Received: 28 September 2021 | Revised: 10 November 2022 | Accepted: 15 November 2022

DOI: 10.1111/jscm.12294

ORIGINAL ARTICLE

Journal of Supply Chain Management WILEY

Using supply chain databases in academic research: A methodological critique

Giovanna Culot[1] | Matteo Podrecca[1] | Guido Nassimbeni[1] | Guido Orzes[2] | Marco Sartor[1]

[1]Polytechnic Department of Engineering and Architecture, University of Udine, Udine, Italy

[2]Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy

Correspondence
Matteo Podrecca, Polytechnic Department of Engineering and Architecture, University of Udine, Via delle Scienze 206, 33100 Udine, Italy.
Email: podrecca.matteo@spes.uniud.it

Funding information
Regione Autonoma Friuli Venezia Giulia, Grant/Award Number: 89/2019; Autonomous Province of Bozen/Bolzano (Open Access)

Abstract

This article outlines the main methodological implications of using Bloomberg SPLC, FactSet Supply Chain Relationships, and Mergent Supply Chain for academic purposes. These databases provide secondary data on buyer–supplier relationships that have been publicly disclosed. Despite the growing use of these databases in supply chain management (SCM) research, several potential validity and reliability issues have not been systematically and openly addressed. This article thus expounds on challenges of using these databases that are caused by (1) inconsistency between data, SCM constructs, and research questions (*data fit*); (2) errors caused by the databases' classifications and assumptions (*data accuracy*); and (3) limitations due to the inclusion of only publicly disclosed buyer–supplier relationships involving specific focal firms (*data representativeness*). The analysis is based on a review of previous studies using Bloomberg SPLC, FactSet Supply Chain Relationships, and Mergent Supply Chain, publicly available materials, interviews with information service providers, and the direct experience of the authors. Some solutions draw upon established methodological literature on the use of secondary data. The article concludes by providing summary guidelines and urging SCM researchers toward greater methodological transparency when using these databases.

KEYWORDS

Bloomberg, endogeneity, FactSet, Mergent, secondary data, supply chain relationships

INTRODUCTION

Over the last decade, commercial information service providers (ISPs), such as Bloomberg, FactSet, and Mergent, have introduced supply chain databases to support investors in evaluating financial opportunities and risks coming from upstream and downstream partners (Bloomberg, 2011, 2019a; FactSet, 2021a). Increasing scholarly access to these databases has opened the doors to extensive studies in the context of supply chain management (SCM). The availability of these data allows to overcome some limitations related to primary data collection such as perceptual, acquiescence, habituation, and informant biases. Researchers can now analyze large numbers of buyer–supplier dyads and conduct large-scale investigations on supplier/customer bases and extended supply networks (e.g., Bellamy et al., 2020; Dong et al., 2020; Kumar et al., 2019). The relevance of the issues that can be approached using Bloomberg, FactSet, and Mergent supply chain data is testified by the ever-growing number of articles (three in 2014 and 17 in 2021) being published in leading SCM journals.

Although these databases cover an extensive number of buyer–supplier relationships, academic studies that use them, especially those based on statistical inference, might incur nontrivial biases and errors, also because analyses are conducted exclusively on relationships that have been publicly disclosed. As amply documented in the methodological literature on secondary data (Ali et al., 2008; Cook et al., 2012; Cowton, 1998; Houston, 2004; Liu, 2020; Miller et al., 2021), appropriate procedures can be applied to mitigate some shortcomings in data sources (e.g., representativeness with respect to the population). Rigorous application of such procedures is a prerequisite for ensuring comparability and replicability of research that uses these data sources.

The SCM community has not agreed upon best practices regarding how to use these supply chain databases. The aims, contents, and features (e.g., data aggregation, access, and retrieval) of the databases have not been sufficiently clarified by previous studies, potentially leading to overoptimistic expectations. As also highlighted in the *Journal of Supply Chain Management* (Miller et al., 2021), there is a need to establish a common ground on the methods for secondary data analysis. Although some articles outline the key methodological aspects in using secondary data in SCM research, it is important to formulate specific recommendations considering the peculiarities of the various data sources (Bansal et al., 2020; Boyer & Swink, 2008; Calantone & Vickery, 2010; Ellram & Tate, 2016; Miller et al., 2021; Roth et al., 2015).

This article clarifies the key methodological issues when using Bloomberg SPLC, FactSet Supply Chain Relationships, and Mergent Supply Chain in academic research. The characteristics of the three databases are illustrated based on publicly available materials and interviews with the ISPs. Building on a literature review and the authors' first-hand experience, this article discusses potential challenges and solutions when using these databases. Some of the suggested data preparation strategies also draw upon established methodological articles on secondary data analysis and empirical studies using different databases that have similar methodological issues. Overall, this research provides methodological guidelines for future studies that plan to use these databases. However, when using these guidelines, researchers should always assess the relevance of the challenges and the appropriateness of the suggested countermeasures with respect to their research questions and units of analysis.

OVERVIEW OF THE DATABASES

This section illustrates the aims, content, and key characteristics of the supply chain databases mostly used in SCM research: Bloomberg SPLC, FactSet Supply Chain Relationships, and Mergent Supply Chain.[1] This overview is based on publicly available information and on five 45- to 60-min semistructured interviews with analysts and sales representatives.[2] The interview protocol aimed at clarifying the distinctive features and the methods for data collection and analysis of the three ISPs. The interviewees also shared some supplementary materials and details on the data structure and approaches (e.g., classification and estimation procedures) (Bloomberg, 2011, 2019b; FactSet, 2021a, 2021b, 2021c; Mergent, 2014, 2021). Further questions after analyzing the data (e.g., clarifications on the presence of errors) were posed through helpdesk queries.

Aims, content, and features

Major ISPs have recently included supply chain databases on their portals. This allows them to offer financial investors visibility on firms' supply chains for a better understanding of opportunities and risks coming from upstream and downstream partners (Bloomberg, 2011, 2019a; FactSet, 2021a, 2021b; Mergent, 2014, 2021). Bloomberg made the SPLC module available in 2010 with the acquisition of Connexiti, a start-up that offered research and software products on supply chains (Basole

[1] Compustat Customer Segment Files was also used for similar purposes (e.g., Fee & Thomas, 2004; Hertzel et al., 2008). However, it differs in terms of aims, sources, and data availability. Compustat data are limited to firms that are publicly traded in the United States and include only relationships that are mandatorily disclosed according to the Security and Exchange Commission's regulation (i.e., disclosure of any single customer representing more than 10% of annual revenues; SEC, 2018). These relationships—approximately 2000 per year (Agca et al., 2022)—are also included in Bloomberg, FactSet, and Mergent. Further, these have global coverage and monitor also other sources. As a result, they identify at least seven times more buyer–supplier relationships than Compustat (Kumar et al., 2020; Wang et al., 2015). For these reasons, Compustat was not considered here.

