. The accuracy percentage indicates the number of correct records divided by the total number of records. In order to calculate the error rate or the accuracy percentage, one needs to have a precise definition of what is considered to be correct or wrong for the data. In addition, as databases may maintain millions of records, it is difficult and inefficient to count all the errors in the entire database. Therefore, a sampling technique is vital to take appropriate samples to measure accuracy in specific databases. Although this metric measures the basics, it does not measure all aspects of data accuracy.

More recently, Fisher et al. [38] proposed an accuracy metric by changing

the simple ratio scale to a vector approach which includes percentages, a randomness measure, and a probability distribution. The metric combines a simple ratio, number of cells in error to total number of cells, with a randomness measure computed using the Lempel-Ziv complexity measure algorithm[2]. This algorithm is used to differentiate whether the errors in a database are random or systematic in nature. Once the randomness of the errors is determined, a probability distribution can be used to address different managerial questions.

Similarly, Han et al. [52], assessed accuracy in three phases. The first phase identifies the context or highly relevant data sources with which the data are to be compared. For identifying the context, they use a q-gram metric space. The second phase extracts the most approximate data values for different data sources using a vote-fusion policy [73]. And finally, accuracy is measured using Jensen-Shannon divergence (JSD).

As mentioned above, one difficulty to measure accuracy is finding reference data which are considered or known to be accurate. To alleviate this difficulty, Sessions et al. [127] developed an Accuracy Assessment Algorithm (AAA) for measuring the accuracy level using bayesian networks[3]. The AAA assesses the quality of the data with no prior knowledge of the dataset. The Law Enforcement (LE) datasets was used to develop the algorithm because of the difficulty of assessing the accuracy of LE data without sampling and checking the data against other reliable sources. The algorithm determines the accuracy level of the data by assessing the effect of erroneous data on the algorithm, i.e. the effect of erroneous data on the algorithm is the dependent variable to be measured. The authors set four different significance levels between 0.05 and 0.00005 to a higher cross-entropy. The significance 0.05 level indicates stronger set of network dependencies and edges. The algorithm learns different correlations in the data with fully connected networks (all nodes/fields connected via edges). Then it eliminates the edges that do not exist. This algorithm is found to have difficulty in eliminating non-correlated edges under inaccurate datasets. It was therefore hypothesized that the accuracy level of datasets can be approximated by examining the number

[2]Lempel Ziv is an algorithm for lossless data compression. In fact, it is not a single algorithm, but a whole family of algorithms, stemming from the two algorithms proposed by Jacob Ziv and Abraham Lempel in 1978 [158].

[3]A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) [127].

of learned edges, or average degree of nodes, in a network at the default significance level of 0.05.

On the other hand, different statistical methods have been used to identify outliers or anomalies based on a certain standard deviations. Maletic and Marcus [82] found that five standard deviations from the mean were optimal in order to detect outliers. Similarly, clustering methods have been used to identify outlier data values based on some distance measures such as Euclidian distance. Then, based on the cluster results, the method can detect clusters of outlier data points [28]. Similar to clustering, pattern recognition approaches also identify records with similar characteristics and categorize them into groups [58]. Records which do belong to the patters are grouped based on different measures such as distance from mean.

Association rules can also be used to find different associations (e.g. if the COUNTRY is Belgium, the CITY must not be London) in different records. These associations can help to determine the outliers, such as, records which are often associated with themselves but for some reasons they are not. Alpar and Winkelstrater [1] used association rules to determine the accuracy level of an accounting data set. They first mined different association rules on cost accounting transactions. Based on those association rules, a transaction can be classified as correct or wrong. Finally, they incorporated the cost of misclassification using expert opinions. For example, experts identified that the cost of correcting erroneous data can be five to ten times higher than costs of preventing DQ errors. Although the author included the impact of poor DQ in their model using expert opinion, measuring the impact of poor DQ is different from measuring the DQ level. In contrast, Laure-Berti-Equille developed a method which predicts the cost or impact of poor quality data on the quality of discovered association rules [8]. The author further suggested the merging of DQ scores for DQ dimensions relevant to the application considered in order to mine the new association rules.

4.3.2.3 Completeness metrics

Completeness is another important DQ dimension. It is often defined as the extent to which all data values are present or not missing in a specific dataset [149].

Different methods to measure and replace missing values have been discussed in literature. Deleting all the records with missing values in a file can be

identified as the first method to treat missing data [117]. Indeed, this is the simplest way to deal with completeness issues, but obviously it is not universally applicable because of several reasons. The main reason is the high possibility of losing a considerable amount of knowledge with disregarded observations. Similar to disregarding all missing values, replacing them with null or constant values was suggested as another approach to handle missing data. This approach may solve some database problems, for example it can solve referential integrity issues [20]. However, like disregarding the missing values, replacing them with null values creates the loss of a considerable amount of data or significantly wrong distributions if the missing values are manifold. Therefore, the applicability of these two ways of replacing missing data is minimal. Another approach is to replace the missing values using simple estimates. For numeric types of data, usually the mean of non-missing values of an attribute can be used to replace the missing observations. For categorical data, usually the mode of the attribute value can be used to replace the missing observations.

More advanced, Li [77] proposed estimating and replacing missing categorical data using a Bayes method. The author suggested using two approaches to estimate and replace the missing values. The first approach is to replace the missing value of an attribute with the value which has a maximum posterior probability calculated using all recorded values of an attribute. This approach is adopted from the method which the simple Bayes classifier uses to assign a class value. The second approach suggested is to replace the missing value with a value that is selected with probability proportional to the estimated posterior distribution. Similarly, Shen and Chen [133] proposed a method using Association Rule Mining (ARM) techniques to estimate and replace missing values. Likewise, Horton and Kleinman [59] provide a survey of different statistical methods to estimate and replace missing observations.

Although most of the methods deal with how to estimate or impute missing values, there are few studies on how to measure or estimate the impact of missing data on operational and strategic business activities.

The simplest approach to measure the number of missing values proposed is the simple ratio. Similar to the accuracy dimension, completeness can be measured using a simple ratio of number of missing fields or records divided by total number of fields or records [112]. Although calculating the ratio seems simple, it can't measure all aspects of the completeness dimension.

More advanced, Horton and Kleinman [59], Parssian [104] and Parssian et al. [102] provided different machine learning techniques for measuring missing values and their impact on the quality of strategic decisions.

Some examples of the different metrics discussed above are given in Table 4.2.

4.4 Methodology

This section will elaborate on the research questions and the approach used for assessing the maturity level of the DQ management activities in five financial institutions.

4.4.1 Research Context and Questions

Prior research has indicated that many operational and decision support activities in financial institutions are negatively impacted by poor quality data [42]. It has also been argued that inaccurate and incomplete data may adversely affect the competitiveness of an organization because operational activities and strategic decisions are based on the analysis and interpretation of the data available [89]. Generally, it was predicted that organizations that ignore information and DQ management will have difficulty in maintaining their status in business environment because of the many impacts that poor DQ may have on their business activities [122]. However, although DQ should be an integral part of any information and communication technologies (ICT), many companies either do not have a concept of DQ or ignore DQ problems because of lack of perceived values [10, 43]. Therefore, most DQ projects are reactive; only addressing DQ issues as they occur but not addressing the underlying process issues which created the DQ problems [43]. As a result, most DQ initiatives only give sub-optimal benefits [43]. However, it is clear that failure to identify and continuously improve weak internal processes creates many of the DQ problems in many organisations [89]. Therefore, the adoption of DQ improvement programmes and conducting DQ management activities in an organized way have been suggested to minimize DQ issues and their impact on business performance (e.g. high costs, low productivity and profitability) [27]. In line with this, the major aim of this research is to determine the key and general process areas for improvement by assessing

their maturity levels, and to suggest a framework based on the best DQ management practices in the financial institutions with relatively high DQ management maturity levels to enhance DQ management activities in the financial institutions with lower maturity levels.

4.4.2 Empirical Study

4.4.2.1 Case study

Case study is defined as an empirical inquiry that investigates different situations within their real-life context [157]. Case studies provide an opportunity for the researcher to gain a clear view of the research problem and may facilitate describing, understanding and explaining the different facets of the problem. Case study can be used to describe or explain certain phenomena so that casual relationships may be determined and different theories can be developed. As such, there are many case study types such as descriptive, exploratory and instrumental case studies. Similarly, there are many approaches to conduct a case study such as positivist, interpretative or critical. However, non of the types and the approaches is superior given that the adoption of one of the types or the approach highly depends on the research objective. Therefore, we adopted an exploratory case study which can be used to investigate causal relationships and it is highly characterized by "how" and "why" research questions [136].

4.4.2.2 Data collection

The number of case studies to be conducted is not always clear. Therefore, prior research analyzed the suggestions in the literature and proposed that 2 to 4 as a minimum and 10 to 15 cases as the maximum [110]. In this research, 5 exploratory case studies were conducted to answer the research questions in Section 4.4.1. To this end, we selected five major Belgian financial institutions where credit risk management represents a key activity. Credit risk analysts who use similar data types to accomplish their daily tasks were selected as subjects for the interview. The fact that all respondents are credit risk analysts from the same sector decreases the variability and the effect of regulatory compliances that different sectors must fulfil. The interview was conducted with physical presence and took approximately $1\frac{1}{2}$ hours.

4.4.2.3 Interview Questionnaire

A structured interview was developed based on the IQM-CMM's appraisal criteria. The interview questionnaire consists of 48 major questions, not including the sub-questions under each of the major questions. Each of the questions assessed the satisfaction level of a specific CSF. The full questionnaire is shown in the Appendix. The questionnaire was validated for its relevancy and understandability by two data governance officials in two Belgian banks, who are not included into the final study.

4.4.2.4 Qualitative analysis

A qualitative analysis is used to understand the cases and to infer the underlying theories. This analysis technique generalizes the study or develops a theory by comparing the results from the empirical case studies with previously developed theories [157]. As such, we compared the results of the actual case studies with the IQM-CMM model theories.

4.5 Results and Discussions

In this section, the IQM-CMM capability maturity assessment results for the five financial institutions will be presented. Thus, the key areas for improvement will be determined and best practices will be suggested based on the DQ management activities in the financial institutions with relatively high IQM-CMM capability maturity levels.

4.5.0.5 Maturity Assessment Results

The IQM-CMM level is determined by the extent to which the CSFs are satisfied. The ordinal scale (Fully, Partially and Not Satisfied) is used to assess the level of satisfaction of each CSF (see Table 4.3).

Analyzing an ordinal scale using statistical techniques requires a clear translation of the ordinal scale to numbers [40]. However, the ordinal scale "partially satisfied" cannot be quantified in numbers as it can represent a CSF's satisfaction level less than, equal to or greater than 50%. As a result,

Rating	Description	Comparable SCAMPI a, v1.2 Rating
Not Satisfied (NS)	There is no documentation and there is limited or no evidence to confirm the implementation	**Not implemented (NI)** • Direct artefacts are absent or judged to be inadequate. • No other evidence (indirect artefact or affirmations) supports the practice implementation. • One or more weaknesses are noted.
Partially Satisfied (PS)	Some documentation exists, however there is inconsistent implementation through ad-hoc processes	**Partially Implemented (PI)** • Direct artefact are absent or judged to be inadequate. • One or more indirect artefact or affirmations suggest that some aspects of the practice are implemented. • One or more weaknesses are noted. OR • One or more direct artefacts are present and judged to be adequate. • No other evidence (indirect artefact, affirmations) supports the direct artefact). • One or more weaknesses are noted.
Fully Satisfied (FS)	Entirely documented, consistently implemented, effective and efficient, with above expectations results, utilizing industry best practices.	**Fully implemented (FI)** • One or more direct artefacts are present and judged to be adequate. • At least one indirect artefact and/or affirmation exists to confirm the implementation. • No weaknesses are noted.

Table 4.3: *The standard appraisal criteria to determine "Fully", "Partially" and "Not satisfied" CSFs as adopted from SCAMPI 2006 [141].*

Woodall et al. [156] developed two measures to aggregate the values for the Critical Success Factors (CSFs) into maturity levels when they determine the IQM-CMM rating of ten asset management organizations. We borrowed these measures in order to approximately place the five financial institutions in one of the IQM-CMM levels. The notations in Table 4.4 are used to illustrate the measures.

$$F_M = \frac{cf}{N}$$

$$FP_M = \frac{cf + cp}{N}$$

$$IQMCMM_I = \sum_{M=1}^{5} \frac{FP_M + F_M}{2}$$

If F_M is greater than 50% and FP_M is greater than 80%, then the level is said to be matured. Finally, The value of $IQMCMM_I$ is used to approximately place the institutions in one of the maturity levels.

