Computers & Industrial Engineering 191 (2024) 110108

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Dimensions of data sparseness and their effect on supply chain visibility

Isabelle M. van Schilt [a,*], Jan H. Kwakkel [a], Jelte P. Mense [b], Alexander Verbraeck [a]

[a] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands
[b] National Policelab AI, Utrecht University, Princetonplein 5, 3584CC Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Data sparseness
Supply chain visibility
Data quality
Supply chain
Classification
Simulation

ABSTRACT

Supply chain visibility concerns the ability to track parts, components, or products in transit from supplier to customer. The data that organizations can obtain to establish or improve supply chain visibility is often sparse. This paper presents a classification of the dimensions of data sparseness and quantitatively explores the impact of these dimensions on supply chain visibility. Based on a review of supply chain visibility and data quality literature, this study proposes to characterize data sparseness as a lack of data quality across the entire supply chain, where data sparseness can be classified into three dimensions: noise, bias, and missing values. The quantitative analysis relies on a stylized simulation model of a moderately complex illicit supply chain. Scenarios are used to evaluate the combined effect of the individual dimensions from actors with different perspectives in the supply chain, either supply or demand-oriented. Results show that when a data sparseness of 90% is applied, supply chain visibility reduces to 52% for noise, to 65% for bias, and to 32% for missing values. The scenarios also show that companies with a supply-oriented view typically have a higher supply chain visibility than those with a demand-oriented view. The classification and assessment offer valuable insights for improving data quality and for enhancing supply chain visibility.

1. Introduction

The COVID-19 pandemic caused a steep rise in the worldwide demand for Personal Protective Equipment (PPE) such as face masks, gloves, goggles, and glasses (Omar, Debe, Jayaraman, Salah, Omar, & Arshad, 2022). To enable proper planning for purchasing and producing PPE in such a high-demand situation, a good insight into the overall supply chain is required. There is a range of PPE products available, which can generally be classified as medical PPE or non-medical PPE. Medical PPE is certified and has a higher price and profit margin. This made it attractive for fraudulent organizations to enter the market, and sell non-medical PPE as medical (Ippolito, Gregoretti, Cortegiani, & Iozzo, 2020). Hashemi, Huang, and Shelley (2022) found that during the initial stages of the COVID-19 pandemic, the majority of fraudulent PPE manufacturers emerged in Asia. However, counterfeit PPE activities and related logistics operations remained largely invisible due to little historical data on COVID-19 and on fraudulent organizations trying to obfuscate their data (van Schilt, Kwakkel, Mense, & Verbraeck, 2023). This counterfeit PPE case exemplifies a scenario where supply chain visibility is of the utmost importance, but it is hampered by sparse data (Zhao, Hong, & Lau, 2023).

Supply chain visibility focuses on the ability to track parts, components, or products in transit from supplier to customer, addressing the actors' capability to monitor and trace the movement of goods with accurate and timely information (Kalaiarasan, Olhager, Agrawal, & Wiktorsson, 2022; Saqib, Saqib, & Ou, 2019). When supply chain visibility increases, logistical processes within the supply chain can be more effectively aligned (Kalaiarasan et al., 2022; Srinivasan & Swink, 2018). For example, hospitals can more effectively prepare for stock-outs of medical PPE, or align with trustworthy organizations from whom they can buy legitimate medical PPE.

Even in this digital era, many supply chain organizations still face challenges in processing and retrieving visibility data. Tiwari, Wee, and Daryanto (2018), Wang, Gunasekaran, Ngai, and Papadopoulos (2016) and Wang and Zhuo (2020). Additionally, the data required to improve supply chain visibility, such as data on demand, inventory levels, processing times of a manufacturer, and transportation times, is often sparse (Kuipers, 2021; Somapa, Cools, & Dullaert, 2018). One of the causes is reluctance among actors within a supply chain to share (correct) data for various reasons such as competition and high costs of data solutions, or because of illegal behavior in case of fraudulent supply chain partners (Boone, Ganeshan, Jain, & Sanders, 2019). Other potential problems in data collection and sharing are malfunctioning sensors leading to biased values or missing data points, inconsistency in data formats between different systems, or simply typos (Oliveira & Handfield, 2019).

* Correspondence to: Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Room C2.020, 2628 BX Delft, Netherlands.
E-mail address: I.M.vanschilt@tudelft.nl (I.M. van Schilt).

https://doi.org/10.1016/j.cie.2024.110108
Received 22 March 2023; Received in revised form 11 December 2023; Accepted 29 March 2024
Available online 1 April 2024

Table of Notation

a_n	Quantity percentage of data for node n
u_{tn}	Bias value of time t and node n for the ground truth data
v_{tn}	Value of time t and node n for the ground truth data
v'_{tn}	Value of time t and node n for the sparse data
scv_n	Supply chain visibility for node n
scv	Global supply chain visibility
q_n	Quality percentage of data for node n
w_n	Weight for node n based on average inventory
A_n	Set of values that are not *NaN* for each node n, $\forall t \in T$
N	Set of nodes in the data of the supply chain model, $n \in N$, where each node represents an actor in the supply chain network
T	Set of elements in the time domain in the ground truth data of the supply chain model, $t \in T$
T^*	Set of elements in the time domain in the ground truth data indicating bias

Gaining more insight into the effect of data sparseness on supply chain visibility is essential for making improvements. A first step is to define data sparseness for supply chains. Unfortunately, there is no clear and agreed definition of data sparseness in the context of supply chain management. Various data quality issues can be seen as data sparseness, such as noise, bias, missing values, out-of-date information, different representations of the same data, or data that is not relevant for its use (Laranjeiro, Soydemir, & Bernardino, 2015; van Schilt et al., 2023). Laranjeiro et al. (2015) presents a large variety of poor data instances and how they impact data quality. Although a large variety of poor data instances is presented in the literature, a clear and concise formalization of data sparseness is still lacking, especially in the field of supply chain management. Oliveira and Handfield (2019) found that information quality plays a key role in supply chain visibility, and poor data resulting from data errors impacts decision-making. For example, when supply chain partners act on incomplete, inaccurate, and outdated data, this can lead to forecasting errors and supply chain disruptions (Agrawal, Kalaiarasan, Olhager, & Wiktorsson, 2022). Certain errors may have a more significant impact on supply chain visibility than others. For example, missing data values can result in a complete lack of knowledge of the supply chain, while noisy observations provide some indication of the value's magnitude in the supply chain (Laranjeiro et al., 2015). The exact effect of these different types of data errors (i.e., data sparseness) on supply chain visibility is still poorly understood.

This paper, therefore, focuses on the conceptualization of data sparseness in the context of supply chain management and the impact of data sparseness on supply chain visibility. Based on a review of supply chain visibility and data quality literature, a 3-dimensional classification of data sparseness is derived. Next, the effects of these dimensions of data sparseness on supply chain visibility are quantitatively assessed through a case study. A simulation model of a stylized supply chain of counterfeit PPE is used as ground truth. Complete data is extracted from this model, and then this data is systematically modified to increase sparseness along each of the three dimensions. Next, we assess how supply chain visibility changes. To evaluate the role of interaction effects between the three dimensions, we use scenarios to investigate the combined effect of the three dimensions of data sparseness. These scenarios describe data sparseness situations that could occur in real-life supply chains from the perspective of different actors, such as those positioned at the beginning of the supply chain (supply-oriented) or at the end (demand-oriented).

The contribution of this research is two-fold: (i) to provide a classification of data sparseness, and (ii) to assess its impact on supply chain visibility. By explicitly including data sparseness, our study is novel compared to the most recent systematic literature reviews on supply chain visibility of Kalaiarasan et al. (2022) and Somapa et al. (2018). Although both studies discuss data quality, they do not specifically focus on the dimensions of data sparseness and their impact on supply chain visibility. As for managerial implications, it is important for companies in a supply chain to be aware of the different dimensions of data sparseness and the differences in their impact on supply chain visibility. This might help companies to prioritize how to improve their data and, thereby, their visibility. Supply chain visibility is key for making the supply chain operations more efficient (Sodhi & Tang, 2019; Srinivasan & Swink, 2018).

The paper is structured as follows. Section 2 presents the method for performing the literature review. Section 3 discusses the current state-of-the-art for supply chain visibility. Section 4 reviews the literature on data quality. Section 5 combines these two bodies of literature and presents a classification of data sparseness. Section 6 formalizes data sparseness and supply chain visibility, explains the design of the simulation experiment, and introduces the case study. Section 7 presents the effects of an increasing degree of sparseness for each of the identified dimensions of data sparseness on supply chain visibility, and evaluates the effect of data sparseness on supply chain visibility for plausible real-life scenarios. Section 8 discusses the results. Section 9 concludes this study and provides directions for further research.

2. Literature review method

A literature review was conducted for papers in the fields of supply chain visibility and data quality. For a comprehensive overview, the authors have executed a systematic search for relevant literature following the method described by van Wee and Banister (2016). Database engines such as Scopus and Google Scholar were used to identify the relevant literature. The scope of this literature review was restricted to academic papers and books in English. Papers were selected based on the number of citations, while taking into account how recently the papers were published, to not miss recent contributions. Papers had to meet a minimum threshold of 50 citations, subject to their year of publication and relevance to the topic, with the exception of a few papers that provided a key insight into the literature but had fewer citations. The specified publication date range is from 2000 to 2023, permitting a few exceptions for older literature that is still heavily cited. This research examines two bodies of literature: supply chain visibility, and data quality. In searching for the papers, we explicitly looked for different viewpoints and approaches for supply chain visibility and data quality over the years.

For supply chain visibility, the first step was to search for current state-of-the-art literature defining supply chain visibility using the search keywords: "supply chain visibility", "supply chain transparency", "supply chain visibility definition", and "supply chain management and visibility". The date range for filtering the literature is from 2000 to 2023. Papers were selected based on the number of citations. In the second step, additional papers were found using snowballing. In the third step, the search was focused on the literature for measuring supply chain visibility with the date range of 2000 to 2023 using the search keywords: "measure", "calculate supply chain visibility", "assessing supply chain visibility", and "operationalize". We limited the papers to those that include the calculation of supply chain visibility, and rejected papers that only mention the characterization of supply chain visibility. For all papers, the title, keywords, introduction, conclusion, and approach section were scanned. Papers were selected

based on the number of citations, taking into account the publication date of the paper. In the fourth step, related papers were searched using snowballing.

For data quality, the first step was to search for current state-of-the-art literature on data quality with a date range from 2000 to 2023 using the specific search keywords: "data quality", "data quality dimensions", "characterize data quality", and "sparse data quality". In the second step, related papers were searched using snowballing. Some papers from before the year 2000 were also included in the literature search as there is a relatively older body of literature about data quality. In the third step, the search was targeted toward literature on data quality issues from year 2000 onwards using the keywords: "data issues", "degraded data", "data completeness", and "poor data". More in-depth papers on the definition of data quality issues were searched in the fourth step using snowballing and the specific keywords "measuring data quality issues" and "calculate noise/bias/missing values". In addition to recent research from the years 2000 to 2023, literature from before 2000 has also been included as a basis of reference.

Through this systematic literature review of the two bodies of literature, the overlap between the topics was examined, facilitating the identification and classification of dimensions of data sparseness in relation to supply chain visibility. The evaluation of the obtained publications involved assessing their quality and comprehensiveness through the application of a quality filter at the beginning of the search and during snowballing. The quality filter checked the relevance of the literature based on the publication's keywords, title, and abstract, as well as the impact factor of the journal of publication. The filter has been applied for the initial list of literature and for the literature resulting from snowballing. To obtain extra feedback, the results were presented and discussed by the researchers during a conference in the field of transport and logistics.

3. Supply chain visibility

In recent years, supply chain visibility has become key for improving supply chain management and design (Busse, Schleper, Weilenmann, & Wagner, 2017; Roy, 2021). Successful supply chain management is heavily dependent on the availability of information shared by multiple actors within the supply chain (Brun, Karaosman, & Barresi, 2020). Research shows that to improve competitiveness by reducing costs, fulfilling demand, enhancing operational efficiency, or increasing customer service, it helps to have a more visible supply chain (Lavastre, Gunasekaran, & Spalanzani, 2014; Swift, Guide, & Muthulingam, 2019). Supply chain visibility creates a valuable opportunity to gain insights and exchange knowledge with other stakeholders in the network, which in turn is beneficial for designing an efficient supply chain system (Somapa et al., 2018; Wei & Wang, 2010). Moreover, it facilitates action and reduces (decision) risk, making the supply chain more resilient (Rogerson & Parry, 2020; Saqib et al., 2019).

