CIRRELT
1971 – 2021

CIRRELT-2022-10

A Method to Classify Data Quality for Decision Making under Uncertainty

Vanessa Simard
Mikael Rönnqvist
Luc Lebel
Nadia Lehoux

April 2022

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

A Method to Classify Data Quality for Decision Making under Uncertainty

Vanessa Simard[1,*], Mikael Rönnqvist[1,2], Luc Lebel[2,3], Nadia Lehoux[1,2]

1. Department of Mechanical Engineering, Université Laval
2. Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)
3. Department of Wood and Forest Science, Université Laval

Abstract. Every decision-making process is subject to a certain degree of uncertainty. In sectors where the outcomes of planned activities are uncertain and difficult to control such as in forestry, data describing the available resources can have a large impact on productivity. When planning operations, it is often assumed that such data are accurate, which causes a need for more replanning efforts. Data verification is kept to a minimum even though using erroneous information increases the level of uncertainty. In this context, it is relevant to develop a process to evaluate whether the data used for planning decisions are appropriate, so as to ensure the decision validity and provide information for better understanding and actions. However, the level of data quality alone can sometimes be difficult to interpret and needs to be put into perspective. This paper proposes an extension to most data quality assessment techniques by comparing data to past quality levels. A classification method is used to evaluate the level of data quality in order to support decision making. Such classification provides insights into the level of uncertainty associated with the data. The method developed is then exploited using a theoretical case built from data quality assessments from the literature and a practical case study from the forest sector. An example of how classified data quality can improve decisions in a transportation problem is finally shown.

Keywords: Data quality, uncertainty, decision-making process, forest industry.

Acknowledgments: The authors would like to acknowledge the financial support from the Fonds de Recherche du Québec and the private and public partners of the FORAC Research Consortium.

---

* Corresponding author: vanessa.simard.4@ulaval.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2022

1 Introduction

The efficiency of a company is greatly affected by its capacity to forecast future events and prepare accordingly (Mula *et al.*, 2006). While some events can be anticipated with algorithms and 'experience', others are more difficult to predict with absolute certainty. Zhu *et al.* (2012) explained that forecasts will not represent reality if the data on which the decisions are based are erroneous. Poor data quality can cause unexpected problems and induce major consequences on supply chains. Based on 40 case studies, Strong *et al.* (1997) presented 10 types of problem that can lead to poor data quality, from multiple data collection sources to a subjective evaluation. Redman (1998) shared the results of three cases where the cost of poor data quality was estimated to have decreased revenues by 8% to 12%. The article also listed other typical impacts, such as lower satisfaction from employees and customers, and a higher difficulty in changing or improving processes that involve poor data quality. Redman (1998) also estimated a typical database to have at least 1% to 5% inaccuracies. Moges *et al.* (2013) surveyed 150 financial institutions about their knowledge of data quality. Most of them estimated that 10% to 20% of data used in their decision-making process were erroneous.

Parssian *et al.* (2004) proposed two factors that could explain why organizations do not consider data quality, despite its importance in the decision-making process. The first factor concerns the cost of controlling and managing data quality over time. The level of data quality is not fixed, it evolves and should be re-evaluated often, which implies constant work. A second factor is the context-specific nature of data quality. While various assessment techniques are proposed in the literature, an evaluation should always be adapted to its context (Huh *et al.* (1990), Chiang and Miller (2008)).

The motivation behind this research originates from a forest products company located in Quebec, Canada. This company collected data describing the characteristics of its harvestable trees on a five-year cycle and then used this information to plan the transportation of the trees from the forests to its mills as well as the production at these mills. As the company used the data 'as is', the predicated wood supply was typically different from what was really available, leading to constant modifications of transportation and production plans. The company was therefore looking for a way to better know how to navigate with this insufficient data quality, to compare the accuracy of its forest sectors, and to better understand the data they had in hand.

This article thus focuses on data quality classification to offer insights on the result of a data quality assessment and to support decision making under uncertainty. The two main research questions investigated are summarized as follows:

RQ1: How can the quality of data be classified?

RQ2: How can quality data classes be used to benefit decision making under uncertainty?

To answer these research questions, the Action Design Research (ADR) methodology (Sein *et al.* 2011) was applied. The classification method was designed in response to the needs of partnering forest products companies. According to ADR principles, all decisions were made with their collaboration and followed guidelines from the literature. The classification method thus allows us to answer the first research question by taking the results of a data quality assessment and gives insight on the level of quality according to the organization standard (RQ1). The method is validated with an application to a theoretical case. To demonstrate how data quality classes can be used in decision making (RQ2), the quality classes of a practical case study are used to solve a transportation problem. These last tests allow us to assess the benefits of including data quality classes in decision making.

The contribution of this paper can be summarized as follows: First, the proposed method enables using data quality classification as a qualitative representation instead of data quality metrics. This is in line with the research of Moody (2003), who suggested that using classes of data quality is more intuitive and easy to use than quantitative measures. The second contribution is the concept that data classification should not be static. The case studied in this research suggests that the level of data quality is dependent on the current level of data and should evolve over time, highlighting the need for an adaptive classification method. The data quality assessment and classification should be viewed as a recurrent process. The final contribution is to present how levels of data quality could be used to assess uncertainty in decision-making processes. Lower quality suggests

higher uncertainty as pointed out by Zhu *et al.* (2012). Including data quality level in decision making could help in reducing the impact of the uncertainty coming with poor data quality.
The article is structured as follows. First, a literature review on data quality is presented, covering the uncertainty concept as it relates to poor information. The research methodology is then detailed to highlight how the method has been developed. The proposed adaptive classification method is then described, followed by application to a theoretical case and to supply data in forestry. The result of the classification method is then demonstrated with a transportation problem. The article concludes with recommendations and managerial insights for future use.

2 Literature Review

Before considering the classification of data quality in an uncertain context, there is a need to review the literature and see the state of research in the related fields. Data quality is usually described based on different dimensions of quality. Since the research discussed in this article focuses on the classification of data quality to improve decision outcomes, it is important to consider how data quality has been used in decision making in previous studies. Furthermore, the relation between data quality and planning uncertainty has captured researchers' attention in recent years. The literature review proposed here therefore summarizes these key concepts.

2.1 Quality Dimensions

Ballou and Pazer (1985) were among the first to introduce a set of dimensions to describe data quality, which were accuracy, timeliness, completeness, and consistency. Since then, additional dimensions have been proposed. Erroneous data can be described as inaccurate, inconsistent, outdated, incomplete, difficult to understand, duplicated or irrelevant. These are a few of the quality dimensions on which data quality can be evaluated. Wang *et al.* (1995) reported 25 such dimensions. There have been many attempts to organize and define those dimensions. For their part, Wang and Strong (1996) proposed four categories of data quality: intrinsic, contextual, representation, and access. The intrinsic category includes dimensions inseparable from the data itself, such as accuracy and believability. Contextual dimensions are related to the intended use. This includes completeness of the dataset, timeliness, and whether the amount of data is appropriate. The representational category regroups all dimensions describing the meaning of the data, such as consistency and understandability. The last category represents the ease with which data can be accessed. Chu *et al*. (2001) proposed a change to this division of quality. Instead of the intrinsic category, they argued that 'believability' incorporates accuracy, completeness, and consistency. Merino *et al.* (2016) grouped 15 dimensions into three categories related to the format, the period of time and the intended use. The major difference here is that some dimensions, like accuracy and consistency, are present in more than one category. Fox *et al.* (1994) proposed evaluating data quality based on four principal dimensions: accuracy, completeness, consistency, and timeliness. To enhance existing approaches, Heinrich *et al.* (2018) presented five requirements that metrics should fulfill, which can be applied to different dimensions.
Accuracy is the closeness between reality and the information in the system (Ballou and Pazer, 1985). Thus, performing a quantitative study on accuracy aims to evaluate the number of different values between two sets of data describing the same elements. Another way to represent the level of accuracy is to evaluate how many acceptable data are present in a dataset (Ballou *et al.*, 2006). However, if the true values are unknown or no longer available, accuracy can then be difficult to measure (Fan, 2015). To solve this issue, Cao *et al.* (2013) proposed a method to determine which of the two values is more accurate by using a chase algorithm to deduce accuracy rules.
Completeness is the dimension used to verify if a set of estimated data contains all the relevant information, so there will be no need for additional information. Here, the question is not if all the characteristics are included in the database but rather if all the entries are valid. Fox *et al.* (1994) identified two possible kinds of incomplete or invalid data. An entire entry can be missing from the database or values can be missing from an entry of the database. In both cases, incomplete information engenders uncertainty since they give rise to doubt for the decision maker. To evaluate completeness, Fan *et al.* (2009) proposed testing the database with

a set of queries or matching rules. The performance will depend on the capacity to respect the rules using only the data. Thus, if the data is missing or if there is more than one entry in response, there is incompleteness.
Consistency, sometimes named 'integrity', is the capacity to have coherent and similar data everywhere. The idea here is to verify if the estimated database contains logical information. When there is inconsistent data, it is difficult to trust information, thus causing uncertainty and insecurity towards the decision. Since it is harder to see if information is consistent, Fan (2015) suggested using data dependencies based on the need of the decision maker to define consistent data. This way, all entries respecting those rules are consistent and all others are not. A lot of different classes of rules can be found in the literature depending on the context studied. For example, the work of Chiang and Miller (2008) presented a discovery algorithm to facilitate the creation of conditional functional dependencies, a widely-used technique based on semantic quality rules.
Timeliness corresponds to the use of up-to-date data. Depending on the decision level, the 'age' or 'currency' of the information should be appropriate. Strategic decisions can be based on older information, while operational planning needs more recent data (Ballou and Tayi, 1999). If the data are too old, it is likely that the situation has since changed and the result emerging from the decision-making process may be wrong. Timeliness can be defined using three events (Blake and Mangiameli, 2011): when the change happens in the real world, when the data is entered in the system and when the information is used. It is difficult to know which data are out-of-date before they are updated. As stated by Heinrich and Hristova (2016), this makes this dimension quite different from the others mentioned in this section. In fact, Heinrich and Klier (2015) presented timeliness as a probability and Wechsler and Even (2012) described timeliness as the likelihood for certain data attributes to transition between states within a given time period. Another particularity of timeliness is the natural quality decline since the information may become outdated with time (Zak and Even, 2017).