Denotation	Meaning
FS	Fully Satisfied
PS	Partially Satisfied
NS	Not Satisfied
M	The level in IQM-CMM
KPA	Key Process Area
$cf - KPA$	The number of fully satisfied CSFs in each KPA
$cp - KPA$	The number of partially satisfied CSFs in each KPA
cf	The number of fully satisfied CSFs in each M
cp	The number of partially satisfied CSFs in each M
N	The total number of CSFs in each M
n	The total number of CSFs in each KPA
F_M	The satisfaction level of each M using only the cf
FP_M	The satisfaction level of each M using both cf and cp
$IQMCMM_I$	The IQM-CMM rating for a financial institution

Table 4.4: *Notations*

If $IQMCMM_I < 2$, the IQM-CMM level is Chaotic

If $3 > IQMCMM_I \geq 2$, the IQM-CMM level is Reactive

If $4 > IQMCMM_I \geq 3$, the IQM-CMM level is Measuring

If $5 > IQMCMM_I \geq 4$, the IQM-CMM level is Managed

If $IQMCMM_I \geq 5$, the IQM-CMM level is Optimizing

4.5.1 CASE A

CASE A is approximately placed on IQM-CMM level 3 (Measuring) as shown in Figure 4.2 because it only partially satisfied five KPAs such as IQ assessment, IQM Roles and Responsibilities, IQM Governance, Continuous IQ Improvement, and IQM Performance Monitoring (see Table 4.6).

4.5.2 CASE B

The result in Figure 4.3 indicates a low level of DQ management capability maturity, thus approximately placing CASE B on IQM-CMM Level 2 (Reactive). CASE B is reacting to DQ problems whenever they occur. DQ management

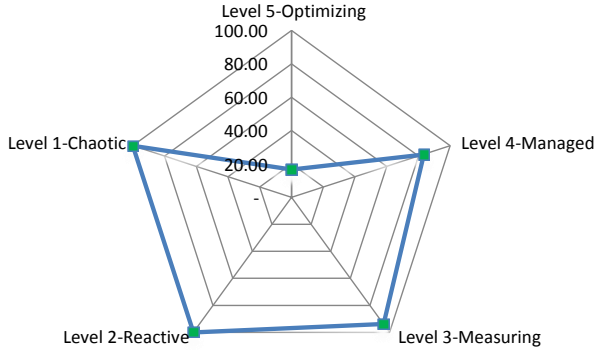


Figure 4.2: *Maturity level- CASE-A*

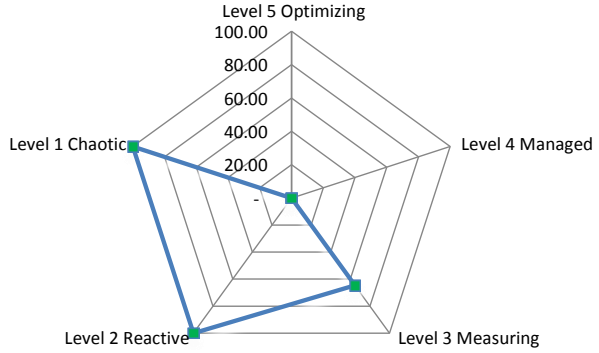


Figure 4.3: *Maturity level - CASE B*

activities are at their early phase. This can also be explained by the fact that CASE B is on the process of merging with other institution. The institution didn't satisfy many of the KPAs such as IP Management, IQ Needs Analysis, IQ Assessment, IQM roles and responsibilities, IQM Governance, Enterprise Information Architecture, Continuous IQ Improvement, IQM Performance Monitoring, and IQM Continuous Improvement (see Table 4.6).

4.5.3 CASE C

The result in Figure 4.4 places CASE C on IQM-CMM level 3 (Measuring). It either partially or not satisfied nine KPAs such as IP Management, IQ Needs Analysis, IQ Assessment, IQM Roles and Responsibilities, IQM Governance,

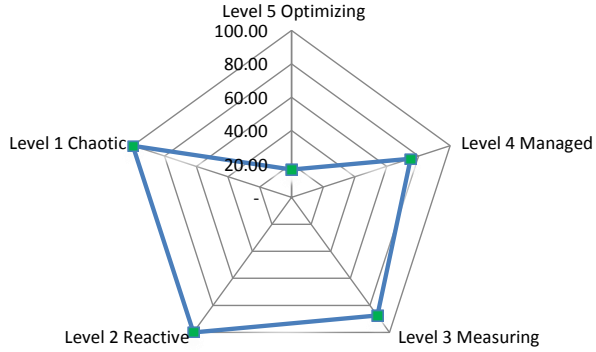


Figure 4.4: *Maturity level - CASE C*

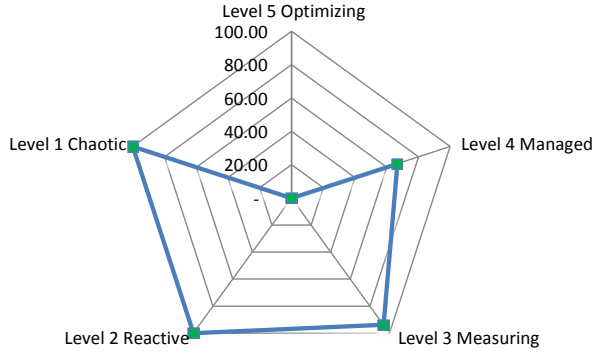


Figure 4.5: *Maturity level - CASE D*

Enterprise Information Architecture Management, Continuous IQ Improvement, IQM Performance Monitoring, and IQM Continuous Improvement (see Table 4.6).

4.5.4 CASE D

The result in Figure 4.5 places CASE D on IQM-CMM level 3 (Measuring). Similar to CASE C, CASE D either partially or not satisfied nine KPAs such as IP Management, IQ Needs Analysis, IQ Assessment, IQM Roles and Responsibilities, IQM Governance, Enterprise Information Architecture, Continuous IQ Improvement, IQM Performance Monitoring, and IQM Continuous Improvement (see Table 4.6).

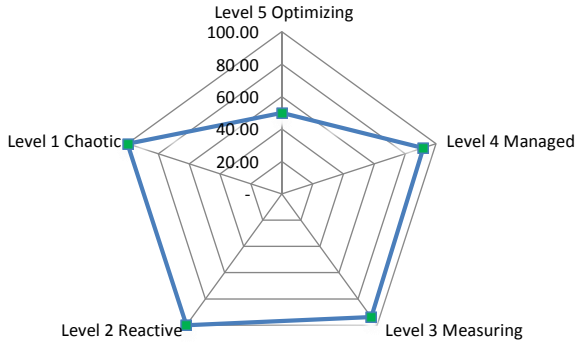


Figure 4.6: *Maturity level - CASE E*

4.5.5 CASE E

The result in Figure 4.6 places CASE E on IQM-CMM level 4 (Managed). Most of the DQ management activities are matured enough to create a good DQ level. We believe that CASE E has achieved the Managed level exceptionally from the other four cases because of its DQ measuring and analysis activities. The DQ assessment activities are being conducted differently than observed in other institutions. In general, it was noted that some of the DQ measuring activities are outsourced to an external DQ management company. This obviously indicates the institution's capability in developing a clear economic model for DQ improvement activities unlike the other four.

In this institution, DQ problems are identified and mapped to the corresponding DQ dimensions with the agreements of all data stakeholders. Thus, accuracy, completeness, timeliness and consistency DQ dimensions are selected to assess the DQ level. Furthermore, clear definitions of the dimensions are reconciled and documented. Two databases are prioritized to be continuously monitored for their quality. These are customer and credit risk (FERMAT) databases.

Ensuring the DQ level of the customer database is outsourced to an external DQ management company. For the customer database, three DQ dimensions are mapped with the recurring DQ errors. Accuracy includes the wrong spelling of customer and address street names. These problems are identified as not recurring. However, the consistency issues which include duplicate entries, the completeness issues which include missing birth dates and the timeliness issues which include outdated addresses are accounted for 90% of

the DQ errors in the identified database. The DQ metric used to measure the amount of DQ problems in the databases is often "database bashing" using set operators (INTERSECT, UNION, UNION ALL and others). The database in reference is identified as containing true or standard values for comparing it to the other databases. The external company provides the DQ assessment report as exemplified in Table 4.5. The reports from the metrics include a unique key identification of each record and quality indicator scores. For example, if a record has a spelling error and missing data, the score for the record is less than 100%. Similarly, if two records are identified to have 50 or more % of similarities, they are considered to be duplicate records. For example, line 1 and line 3 are duplicate records for one customer but with different addresses. The correct address is indicated by a 100% score for line A. Therefore, based on the reports, the high score records will be kept, yet, the low score records will be removed. These DQ measuring processes are conducted by the external DQ management company on a quarterly base and the costs are estimated to be 150000 euros per year.

For credit risk modeling, DQ is defined according to whether the value in the credit request form is real and whether the form includes all the necessary information to build three models such as credit norm assessment, budget analysis and credit scoring. Each of these models independently determines the credit worthiness of the credit risk applicant. Therefore, all the elements of the models or important attributes are identified as the critical data. The DQ checks are conducted in three layers. The first layer mainly ensures whether all the required data are present and it mainly involves business analysts from credit risk and marketing departments. The second layer is mainly conducted through business processes controls and software applications. Finally, the third layer DQ checks are being conducted by the audit department and they include identifying the sources of discrepancies between similar reports from different departments. In general, business rules are mostly pre-defined and used as DQ firewall, they are also subject to different updates to integrate the new needs. Although, some of the DQ activities are manual and involve business experts, the institution managed to build trust for the data and employee satisfaction.

Defining and enforcing DQ management controls using a layered approach is assessed to be a best practice for DQ improvement [13]. In addition, the cost of all DQ related activities could be motivated because the economic model (e.g. increased costs 8-12% of revenue because of poor quality data) for a high DQ level has been defined and accepted by the management.

CASE E's customer data			DQ reporting from DQ company	
Internal key	Family name	Address	Unique Key	Score
1	De Roy	Vlamingenstraat 37, B-3000, Leuven	A	100
2	Talbom	Emile Jacqmainlaan 30, B-1000, Brussels	B	100
3	De Roy	Havenlaan 50, B-1080 Brussels	A	50

Table 4.5: *Reporting from DQ checks by the external company for CASE E*

Level	**KPA**	**CASE**					**Shared Problem and Success Areas**	**For specific cases**
		A	**B**	**C**	**D**	**E**		
Optimizing	IQM Continuous Improvement	N	N	N	N	N	Shared by CASE A, B, C, D and E • There is a plan to continuously improve DQ but there is no documentation for activities planned to improve DQ management and the DQ level	
	IQM Performance Monitoring	P	N	P	N	P	Shared by CASE A, B, C, D and E • DQ management activities are not benchmarked against DQ management best practices in or outside the institutions • DQ management evaluation metrics do not exist • DQ management key performance indicators are not identified or determined	Specific to CASE B • DQ problem analysis and reporting activities do not exist
Managed	Continuous IQ Improvement	P	N	P	P	P	Shared by CASE A, B, C, D and E • The cost of poor DQ is implicitly understood but not estimated • The benefits of DQ improvement initiatives are not explicitly known • Root-cause DQ problem analysis is not in place	Specific to CASE B • Changes of DQ level improvement are not planned
	Enterprise Information Architecture Management	P	N	P	P	P	Shared by CASE A, B, C, D and E • There is no dynamic generation of DQ rules in place • Some master data are not centrally consolidated, stored and managed • Single Version of The Truth (SVOT) has not been established	
	IQM Governance	P	N	P	P	P	Shared by CASE A, B, C, D and E • DQ management roles and responsibilities are not clearly indicated • DQ is not a criterion for job performance review • DQ related rewards and incentives are not in place	Specific to CASE B • DQ management needs are not aligned with business goals
Measuring	IQM Roles and Responsibilities	P	P	P	P	P	Shared by CASE A, B, C, D and E • Scripted DQ cleansing methods do not exist • Standard procedures are not in place for DQ problem classification • There is no DQ related training or education in place	Specific to CASE B • There is no standard procedure to solve DQ problems
	IQ Assessment	P	P	P	P	F	Shared by CASE A, B, C and D • DQ level assessment is being conducted using some business rules yet the business rules are not comprehensive enough to measure every aspect of DQ.	
	IQ Needs Analysis	F	P	P	P	F	Shared by CASE A, B, C, and D • DQ dimensions are not fully prioritized for improvement, there is no clear documentation on how to improve DQ dimensions	
	IP Management	F	P	P	P	F	• Information product visualization, configuration and taxonomy is not fully documented • There is no standard procedure to record metadata	
Reactive	Information Security Management	F	F	F	F	F	• There is security classification of information products • There is a standard procedure to transmit sensible information • There is a standard procedure to dispose sensible information	
	Access Control Management	F	F	F	F	F	• Authorization of user accounts is documented • Authentication of users privilege is regularly reviewed.	
	Information Storage Management	F	F	F	F	F	• Information is stored in dedicated areas • There are standard backup and recovery procedures	
	Information Needs Analysis	F	F	F	F	F	• Conceptual, Logical and Physical modeling are in place	
Chaotic	No key process areas	F	F	F	F	F		

Table 4.6: *Summary of the DQ management activities maturity assessment for all financial institutions*

4.5.6 Critical Success Factors (CSFs)

The satisfaction level of each Critical Success Factor (CSF) is determined using the interview questionnaire which is based on the appraisal criteria of each CSF. Figure 4.7 and Table 4.7 show the extent to which each CSF is satisfied by the five financial institutions. The last three columns of the Table show the number of times each CSF is fully, partially or not satisfied respectively. The percentages of these columns are used to visualize the satisfaction level as displayed in Figure 4.7. The dark green (black), medium green (gray) and light green (white) bars indicate the number of institutions that fully, partially and not satisfied each CSF respectively. The black bars are saturated at the lower IQM-CMM levels (Reactive and Measuring), implying that most of the DQ management activities in the Managed and Optimizing levels either do not exist or they are at their early phase. Yet, in the Reactive level, the CSF stakeholder management is only partially satisfied in the majority of the cases. This implies that all data stakeholders, their relationships, their roles and responsibilities are not identified and documented. In general, stakeholder taxonomy is not well developed. Similarly, in the Measuring level, there are no DQ related education and trainings in all the institutions, thus, the corresponding CSF is never satisfied. Likewise, although metadata management is present in all of the cases, it is not at the level it needs to be. For example, there is no a standard procedure to record metadata. Many of the metadata recorded are not clear, thus they are not used. Not surprisingly, many CSFs in the Managed and Optimizing levels, including DQ risk management and impact assessment, DQ management cost-benefit analysis, physical, application, information and enterprise tier managements are never fully satisfied.