[2] The interviews involved at least one interviewee for each ISP. In case the initial respondent was not able to answer to all the questions, a second respondent was involved.

et al., 2018). FactSet Supply Chain Relationships data were introduced in 2013 after the acquisition of Revere Data, a provider of industry classification and supply chain information and analytics (FactSet, 2021b; FactSet, 2013). Mergent launched its supply chain module—formerly Mergent Horizon—in 2014 through a partnership with FactSet (Mergent, 2014). The agreement is still in place today, making the two databases virtually interchangeable, with the exception of the access to relationships that are no longer active (available only in FactSet).

All three databases allow users to search by firm name or ticker (i.e., the unique identifier assigned to each publicly traded stock). Information on direct suppliers, customers, and competitors is provided. Moreover, FactSet/Mergent include the strategic partners linked to the focal firm by investments, joint ventures, research collaborations, and integrated product offerings. Whenever possible, buyer–supplier relationships are classified by type of expense in Bloomberg (i.e., cost of goods sold [COGS]; capital expenditures [CAPEX]; selling, general, and administrative expenses [SG&A]; and research and development costs [R&D]) and by activity in FactSet/Mergent (i.e., distribution, manufacturing, marketing, and licensing). In addition, FactSet/Mergent elaborate a series of keywords related to the specific supply agreement (e.g., the name of the product to which they are referred to). The databases also provide information on the industry of the firms.

A fundamental difference between Bloomberg and FactSet/Mergent is the level of data aggregation. Whereas Bloomberg presents the lists of buyers/suppliers, FactSet/Mergent display contracts/supply agreements labeled as "direct" (those disclosed by the focal firm) or "reverse" relationships (those disclosed by the suppliers/customers of the focal firm). The databases include both relationships for which the monetary value[3] is known and those for which it is not known. In the absence of direct information on monetary value, only Bloomberg provides estimates for it. Data access and retrieval depend on the type of subscription. Overall, FactSet/Mergent enable data download, whereas a standard Bloomberg license does not allow the collection of large amounts of data in bulk. By using the Excel add-in or an application programming interface in Bloomberg, it is only possible to download the data regarding the top 20 customers/suppliers at the extraction date. Should authors require more complete data, the information needs to be manually copied, thus rendering the creation of large-scale datasets a demanding process (Shao et al., 2018; Sharma et al., 2020). The main characteristics of the three databases are summarized in Table 1.

Data sources, classification, and estimation methods

The databases are built on firms' disclosures of contracts/supply agreements made on public filings, annual and quarterly reports, transcripts of conference calls with investors and analysts (e.g., earnings calls), capital markets presentations, sell-side conferences, and firm press releases and websites (Bloomberg, 2019b; FactSet, 2021a; Mergent, 2021). Disclosures are processed by trained analysts into single records to ensure consistency in terms of data aggregation and classification.

A large part of the data is captured from periodic reports. Annual reports are the richest source of information. Public firms release them between 1 and 5 months after the end of their fiscal year (which in many cases does not end on December 31). The annual reports of US-traded firms are usually checked within 2 months of release; the priority is given to large firms. It might take up to 1 year for non-US-traded firms because the format of their reports is not standardized, and they might be written in local languages only. With respect to other data sources (e.g., press releases), they are updated on an ongoing basis. As most of the relationships are disclosed by suppliers, for large US-based firms as well, the data are usually more complete in the second part of the year. It is important, however, to acknowledge that the start date of the relationships reported in the databases is the one when the information is processed.

Bloomberg presents the information at the buyer–supplier relationship level. Each relationship is maintained for 4 years after disclosure, a period considered the average contract length across industries. By contrast, FactSet/Mergent include single announcements about supply agreements that refer to different firms' products and/or subsidiaries. End dates are obtained from the announcements; if this information is not explicitly indicated, the analysts verify over time whether the contract is still in place. As far as the monetary value of the relationship is concerned, Bloomberg is the only source that provides estimates, although just for relationships involving one of the 1200 public firms that were selected based on their market capitalization, data availability, and the interest of the database's users (i.e., the number of searches performed for the firm) (Bloomberg, 2011, 2019b).

In Bloomberg, the estimation approaches are based on either simple calculations (i.e., the monetary value is derived from the revenue percentage reported in the

[3] The monetary value of the relationship describes the purchasing spend (from the buyer side) or sales amount (supplier side) over the period.

TABLE 1 Main characteristics of the databases

	Bloomberg SPLC	FactSet Supply Chain Relationships	Mergent Supply Chain
Inception	2010, acquisition of Connexiti	2013, acquisition of Revere Data	2014, partnership with FactSet
Data availability	Since 2006	Since 2003 for North America, global coverage since 2016 (access to relationships that are no longer active depending on the license)	No access to relationships that are no longer active
Type of relationships	- Customers - Suppliers - Competitors	- Customers - Suppliers - Competitors - Strategic partners	
Subcategories	By type of expense (only for quantified relationships): - COGS - SG&A - R&D - CAPEX	By content of the transaction (only in case there is enough information): - Distribution - Manufacturing - Marketing - Licensing List of keywords (e.g., product name and type of supply)	
Data aggregation	Buyer–supplier relationship	Single supply contract/agreement[a] - Presence of multiple records if announced by both parties involved ("direct" and "reverse" relationships)	
Quantification	- Monetary value and weight (% of costs; % of revenues) - Estimates	- Monetary value and weight (% of revenues), disclosed values only[a] - Relationship relevance (ranking)	- Monetary value and weight (% of revenues), disclosed values only
Data access and retrieval	- Data cannot be downloaded in bulk; need to manually copy data - Option to download the data with the Excel add-in or an API (only for the top 20 customers/ suppliers)	- Data can be downloaded - Option to request data feed/transfer through API	- Data can be downloaded

Abbreviations: API, application programming interface; CAPEX, capital expenditures; COGS, cost of goods sold; R&D, research and development costs; SG&A, selling, general, and administrative expenses.

[a] Depending on the type of license.