2.2 Data Quality in Decision Making

As stated by Fisher *et al.* (2003), "decision making is a response to problems". Decision makers need to use available data to predict possible outcomes and then choose the preferable ones. They need the best data possible to have an accurate view of reality and be able to consider all alternatives. However, it has been shown that unless a company has made extraordinary efforts to improve data quality, an inaccuracy rate ranging from 1% to 10% is to be expected (Blake and Mangiameli, 2011). The authors confirmed that such deviation can have a huge impact on the decision-making process, especially when such inaccuracy rates are unknown. In addition to the cost of a 'bad' decision, which will vary depending on the impact of the decision, there are several drawbacks to poor data quality. People using erroneous data will inevitably develop organizational mistrust and frustration with their work. From a managerial point of view, bad data can reduce the ability to develop and implement strategies since the decision-making process becomes ineffective. From a financial point of view, Sheng and Mykytyn (2002) highlighted the importance of data quality in service organizations by showing that inaccuracies and inconsistencies could lead to a 40% to 60% increase in expenses.
An important part of using data quality to support decision-making concerns the choice of the assessment technique. A review presented by Batini *et al.* (2009) described the numerous techniques available in the literature for data assessment. Most of these techniques aim at describing and improving data quality rather than supporting planning decisions. Woodall *et al.* (2013) compared the most popular techniques to present the important steps an assessment technique should include. Such a technique should encompass the selection of the data items to be evaluated, the identification of data dimensions to consider, the metrics (how to measure the dimensions) and reference data (a set of true values). In parallel, the place where data must be measured has to be selected to know where to apply the metrics during the measurement. Finally, the results must be analyzed before any use. Some studies in the literature specifically looked at how measuring and improving data quality (Heinrich *et al.* (2018), Heinrich and Klier (2015), Lee *et al.* (2002), Wechsler and Even (2012)). Some studies furthermore examined the benefits of including data quality in the decision-making process. Chengalur-Smith *et al.* (1999) wanted to verify whether knowing about the quality of data improved the outcomes of a decision or not. They compared the choices made by managers given three levels of information on data quality: a two-point ordinal scale ('good' or 'bad'), a percentage, and no information. While their

survey confirmed their assumption, they also found that more complex situations are more susceptible to information overload. In their study, Parssian (2006) addressed the problem of poor data quality by replacing inaccurate, incomplete or null data with maximum likelihood estimates. Depending on the chosen average value, the sum of the dataset with substitution was close (87%-96%) to the sum of 'real' values. Wechsler and Even (2012) showed the potential of assessing timeliness to support data quality management tasks. They proposed using this dimension to predict accuracy degradation and improve data quality auditing. Furthermore, Heinrich and Hristova (2016) proposed an extended timeliness metric to support decision makers in adjusting their decisions based on indications about the current real-world information at the time of measurement. They also discussed the impact of data quality on the choices made by the decision maker. They described this aspect as extremely important, especially in the case of decision making under uncertainty.

Even though the subject has long been studied (Hoare, 1975), ever-changing technology brings new challenges for research on data quality (Zhu *et al.* 2012). The way data are generated, stored, manipulated, and consumed is constantly evolving (Staegemann *et al.* 2019). One of the challenges in the interpretability of data quality is to describe what is considered 'good' data quality, since good quality greatly depends on the subjective opinion of the decision maker (Huh *et al.*, 1990). A recent study found that more than 51% of surveyed organizations consider their data management technology to be ineffective, which shows the importance of taking data quality into account (Bai et al., 2018). Zak and Even (2017) stated that the potential damage cause by data quality defects is rising, impacting organizations without robust and economically sound data quality management processes. However, deciding whether data quality is acceptable can be a challenge. Since there are many dimensions used to describe data quality, it is generally recommended to base quality criteria on the relevant dimensions of data instead of considering them all.

The challenge in including data quality in a decision-making process resides in the interpretability. The responsibility of defining what is good data quality often falls to the user. However, as studied by many authors (Chengalur-Smith *et al.* (1999), Moody (2003), Holden *et al.* (2005), Merino *et al.* (2016), Plotkin (2020)), good quality is not easily defined. Vaziri et al. (2019) explained that some data may be more significant than others for the organization, which will influence the interpretation. They proposed to use weighted metrics to reflect this particularity. Woodall *et al.* (2019) explained that using tags to inform the decision makers about the quality of the data they use has not been successfully adopted despite their potential. The reason behind this would be the time consumption and expenses of continuously measuring data quality. Their research using surrogate tags showed that decisions makers can avoid problems caused by inaccuracies without having to physically measure accuracy.

When looking at the application of data quality, Moody (2003) suggested that problems with data quality once discovered by users should be prioritized as 'Critical', 'Important' and 'Desirable'. Their survey showed that representing data quality as a metric does not necessarily meet the needs of an analyst. Holden *et al.* (2005) used classification to represent the state of data quality in the evaluation of flavonoids for different kinds of fruits and vegetables based on 475 studies on the subject. They proposed a scale based on the percentage of overall data quality. The classes were 'Exceptional' (75%-100%), 'Above average' (50%-74%), 'Average' (25%-49%) or 'Below average' (below 25%). They found that 64% of studies were in the top two classes, which they considered good. Their work showed that such a classification is an easy way to understand data quality and support decision making. Merino *et al.* (2016) assessed the level of data quality instead of a more precise metric. They suggested comparing the quality level of evaluated data to the quality requirement to know whether the data are appropriate to make decisions or not.

2.3 Data Uncertainty

In this study, the uncertainty considered concerns the availability of forest supply. Precisely, it describes the difference between the supply the company thought would be available and the actual wood volume that can be harvested at the time of resource extraction. As presented by Ferson and Ginzburg (1996), there are two categories of explanation of this difference. The first part is related to variability, mostly inevitable since it is related to uncontrollable events such as weather, natural disasters, disease, etc. The second type is related to ignorance and can be reduced by a company with additional effort or investments. In our case, there is a part

of the uncertainty observed that is related to data quality, which is a type of ignorance since all the necessary information to assess data quality is available in databases, but still unused.

In their study on extending the timeliness metrics, Heinrich and Hristova (2016) considered the uncertainty as the unknown state of nature that will occur after the decision is made. They presented two kinds of such uncertainty: environmental and quality. Environmental uncertainty describes the unknown effect of a decision, which can be reduced by having access to more information. Quality uncertainty is related to the correspondence between the world described by the data and the real world. Heinrich and Hristova (2016) concludes that quality uncertainty is relevant to real-world decision making, which is what was considered in this research.

In their review on data quality research, Zhu *et al.* (2012) pointed out that many authors state that poor data quality leads to uncertainty. The list of impacts described by Redman (1998) suggests a connection between poor-quality data and uncertainty. The paper explained that even though all decisions are subject to uncertainty, companies using good-quality data have a better chance of reaching their intended goal. As described by Klibi *et al.* (2010), any decision made with partial or imperfect information is subject to uncertainty. Certainty can only be attained when perfect information is available. As a matter of fact, data-driven techniques are often used by researchers to illustrate or include uncertainty in the decision-making process. For example, Zhang *et al.* (2016) used data quality to define confidence intervals of uncertain parameters. However, in most cases, such data analysis seems to be limited to the accuracy dimension.

As there are many types of uncertainty, their effects might not necessarily be similar. Snyder and Shen (2006) looked at the difference between supply and demand uncertainty in a simple multi-echelon supply chain. They used simulations to demonstrate that the optimal strategy to cope with the impact of uncertainty will be different depending on the type of uncertainty. Chopra et al. (2007) separated supply uncertainty into two sub-types; disruption uncertainty, related to unpredictable events causing shortage and stopping operation, and recurrent uncertainty, where the effect is less severe but always present. They also concluded that each subtype should be dealt with independently to effectively reduce uncertainty. There are ways to minimize the impact of uncertainty. In regard to demand for example, a contract or certain incentives can help a supplier to have more predictable orders (Hu and Feng, 2017). It is also possible, at a certain cost, to use preventive maintenance actions or more sophisticated forecasting methods and additional equipment so as to improve measurement (Zyngier, 2017). Even though many sources of information used in the decision-making process can induce uncertainty, it is not always needed to consider all of them when modeling a system. To identify the sources worth considering, Pietilä *et al.* (2010) proposed evaluating inoptimality losses. This technique allows evaluating the negative effects occurring when the uncertainty of a particular information is ignored. In their study, they used the inoptimality losses to see the effect of growth uncertainty in harvesting decisions in forestry. They were able to put a price on this uncertainty: 230 euro/ha for 5-year inventory intervals and 860 euro/ha for 60 years. To manage all sources of uncertainty, Walker *et al.* (2003) suggested a general classification method, defining uncertainty based on three factors: location, level and nature. The location depends on where the problem manifests itself during the decision-making process. The level of uncertainty represents how much is unknown on the subject. The nature is similar to the variability/ignorance difference proposed by Ferson and Ginzburg (1996). Those studies support the relation between data quality and data uncertainty.

3 Methodology

The objectives of this study are to evaluate data quality using a classification methodology and to present how such a method could be included in the decision-making process. The methodology followed is based on the ADR methodology (Sein *et al.* 2011). ADR supports the conception and the evaluation of information technology artifacts, or decision tools. The methodology deals with two challenges: (1) addressing a problem situation from a specific organizational setting and (2) constructing and evaluating a IT artifact typical to the identified situation. Figure 1 summarizes the ADR method with the principles behind each stage followed during this research. Activities and results represent how the phases and principles were applied in this specific study.

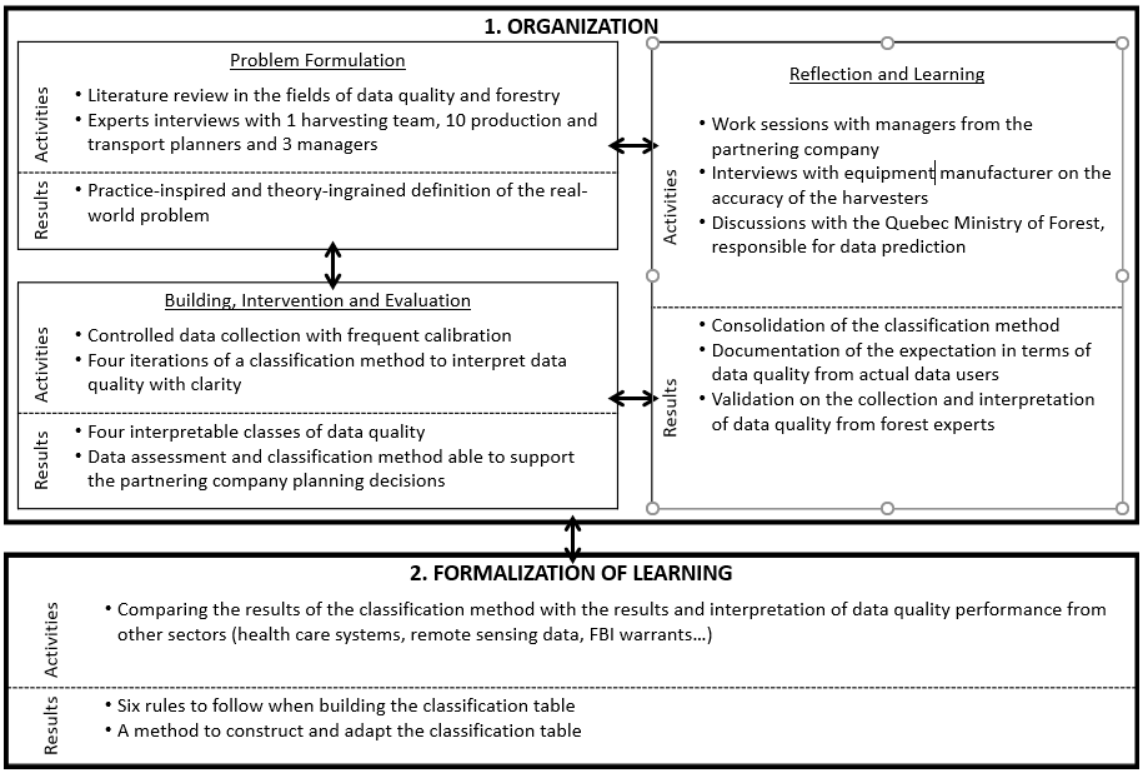

Figure 1: Stages and principles of the ADR method structure based on Dremel *et al.* (2020) and Sein *et al.* (2011).