4.5.7 Key Process Areas for improvement

This Section will present the Key Process Areas (KPAs) which need improvement based on the satisfaction level of each KPA. The satisfaction levels of the KPAs are determined using the following formulas.

$$F_{KPA} = \frac{cf - KPA}{n}$$

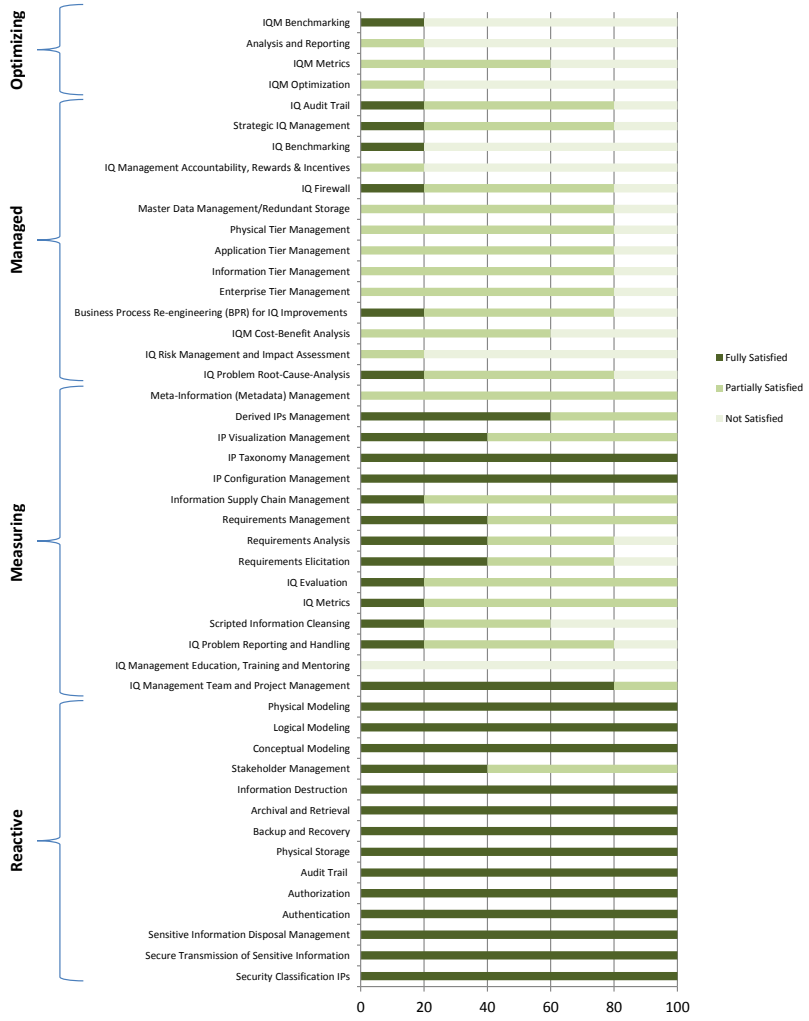


Figure 4.7: *The total satisfaction level of each CSF's in the five financial institutions.*

Level	KPA	CSF	CASE					F. Count	P. Count	N. Count
			A	B	C	D	E			
Optimizing	IQM Continuous Improvement	IQM Optimization	N	N	N	N	P	0	1	4
	IQM Performance Monitoring	IQM Metrics	P	N	P	N	P	0	3	2
		Analysis and Reporting	N	N	N	N	P	0	1	4
		IQM Benchmarking	N	N	N	N	F	1	0	4
Managed	Continuous IQ Improvement	IQ problem Root-Cause-Analysis	P	N	P	P	F	1	3	1
		IQ Risk Management and Impact Assessment	N	N	N	N	P	0	1	4
		IQM Cost-Benefit Analysis	P	N	P	N	P	0	3	2
		Business Process Re-engineering (BPR) for IQ Improvements	P	N	P	P	F	1	3	1
	Enterprise Data Arct. Management	Enterprize Tier Management	P	N	P	P	P	0	4	1
		Information Tier Management	P	N	P	P	P	0	4	1
		Application Tier Management	P	N	P	P	P	0	4	1
		Physical Tier Management	P	N	P	P	P	0	4	1
		Master Data Management/Redundant Storage	P	N	P	P	P	0	4	1
		IQ Firewall	P	N	P	P	F	1	3	1
	IQM Governance	IQ Management Accountability, Rewards & Incentives: IQ is Everyone's Responsibility	P	N	N	N	N	0	1	4
		IQ Benchmarking	N	N	N	N	F	1	0	4
		Strategic IQ Management	P	N	P	P	F	1	3	1
		IQ Audit Trial	P	N	P	P	F	1	3	1
Measuring	IQM Roles and Responsibilities	IQ Management Team and Project Management	F	P	F	F	F	4	1	0
		IQ Management Education, Training and Mentoring	N	N	N	N	N	0	0	5
		IQ Problems Reporting and Handling	P	N	P	P	F	1	3	1
		Scripted Information Cleansing	P	N	N	P	F	1	2	2
	IQ Assessment	IQ Metrics	P	P	P	P	F	1	4	0
		IQ Evaluation	P	P	P	P	F	1	4	0
	IQ Needs Analysis	Requirements Elicitation	F	N	P	P	F	2	2	1
		Requirement Analysis	F	N	P	P	F	2	2	1
		Requirements Management	F	P	P	P	F	2	3	0
	IP Management	Information Supply Chain Management	P	P	P	P	F	1	4	0
		IP Configuration Management	F	F	F	F	F	5	0	0
		IP Taxonomy Management	F	F	F	F	F	5	0	0
		IP Visualization Management	F	P	P	P	F	2	3	0
		Derived IPs Management	F	P	P	F	F	3	2	0
		Meta-Information (Metadata) Management	P	P	P	P	P	0	5	0
Reactive	Information Security Management	Security Classification IPs	F	F	F	F	F	5	0	0
		Secure Transmission of Sensitive Information	F	F	F	F	F	5	0	0
		Sensitive Information Disposal Management	F	F	F	F	F	5	0	0
	Access Control Management	Authentication	F	F	F	F	F	5	0	0
		Authorization	F	F	F	F	F	5	0	0
		Audit Trial	F	F	F	F	F	5	0	0
	Information Storage Management	Physical Storage	F	F	F	F	F	5	0	0
		Backup and Recovery	F	F	F	F	F	5	0	0
		Archival and Retrieval	F	F	F	F	F	5	0	0
		Information Destruction	F	F	F	F	F	5	0	0
	Information Needs Analysis	Stakeholder Management	F	P	P	P	F	2	3	0
		Conceptual Modeling	F	F	F	F	F	5	0	0
		Logical Modeling	F	F	F	F	F	5	0	0
		Physical Modeling	F	F	F	F	F	5	0	0
Chaotic										

Table 4.7: *Each CSF's maturity assessment in the five financial institutions. Although the IQM-CMM model suggests to move the Firewall CSF to Level 5, there are no differences in the maturity levels of the financial organizations between both cases when Firewall CSF is included either in Level 4 or Level 5.*

$$FP_{KPA} = \frac{cf - KPA + cp - KPA}{n}$$

If F_{KPA} is greater than 50% and FP_{KPA} is greater than 80%, then the corresponding KPA is said to be fully satisfied. If F_{KPA} or FP_{KPA} are greater than zero but less than or equal to 50% and 80% respectively, then the corresponding KPA is said to be partially satisfied. If F_{KPA} and FP_{KPA} are zero, then the corresponding KPA is said to be not satisfied. Therefore, all the KPAs which are either partially satisfied or not satisfied by most of the financial institutions are identified to be process areas for improvement (see Table 4.6 and Figure 4.7).

4.5.7.1 IP Management

This process area assesses whether the following maturity indicators are present [5].

- Whether internal information/external suppliers have been identified and documented
- Whether information is never copied manually
- Whether the information flow has been documented
- Whether metadata taxonomy has been developed and documented
- Whether all information products have the same look and feel
- Whether information product dependencies, aggregations, compositions and associations have been fully documented

Although most of the financial institutions are already managing their information as a tangible product, most of these critical success factors are only partially satisfied by the majority of the cases. Because of the lack of some standardization in some of the critical success factors, for example, metadata management, this process area is only partially satisfied by the majority of the cases.

4.5.7.2 IQ Needs Analysis

This process area assesses whether the following maturity indicators are present [5].

- Whether DQ dimensions have been clearly defined, documented and communicated to all stakeholders
- Whether DQ dimensions have been prioritized based on their criticality
- Whether DQ requirements have been collected from a statistically valid representative sample of the critical stakeholder
- Whether DQ dimensions have been mapped to the corresponding entities in the information model
- Whether minimum and desirable levels of DQ have been specified
- Whether DQ requirements are effectively communicated to all stakeholders

In most of the financial institutions, DQ needs analysis is partially satisfied. Although the institutions are in the process of identifying relevant DQ dimensions, mapping those DQ dimensions with the recurring DQ challenges and defining DQ metrics, these activities are not very organized. As such the organizations are unable to fully assess the current quality level of their data. In addition, the minimum DQ level threshold is not defined.

4.5.7.3 IQ Assessment

This process area assesses whether the following maturity indicators are present [5].

- Whether surveys are used to assess information consumers's subjective perceptions of DQ
- Whether the quality of information products is assessed
- Whether qualitative or quantitative DQ key Performance Indicators (KPIs) are identified
- Whether qualitative or quantitative metrics exist to measure those KPIs

Although all the financial institutions are very aware of the importance of identifying different KPIs to measure the quality of their data, the KPIs are not fully and consistently defined. Therefore, comprehensive qualitative or quantitative DQ metrics are not fully realized. In general, we have noticed that the DQ measuring activities in the majority of the organizations are not organized. This is also indicated by the fact that 60% of the cases are at the measuring level in the IQM-CMM model.

4.5.7.4 IQM Roles and Responsibilities

This process area assesses whether the following maturity indicators are present [5].

- Whether DQ governance team or personnel exist
- Whether DQ management project scope and responsibilities are defined
- Whether there is a standard procedure for DQ problems reporting and handling
- Whether there is a standard procedure to classify the DQ problems identified
- Whether DQ related educations and training exist
- Whether there exist DQ metrics to assess the DQ level
- Whether there exist scripted DQ cleansing practices

In most of the financial institutions, the DQ management team originates and operates in the credit risk department. This is motivated by the fact that the percentage of poor DQ is one of the elements to estimate the capital buffer[4]. In other words, the regulatory compliances that the credit risk department must fulfil initiated the DQ management activities. Although the department produces data, it receives most of its data from internal or external suppliers. Yet, DQ checks, if any, for all the data are being conducted in the department. Most of the interviewees described this as a difficult task because they don't have the implicit knowledge concerning how the data are produced. Therefore, they follow an ad hoc procedure of tracing back to the data production process.

[4]"Mandatory capital that financial institutions are required to hold in addition to other minimum capital requirements" [4]

4.5.7.5 IQM Governance

IQM Governance process ensures the existence and maturity of the following activities [5].

- Whether DQ management roles and responsibilities have been transparently and hierarchically defined
- Whether there is a standard procedure or system that ensures stakeholders' accountability for the data they produce
- Whether DQ related rewards and incentives exist
- Whether DQ management is aligned with organizational strategies
- Whether DQ capturing, modification and destruction are recorded and used as audit trial
- Whether DQ benchmarking is being conducted within or outside the organization

Organizational wide DQ management efforts have not been realized in most of the institutions. There are no clear documents which show the roles and responsibilities, and accountability of all data stakeholders. In general, data ownership is not assumed. It is, however, indicated that specifying and documenting the rights and accountability enhances the production of good DQ [43]. Similarly, DQ level benchmarking within or outside the institutions is not yet possible. Likewise, although the capturing, modification and destruction of data are recorded, they are not being regularly analyzed and used as audit trail. Moreover, in most of the financial institutions, DQ management is not yet taken into consideration within the organizational strategies.