Securities and Exchange Commission filing) or more structured algorithms, depending on the level of detail available in the firm disclosures. In the case of the algorithm-based approach, the analysts input the business segment for the supplier's revenues and the relationship type (e.g., COGS). The provisional estimates are checked for consistency with industry-level information and firm disclosures. When the algorithm determines a value considered erroneous by the analyst, it is further adjusted. The role of the analysts is equally important in FactSet, as they define the ranking of the most important suppliers, using both quantitative (e.g., revenue dependency disclosure and stock price correlation) and qualitative information (e.g., disclosure direction and geographic and industry overlap) (FactSet, 2021c).

Furthermore, researchers should be aware that the databases have been subjected to some methodological adjustments over the years in terms of the type of source used (e.g., Bloomberg included conference call transcripts in 2019; there are plans to start monitoring social networks in the near future), languages processed (e.g., Bloomberg discontinued sources in Polish, Greek, and Scandinavian languages in 2013), industries in scope (e.g., from 2020, Bloomberg started monitoring real estate firms and major utility firms in the United States and Japan), and classification procedures (e.g., the first version of FactSet did not specify product-related keywords, and it is, therefore, possible that there is more than one record for the same supply contract when the date ranges overlap).

Data coverage

As of September 2022, Bloomberg includes 490,000 active buyer–supplier relationships, linking together approximately 28,000 public firms (data provided by the Bloomberg helpdesk). FactSet encompasses 270,000 active relationships involving 25,000 public firms (data provided by the FactSet helpdesk). The relationships for which the monetary value is available (i.e., quantified relationships) are 15% in Bloomberg (62% of them are estimates) and 10% in FactSet/Mergent.

To better understand the difference in coverage levels between Bloomberg and FactSet/Mergent, this section presents an analysis of the data for the five largest public firms by total assets in 2020 for each three-digit North American Industry Classification System (NAICS) code. A total of 105 firms in 20 countries were considered for both supplier- and customer-side relationships at three year-ends (2010, 2015, and 2020). A summary of the analysis is presented in Tables 2 and 3. The analysis for each NAICS code is available in Tables S1 and S2.

In both cases, the number of reported relationships has increased since 2010, suggesting higher monitoring efforts after Connexiti and Revere Data were acquired by Bloomberg and FactSet, respectively. On average, the number of reported relationships is substantially higher for Bloomberg, which is also more complete in terms of quantifying the monetary value of the relationship due to the presence of estimates.

To assess the overlap of the information included in the databases, the data regarding the three NAICS codes with the highest number of suppliers in 2020 were compared (i.e., 324 – Petroleum and Coal Products Manufacturing; 334 – Computer and Electronic Product Manufacturing; and 336 – Transportation Equipment Manufacturing). The results show that for these firms, on average, the three databases share 43% of suppliers, whereas Bloomberg includes 40% of unique suppliers and FactSet/Mergent 17%. These differences stem from diverse monitoring efforts, assumptions on relationship duration, and the attribution of relationships to the parent firm/subsidiary.

Within each database, the high variability in the number of customers/suppliers and revenues/costs reported for each focal firm depends on several factors. According to the interviewees at Bloomberg and FactSet, geographical differences are related to monitored languages and disclosure propensity (according to the ISPs, firms in Asia are more reluctant to reveal customer relationships than those based in large Western countries). In terms of industry-specific factors, coverage levels could be affected by the concentration of players in the segment. This is exemplified by the relationships reported for car manufacturers (NAICS 336 – Transportation Equipment Manufacturing). The number of suppliers is comparatively higher and has a lower variance than in other sectors; conversely, the percentage of known revenues is limited, probably because of the dealership model of the automotive industry, which involves several small-

TABLE 2 Supplier side data coverage over time, like-for-like analysis

	Number of suppliers included		Number of suppliers with a quantified relationship included		% known COGS
	Bloomberg	FactSet	Bloomberg	FactSet	Bloomberg
Year	Mean (STD.DEV)	Mean (STD.DEV)	Mean (STD.DEV)	Mean (STD.DEV)	Mean (STD.DEV)
2020	162.32 (197.37)	87.30 (125.42)	44.86 (70.78)	8.25 (14.14)	10.20 (17.18)
2015	161.42 (213.29)	57.94 (87.54)	65.44 (99.04)	5.90 (12.36)	11.46 (17.78)
2010	98.13 (134.52)	22.00 (34.11)	7.14 (11.66)	2.59 (5.50)	3.20 (6.56)

Note: The analysis does not explicitly cover Mergent, which includes the same data as FactSet. FactSet and Mergent do not provide the buyer's dependency on the supplier (FactSet, 2021a).

Abbreviations: COGS, cost of goods sold; STD.DEV, standard deviation.

TABLE 3 Customer side data coverage over time, like-for-like analysis

	Number of customers included		Number of customers with a quantified relationship included		% known revenues	
	Bloomberg	FactSet	Bloomberg	FactSet	Bloomberg	FactSet
Year	Mean (STD. DEV)	Mean (STD. DEV)	Mean (STD. DEV)	Mean (STD. DEV)	Mean (STD. DEV)	Mean (STD. DEV)
2020	89.78 (96.52)	46.35 (68.44)	41.19 (57.87)	0.35 (1.00)	15.07 (19.43)	2.42 (7.91)
2015	94.36 (113.52)	27.57 (37.93)	46.94 (69.53)	0.31 (0.83)	17.33 (20.86)	2.98 (8.79)
2010	52.98 (78.32)	11.27 (16.79)	2.54 (4.15)	0.15 (0.48)	3.93 (9.12)	1.75 (6.31)

Note: The analysis does not explicitly cover Mergent, which includes the same data as FactSet.

Abbreviation: STD.DEV, standard deviation.

and medium-sized enterprises. The popularity of the focal firm also plays a role. This appears clearly in the articles by Kim and Davis (2016) and Sharma et al. (2020), where a statistically significant correlation emerges between the number of known suppliers and the media attention and reputation of the buying firm, respectively.

It is important to keep in mind that these databases are not specifically designed for academic research; no claim is made by the ISPs on how exhaustive and representative the data are. Rather, the providers present them as a collection of disclosed information elaborated based on a series of assumptions, data triangulation methods, proprietary algorithms, and analysts' experience. Under this premise, the next section discusses potential methodological issues of using these databases for SCM research purposes.