3.1 Problem Formulation

The problem formulation stage is based on two principles: practice-inspired research and theory-ingrained artifacts. On the one hand, the classification method presented in this research is greatly influenced by the practical aspect of data quality. On the other hand, the method is constructed based on data quality assessment from past studies and other theories from the literature.

As mentioned previously, the origin of the project comes from a forest products company. The company collected data describing its uncertain supply activities for years but did not know how to use it efficiently. Even after using a data quality assessment technique adapted to its specific need, how to interpret the data quality obtained was still an issue. This led to the definition of the problem and the research questions presented in the introduction section (RQ1 and RQ2).

Regarding the principle of "Theory-Ingrained Artifact", the literature review helped to understand what had been done on the subject and to define the direction of the study. The review showed a gap in the literature that could be filled by a support tool regarding the interpretation of data quality.

An important part of the ADR method concerns the organization's implication. During this research, two employees from the forest products company were actively involved: the forest operations coordinator and the superintendent of forest analytics. They helped to prepare the experimental dataset. In particular, the forest operations coordinator was responsible for supervising measurement technique and calibration and the superintendent was responsible for preparing the predicted supply data. In addition, interviews with employees from a harvesting team were performed to validate the quality of historical data, collected at the time of harvesting by the equipment. Further interviews with 10 production and transport planners helped better understand the direct impact of uncertainty over their planning decisions, which they had to adjust daily because of unreliable predictions. Discussions with managers brought other impacts forward such as the cost related to the difficulty to know what will be produced or harvested in advance.

3.2 Building, Intervention and Evaluation

The second stage involves the development of the initial design of the method. This is an iterative process where the method is being continually evaluated based on the definition of the problem (previous stage). Three principles were followed in the building, intervention, and evaluation stage, as presented in Figure 1. The method was shaped reciprocally by the organization context and the data quality concepts. The interpretation of data quality should always depend on the context; thus, the method was developed jointly with the partnering organization. In the same manner, the researchers, and the forest products company representatives both had influential roles in the process. The two employees from the company were present in every step of the project. They shared their experience to improve the method and make sure it integrated their knowledge of the supply data. This is also part of the third principle, authentic and concurrent evaluation. As prescribed by the ADR method, the method was evaluated while it was being built. This was executed by discussing every change with the forest products company employees. The company greatly contributed to the research by assuring a controlled data collection including frequent calibration and supervision over 3 selected harvesting teams for over a year.

The first iteration of the project focused on the analysis and correction of the data quality assessment technique. The first set of data evaluated with the company had to be corrected before being analyzed. Some sets had no GPS coordinates, so they could not be grouped geographically, while in other sets the types of trees were erroneous. After identifying and correcting inconsistencies, both the research team and the company were satisfied with the quality performance. In the second iteration, the average data quality was compared with the classification table proposed by Blake and Mangiameli (2011). According to the organization, the general classification did not represent the state of their data quality. This experiment showed a need to construct quality classes based on past data qualities. Thus, in the third iteration the rules of a good classification table were defined based on the literature and on interviews with the employees. Their experience on how they see their own data, they predict the classification results, and they deal with poor quality inspired the rules proposed to create the quality classification table of the classification method. In the last iteration, the method was defined by testing the classification table on the results of past research papers performing data quality on real case studies. The last version of the classification method was presented to the company to be sure that it was still in line with their needs.

This phase resulted in a classification method offering four interpretable classes of data quality able to support the partnering company in their planning decisions. The data assessment methodology followed to evaluate the data collected by the company is also made available for future uses (Simard *et al*, 2019).

3.3 Reflection and Learning

The third stage of the ADR method is conducted in parallel with the first two stages. Throughout the research, reflection occurs on past and future decisions. There is also continuous learning since the literature review grows with each question raised. This is commonly known as guided emergence. Answering a problem specific to the situation of the forest products company and including the organization in the conception ensured that the classification method would be adapted to the intended use.

After each iteration, the design of the classification was challenged: Is the classification easy to understand? Does it reflect the reality of the organization? How is it different from other classifications found in the literature? And especially, are the results useful for decision makers?

The questions were brought up in work sessions with managers from the partnering company where their answers resulted in growing documentation on their expectation. Interviews with the harvesting equipment manufacturer were performed to further validate the data collection technique. The classification method was also discussed with a representative from the Quebec Ministry of Forest, the entity responsible for the forest data predictions. These discussions and interviews helped to consolidate the classification method.

3.4 Formalization of Learning

The objective in the second part of the methodology is to use what was developed specifically for the organization and to generalize the outcome. It is an important step of the ADR method to ensure that the solution of a unique problem leads to a concept that can be applied to other cases.

In this research, the state of the overall data quality is always changing. It can improve following investment or training on data measurements, but it can also worsen with relaxed calibration protocol or with worn-out machinery. Therefore, the classification is constructed to evolve with the company and be re-evaluated when needed.
In an effort to define a classification method that can be adapted to different contexts, six logical rules were proposed on how the quality classes should be defined. The classification method was used to define quality thresholds for a set of data quality assessed in 29 studies presenting real data quality assessment, tested over a different set of data.

4 Adaptive Classification Method

The key results of this research are a classification method to analyze data quality and the exploitation of the method in decision-making processes. The method developed (Figure 2) is an iterative process, including five steps which should be repeated when needed (e.g., if the state of data quality changes, when the decision-making process is executed, etc.). The method is intended for decision makers or data analysts and should be part of the decision process. According to the guidelines of a typical data quality assessment technique (Woodall *et al.* 2013), the classification technique would be used in the 'analysis' step. In an organization, it is assumed that changes in data quality will be rare. The typical use of the method should thus involve the key steps appearing in the gray area in Figure 2.

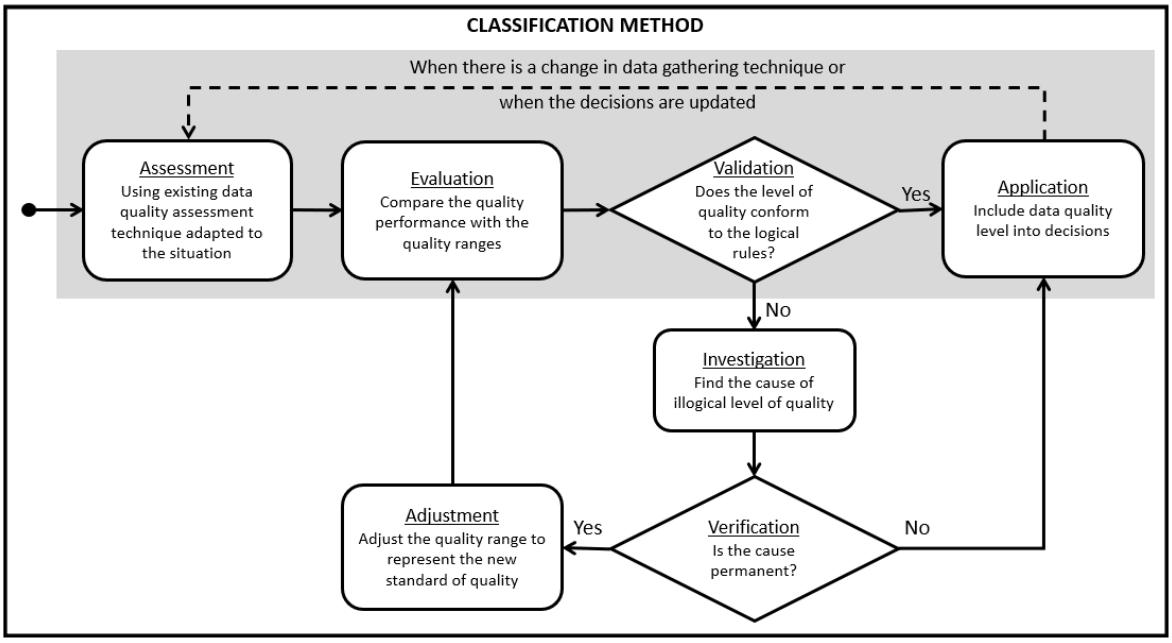

Figure 2 : Method to classify data quality.

The proposed method begins with the **assessment of data quality**. Multiple techniques exist as presented in the literature review section to conduct this type of assessment. It can, moreover, be sometimes preferable to create a custom technique to better reflect the needs and reality of the context considered. Woodall *et al.* (2013) presented many guidelines on how to build a data quality assessment technique. One important part of the assessment concerns the choice of the quality dimensions. This choice will affect the classification technique since there has to be one set of quality classes for each dimension selected. Another important factor is data clustering. The objective of the method is ultimately to support the decision-making process. If the information is too aggregated, it may not be suited for targeted decisions. For example, knowing the overall accuracy of 20 types of products may be too aggregated to represent the actual accuracy of each individual

product. In this situation, the decision maker should evaluate the data quality of each product and compare the 20 resulting quality performances in the classification.

Once the data quality is known, it is then possible to **evaluate the level of quality** by comparing the results with the quality table. This is a procedure where the users will consider the percentage of one dimension of data quality and see the associated class according to the latest version of the quality table. The first time the method is applied, it is suggested that a general definition of quality classes be used like the one presented in Table 5. The quality table considers four principal dimensions (Fox *et al.* 1994), which are accuracy, completeness, consistency, and timeliness. The method could be applied to other quality dimensions as long as they are presented as percentage scale, where 100% is the best quality possible. The first time the method is followed, beginning with the classes from Holden *et al.* (2005) is suggested, which is the same for all dimensions and covers all possible value (from 0 to 100%). Blake and Mangiameli (2011) also proposed a general classification table, although their evaluation is much more restrictive. For example, their table does not consider accuracy below 80%, which can be difficult to achieve in certain settings. The general classification tables found in the literature may be based on theoretical guidelines (Blake and Mangiameli, 2011) or on subjective views (Holden *et al.* 2005), but they are still a good base to construct a table adapted to the specific situation of an organization.

After each evaluation, results should be **validated**. Data quality can change at any time and the classification should not be considered an immutable table. Plotkin (2020) presented the importance of defining a set of rules as an important part of establishing the state of data quality. Based on the guidelines of similar studies (Mezzanzanica *et al.* (2015), Hartig and Zhao (2009), Holden *et al.* (2005), Blake and Mangiameli (2011)), we formulate six logical rules to validate the process, presented in Table 3. If one or more rules are not true, then the classification may need adjustments.