The outbreak of COVID-19 showed the vulnerabilities of supply chains with low visibility, leading to a vast array of distribution issues and shortages (Junaid, Zhang, Cao, & Luqman, 2023; Zhao et al., 2023). Both a lack of upstream visibility to the suppliers and downstream visibility to the customers existed (Busse et al., 2017; Kalaiarasan et al., 2022).

Our literature overview focuses on the definition of supply chain visibility, and the methods for assessing and measuring it. The current state-of-the-art papers on these topics are used for operationalizing supply chain visibility for this research.

3.1. Definition

Supply chain visibility is a commonly and broadly used term in supply chain and logistics with a variety of meanings. Francis (2008, p. 182) proposes a general definition based on a literature review: *"Supply chain visibility is the identity, location and status of entities transiting the supply chain, captured in timely messages about events, along with the planned and actual dates/times for these events".* Similar to Saqib et al. (2019), this definition assumes that a detailed picture of the entities, i.e., any object moving through the supply chain, is needed. Providing complete information about all objects in the supply chain presents a challenge for the stakeholders in the supply chain, who might need to provide confidential and competitive information, and as a result, they are often reluctant to share such information (Pero & Rossi, 2014; Wang & Zhuo, 2020). Second, not all stakeholders benefit from improved supply chain visibility: having too much information without a clear use case can be a distraction. Barratt and Oke (2007) includes the extent to which data is key or useful for supply chain visibility according to their definition. This definition is often referred to by other authors Kalaiarasan et al. (2022). Concluding, a weakness in the general definition offered by Francis (2008) is the absence of the relevance of the information for the stakeholders. Barratt and Oke (2007), McCrea (2005) and Schoenthaler (2003) do include this relevance in their definitions of supply chain visibility.

Later, Williams, Roh, Tokar, and Swink (2013) adds the quality of supply and demand information on accuracy, timeliness, completeness, and usability in their definition of supply chain visibility. Kalaiarasan et al. (2022, p. 4) takes this a step further by defining supply chain visibility as *"the extent to which actors within a supply chain have visual access to the timely and accurate demand and supply information that they consider to be key or useful to their operations and supply chains."*

Most literature indicates that supply chain visibility is dependent on good data, either stating usefulness or data quality dimensions. Some definitions require a detailed picture of the entire supply chain (Francis, 2008), while other definitions are more aggregated on either the supply or the demand side (Barratt & Oke, 2007; Kalaiarasan et al., 2022; Williams et al., 2013). Combining the major insights from the literature, supply chain visibility for this research is defined as:

> *Supply chain visibility refers to the ability of tracking parts, components or products in transit from supplier to customer through relevant data of stakeholders.*

Next to the dependence on good quality data, supply chain visibility also depends on the willingness of organizations to share this data. Bartlett, Julien, and Baines (2007) uses transparency as a measure of visibility, and combines it with a degree of obscurity. Sodhi and Tang (2019) refers to supply chain visibility as the company's effort to gather information and data, and supply chain transparency as the company's willingness to share information with the public. Brun et al. (2020) notes that collaboration amongst supply chain partners and the level of trust should increase to achieve supply chain visibility. Since our study does not focus on the general public but on supply chain partners, transparency in the context of supply chain visibility is defined as the willingness to share relevant data with stakeholders.

3.2. Methods for assessing supply chain visibility

Somapa et al. (2018) is the most recent literature review that discusses the characterization and the quantification of supply chain visibility in a network. They define three characteristics to capture supply chain visibility: (1) accessibility of information, (2) quality of information, and (3) usefulness of information. The first characteristic focuses on the capability of information and communication technology (ICT) systems to collect data, whereas the other two characteristics focus on the quality of information for obtaining the organization's goal. In recent years, a new generation of ICT systems has arisen to collect data for improving supply chain visibility. One of the most interesting recent concepts is the Internet-of-Things (IoT), consisting of Internet-embedded sensors and ICT components to provide data on supply chain and logistics activities (Calatayud, Mangan, & Christopher, 2019). IoT

can make more data accessible such that the supply chain is more visible for all actors in real-time (Kumar, Singh, Mishra, & Wamba, 2022). Another useful concept is the Radio-frequency identification transponder (RFID), an auto-identification system for detecting objects and elements while they move along the supply chain (Pero & Rossi, 2014). Kalaiarasan, Agrawal, Olhager, Wiktorsson, and Hauge (2023) notes that many studies show how these concepts can be used for improving supply chain visibility. The authors show the potential of this new generation of ICT systems, including IoT, RFID and blockchain, for collecting data in real-time from all stakeholders. One of the key aspects for these systems is the collaboration between actors within the supply chain (Kalaiarasan et al., 2023; Pero & Rossi, 2014). As mentioned before, competition can limit the necessary collaboration between actors. In our example case of counterfeit PPE supply chains, the necessary collaboration and sharing of data are out of the question, since data can give away information about illegal activities. Collaboration between supply chain partners is therefore not always a given.

Somapa et al. (2018) gives an overview of quantitative and qualitative approaches for measuring supply chain visibility. Common quantitative methods are regression analysis, visibility scorecards, utilization ratios, and mathematical models rooted in, e.g., set theory. Only a few methods consider the global supply chain level instead of the firm level (Somapa et al., 2018). One of these methods is presented by Zhang, Goh, and Meng (2011) who measure supply chain inventory visibility by using set theory. They define visibility as the capability to access and provide information among several companies. Lee and Rim (2016) uses the Six Sigma method to evaluate the end-to-end supply chain visibility with a focus on operational capabilities. In contrast to studies that focus on the information perspective of visibility, they focus on the visibility of processes to asses whether the supply chain has the capability to execute the supply chain plan (Somapa et al., 2018). Lee and Rim (2016) calculate the mean and standard deviation of individual processes for lead time, yield, quality, and utilization.

Another method to determine supply chain visibility that includes the end-to-end supply chain is the calculation of geometric means of information quantity and quality shared between the other actors and the focal company, as designed by Caridi, Crippa, Perego, Sianesi, and Tumino (2010), Caridi, Perego, and Tumino (2013). A strength of this paper is that the authors focus on measuring supply chain visibility in complex networks, which is particularly challenging. In contrast, most of the literature focuses on relatively simple two-tier or linear supply chains. Caridi et al. (2010, 2013) is a notable exception by giving a quantitative approach to assess the degree of supply chain visibility in complex systems for inbound and outbound logistics. They distinguish four types of information flows for supply chain visibility: (1) transactions/events, (2) status information, (3) master data, and (4) operational plans. They measure visibility as the amount and the quality of information the focal company possesses, compared to the total information that could be obtained. First, the visibility that the focal company has of each individual actor in the supply chain is measured by supply chain managers who judge the quality and the quantity of information available for providing visibility. These judgments are collected for each type of information flow and for each supply chain actor on a relative scale from 1 (lowest) to 4 (best). An argument against this technique is that it is subjective. After obtaining the judgments, the individual visibility measures are combined to calculate the global visibility. The global measure is the weighted average of visibility for each actor. The weight for each actor is based on how much the focal company purchases from an actor, and how much an actor buys from the focal company, and the distance between the companies in terms of the number of tiers and vertical integration. So, the more an actor sells to or buys from the focal company or the closer it is to the focal company in the supply chain, the higher the weight.

3.3. Operationalization

Combining the insights of Calatayud et al. (2019), Caridi et al. (2010, 2013), Kalaiarasan et al. (2022) and Somapa et al. (2018), this research measures supply chain visibility as *the weighted average of the available information quantity and quality divided by their theoretical maximum for all actors in the supply chain given the goods, information, and financial flows.* The characteristics for measurement can be captured by the quality and the quantity of information (Caridi et al., 2010, 2013; Somapa et al., 2018). This means that accessibility of information (e.g., the capability of IT systems, IoT, RFID) will be out of scope. Three types of flows can be distinguished for measuring supply chain visibility: the goods flow, the information flow, and the financial flow (Min & Zhou, 2002; Stadtler & Kilger, 2002). For each of these flows, data can be extracted to assess visibility. This paper primarily focuses on the goods flow.

To measure supply chain visibility, the available quantity and quality of the information are compared to their theoretical maximum (Caridi et al., 2010). Instead of using expert judgments, quantitative measures are used to calculate the quantity and the quality. Quantity is measured as the percentage of the number of data points that are available to the actor in comparison to the full data set. Quality is measured as the mean absolute percentage error of the data set of the actor compared to the full data set. Along the lines of Caridi et al. (2010), these percentages are combined into a geometric mean to determine the supply chain visibility of an individual actor.

Similar to Caridi et al. (2010), supply chain visibility is first measured for each actor, but without the presence of a focal company. Next, the supply chain visibility scores of individual actors are aggregated into a global measure using a weighted average. The weight of an actor is determined by the number of orders and the costs they represent. The weight is assigned to each corresponding actor to determine the weighted visibility of the actor. The sum of the visibility scores of all actors in the supply chain results in a percentage value for the global supply chain visibility.

4. Data quality

Data quality is a topic that has been researched for many years and in various disciplines (Ehrlinger & Wöß, 2022). Data quality management involves data collection (data profiling), the characterization of data quality, the measurement of data quality, and data quality monitoring (Bronselaer, 2021). Our research focuses on sparse data with a low volume, whereas big data literature focuses on high volumes of data (see e.g., Günther, Mehrizi, Huysman, & Feldberg, 2017; Jeble, Dubey, Childe, Papadopoulos, Roubaud, & Prakash, 2018). Therefore, literature on big data, e.g., the 5 V's for the quality of data: Volume, Variety, Velocity, Veracity, and Value (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015) is kept out of scope.

4.1. Data quality dimensions

Several papers have provided a categorization of data quality. Wang and Strong (1996) defines data quality as "the fitness of use" and presents a framework of data quality aspects that are important to data customers. They identify four main categories: (1) accuracy, (2) relevance, (3) representation, and (4) accessibility. Pipino, Lee, and Wang (2002) defines a detailed list of sixteen data quality dimensions based on a survey of healthcare, finance, and consumer product companies. Most of them fall into the categories of Wang and Strong (1996). One new dimension has been added: ease of modification, i.e., the level to which data is easy to modify. Fan and Geerts (2012) states five central issues for data quality: (1) data consistency, i.e., the validity of data to the real-world, (2) data deduplication, i.e., multiple points referring to the same real-world entity (3) data accuracy, i.e, closeness of a value

to its true value, (4) information completeness, i.e., complete data to answer the question, and (5) data currency, i.e., timeliness.

Huang (2013) aggregates data quality into three main categories: (1) syntactic quality, the level to which data follows the rules of a data model, with subcriteria including accuracy and consistency, (2) semantic quality, the level to which data is relevant and required for the purpose, with subcriteria including accuracy, completeness, and mapping consistency, and (3) pragmatic quality, the level to which data is suitable for a given application, with subcriteria including completeness, timelineness, and presentation suitability. Hazen, Boone, Ezell, and Jones-Farmer (2014) defines four dimensions of data quality in the context of supply chain management: (1) accuracy, the degree to which data has errors, i.e., the degree to which it is similar to the "real" value, (2) timeliness, the degree to which data is up-to-date, (3) consistency, the degree to which similar data is presented in the same format, and (4) completeness, the degree to which necessary data is available. These dimensions are similar to the subcriteria of Huang (2013) and the taxonomy presented in Gao, Xie, and Tao (2016). In a comparison study on data quality frameworks, Cichy and Rass (2019) shows that accessibility, accuracy, completeness, consistency, and timeliness have the highest number of occurrences as data quality criteria. Although there is an ongoing discussion on the dimensions of data quality in literature, the criteria identified by the above authors (accuracy, timeliness, consistency, completeness) are the most frequently used ones to describe data quality (Ehrlinger & Wöß, 2022).

