Vliegen, Lea; Moroff, Nikolas Ulrich; Riehl, Katharina

Conference Paper

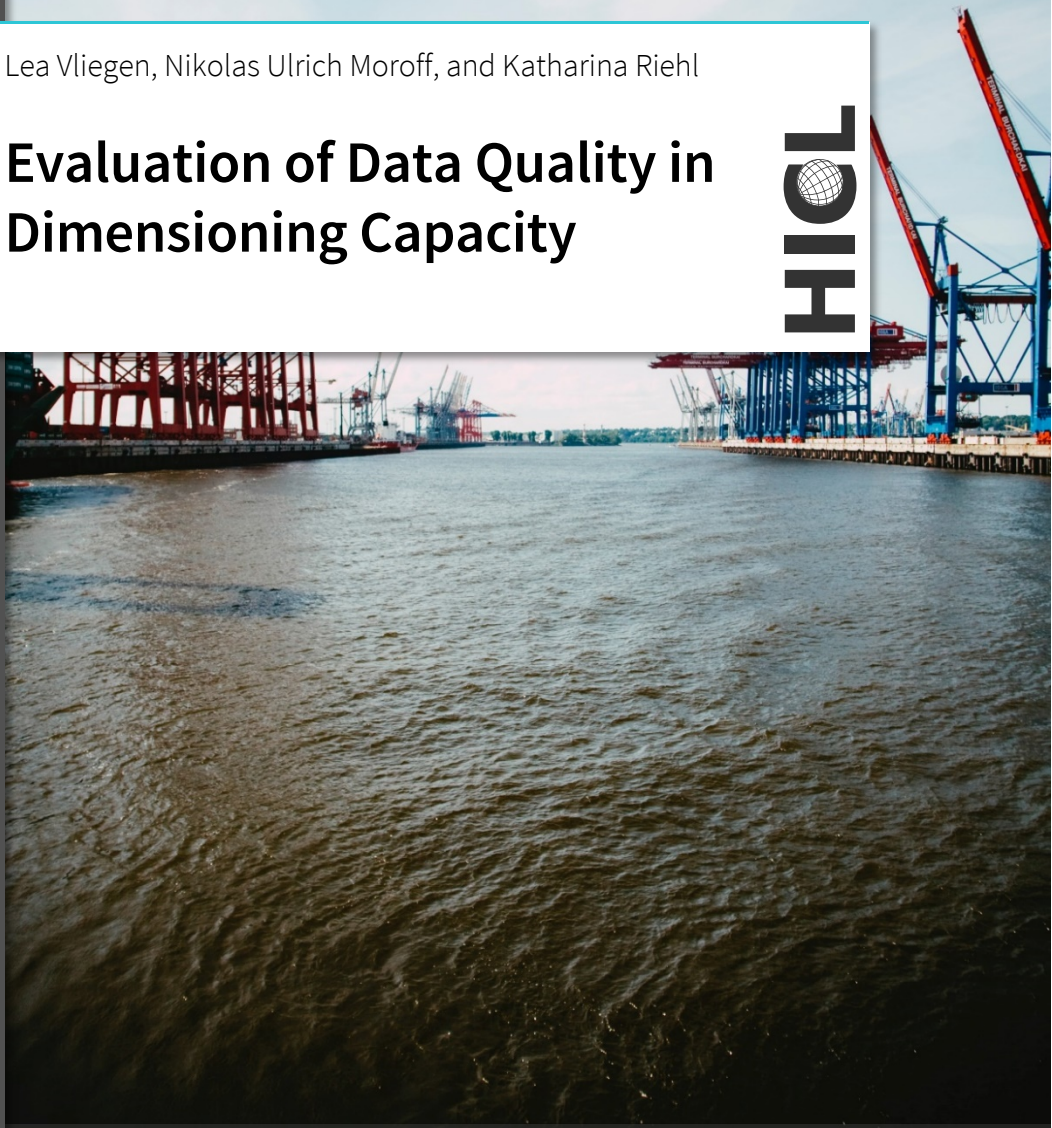
Evaluation of data quality in dimensioning capacity

Proceedings of the Hamburg International Conference of Logistics (HICL) – 29

Lea Vliegen, Nikolas Ulrich Moroff, and Katharina Riehl

Evaluation of Data Quality in Dimensioning Capacity

HICL



Published in: Data science and innovation in supply chain management
Wolfgang Kersten, Thorsten Blecker and Christian M. Ringle (Eds.)
ISBN: 978-3-753123-46-2 , September 2020, epubli

Evaluation of Data Quality in Dimensioning Capacity

Lea Vliegen [1], Nikolas Ulrich Moroff [1], and Katharina Riehl [1]

1 – Fraunhofer Institute for Material Flow and Logistics IML

Purpose: This paper aims to give an overview of the current state of research on measuring data quality. The identified methods will be applied to the task of dimensioning capacities (e.g. warehouse capacities) in the field of supply chain design (SCD) to further increase trust in decision support and to make full use of the potential of analytics.

Methodology: The data requirements for SCD decisions are identified through the combination of findings of a research project and additional literature research. Moreover, an overview on measuring data quality will be given according to a literature study. Based on the required data, the applicability of methods to measure data quality will be analyzed and an application concept developed.

Findings: The quality of decisions can only be as good as the quality of the data they are based on. The article provides an overview of methods for evaluating datasets and develops an approach for measuring and evaluating data quality for the specific case of capacities in the SCD process.

Originality: The adaption of approaches of measuring data quality to the problem of dimensioning capacities in SCD ensures an adequate evaluation of whether the data fulfills the required quality for the planning tasks.

First received: 13. Mar 2020 Revised: 16. Jun 2020 Accepted: 12. Aug 2020

1 Motivation

Supply chains of companies have changed significantly in the last decades due to the advancing globalization. Company networks become more and more complex in order to serve the growing and changing market requirements. This makes the planning of supply networks, capacities, and inventories increasingly complex.