Table 1 Logical rules and their motivations

ID	Logical rules	Motivation
R1	The majority of data should be judged ‘Good’	The ‘Good’ class represents a stable acceptable level of data quality.
R2	There should be more data judged ‘Good’ than ‘Excellent’	The ‘Excellent’ class should represent the highest level of quality.
R3	There should be more data judged ‘Good’ than ‘Sufficient’	The ‘Sufficient’ class represents less than desirable but still acceptable level of quality.
R4	There should be more data judged ‘Sufficient’ than ‘Excellent’	The ‘Excellent” class should be an exceptionally high level of quality, thus smaller than the ‘Sufficient’ class which is closer to the normal quality.
R5	There should be less data judged ‘Insufficient’ than in any class	The ‘Insufficient’ class should represent a problematic level of quality and be the subject of improvements.
R6	The ‘Excellent’, ‘Good’ and ‘Sufficient’ classes should not be empty	Only the ‘Insufficient’ class can be empty if the organization judges that there is absolutely no need for improvements.

In addition to the logical rules presented in Table 3, guidelines on data accuracy percentile can help in deciding on the quality thresholds. Data with accuracy in the 80^{th} percentile and over could be considered as ‘Excellent’ while accuracy between the 40^{th} and the 79^{th} percentile could be perceived as ‘Good’. This way, most data at the time of the evaluation would represent the best quality available for the company. Following the logical rules, it is suggested that data entry with quality between the 5^{th} and the 39^{th} percentile be considered ‘Sufficient’ and the rest of the dataset ‘Insufficient’. It would be preferable to adjust the proposed percentiles than to separate similar data. For example, two data entries with accuracy performance of 75.6% and 75.9% should be classed in the same category even though they represent the 79^{th} and the 80^{th} percentile.

The main idea behind the classification method is to follow the general state of data quality. When the results are incoherent or stop following logical rules, it is time to **investigate** the data-gathering process. This does

not mean that the classification table will have to be adjusted systematically. The idea is rather to **verify** if the difference in data is permanent and will impact the planning decisions. For example, if the data quality becomes worse because there is an equipment malfunction, the situation can be resolved without having to adjust the quality classes. Following the same idea, if there was an investment to improve data quality but the change did not contribute to upgrading the data quality level from 'Good' to 'Excellent', then there is a need for investigation. In this case, if the situation is improved, the classification should also be **adjusted** to reflect the new range for 'Excellent', 'Good', 'Sufficient' and 'Insufficient' data.
If the validation shows that the classification respects the logical rules, then the quality levels can be **included in the decision-making process**. It can be difficult for decision makers to know how to interpret quantitative data quality regarding their own needs. As can be seen in the theoretical case in Section 5.1, an 80% accuracy can be considered as poor quality in certain contexts while be seen as exceptional in others. The classification method is used to present which set of data can be judged 'Good' to decision makers according to their own standards. When facing uncertainty, this can show which data they can 'trust' more. This way, decision makers could use suppliers with 'Good' or 'Exceptional' data to assure the delivery of critical orders. They could decide to distribute tasks so that each member of a supply chain deals with a similar level of data quality. They could also work with their suppliers to improve data quality by using a qualitative scale that can be easily understood. Those are only a few examples of how knowing and using data quality can help decision makers to have a better understanding and a better control over their decisions. To be effective, the classification method should become an automated data processing step for formatting the information in a compatible format. An example of quality level applied to the stochastic optimization of a simple transportation problem subject to uncertainty is presented in Section 5.3.

5 Demonstration

To better illustrate how the proposed classification method could be exploited, we present three examples of applications. The first one concerns a theoretical case where the classification method is applied to the results of data quality assessment found in the literature. This case represents a normal application of the method, following the steps in the gray area in Figure 2. Secondly, a practical case study where the method is applied to historical data describing the uncertain supply of a forest products company is proposed. This case demonstrates a thorough application of the method where the classification needs to be adjusted. Finally, a demonstration of how the level of quality obtained with the method could be included in a transportation problem reducing costs is presented as a third application.

5.1 Theoretical Case

Using the classification method on a theoretical case was seen as a proper way to validate the method. For each of the four principal dimensions presented in the literature review (i.e., accuracy, completeness, consistency, and timeliness), articles on the application of data quality assessment were selected in order to construct a theoretical case. The resulting dataset is used to demonstrate the importance of an adaptive classification method over a general table. This is achieved by comparing the results of the proposed method to two general classification tables (Blake and Mangiameli (2011), Holden et al. (2005)) and by taking into account the experts' insights on the expected level of data quality.

Assessment: The studies considered here come from different contexts, from patient records in health care systems to remote sensing data and FBI warrants. Although the four dimensions were all considered during the theoretical case, only the accuracy evaluation is presented in detail in Table 4. During the assessment of data quality, a chosen technique is used to evaluate the quality of one or more sets of data. In the theoretical case, the dataset comes from the assessment of past studies, as shown by the 'Accuracy' column in Table 4. The next step of the classification method uses a quality table to evaluate the quality level each accuracy value represents, which can also be seen in Table 4. A description of the evaluation step follows.

Table 2 Levels of quality for past accuracy assessments

Source	Accuracy	Quality level		
		Proposed Classification	Blake and Mangiameli (2011)	Holden *et al.* (2005)
Baker *et al.* (2007)	92.3%	Good	High quality	Exceptional
Barlow *et al.* (1994)	98%	Excellent	High quality	Exceptional
Barrie and Marsh (1992)	92.9% (worst)	Good	High quality	Exceptional
	97.8% (best)	Excellent	High quality	Exceptional
Congalton (1991)	74%	Sufficient	---	Above average
	78%-89% (improved)	Good	Medium quality	Exceptional
Faulconer and de Lusignan (2004)	75%	Good	---	Exceptional
Goodyear-Smith *et al.* (2007)	89% (worst)	Good	Medium quality	Exceptional
	97% (best)	Excellent	High quality	Exceptional
Hohnloser et al (1994)	74.1%	Sufficient	---	Above average
Lauren (1986)	93.7%	Good	High quality	Exceptional
Maresh et al (1986)	95%	Excellent	High quality	Exceptional
Morey (1982)	75%	Good	---	Exceptional
	84.7% (improved)	Good	Low quality	Exceptional
Persell *et al.* (2009)	72%	Sufficient	---	Above average
Powell *et al.* (2006)	68%	Sufficient	---	Above average
Pringle *et al.* (1995)	47.4%(worst)	Sufficient	---	Average
	96.7% (best)	Excellent	High quality	Exceptional
Ricketts *et al.* (1993)	44% (worst)	Sufficient	---	Above average
	67% (best)	Sufficient	---	Average
Saigh et al (2006)	45%	Sufficient	---	Average
Sigurdardottir *et al.* (2012)	96.4%	Excellent	High quality	Exceptional
Viscusi *et al.* (2014)	44% (overall)	Sufficient	---	Average
	78.4% (best)	Good	---	Exceptional
	5.4% (worst)	Insufficient	---	Below average
Wagner and Hogan (1996)	83%	Good	Low quality	Exceptional
Wilton and Pennisi (1994)	89.8%	Good	Medium quality	Exceptional
Yarnall *et al.* (1995)	62%	Sufficient	---	Above average
	82% (improved)	Good	Low quality	Exceptional

Regarding the other dimensions, the completeness evaluation presented less variability than the level of accuracy while for the consistency and the timeliness, only a few articles were found with the assessment of real case study. The complete table can be found in Annex 1.

Evaluation: Since the classification method is used for the first time for this set of data, the quality table needs to be created by following the logical rules from Table 3. First, the accuracy levels are evaluated using the Holden *et al.* (2005) classification. They considered 'Exceptional' quality between 75% and 100%, 'Above average' quality between 50% and 74%, 'Average' quality between 25% and 49%, and 'Below average' quality between 0% and 24%. Since this quality table does not follow the logical rules, the classes are adjusted. First, the threshold between 'Above Average' and 'Exceptional' is augmented to respect rules R1 and R2, from 75% to 95%. This will become the threshold between 'Good' and 'Excellent'. Then, for the 'Sufficient'

class, the 'Average' thresholds must be adjusted to respect rules R3, R4, and R5. This way, the 50% threshold becomes 75% and 25% becomes 40%. The thresholds should also assure that the fourth class, 'Insufficient' respects rules R6. This reasoning is repeated for all dimensions. The resulting range is presented in Table 5.

Table 3 : Proposed range of quality classes

Quality dimensions / Quality classes	Accuracy	Completeness	Consistency	Timeliness
Excellent	95-100%	95-100%	85-100%	85-100%
Good	75-94%	75-94%	50-84%	50-84%
Sufficient	40-74%	25-74%	25-49%	18-49%
Insufficient	0-39%	0-24%	0-24%	0-17%

The result of the evaluation based on the quality range in Table 5 can be seen in Table 4. To give perspective to the reader, it is compared with the evaluation using existing classification. In this study, two other comparable classifications were found, although they are not presented as an iterative way of including data quality into decision-making processes. Blake and Mangiameli (2011) considered for accuracy, completeness, and consistency 'High' quality between 92% and 100%, 'Medium' quality between 88% and 91%, and 'Low' quality between 80% and 87%. They did not consider the possibility of data quality being below 80%. At first glance, this classification range is inapplicable (indicated as '---') for 14 studies out of 29. It seems unrealistic to expect data quality to be above 80% in all cases. In comparison, the classification proposed by Holden *et al.* (2005) judged 18 out of 29 studies as having exceptional quality (62%). The classification proposed seems a reasonable middle ground between the two. For instance, it is interesting to look at the classification of improved data quality. After an improvement on data quality, it is assumed that the level of quality will also improve. For two of the three studies presenting improved data quality and presented in Table 5 (Congalton (1991), Morey (1982) and Yarnall *et al.* (1995)), the proposed classification reflects the benefit of the improvement by showing an increased quality class (from 'Sufficient' to 'Good').

Validation: Following the classification method, the results of past data quality assessment were validated using Table 3. The evaluation respected the logical rules. By defining the 'Good' accuracy class between 75% and 94%, as shown in Table 5, it includes the majority (41%) of studies in Table 4. The 'Excellent' accuracy class contains 21% of the studies in Table 4, so less than the 'Sufficient' accuracy with 34%, as supported by the logical rules. The only 'Insufficient' entry represents 3% of the total, as expected regarding the logical rules.

Application: Because this is a theoretical case and all contexts assessed are very different, there is no following decision-making process to put in place. However, this section of the method will be explored in Subsection 5.3.