4.2. Data quality issues

Data quality issues such as data sparseness or errors in the data for one or more of the dimensions of data quality lead to poor decision quality (Bronselaer, 2021; Heinrich, Hristova, Klier, Schiller, & Szubartowicz, 2018). Accuracy decreases when data deviates from the "real" value; timeliness decreases when the data is outdated; consistency decreases when different data points are not presented in the same format; completeness decreases when there is missing data (Souibgui, Atigui, Zammali, Cherfi, & Yahia, 2019). Additionally, in the case of (partly) illicit supply chains, data can be manipulated or masked to avoid detection (van Schilt et al., 2023). In terms of the data quality dimensions, this study identifies and addresses three main data quality issues that are relevant for decision-making: noise, bias, and missing values (Janssen, van der Voort, & Wahyudi, 2017; Oliveira, Rodrigues, & Henriques, 2005).

Noise in data results in corrupted or distorted data, potentially rendering it meaningless (Sáez, Galar, Luengo, & Herrera, 2014). Noise in a data point is generally defined as a deviation of that particular data point, where the distribution of the deviation has a mean and a noise width (Gaussian noise) (Teng, 1999; Zhu & Wu, 2004; Zhu, Wu, & Yang, 2004). So, a data point with noise results in the original value plus or minus a deviation. There is a difference between noise that is inherent (natural), and injected (artificial). When analyzing noise, it is important to take this distinction into account (Seiffert, Khoshgoftaar, Van Hulse, & Folleco, 2014).

Bias in data means that the data is not representative of the population or the phenomenon of study (Tripepi, Jager, Dekker, & Zoccali, 2010). Bias means that some members are more likely to be included than others, thus the probability of a member being included is unequally distributed. Data produced by humans may contain bias as a result of human preferences or human observational capabilities. The most common types of bias are (i) selection bias, i.e., group representation, (ii) reporting bias, i.e., some observations are more likely to be reported than others, and (iii) detection bias, i.e., a phenomenon is more likely to be observed than others (Ntoutsi et al., 2020).

Missing values relate to the completeness of a data set. Peng, Hahn, and Huang (2023) presents a review and notes that missing values are a widespread data quality problem. They categorize missing values into three categories, building on research by Rubin (1976). This first category addresses data that is missing completely at random, meaning the absence of a data value is based on a random sample of the complete data set. The second category of missing values is missing at random, meaning the absence of a data value is related to some properties of the observed data (the data set without the missing values) but not to the missing data. The third category is missing not at random, meaning the absence of a data value is systematically correlated to properties of the missing data itself (Fox, 2015). As an example of the second category, people with a higher age are more likely to withhold information on their income, meaning that the probability of missing data depends on the age (a property of the observed data). As an example of the third category, people with a higher income are more likely to withhold information on their income, meaning the probability of missing data depends on the income level itself (a property of the missing data).

5. Classification of data sparseness

In this section, the literature on supply chain visibility is combined with the literature on data quality for classifying data sparseness in the field of supply chain management. First, the overlap between the two bodies of literature is discussed. Next, the classification of data sparseness based on the literature review is presented.

Supply chain visibility is primarily determined by the quality and the quantity of the data (Caridi et al., 2010; Kalaiarasan et al., 2022). For quality, the data quality criteria of Ehrlinger and Wöß (2022), Gao et al. (2016) and Huang (2013) are used as these are the most frequently used ones to describe data quality. Quality and quantity of data are specified by the syntactic and semantic criteria, more specifically by their accuracy, consistency, and completeness. In the field of supply chain management, these data quality criteria are of relevance for enhancing supply chain visibility, and for informed decision-making (Kalaiarasan et al., 2022; Munir, Jajja, Chatha, & Farooq, 2020). Accuracy ensures precise information on important supply chain variables such as inventory, order status, and lead times. This helps, for example, to make accurate demand forecasts to prevent excess inventory, which is important for the predictive real-time nature of the supply chain. Consistency ensures reliable data of supply chain operations that is shared between stakeholders. For example, a consistent data format between stakeholders helps to track the movement of goods. Considering the multi-sourced and geospatial characteristics of a supply chain, a consistent data set shared among the many stakeholders is of high importance to enhance supply chain visibility by enabling the tracking of the movement of goods and inventory levels. Completeness ensures comprehensive data of the supply chain operations. For example, this helps to anticipate demand and avoid stockouts, or to enable efficient planning when looking at the temporal characteristics of the supply chain.

For data *sparseness*, three main issues for data quality are distinguished: noise, bias, and missing values (Janssen et al., 2017; Oliveira et al., 2005; van Schilt et al., 2023). These issues are classified as the dimensions of data sparseness. The logical relationships between the dimensions of data sparseness and data quality criteria including their impact on supply chain visibility are illustrated as follows: Noise impacts the accuracy and the consistency of data quality. For example, in a case where the inventory of medical PPE is monitored manually, a typo in the data leads to noise. Inaccurate data on the inventory levels affects the accuracy and reliability of the supply chain visibility. Bias impacts the consistency and completeness of the data. For example, using the PPE case again, large hospitals could be overrepresented in the supply chain data, making small hospitals invisible. This would make the supply chain data skewed and incomplete as there is less information on small hospitals. As a result, fewer resources could be allocated to smaller hospitals, leading to stockouts. Missing values impact the completeness criteria. As an example in the PPE case, there could be no data on the lead times from the supplier to the hospitals,

Table 1
Classification of data sparseness in three dimensions.

	Description	Level of data quality	Fraction of intentional sparseness
Noise	Distortedness.	Value is modified by adding a deviation following a distribution in x% of original data elements.	Noise is for y% intentionally sparse in the data.
Bias	Representativeness.	Value is structurally more likely to be present in x% of the original data elements.	Bias is for y% intentionally sparse in the data.
Missing values	Completeness.	Value is missing in x% of the original data elements.	Missing values is for y% intentionally sparse in the data.

meaning that the hospitals have no visibility on how to manage their stock.

Other criteria, such as the pragmatic criteria of Huang (2013) and the timeliness of Ehrlinger and Wöß (2022), are not included in our classification. These criteria describe the relevance of the data, and indicate whether it is suitable and up-to-date for a given application. However, relevance is a very different kind of criterion than noise, bias, and missing values. Relevance concerns the applicability of the data set as a whole given a specific type of analysis or decision, whereas the other dimensions concern the modification of values within the data set for any analysis or decision purpose (Bronselaer, 2021). Especially considering the temporal and dynamic nature of a supply chain, the relevance of the data is subject to time-sensitive and up-to-date information. For example, accurate demand forecasting needs a relevant and up-to-date data set but still faces challenges when some data values with the data set are inaccurate, inconsistent, and incomplete.

Data in a supply chain can either be sparse by itself (i.e., unintentional sparseness) or sparse by manipulation (i.e., intentional sparseness) (Bartlett et al., 2007). Intentional manipulation of data is also a data quality issue (Janssen et al., 2017). The willingness of stakeholders to share this sparse data is the primary factor that determines transparency in the context of supply chain visibility (Baah et al., 2022; Wang & Zhuo, 2020). Combining (un)intentional sparseness and (non-)transparency leads to four cases, where stakeholders are either (1) willing to share unintentionally sparse data to *improve* supply chain management, (2) unwilling to share unintentionally sparse data to *hide* data, (3) willing to share intentionally sparse data to *mislead* other stakeholders, or (4) unwilling to share intentionally sparse data to *prevent* poor data availability. The fraction of intentional sparseness of the data has an impact on how to cope with data in supply chain management and how to use it in decision-making (Bronselaer, 2021; Oliveira & Handfield, 2019). For example, if a supplier intentionally withholds key production data about the fabric of medical PPE to gain a competitive advantage, the manufacturer may make sub-optimal decisions on inventory levels, leading to potential disruptions in the supply chain and increased costs.

This literature study led to the following definition of sparse data in relation to supply chain visibility:

> *Sparse data in supply chain management refers to the lack of data quality across the entire supply chain for the quality dimensions: noise, bias, and missing values, where a certain fraction of data sparseness is intentional.*

Table 1 presents the classification of sparse data in the context of supply chain management. In summary, there are three dimensions of data sparseness: (1) **noise**, i.e., values in the data set are distorted; (2) **bias**, i.e., values in the data set are not representative of the population or the phenomenon of study; (3) **missing values**, i.e., values in the data set are missing. Each dimension has a certain fraction of intentional sparseness. Thus, each dimension of data sparseness consists of (i) the level of data quality, and (ii) the fraction of intentional sparseness.

6. Methods

In this research, the effect of the identified dimensions of data sparseness on supply chain visibility is assessed by systematically increasing the degree of sparseness in the data. First, the quantification of the dimensions of data sparseness is described. Second, the formalization of global supply chain visibility is discussed. Next, the design of experiments using a ground truth simulation is explained. Last, the case study used in this research for performing experiments is presented.

6.1. Formalization of dimensions

Let $t \in T$ be an index t in the set of elements of the time domain T in the data. Let $n \in N$ be a node n (in this case, an actor) in the set of nodes N in the data. Let v_{tn} be a value of time t and node n for the ground truth data, and let v'_{tn} be a value of time t and node n for the sparse data. The degree of data sparseness is systematically increased on the three identified dimensions of data sparseness as follows:

Noise level of x% is defined as x% of original data elements are modified by adding a deviation following a distribution. This means that, over the entire data set, x% of the data has noise. It is randomly determined, using a discrete Uniform distribution, which elements of the data set have noise. The deviation of the noise follows a Gaussian distribution with a standard deviation of 1. A value with noise can be defined as:

$$v'_{tn} \sim v_{tn} + \mathcal{N}(\mu = 0, \sigma = 1) \tag{1}$$

Bias level of x% is defined as values that are structurally more likely to be present in x% of the original data elements. A sample of x% of the rows is randomly drawn to represent bias. Every row is allocated a weight through a log-normal distribution with $\mu = 0$ and $\sigma = 1$, and a sample is selected based on these weights. For example, there are 100 data rows with 25% bias. This means that on average 25 rows are sampled using the weights resulting from the log-normal distribution, and replace a selected row from the ground truth data set. The other 75 rows remain the same as the ground truth data set. Let $T^* \subset T$ be the set of elements in the time domain indicating bias. Let u_{tn} be a historical value that is already present in the data set, and used to create bias:

$$u_{tn} \in \{v_{t'n'} : t' \in T, n' \in N\} \tag{2}$$

A value with bias can be defined as:

$$v'_{tn} = \begin{cases} v_{tn}, & \forall t \notin T^*, \\ u_{tn}, & \forall t \in T^* \end{cases} \tag{3}$$

Missing values level of x% means that a value is missing in x% of the original data elements. Similar to the noise level, it is randomly determined which x% of data points are missing over the entire data set by following a discrete Uniform distribution. A missing value can be defined by a non-value (*NaN*, indicating *Not a Number*) as follows:

$$v'_{tn} = NaN \tag{4}$$

Important to note is that a non-value differs from a zero value. In the context of supply chain management, many true values can be zero, such as zero inventory of a product, so a missing value is encoded as *NaN* rather than as zero (Heinrich et al., 2018).

6.2. Formalization of supply chain visibility

Supply chain visibility is measured by comparing the available quantity and quality of the information to its theoretical maximum, as described in Section 3.3. The calculation of supply chain visibility in our research is as follows: first, the quantity and the quality of the information at each node are measured. For each node $n \in N$, the quality as a percentage is defined as follows:

$$q_n = 100 - MAPE(v_n, v'_n) \tag{5}$$

where MAPE is the mean absolute percentage error relative to the average of the data elements of the node. Hereby, the magnitude of the mean absolute percentage error is taken into account. MAPE is defined as,

$$MAPE(v_n, v'_n) = \frac{100}{\#T} \sum_{t \in T} \left| \frac{v_{tn} - v'_{tn}}{\overline{v_n}} \right| \tag{6}$$

For each node $n \in N$, the quantity as a percentage is defined as follows:

$$a_n = 100 \times \frac{\#A_n}{\#T}, A_n = \{v'_{tn} \neq NaN, \forall t \in T\} \tag{7}$$

The supply chain visibility for each $n \in N$ is calculated as:

$$scv_n = \sqrt{q_n \times a_n} \tag{8}$$

Second, the weight of each node in the supply chain is determined. The weight is based on the average number of orders w_n of each node n. The average number of orders is normalized over all nodes. This gives,

$$w_n = \frac{\overline{v_n}}{\sum_{n \in N} \overline{v_n}} \tag{9}$$

The global supply chain visibility as a percentage can be calculated as follows:

$$scv = \sum_{n \in N} w_n \times scv_n \tag{10}$$

6.3. Design of experiments

This research uses a ground truth simulation model to evaluate and compare the supply chain visibility for varying degrees of data sparseness in each of the dimensions. The simulation model calculates the ground truth values to obtain the theoretical maximum quality and quantity of information. This set-up allows for correctly assessing how supply chain visibility changes as the true maximum is known which is often not the case in real life (Khondoker, Dobson, Skirrow, Simmons, & Stahl, 2016).