The services and products offered by the companies have become largely interchangeable, therefore there is an increased focus on flexible customer service, speed and adherence to delivery dates at the lowest possible prices (Wassermann, 2013). This development can be favored by shortened product life cycles, fluctuating customer behavior and increasingly complex data structures in the supply chain (SC). As a result, the entire logistics SC, production capacities and shipping processes must react immediately to market fluctuations, when these cannot be planned in advance using forecasting methods (Erben and Romeike, 2003).

To improve the quality of planning despite challenging environmental influences, methods from the field of data analytics are increasingly used. Especially the areas of forecast demand, production, promotion, pricing and delivery can be optimized with the help of new methods to thus meet the growing requirements of the market (Dash, et al., 2019).

Nevertheless, the basis for the use of data driven methods is a valid database of adequate quality. For this reason, a strong focus is placed on the preprocessing of the data base before the modeling of the data driven approach can be started (Gudivada, Apon and Ding, 2017).

To keep the data preparation effort as low as possible, the data quality has to be measured in advance and assessed for the specific case of application. Due to the increased complexity in SCs, data is gathered at various points in the SC in more detail. The availability of large amounts of data offers potential for the planning process of SCs as well as in operation and for optimizations. In a study from Statista on big data analytics and its SC outcomes for companies it was indicated that 41% of the considered companies had faster and more efficient reaction times and 36% had an improvement of efficiency in their SC exceeding 10% (Statista, 2014). The availability of data is an opportunity and a challenge at the same time for SCs. New approaches and methods have to be adapted and developed to make use of their potential and make data-backed decisions (Waller and Fawcett, 2013). This potential is most promising on a strategic level when the SCs are designed, since the basic structure is set up with its strategic partners, locations, and capacities. The dimensioning of capacities of areas like a warehouse or in production are crucial for the operation of a SCs. If these decisions are based on data with a poor data quality, adjustments demand enormous efforts.

In the research project E^2-Design the focus is to design a toolbox for companies enabling them to include energy efficiency as an additional parameter in the strategic and tactical planning of SC networks. Thereby, energy efficiency extends the currently mainly used target parameters of the magic triangle: Time, costs, and quality/performance. One research question being addressed is dimensioning warehouse and production capacities under ecological aspects. Within the project it became clear that the optimization

depended to a high degree on the data quality, thus a concept was developed to determine data quality for the specific application of capacity dimensioning.

This paper presenting the developed concept is structured into four sections. First, the basics of capacity dimensioning and data quality are introduced. This is followed by the results of the literature search on the topic of methods for measuring data quality. In order to select a method, the specific requirements of dimensioning capacities were evaluated in this paper using a pair comparison and assigned to individual quality dimensions. Based on the resulting requirement profile a new concept was developed to determine the data quality in the best possible way, by connecting existing methods to fulfill the specific requirements of the use case. In the following chapter the SCD task model is described and the use case will be illustrated with a focus on dimensioning capacities to better understand the challenges of the research project.

2 Description of Use Case - Capacity dimensioning

Due to globalization supply networks become more widespread leading to longer lead times. This increases the importance of an efficient SC, making it a decisive competitive factor and therefore, more emphasis is placed on the design on the SC. A SC is characterized as a network of suppliers, production, warehouses, and distribution that transforms an input such as raw materials into finished goods, which are delivered to the customer supplies (Santoso, et al., 2005; Ketchen and Hult, 2007). In the SCD process the basis and long-time structure of the SC are planned and determined. The design process can be structured into planning levels and different tasks (Baghalian, Rezapour and Farahani, 2013; Fattahi, et al., 2015). Based on a literature study by Parlings, Cirullies and Klingebiel (2013) vital tasks for the SCD process were identified, classified, and structured into a reference model. The model is structured hierarchically into three levels: superordinate SCD tasks, SC structure design and SC process design (see Figure 1).

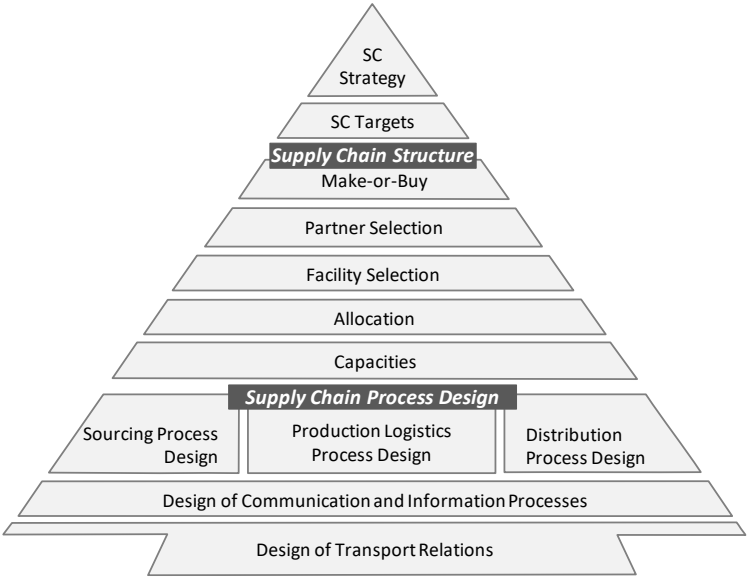


Figure 1: SCD task model (Parlings, Cirullies and Klingebiel, 2013)

In the superordinate tasks the most extensive choices for the SC are made (deciding on the SC strategy and targets). They must be aligned with the overall company strategies and goals. In the SC structure design the decisions for make-or-buy must be done as well as the selection of strategic partners and facilities. Additionally, the allocation of products to locations for production and warehousing and dimensioning of their capacities is a crucial parameter for efficient processes. Especially in manufacturing the

capacities are a key driver for capital costs. Higher capacities allow economies of scale, but when already produced quantities cannot be sold due to a lack in demand, utilization is low, and costs increase (Hsu and Li, 2009).
In the SC process design the strategic decisions for the sourcing, production and distribution are synchronized with the communication process and transport relations. Within the three planning levels there is no hierarchy of tasks, as they are highly correlated. For the network to function holistically, integrated choices must be made on all levels (Parlings, Cirullies and Klingebiel, 2013). A holistic approach enables fast reactions when adjustments of goals and strategies are necessary to comply with political or legislative changes. With alignments such as designing a SC more energy efficient, but still cost effective, new models and planning tools are being developed (Schreiber, 2019). Simulation is a useful tool to allow SC planners at strategic level to try out different priorities and see the impact before implementation. However, in distributing capacities for e.g. warehouses the dependencies must be clarified. One of the main challenges is to find the appropriate level of abstraction for the use case so that data from the operational level can be used effectively on the strategic level. This occurs especially with dimensioning capacities. The use case is from a company trading raw and processed materials and delivering the service to bring them customized to their client. The materials provided vary greatly in shape, dimension and weight. In all three characteristics restrictions may apply leading to a different need of warehousing and later different processing steps. Due to the variety of products there are around 200 product subgroups, which have different volume parameters. This increases the challenge of selecting ideal warehouse systems.