However, unlike the four cases, CASE E has implemented an organization wide DQ management effort. It is also using an internal data quality maturity model for benchmarking. Also, a dashboard is being used to visualize and communicate the DQ level to the top managers, and DQ is included in the corporate scorecard to ensure the alignment of DQ management with strategic and business goals. Although there are no DQ related incentives and rewards in place, every data set has an identified owner.

4.5.7.6 Enterprise Information Architecture Management

The Information Architecture Management process ensures the maturity and existence of the following DQ management activities [62].

- Whether Single Version of the Truth (SVOT) is established
- Whether the software architecture provides the necessary support for the information/enterprise tiers
- Whether the hardware architecture provides the necessary support for the enterprise tier
- Whether master data are centrally consolidated
- Whether an automatic or manual DQ firewall exist

Appropriately managing the data architecture is essential in order to enhance the DQ level. Most of the financial institutions' data architecture documents do not show the logical, physical and application tier management flows, and if they do, the flow is too complicated to understand. Furthermore, the architecture document does not clearly indicate how the business processes and work flows are modeled. Yet, it is obvious that process issues contribute to many of the DQ problems. Also, heterogeneous DQ sources are not combined in a single representation. Thus, Single Version of The Truth (SVOT) is not yet possible as business officers are still using fragmented excel files for decisions. Some master data are not centrally consolidated, stored and managed. In addition, although there are some business rules as DQ firewall, they are not comprehensive and sometimes outdated.

4.5.7.7 Continuous IQ Improvement

The process Continuous IQ Improvement should ensure the existence and maturity of the following activities [5].

- Whether there is an economic model for DQ improvement activities (i.e., whether the costs and benefits of DQ improvement activities are estimated)
- Whether root-cause analysis is conducted for DQ problems
- Whether the associated risks of poor DQ is identified

- Whether different changes are planned and documented to improve the DQ level

Any DQ management effort aims at improving the DQ level continuously. After the DQ level has been assessed and evaluated, and problems have been singled out, the next step is aiming to improve the DQ level. The first process to improve DQ is determining the causes and related risks of the DQ problems [145]. However, in most of the financial institutions, DQ problems' root-cause analysis does not exist or it is not comprehensive. Therefore, mitigating DQ problems from their source is still a difficult task. The most practiced approach is correcting the DQ problems as they occur. Similarly, there are no economic models for DQ improvement activities defined. In other words, the costs of poor DQ are implicitly assumed as well as the benefits of high DQ levels. Therefore, it is somehow difficult to suggest and motivate continuous DQ improvement actions, instead, ad hoc DQ problem fixing procedures are being practiced. In addition, although business process re-engineering is intended, there are no explicit documentations which show what changes are planned and how they will be practiced.

4.5.7.8 IQM Performance Monitoring

The IQM Performance Monitoring includes processes which define the key process areas to assess the DQ management activities in place. The DQ management practices should be benchmarked against best practices within or outside the organizations. Therefore, the DQ problem analysis should be included in the reports to management. However, in the five financial institutions there are no DQ management metrics to assess the level of the DQ management practices. In addition, DQ problem analysis and reporting is an ad hoc process.

4.5.7.9 IQM Continuous Improvement

DQ management continuous improvement analyzes the process that institutions plan to change or implemented to enhance their DQ management activities. As such, any DQ improvement changes should be documented. Although DQ improvement changes are planned, they are not explicitly documented.

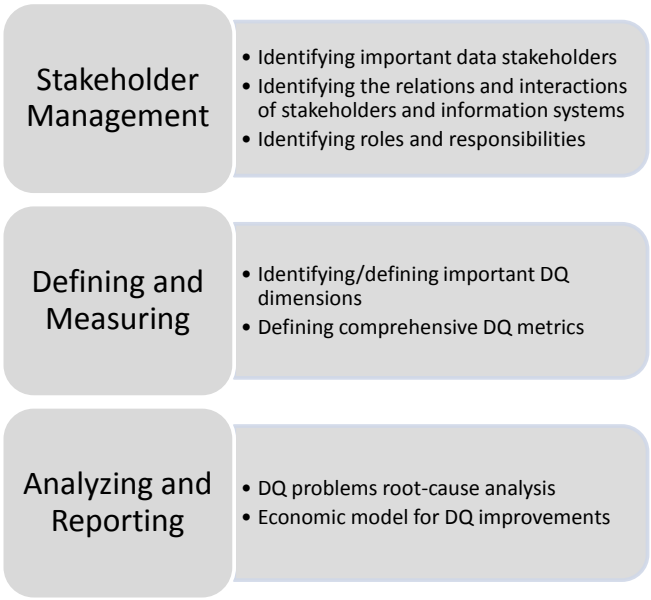


Figure 4.8: *DQ measuring framework.*

4.5.8 DQ Measuring/Assessing Framework

The institution with IQM-CMM level 4 (Managed) is characterized to be different in its DQ assessment practices from the other 4 institutions as discussed in Section 4.5.5. In addition, the three institutions are approximately assessed to have IQM-CMM level 3 (Measuring) where they are still in a phase of DQ assessment and measuring activities. Moreover, previous studies have acknowledged the importance of DQ measuring activities for DQ improvement with a saying "You can't manage what you don't measure" [125]. As such, a standard DQ assessment activity is believed to be relevant in order to enhance the maturity levels. Therefore, this Section presents a three level DQ assessment framework inferred from the mature institution's DQ management activities and literature studies.

Although DQ measurement is not the only DQ management activity, it is an integral part of it. Therefore, a thorough DQ measurement leads to a high DQ level [146, 76, 155].

4.5.8.1 Stakeholder Management

Data production processes include different stakeholders such as data suppliers (who create or collect data), data users (who use the data to accomplish their daily routines) and data product manufacturers (who design, develop and maintain the information system) [57]. The first step in stakeholder management is identifying all the stakeholders, defining their roles and responsibilities, and outlining their relationship. This helps to understand the data flow and production process, and enhances the possibility of tracing back to every step of the data production processes. This, in turn, ensures the identification of recurring DQ problems, and encourages the production and maintaining of high quality data. In addition, a comprehensive and an objective DQ definition by incorporating all the DQ requirements of the stakeholders can be realized. In this way, DQ assessing and improvement activities can be facilitated. Depending on the recurring DQ problem areas and the purposes of DQ measurement, organizations can choose where in the data production stages DQ measurement and improvement should be conducted. Table 4.8 shows the most common data production stages and the benefits of conducting DQ measurement on them [146].

In addition to stakeholder identification, assigning responsibility (i.e., assembling a DQ team) is one of the many DQ activities identified for the DQ success in the matured financial institution. DQ teams or assigned experts can signify the importance of quality data and therefore other employees can associate business re-works and delayed reports with DQ problems instead of considering them as part of the routine business processes. This improves efficiency and effectiveness of business performance. Moreover, DQ activities can be handled in a very organized way.

However, stakeholder management is noted to be the only CSF only partially satisfied by most of the institutions in the Reactive level.

4.5.8.2 Defining and Measuring

The lack of a comprehensive and an objective DQ definition and DQ measuring device has been realized in most of the financial institutions.

Where DQ can be measured	Reasons & Benefits
When new data enters into databases	To understand and avoid DQ challenges at the data entry level.
In databases	Provides an opportunity to correct DQ errors in the entire database.
When the output data is delivered to data users	It is economical to do DQ measurement at this stage because data users usually use a small fraction of the data in databases.
In the entire information chain	This helps to correct all the problems in the entire information chain.

Table 4.8: *Examples of data production stages where DQ can be measured [125].*

Define

In principle, all the data available in any database should be correct or assessed for their quality because all the data are assets of the organization. However, some data are more critical than others for a certain task. DQ management is a costly process and therefore, identifying the important data for each stakeholder is important. In this way, the DQ requirements can be reconciled and the comprehensive DQ definitions can be listed. For example, for credit risk assessment, the employment history, income, monthly expenses and age of a customer are critical data.

Then, the important DQ dimensions, and their comprehensive and objective definitions should be outlined by reconciling all the requirements of important stakeholders to achieve the fitness-for-use level of the data [146]. Therefore, the quality of the data can be improved based on the specified goals. Translating the DQ dimensions into more objective and measurable characteristics is the next step. For example, timeliness and credibility DQ dimensions can be indicated by the age and the data collection method respectively. Similarly, considering the purpose of the task, for a customer database, for example, inaccuracy can be characterized by incorrect BIRTH-DATE values. These values may include missing (NULL and BLANK values), non-valid values or outliers (e.g. 0 or 150), and representationally inconsistent values (e.g. if the correct BIRTH-DATE format is DD-MM-YYYY, the values recorded in the format MM-DD-YYYY are incorrect). In general, all the incorrect values can be summarized under the accuracy dimension. It is also important to note the correlation between the DQ dimensions [92]. The NULL and BLANK values also characterize the completeness DQ dimensions. Similarly, the format differences indicate the consistency problems. For example, in this case, ensuring the accuracy of the BIRTH-DATE values also ensures its completeness

and consistency.

Measure

The first step in DQ measurement is developing or adopting the appropriate DQ metrics to detect the defined DQ problems. DQ metrics or assessing methods range from simple to complex as discussed in Section 4.3. Yet, choosing the appropriate one is an important aspect. For example, the degree of importance of DQ assessing techniques such as data tracking (assessing DQ from data entry to different stages in the information chain), inspection by experts (DQ teams can identify errors at different stages in the data production process), business rules, comparing data values with other standards and user complaints can be different depending on the intended use. Choosing the appropriate DQ measuring device relies on many factors such as the cost of acquiring the device, the types of data sets to be measured and the business requirements. Usually, DQ level (e.g. accuracy and completeness) is measured based on real world data that are credible or presumed to be correct. However, such data do not always exist [125]. Therefore, most financial institutions are basically relying on two methods: business expert analysis and business rules. However, the experts and the business rules are no substitute for real world data. Experts can indicate wrong values as correct or vice versa. Business rules may not be documented in detail, may be fuzzy, outdated or too general. Therefore, the DQ measurement based on experts and business rules may not indicate the real level of the quality of data, but at least it may provide an estimate. Moreover, although generally business experts and business rules can be good DQ measuring devices, they are not comprehensive. However, the representativeness and comprehensiveness of these devices, and any DQ metric for the needed measurement is crucial. Comprehensive and representative DQ metrics should identify all DQ errors in a specified data set, indicate the types of errors and how the errors are distributed, and hints how to correct the errors [125]. For example, a DQ metric which is defined by the ratio $\frac{numberofincorrectvalues}{totalnumberofvalues}$ indicates the percentage of wrong values in the total value set. However, this ratio doesn't indicate whether the errors are distributed randomly or systematically in the database. Yet, this information is important for DQ improvement actions. In addition, the output of the metrics should be clear and easy to enhance improvement actions. However, there is a lack of sophisticated DQ metrics which are comprehensive enough to detect all kinds of DQ errors in the five

financial institutions.

Some other examples of non-representative DQ metrics for the intended task:-

- If we want to measure the accuracy of an INCOME attribute using a business rule, the value of INCOME must be between 0 and 10000, so we can only check whether the values are in the specified domain. However, we cannot check if a value in the specified domain is correct or not. The business rule is unable to identify for example whether the INCOME value of a CUSTOMER A is switched with the INCOME value of CUSTOMER B. Yet, this inaccuracy has a detrimental effect in the loaning decision. For such reasons, the business rule is not comprehensive enough to measure the accuracy of the INCOME account.
- If we want to measure the accuracy of a PRODUCT-LEDGER account with reference to the INVENTORY of the product, we must assume that the INVENTORY shows the real value. However, the INVENTORY cannot be considered as real because of the possible under/over deliveries or theft of physical products.

4.5.8.3 Analysis and Reporting

Analyzing the types of DQ problems and their causes follows the DQ measuring processes. The analysis step should identify the causes of the DQ problems and their impact on business performance. Generally, identifying the causes of poor DQ helps to determine the appropriate methods to solve the problems. Likewise, identifying the costs of poor quality data may help to focus improvement actions. Yet, the costs of poor quality data are often difficult to quantify because they involve both tangible and intangible aspects [125]. However, without clear cost estimates, organizations may not realize the impact of poor quality data on their business performances, therefore, improvement actions can not be priorities. Therefore, the costs of DQ cleansing technologies or human resource assignments may not be motivated. The cost of poor quality data without an active DQ management program in place is estimated to be 20% of the revenue [122]. In addition, identifying the impact of these DQ problems on business performance helps to prioritize DQ management activities. For example, a mistake in one decimal point can have a disastrous effect in medicine prescription. Similarly, an insurance bill sent to the wrong customer may cost the customer an additional sum of money if

he/she paid it without noticing the name differences in the envelop. It may also cost the company when the customer discovers the extra money he/she paid and decides to churn. Yet, there are DQ problems which do not have a disastrous effect. Therefore, analyzing DQ problems in association with their risk on business activities allows the allocation of resources to address the most critical DQ problems.

Statistical process analysis (flow charts, control charts, histograms or scatter diagrams) or pattern recognition technologies are very common to analyze DQ problems [58]. For example, a dummy account can also be introduced in the information systems to identify sources which cause poor DQ [6].