5.2 Case Study in the Forest Industry

To demonstrate how to use the method for a practical case, data from a Canadian forest products company were collected and evaluated. Production in forest-based processes is based on a natural resource, therefore data describing the characteristics of the available supply source has a huge impact on the planning process. As stated by Duvemo and Lämås (2006), most planning activities in the entire value chain depend on forest-related data. Those will typically describe the species (e.g., fir, spruce, etc.) and dimensions (i.e., diameter and length) of harvestable trees from which the volume and other relevant information will be estimated. All decisions based on supply information are subject to multiple sources of uncertainty. First, the predicted volume depends on tree growth which is affected by uncontrollable events such as disease, insects, and weather conditions. Second, the trees are located in large areas that are difficult to measure with details (Pasalodos-Tato *et al.*, 2013). In addition, planning decisions in forestry are also affected by the uncertainty

caused by poor data quality. The risk of error is even higher when there is no direct control over data collection, as is the case for forest products companies operating on public land (i.e., forests owned by the government). Even though there is considerable research on forest harvesting, forest data acquisition and remote sensing, only a few articles have looked at the data quality aspect. In their review, Yousefpour *et al.* (2012) were looking at forest management, uncertainty from hazards, market fluctuation and climate change, but not at data quality. Yet, the evaluation of information could help cover the trade-off between data acquisition cost and the effect of better planning decisions in forestry. According to Holmström *et al.* (2003), the lack of interest towards data evaluation in forestry comes from the difficulty in evaluating the cost of poor quality of information. In addition, Kangas and Kangas (2004) described uncontrollable uncertainty with emphasis on the effect it can have on any decision of the forest sector and the importance of good information. As such, forestry is a perfect case study for the classification of data quality.

Assessment: To conduct the classification, three types of data were collected under the supervision of a forest operations coordinator. There were data collected by production machinery, data from computations, and data manually collected by employees. There were 600,000 entries, each representing a single tree (species, location, dimension, and class types), used to predict the outcome of forest harvesting operations. Those trees were clustered into 63 forests, each representing a harvesting area. Multiple characteristics provided by the company influence the upcoming wood processing activities. It was decided to focus on the harvestable volume available in the forest because this uncertain information greatly affects sawmill productivity. Out of the four-principal dimension of Fox *et al.* (1994), accuracy was the one selected. In fact, since each tree was only measured once, it was impossible to compare the evolution of quality and assess timeliness. Controlling the data gathering environment meant that all data entries were complete, thus the assessment of completeness was meaningless. Concerning consistency, the organization had no standard for their data and did not establish rules to know whether the data were consistent or not. Figure 3 presents the frequency distribution of accuracy for the 63 clusters of forests. Because the choice of data quality assessment is greatly affected by the context, a specific technique was created to assess the data describing the supply volume subject to uncertainty. The data quality assessment followed in this case was developed by Simard *et al.* (2019) for supply data in forestry.

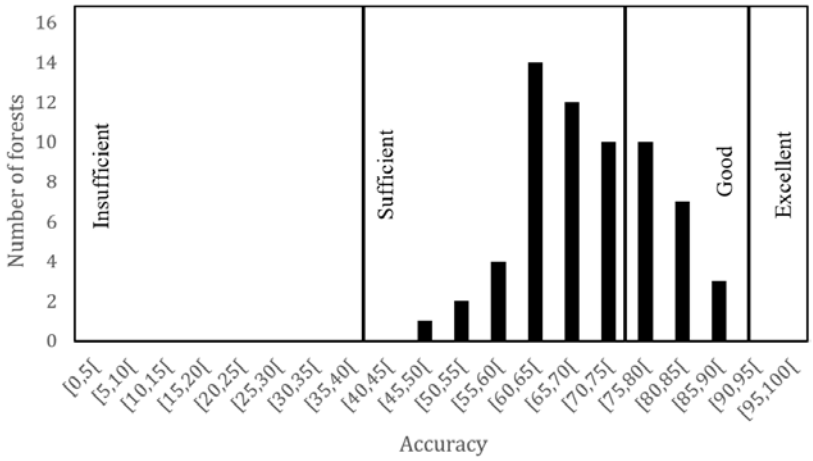

Figure 3 Frequency distribution of data accuracy divided by quality classes.

Evaluation: Following the assessment, the quality of each of the 63 data clusters was interpreted. By applying the range in Table 5 constructed in the theoretical case, it can be noticed that 68% of data are 'Sufficient' and 32% are 'Good'.

Validation: The evaluation showed that the classification does not respect the logical rules. The majority of data were judged 'Sufficient', which goes against two rules (R1, R3). No data were considered 'Excellent' or 'Insufficient', which once again does not respect the logical rules (R6). Although all other rules were respected, there is a need for investigation before including data in a decision-making process.

Investigation: The investigation involved interviews with company employees. The data gathering method and data assessment technique were inspected to identify any problem explaining the low quality. However,

the forest products company experts did not find their level of data quality illogical or surprising. An inherent problem in their situation is the data collected by harvesting contractors hired by the company to supply the raw material. They are not directly affected by the impact of low data quality and thus their dedication is not assured. This, in addition to the mentioned challenge of uncertain supply in forestry (predicting tree growth, large area, publicly owned land), explains why the theoretical quality class may not represent the reality of the forest products company and needs to be adjusted.

Verification: The cause of the change is permanent so the quality table should be modified. As mentioned previously, the first time the method is applied, an adjustment is expected to create a customized quality table.

Adjustment: Here, quality classes have to be adjusted to make sure that all the logical rules will be respected. The 'Good' class threshold was therefore reduced so the highest quality forests could be judged as 'Excellent'. Although it is possible that no data are judged as of 'Insufficient' quality, feedback from the organization showed that the actual level of quality would certainly need improvement. The 'Sufficient' class was thus adjusted so the lowest quality could be judged as 'Insufficient'. The theoretical and practical classifications for the accuracy dimension are presented in Table 6.

Table 4 Theoretical and revised accuracy classes

Quality dimensions / Quality classes	Theoretical accuracy	Revised accuracy
Excellent	95%-100%	80%-100%
Good	75%-94%	65%-79%
Sufficient	40%-74%	55%-64%
Insufficient	0%-39%	0%-54%

Evaluation and Validation: The data were evaluated a second time using the revised quality classes. Figure 4 presents the comparison with the previous classification. As the current classification respects all logical rules, data can be used in the decision-making process. Looking at the percentile, the results are not far from the one suggested in Section 4. All forest sectors in the 84th and above percentile are considered 'Excellent' and sectors between the 34th and the 83rd percentile considered 'Good'. Forest sectors between the 5th and the 33rd percentile are 'Sufficient' and anything below the 5th percentile is Insufficient'.

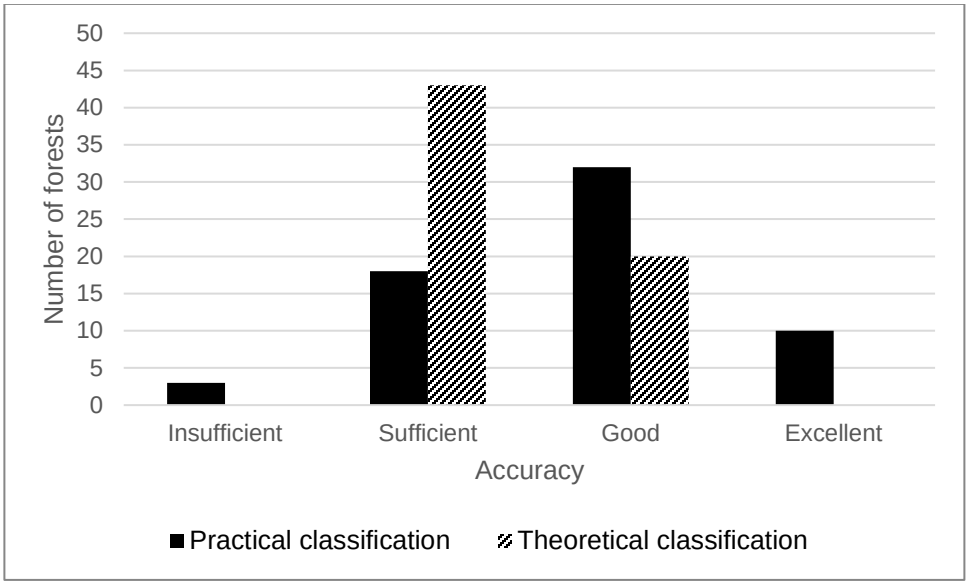

Figure 4: Evaluation of data accuracy with theoretical and revised classification

Application: With the classified data, users now have more insights on their data quality, which can support the decision-making process. For the organization, data accuracy becomes an indicator of the level of uncertainty related to their data. As this company sells its products following a push system before even producing them, it will now be able to use data accuracy to generate more representative production prediction. Knowing which forests are more reliable, the company will also be able to adjust its commitment to its clients accordingly. Regarding supply decisions, the company also planned on using quality levels to distribute uncertainty among the wood harvesting teams as some teams are never sent to harvest 'Good' forests. This may seem like a minor problem, but it causes dissatisfaction within the company and an unfair repartition of profit since the teams are paid based on their productivity. By knowing which forest has 'Sufficient' or 'Insufficient' quality, the company can make sure to divide the worst forests equally between harvest teams. Another example of application is presented in detail in the following subsection.

5.3 Applying the Classification Method for a Transportation Problem

The theoretical case and the practical case have shown how to classify data quality. To answer our second research question, we present an example of application which demonstrates how the classification method could be used to support planning decisions in a typical transportation problem. We also compare the use of the method with the cases where detailed data accuracy is considered and no strategy at all is put in place. The objective of the planner is to minimize the transportation cost based on an estimation of available supply volumes and a known demand. Linear optimization methods are used to simulate the decisions made by the planner. The network is a subset of the partnering forest products company, including the operations of 14 forests and 5 sawmills. Figure 5 is a representation of this problem.

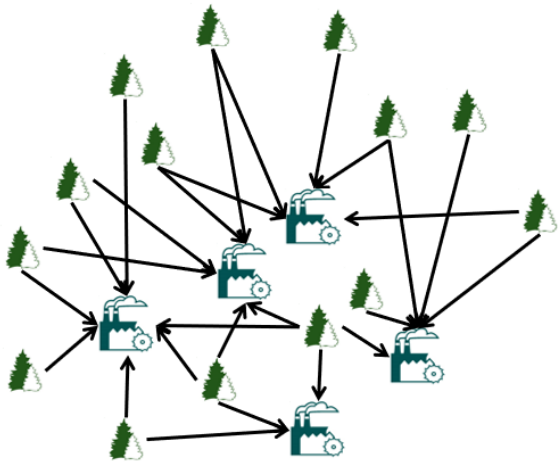

Figure 5 Graphic representation of a forest products company supply network.

Three types of instances are considered, each with a different way of using the predicted supply availability. First (I) the estimated volumes are used directly without consideration for data quality. The estimated data comes from the Quebec Ministry of Forests, the entity normally responsible for forest data collection. This first instance is represented by a deterministic model since the uncertainty is not taken into account. This instance is the closest to the actual planning process of the partnering forest products company. The second instance (II) uses the frequency distribution of the data quality accuracy to represent the behavior of the uncertain supply availability. Each forest is associated to a unique probability distribution of possible available supply based on the historic difference between what was estimated and what was harvested. For the last instance (iii), a probability profile is created for each quality class based on a normal distribution, where better quality offers smaller deviations and vice-versa. Each forest with the same quality is associated to the same probability distribution of possible available supply. Considering the supply as a distribution of potential available volume made the second and third instances impossible to solve using the same deterministic model developed for the first instance. A two-stage stochastic problem was therefore formulated based on the

deterministic model to solve these instances. It was solved by applying the Sample Average Approximation method which comes from Monte Carlo sampling (Zanjani *et al.* (2013), Santoso *et al.* (2004), Kim *et al.* (2011)). Results of all three instances are presented in Table 7. All costs coming from stochastic optimization are represented as an average over 100 scenarios.