Fig. 1 presents the method used for calculating supply chain visibility for various degrees of data sparseness using the ground truth. First, the ground truth data for each time element $t \in T$ and each node $n \in N$, v_{tn}, is extracted from the simulation model. This ground truth data does not include any sparseness. Then, a certain percentage of data sparseness is added: noise, bias, and missing values. Next, the ground truth data values (v_{tn}) and the sparse data values (v'_{tn}) are used to calculate the supply chain visibility. First, the supply chain visibility is calculated for each node, $n \in N$, using the ground truth data and the sparse data. The quantity and the quality of the sparse data is compared to the ground truth. Next, the weights of each node are determined based on the average number of orders. Then, these measures are combined to a global supply chain visibility (indicated by *SCV* in Fig. 1) as a percent value.

Two experiments are performed in this study: (1) systematically increase the degree of data sparseness for each individual dimension, and (2) design and evaluate plausible real-life scenarios with regard to data sparseness. First, the degree of data sparseness is systematically increased by 10% for each dimension. More specifically, the experiments are 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%. For the ground truth data, i.e., the base case, there is no data sparseness in any dimension so the supply chain visibility is always 100%.

Second, the effect of the three dimensions of data sparseness on stylized scenarios that are theoretically plausible in real-life supply chains is evaluated. In these stylized scenarios, all three dimensions of data sparseness are used as most real-world data sets include all these dimensions of data sparseness. The dimensions are added in the following order to the data set: (1) add bias, so bias only exists for true values of the data, (2) add noise to this biased data set, (3) delete values to create missing values. The configuration of these stylized scenarios is presented in Section 7.2.

Each experiment is performed with 200 unique seeds to account for the effect of stochasticity on supply chain visibility. By transforming the ground truth data using the same seeds for each experiment, it is ensured that the exact same observations are modified for each dimension of data sparseness.

6.4. Case study

In this research, a stylized counterfeit PPE supply chain is used as a case study for performing experiments. This supply chain is characterized by sparse data since (1) the production of counterfeit PPE during COVID-19 presented a new and unexpected phenomenon with little historical data, and (2) counterfeit PPE supply chains are operated by fraudulent organizations that obfuscate as much data as possible (Hashemi, Jeng, Mohiuddin, Huang, & Shelley, 2023).

Fig. 2 visualizes the stylized counterfeit PPE supply chain simulation model. The symbols in the figure represent the main actors in the supply chain, and the arrows represent the transportation flows. The supply chain starts with the raw materials supplier, placed in this stylized case in Vietnam, who supplies products for PPE such as fabrics. Next to China and India, many PPE come from Vietnam (Nikkei Asia, 2020). These products are transported over land to one of the two manufacturers in the same country, Vietnam. These manufacturers produce counterfeit PPE in the factory and pack them in batches for transport. Each batch has a certain quantity of counterfeit PPE. For example, a batch consists of 2000 boxes of 200 PPE which equals a quantity of 200,000 PPE in total. Next, a batch of finished counterfeit PPE is transported from the manufacturers' location via a truck to the export port in Hai Phong, Vietnam. The batch is loaded into a 40 ft container and transported by a small container ship to the transit port, Tanjung Pelepas, Malaysia. The small container ship unloads the container with counterfeit PPE at the transit port. At the same port, the container is loaded onto a larger container ship for overseas transport. The destination of this ship, also the import port, is either the Port of Rotterdam, The Netherlands, or the Port of Antwerp, Belgium. The container is unloaded at one of these ports and waits for inland transport to the (illegal) wholesales distributor in Eindhoven, The Netherlands. This means when the container arrives in Antwerp, the truck crosses a land border to arrive at the wholesales distributor. At the wholesales distributor, the batch of counterfeit PPE in the container is equally divided into three smaller batches for the three retailers. These smaller batches are transported by small trucks to the retailers in Amsterdam, Utrecht, and Venlo in The Netherlands. When the counterfeit PPE arrives at the retailer, the products are being sold with or without knowing that they are counterfeit.

A discrete event simulation model of this stylized configuration of a counterfeit PPE supply chain from Vietnam to stores in the Netherlands is used to gather the ground truth data. Table 2 shows the input parameters for the actors and the links used in the stylized simulation model.

In the simulation model, most uncertainties such as delays of transport modalities and speed of transport modalities follow triangular distributions inspired by real-world data of a fashion retailer (Kuipers, 2021). Table 3 shows the input parameters and the distributions of the speed and the delays of the transport modalities for the simulation

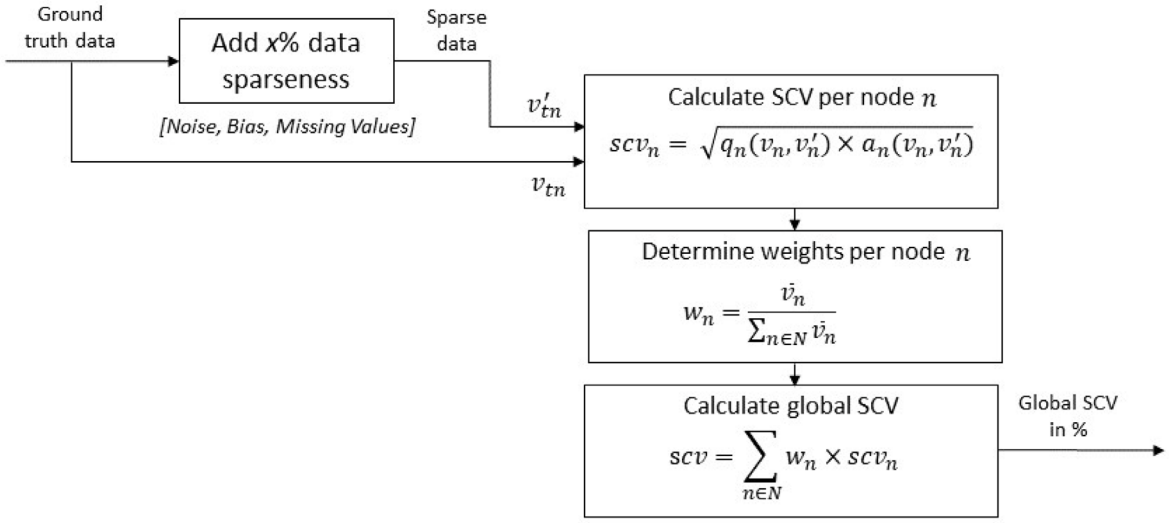


Fig. 1. Method for calculating global Supply Chain Visibility (SCV).

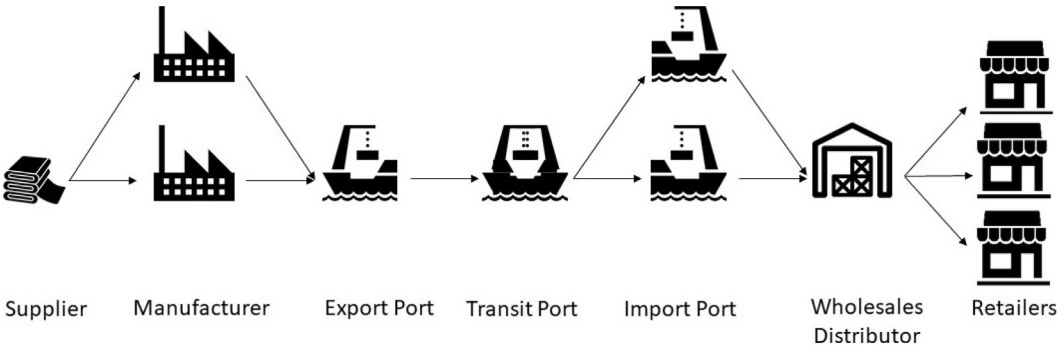


Fig. 2. Stylized supply chain of counterfeit PPE.

Table 2
Input parameters of actors and links for the simulation model of the stylized counterfeit PPE supply chain.

Actors				Links		
Input parameter	Distribution	Value	Unit	Name	Value	Unit
Interarrival time of product at supplier	Exponential	1.5	days	Supplier to manufacturer 1	50	km
Time at manufacturer	None	2.5	days	Supplier to manufacturer 2	45	km
Time at ports	Triangular	1, 2, 2	days	Manufacturer 1 to export port	125	km
Time at wholesales distributor	Triangular	0.5, 1, 2	days	Manufacturer 2 to export port	100	km
Time at retailers	Exponential	0.2	days	Export port to transit port	1656	nautical miles
				Transit port to import port Rotterdam	9286	nautical miles
				Transit port to import port Antwerp	9195	nautical miles
				Import port Rotterdam to wholesales distributor	135	km
				Import port Antwerp to wholesales distributor	100	km
				Wholesales distributor to retailer Amsterdam	125	km
				Wholesales distributor to retailer Utrecht	92	km
				Wholesales distributor to retailer Venlo	60	km

model of this study. This case study represents a complex network suitable for our study due to the many uncertainties in the supply chain simulation model (e.g., delay in transport modalities, loading and unloading times). For example, the retailer's inventory can fluctuate very much, depending on whether a vessel has a 1-day delay or a 7-day delay.

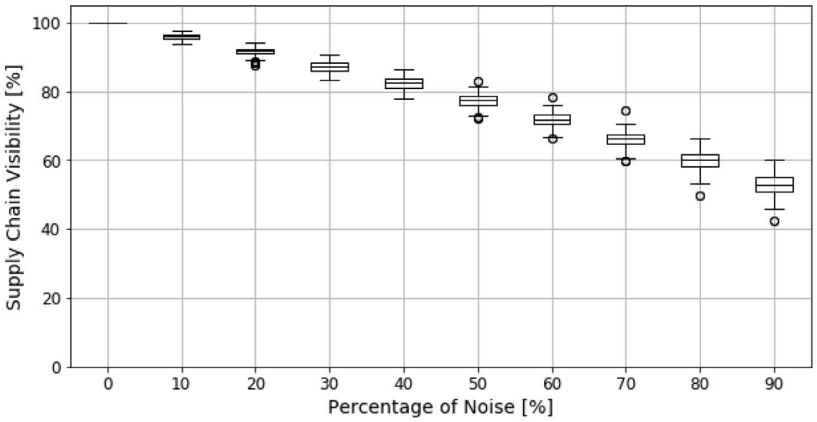
From this simulation model, time series data on the stylized supply chain is extracted as ground truth data. The time series data entails data on the inventory that is located at each actor (e.g., manufacturer, export port, import port) in the supply chain per day. Each data element in the time series data is thus the inventory of an actor at a specific time. The mean inventory value per day is calculated for the multiple replications of the simulation model. A simulation time of 52 weeks with 20 unique replications is used. The simulation model has been developed with the library *pydsol* in Python. This library is a Python implementation of the Distributed Simulation Object Library (DSOL), originally implemented in Java (Jacobs, 2005).