The dimensions, shape and product subgroup are the basis for the planning process and input for the dimensioning. Therefore, as they are part of the article master data, they have to be correct. Otherwise wrong areas of storage types are defined, and the allocated products cannot be distributed accordingly in the warehouse. Not to mention the fact that the necessary equipment for processing might not be available at the dedicated location. The whole network is planned with locations all over Germany with different warehouse systems including the capacities, transport between locations and also specialized locations. The dimensioning of warehouse and production capacities for each cluster is crucial. To further understand the challenges of the SCD task of dimensioning capacities the process is outlined in the next section and the importance of data quality is further detailed in this use case.

2.1 SCD Task: Dimensioning Capacities

In the group of tasks defining the SC structure the location of production sites and the allocation of raw materials and products to these locations are decided together with dimensioning capacities in warehousing and production. These strategic decisions influence the SC long-term and adjustments are likely to be cost intensive. Network design is often only considered as a definition of the locations however the allocation of the variety of products to the sites and decisions on capacities and technology at each site are more complex (Fleischmann and Koberstein, 2015).

Capacity is defined as the maximum performance of a system. In the case of warehouse and production capacities it is the number of products and

components stored or produced in one time period (Minner, 2018). Production and storage capacity planning are closely linked due to their interaction (Friemann, 2015). Capacity planning is typically on a long to medium term basis and is part of the corporate infrastructure planning. The decision between a few large and several small capacity adjustments is significantly influenced by economies of scale of dimensioning costs on the one hand and idle costs of unused capacity on the other hand. The strategic definition of a proactive (lead) strategy must be distinguished from a reactive (lag) strategy in the case of changing demands (Slack and Lewis, 2017). The planning of production capacities in cross-company SCs is particularly difficult if legally independent players cooperate with each other only temporarily (Werner, 2017). A lack of information exchange and communication within the SC makes capacity dimensioning for production and warehousing difficult (Baumgärtel, 2008). For example, even slight fluctuations in demand at upstream stages of the value chain can lead to large increases in demand. A small change triggers an ever-increasing change in final requirements in a downward direction, so that an inventory build-up occurs within the SC (Werner, 2017). A high number of different factors influence the level of capacity (see Figure 2).

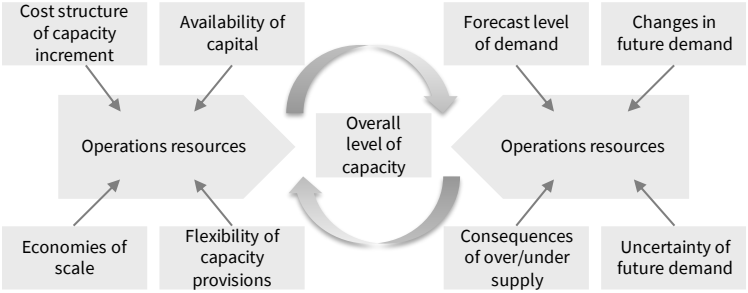


Figure 2: Influencing factors for overall level of capacity (Slack and Lewis, 2017)

Due to the high uncertainty in the long-term data (e.g. in the demand of future products in specific markets, investment volumes, labor costs or exchange rates) flexibility and robustness of the SC have to be considered to reduce risks (Fleischmann and Koberstein, 2015). These factors are linked closely to variables on the tactical and sometimes operational planning levels. This poses the challenge of selecting appropriate levels of abstraction (Friemann, 2015).

On a strategic level one main input for dimensioning capacities is the demand forecast based on potential markets to be served in the future often on an aggregated, annual basis (Friemann, 2015). On this basis, the capacity configuration is carried out along with the decisions on the total capacity required and it is distribution (Slack and Lewis, 2017).

In the research project several challenges occurred in practical experience concerning the data. One obstacle is that process knowledge is in people's minds in different locations and not digitally available and editable. To

gather the required data, templates have to be developed so that all locations provide the data in the same structured way. Then the applicability of the template has to be approved by one business division, before it can be distributed to other divisions and locations. This process is time- and labor-consuming, especially if questions occur.

Another challenge is the wide portfolio of different products with diverse requirements. Additionally, planned products for the future should be considered. This means that either more flexible warehousing solutions have to be found or different systems have to be designed to accommodate all needs. Furthermore, for each product the master data must be filled in correctly in a quantified and understandable way. This includes a clear identifier per product and the dimensions as well as all applicable restrictions with units. Preferably only relevant data for dimensioning capacities is included in the dataset.

Before starting the planning process the dataset has to be complete with an adequate quality for dimensioning capacities. To avoid the repetition of planning due to lacking data quality during the planning process, it should be checked beforehand whether the data quality requirements are fulfilled. Therefore, a systematic approach is needed for the use case of capacity dimensioning. To determine the required level of data quality, the theoretical background of data quality will be outlined in the next chapter. Additionally, an overview of existing methods for measuring data quality will be given.

3 Foundations in Data Quality

The aim of this chapter is to define data quality and to present and compare suitable methods for measuring data quality in the context of SCM. The high data density in the SCM area leads to a high potential in the areas of operational efficiency, customer experience and new product development. This means that a high level of data quality and the measurement of data quality provides a decisive competitive advantage in various fields of activity (Addo-Tenkorang and Helo, 2016).
In order to measure the quality of a dataset, the term data quality must be defined and delimited in order to create a common basis. For this reason, the following section presents established definitions of data quality and introduces a list of existing assessment procedures.