SCM STUDIES USING SUPPLY CHAIN DATABASES: TOPICS AND METHODS

An overview of published articles that have used Bloomberg, FactSet, or Mergent supply chain databases is provided to identify the research questions, units of analysis, and variables. Although these databases have been broadly used in finance and engineering (e.g., Agarwal et al., 2017; Bae et al., 2019; Basole, 2009; Basole & Karla, 2011, 2012; Cao et al., 2021; Chen et al., 2021; Gofman et al., 2020; Piraveenan et al., 2019, 2020; Sugrue et al., 2021; Yamamoto et al., 2021), the review presented in this section considers only articles published in SCM journals or those with a clear focus on this discipline.

A total of 56 articles were identified (see Table S3 for a description of the review approach and the full classification of the articles). The use of the databases (i.e., Bloomberg, FactSet, and Mergent) has been gaining momentum over time, from three articles published in 2014 to 17 in 2021, with an average of 6.22 articles per year. Bloomberg is the most common source (36 articles from 2014 to 2021, including those using Connexiti prior to acquisition). 40 out of 56 articles were published in operations and SCM journals, among which 11 articles were published in the *Journal of Operations Management*, six articles in the *International Journal of Production Economics*, and five articles in the *International Journal of Operations & Production Management* and in the *Journal of Business Logistics*. One article was published in the *Journal of Supply Chain Management*.

Most of the articles (35 out of 56) applied regression models to investigate the relationships between some characteristics of the supply chain and operational, environmental, social, or innovation performance. Both direct and moderating effects on firm performance have been investigated (Bellamy et al., 2020; Ben-Jebara & Modi, 2021; Elking et al., 2017; Lu & Shang, 2017). The databases are used either to identify buyer–supplier dyads/triads or the relationships that constitute the network in which a focal firm is embedded, namely, the supplier/customer base and the extended supply network. Additional data (e.g., financial indexes and sustainability scores) are drawn by matching the firms' names with other databases.

In dyadic/triadic settings, the regressors reflect supply chain partners' characteristics, such as size (Kumar et al., 2019); cultural, demographics, and power differences between the buyer and the supplier (e.g., Kumar et al., 2020); and the tenure and strength of the relationship (e.g., Chae et al., 2020). In studies that consider the extended supply network, variables are defined through a network analysis perspective (see Borgatti & Li, 2009, for an overview) to derive both global measures (e.g., clustering and density) and node-level dimensions (e.g., degree and betweenness centrality) (Adhikary et al., 2020; Basole et al., 2018; Falcone et al., 2022; Sharma et al., 2019). Moreover, several articles use variables that involve the number of suppliers/customers related to each focal firm and the sum of all revenues/costs of the identified suppliers/customers (Dong et al., 2020; Schwieterman et al., 2018, 2020; Sharma et al., 2020).

To a lesser extent, the data have been used to build realistic settings in agent-based simulation studies (Basole et al., 2016; Chen et al., 2022; Li et al., 2020, 2021; Li & Zobel, 2020), to understand the impact of external events on supply chain configuration (Son et al., 2021), and to develop analytical and visualization tools for SCM. Examples of these tools are those proposed by Shao et al. (2018) for identifying the most critical suppliers in the extended supply network, Mizgier et al. (2017) for allocating capital, and Basole et al. (2017) for visualizing networks. By contrast, no study has analyzed the origin and evolution of links between firms, although methods have been outlined in the social network analysis literature (Graham, 2020), and the topic has been investigated for joint ventures, alliances, and online exchanges (Carnovale & Yeniyurt, 2014; Dhanorkar et al., 2019; Park et al., 2018). This gap is probably due to the relevant effort required to collect longitudinal data.

Some remarks on the characteristics of the datasets used in these studies are also necessary. Several, but not all, articles explicitly state their focus on publicly listed firms (49 articles out of 56). Regression-based studies sample cross-industry (22 articles) or broadly within the manufacturing sector (11 articles). Industry-specific studies refer to electronics (eight), automotive (seven), and pharmaceutical (one). Among the agent-based modeling contributions, Honda represents the most used case study (four articles). Many articles (22) considered only US-based or US-listed firms.

USING THE DATA IN ACADEMIC RESEARCH: CHALLENGES AND SOLUTIONS

Although these supply chain databases represent a rich source of data for SCM research, caution is needed in using them, as they were developed for a different purpose (i.e., supporting investors in evaluating opportunities and risks coming from upstream and downstream partners). In this section, possible challenges and solutions are identified. These were drawn from SCM studies using supply chain databases and from other articles dealing with similar issues, albeit with different databases.

The section is organized around three core aspects: *data fit*, *data accuracy*, and *data representativeness*. These are aligned with current methodological debates on the academic use of secondary data (Ali et al., 2008; Kahle & Walkling, 1996; Liu, 2020; Roth et al., 2015; Stewart & Kamins, 1993). To this end, potential peculiarities related to the different research questions and units of analysis of previous research are highlighted. Overall, these aspects ought to be assessed in a specific research context and in relation to the phenomenon of interest. Correcting all potential errors might not always be possible for a particular research project, and yet, researchers are always responsible for fully disclosing limitations in the datasets and issues related to using the chosen databases.

Data fit

When approaching new data sources, it is always important to consider the potential validity and reliability implications with respect to the phenomenon that researchers intend to investigate (Miller et al., 2021). The fit between the data, research questions, and supply chain constructs should be carefully assessed to avoid considering *relationships that are not relevant for the research purpose* (see Table 4).

The databases provide a classification of relationships (by *type of expense* in Bloomberg, such as COGS, CAPEX, and R&D, and according to the *nature of the transaction* in FactSet/Mergent, such as manufacturing and marketing), but this is limited to cases for which sufficient information is available. Moreover, these classifications might not reflect the supply chain construct under investigation and, thus, serve as an improper proxy.

For example, researchers might be interested in relationships that refer to the "traditional" understanding of the physical supply chain—that is, a set of nodes where transformation activities take place and of links that represent physical movements of products between these nodes (Carter et al., 2015; Choi & Hong, 2002). In this case, logistics services and equipment providers, as well as R&D collaborations, should be excluded. None of the reviewed studies that used FactSet/Mergent selected a specific relationship category. The common approach adopted with Bloomberg data for focusing on "traditional" physical supply chains is to select all COGS relationships and exclude SG&A, CAPEX, and R&D (e.g., Wetzel & Hofmann, 2019). Although this choice is intended to factor out relationships that are not directly related to the manufacturing process, it should be noted that relationships classified as COGS also include service providers, thus potentially leading to misinterpretations.