7. Results

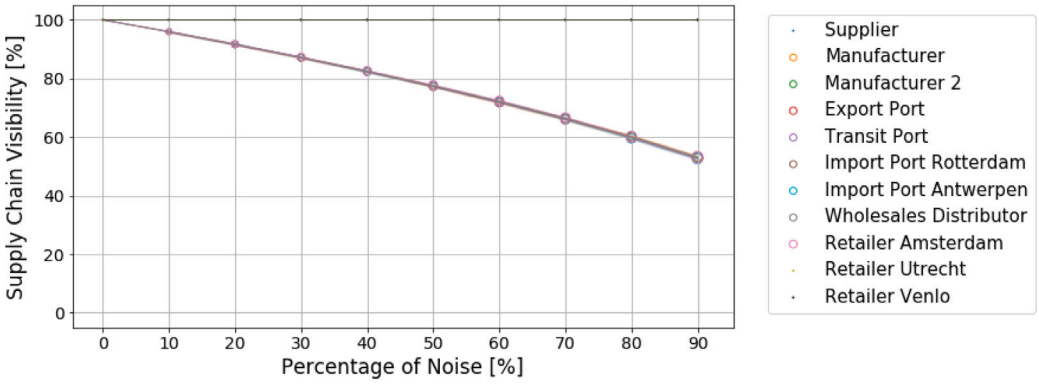
This section presents the results of variations in supply chain visibility given an increasing degree of sparseness for each of the identified dimensions of data sparseness using the case study. Next, the plausible scenarios that could theoretically occur in real life are described. These

Table 3
Input parameters of speed and delay of the transport modalities for the simulation model of the stylized counterfeit PPE supply chain.

Transport modalities							
Input parameter	Distribution	Value	Unit	Input parameter	Distribution	Value	Unit
Speed of small truck	Triangular	0, 100, 120	km/h	Delay of small truck	Triangular	0, 0.2, 0.5	days
Speed of large truck	Triangular	0, 80, 120	km/h	Delay of large truck	Triangular	0, 0.5, 1	days
Speed of feeder	Triangular	10, 18, 25	knots	Delay of feeder	Triangular	0, 4, 16	days
Speed of vessel	Triangular	10, 18, 25	knots	Delay of vessel	Triangular	0, 7, 16	days



(a) Boxplot of global supply chain visibility.



(b) Supply chain visibility per node.

Fig. 3. Results for dimension noise for 200 seeds for various degrees of data sparseness.

scenarios are evaluated for the impact of data sparseness on supply chain visibility.

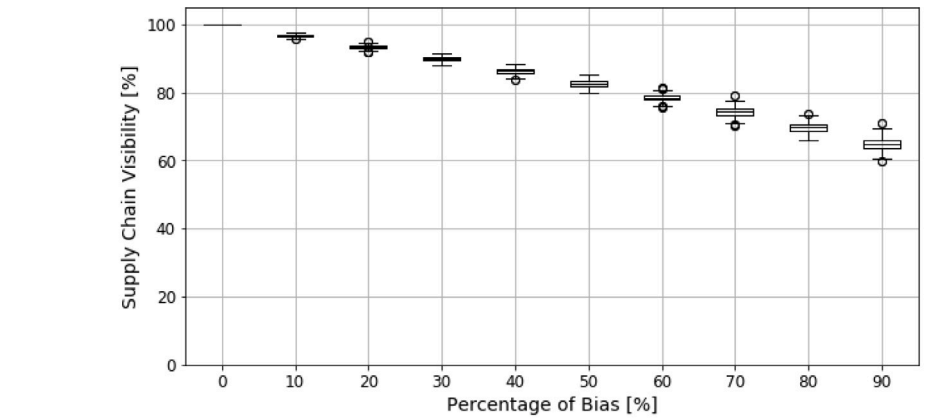
7.1. Effect of the individual dimensions

Figs. 3, 4, and 5 show, for each individual dimension of data sparseness, a boxplot of global supply chain visibility for various degrees of data sparseness. The boxplot displays the minimum, the 1st quartile (i.e., 25th percentile), the median, the 3rd quartile (i.e., 75th percentile), and the maximum of the percentage of supply chain visibility for each degree of data sparseness. Also, the average supply chain visibility of each actor in the supply chain over various degrees of data sparseness including a 95% confidence interval is shown. The size of the markers in the plot is equal to the size of the 95% confidence interval.

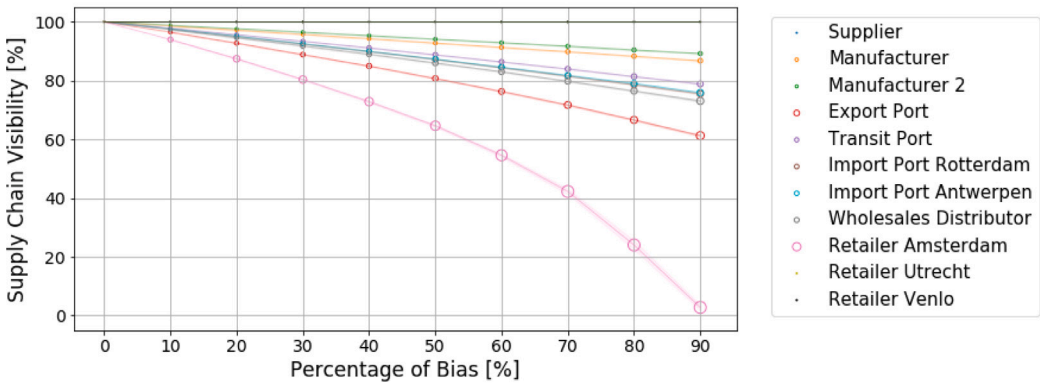
Fig. 3(a) presents the boxplot of supply chain visibility when adding noise to the ground truth data. It shows that the global supply chain visibility gradually decreases when more noise is present in the data with average steps of 4% to 7% per 10% of extra noise. The highest median value of supply chain visibility, excluding the base case, is 95.9% at 10% noise. The lowest median value of supply chain visibility is 52.9% at 90% noise. The spread of the supply chain visibility over the 200 seeds becomes wider with a higher degree of noise, meaning that the interquartile distance (i.e., the distance between the 1st and 3rd quartiles) becomes wider. However, this distance stays limited to at most 4.3%. The distance between the minimum and the maximum value of supply chain visibility becomes even wider over the various degrees of noise with the largest distance of 14% at 90% data sparseness.

When looking more closely at which actors contribute to this spread, Fig. 3(b) shows that most actors follow the same decreasing trend over the various degrees of noise regarding their supply chain visibility. Represented by the size of the marker in this figure, the retailer in Amsterdam has the widest confidence interval of 1.2% when increasing the degree of noise in the data. Other actors have a confidence interval between 0.8% to 1.0% at the highest degree of data sparseness (90%).

Fig. 4(a) shows the boxplot of global supply chain visibility when adding bias to the ground truth data. The boxplot shows that supply chain visibility decreases when more bias is present in the data with on average steps of 3% to 5% per 10% of extra bias. The highest median value of supply chain visibility, excluding the base case, is 96.8% at 10% bias. The lowest median value of supply chain visibility is 64.9% at 90% bias. The spread of supply chain visibility becomes wider up to 50% bias with an interquartile distance from 0.5% to 1.4%, and the distance between the minimum and the maximum values from



(a) Boxplot of global supply chain visibility.



(b) Supply chain visibility per node.

Fig. 4. Results for dimension bias for 200 seeds for various degrees of data sparseness.

1.8% to 5.5%. At 60% bias, the spread becomes smaller (4.5%) and afterwards, it increases by 2% for 70% data sparseness. After 70%, the spread becomes wider with the widest spread at 90% bias with an interquartile distance of 4.9% and a distance between the minimum and the maximum of 8.8%.

When looking at the supply chain visibility per actor including the 95% confidence interval in Fig. 4(b), it shows that the average supply chain visibility percentage of the actors retailer in Amsterdam converges to 2.7% at 90% bias. From 60% onwards, the average supply chain visibility of retailer in Amsterdam is decreasing steeply with steps of 10% to 20%, and with a confidence interval higher than 1.2%. This could explain why the spread of the global supply chain visibility is smaller at 60% bias, and becomes considerably wider afterwards. Also, for this actor, the average supply chain visibility percentage decreases relatively steeply compared to the other actors. The percentage of supply chain visibility of the manufacturers gradually decreases with on average steps of 1% to 2% per 10% bias increase over the various degrees of bias as data sparseness. For the transit port, import ports, and wholesales distributor, supply chain visibility decreases with average steps of 2% to 3%. The percentage of supply chain visibility of the actor export port decreases slightly more steep with average steps of 4% to 5% when increasing bias in the data.

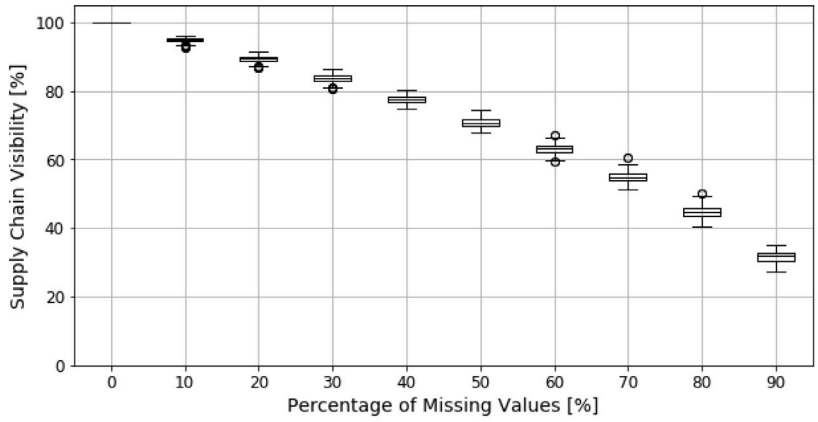
Fig. 5(a) presents the boxplot of global supply chain visibility when adding missing values to the ground truth data. It shows that the supply chain visibility decreases when more missing values are present in the data. The decrease starts with steps of 5% to 6% per 10% increase in missing values. From 50% missing values onwards, the median value of supply chain visibility decreases with 7% to 13% per 10% step. The highest median value of supply chain visibility, excluding the base case, is 94.8% at 10% missing values. The lowest median value of supply chain visibility is 31.9% at 90% missing values in the data. The spread of the supply chain visibility is relatively small but increases over the various degrees of missing values. The interquartile distance is 0.8% at 10% missing values and is gradually increasing to 2.4% at 90% missing values. The distance between the minimum and the maximum value of supply chain visibility is increasing from 2.8% to 8.6%.

When looking at the supply chain visibility for each actor in Fig. 5(b), it shows that most actors in the supply chain follow the same trend regarding the average percentage of supply chain visibility over the various degrees of missing values. The 95% confidence interval of all the actors, except for the retailer in Utrecht and the retailer in Venlo, becomes slightly wider when the percentage of data sparseness increases. However, this is still not more than 0.7%.

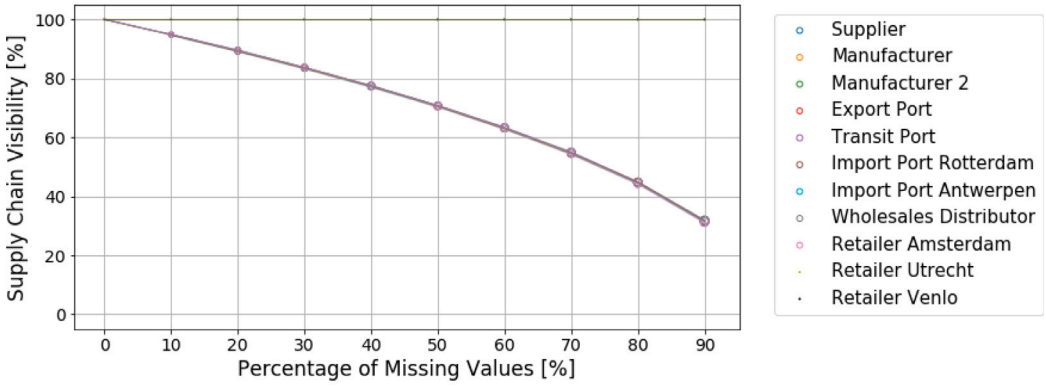
Table 4 shows the mean and the standard deviation, i.e., the spread, of the supply chain visibility as a percentage for each dimension in more detail. It can be observed that the standard deviation of the supply

Table 4
Supply chain visibility (%) mean and standard deviation for each dimension of data sparseness and for various degrees of data sparseness.