3.1 Definition of Data Quality

An essential prerequisite for the use of innovative methods is high-performance data management since data is understood as the basic framework of digital development. Only through further processing and preparation does the data become information, which can be integrated into planning processes (Oppenheim, Stenson and Wilson, 2003).
An effective data management can be characterized by three essential aspects:

1. Control of data volumes
2. Decentralized data processing
3. Definition of data standards

In a survey, data managers from various industries were asked about the greatest challenges in the field of data management. The results show that data quality is regarded as one of the greatest challenges (Österle and Otto, 2014). In order to make an appropriate assessment of data quality, the particular application must be taken into account (Jayawardene, Sadiq and Indulska, 2015). Basically, two concepts can be distinguished in the characterization of data quality: Information technology focus and user-related focus.

The approach of Information Technology Assessment of data quality focuses on the assessment of the data definition, the quality of the dataset content and the data presentation. These three modules form the basic framework for the definition of information technology data quality and were further detailed by English (1998). The detailing of the three quality modules are displayed in Figure 3. The first module focuses on the framework conditions of the data collection. Only data that has been sufficiently specified can be used to measure quality. The second module concentrates on the correctness of the content in terms of unambiguity and completeness. The last module deals with the availability of data. Parameters for this part are e.g. the time of availability and compliance with the format.

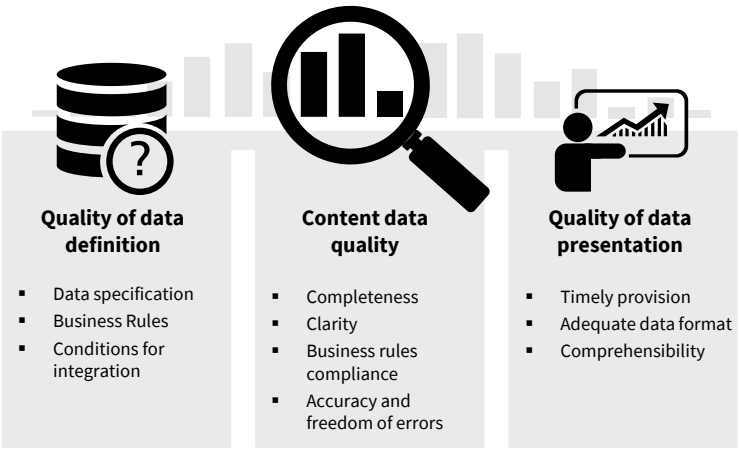


Figure 3: Information Technology Assessment of data quality (English, 1998)

In contrast to the information technology focus, data can also be evaluated on a user-related basis. Here, the focus is on the properties of the dataset and surrounding data models (e.g. definitions and frameworks) are not further considered. Based on the work of Wang and Strong (1996), Sidi, et al. (2012) defined four main components for the evaluation of user-related data quality with the help of an extensive literature research: Timeliness, Accuracy, Completeness and Consistency. These main components have been further detailed in numerous models, resulting in many subcategories.

Especially well known is the model by Rohweder, et al. (2018) which is divided into four quality categories based on 15 dimensions. The key difference between their model to Wang and Strong (1996) is that they do not

consider security as a central quality dimension. Instead, they require security as a necessary basis for measuring data quality. In addition to the base model of Wang and Strong (1996), they introduce the usability (ease of manipulation) dimension. These fifteen quality dimensions can be assigned to four criteria: system-supported, inherent, presentation-related, and purpose-dependent. The following Figure 4 presents the model after Rohweder et al. (2018) (based on Wang and Strong (1996)) in detail with the four quality criteria and their focus for determining data quality.

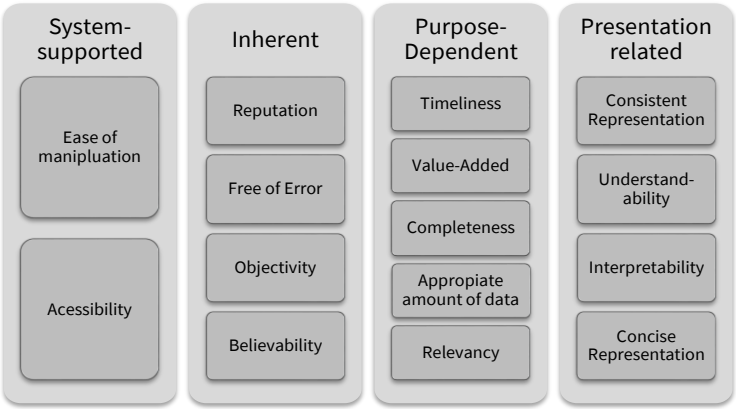


Figure 4: Data quality dimension connected to their quality criteria (Rohweder, et al., 2018)

The previous section provided an overview of possible dimensions of data quality. It becomes clear that due to different perspectives it is not possible to give a general definition of data quality independent of the specific use case. For this reason, the following section presents existing methods for measuring data quality and examines their applicability to the specific use case of capacity dimensioning.

3.2 Methods for Measuring Data Quality

The literature offers a great variety of methods to measure and evaluate data quality. Since the focus of this paper is the application of data quality to the problem of dimensioning capacities in SCD, this paper does not give a complete overview about all existing methods for measuring data quality. Our research is based on the findings of Batini, et al. (2009), who compared many methods for measuring data quality and developed their own. In this paper, Batini, et al. (2009)'s overview is extended with more methods and metrics for measuring data quality. In our research we focused on the quality dimensions that were considered in each method and examined to what extent metrics were used or developed to determine quality. Based on the results, it can be said that there are very general methods for determining data quality that can be adapted to a wide range of applications. Many of them do not contain any metrics and consequently are always a subjective classification. Those methods often aim to improve data quality, rather than exactly measuring the quality. On the other hand, there are procedures that objectively evaluate a single quality dimension in great detail using metrics, but do not consider the context of the use case.

Table shows selected results from the literature review which are connected to the presented use case capacity dimensioning: Name of the methodology and reference, abbreviation and main characteristics. In addition to the main characteristics the included quality dimensions and metrics are important criteria, illustrated in Table 2.