In general, the three databases examined in this study do not provide visibility of multitier product flows, making it difficult to ascertain whether a supplier of a focal firm's first-tier supplier is also a second-tier supplier of the same focal firm. This should be considered in studies focusing on multiple supply chain tiers. For example, the chip manufacturer Intel is linked (as a COGS relationship) with the firm FedEx, which provides Intel with logistics services; FedEx is in turn linked (as a COGS relationship) with airports (e.g., Aéroports de Paris) and airlines (e.g., Lufthansa). Without an appropriate cleaning process, the data would show all these players as part of Intel's extended supply network. The presence of industrial conglomerates operating across different business segments (e.g., Merck Group, 3M, and Henkel) makes it even more difficult to identify multitier relationships. For instance, the science and technology firm Merck Group has three business units: Healthcare, Life Science, and Performance Materials. The Performance Materials unit is typically a first-tier supplier of semiconductor manufacturers, and Healthcare produces and commercializes drugs and medical treatments. The suppliers of this business unit are often other pharmaceutical firms that are unrelated to the semiconductor supply chain.

Overall, these examples indicate that researchers should always reflect on how the relationships sampled from the databases relate to their research questions and the unit of analysis. As a preliminary step, relationship categories should be considered. For instance, Elking et al. (2017) and Adhikary et al. (2020) focused only on COGS relationships, assuming that these suppliers were the most involved in firms' day-to-day activities and, therefore, the most relevant for the investigated

TABLE 4 Potential challenges related to data fit

Challenge	Description	Implications by unit of analysis	Possible solutions and references (if any)
Relationships not relevant for the research purpose	Potential inconsistencies between the included buyer–supplier relationships and the supply chain construct under investigation Difficult identification of product flows between firms (not available in Bloomberg, limited to keyword search in FactSet/Mergent)	Dyads; triads: Sampling not in scope with the study. Errors in the identification of dyads/triads (e.g., in the case of a firm operating in two or more different business segments, suppliers in one business segment might be unrelated to customers in the other segment) Supplier/customer base; extended supply network: Measures affected by the inclusion of firms that are not part of the supplier/customer base or network under investigation. This issue becomes severe in the case of industrial conglomerates among first-tier suppliers	**Limit data extraction (in the main analysis or robustness tests) to** (1) the categories that are of primary interest of the study (e.g., for Bloomberg SPLC, several articles use only COGS relationships) (Adhikary et al., 2020; Gualandris et al., 2021; Li et al., 2021; Li & Zobel, 2020; Wetzel & Hofmann, 2019) **In case researchers are interested in the flow of goods along the supply chain (or in the different business segments the firm operates in), clean the data, insert control variables, and run robustness tests based on** (2) the exclusion of relationships with firms operating in industries not involved in material/product flows (e.g., financial services) (Wang et al., 2021); (3) the comparison with industry-level flows (e.g., TiVA and BEA) (Chen et al., 2016; Kim et al., 2020); (4) the identification of suppliers/customers with a considerable amount of business in the industry of the focal firm (e.g., at least 50% of revenues in the industry) (DeCampos et al., 2022); (5) the number and relevance of the business segments of the focal firm (Lu & Shang, 2017); or (6) the identification of industrial conglomerates (Chen et al., 2016; Gofman & Wu, 2022; Kim & Swink, 2021). *Overall, it is difficult to ensure that a supplier of a focal firm's first-tier supplier is also a second-tier supplier of the focal firm. Despite possible countermeasures, researchers should be cautious in using the data to study material flows in multitier settings or—at a minimum—acknowledge this as a relevant limitation.*

Abbreviations: BEA, US Bureau of Economic Analysis; TiVA, OECD Trade in Value Added.

constructs (i.e., financial dependence and sustainability performance, respectively). Shao et al. (2018) included all relationship categories in their analyses of suppliers' risk.

Besides the use of the classifications provided by the ISPs, there are further opportunities to select relationships based on the industry of the supplier. For example, studies on innovation-related topics might focus on suppliers cooperating effectively in product/process development. In this respect, Potter and Paulraj (2020) excluded logistics service providers from their sample, as they are not usually associated with an intensive patenting activity. Similarly, Wang et al. (2021) removed suppliers belonging to nonmanufacturing industries, such as financial service providers, to capture material flows. Furthermore, whenever researchers investigate firms operating in the same business environment, it seems reasonable to select only suppliers that have a considerable share of revenues in the industry of the focal firm (DeCampos et al., 2022).

An approach adopted by studies focusing on material flows relies on the triangulation of the extracted relationships with industry-level information. Based on aggregated input and output flows between industries, which are available through the Organization for Economic Co-operation and Development (OECD) Trade in Value Added/STructural ANalysis or the US Bureau of Economic Analysis databases, the initial sample can be cleaned by checking the presence of material trades between the industry of the focal firm and those of the reported customers/suppliers (Chen et al., 2016; Kim et al., 2020). Controlling for the effects of industrial conglomerates might further strengthen the analysis. Kim and Swink (2021) inserted, for instance, a dummy variable for firms with more than one business segment. Chen et al. (2016) ran a robustness test based on firms operating in only one business segment. Similarly, Lu and Shang (2017) operationalized supplier base complexity by considering the number and relevance of the business segments of the focal firm.

Although the procedures described above support a more accurate identification of the relationships of interest, a precise reconstruction of multitier material flows cannot be done through the three databases, as they do not indicate whether a supplier of a focal firm's first-tier supplier is also a second-tier supplier of the focal firm (i.e., provides products or services relevant for the focal firm). Any assumption in this sense should therefore be openly stated, and the relevant limitations should be properly acknowledged.

Data accuracy

Issues related to estimated values and manual tabulation have been amply documented for several commercial databases (Cook et al., 2012; Elton et al., 2001; Liu, 2020). A common approach to assess the presence of errors, their magnitude, and possible systematic biases is to perform validation studies, namely, the comparison between the values reported in the database and actual primary data for a sample of entries (Bound et al., 2001; Schennach, 2020). Access to firms' internal records is, however, limited, hindering such validation studies from being conducted (Kim & Davis, 2016; Wang et al., 2015). Articles should therefore rely on

- a set of indirect procedures to assess and/or mitigate the abovementioned issues;
- a clear and transparent explanation of data cleaning and data analysis procedures to allow replication; and
- a detailed discussion of the limitations and generalizability of the findings.