Percentage of	Noise		Bias		Missing	
data sparseness	Mean	Std	Mean	Std	Mean	Std
0%	100.0	0.0	100.0	0.0	100.0	0.0
10%	95.9	0.8	96.8	0.4	94.8	0.6
20%	91.7	1.2	93.4	0.5	89.5	0.9
30%	87.2	1.4	89.9	0.7	83.7	1.1
40%	82.4	1.6	86.3	0.8	77.5	1.2
50%	77.3	1.8	82.5	1.0	70.7	1.4
60%	71.8	2.0	78.5	1.0	63.2	1.5
70%	66.1	2.3	74.3	1.4	54.8	1.5
80%	59.9	2.7	69.8	1.5	44.9	1.7
90%	52.8	3.0	65.0	1.9	31.7	1.7



(a) Boxplot of global supply chain visibility.



(b) Supply chain visibility per node.

Fig. 5. Results for dimension missing values for 200 seeds for various degrees of data sparseness.

chain visibility increases when more noise is added to the data. The increase of the standard deviation from 0.8% at 10% sparseness to 3.0% at 90% sparseness is the highest of all dimensions. The table also makes clear that missing values assert the most influence on supply chain visibility. For missing values, the average percentage of supply chain visibility decreases all the way down to 31.7% for 90% data sparseness. Missing values has the lowest standard deviation over most degrees of data sparseness compared to noise and bias.

7.2. Scenario analysis

To compare the effect of data sparseness on real-life supply chain cases, plausible scenarios that theoretically could occur in a supply chain for assessing supply chain visibility are developed. Table 5 presents the configuration of the percentages of noise and missing values of four stylized scenarios: (i) competitor, (ii) key actor, (iii) supply-oriented, and (iv) demand-oriented. For each scenario, a bias of 25% over the entire data set is added as real-life data often includes values that are structurally more present than others. For example, companies have structurally more information on their own inventory than on the inventory of other actors.

The first scenario, competitor, reconstructs the case where only one of two actors in a competitive position in a supply chain is willing to share data. A possible reason is that an actor is reluctant to share good data for competitive reasons. In our case, this is a manufacturer (referred to as manufacturer 2) with a noise of 95% and missing values of 95%. In real life, it is unlikely that the data of the other actors is perfect. To account for this, the other actors have a noise of 10% and 25% missing values.

The second scenario, key actor, shows the case where an actor at a key position, i.e., in the middle of the supply chain, only provides sparse data to the rest of the supply chain with noise and missing values of 95%. Similar to the competitor scenario, the other actors have a noise of 10% and 25% missing values.

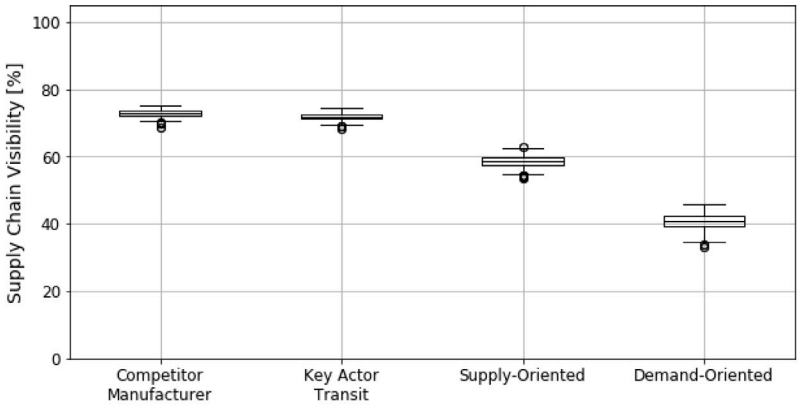
The third scenario, supply-oriented, represents the case where much is known on the supply side (starting with only 10% noise and 25% missing values for the supplier), and less is known on the demand side (ending with 80% noise and 95% missing values for the retailers). This often holds for suppliers as they have more high-quality information on actors upstream than downstream, represented by gradually degrading data over the actors in the supply chain.

The fourth scenario, demand-oriented, represents the case where much is known on the demand side (starting with only 10% noise and 25% missing values for the retailers), and less is known on the supply side (ending with 80% noise and 95% missing values for the supplier). The retailers have a higher quality and quantity of information on the actors close to them, i.e., downstream. Similar to the supply-oriented scenario, this is represented by a gradual increase in the percentage of noise and missing values following the sequential ordering of the upstream actors in the supply chain.

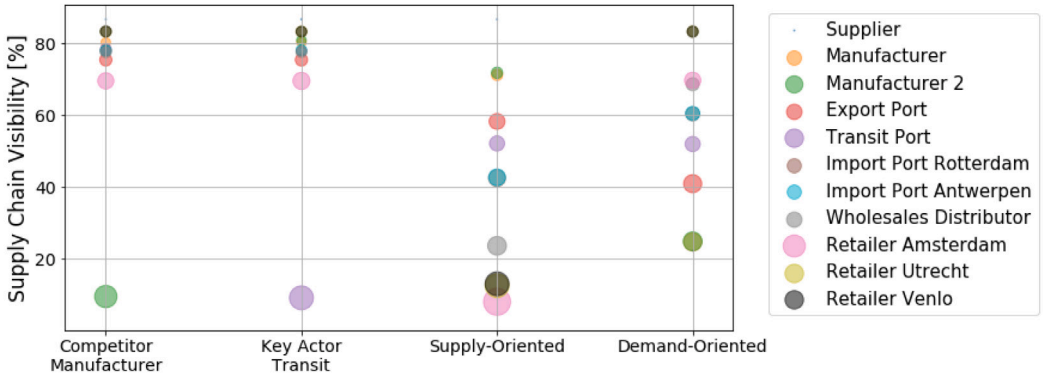
Fig. 6(a) shows the boxplot of the average global supply chain visibility percentage for each scenario. In the scenarios where only one actor provides sparse data, the competitor and the key actor, the global supply chain visibility is 72.9% and 71.9% respectively. The spread of these two scenarios over the 200 seeds is small as the interquartile distance for both scenarios is only 1.3%, and the distance between the minimum and the maximum values is at most 5.1%. Fig. 6(b) presents the average supply chain visibility percentage per actor with the size of the marker representing the 95% confidence interval. From this, it can be observed that the decrease in global supply chain visibility is directly correlated with a low average supply chain visibility of the particular actor that has a high noise and a high number of missing

Table 5
Configuration of percentage of noise and missing value for each actor in % for four scenarios.

	Scenarios							
	Competitor		Key actor		Supply-oriented		Demand-oriented	
	Noise	Missing	Noise	Missing	Noise	Missing	Noise	Missing
Supplier	10	25	10	25	10	25	80	95
Manufacturer 1	10	25	10	25	20	35	70	85
Manufacturer 2	95	95	10	25	20	35	70	85
Export port	10	25	10	25	30	45	50	65
Transit port	10	25	95	95	40	55	40	55
Import port 1 & 2	10	25	10	25	50	65	30	45
Wholesales	10	25	10	25	70	85	20	35
Retailer 1, 2 & 3	10	25	10	25	80	95	10	25



(a) Boxplot of global supply chain visibility.



(b) Supply chain visibility per node per scenarios. The size of the markers represents the 95% confidence interval of the supply chain visibility for each actor.

Fig. 6. Results of the scenarios (1) Competitor, (2) Key actor, (3) Supply-Oriented, (4) Demand-Oriented over 200 seeds.

values in each of the two scenarios. The average supply chain visibility of Manufacturer 2 and the transit port is around 9.2% with a 95% confidence interval of 0.6%. Other actors have an average supply chain visibility between 69.5% to 86.6%.

In the scenarios where noise and missing values are gradually added to the actors in the supply chain, either supply-oriented or demand-oriented, the global supply chain visibility is 58.5% and 40.6% respectively (see Fig. 6(a)). For the supply-oriented scenario, the spread is small with an interquartile distance of 2.0% and a distance between the minimum and maximum values of 7.8%. For the demand-oriented scenario, the spread is wider with an interquartile distance of 3.0% and a distance between the minimum and the maximum of 11.1%.

When looking at the supply-oriented scenario, the average supply chain visibility per actor in Fig. 6(b) decreases over the supply chain. Given the sequential ordering of the actors in the supply chain, the supplier and the manufacturers have the highest average supply chain visibility between 86.6% and 89.7%. The actors with the lowest average supply chain visibility in the supply-oriented scenario are the retailers; between 7.9% to 12.9% with a relatively wide confidence interval.

For the demand-oriented scenario, a similar pattern of sequentially decreasing average supply chain visibility over the actors in the supply chain is present but then reversed in comparison to the supply-oriented scenario. The actors with the highest average supply chain visibility are the retailers in Utrecht and Venlo with 83.5%, and the supply chain visibility of the retailer in Amsterdam is 69.6%. Actors with the lowest average supply chain visibility are the manufacturers with 24.7% and a relatively small confidence interval. Fig. 6(b) also shows that the global supply chain visibility of this scenario is the lowest and it has the widest spread.

8. Discussion

Six main elements are addressed that are essential for properly understanding and interpreting the results of this study: (1) impact of

artifacts of the simulation model, (2) use of sampling method, (3) way of calculation of supply chain visibility, (4) specificity to a sequential supply chain, (5) lack of including intentionality, and (6) limitation on the incorporation of the data collection process.

First, the results show that in all three individual dimensions of data sparseness, the actors at the outer end of the supply chain, i.e., the supplier, the retailer in Utrecht, and the retailer in Venlo, have zero to little spread in their supply chain visibility or are not influenced (i.e., the visibility remains at 100%) when adding data sparseness. A reason is that the inventory of these actors is often zero as this is the starting or the ending node of the supply chain. This is an artifact of the simulation model as the product does not stay at the supplier for long (e.g., not longer than 1 day), and products are assumed to be sold or used quite quickly after arriving at the retailer. Since the average inventory of these actors is low, the weights for calculating the global supply chain visibility are also low (Caridi et al., 2010, 2013). Therefore, these outliers have little impact on the resulting global supply chain visibility. Interestingly, when adding all three dimensions of data sparseness to each actor in the scenarios, the supply chain visibility of the supplier, the retailer in Utrecht, and the retailer in Venlo, are somewhat affected by data sparseness, but the effects are very limited.

Second, the sampling method for the dimensions of data sparseness affects the results depending on the data quality criterion for which data sparseness is introduced (Laranjeiro et al., 2015). The missing values dimension results in a small 95% confidence interval and the lowest standard deviation (not more than 1.7%) for the supply chain visibility. An explanation is that the missing values dimension only impacts the quantity of the data. It is more straightforward in which way the data is transformed, so the spread is low. For noise and bias, dimensions that affect the quality of the data, the ranges on how the data can be transformed are wider and, therefore, the spread in supply chain visibility outcomes is larger. Also, as bias is sampled using a log-normal distribution, there is a higher probability that the correct information of some actors is more often present in the data than the correct information of others. This leads to a higher quality of the data of those actors and, therefore, a higher supply chain visibility (Kalaiarasan et al., 2022). It explains that, in the bias dimension, the average supply chain visibility of most actors is relatively high in comparison to the fast-decreasing supply chain visibility of the retailer in Amsterdam. The data is sampled using a Uniform distribution for the dimensions noise and missing values. As the data of all actors are equally likely to be modified following this Uniform distribution, the actors logically follow the same trend regarding the impact on visibility in these dimensions.

Third, the way of calculating supply chain visibility is of importance when interpreting the results. For the scenarios where only one actor is impacted, the competitor scenario and the key actor scenario, the average supply chain visibility percentage is approximately the same. However, in supply chain theory, a "bull-whip" effect of information would be expected in the key actor scenario, i.e., every actor upstream of the key actor would also be less visible due to the sparse data of the key actor (Lee, Padmanabhan, & Whang, 1997). This means that, theoretically, degrading data in the key actor scenario leads to a lower global supply chain visibility than in the competitor scenario. However, this effect is not represented in the formulas of global supply chain visibility, and therefore, the results of these two scenarios are the same. This lack of including the "bull-whip" effect is a limitation for the calculation.