Table 1: Selected methods and main characteristics

Methodology & Reference	Abbre-via-tion	Main characteristics
Total Data Quality Management Wang (1998)	TDQM	- Systematic application of Total Quality Management with for phases: Definition, Measurement, Analysis, Improvement - Continuous improvement of data quality in operational processes within information systems
Data Warehouse Quality Jeusfeld, Quix and Jarkeet (1998)	DWQ	- Measurement of quality objectives and design options in data warehousing - Perspectives: Conceptual, Logical and Physical - Classification of quality goals according to different stakeholder groups - Quality meta model provides notation for formulating quality goals, queries, and measurements
Total Information Quality Management English (1998)	TIQM	- Processes and techniques for evaluating, optimizing, and controlling the quality of data and information through continuous quality management - Phases: Assessment, Improvement; Improvement Management and Monitoring

Methodology & Reference	Abbreviation	Main characteristics
A methodology for information quality assessment Lee, et al. (2002)	AIMQ	- Information quality measurement based on subjective assessment of quality (carried out by: Surveys and benchmarks) - Components: Product-Service-Performance-Model, quality of data products, Benchmark-Gap-Analysis/Role-Gap-Analysis
Data Quality Assessment Pipino, Lee and Wang (2002)	DQA	- Developing general definition of data quality metrics (subjective and objective) - Comparing the results of the assessments, identifying discrepancies and taking necessary actions for improvement
Comprehensive methodology for Data Quality management Batini and Scannapieco (2006)	CDQ	- Combination of data- and process-driven strategies for data and information quality optimization - Selection of optimal quality improvement process that maximizes benefits for set budget - Phases: State reconstruction, Assessment, Choice of the optimal improvement process

Methodology & Reference	Abbreviation	Main characteristics
Control Charts Jones-Farmer, Ezell and Hazen (2014)	CC	- Control charts for monitoring data quality in aircraft maintenance - Multiple measures of the intrinsic dimensions of data quality
Data Quality Management in Data Warehouse Systems Hinrichs (2002)	DQDWS	- Metrics for selected data quality dimensions to evaluate quality of data stock - Procedure for quantification of data quality aims for objectifiable, target-oriented evaluation - Enables largely automated measurement
Metrics and measurement methods for Data Quality Rohweder, et al. (2018)	MMDQ	- Metrics for dimensions: Completeness, Accuracy, Consistency, and Timeliness - Focus on the requirement of cardinality of metrics
Metrics for Data Quality Assessment	MDQA	- Metrics for dimensions: Completeness, Accuracy, Consistency, and Timeliness

Methodology & Reference	Abbre-via-tion	Main characteristics
Blake and Mangiameli (2011)		
Measuring Data Believability Prat and Madnick (2008)	MDB	- Metric for believability measured by trustworthiness, reasonableness, and temporality - Provenance-based
Health Data Quality Indicator van Deursen, Koster and Petković (2008)	HDQI	- Metric for reputation in healthcare - Considers reputation of information provider and metadata
EigenTrust Algorithm Kamvar, Schlosser and Garcia-Molina (2003)	ETA	- Metric for reputation in peer-to-peer file-sharing network with unique global trust value for each peer

Table 2: Dimensions and metrics of methods

Abbre-viation	Dimensions	Metrics
TDQM	Accuracy, Objectivity, Believability, Reputation, Access, Security, Relevancy, Value-Added, Timeliness, Completeness, Amount of data, Interpretability, Ease of understanding, Concise representation, Consistent representation	-
DWQ	Can be set as objectives	-
TIQM	Inherent dimensions: Consistency, Completeness, Accuracy, Precision,	-
TIQM	Nonduplication, Equivalence of redundant data, Concurrency of redundant data Pragmatic dimensions: Accessibility, Timeliness, Contextual clarity, Derivation integrity, Usability, Rightness, Cost	

Abbreviation	Dimensions	Metrics
AIMQ	Free-of-error, Appropriate amount of data, Concise representation, Relevancy, Completeness, Understandability, Consistent representation, Interpretability, Objectivity, Timeliness, Believability, Security, Accessibility, Ease of operation, Reputation	-
DQA	Accessibility, Appropriate amount of Data, Believability, Completeness, Concise Representation, Consistent Representation, Ease of Manipulation, Free-of-error, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability, Value-Added	Suggested percentage ratio
CDQ	Schema: Correctness with respect to the model, Correctness with respect to Requirements, Completeness, Pertinence, Readability, Normaliza-	Accuracy, Completeness, Currency, Timeliness, Volatility, Consistency

Abbreviation	Dimensions	Metrics
	tion - Data: Syntactic/Semantic Accuracy, Semantic Accuracy, Completeness, Consistency, Currency, Timeliness, Volatility, Completability, Reputation, Accessibility, Cost	
CC	Accuracy, Timeliness, Consistency, Completeness	Accuracy, Completeness, Consistency
DQDWS	Accuracy, Objectivity, Believability, Reputation, Relevancy, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Ease of Understanding, Concise Representation, Consistent Representation, Accessibility, Access Security	Accuracy, Consistency, Completeness, Amount of Data, Relevancy, Timeliness, Interpretability, Ease of Understanding, Consistent Representation
MMDQ	Ease of Manipulation, Accessibility, Reputation, Free of Error, Objectivity, Believability, Timeliness, Value-Added, Completeness. Appropriate	Completeness, Free of error, Concise Representation, Timeliness

Abbreviation	Dimensions	Metrics
	amount of data, Relevancy, Consistent Representation, Understandability, Interpretability, Concise Representation	
MDQA	Accuracy, Completeness, Consistency, Timeliness	Accuracy, Completeness, Consistency, Timeliness
MDB	Believability	Believability
HDQI	Reputation	Reputation
ETA	Reputation	Reputation

Due to the wide range of different methods, the optimal determination of data quality must always be based on the specific application. To connect the data quality requirements from dimensioning capacities to the methods for measuring data quality, the requirements will be selected in form of statements with assigned quality dimensions and later matched to the data quality measurement methods from this chapter. Based on the prioritization of the quality dimensions, an own method will be developed for the use case of capacity dimensioning on the base of existing methods.