Table 5 Average transportation cost and volume for the transportation problem investigated.

Scenario	I. Harvesting volume predicted manually (Without data quality)	II. Supply predicted with detailed accuracy (With data quality assessment)	III. Supply predicted with quality level (With classification method)
Predicted costs (Volume)	$3,733,030 (387,176m3)	$4,067,916 (386,157m3)	$3,855,333 (387,132m3)
Committed costs (Volume)	$1,390,480 (211,966m3)	$1,406,607 (213,151m3)	$1,438,043 (216,640m3)
Adjustment costs (Volume)	$2,345,060 (175,209m3)	$2,109,862 (174,024m3)	$2,202,228 (170,535m3)
Total costs (Volume)	$3,735,540 (387,175m3)	$3,516,468 (387,175m3)	$3,640,270 (387,175m3)
Comparison with scenario I	---	-5.86%	-2.55%

The **predicted costs** in Table 7 represent the predicted transportation cost between the forests where wood is harvested and the sawmills where the wood is processed. Although the forest products company may try to respect the transportation planning, the uncertain supply can make it impossible to do so because the estimated available wood volume is not enough to produce what was planned. There is a part of the predicted transport that will be executed without any issue, which is presented in Table 7 as the **committed costs**. For the missing supply, the company will have to pay **adjustment costs** so the sawmills will not run out of wood. The **total costs represent** the sum of the committed and the adjustment costs.
The planner usually prefers a solution in which more committed volume is considered, meaning fewer volume adjustments as they come at a higher cost. Looking at the results in Table 7, the scenario using data quality classification (III) seems to perform well. It is the most stable option since it offers more committed volume and less volume adjustment. However, because the adjustments cost more than the instance directly using data accuracy (II), it is not the option offering the lowest total costs. It is not surprising that the instance offering the lowest total cost is the second, since the supply is based on the detailed accuracy. While it may be possible to use past accuracy in this application based on historic data where the actual supply availability is known, it is a challenge to predict with certainty the accuracy of multiple forests. On the other hand, classification can be an easier way to predict supply variation and still show more potential than using estimated volume directly.

As an extension to the transportation problem, classified data quality could be used to create realistic scenarios over more than one company. Using data quality to describe clusters of similar trees could thus become a way of describing large, harvested area in a more detailed manner. Instead of expecting ±20% volume overall, each forest or area, could have its own uncertainty measure. An example of this concept is presented in Figure 6. There, each of the four identified areas has its own level of uncertainty, as shown in the legend. This differentiation could be used in governmental study and the resource distribution.

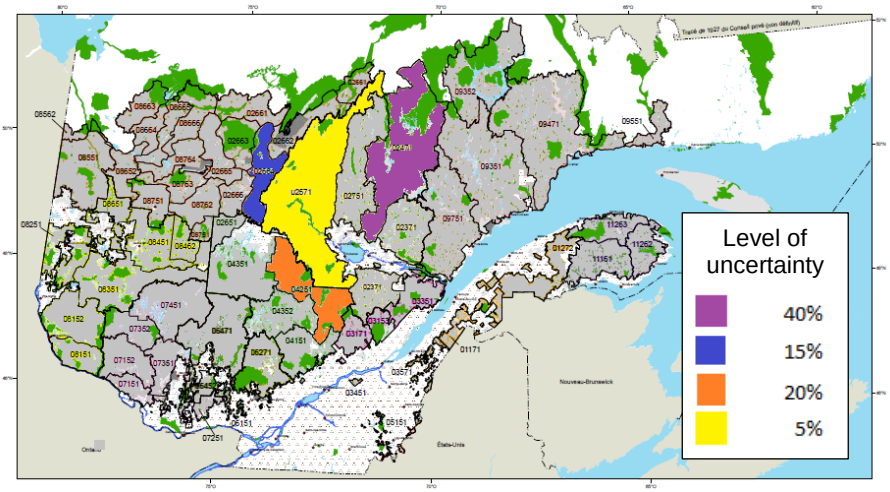

Figure 6 : Map of the forest management units in Quebec[1] as an example of how to use data quality to model uncertainty

6 Discussion

The article attempted to answer two research questions, conceding how to classify data quality and how to use it to benefit decision-making processes in uncertain context. The classification method presented in Section 4 encompasses steps to follow so as to classify data quality. An example of application of the classification method for a simple transportation problem then showed how the quality classes could be used to represent probability distributions of errors on uncertain data and how it could lead to a higher committed transportation cost (3.42%), a lower adjustment cost (-6.09%), and a reduced total cost (-2.55%).

Peffers *et al.* (2006) identified three objectives to evaluate a method like the one presented in this article. First, the research should produce a design consistent with prior research theory and practice. The method developed with the practical point of view of a partnering organization and with guidelines from the literature met this objective. Second, it should provide a nominal process. The method was indeed applied to a theoretical and a practical case. The comparison to a similar classification also showed that the method is effective for its intended purpose. Thirdly, the research should provide a mental model for the characteristics of research outputs. The method presented in Section 4 represents all steps to follow and could be applied without modification. This was demonstrated by the demonstration of application to a theoretical case and a practical case. The results of data quality assessment technique could create an overload of information (Fisher and Kingma, 2001). The addition of a classification phase can help the interpretation of data quality and simplify the inclusion into the decision-making process. The proposed classification model thus satisfies the three objectives and is considered a valid design.

The application of the proposed classification method to a theoretical case as presented here, confirms the need for an adaptive method. The quality classes constructed for the theoretical dataset (Figure 5) were a better choice for the dataset in Table 4 when compared with general tables. It covered all possible accuracy performance as opposed to Blake and Mangiameli (2011)'s table. It was also more aligned with the experts' insights, indicating when some accuracy performance was 'best' or 'worst' as opposed to Holden *et al.* (2005)'s table. However, when applied to the practical case study in Section 5.2, the classes did not reflect the current state of data quality for the context and a new table was constructed. This not only demonstrates that data quality is context-specific, but also that it should evolve with a company.

[1] The original map can be found on the government website: https://mffp.gouv.qc.ca/les-forets/amenagement-durable-forets/les-droits-consentis/lunite-damenagement-ua/

In this article, we presented one example where data quality can improve decisions, but there are multiple other opportunities to be explored. The quality levels could be used as a surrogate data quality tag to give information on data quality level to the decision makers without the cost and time necessary to keep typical tags updated (Woodall *et al.* 2019). For example, knowing the level of accuracy of data describing the resource availability can lead to better control over logistics. Fewer errors in supply and inventory data could also result in fewer changes and improved efficiency in production activities. Accounting for data accuracy in production systems allows better adjusting to upcoming changes. It could be directly linked to the level of safety stock that should be maintained at different points in the supply chain. Furthermore, more reliable planning means more accurate information to support sales activities, entailing a better customer service and possibly more revenue.

Regarding the limitation of our work, the classification method was constructed for the purpose of supporting decisions in procurement planning processes subject to uncertainty related to the predicted supply. It is possible that such method would be of benefit for other problems, such as decision in production planning or data describing demand subject to uncertainty. The research would also benefit from tests on case studies from other sectors where the uncertain supply has to be predicted, such as agriculture, bioethanol fuel, or textile industries.

7 Concluding Remarks

The objective of this paper was to propose a classification method for data quality that could also be used as a support for the decision-making process. By following the ADR methodology, the resulting iterative method is designed to classify the quality of data as 'Excellent', 'Good', 'Sufficient' or 'Insufficient' compared to typical organization standards. The method was applied to a theoretical case using data quality performance from the literature and to a practical case using historical data from a forest products company. A demonstration on the application of quality case to a transportation case showed the benefits of including data quality into decision-making processes. There are many approaches to data quality and decision-making under uncertainty. The method proposed in this article connects the two perspectives.

This article contributes to data-quality research in three ways. First, the proposed classification method provides an interpretation of the results of typical assessment techniques. Even though there are many ways to quantify data quality, a few articles propose efficient ways to include this information in the decision-making process. In today's context of big data, it can become overwhelming for decision makers to know the level of quality. Secondly, the adaptive classification method is developed so as to represent the state of data quality for an organization compared to its own standard. By comparing general classification tables with the proposed adaptive classification method in a theoretical case, it showed that data quality must be evaluated based on past quality levels. Quality levels are specific to each process and should evolve with a company, which becomes possible by using the proposed classification method. Applying the proposed classification method to a real-life dataset supplied by a forest products company also supported the need for quality levels which evolve according to their reality. The proposed classification method applied to the theoretical case and the practical one thus leads to different quality levels. A third contribution of this paper is to propose data quality levels to assess uncertainty in decision making. The qualitative representation can be used to facilitate the inclusion of data quality into algorithms supporting decision making, as demonstrated with the application to a transportation problem. Furthermore, the partnering company aims to use data quality levels to better distribute uncertainty among its harvesting teams, so that the negative impact of poor data quality will be spread between them. Two employees who have to deal with data uncertainty daily followed the project to confirm that the results obtained properly reflected their reality.

Further research could aim to expand the relationship between data quality and decision making. We presented an example of application to a transportation problem, but many other opportunities exist. The quality may be an asset in other types of problem like production planning, procurement planning, inventory management, etc. It would also be beneficial to expand the tests presented in this paper and consider a more complex transportation problem. Since the focus of this article was mainly the quality of supply data, it could be interesting to look at the demand. Regarding the chosen case study, the quality table used for the forestry data could be compared to the quality table constructed for other sectors dealing with supply uncertainty to verify if there is a trend in the quality of supply data. Another possibility of research could be to use data quality classes in a predictive model.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support from the Fonds de Recherche du Québec and the private and public partners of the FORAC Research Consortium.

REFERENCES

Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rules', 94 Proceedings of the 20th International Conference on Very Large Data Bases, 1215, pp. 487–499.

Bai, L., Meredith, R., & Burstein, F. (2018). A data quality framework, method and tools for managing data quality in a health care setting: an action case study. Journal of Decision Systems, 27(S1), 144–154.

Baker, D. W., Persell, S. D., Thompson, J. A., Soman, N. S., Burgner, K. M., Liss, D., *et al.* (2007). Automated review of electronic health records to assess quality of care for outpatients with heart failure. Annals of Internal Medicine, 146, 270-277.

Ballou, D., and Pazer, H. (1985) 'Modeling data and process quality in multi-input, multi-output information systems', Manage Science, 31(2).

Ballou, D. P., Chengalur-Smith, I. N., and Wang, R. Y. (2006) 'Sample-based quality estimation of query results in relational database environments'.