In addition, the way in which the measurement results are reported is also critical for ensuring DQ improvement; clear and goal-oriented reporting is essential [30]. Finally, based on the DQ problem analysis, continuous DQ improvement actions, such as a quick fix or a long term plan, can be suggested. For example, automated DQ cleansing methods can be developed and used quickly, whereas, aligning DQ needs to strategic and business goals may need time to be realized.

To summarize, a good DQ measurement process should include the process of identifying key databases or attributes, defining DQ (e.g. identifying relevant DQ dimensions), developing comprehensive and representative DQ metrics, assigning responsibility for DQ measurement, and determining the data production stages where to measure the DQ level. Most importantly, the types of errors under each DQ dimension should be listed in association with the metrics to be used.

The DQ measuring framework addresses CSF (Stakeholder Management) and KPAs (IQ Needs Analysis and IQ Assessment). The Stakeholder Management CSF identifies critical stakeholders and assigns roles and responsibilities to these stakeholders. As such important DQ dimensions can be identified and mapped to the recurring DQ challenges at the IQ Needs Analysis KPA, and a rigorous DQ assessment and DQ level communication to all stakeholders can be easily realized at the IQ Assessment KPA.

4.5.9 Illustration of the three layered DQ Measuring Framework

In order to successfully answer the business requirement, for example, what is the quality of the Credit Risk Request (CRR) in FERMAT database requires the implementation of the DQ measuring framework. The following bullets illustrate the DQ measuring framework which is being implemented in CASE E.

- The first step includes identifying the stakeholders responsible for the DQ measurement processes and the data production stages where the data need to be assessed for their quality. For example, in CASE E, two business analysts were identified to assess the DQ level of the CRR because they are believed to have knowledge more than other business analysts on how the data are being produced. Similarly, the CRR are assessed for their quality when they are supplied to the data customer (the credit risk department).
- The second step includes identifying the data sets to be measured, identifying the relevant DQ dimensions to assess the DQ level, defining each dimension objectively and identifying the appropriate DQ metrics. For example, in this illustration, Credit Risk Requests (CRRs) are identified to be critical to be measured for their quality, and accuracy, completeness and consistency were identified as relevant DQ dimensions to assess the DQ level of the CRRs. Accuracy is defined as whether the income, the expense and age of the applicant in a specific CRR are valid and correct. Completeness is defined as whether there is no missing or null values for all the predefined required fields. Similarly, consistency is defined as whether the data in the CRR are consistent to each other. (For example, if the sum of the expenses is greater than the income mentioned in the CRR, then the CRR is said to be inconsistent). Finally, comparison of the data values to their domain and accepted values is used to assess the CRR's accuracy level using a pre-written java-code. Completeness is also checked using pre-defined business rules whether all the required fields are complete with their appropriate domain values. Likewise, consistency is checked using pre-defined business rules. If the quality of CRR falls in the acceptable range, the CRR' quality will be labeled as acceptable quality level. Conversely, if the quality of the CRR doesn't fall in the acceptable range, the quality

is labeled as not acceptable and the CRR will be send to further checks and improvements. Thus, the data values in the CRR will be compared with the data-values with their real-world counterparts to confirm the DQ problems identified from the pre-defined business rules.

- The third step includes analyzing and reporting the DQ problems to all the stakeholders of the CRR which include the commercial agents in the marketing department, the marketing department manager, credit risk modelers and the credit risk department manager. The analysis step includes, identifying the causes of the DQ problems by tracing back to the data production process. The causes range from simple recording errors to exaggerated personal judgements. These analyses are being done manually by business experts. Finally, the reports of the analysis include, how many CRRs are checked for their quality, how many of them have passed the first business rule test, how many of them have failed the first business rule test and was checked for their quality using the second method, the causes of the DQ errors if they are known, and the number and types of the identified errors per CRR and in total. This report is being produced every month.

4.5.10 Limitations

Although, a well-organized assessment method has been applied to assess the maturity levels, no quantitative measures have been used. Therefore, the assigned IQM-CMM levels are approximate. In addition, as every interpretive research paradigm, all the results have been interpreted by the author thus, they might have been influenced by any biases of the author. Finally, even though all the interviewees are highly aware of the DQ related activities in the credit risk department and in general in the financial institutions, it is possible that some key players may not be included, especially when considering the fact that the financial institutions have many employees involved in different roles related to DQ.

4.6 Conclusions and Future research ideas

Mature DQ management activities lead to a good DQ level [64]. In addition, the maturity level of the DQ activities is a good indicator of the DQ level in any

organization. Moreover, understanding how the most mature organizations are conducting data and DQ management activities can help in identifying best practices by which organizations with lower maturity levels can improve their DQ management activities.

Therefore, we assessed the maturity level of data and data quality management activities in five financial institutions based on the IQM-CMM maturity model. The maturity level is determined by the extent to which the CSFs are satisfied by the organizations. A questionnaire was developed based on the IQM-CMM appraisal criteria to determine the satisfaction level of each CSF. Finally, the aggregated maturity level was determined. The results indicate that only one of the five financial institutions is in IQM-CMM level 4. Therefore, based on the maturity assessment results, we identified process areas which should get special attention in order to improve the DQ level. In addition, a DQ measuring framework with three steps was inferred from the literature and the DQ measuring activities of the financial institution with a relatively high maturity level so that other financial institutions with lower maturity levels can follow to enhance their DQ levels. Although DQ measuring activities as identified in the framework can be standardized to a certain extent, different business directions and credit risk model elements create large differences in DQ measuring activities. Therefore, this paper gives insights in the maturity level of DQ activities in the four financial institutions, but the results may not be conclusive to the entire sector.

Although the results of this paper indicate the approximate maturity level of different data and DQ management activities in financial institutions and identify different key process areas for improvement using credit risk officers (because most of the DQ management activities are being conducted in credit risk departments), a role gap-analysis can be done to validate the results. As such, different roles in financial institutions can assess the maturity level of different data and DQ management activities. In this way, the key process areas for improvement can be prioritized by reconciling the assessment results from the different roles. In addition, identifying a standardized method to enhance the key process areas can be an interesting idea for future research.

Acknowledgements

All the researches in this PhD thesis were supported by the Odysseus program (Flemish Government, FWO) under grant G.0915.09.

5 Conclusions

5.1 Conclusions and Future research ideas

This last section provides some general conclusions and the limitations of the approach taken.

Corporate databases contain plenty of data linked in various batch or real-time data feeds. The data move and change, the databases continuously re-designed and upgraded. Consequently, information technology gets better, yet, the data quality deteriorates. However, data quality highly determines the value of the data for businesses. Therefore, organizations which do not consider the importance of data quality struggle to survive in the business environment. In this information based economy, well organized information and information quality management activities are the major requirements to a healthy survival of any organization. This PhD thesis mainly focused on the management of information quality in financial sector specifically in the context of credit risk assessment tasks. As such, different views and insights are proposed.

Data quality is broadly defined as "fitness for use" [149]. In line with the definition, Chapter 2 defined DQ and identified different characteristics or dimensions which fulfill the quality requirements of the data for credit risk assessment. Thus, it was possible to assess the quality level of the credit risk assessment data using those identified dimensions. Furthermore, a scorecard

approach was suggested in order to benchmark the quality level of the data within or outside an organization. In addition, different process areas were identified as causes of different DQ problems in the sector. For example, despite the high automation of the information technology being used, manual data entry processes are causing the majority of the DQ problems.

In order to mitigate the DQ problems and their impact on the business activities, different approaches can be pursued. The approaches may range from fixing the DQ problems as they are detected to changing or improving the respective processes that cause the DQ problems. As such, Chapter 3 proposed the inclusion of the assessed quality level of the data as data quality metadata in databases so that decision makers can take into consideration the quality level of the data in their decision making processes. This is believed to have a positive impact on these decision making processes. However, as data quality metadata are additional information, they can create an information overload and may have an adverse effect. Therefore, an experiment was conducted to identify the impact of these metadata. The results indicated that data quality metadata can be useful for decision making processes depending on the characteristics of the decision makers and the task context.

Although quick fixes to DQ problems as they are identified can mitigate some DQ problems, this doesn't prevent the repeated occurrence of such problems. Yet, improving different data related processes, and installing mature data and DQ management activities are indicated to mitigate DQ problems from their source and ensure that those DQ problems will not occur again. Therefore, Chapter 4 assessed the maturity level of different data and DQ management processes in the sector in order to identify weak process areas which should be improved to create an acceptable DQ level and to identify best practices in the highly maturated financial institutions. As such, key process areas for improvement were identified. Also, a DQ measuring framework was suggested based on the best DQ measuring practices in financial institutions so that DQ analysis and improvement would be feasible.

In general, this PhD thesis implemented an empirical study with quantitative and qualitative analysis in order to answer the research questions indicated in Chapter 1. The data are collected in different ways. Chapter 2 implemented a survey instrument. Chapter 3 used an experiment while Chapter 4 used a case study methodology. The advantage of adopting an empirical approach is that it captures task specific user's requirements [149]. Furthermore, it may reveal characteristics that theoretical researchers have not defined as

part of DQ and DQ management activities. However, the disadvantage is that the accuracy or completeness of the results found can not be proven by fundamental principles.

5.2 Future Research ideas

Although this thesis presented the contextual DQ management from the perspective of domain experts, there should be an intrinsic DQ evaluation framework for the credit risk assessment data. However, although such frameworks or techniques are abundantly available in literature, they are not being used or implemented in practice. This may imply that these techniques are either too complex to implement or too expensive to acquire. For example, despite the fact that DQ level measuring is the most important activity for continuous DQ improvement, there are no rigorous methods being used to measure the DQ level. As such, we believe that comprehensive DQ metrics should be a priority for future research by considering the elements described in Chapter 4.

Although we believe that the results in this thesis indicate the current DQ requirements and DQ management activities in financial institutions because of the well structured study materials used and rigorous analysis conducted, the validation of the methods and the results can be a future research idea.

6
Appendix

Study Materials - Chapter 2

Below, the questionnaire used in this study is included. As there is a considerable overlap between the pilot and final study questionnaires, only the final study questionnaire is presented. The questions unique to the pilot study are indicated each time.

Final Study Questionnaire

General Questions

1. The sector in which your company operates?—-
2. The country in which your company is?—
3. The primary type of data you are reporting in this questionnaire are?
    a. Financial or Accounting Data
    b. Credit Risk Management Data
    c. Marketing or Sales Data
    d. Human Resource Data

e. Patient, Clinical Data

f. Other (Please specify)

4. Your main role relative to these data; do you primarily:

   a. Collect these data

   b. Use these data in tasks

   c. Work as an information systems professional

   d. Manage those who collect these data

   e. Manage those who use these data in tasks

   f. Manage information systems professionals

   g. Other (Please specify)

5. Your department is:

   a. Financial, Accounting

   b. Risk management

   c. Production, Manufacturing

   d. Marketing, Sales

   e. Human Resource

   f. Information Systems (MIS)

   g. Legal

   h. Senior Executive

   i. Other (Please specify)

6. How long have you worked for this company?

   a. Less than 1 year

   b. 1 to 5 years

   c. 6 to 10 years

   d. More than 10 years

7. How many years of experience do you have?

   a. Less than 1 year

   b. 1 to 5 years
   c. 6 to 10 years
   d. More than 10 years
8. How long have you held your current job?
   a. Less than 1 year
   b. 1 to 5 years
   c. 6 to 10 years
   d. More than 10 years
9. What is your current job title?
10. Highest educational level or degree that you hold?
   a. High school
   b. College degree
   c. Graduate Degree
   d. Other (Please specify)
11. Gender
   a. Female
   b. Male

Part I of the study

1. When you think of data quality, what attributes/dimensions other than accuracy which are necessary for your task come to mind? Please list as many as possible with their meaning?—-(***Note: this question is only asked in the pilot study***)
2. After reviewing the following list, do any other data quality attributes or dimensions which are necessary for your task come to mind? If so, please list them with their meaning. The definitions of all the listed DQ dimensions are given in Question No.3 (***Note: this question is only asked in the pilot study***)

   Accuracy, Relevance, Objectivity,

Reputation, Completeness, Appropriate-amount,

Value-added, Timeliness, Interpretable,

Easily-understandable, Representational-consistency,

Concisely-represented, Accessibility, Security

3 If you are given the following four DQ categories, in which category you will place the newly identified DQ dimensions? (***Note: this question is only asked in the pilot study***)

   a. Access: The extent to which data are available or obtainable.

   b. Contextual: The extent to which data are applicable to the task of the data user.

   c. Intrinsic: The extent to which data values are in conformance with the actual or true values.

   e. Representation: The extent to which data are presented in an intelligible and clear manner.

The above three questions are asked only in the Pilot survey because the main aim of the final survey is not identifying new DQ dimensions rather the aim is assessing the importance level of the already existing DQ dimensions for the task reported.