4 Concept of Measuring Data Quality in Dimensioning Capacities

The requirements for data quality in the use case of dimensioning capacities from chapter 2 are summarized as statements in Table. Since the measurement of data quality and the weighting of the individual quality dimensions is strongly dependent on the case of application, the requirements of the use case are compiled to select a suitable procedure. The statements were collected within the research project to detail the requirements for determining data quality. Each statement is assigned the relevant data quality dimensions and clustered in one of the groups: Master data (MD), context (C) and framework (F).

Table 3: Statements of data quality requirements for dimensioning capacities

Statement	Dimension	Cluster
Digital form	Accessibility, Ease of manipulation	F
All locations have structured data in same way	Consistent representation, Interpretability, Objectivity	F
Centrally available dataset	Accessibility	F
Editable data format	Ease of manipulation	F

Statement	Dimension	Cluster
Content relevant datasets only	Appropriate amount of data, Relevancy	F
Compressed, complete dataset	Appropriate amount of data, Completeness	F
Master data of products must be maintained/ filled	Completeness, Timeliness	MD
Correct master data/ reliable data source	Believability, free of error, Reputation	MD
Consistency of master data (target/ actual)	Believability, Reputation	MD
Levels of aggregation of products (product key)	Completeness, Appropriate amount of data	C
All products from location must be listed	Appropriate amount of data, Completeness, Relevancy	C

Statement	Dimension	Cluster
Unique identifier for each product (e.g. material number)	Appropriate amount of data, Concise representation, Interpretability	C
Future products are included	Appropriate amount of data, Completeness	C
Current time horizon	Timeliness	C
Units are clearly defined	Concise representation	C
Restrictions for relation product - warehouse/handling/machine	Appropriate amount of data, Concise representation	C
No interpretation for attributes (e.g. material)	Interpretability, Understandability	C
Quantifiable dataset	Objectivity, Value added	C

For a structured comparison of these subjective statements there are two popular methods: Single stimulus and pairwise comparison method. In recent literature it was shown that the pairwise comparison method leads to more accurate and reliable results (Mantiuk, Tomaszewska and Mantiuk, 2012). In this method every object (criteria, alternatives, etc.) is compared

to all other objects on a scale from -2 to+2 (-2 meaning row is much less important than column, -1 row is less important than column, 0 both are equally important, 1 row is more important than column and 2 row is much more important than column) (Abdi and Williams, 2010; Zhang, et al., 2017). A decision per pair makes the choice easier than handling all choices simultaneously. After ranking each pair, the results can be displayed in a matrix, sum totals can be formed per row, the characteristics are weighed and ranks can be assigned in the proposed order. Especially where direct measurements are impractical the pairwise comparison method is of great value. The statements with the assigned dimensions are weighted, ranked and displayed by quality dimension in Table 4. The sum of the weighted points does not necessarily have to be zero, because the statements examined were assigned to different numbers of quality criteria.

Table 4: Ranked quality dimensions for dimensioning capacities

Rank	Weighted points	Quality dimension
1	27	Free of error
2	13	Concise representation
3	12	Believability, Reputation
4	4	Timeliness
5	3	Completeness

Rank	Weighted points	Quality dimension
6	0	Interpretability
7	-1	Ease of manipulation, Appropriate amount of data
8	-2	Value added
9	-7	Relevancy
10	-8	Understandability
11	-10,5	Objectivity
12	-12,5	Accessibility
13	-19	Consistent representation

This shows the five most relevant data quality dimensions for capacity planning: Free of error, concise representation, believability, reputation and timeliness. Applying the pairwise comparison also to the clusters, gives additional insights (see Table 5).

Table 5: Ranked clusters for dimensioning capacities

Cluster	Weighted points
Master data	11,33

Cluster	Weighted points
Context	2,11
Framework	-8,83

This leads to the conclusion that for dimensioning capacities in the SCD process clear and correct master data is more relevant than the context and the least relevant, the framework.

To identify a suitable method for dimension capacities in SCD the findings from this chapter are applied to the methods of assessing data quality, concerning the five key dimensions as well as the clusters: Master data and context.

4.1 Framework for Measuring Data Quality in Dimensioning Capacities

In the overview from Batini, et al. (2009) and in the additional literature research the main findings were qualitative methods and approaches with a focus on specialized metrics. Also a few hybrid methods containing both aspects were found. The five most significant dimensions for the presented use case are: Free of error (Fe), Concise representation (Cr), Believability (B), Reputation (R) and Timeliness (T). Since these are only partially included in the methods an overview of the approaches containing metrics is given in Table .

Table 6: Methods with metrics for top five quality dimensions of use case

Method	Fe	Cr	B	R	T
CDQ	X	X			X
CC	X	X			
DQDWS	X	X			X
MMDQ	X	X			X
MDQA	X	X			X
MDB			X		
HDQI				X	
ETA				X	

Four methods contain metrics for the dimensions free of error, concise representation and timeliness, but believability and reputation are not included. Hinrichs (2002) includes these three metrics and additionally most other metrics as stated in

Table (further metrics: Consistency, completeness, amount of data, relevancy, interpretability and consistent representation). For better comparison of the individual formulas, the metrics are normalized to the interval 0-1. Adapted metrics are presented to measure the dimensions on different levels: Attribute value level, tuple level, database level and relation level (Hinrichs, 2002).

After selecting the metrics from Hinrichs (2002) only three out of the five dimensions are provided. Therefore, the selection must be supplemented for the missing dimensions believability and reputation. These dimensions often tend to be estimated subjectively, but metrics can be found. These are presented in the following.

For the dimension believability the metric of Prat and Madnick (2008) is applicable. In this method trustworthiness, reasonableness and temporality are identified as the three components of believability. The values for each component are calculated regarding their data provenance (Prat and Madnick, 2008).

For the dimension reputation the metric proposed by van Deursen, Koster and Petković (2008) is fitting the requirements best. The method was developed for the application in the healthcare sector as a reputation-based health data quality indicator. This is especially relevant when patients provide their own information to health care providers and the quality cannot be guaranteed. The method was developed to address this problem and considers the reputation of the information provider and of the metadata provided by measurement systems (van Deursen, Koster and Petković, 2008).