Ballou, D. P. and Tayi, G. K. (1999) 'DataQuality in DataWarehouse Environments', Communications of the ACM, 42(1), pp. 73–78.

Barlow, I.W., Flynn, N.A. and Britton, J.M. (1994) 'The Basingstoke Orthopaedic Database: a high quality accurate information system for audit' Annals of the Royal College of Surgeons of England, 76(6 suppl) pp.285-287

Barrie, J.L. and Marsh, D.R. (1992) 'Quality of data in the Manchester orthopaedic database', British Medical Journal, 304(6820), pp. 159-162

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009) 'Methodologies for data quality assessment and improvement', ACM Computing Surveys, 41(3), pp. 1–52.

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

Cao, Y., Fan, W., & Yu, W. (2013). Determining the relative accuracy of attributes. Proceedings of the 2013 International Conference on Management of Data - SIGMOD '13, 565.

Chengalur-Smith, I. N., Ballou, D. P., and Pazer, H. L. (1999) 'The impact of data quality information on decision-making: an exploratory analysis', IEEE Transactions on Knowledge and Data Engineering, 11(6), pp. 853–864.

Chiang, F. and Miller, R. J. (2008) 'Discovering data quality rules', Proceedings of the VLDB Endowment, 1(1), pp. 1166–1177.

Chopra, S., Reinhardt, G., Mohan, U., & Kellogg, J. L. (2007). 'The Importance of Decoupling Recurrent and Disruption Risks in a Supply Chain'. Naval Research Logistics, 54, 544–555.

Chu, Y., Yang, S., and Yang, C. (2001) 'Enhancing data quality through attribute-based metadata and cost evaluation in data warehouse environments', Journal of the Chinese Institute of Engineers, 24(4), pp. 497–507.

Congalton, R.G., (1991) 'A review of assessing the accuracy of classifications of remotely sensed data' Remote sensing of environment, 37(1), pp.35-46.

Conroy, M. B., Majchrzak, N. E., Silverman, C. B., Chang, Y., Regan, S., Schneider, L. I., *et al.* (2005). Measuring provider adherence to tobacco treatment guidelines: A comparison of electronic medical record review, patient survey, and provider survey. Nicotine and Tobacco Research, 7(Suppl. 1), S35-S43.

Dremel, C., Stoeckli, E. and Wulf, J., (2020) 'Management of analytics-as-a-service-results from an action design research project' Journal of Business Analytics, 3(1), pp.1-16.

Duvemo, K. and Lämås, T. (2006) 'The influence of forest data quality on planning processes in forestry', Scandinavian Journal of Forest Research, 21(4), pp. 327–339.

Edsall, D.W., Deshane, P., Giles, C., Dick, D., Sloan, B., and Farrow, J.,(1993) 'Computerized patient anesthesia records: less time and better quality than manually produced anesthesioa records' Journal of clinical anesthesia, 5(4), 275-283

Fan, W. (2015) 'Data Quality: From Theory to Practice', ACM SIGMOD Record, 44(3), pp. 7–18.

Fan, W., Jia, X., Li, J., & Ma, S. (2009). Reasoning about record matching rules. Proceedings of the VLDB Endowment, 2(1), 407–418.

Faulconer, E. R., & de Lusignan, S. (2004) 'An eight-step method for assessing diagnostic data quality in practice: Chronic obstructive pulmonary disease as an exemplar'. Informatics in Primary Care, 12, 243-253

Ferson, S. and Ginzburg, L. R. (1996) 'Different methods are needed to propagate ignorance and variability', Reliability Engineering and System Safety, 54(2–3), pp. 133–144.

Fisher, C. W., Chengalur-Smith, I., & Ballou, D. P. (2003). The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. Information Systems Research, 14(2), 170–188.

Forster, M., Bailey, C., Brinkhof, M. W., Graber, C., Boulle, A., Spohr, M., *et al.* (2008). Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. Bulletin of the World Health Organization, 86,

Fox, C. (1994) 'The notion of data and its quality dimensions', Information Processing and Management, 30(I), pp. 9–19.

Goodyear-Smith, F., Grant, C., York, D., Kenealy, T., Copp, J., Petousis-Harris, H., *et al.* (2007). Determining immunisation coverage rates in primary health care practices: a simple goal but a complex task. International Journal of Medical Informatics, 77, 477-485.

Gouveia-Oliveira, A., Raposco, V.D., Salgado, N.C., Almeida, I., Nobre-Leitao, C., and Galvao de Melo, F., (1991) 'Longitudinal comparative study on the indluence of computers on reporting of clinical data' Endoscopy, 23(06), 334-337

Hartig, O., and Zhao, J. (2009), 'Using web daata provenance for quality assessment' CEUR Workshop ProceedingsHeinrich, B., & Hristova, D. (2016). A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. Journal of Decision Systems, 25(1), 16–41.

Heinrich, B., Hristova, D., Klier, M., Schiller, A. and Szubartowicz, M., (2018) 'Requirements for data quality metrics'. Journal of Data and Information Quality, 9(2), 1-32. Heinrich, B., & Klier, M. (2015) 'Metric-based data quality assessment - Developing and evaluating a probability-based currency metric'. Decision Support Systems, 72, 82-96.

Hoare, C. A. R. (1975). Data reliability. acm sigplan Notices, 10(6), 528-533.

Hohnloser, J.H., Fischer, M.R., König, A., and Emmerich, B., (1994) 'Data quality in computerized patient records' International Journal of clinical monitoring and computing, 11(4), 233

Holden, J. M., Bhagwat, S. A., Haytowitz, D. B., Gebhardt, S. E., Dwyer, J. T., Peterson, J., Beecher, G. R., Eldridge, A. L., and Balentine, D. (2005) 'Development of a database of critically evaluated flavonoids data: Application of USDA's data quality evaluation system', Journal of Food Composition and Analysis, 18(8), pp. 829–844.

Holmström, H., Kallur, H., and Ståhl, G. (2003) 'Cost-plus-loss analyses of forest inventory strategies based on kNN-assigned reference sample plot data', Silva Fennica, 37(3), pp. 381–398.

Hu, B. and Feng, Y. (2017) 'Optimization and coordination of supply chain with revenue sharing contracts and service requirement under supply and demand uncertainty', International Journal of Production Economics. Elsevier, 183(October 2016), pp. 185–193.

Huh, Y. U., Keller, F. R., Redman, T. C., and Watkins, A. R. (1990) 'Data Quality', Information and Software Technology, 32(8), pp. 559–565.

Johnson, N., Mant.,D., Jones, L., and Randall, T. (1991) 'Use of computerised general practice data for population surveillance: comparative surdy of influenza data' Bristich Medical Journal, 302(6779), 763-765

Jones, R.B. and Hedley, A.J. (1986) 'A computer in the diabetic clinic. Completeness of data in a clinical information system for diabetes' Practical Diabetes International, 3(6), 295-296

Kangas, A. S. and Kangas, J. (2004) 'Probability, possibility and evidence: approaches to consider risk and uncertainty in forestry decision analysis', Forest Policy and Economics, 6 (2), pp. 169-188

Kim, S., Pasupathy, R., & Henderson, S. G. (2011) 'A Guide to Sample-Average Approximation'. In Handbook of Simulation Optimization. 207-243

Klibi, W., Martel, A., and Guitouni, A. (2010) 'The design of robust value-creating supply chain networks: A critical review', European Journal of Operational Research. Elsevier B.V., 203(2), pp. 283–293.

Kuhn, K., Swobodnik, W., Johannes, R.S., Zemmler, T., Stange, E.F., Ditschuneit, H. and Classen, M. (1991) 'The quality of gastroenterorological reports based in free text dictation: an evaluation in endoscopy and ultrasonography' Endoscopy, 23(05), 262-264

Lakshminarayan, K., Harp, S.A. and Samad, T., (1999). 'Imputation of missing data in industrial databases'. Applied intelligence, 11(3), pp.259-275.

Laudon, K. C. (1986). Data quality and due process in large interorganizational record systems. Communications of the ACM, 29(1), 4–11.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. 'AIMQ: a Methodology for Information Quality Assessment'. in Proceedings of the 7th International Conference on Information Quality, C. Fisher and B. Davidson (eds.), Cambridge, MA: Elsevier BV, pp. 133-146.

Linder, J. A., Kaleba, E. O., & Kmetik, K. S. (2009). Using electronic health records to measure physician performance for acute conditions in primary care: Empirical evaluation of the community-acquired pneumonia clinical quality measure set. Medical Care, 47, 208-216.

Maresh, M., Dawson, A.M. and Beard, R.W. (1986) 'Assessment of an on-line computerized perinatal data collection and information system' BJOG: An International Journal of Obstetrics & Gynaecology, 93(12) 1239-1245

Merino, J., Caballero, I., Rivas, B., Serrano, M., and Piattini, M. (2016) 'A Data Quality in Use model for Big Data', Future Generation Computer Systems. Elsevier B.V., 63, pp. 123–130.

Mezzanzanica, M., Boselllli, R., Cesarini, M., and Mercorio, F. (2015) 'A model-based evaluation of data quality activities in KDD' Information Processing & Management, 51(2), 144-166

Moges, H. T., Dejaeger, K., Lemahieu, W., and Baesens, B. (2013) 'A multidimensional analysis of data quality for credit risk management: New insights and challenges', Information and Management. Elsevier B.V., 50(1), pp. 43–58.

Moody, D. (2003) 'Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice.', Ecis, (2003).

Morey, R.C., (1982) 'Estimating and improving the quality of information in a MIS', Communications of the ACM, 25(5), pp.337-342.

Mula, J., Poler, R., García-Sabater, G. S., and Lario, F. C. (2006) 'Models for production planning under uncertainty: A review', International Journal of Production Economics, 103(1), pp. 271–285.

Parssian, A. (2006) 'Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions'.

Parssian, A., Sarkar, S., and Jacob, V. S. (2004) 'Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product', MANAGEMENT SCIENCE, 50(7), pp. 967–982.

Pasalodos-Tato, M., Mäkinen, A., Garcia-Gonzalo, J., Borges, J. G., Lämås, T., and Eriksson, L. O. (2013) 'Review. Assessing uncertainty and risk in forest planning and decision support systems: review of classical methods and introduction of new approaches', Forest Systems, 22(2), p. 282.

Persell, S. D., Dunne, A. P., Lloyd-Jones, D. M., & Baker, D. W. (2009). Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. Medical Care, 47, 418-424.

Pietilä, I., Kangas, A., Mäkinen, A., and Mehtätalo, L. (2010) 'Influence of growth prediction errors on the expected losses from forest decisions', Silva Fennica, 44(5), pp. 829–843.

Plotkin, D., (2020) 'Data stewardship: An actionable guide to effective data management and data governance', Academic press

Powell, J., Fitton, R., & Fitton, C. (2006). Sharing electronic health records: The patient view. Informatics in Primary Care, 14, 55-57.