3.	**How important is it to your task that the data you reported in Question No. 3 in the General Questions section are:**	
		0 1 2 3 4 5 6 7 8 9 10
	accurate: data are certified, error-free, correct, flawless, reliable,	
	complete: data are not missing and cover the needs of tasks	
	value-added: data give you a competitive edge, add value to your operations	
	Timeliness: data are sufficiently up-to-data	
	Interpretable: data are in appropriate language and symbols and the definitions are clear	
	.	.
	.	.
	.	.
	Note: the importance rate increases from 0 to 10	
	Note: this question is asked for all DQ dimensions in Table 2	

Part II of the study

The Part II of the questionnaire includes controlling questions for each DQ dimension. Each DQ dimension has three or four controlling questions. Therefore the consistency of the answers for the controlling questions has been checked using the Cronbach's alpha measure.

1.	**For each statement, indicate the extent to which it is true for the data that you reported in Question No. 3 in the General Questions section.**	
		0 1 2 3 4 5 6 7 8 9 10
	The sources of these data are possible to trace.	
	These data are accurate.	
	Access to these data is sufficiently restricted.	
	These data are formatted compactly.	
	The amount of the data are neither too much nor too little.	
	These data come from good sources.	
	These data are consistently presented in the same format.	
	It is easy to interpret what these data mean.	
	These data are complete.	
	These data are objectively collected.	
	.	.
	.	.
	.	.
	Note: 0 is not at all and 10 is completely true	
	Note: this question is asked for all DQ dimensions in Table 2	

Part III of the study

1. Why is data quality a concern for your task?
    a. Because of regulatory compliance (e.g. Basel II, Solvency II)
    b. Because data quality is becoming a bottleneck for my operational analysis
    c. Because data quality is becoming a bottleneck for my strategic decisions

   d. In order to get a competitive advantage over other competitors
   e. Other (Please specify)

2. What are the major data quality problems for your data?
   a. Getting data consistently represented across business departments
   b. Incomplete data
   c. Wrong values
   d. Diversity of data sources
   e. Making use of available data
   f. Outdated data
   g. Insecurity of the data
   h. Inconsistencies between different copies of the same data
   i. Other (Please specify)

3. What portion of the database where your primary data comes from suffers from data quality problems?
   a. Less than 5%
   b. Between 5-10%
   c. Between 10-20%
   d. Greater than 20%
   e. Not applicable (Specify the reason)
   f. Other (Please specify)

4. Does your organization have a cross-functional data management effort in place?
   a. Yes, Please describe the activities of this cross-functional data management effort
   b. No

5. Do you have a data quality team in your department?
   a. Yes, Please specify the activities and the number of employees working in this team

b. No

6.	**Can you indicate the major and minor causes of the data quality problems you reported in Question No. 2 from the data processes listed below?**	Major Cause	Minor Cause	NA
	Initial data conversion: Data conversion from some previously existing old system to the new databases.	-	-	-
	System consolidation: Database consolidations after corporate mergers.	-	-	-
	Manual data entry: Entering data into system manually.	-	-	-
	Batch feeds: Regular data exchange between systems through batch interfaces.	-	-	-
	Real-time interfaces: Data exchanged between the systems through real-time interfaces.	-	-	-
	Data Processing: The change in the programs responsible for regular data processing.	-	-	-
	Data cleansing: Using automated data cleansing rules to make corrections in mass	-	-	-
	Data purging: Deleting old data routinely from the system to make way for more new data.	-	-	-
	Changes not captured: Different organizational changes but not captured in the system.	-	-	-
	System upgrades: Systems software are often upgraded every few years.	-	-	-
	New data uses: The data may be good enough for one purpose but inadequate for another.	-	-	-
	Loss of expertise: Much of the data knowledge exists in people's minds rather than metadata documents.	-	-	-
	Process automation: With the progress of technology, more and more tasks are automated.	-	-	-

Study Materials - Chapter 3

The complex decision task is given as an example of the experiment.

Financial health or bankruptcy prediction Task: Genet is working as a consultant in an accounting firm. She has been given 8 firms and was asked to rank them according to their solvency (financial health). Hence, she has begun the decision process of examining the firms. First, she identified four solvency determinant criteria and indicated the importance of each criterion using a weight based on an Altman Z-score model for non-manufacturing firms. The weight indicates the relative importance of each criterion in predicting the solvency of a company. The higher the weight, the higher the importance of a particular criterion in predicting the solvency of the company. Next, she represented the value for each criterion in euros, where higher values refer to more healthy firms. For example, a value of 90 euros for $\frac{working\ capital}{Total\ assets}$ ratio indicates a firm which is more healthy compared to a firm with a value of 50 euros.

Yet, she realized that the values may not be completely accurate as they are not consistent among different databases she checked. Thus, she decided to incorporate this uncertainty into her decision making process by using a $[0, 1]$ accuracy measure where 0 indicates an inaccurate value and 1 indicates a perfectly accurate value. For example, an accuracy of 0.8 for a criterion's value indicates a 80% chance for the value to be correct.

However, because she is assigned to other work, she was unable to finish her ranking decision. Hence, her supervisor asked you to continue her work and to decide upon the ranking of the firms. You can assume that the accuracy, the value and the weight of the firms are correctly retrieved by Genet. Please rank the firms below according to their solvency from the most healthy firm (Rank 1) to the least healthy firm (Rank 8). Also, please explain why.

Firm	Criterion	Accuracy	Value	Weight
Firm A	$\frac{Retained\ earnings}{total\ assets}$	0.8	84	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.8	24	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	0.5	80	7
	$\frac{working\ capital}{Total\ assets}$	0.5	16	6.5
Firm B	$\frac{Retained\ earnings}{total\ assets}$	0.8	20	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.8	16	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	1	48	7
	$\frac{working\ capital}{Total\ assets}$	1	30	6.5
Firm C	$\frac{Retained\ earnings}{total\ assets}$	0.4	100	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.2	80	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	0.1	90	7
	$\frac{working\ capital}{Total\ assets}$	0.1	80	6.5
Firm D	$\frac{Retained\ earnings}{total\ assets}$	0.6	52	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.8	48	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	0.8	54	7
	$\frac{working\ capital}{Total\ assets}$	0.8	26	6.5
Firm E	$\frac{Retained\ earnings}{total\ assets}$	0.7	76	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.8	24	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	1	55	7
	$\frac{working\ capital}{Total\ assets}$	1	40	6.5
Firm F	$\frac{Retained\ earnings}{total\ assets}$	0.8	24	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.5	18	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	0.2	20	7
	$\frac{working\ capital}{Total\ assets}$	0.2	56	6.5
Firm G	$\frac{Retained\ earnings}{total\ assets}$	1	50	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.8	40	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	1	50	7
	$\frac{working\ capital}{Total\ assets}$	1	28	6.5
Firm H	$\frac{Retained\ earnings}{total\ assets}$	0.7	52	3
	$\frac{Market\ value\ equity}{Book\ value\ of\ total\ liabilities}$	0.3	48	1
	$\frac{Earnings\ before\ interest\ \&\ taxes}{Total\ assets}$	0.3	51	7
	$\frac{working\ capital}{Total\ assets}$	0.2	29	6.5

6.1 Exit Survey

1. Please explain how you reached to the ranking decision clearly.
2. Please write the formula you used and the variables you included in the formula.
3. Please also indicate the assumptions you made when you are solving this task if there is any.
4. Is the experiment completely clear ?
    a. Yes
    b. No
5. I am confident that my financial health prediction ranking is correct
    a. Strongly agree
    b. Agree
    c. Neither agree/disagree
    d. Disagree
    e. Strongly Disagree
6. The factors that contribute to my degree of confidence (or lack of confidence) are:
7. Gender
    a. Female
    b. Male
8. Please indicate the highest educational level that you have achieved
    a. High School
    b. Bachelors Degree
    c. Masters Degree
    d. Post Masters Degree (please specify)
9. My occupation may be described as
    a. Professor

b. Full time graduate student

c. Engineer

d. Programmer

e. Other (please specify)

10. My age is

a. 17-20

b. 21-30

c. 31-40

d. 41-50

e. 51-60

f. Greater than 60

11. The number of years that I have been a full-time employee is

a. 0

b. 1-10

c. 11-20

d. 21-30

e. Greater than 30

12. In the financial health prediction task, what data was most useful to you?

13. In the financial health prediction task, what data would you like to have had that you did not have?

14. How many times have you experienced such decision?

a. 0

b. 1-3

d. 4-6

e. 7-9

f. 10 or more

15. Have you heard of data quality?
    a. Yes
    b. No
16. Have you ever attended any data quality related training?
    a. Yes
    b. No
17. What is data quality to you?
18. How do you define accuracy in the data quality context?
19. What does accurate data mean to you?
20. Can you please give one example of inaccurate data?
21. Do you know what is meant by 'data quality dimensions'?
22. Can you please mention some data quality dimensions?

Study Materials - Chapter 4

Questionnaire to assess the maturity level of DQ measuring activities in four financial institutions.

Preliminary

- Can you please mention the most important purposes (types of decisions, processes,...) for which the data are used in your department?
- Can you please tell me what your position is in the organization?

Part I and II - maturity model and DQ metric framework

1. Can you please tell me which data types you are using to complete your work?
    a. Database types/excel files or word files
2. Do you have a set of important data characteristics identified and defined for the data you use? (For example, the characteristics of a customer account database include items such as account number, name and so on.)
    a. If the data characteristics are defined and documented for the data you use, can you please give some example of those?
    b. Are you satisfied by all the characteristics which the data have by now for your use? In other words, do you need the data to have more characteristics to fully accomplish your task than what they have now? If so, can you please mention those?
3. Did you identify the information production systems which are used to produce the data in question 1? can you please mention those?
4. Do you know all the stakeholders of (data producer, users, suppliers and data production managers) the data you use? For example, the users of the customer account database include financial controllers, accountants, customer representatives and so on.

   a. Can you please mention the stakeholders for the data in question 1?

5. How are the data for your usage produced? Do you have a documentation of an information production systems which describes how the data is produced and the interaction among information suppliers, producers, users and others.

   a. Is it possible, easy and clear to trace back to every step of the data production process?

   b. Are there information which is derived (e.g. results of calculations, aggregated information) documented so that it is possible to know what is the source information?

   c. What is the process used to determine the source of aggregated (e.g. results of calculations)information?

6. What is the general definition of data quality and data quality dimensions in your department? What are the most recurring DQ problems identified in your department?

7. Did you already identify the most important DQ dimensions in order to fulfill the quality requirement for the data you use? Can you please mention those and their definitions (or requirements under each of those dimensions)? For example, timeliness for the stock trader indicates how old the data is?

   a. How are these DQ dimensions identified? For example, is there a discussion between all the stakeholder of the data to identify and reconcile those important DQ dimensions?

   b. What methods/technologies or experts used to identify these DQ dimensions?

8. Did you identify which data to be measured for their quality?

   a. Databases

   b. Key attributes - such as client addresses, sales per client and so on

9. Did you identify the maximum number of errors (a threshold for the errors) to be tolerated per each row/column or relational table in the databases or in general in the database in question nr. 6

   a. How is the threshold developed?

10. Do you know the quality level of the data you are using?
    a. Do you measure the level of data quality for the data used by your department?
    b. Which data are being measured for their quality?
    c. Where in the data production process are the data being measured for their quality?
    d. Is there a responsible stakehoder (or stakeholders) to measure the quality level of the data?
    e. Do you use in house developed or purchased data quality metrics to measure the quality level of the data in question? Can you please give the formula of the metrics?
    f. Have the types of errors to be measured been identified and documented?
    g. Can you please indicate the types of errors and to which DQ dimensions they belong? For example, the number of incorrect customer addresses belong to the accuracy dimension, the time when the customer account was last updated belongs to timeliness dimension, and so on.
11. Are the metrics used based on business rules? Such as, a total risk exposure of a client should not exceed a certain limit?
    a. If so, can you please indicate some of the business rules used and how they are developed?
12. Is it easy to interpret the results from the metrics?
    a. How are the results of the DQ measurement results being reported? at what scale? at the record, filed or relational table scale.
13. How often is the quality of the data is question being measured?
    a. Weakly
    b. Monthly
    c. Whenever the need is raised by the data users
    d. Please indicate if there are other?

14. Has the data quality level been communicated to the stakeholders of the data? And how often?
15. Have the root cause of different data quality problems such as inaccuracies and missing values been identified?
    a. What are these causes?
16. Has the economic model for the value of quality data been identified?
    a. How and by whom is it identified?
17. Do you clearly know the cost of poor quality data on your department's performance?
    a. What are those costs identified? And how they are identified?
    b. Are these costs communicated to the stakeholders (data users, producers or suppliers) of the data? Is there a documented way of communicating those costs? What is the channel of communicating those costs?
18. What are the methods or technologies used to analyze the DQ problems identified?
    a. Statistical process control
    b. Pattern recognition methods
    c. Pareto chart analysis for poor DQ dimensions over time
19. Did you evaluate how representative or comprehensive the DQ metrics used are?
    a. Whether those DQ metrics are the right set of metrics?
    b. How the DQ metrics link to the factors that are critical to the quality of the data?
20. Did you identify key areas for improvement? Such as;
    a. Aligning information flow and work flow with the corresponding information manufacturing systems
    b. Re-aligning the key characteristics of the data with business needs
21. Are employees rewarded for their efforts in creating very good quality data?