Metrics have been identified for the five key dimensions. This addresses the challenge portrayed in Table with the most important cluster master data. Additionally, the second most important cluster, the context, should also be considered in this method. To meet this challenge a more general measurement method for data quality must be identified, in which the metrics can be embedded. The best combination of existing methods depends on the use case and the focus of the important dimensions. For this use the

applicable general method called Data Warehouse Quality Method from Jeusfeld, Quix and Jarkeet (1998) seems appropriate to apply. All data quality dimensions can be considered, as the first phase of the methodology is that the relevant objectives in the form of quality dimensions are set. The method briefly described in Table is a general approach for measuring data quality. The core of the method is assessing heterogeneous information from different sources to be able to integrate the information uniformly into a data warehouse. One step is to set objectives according to stakeholder groups that can also be quality dimensions. A quality meta model provides notation for formulating quality goals, queries, and measurements (Jeusfeld, Quix and Jarkeet, 1998; Batini, et al., 2009).

To combine the requirements from the use case dimensioning capacities and from the quality dimensions to cluster them into one solution, a two-stage method was developed (see Figure 5). For the general data quality assessment, the method is strongly inspired by the Data Warehouse Quality Method from Jeusfeld, Quix and Jarkeet (1998). First, the context for the assessment of data quality is defined. To formulate the quality goal, the purpose of the project and the different stakeholders must be considered. A focus has to be set for at least the five key quality dimensions in the use case: Free of error, Concise representation, Believability, Reputation and Timeliness. In order to measure the quality goal a quality query is needed against which the goal is calculated by the measuring agent. This value is saved as the expected value, which marks the starting point of the allowed range of values. The next step is the quality metric, the formula used for measuring the quality dimensions set as goals. In this use case the metrics needed for the five key quality dimensions are from Hinrichs (2002), Prat and Madnick (2008) and van Deursen, Koster and Petković (2008). When the

metrics are calculated a time stamp will be saved along with the actual measurement. The result from the measurement provides a value, which by itself has no meaning, but it can be evaluated with the quality query to check if the value is permitted or not, which is the quality domain. The quality domain will be reviewed in continuous intervals.

With the five key quality dimensions and master data and context considered a method for measuring data quality for dimensioning capacities was developed in this paper.

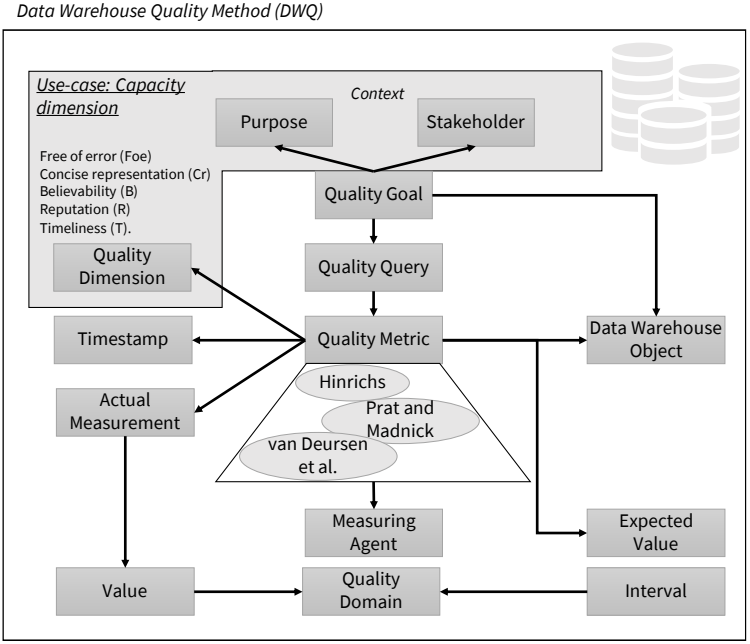


Figure 5: Two step methodology

5 Conclusion

This paper describes the use case of designing SC networks regarding integrating operative indicators in strategic planning. The research project's challenges occurred in the SCD task of dimensioning capacities with the quality of the available data for strategic planning. The challenge to measure data quality in the use case of dimensioning capacities was addressed in this paper. First the SCD tasks were outlined to further understand the context. Then the specific case of application was described and the occurred challenges with data quality. After exploring the theoretical foundation of data quality and methods for measuring and metrics were displayed it became clear, that the methods can be divided into two groups. On one side there are general methods that are defining guidelines, mostly with a focus on improving data quality, which tend to be subjective. On the other side there are specified methods and metrics that mostly focus on one dimension and consider a maximum of nine metrics. To measure the five most important dimensions from the use case along with the clusters of master data and the context, a new two-stepped methodology to assess data quality was developed. The integration of the general method and needed metrics to specifically meet the requirements of the use case provides a benefit for future planning.

In the next step the developed two step method must be validated with a use case to be able to evaluate its applicability. After that it can be assessed if the model can be applied to a wider spectrum of use cases by changing the key metrics. Additionally, it is advisable to collect metrics for the dimensions that are currently not considered in the developed method.

Future research is needed to adapt the metrics to the respective use cases and to develop an evaluation scale for classifying data quality in terms of the benefits that can be derived from the data (cost-benefit estimation).

Financial Disclosure

The results of this paper are based on the research project E^2-Design, funded by the German Federal Ministry for Economic Affairs and Energy (FKZ 03ET1558A).

References

Abdi, H. and Williams, L. J., 2010. Newman-Keuls test and Tukey test. Encyclopedia of Research Design. Thousand Oaks, CA: Sage, pp. 1–11.

Addo-Tenkorang, R. and Helo, P. T., 2016. Big data applications in operations/supply-chain management: A literature review. Computers & Industrial Engineering, 101, pp. 528–543.

Baghalian, A., Rezapour, S. and Farahani, R. Z., 2013. Robust supply chain network design with service level against disruptions and demand uncertainties: A real-life case. European Journal of Operational Research, 227(1), pp. 199–215.

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

Batini, C. and Scannapieca, M., 2006. Data Quality: Concepts, Methodologies and Techniques. [e-book]. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10145282>.