Based on the authors' experience with the data and the review of previous SCM studies, two sets of potential areas of concern are particularly relevant (see Table 5). The first is related to *inaccuracies in classification, approximation, and estimation* encompassing both random and systematic errors. Random errors occur mainly due to oversight in data tabulation. The impacts of these random errors in large-scale statistical analyses are usually rather limited (Craighead et al., 2011; Schwab, 2013). Examples include erroneous classifications of buyer–supplier relationships (e.g., in Bloomberg, Leonardo was linked as COGS with FedEx for the provision of aircrafts but should have been considered a CAPEX relationship) and mistakes due to similar firm names (e.g., the carmaker Mitsubishi was confused in one database with the unrelated producer of stationery items Mitsubishi Pencil and thus appeared in the supplier lists of Tesco and Metro). Besides the comparison of information included in the different sources (Bloomberg vs. FactSet/Mergent), there are opportunities to check the data through visual inspections by looking through industry codes for consistency. Potential doubts should then be cleared with the ISP's helpdesk and the original source (if available).

By contrast, systematic errors can cause more relevant issues, leading to upward or downward biases in the results (i.e., overestimation or underestimation, respectively). These errors are caused by the assumptions and procedures used to enter public disclosures into database records. One of the main issues in this respect is the lack of consistency in the level of attribution of the relationship (e.g., parent, subsidiary, or operating unit). For example, in Bloomberg and FactSet/Mergent, the relationships involving private subsidiaries are normally reported as such, whereas they are attributed (i.e., rolled-up) to the ultimate parent whenever this is a publicly

TABLE 5 Potential challenges related to data accuracy

Challenges	**Description**	**Implications by unit of analysis**	**Possible solutions and references (if any)**
Inaccuracies in classification, approximation, and estimation	Human errors in data tabulation (e.g., firm name and classification of the relationship) Inconsistencies in how disclosures are processed into single records: - attribution of the relationship to the parent firm/private subsidiary/operating unit (e.g., rolling-up of private subsidiaries to the ultimate publicly listed parent); - attribution of the relationship to the focal firm/contract manufacturer (e.g., double counting of the cost of the component as the supplier reports a direct relationship with the focal manufacturer); and - attribution of relationships in the case of mergers and acquisitions	Dyads; triads: Inclusion of nonexisting relationships and miscalculation of the characteristics of the relationship Supplier/customer base; extended supply network: Measures affected by systematic bias (especially in industries with a high presence of financial holdings and contract manufacturers)	**Identify and correct random errors in the datasets by** (1) visual inspection for industry consistency, check with the original source; and (2) comparing data from multiple databases **For systematic errors, clean the data, insert control variables, and run robustness tests based on** (3) the identification of firms potentially affected by the issue (e.g., financial holdings, contract manufacturers, firms operating in industries that resort massively to contract manufacturing, firms whose reported COGS exceed 100%, firms affected by mergers and acquisitions) (Chellappa & Saraf, 2010)
Time inconsistencies	Short-run variances in the data determined also by disclosure lags and assumptions regarding the length of the relationships (e.g., 4-year standard period in Bloomberg) In the long run, variances in the data also explained by a change in country, source, and language coverage	Dyads; triads: Exclusion of relevant relationships not reported in the specific period for which data are collected Supplier/customer base; extended supply network: Variation of measures over time affected by incomplete data rather than actual changes (especially in studies investigating the top *N* suppliers/customers, or focusing on quantified relationships only)	**Assess the impact of time inconsistency by** (1) collecting and combining multiple rounds of data to analyze the difference between extracted relationships/calculated metrics at different times (Kim & Davis, 2016; Gualandris et al., 2021); (2) replicating the analysis with data for different time periods; (3) building subsamples of active/inactive relationships (Agca et al., 2022); or (4) comparing data from multiple databases **Run the main analysis/robustness tests stabilizing the data for the period of interest (year/quarter) by** (5) considering rolling periods (i.e., all suppliers that appear in one given period or the previous one) (Dong et al., 2020; Osadchiy et al., 2021); (6) building a time-invariant dataset (i.e., pooling together all buyer–supplier relations for the period/averaging the monetary value of the relationship) (Dong et al., 2020); (7) including only relationships that are stable over a long time horizon (number of years depending on the industry) (Hofer et al., 2021; Kim & Swink, 2021); or (8) considering relationships active if the time gap between two consecutive reported relationships is no longer than a certain threshold (e.g., 6 months/1 year) (Agca et al., 2022; Gofman & Wu, 2022) *Potential countermeasures mainly address time inconsistencies in the short run. The effects of changes in coverage methodology can hardly be mitigated. Longitudinal studies—especially over many years—should acknowledge this among relevant limitations.*

Abbreviation: COGS, cost of goods sold.

listed firm. Especially in the case of financial holdings, this results in an imprecise characterization of the buyer–supplier relationships as well as in inaccurate metrics of the supplier base and the extended supply network. Similarly, data from firms involved in mergers and acquisitions have potential problems. FactSet/Mergent transfer only the "reverse" relationships (i.e., those that are not disclosed by the focal firm) to the acquirer or the new entity. In Bloomberg, the relationships of the two original firms are transferred to the new entity only in the case of mergers, but not for acquisitions.

Further issues relate to the presence of contract manufacturers and distributors. Component suppliers (e.g., Samsung and Intel) might report a direct relationship with a focal manufacturer (Apple), whereas they actually sell to its contract manufacturers (e.g., Foxconn) or distributors. The same cost is thus reported twice (double counting)—the first time for the cost of the component and the second time for the total assembly cost (e.g., in 2021, the monetary value of Apple's COGS relationships reported by Bloomberg accounts for 167% of the firm's COGS). SCM studies that do not acknowledge this would miscalculate the strength of dyadic relationships, erroneously place assemblers and component suppliers at the same level of the supply chain, and improperly calculate the structural characteristics of the supplier base/extended supply network. The issue is particularly relevant in sectors where contract manufacturers and distributors are more common (e.g., Computer and Electronic Product Manufacturing – Apple and Retail Trade – Walmart).

In many cases, inaccuracies in classification, approximation, and estimation can be addressed through data cleaning procedures. Potentially problematic data entries can be screened by filtering the dataset by industry of the suppliers/customers—for example, selecting relationships involving financial holdings and accessing the original disclosures to check whether they refer to a private subsidiary rather than the ultimate public parent firm. Semi-automatic methods for scanning the data are not applicable, as no algorithms or precise guidelines are available to identify and solve these issues; the experience and judgment of the researcher are needed. Data comparisons among different databases can further help identify misreported relationships. Once detected, it is possible to use control variables (e.g., specific dummies) or remove the firms/relationships with potentially more affected data from the sample (e.g., specific industries and firms involved in mergers and acquisitions) (Chellappa & Saraf, 2010). This choice depends on the number of observations falling into these categories and, thus, implications in terms of generalizability.