22. Are DQ cleansing activities being carried out?
23. How do you correct those DQ deficiencies in your department?
24. Do you keep data quality metadata? If so what it constitutes and at what level (data item level, record level, attribute level and relational table level)
25. Inaccuracies and incompleteness are mitigated from their sources (Fully, partially or not at all).
26. Do you have a plan in place to re-define and include more DQ dimensions than the one mentioned?
27. Do you have a plan in place to measure, analyze and improve the DQ level of the data in question continuously?
28. Are all the credit risk information needs translated into specifications for the information systems? (Data modeling: conceptual, logical and physical)
    a. How are they translated? Can you please give some examples?
29. For all your credit risk information systems, are there documented processes in place for the physical storage, the backup of critical information, the archival and destruction of information.
30. Are there authorization and authentication in place to access credit risk information?
31. Do you have confidential credit risk information? Are there specific practices in place for the management and disposal of these information? if yes?
    a. What are these practices?
    b. How is confidential information transmitted between parties and disposed of?
32. Are there standardized templates for the Visualization of the information in the different credit risk information systems? are standardized templates similar to all credit risk information systems?
33. Do you use the concept of information products similar to manufacturing of tangible goods?

34. Have the relationships between the different information products been identified and documented?
35. Is the information flows in the credit risk management department documented?
36. What techniques are used to assess information quality for the different credit risk information systems? for example, automated tools, surveys, data profiling techniques, business rule violation?
37. Are there information cleansing scripts in use in the credit risk management information systems?
38. Is there a standardized procedure in use for handling credit risk information quality problems?
    a. Can you please describe what it covers?
    b. Is there a standard procedure for reporting IQ problems?
39. Is there any information quality training available for credit risk employees or in general in the institution? if yes, what it covers?
40. Is there an information quality team in the organization or in credit risk department?
41. Is the credit risk management strategy explicitly take into account information quality?
    a. What does the strategy include with respect to information quality?
42. Is information quality benchmarked against other departments or organizations?
43. Do you use master data management?
44. Are information quality initiatives supported by current information system deployment?
45. How well are information quality initiatives aligned with the enterprise architecture requirements?
    a. Not at all
    b. Not well
    c. Well

d. Very Well

46. How is information systems architecture integrated across the different systems in the credit risk department? (systems are completely independent, systems are independent but communicate, systems are synchronized and master information is consolidated into one system)

47. Do you benchmark the information quality management practises in your department against information quality management best practices or standards?

48. Do you have automated information quality checks so that information is checked for quality before it is allowed to propagate through to all the organizational systems?

Critical Success Factors (CSFs)	Mapping of the questions in Section 6.1 which assess the satisfaction level of each CSF
IQM Optimization	25,27
IQM Metrics	13,19,19a,19b,
Analysis and Reporting	18a,18b,18c
IQM Benchmarking	9,9a,47
IQ Problem Root-Cause-Analysis	15,15a,25
IQ Risk Management and Impact Assessment	17,17a,17b
IQM Cost-Benefit Analysis	16,16a
Business Process Re-engineering (BPR) for IQ Improvements	20a,20b
Enterprise Tier Management	45
Information Tier Management	46
Application Tier Management	44
Physical Tier Management	44
Master Data Management/Redundant Storage	43
IQ Firewall	48
IQ Management Accountability, Rewards & Incentives: IQ is Everyone's Responsibility	21
IQ Benchmarking	9,9a,47
Strategic IQ Management	41,41a
IQ Audit Trail	5,5a
IQ Management Team and Project Management	40
IQ Management Education, Training and Mentoring	39
IQ Problem Reporting and Handling	38,38a,38b
Scripted Information Cleansing	22,23,37
IQ Metrics	10a,10d,11,11a,13
IQ Evaluation	10, 10g
Requirements Elicitation	7,7a
Requirements Analysis	7a, 7b
Requirements Management	7
Information Supply Chain Management	33,34,35
IP Configuration Management	33,34
IP Taxonomy Management	33,34
IP Visualization Management	32
Derived IPs Management	5b,5c
Meta-Information (Metadata) Management	24
Security Classification IPs	31
Secure Transmission of Sensitive Information	31,31a,31b
Sensitive Information Disposal Management	31a
Authentication	30
Authorization	30
Audit Trail	5a
Physical Storage	29
Backup and Recovery	29
Archival and Retrieval	29
Information Destruction	29
Stakeholder Management	3,4,4a,5
Conceptual Modeling	28
Logical Modeling	2,2a,2b,28
Physical Modeling	28

List of Figures

List of Tables

Bibliography

[1] Paul Alpar and Sven Winkelsträter. Assessment of data quality in accounting data with association rules. *Expert Systems with Applications*, 41(5):2259–2268, 2014.

[2] E.I. Altman et al. Predicting financial distress of companies: Revisiting the z-score and zeta models. *Stern School of Business, New York University*, 2000.

[3] B. Baesens, C. Mues, D. Martens, and J. Vanthienen. 50 years of data mining and or: upcoming trends and challenges. *Journal of the Operational Research Society*, 60:16–23, 2009.

[4] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards. *Technical report, Bank of international settlements*, 2006.

[5] Sasa Baskarada. *IQM-CMM: Information quality management capability maturity model*. Springer, 2010.

[6] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*, pages 20–50. Springer, New York, 2006.

[7] Lisa Belkin. How can we save the next victim. *The New York Times Magazine*, page 13, 1997.

[8] Laure Berti-Equille. Data quality awareness: a case study for cost optimal association rule mining. *Knowledge and information Systems*, 11(2):191–215, 2007.

[9] A Bitterer. Gartners data quality maturity model. *Gartner, Stamford*, 2007.

[10] Andreas Bitterer and D Newman. Organizing for data quality. *Gartner Research, Stamford, CT*, 2007.

[11] Columbia Accident Investigation Board. Columbia accident investigation board report. *online report*, 1, 2003.

[12] H. Butcher. Information overload in management and business. In *Information Overload, IEE Colloquium on*, pages 1–1. IET, 1995.

[13] Ismael Caballero, Angélica Caro, Coral Calero, and Mario Piattini. Iqm3: Information quality management maturity model. *J. UCS*, 14(22):3658–3685, 2008.

[14] P. Caplan. *Metadata fundamentals for all librarians*. American Library Association, 2003.

[15] C. Cappiello, P. Giciaro, and B. Pernici. Hiqm: A methodology for information quality monitoring, measurement, and improvement. *ER Workshops*, LNCS 4231:339–351, 2006.

[16] Presidential Commission On Space Shuttle Challenger and WP Rogers. Report of the presidential commission on the space shuttle challenger accident, 1986.

[17] C.C. Chen and Y.D. Tseng. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, in press, 2010.

[18] C.C. Chen and Y.D. Tseng. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 50(4):755–768, 2011.

[19] I.N. Chengalur-smith, D.P. Ballou, and H.L. Pazer. The impact of data quality information on decision making: An exploratory analysis. *IEEE Transactions of Knowledge and Data Engineering*, 11(6), 1999.

[20] Edgar F Codd. Extending the database relational model to capture more meaning. *ACM Transactions on Database Systems (TODS)*, 4(4):397–434, 1979.

[21] Edgar F Codd. Data models in database management. In *ACM Sigmod Record*, volume 11, pages 112–114. ACM, 1980.

[22] Philip B Crosby. *Quality is free: The art of making quality certain*, volume 94. McGraw-Hill New York, 1979.

[23] Thomas H Davenport and Laurence Prusak. *Information ecology: Mastering the information and knowledge environment*. Oxford University Press, 1997.

[24] K. Dejaeger, B. Hamers, J. Poelmans, and B. Baesens. A novel approach to the evaluation and improvement of data quality in the financial sector. In *Proceedings of the 15^{th} International Conference on Information Quality*, Little Rock, USA, 2010.

[25] W.H. Delone and E.R. McLean. Information systems success: The quest for the dependant variables. *Information Systems Research*, 3(1):60–95, 1992.

[26] William Edwards Deming and Deming W Edwards. *Quality, productivity, and competitive position*, volume 183. Massachusetts Institute of Technology, Center for Advanced Engineering Study Cambridge, MA, 1982.

[27] Kevin J Dooley and Richard F Flor. Perceptions of success and failure in tqm initiatives. *Journal of Quality Management*, 3(2):157–174, 1998.

[28] Lian Duan, Lida Xu, Ying Liu, and Jun Lee. Cluster-based outlier detection. *Annals of Operations Research*, 168(1):151–168, 2009.

[29] Janet M Dukerich and Mary Lippitt Nichols. Causal information search in managerial decision making. *Organizational Behavior and Human Decision Processes*, 50(1):106–122, 1991.

[30] Wayne W Eckerson. Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*, 2002.

[31] A. Edmunds and A. Morris. The problem of information overload in business organisations: a review of the literature. *International journal of information management*, 20(1):17–28, 2000.

[32] L.P. English. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*, volume 1. Wiley New York, 1999.

[33] M.J. Eppler and D.Wittig. Conceptualizing information quality: A review of information quality frameworks from the last ten years. In *Proceedings of the 2000 Conference on Information Quality*, 2000.

[34] A. Even, G. Shankaranarayanan, and S. Watts. Enhancing decision making with process metadata: Theoretical framework, research tool, and exploratory examination. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 8, pages 209a–209a. IEEE, 2006.

[35] G. Shankaranarayanan & A. Even. Metadata: A great pomise of a sisphean torture. *Communication of the ACM*, 49(2):88–94, 2006.

[36] C.W. Fisher and D.P. Ballou. The impact of experience and time on use

of data quality information in decision making. *Information Systems Research*, 14(2):170–188, 2003.

[37] C.W. Fisher and B.R. Kingma. Criticality of data quality as exemplified in two disasters. *Information & Management*, 39(2):109–116, 2001.

[38] C.W. Fisher, E.J.M. Lauria, and C.C. Matheus. An accuracy metric: Percentages, randomness, and probabilities. *Journal of Data and Information Quality (JDIQ)*, 1(3):16, 2009.

[39] European Foundation. Efqm. introducing excellence. technical report. *European Foundation for the Quality Management*, 2003.

[40] Floyd J Fowler. *Survey research methods*, volume 1. Sage publications, 2014.

[41] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[42] Nelson Friedman. Radcliffe (2004). *CRM Demands Data Cleansing. Gartner Research*.

[43] T Friedman, D Feinberg, MA Beyer, B Gassman, A Bitterer, D Newman, J Radcliffe, A White, R Paquet, C DiCenzo, et al. Hype cycle for data management. *GartnerGroup Research, Stamford, CT, July*, 6, 2006.

[44] G.C.Cawley and N.LC.Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, 2003.

[45] T.V. Gestel and B. Baesens. *Credit Risk Management*. Oxford University Press Inc. New York, 2009.

[46] Gerd Gigerenzer and Reinhard Selten. *Bounded rationality: The adaptive toolbox*. Mit Press, 2002.

[47] T. Gill, A. J. Gilliland-Swetland, M. Whalen, M. S. Woodley, and M. Baca. *Introduction to metadata*. Getty Research Institute, 2008.

[48] S.W. Gilliland, L. Wood, and N. Schmitt. The effects of alternative labels on decision behavior: The case of corporate site selection decisions. *Organizational behavior and human decision processes*, 1994.

[49] C James Goodwin. *Research in psychology: Methods and design*. John Wiley & Sons, 2009.

[50] M.B. Gordy. A comparative anatomy of credit risk models. *Journal of Banking and Finance*, 24:119–149, 2000.

[51] J. Greenberg. Understanding metadata and metadata schemes. *Cataloging & classification quarterly*, 40(3-4):17–36, 2005.

[52] Jingyu Han, Dawei Jiang, and Lingjuan Li. Automatic accuracy assessment via hashing in multiple-source environment. *Expert Systems With Applications*, 37(3):2609–2620, 2010.

[53] Hervé Hannoun. The basel iii capital framework: a decisive breakthrough. *discurso pronunciado en el seminario de alto nivel BoJ-BIS Financial Regulatory Reform: Implications for Asia and the Pacific, www. bis. org/speeches/sp101125a. pdf*, 2010.

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer, 2001.

[55] B. Heinrich and M. Klier. Assessing data currency a probabilistic approach. *Journal of Information Science*, 37(1):86, 2011.

[56] M.A. Hernández and S.J Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1):9–37, 1998.

[57] Rudy Hirschheim, Heinz K Klein, and Kalle Lyytinen. *Information systems development and data modeling: conceptual and philosophical foundations*, volume 9. Cambridge University Press, 1995.

[58] Victoria J Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[59] Nicholas J Horton and Ken P Kleinman. Much ado about nothing. *The American Statistician*, 61(1), 2007.

[60] SAS Institute. *SAS/STAT 9. 22 User's Guide: Statistical Graphics Using ODS (Book Excerpt)*. SAS Institute, 2010.

[61] M. Jarke and Y. Vassiliou. Data warehouse quality: A review of the DWQ project. In *Proceedings of the Confrenece on Information Quality*, pages 299–313, Cambridge, MA, 1997.