Baumgärtel, H., 2008. TSCP-Ein taktisches Supply Chain Planungsmodell für Kollaboratives Planen. Informations-und Kommunikationssysteme in Supply Chain Management, Logistik und Transport, DSOR Beiträge zur Wirtschaftsinformatik, 5, pp. 21–38.

Blake, R. and Mangiameli, P., 2011. The effects and interactions of data quality and problem complexity on classification. Journal of Data and Information Quality (JDIQ), 2(2), pp. 1–28.

Dash, R., McMurtrey, M., Rebman, C. and Kar, U. K., 2019. Application of Artificial Intelligence in Automation of Supply Chain Management. Journal of Strategic Innovation and Sustainability, 14(3).

English, L., 1998. Data quality: Meeting customer needs. Pitney Bowes white paper.

Erben, R. F. and Romeike, F., 2003. Komplexität als Ursache steigender Risiken in Industrie und Handel. In: 2003. Erfolgsfaktor Risiko-Management: Springer, pp. 43–63.

Fattahi, M., Mahootchi, M., Govindan, K. and Husseini, S. M. M., 2015. Dynamic supply chain network design with capacity planning and multi-period pricing. Transportation Research Part E: Logistics and Transportation Review, 81, pp. 169–202.

Fleischmann, B. and Koberstein, A., 2015. Strategic network design. In: 2015. Supply chain management and advanced planning: Springer, pp. 107–123.

Friemann, F., 2015. Strategische Lagerkapazitätsplanung: Ein Konzept zur stärkeren Integration in den strategischen Supply Chain Planungsprozess am Beispiel der pharmazeutischen Industrie. ETH Zurich.

Gudivada, V., Apon, A. and Ding, J., 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal on Advances in Software, 10(1), pp. 1–20.

Hinrichs, H., 2002. Datenqualitätsmanagement in data warehouse-systemen. Universität Oldenburg.

Hsu, C.-I. and Li, H.-C., 2009. An integrated plant capacity and production planning model for high-tech manufacturing firms with economies of scale. International Journal of Production Economics, 118(2), pp. 486–500.

Jayawardene, V., Sadiq, S. and Indulska, M., 2015. An analysis of data quality dimensions.

Jeusfeld, M. A., Quix, C. and Jarke, M., 1998. Design and analysis of quality information for data warehouses. In: Springer. International Conference on Conceptual Modeling, pp. 349–362.

Jones-Farmer, L. A., Ezell, J. D. and Hazen, B. T., 2014. Applying control chart methods to enhance data quality. Technometrics, 56(1), pp. 29–41.

Kamvar, S. D., Schlosser, M. T. and Garcia-Molina, H., 2003. The eigentrust algorithm for reputation management in p2p networks. In: Proceedings of the 12th international conference on World Wide Web, pp. 640–651.

Ketchen Jr, D. J. and Hult, G. T. M., 2007. Bridging organization theory and supply chain management: The case of best value supply chains. Journal of Operations Management, 25(2), pp. 573–580.

Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y., 2002. AIMQ: a methodology for information quality assessment. Information & management, 40(2), pp. 133–146.

Lewis, M. and Slack, N., 2017. Operations strategy. 5th ed.: Pearson Education.

Mantiuk, R. K., Tomaszewska, A. and Mantiuk, R., 2012. Comparison of four subjective methods for image quality assessment. In: Wiley Online Library. Computer graphics forum, pp. 2478–2491.

Minner, S. (2018): Kapazitätsdimensionierung von Produktionssystemen. In: Corsten: Handbuch Produktions- und Logistikmanagement in Wertschöpfungsnetzwerken.

Oppenheim, C., Stenson, J. and Wilson, R. M. S., 2003. Studies on information as an asset I: definitions. Journal of Information Science, 29(3), pp. 159–166.

Österle, H. and Otto, B., 2014. Das datenzentrierte Unternehmen: Eine Business-Engineering-Perspektive. In: 2014. Enterprise-Integration: Springer, pp. 91–105.

Parlings, M., Cirullies, J. and Klingebiel, K., 2013. A literature-based state of the art review on the identification and classification of supply chain design tasks.

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

Prat, N. and Madnick, S., 2008. Measuring data believability: A provenance approach. In: IEEE. Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), p. 393–393.

Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A. and Schmid, J., 2018. Informationsqualität-Definitionen, Dimensionen und Begriffe. In: 2018. Daten-und Informationsqualität: Springer, pp. 23–43.

Santoso, T., Ahmed, S., Goetschalckx, M. and Shapiro, A., 2005. A stochastic programming approach for supply chain network design under uncertainty. European Journal of Operational Research, 167(1), pp. 96–115.

Schreiber, L., 2019. Optimization and simulation for sustainable supply chain design. In: Berlin: epubli GmbH. Digital Transformation in Maritime and City Logistics: Smart Solutions for Logistics. Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 28, pp. 271–298.

Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H. and Mustapha, A., 2012. Data quality: A survey of data quality dimensions. In: IEEE. 2012 International Conference on Information Retrieval & Knowledge Management, pp. 300–304.

Statista, 2014. Supply Chain results using big data analytics. [online] Available at: <https://www.statista.com/statistics/491211/supply-chain-results-using-big-data-analytics/> [Accessed 18 May 2020].

van Deursen, T., Koster, P. and Petković, M., 2008. Hedaquin: A reputation-based health data quality indicator. Electronic Notes in Theoretical Computer Science, 197(2), pp. 159–167.

Waller, M. A. and Fawcett, S. E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), pp. 77–84.

Wang, R. Y., 1998. A product perspective on total data quality management. Communications of the ACM, 41(2), pp. 58–65.

Wassermann, O., 2013. Das intelligente Unternehmen: Mit der Wassermann Supply Chain Idee den globalen Wettbewerb gewinnen: Springer-Verlag.

Werner, H., 2017. Supply Chain Management: Grundlagen, Strategien, Instrumente und Controlling. 6th ed.: Springer.

Zhang, Z., Zhau, J., Liu, N., Gu, X. and Zhang, Y., 2017. An improved pairwise comparison scaling method for subjective image quality assessment. In: IEEE. 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–6.