Another set of potential challenges relates to *time inconsistencies*. Researchers should be aware that the databases do not offer a real-time picture of active relationships. As illustrated in Section 2.2, the relationships are mostly disclosed in firms' annual reports, and processing time varies depending on the kind of firm (large firms have priority) and country (US- vs. non-US-listed firms). Other sources (e.g., press releases) are processed on an ongoing basis. Therefore, the data extracted for a period include relationships that correspond to the same year up to 2 years before (i.e., the year referenced by the annual report). Researchers conducting analyses that must unequivocally capture a specific period of time (e.g., event studies that assess the impact of external events on the characteristics of the supply chain) should acknowledge this. Otherwise, this does not represent a major issue under the premise of a relative stability of the supply chain structure over time (Gualandris et al., 2021; Kim & Davis, 2016). In regression-based studies, it is, however, important to check that the time overlap between the supply chain data and the dependent variable is not a cause of simultaneity (i.e., the explanatory variable is jointly determined with the dependent variable; Wooldridge, 2015).

Moreover, given that the ISPs have partially changed their data collection methods over the years (e.g., which sources and firms are considered in the data collection), longitudinal studies might be subject to distortions, especially those spanning over long periods (for which changes in the data collection methods are more likely). As far as cross-sectional studies are concerned, researchers should be aware that over contiguous time periods, some buyer–supplier relationships might appear and disappear from the records.

These "holes" in the databases may be explained by the disclosure lags (i.e., the time between the agreement and its disclosure) and the assumptions made by the ISPs with respect to the length of the relationships (e.g., Bloomberg posits a 4-year standard duration of a contract). The contractual agreements may not have changed. By the same token, studies that use Bloomberg's data limited to relationships for which the monetary value is available should recognize that there might not be enough information to run estimates for some years (e.g., for 2014 and 2016, the database contains a quantified COGS relationship between Kubota Corporation and Kitagawa Corp., whereas no information on the type and the monetary value of such relationship is reported for 2015).

Only a few of the reviewed studies explicitly mentioned the potential time inconsistencies described above and proposed solutions, despite the serious implications such inconsistencies can have for calculating network-level variables. In dyadic/triadic settings, there is, moreover, a risk of disregarding relationships that are not reported in the specific period for which data are extracted (Orenstein, 2021). In general, it is appropriate to collect data from multiple time periods from the databases and assess the stability of the variables/results (Gualandris et al., 2021; Kim & Davis, 2016). Whenever the structure of the firms' network substantially changes over time, it is advisable to perform additional news searches or consider the support material included in the databases. Another approach is to verify possible changes in the results using subsamples of active and inactive relationships (such as the article of Agca et al., 2022, on shock propagation along the supply chain). Moreover, data comparisons among Bloomberg and FactSet/Mergent might further help to identify temporal inconsistencies.

There are opportunities to stabilize the characteristics of datasets over time. For instance

- using rolling periods (i.e., considering all suppliers/customers that appear in the previous period or in the current period as suppliers/customers for the current period; Dong et al., 2020; Osadchiy et al., 2021);
- building a time-invariant dataset (i.e., a pooled dataset in which a supplier/customer that appeared in one period is also included in all other periods; Dong et al., 2020);
- focusing only on stable relationships (i.e., only relationships that are reported in all the considered periods; Hofer et al., 2021; Kim & Swink, 2021); or
- assuming continuity between reported relationships in the short term (i.e., combining multiple relationships between the same supplier–customer pair over different time periods into a single continuous relationship if the time gap between two relationships is shorter than a certain time; Agca et al., 2022; Gofman & Wu, 2022).

Instead of performing these procedures up front, researchers can also apply them as robustness tests. The appropriateness of these solutions depends mainly on the focus of the study. For instance, it can be assumed that managerial practices (e.g., inventory leanness) are more likely to spread through strong ties, thus prompting researchers to select only stable relationships, as done by Hofer et al. (2021). The main issue of data stabilization is, however, a certain arbitrariness in how start dates, thresholds, and rolling periods are defined. To avoid this, it is possible to use industry-specific information concerning the average length of supply contracts and alternative temporal cutoffs in robustness tests (Wang et al., 2021; Zhong et al., 2021).

Data representativeness

Representative data are important for the validity and reliability of research results, a prerequisite often not given when using nonrandom samples (Ali et al., 2008; Banz & Breen, 1986; Barnes et al., 2014; Keil, 2017). Systematic data omission might, in fact, give rise to two different issues: *bias in the selection of focal firms* and *bias in the relationships reported for each focal firm* (see Table 6).

A *bias in the selection of focal firms* potentially occurs when firms with available data represent only a subset of the population of interest, as explained by Lu and Shang (2017) and Sharma et al. (2019). In fact, the ISPs focus mainly on publicly traded (and usually large) firms; others are included only if they report a relationship with one of those. Although the average coverage levels of the databases (i.e., the number of included firms and the relationships available for each of them) appear to be substantial, the ISPs do not pretend their data to be exhaustive or representative. This represents a clear limitation to the generalizability of the results, especially for nonpublicly listed firms. Endogeneity problems might arise whenever a variable affecting the inclusion of firms in the database is correlated with the dependent variable of the study (Antonakis et al., 2010; Certo et al., 2016). Further issues may appear when data drawn from supply chain databases are combined with those from other databases (e.g., ESG, financial, or patent databases) as this further reduces the number of firms with available data (e.g., Sharma et al., 2019).

A carefully developed research design and an appropriate definition of the sampling criteria should be a priority. Private firms are not in scope with the ISPs monitoring efforts and thus should not be sampled. Further sample selection criteria need to account for the size, geography, and reputation of the firms. For example, Gualandris et al. (2021) performed a stratified random sampling among Forbes 2000 firms; Chedid et al. (2021) focused on firms with a market capitalization higher than US $6 billion. Despite some drawbacks in terms of generalizability, these strategies strongly reflect the nature of the databases.