Pringle, M., Ward, P. and Chilvers, C. (1995) 'Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer' British Journal of General Practice, 45, pp. 537-541

Redman, T. C. (1998) 'The impact of poor data quality on the typical enterprise', Communications of the ACM, 41(2), pp. 79–82.

Ricketts, D., Newey, M., Patterson, M., Hitchin, D. and Fowler, S. (1993) 'Markers of data quality in computer audit: the Manchester Orthopasedic Databse' Annals of the Royal College of Surgeons of England, 75(6) p.393

Santoso, T., Ahmed, S., Goetschalckx, M., & Shapiro, A. (2004) 'Production, Manufacturing and Logistics A stochastic programming approach for supply chain network design under uncertainty'. European Journal of Operational Research, 167(1). 96-115.

Saigh, O., Triola, M. M., & Link, R. N. (2006). Brief report: Failure of an electronic medical record tool to improve pain assessment documentation. Journal of General Internal Medi- cine, 21, 185-188

Sein, M.K., Henfridsson, O., Purao, S., Rossi, M. and Lindgren, R., 2011. 'Action design research'. *MIS quarterly*, pp.37-56.

Sheng, Y. P., & Mykytyn, P. P. (2002). Information Technology Investment and Firm Performance: A Perspective of Data Quality. Seventh International Conference on Information Quality, 132–141.

Sigurdardottir, L. G., Jonasson, J. G., Stefansdottir, S., Jonsdottir, A., Olafsdottir, G. H., Olafsdottir, E. J., & Tryggvadottir, L. (2012). Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness. Acta oncologica, 51(7), 880-889.

Simard, V., Rönnqvist, M., Lebel, L., & Lehoux, N. (2019). A General Framework for Data Uncertainty and Quality Classification. IFAC-PapersOnLine, 52(13), 277-282.

Snyder, L. V. and Shen, Z.-J. M. (2006) 'Supply and Demand Uncertainty in Multi-Echelon Supply Chains'.

Sategemann, D., Volk, M., Jamous, N. and Turowski, K. (2019) 'Understanding issues in big data applications - A multidimentional endeavor'. AMCIS 2019 Proceedings, 19

Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997) '10 Potholes in the road to information quality', Computer, 30(8), pp. 38–46.

Vaziri, R., Mohsenzadeh, M., & Habibi, J. (2019). 'Measuring data quality with weighted metrics'. Total Quality Management & Business Excellence, 30(5–6), 708–720.

Viscusi, G., Spahiu, B., Maurino, A. and Batini, C., 2014. 'Compliance with open government data policies: An empirical assessment of Italian local public administrations'. Information polity, 19(3, 4), pp.263-275.

Wagner, M.M. and Hogan, W.R. (1996) 'The accuracy of medication data in an outpatient electronic medical record' Journal of the American Medical Informatics Association, 3(3), pp. 234-244

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., and Krayer von Krauss, M. P. (2003) 'Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support', Integrated Assessment, 4(1), pp. 5–17.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers', Journal of Management Information Systems, 12(4), pp. 5–33.

Wang, Y., Storey, V. C., and Firth, P. (1995) 'A framework for analysis of data quality research', IEEE Transactions on Knowledge and Data Engineering, 7(4), pp. 623–640.

Wechsler, Alisa, and Adir Even. (2012) 'Using a Markov-Chain model for assessing accuracy degradation and developing data maintenance policies'. AMCIS 2012 Proceedings. 3

Wilton, R. and Pennisi, A.J. (1994) 'Evaluating the accuracy of transcribed computer-stored immunization data' Pediatrics, 94, pp.902-6

Woodall, P., Borek, A., and Parlikad, A. K. (2013) 'Data quality assessment: The Hybrid Approach', Information and Management. Elsevier B.V., 50(7), pp. 369–382.

Woodall, P., Giannikas, V., Lu, W. and McFarlane, D. (2019) 'Potential Problem Data Tagging: Augmentation information systems with the capability to deal with innacuracies'. Decision Support Systems, 121, 72-83

Yarnall, K.S., Michener, J.L., Broadhead, W.E., Hammond, W.E and Tse, C.K.J. (1995) 'Computer-prompted diagnostic codes' Journal of Family Practice, 40(3), pp.257-262

Yousefpour, R., Bredahl Jacobsen, J., Jellesmark Thorsen, B., Meilby, H., Hanewinkel, M., and Oehler, K. (2012) 'A review of decision-making approaches to handle uncertainty and risk in adaptive forest management under climate change', Annals of forest science, 69, p. 15.

Zak, Y., and Even, A. 2017. 'Development and Evaluation of a Continuous-Time Markov Chain Model for Detecting and Handling Data Currency Declines', Decision Support Systems (103), pp. 82-93.

Zanjani, M. K., Ait-Kadi, D., & Nourelfath, M. 2013. 'A stochastic programming approach for sawmill production planning'. International Journal of Mathematics in Operational Research, 5(1), 1–18.

Zhang, Y., Feng, Y., and Rong, G. (2016) 'Data-Driven Chance Constrained and Robust Optimization under Matrix Uncertainty', Industrial and Engineering Chemistry Research, 55(21), pp. 6145–6160.

Zhu, H., Madnick, S., Lee, Y., and Wang, R. Y. (2012) 'Data and Information Quality Research: Its Evolution and Future', Computing Handbook Set, pp. 1–22.

Zyngier, D. (2017) 'An uncertainty management framework for industrial applications', Optimization and Engineering. Springer US, 18(1), pp. 179–202.

Annex 1: List of data quality assessments and their level of quality

Source	Accuracy	Quality level		
		Proposed Classification	Blake and Mangiameli (2011),	Holden *et al.* (2005)
Baker *et al.* (2007)	92.3%	Good	High quality	Exceptional
Barlow *et al.* (1994)	98%	Excellent	High quality	Exceptional
Barrie and Marsh (1992)	92.9% (worst)	Good	High quality	Exceptional
	97.8% (best)	Excellent	High quality	Exceptional
Congalton (1991)	74%	Sufficient	---	Above average
	78%-89% (improved)	Good	Medium quality	Exceptional
Faulconer and de Lusignan (2004)	75%	Good	---	Exceptional
Goodyear-Smith *et al.* (2007)	89% (worst)	Good	Medium quality	Exceptional
	97% (best)	Excellent	High quality	Exceptional
Hohnloser et al (1994)	74.1%	Sufficient	---	Above average
Lauren (1986)	93.7%	Good	High quality	Exceptional
Maresh et al (1986)	95%	Excellent	High quality	Exceptional
Morey (1982)	75%	Good	---	Exceptional
	84.7% (improved)	Good	Low quality	Exceptional
Persell *et al.* (2009)	72%	Sufficient	---	Above average
Powell *et al.* (2006)	68%	Sufficient	---	Above average
Pringle *et al.* (1995)	47.4%(worst)	Sufficient	---	Average
	96.7% (best)	Excellent	High quality	Exceptional
Ricketts *et al.* (1993)	44% (worst)	Sufficient	---	Above average
	67% (best)	Sufficient	---	Average
Saigh et al (2006)	45%	Sufficient	---	Average
Sigurdardottir *et al.* (2012)	96.4%	Excellent	High quality	Exceptional
Viscusi *et al.* (2014)	44% (overall)	Sufficient	---	Average
	78.4% (best)	Good	---	Exceptional
	5.4% (worst)	Insufficient	---	Below average
Wagner and Hogan (1996)	83%	Good	Low quality	Exceptional
Wilton and Pennisi (1994)	89.8%	Good	Medium quality	Exceptional
Yarnall *et al.* (1995)	62%	Sufficient	---	Above average
	82% (improved)	Good	Low quality	Exceptional

Source	Completeness	Quality level		
		Proposed Classification	Blake and Mangiameli (2011),	Holden *et al.* (2005)
Barlow *et al.* (1994)	92.5%	Good	High quality	Exceptional
Barrie and Marsh (1992)	45.9% (worst)	Sufficient	---	Average
	82% (best)	Good	Low quality	Exceptional
Conroy *et al.* (2005)	38% (worst)	Sufficient	---	Average
	98% (best)	Excellent	High quality	Exceptional
Edsall *et al.* (1993)	87%	Good	Medium quality	Exceptional
Faulconer and de Lusignan (2004)	90%	Good	Medium quality	Exceptional
Forster *et al.* (2008)	89.1%	Good	Medium quality	Exceptional
	94.7% (improved)	Good	High quality	Exceptional
Goodyear-Smith *et al.* (2007)	20% (worst)	Insufficient	---	Below average
	70% (best)	Sufficient	---	Above average
Gouveia-Oliveira et al (1991)	81.6% (average)	Good	Low quality	Exceptional
Hohnloser et al (1994)	54.5%	Sufficient	---	Above average
Johnson *et al.* (1991)	28.2%	Sufficient	---	Average
Jones and Hedley (1986)	90.3% (worst)	Good	Medium quality	Exceptional
	100% (best)	Excellent	High quality	Exceptional
Kuhn et al (1991)	90.7%	Good	Medium quality	Exceptional
Lakshminarayan *et al.* (1999)	50%	Sufficient	---	Above average
	90.9% (improved)	Good	Medium quality	Exceptional
Linder *et al.* (2009)	36% (worst)	Sufficient	---	Average
	93% (best)	Good	High quality	Exceptional
Persell *et al.* (2009)	70.1% (worst)	Sufficient	---	Above average
	98.1% (best)	Excellent	High quality	Exceptional
Pringle *et al.* (1995)	27.2%(worst)	Sufficient	---	Average
	100% (best)	Excellent	High quality	Exceptional
Ricketts *et al.* (1993)	17% (worst)	Insufficient	---	Below average
	53% (best)	Sufficient	---	Above average
Sigurdardottir *et al.* (2012)	99.15%	Excellent	High quality	Exceptional
Viscussi *et al.* (2015)	5% (worst)	Insufficient	---	Below average
	91.6% (best)	Good	Medium quality	Exceptional
Wagner and Hogan (1996)	93%	Good	High quality	Exceptional
Wilton and Pennisi (1994)	88.4%	Good	Medium quality	Exceptional
Yarnall *et al.* (1995)	79%	Good	---	Exceptional
	84% (improved)	Good	Low quality	Exceptional

Source	Consistency	Quality level		
		Proposed Classification	Blake and Mangiameli (2011),	Holden *et al.* (2005)
Ronveaux *et al.* (2005)	31% (worst)	Sufficient	---	Average
	53% (overall)	Good	---	Above average
	73% (best)	Good	---	Above average
Mezzanzanica *et al.* (2015)	33%	Sufficient	---	Average
	89.2% (improved)	Excellent	Medium quality	Exceptional
Source	Timeliness	Quality level		
		Proposed Classification	Blake and Mangiameli (2011),	Holden *et al.* (2005)
Hartig and Zhao (2009)	75%	Good	Medium quality	Exceptional
Sigurdardottir *et al.* (2012)	84.8% (worst)	Good	High quality	Exceptional
	96.9% (best)	Excellent	High quality	Exceptional
Viscusi *et al.* (2014)	32.7%	Sufficient	Low quality	Average