When comparing demand-orientation and supply-orientation, the results show that the demand-oriented scenario has a lower global supply chain visibility than the supply-oriented scenario. Additionally, the supply-oriented scenario includes more actors with low supply chain visibility. A cause for this phenomenon is that the weights assigned to each actor for calculating global supply chain visibility are based on average order quantity in units (i.e., inventory levels), following Caridi et al. (2010). Actors upstream in the supply chain generally have more average inventory than those downstream as they use a make-to-stock approach. More specifically, the PPE supply chain is a push supply chain where the supplier and manufacturer create inventory for the long-term demand instead of a pull supply chain where they respond to real-time demand (Nag, Han, & Yao, 2014). This entails that the supplier and the manufacturer have a high weight, contributing more to the global supply chain visibility according to the formula used in our study. Thus, the results hold for cases where the average inventory is a key indicator for determining global supply chain visibility. In other words, the supply chain characteristics are important for calculating the average inventory and, therefore, for the validity of our results. Next to the push and pull characteristic, the structure of the supply chain plays a crucial role in determining the average inventory of actors (Li, Zobel, Seref, & Chatfield, 2020). For example, if an assembly supply chain of a car were studied with many suppliers of small products like windows and steering wheels, the inventory load might be differently distributed than in the case of PPE. It would be interesting to examine whether these results hold for different types of complex supply chains where inventory is distributed differently.

Fourth, the results are specific to the linear counterfeit PPE supply chain model used in our study. A supply chain is often represented as a sequential network, meaning that, for example, there is a one-directional flow between supplier and manufacturer. On the one hand, this direct and linear dependency between the actors could lead to a more straightforward calculation of supply chain visibility, being a limitation to the generalizability of the results. On the other hand, many supply chains are characterized by a sequential network, even when there are more actors involved. Thus, the effect of the dimensions of data sparseness on supply chain visibility is generalizable to other supply chains with similar complexity.

Fifth, the quality and the quantity of the data, hence the supply chain visibility, are not directly affected by the intentionality of data sparseness as it does not matter whether the actor intentionally transformed the data for calculating supply chain visibility in this study. Therefore, the intentionality aspect of data sparseness is not included in our analysis. However, coping with sparse data and using it for decision-making is different when data is intentionally transformed (Janssen et al., 2017; Oliveira & Handfield, 2019). For example, when bias is intentionally added to the data of counterfeit PPE, it is most likely that fraudulent organizations try to mask their real activities, and planning effective interventions on this biased data is difficult. Whereas, if data is unintentionally sparse, masking of data for one specific actor in the supply chain does not take place, and effective interventions can still be planned on the biased data. The studied scenarios for a key actor hiding information and a competitor hiding information could be seen as first experiments with intentional data sparseness. As the fraction of intentional sparseness impacts how to cope with data and how to use it in decision-making, it would be interesting to examine the impact of intentionality on data sparseness for decision-making (Bronselaer, 2021).

Last, a limitation of the systematic literature review on supply chain visibility and data quality is that the data collection phase was kept out of scope. For the purpose of this research, only the impact of data sparseness on supply chain visibility has been studied. The literature study provided some possibilities on decreasing data sparseness during the data collection phase, such as the use of IoT, RFID, and blockchain (Kumar et al., 2022; Pero & Rossi, 2014). Extending this research by analyzing how to improve data quality for all phases of the data management process and how to rank these solutions would be interesting for academics and practitioners.

9. Conclusion

Improving data quality is crucial for enhancing supply chain visibility, because accurate and comprehensive data allows for informed

decision-making, monitoring operations, enhancing resilience, and mitigating potential inefficiencies (Bronselaer, 2021; Munir et al., 2020). Poorly informed supply chain management decisions may result from data sparseness, creating challenges for stakeholders to coordinate effectively, and potentially resulting in shortages of products (Janssen et al., 2017; Kalaiarasan et al., 2022). Therefore, it is important to make supply chain practitioners aware of the different dimensions of data sparseness and how these dimensions impact supply chain visibility. However, no clear and concise formalization of data sparseness exists in the current state-of-the-art literature on supply chain management. Additionally, a knowledge gap exists in understanding the extent of the impact caused by different dimensions of data sparseness. Addressing these knowledge gaps is essential for enhancing the ability of supply chain practitioners to deal with data sparseness, and for contributing to further developments in the supply chain field by explicitly including the notion of data sparseness and its impact.

This research addresses the gaps in the existing literature by providing a classification of data sparseness in the context of supply chains and assessing its impact on supply chain visibility. First, using a systematic literature review, data sparseness is classified into three dimensions: (1) noise, i.e., values in the data set are distorted, (2) bias, i.e., data is not representative of the population or the phenomenon of study, (3) missing values, i.e., values are missing in the data. Each dimension has a certain fraction of intentional sparseness. Thus, sparse data in relation to supply chain visibility is referred to as: *"lack of data quality across the entire supply chain for the quality dimensions: noise, bias, and missing values, where a certain fraction of data sparseness is intentional"*.

Next, the impact of these dimensions on the supply chain visibility is evaluated for an increasing degree of data sparseness. A stylized counterfeit PPE supply chain simulation model is used as ground truth. Data is extracted from this model, and then data sparseness for the three dimensions is systematically added to this data. Hereby, the magnitude of change in supply chain visibility for an increasing degree of data sparseness on each individual dimension is assessed. Four stylized scenarios that could occur in real life regarding data sparseness and their effect on supply chain visibility are also examined.

The main research findings demonstrate that data sparseness greatly affects the visibility of the counterfeit PPE global supply chain. More specifically, data sparseness impacts supply chain visibility, leading to a reduction of up to 52.8% for noise, 65.0% for bias, and 31.7% for missing values. For all three individual dimensions, the average percentage of global supply chain visibility decreases when more sparseness is added to the data, and the visibility values have a small 95% confidence interval. The missing values dimension has the largest impact on the decrease in supply chain visibility, whereas bias has the least impact. The results show the relative importance of the dimensions of data sparseness for actors in the supply chain. The scenario analysis shows that the location of an actor who is unwilling to share data (either a competitor or a key actor) makes no difference for the global supply chain visibility percentage when using the current formulas. The scenario analysis also shows that the demand-oriented scenario has the lowest average global supply chain visibility at 40.6%. A reason is that the global supply chain visibility percentage decreases more when actors with a high average inventory provide sparse data. It also shows that companies with a supply-oriented view will have a better insight into the supply chain visibility than those with a demand-oriented view.

To provide practical advice, this study helps supply chain practitioners by providing information on the relationship between dimensions of data sparseness and supply chain visibility. The primary impact on supply chain visibility appears to be missing data, suggesting that supply chain practitioners should prioritize addressing missing values to improve supply chain visibility. Additionally, companies with a demand-oriented view should prioritize collecting data from upstream as much as possible. This would enhance their decision-making capabilities.

Future research should focus on evaluating the impact of data sparseness on different supply chain configurations in the context of supply chain visibility, e.g., non-sequential supply chain networks. Following on this, expanding the complexity of the simulation model (e.g., including more actors), and therefore, the complexity of the data set is also a direction for future research. Another research direction is to investigate the inclusion of the "bull-whip" effect in the calculation of supply chain visibility, and to include intentionality for evaluating decision-making with data sparseness. A final research direction is to research methods to enhance the data quality management process.

CRediT authorship contribution statement

Isabelle M. van Schilt: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. **Jan H. Kwakkel:** Conceptualization, Methodology, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Jelte P. Mense:** Conceptualization, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Alexander Verbraeck:** Conceptualization, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Agrawal, T. K., Kalaiarasan, R., Olhager, J., & Wiktorsson, M. (2022). Supply chain visibility: A delphi study on managerial perspectives and priorities. *International Journal of Production Research*, 1–16. http://dx.doi.org/10.1080/00207543.2022.2098873.

Baah, C., Opoku Agyeman, D., Acquah, I. S. K., Agyabeng-Mensah, Y., Afum, E., Issau, K., et al. (2022). Effect of information sharing in supply chains: Understanding the roles of supply chain visibility, agility, collaboration on supply chain performance. *Benchmarking: An International Journal*, *29*(2), 434–455. http://dx.doi.org/10.1108/BIJ-08-2020-0453.

Barratt, M., & Oke, A. (2007). Antecedents of supply chain visibility in retail supply chains: a resource-based theory perspective. *Journal of Operations Management*, *25*(6), 1217–1233. http://dx.doi.org/10.1016/j.jom.2007.01.003.

Bartlett, P. A., Julien, D. M., & Baines, T. S. (2007). Improving supply chain performance through improved visibility. *The International Journal of Logistics Management*, *18*(2), 294–313. http://dx.doi.org/10.1108/09574090710816986.

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, *35*(1), 170–180. http://dx.doi.org/10.1016/j.ijforecast.2018.09.003.

Bronselaer, A. (2021). Data quality management: An overview of methods and challenges. In T. Andreasen, G. De Tré, J. Kacprzyk, H. Legind Larsen, G. Bordogna, & S. Zadrożny (Eds.), *Flexible query answering systems* (pp. 127–141). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-86967-0_10.

Brun, A., Karaosman, H., & Barresi, T. (2020). Supply chain collaboration for transparency. *Sustainability*, *12*(11), 4429. http://dx.doi.org/10.3390/su12114429.

Busse, C., Schleper, M. C., Weilenmann, J., & Wagner, S. M. (2017). Extending the supply chain visibility boundary: Utilizing stakeholders for identifying supply chain sustainability risks. *International Journal of Physical Distribution and Logistics Management*, *47*(1), 18–40. http://dx.doi.org/10.1108/IJPDLM-02-2015-0043.

Calatayud, A., Mangan, J., & Christopher, M. (2019). The self-thinking supply chain. *Supply Chain Management: An International Journal*, *24*(1), 22–38. http://dx.doi.org/10.1108/SCM-03-2018-0136.

Caridi, M., Crippa, L., Perego, A., Sianesi, A., & Tumino, A. (2010). Measuring visibility to improve supply chain performance: A quantitative approach. *Benchmarking: An International Journal*, *17*(4), 593–615. http://dx.doi.org/10.1108/14635771011060602.

Caridi, M., Perego, A., & Tumino, A. (2013). Measuring supply chain visibility in the apparel industry. *Benchmarking: An International Journal*, *20*(1), 25–44. http://dx.doi.org/10.1108/14635771311299470.

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. http://dx.doi.org/10.1109/ACCESS.2019.2899751.

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, *5*, Article 850611. http://dx.doi.org/10.3389/fdata.2022.850611.

Fan, W., & Geerts, F. (2012). *Foundations of data quality management* (3rd ed.). Switzerland: Springer Nature, http://dx.doi.org/10.1007/978-3-031-01892-3.

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd ed.). CA, USA: Sage Publications.

Francis, V. (2008). Supply chain visibility: lost in translation? *Supply Chain Management: An International Journal*, *13*(3), 180–184. http://dx.doi.org/10.1108/13598540810871226.

Gao, J., Xie, C., & Tao, C. (2016). Big data validation and quality assurance–Issuses, challenges, and needs. In *Symposium on service-oriented system engineering* (pp. 433–441). Oxford, UK: IEEE, http://dx.doi.org/10.1109/SOSE.2016.63.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191–209. http://dx.doi.org/10.1016/j.jsis.2017.07.003.

Hashemi, L., Huang, E., & Shelley, L. (2022). Counterfeit ppe: Substandard respirators and their entry into supply chains in major cities. *Urban Crime. An International Journal*, *3*(2), 74–109. http://dx.doi.org/10.26250/heal.panteion.uc.v3i2.290.

Hashemi, L., Jeng, C. C., Mohiuddin, A., Huang, E., & Shelley, L. (2023). Simulating counterfeit personal protective equipment (PPE) supply chains during COVID-19. In B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, & P. Lendermann (Eds.), *Proceedings of the 2022 winter simulation conference* (pp. 522–532). Singapore: Institute of Electrical and Electronics Engineers, Inc, http://dx.doi.org/10.1109/WSC57314.2022.10015398.

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72–80. http://dx.doi.org/10.1016/j.ijpe.2014.04.018.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality*, *9*(2), 1–32. http://dx.doi.org/10.1145/3148238.