[62] Henk Jonkers, Marc M Lankhorst, Hugo WL ter Doest, Farhad Arbab, Hans Bosma, and Roel J Wieringa. Enterprise architecture: Manage-

ment tool and blueprint for the organisation. *Information Systems Frontiers*, 8(2):63–66, 2006.

[63] J.R.A.Santos. Cronbachs alpha: A tool for assessing the reliability of scales. *Journal of Extension*, 37(2):1–5, 1999.

[64] J. S. Park K. S. Ryu and J. H. Park. A data quality management maturity model. *ETRI Jouranl*, 28(2):191–204, 2006.

[65] B.K. Kahn, D.M. Strong, and R.Y. Wang. Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45(4), 2002.

[66] Stephen H Kan. *Metrics and models in software quality engineering*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[67] Thomas Kean. *The 9/11 commission report: Final report of the national commission on terrorist attacks upon the United States*. Government Printing Office, 2011.

[68] Mohamed Khalifa and June M Verner. Drivers for software development method usage. *Engineering Management, IEEE Transactions on*, 47(3):360–369, 2000.

[69] D.E. Kieras and D.E. Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human-computer interaction*, 12(4):391–438, 1997.

[70] B.R. Kingma. *The Economics of Information: A Guide to Economic and Cost-Benefit Analysis for Information Professionals. Library and Information Science Text Series.* ERIC, 1996.

[71] B.D. Klein, D.L. Goodhue, and G.B. Davis. Can humans detect errors in data? impact of base rates, incentives, and goals. *Management Information Systems Quarterly*, 21:169–194, 1997.

[72] Rita Kovac, Yang W Lee, and Leo Pipino. Total data quality management: The case of iri. In *IQ*, pages 63–79, 1997.

[73] Karen Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439, 1992.

[74] Ranjit Lall. Why basel ii failed and why any basel iii is doomed. *Global Economic Governance Programme, GEC Working Paper*, 52, 2009.

[75] Y.W. Lee, L.L. Pioino, J.D. Funk, and R.Y. Wang. *Journey to Data Quality*, pages 67–108. The MIT Press, London, 2006.

[76] Y.W. Lee, D.M. Strong, B.K. Kahn, and R.Y. Wang. A methodology for information quality assessment. *Information and Management*, 40:133–146, 2002.

[77] Xiao-Bai Li. A bayesian approach for estimating and replacing missing categorical data. *Journal of Data and Information Quality (JDIQ)*, 1(1):3, 2009.

[78] E.K. Macdonald and B.M. Sharp. Brand awareness effects on consumer decision making for a common, repeat purchase product:: A replication. *Journal of Business Research*, 48(1):5–15, 2000.

[79] S Madnick and RY Wang. Introduction to total data quality management (tdqm) research program. Technical report, TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management, 1992.

[80] S. Madnick and H. Zhu. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, 59:460–475, 2006.

[81] S.E. Madnick, R.Y Wang, Y.W. Lee, and H. Zhu. Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)*, 1(1):2, 2009.

[82] Jonathan I Maletic and Andrian Marcus. Data cleansing: Beyond integrity analysis. In *IQ*, pages 200–209. Citeseer, 2000.

[83] Yogesh Malhotra. Business process redesign: an overview. *IEEE Engineering Management Review*, 26:27–31, 1998.

[84] N. Mantel. Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700, 1963.

[85] J.Y. Mao and I. Benbasat. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, 17(2):153–180, 2000.

[86] D. Marco. *Building and managing the meta data repository*. Wiley New York, 2000.

[87] A. Maydanchik. *Data Quality Assesment*, pages 5–30. Technics publications, 2007.

[88] Helinä Melkas. Analyzing information quality in virtual service networks with qualitative interview data. In *IQ*, pages 74–88, 2004.

[89] Michael E Milakovich. Rewarding quality and innovation: awards, charters, and international standards as catalysts for change. In *Knowledge Management in Electronic Government*, pages 80–90. Springer, 2004.

[90] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. Data quality for credit risk management: new insights and challenges. In *International Conference on Information Quality (ICIQ)*. University of South Australia, Adelaide (Australia)., 2011.

[91] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. A total data quality management for credit risk: new insights and challenges. *International Journal of Information Quality*, 3(1):1–27, 2012.

[92] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Journal of Information & Managmenet*, 50(1):43–58, 2013.

[93] H.T. Moges, W. Lemahieu, and B. Baesens. The use of data quality information (dqi) for decision-making: an exploratory study. In *Proceeding of the 2008 conference on Collaborative Decision Making: Perspectives and Challenges*, pages 233–244. IOS Press, 2008.

[94] C. Moraga, M.A. Moraga, C. Calero, and A. Caro. Towards the discovery of data quality attributes for web portals. *ICWE*, LNCS 5648:251–259, 2009.

[95] P. C. Morrow. Physical attractiveness and selection decision making. *Journal of Management*, 16(1):45–60, 1990.

[96] R.R. Nelson, P.A. Todd, and B.H. Wixom. Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems*, 21(4):199–235, 2005.

[97] P.C. Nutt. Types of organizational decision processes. *Administrative Science Quarterly*, pages 414–450, 1984.

[98] John S Oakland. *Total quality management: text with cases*. Routledge, 2003.

[99] Ken Orr. Data quality and systems theory. *Communications of the ACM*, 41(2):66–71, 1998.

[100] A. Paul P. Cykana and M. Stern. Dod guidelines on data quality management. In *Proceedings of the Conference on Information Quality*, pages 154–171, Cambridge, MA, 1996.

[101] F. Panse and N. Ritter. Completeness in databases with maybe tuples. *ER workshops*, pages 202–211, 2009.

[102] A. Parssian and V.S. Jacob. Assessing data quality for information products: Impact of selection, projection, and cartesian product. *Management Science*, 50(7):967–982, 2004.

[103] A. Parssian, S. Sarkar, and V.S. Jacob. Assessing data quality for information products: impact of selection, projection, and cartesian product. *Management Science*, 50(7):967–982, 2004.

[104] Amir Parssian. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decision Support Systems*, 42(3):1494–1502, 2006.

[105] Amir Parssian, Sumit Sarkar, and Varghese S Jacob. Assessing information quality for the composite relational operation join. In *IQ*, pages 225–237, 2002.

[106] Mark Paulk. *Capability maturity model for software*. Wiley Online Library, 1993.

[107] J.W. Payne. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387, 1976.

[108] J.W. Payne, J.R. Bettman, and E.J. Johnson. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.

[109] J.W. Payne, J.R. Bettman, and E.J. Johnson. *The adaptive decision maker*. Cambridge Univ Pr, 1993.

[110] Chad Perry. Processes of a case study methodology for postgraduate

research in marketing. *European journal of marketing*, 32(9/10):785–802, 1998.

[111] M. Pinto. Data representation factors and dimensions from the quality function deployment (qfd) perspective. *Journal of information science*, 32(2):116–130, 2006.

[112] L.L. Pipino, Y.W. Lee, and R.Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4), 2002.

[113] Les Porter and Steve Tanner. *Assessing business excellence*. Routledge, 2012.

[114] R. Price and G. Shanks. Representing data quality information usably. *Clayton School of Information Technology, Monash University, Technical report*, 243:1–19, 2009.

[115] R. Price and G. Shanks. DQ tags and decision-making. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[116] R. Price and G. Shanks. The impact of data quality tags on decision-making outcomes and process. *Journal of the Association for Information Systems*, 12(4):1, 2011.

[117] J Ross Quinlan. Unknown attribute values in induction. In *ML*, pages 164–168, 1989.

[118] Ronald A. Radice, Norman K. Roth, AC O'Hara, and William A Ciarfella. A programming process architecture. *IBM Systems Journal*, 24(2):79–90, 1985.

[119] S. Raghunathan. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support Systems*, 26:275–286, 1999.

[120] Srinivasan Raghunathan. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support Systems*, 26(4):275–286, 1999.

[121] E. Rahm and H.H. Do. Data cleaning: problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2000.

[122] T.C. Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 1998.

[123] Thomas C Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2):79–82, 1998.

[124] Jack B ReVelle and Robert Alan Kemerling. Total quality management, six sigma, and continuous improvement. *Mechanical Engineers' Handbook: Manufacturing and Management, Volume 3, Third Edition*, pages 583–615, 2006.

[125] S. E. Madnick R.Y. Wang, E. M. Pierce and C. W. Fisher. *Advances in Management Information Systems*. Armonk, NY: M. E. Sharpe, 2005, 2005.

[126] A.H. Schoenfeld and D.J. Herrmann. Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):484, 1982.

[127] V. Sessions and M. Valtorta. Towards a method for data accuracy assessment utilizing a bayesian network learning algorithm. *Journal of Data and Information Quality (JDIQ)*, 1(3):14, 2009.

[128] G. Shankaranarayanan and Y. Cai. Supporting data quality management in decision-making. *Decision Support Systems*, (42):302–317, 2006.

[129] G. Shankaranarayanan and B. Zhu. Data quality metadata and decision making. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1434–1443. IEEE, 2012.

[130] G. Shankaranarayanan, M. Ziad, and R.Y. Wang. Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management*, 14(4):14–32, 2003.

[131] G. Shanks. The impact of data quality tagging on decision outcomes. 2001.

[132] G. Shanks and E. Tansley. Data quality tagging and decision outcomes: An experimental study. In *IFIP Conference on Decision Making and Decision Support in the Internet Age*, pages 399–410, 2002.

[133] Jau-Ji Shen and Ming-Tsung Chen. A recycle technique of association rule for missing value completion. In *Advanced Information Networking*

and Applications, 2003. AINA 2003. 17th International Conference on, pages 526–529. IEEE, 2003.

[134] H.A. Simon. Administrative behaviour. *Australian Journal of Public Administration*, 9(1):241–245, 1950.

[135] Heather A Smith and James D McKeen. Developments in practice xxxvi: How to talk so business will listen and listen so business will talk. *Communications of the Association for Information Systems*, 27(1):13, 2010.

[136] Robert E Stake. *The art of case study research*. Sage, 1995.

[137] D.N. Stone and D.A. Schkade. Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes*, 49(1):42–59, 1991.

[138] D.M. Strong, Y.W. Lee, and R.Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[139] O. Svenson and A. Edland. Change of preferences under time pressure: Choices and judgements. *Scandinavian Journal of Psychology*, 28(4):322–330, 1987.

[140] G.K. Tayi and D.P. Ballou. Examining data quality. *Communications of the ACM*, 41(2), 1998.

[141] SCAMPI Upgrade Team. Standard cmmi appraisal method for process improvement (scampi) a. *Version 1.2: Method Definition Document*, 2006.

[142] J.D. Thompson. *Organizations in action: Social science bases of administrative theory*. Transaction Pub, 1967.

[143] Y. Wand and R.Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 1996.

[144] Richard Y Wang, Yang W Lee, Leo L Pipino, and Diane M Strong. Manage your information as a product. *Sloan Management Review*, 39:95–105, 1998.

[145] Richard Y Wang, Martin P Reddy, and Henry B Kon. Toward quality data: An attribute-based approach. *Decision Support Systems*, 13(3):349–372, 1995.

[146] R.Y. Wang. A product perspective on data quality management. *Communications of the ACM*, 2, 1998.

[147] RY Wang, K Chettayar, F Dravis, J Funk, R Katz-Haas, C Lee, Y Lee, X Xian, and S Bhansali. Exemplifying business oppurtunities for improving data quality from corporate household research. *Advances in Management Information Systems-Information Quality (AMIS-IQ) Monograph*, 2005.

[148] R.Y. Wang, V.C. Storey, and C.P. Firth. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4), 1995.

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

[150] J.E. Ware and B. Gandek. Methods for testing data quality, scaling assumptions, and reliability. *Journal of Clinical Epidemiology*, 51(11):945–952, 1998.

[151] S. Watts, G. Shankaranarayanan, and A. Even. Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48:202–211, 2009.

[152] S. Watts and W. Zhang. Knowledge adoption in online communities of practice. *system d'information et Management*, 1:9, 2004.

[153] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[154] R.E. Wood. Task complexity: Definition of the construct. *Organizational behavior and human decision processes*, 37(1):60–82, 1986.

[155] Philip Woodall, Alexander Borek, and Ajith Kumar Parlikad. Data quality assessment: The hybrid approach. *Information & Management*, 50(7):369–382, 2013.

[156] Philip Woodall, Ajith Kumar Parlikad, and Lucas Lebrun. Approaches to information quality management: State of the practice of uk asset-intensive organisations. In *Asset Condition, Information Systems and Decision Models*, pages 1–18. Springer, 2012.

[157] Robert K Yin. *Case study research: Design and methods*. Sage publications, 2014.

[158] C. Zeeh. The Lempel Ziv Algorithm. Technical report, University of Munich, 2003.

[159] H. Zhu and R.Y. Wang. Information quality framework for verifiable intelligence products. *Data Engineering*, pages 315–333, 2010.

[160] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wode web. In *Proceedings of the 3rd Annual Internation ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 288–295, 2000.

Doctoral dissertations from the faculty of business and economics

A full list of the doctoral dissertations from the Faculty of Business and Economics can be found at:

www.kuleuven.ac.be/doctoraatsverdediging/archief.htm.