Huang, Y. (2013). *Automated simulation model generation* (Doctoral thesis), Delft University of Technology, http://dx.doi.org/10.4233/uuid:dab2b000-eba3-42ee-8eab-b4840f711e37.

Ippolito, M., Gregoretti, C., Cortegiani, A., & Iozzo, P. (2020). Counterfeit filtering facepiece respirators are posing an additional risk to health care workers during covid-19 pandemic. *American Journal of Infection Control*, *48*(7), 853. http://dx.doi.org/10.1016/j.ajic.2020.04.020.

Jacobs, P. H. M. (2005). *The DSOL simulation suite* (Doctoral Thesis), Delft University of Technology, http://dx.doi.org/10.4233/uuid:4c5586e2-85a8-4e02-9b50-7c6311ed1278.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338–345. http://dx.doi.org/10.1016/j.jbusres.2016.08.007.

Jeble, S., Dubey, R., Childe, S. J., Papadopoulos, T., Roubaud, D., & Prakash, A. (2018). Impact of big data and predictive analytics capability on supply chain sustainability. *The International Journal of Logistics Management*, *29*(2), 513–538. http://dx.doi.org/10.1108/IJLM-05-2017-0134.

Junaid, M., Zhang, Q., Cao, M., & Luqman, A. (2023). Nexus between technology enabled supply chain dynamic capabilities, integration, resilience, and sustainable performance: An empirical examination of healthcare organizations. *Technological Forecasting and Social Change*, *196*, Article 122828. http://dx.doi.org/10.1016/j.techfore.2023.122828.

Kalaiarasan, R., Agrawal, T. K., Olhager, J., Wiktorsson, M., & Hauge, J. B. (2023). Supply chain visibility for improving inbound logistics: A design science approach. *International Journal of Production Research*, *61*(15), 5228–5243. http://dx.doi.org/10.1080/00207543.2022.2099321.

Kalaiarasan, R., Olhager, J., Agrawal, T. K., & Wiktorsson, M. (2022). The abcde of supply chain visibility: A systematic literature review and framework. *International Journal of Production Economics*, *248*, Article 108464. http://dx.doi.org/10.1016/j.ijpe.2022.108464.

Khondoker, M., Dobson, R., Skirrow, C., Simmons, A., & Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical Methods in Medical Research*, *25*(5), 1804–1823. http://dx.doi.org/10.1177/0962280213502437.

Kuipers, L. (2021). *Increasing supply chain visibility with limited data availability: data assimilation in discrete event simulation* (M.Sc. Thesis), Delft University of Technology, https://resolver.tudelft.nl/uuid:5f68b82f-205e-4509-9a64-22082c46065f.

Kumar, D., Singh, R. K., Mishra, R., & Wamba, S. F. (2022). Applications of the internet of things for optimizing warehousing and logistics operations: A systematic literature review and future research directions. *Computers & Industrial Engineering*, *171*, Article 108455. http://dx.doi.org/10.1016/j.cie.2022.108455.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: Classifying poor data. In G. Wang, T. Tsuchiya, & D. Xiang (Eds.), *2015 IEEE 21st Pacific Rim international symposium on dependable computing* (pp. 179–188). Zhangjiajie, China: IEEE, http://dx.doi.org/10.1109/PRDC.2015.41.

Lavastre, O., Gunasekaran, A., & Spalanzani, A. (2014). Effect of firm characteristics, supplier relationships and techniques used on supply chain risk management (SCRM): an empirical investigation on french industrial firms. *International Journal of Production Research*, *52*(11), 3381–3403. http://dx.doi.org/10.1080/00207543.2013.878057.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*(4), 546–558. http://dx.doi.org/10.1287/mnsc.43.4.546.

Lee, Y., & Rim, S.-C. (2016). Quantitative model for supply chain visibility: Process capability perspective. *Mathematical Problems in Engineering*, *2016*, http://dx.doi.org/10.1155/2016/4049174.

Li, Y., Zobel, C. W., Seref, O., & Chatfield, D. (2020). Network characteristics and supply chain resilience under conditions of risk propagation. *International Journal of Production Economics*, *223*, Article 107529. http://dx.doi.org/10.1016/j.ijpe.2019.107529.

McCrea, B. (2005). EMS completes the visibility picture. *Logistics Management*, *44*(6), 57–61.

Min, H., & Zhou, G. (2002). Supply chain modeling: past, present and future. *Computers & Industrial Engineering*, *43*(1–2), 231–249. http://dx.doi.org/10.1016/S0360-8352(02)00066-9.

Munir, M., Jajja, M. S. S., Chatha, K. A., & Farooq, S. (2020). Supply chain risk management and operational performance: The enabling role of supply chain integration. *International Journal of Production Economics*, *227*, Article 107667. http://dx.doi.org/10.1016/j.ijpe.2020.107667.

Nag, B., Han, C., & Yao, D.-q (2014). Mapping supply chain strategy: an industry analysis. *Journal of Manufacturing Technology Management*, *25*(3), 351–370. http://dx.doi.org/10.1108/JMTM-06-2012-0062.

Nikkei Asia (2020). Vietnam revamps as 'world's mask factory' to offset COVID hit. available at: https://asia.nikkei.com/Economy/Trade/Vietnam-revamps-as-world-s-mask-factory-to-offset-COVID-hit. (accessed on 2022-03-15).

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., et al. (2020). Bias in data-driven artificial intelligence systems — An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), Article e1356. http://dx.doi.org/10.1002/widm.1356.

Oliveira, M. P. V. d., & Handfield, R. (2019). Analytical foundations for development of real-time supply chain capabilities. *International Journal of Production Research*, *57*(5), 1571–1589. http://dx.doi.org/10.1080/00207543.2018.1493240.

Oliveira, P., Rodrigues, F., & Henriques, P. (2005). A formal definition of data quality problems. In F. Naumann, M. Gertz, & S. Madnick (Eds.), *Proceedings of the MIT information quality conference* (pp. 1–14). Cambridge, MA: MIT.

Omar, I. A., Debe, M., Jayaraman, R., Salah, K., Omar, M., & Arshad, J. (2022). Blockchain-based supply chain traceability for COVID-19 personal protective equipment. *Computers & Industrial Engineering*, Article 107995. http://dx.doi.org/10.1016/j.cie.2022.107995.

Peng, J., Hahn, J., & Huang, K.-W. (2023). Handling missing values in information systems research: A review of methods and assumptions. *Information Systems Research*, *34*(1), 5–26. http://dx.doi.org/10.1287/isre.2022.1104.

Pero, M., & Rossi, T. (2014). Rfid technology for increasing visibility in eto supply chains: A case study. *Production Planning and Control*, *25*(11), 892–901. http://dx.doi.org/10.1080/09537287.2013.774257.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. http://dx.doi.org/10.1145/505248.506010.

Rogerson, M., & Parry, G. C. (2020). Blockchain: case studies in food supply chain visibility. *Supply Chain Management: An International Journal*, *25*(5), 601–614. http://dx.doi.org/10.1108/SCM-08-2019-0300.

Roy, V. (2021). Contrasting supply chain traceability and supply chain visibility: are they interchangeable? *The International Journal of Logistics Management*, *32*(3), 942–972. http://dx.doi.org/10.1108/IJLM-05-2020-0214.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. http://dx.doi.org/10.1093/biomet/63.3.581.

Sáez, J. A., Galar, M., Luengo, J., & Herrera, F. (2014). Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition. *Knowledge and Information Systems*, *38*(1), 179–206. http://dx.doi.org/10.1007/s10115-012-0570-1.

Saqib, Z., Saqib, K., & Ou, J. (2019). Role of visibility in supply chain management. In S. A. R. Khan, & S. I. Sümer (Eds.), *Modern perspectives in business applications* (pp. 1–14). IntechOpen, http://dx.doi.org/10.5772/intechopen.87202.

Schoenthaler, R. (2003). Creating real-time supply chain visibility. *Electronic Business*, *29*(8), 12.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Folleco, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*, *259*, 571–595. http://dx.doi.org/10.1016/j.ins.2010.12.016.

Sodhi, M., & Tang, C. (2019). Research opportunities in supply chain transparency. *Production and Operations Management*, *28*(12), 2946–2959. http://dx.doi.org/10.1111/poms.13115.

Somapa, S., Cools, M., & Dullaert, W. (2018). Characterizing supply chain visibility - A literature review. *The International Journal of Logistics Management*, *29*(1), 308–339. http://dx.doi.org/10.1108/IJLM-06-2016-0150.

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. http://dx.doi.org/10.1016/j.procs.2019.09.223.

Srinivasan, R., & Swink, M. (2018). An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. *Production and Operations Management*, *27*(10), 1849–1867. http://dx.doi.org/10.1111/poms.12746.

Stadtler, H., & Kilger, C. (2002). *Vol. 4, Supply chain management and advanced planning: concepts, models, software, and case studies*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Swift, C., Guide, V. D. R., Jr., & Muthulingam, S. (2019). Does supply chain visibility affect operating performance? Evidence from conflict minerals disclosures. *Journal of Operations Management*, *65*(5), 406–429. http://dx.doi.org/10.1002/joom.1021.

Teng, C.-M. (1999). Correcting noisy data. In I. Bratko, & S. Dzeroski (Eds.), *Proceedings of the sixteenth international conference on machine learning* (pp. 239–248). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tiwari, S., Wee, H.-M., & Daryanto, Y. (2018). Big data analytics in supply chain management between 2010 and 2016: Insights to industries. *Computers & Industrial Engineering*, *115*, 319–330. http://dx.doi.org/10.1016/j.cie.2017.11.017.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, *115*(2), c94–c99. http://dx.doi.org/10.1159/000312871.

van Schilt, I. M., Kwakkel, J., Mense, J. P., & Verbraeck, A. (2023). Calibrating simulation models with sparse data: Counterfeit supply chains during covid-19. In B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, & P. Lendermann (Eds.), *Proceedings of the 2022 winter simulation conference* (pp. 496–507). Singapore: Institute of Electrical and Electronics Engineers, Inc, http://dx.doi.org/10.1109/WSC57314.2022.10015241.

van Wee, B., & Banister, D. (2016). How to write a literature review paper? *Transport Reviews*, *36*(2), 278–288. http://dx.doi.org/10.1080/01441647.2015.1065456.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. http://dx.doi.org/10.1016/j.ijpe.2014.12.031.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. http://dx.doi.org/10.1016/j.ijpe.2016.03.014.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. http://dx.doi.org/10.1080/07421222.1996.11518099.

Wang, J., & Zhuo, W. (2020). Strategic information sharing in a supply chain under potential supplier encroachment. *Computers & Industrial Engineering*, *150*, Article 106880. http://dx.doi.org/10.1016/j.cie.2020.106880.

Wei, H.-L., & Wang, E. T. (2010). The strategic value of supply chain visibility: Increasing the ability to reconfigure. *European Journal of Information Systems*, *19*(2), 238–249. http://dx.doi.org/10.1057/ejis.2010.10.

Williams, B. D., Roh, J., Tokar, T., & Swink, M. (2013). Leveraging supply chain visibility for responsiveness: The moderating role of internal integration. *Journal of Operations Management*, *31*(7–8), 543–554. http://dx.doi.org/10.1016/j.jom.2013.09.003.

Zhang, A. N., Goh, M., & Meng, F. (2011). Conceptual modelling for supply chain inventory visibility. *International Journal of Production Economics*, *133*(2), 578–585. http://dx.doi.org/10.1016/j.ijpe.2011.03.003.

Zhao, N., Hong, J., & Lau, K. H. (2023). Impact of supply chain digitalization on supply chain resilience and performance: A multi-mediation model. *International Journal of Production Economics*, *259*, Article 108817. http://dx.doi.org/10.1016/j.ijpe.2023.108817.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, *22*(3), 177–210. http://dx.doi.org/10.1007/s10462-004-0751-8.

Zhu, X., Wu, X., & Yang, Y. (2004). Error detection and impact-sensitive instance ranking in noisy datasets. In *Proceedings of the nineteenth national conference on artificial intelligence* (pp. 378–384). San Jose, CA, USA: American Association for Artificial Intelligence, https://www.aaai.org/Papers/AAAI/2004/AAAI04-061.pdf.