Further, different methods can be used to control for bias due to nonrandom sampling. Among these, the Heckman (1976, 1979) two-stage estimation approach is one of the most used. Although not extensively, the Heckman approach has also been applied in studies using the three databases analyzed in this article. It builds on identifying one or more variables (called exclusion restrictions; Certo et al., 2016) that can explain the inclusion of the firms in the dataset but are uncorrelated with the regressand. For instance, Lu and Shang (2017) verified whether there was a bias resulting from firms with missing data in Mergent. They used the number of years since incorporation as exclusion restriction, assuming that this variable may influence firms' inclusion in

TABLE 6 Potential challenges related to data representativeness

Challenges	Description	Implications by unit of analysis	Possible solutions and references
Bias in the selection of focal firms	The companies included in the databases might be a nonrandom subset of the population of interest Further limitations arise if the databases are used in combination with other sources of firm-level data due to record matching	Dyads; triads; supplier/customer base; extended supply network: Generalizability of the results and potential endogeneity issues if the criteria for a company inclusion are correlated with the dependent variable	**Define sampling criteria that reflect the nature of the database by** (1) focusing on categories that are broadly represented in the databases, such as large public companies, firms headquartered in the large Western countries, highly reputed firms (e.g., Fortune 500, Global Fortune 500, Forbes 2000) (Chedid et al., 2021; Gualandris et al., 2021) **Control and correct for bias due to nonrandom sampling through** (2) the use of approaches such as Heckman two-stage estimation (Lu & Shang, 2017; Sharma et al., 2019, 2020)
Bias in the relationships reported for each focal firm	Buyer–supplier relationships available in the databases are only those that have been disclosed There might be biases due to different propensity to disclose related to the home country, the industry, and the reputation of the company	Dyads; triads: Generalizability of the results, potential endogeneity issues if the data coverage is correlated with the dependent variable Supplier/customer base; extended supply network: Error in the calculation of supplier/customer base measures, asymptotic bias, and potential endogeneity issues if the data coverage is correlated with the dependent variable	**Define sampling criteria that reflect data coverage by** (1) focusing on firms that have higher data availability, such as large public companies, firms headquartered in large Western countries, highly reputed firms (Chedid et al., 2021; Gualandris et al., 2021); and (2) considering only the top *N* suppliers/customers in order to increase the comparability of firms with different data availability (Schwieterman et al., 2018; Wetzel & Hofmann, 2019) **Clean the data/remove outliers based on** (3) the identification of focal firms with limited data availability (i.e., number of reported relationships, percentage of known revenues/costs, and percentage of first-tier suppliers with available data on second-tier relationships) (Gualandris et al., 2021; Kim & Davis, 2016) **Assess the presence of potential issues, introduce fixed/random effects or control variables, and run robustness tests based on** (4) the identification of focal firms with limited data availability (i.e., number of reported relationships, percentage of known revenues/costs, percentage of first-tier suppliers/customers with available data on second-tier relationships) (Kashiwagi et al., 2021; Schwieterman et al., 2020; Wang et al., 2021); and (5) variables that affect firms' disclosure propensity or the database monitoring efforts (e.g., industry, size, ownership, geography, and reputation of the focal firm) (Adhikary et al., 2020; Dai et al., 2021; Lu & Shang, 2017; Wang et al., 2021) **Moreover, it is possible to further validate the results through** (6) the adoption of complementary approaches (DeCampos et al., 2022); (7) the use of alternative measures/data (e.g., supply chain measures built from focal companies' financial statements) (Lam, 2018); or (8) the analysis of aggregated data (e.g., industry) (Bae et al., 2019; Gofman et al., 2020; Osadchiy et al., 2016, 2021)

the database but not the dependent variable of their study (i.e., the firm's financial performance).

The risk of *bias in the relationships reported for each focal firm* comes, for the most part, from the inclusion of mainly voluntarily disclosed relationships. As illustrated in Section 2.3, there is high variability in the number of reported suppliers/customers and the percentage of known costs/revenues available for each focal firm. Against this backdrop, considerations concerning generalizability and potential endogeneity issues are needed. Most crucially, a possible correlation may exist between the firm-level data coverage (e.g., the total percentage of a customer's known spend on its direct suppliers or the total percentage of a supplier's known revenues from its customers) and the dependent variable. Such a correlation would have a significant impact on studies that could be subject to desirability bias, such as those dealing with sustainability, social responsibility, or risk. In fact, an intrinsic endogeneity bias can occur if firms decide not to publicize ties with other supply chain members because of their practices, performance, or reputation, as noted by Gualandris et al. (2021) about suppliers' sustainability and transparency.

The implications of this are more critical in multitier settings. For example, Sharma et al. (2019) found that only 628 firms out of a random sample of 2000 firms from Bloomberg reported data for their second-tier suppliers and customers. A possible reason for this is the limited availability of data for private firms. Whenever first-tier suppliers are not publicly listed, there is a risk that whole portions of the extended supply network will not be included. This can create substantial distortions in the description of the topological structure of the network (Orenstein, 2021). The measures built for each focal firm may therefore be subject to errors, giving rise to asymptotic bias if not properly addressed (i.e., the estimator does not converge to the true value of the parameter) (Guide Jr & Ketokivi, 2015).

The implications of incomplete and potentially biased reporting of buyer–supplier relationships should first be addressed through a carefully developed research design and an appropriate definition of the sampling criteria, such as ownership, size, industry, and geography. For instance, Wang et al. (2021) found smaller coefficients and less significant estimates when repeating their analysis for samples consisting of only (a) non-US/non-European and (b) small firms. Comparisons between firms characterized by different coverage levels should thus be avoided. As suggested for the bias in the selection of focal firms, the first step is to focus on samples that are characterized by high coverage in the databases, such as large public firms (Chedid et al., 2021) or firms included in the Forbes 2000 list (Gualandris et al., 2021).

In studies investigating the supplier/customer base, it is also possible to consider only the top *N* buyers/suppliers (in terms of the monetary value of the relationship) in the main analysis or in robustness tests. This limits potential biases related to the number and share of reported relationships (Schwieterman et al., 2018; Wetzel & Hofmann, 2019), also taking into account that manufacturers tend to source most of their volumes from a few strategic suppliers (Håkansson et al., 2010; Varadarajan et al., 2001). Although this top *N* buyer/supplier approach is suitable for studies investigating financial performance (e.g., Return on Asset - ROA) or manufacturing practices, it might be more problematic when dealing with supply chain risk or resilience-related topics (as it is often not suppliers with the highest purchasing volume that put focal firms at the highest risk; Simchi-Levi et al., 2015). The same logic can be applied when defining the sampling criteria for dyads, including only suppliers/buyers representing a given share of the focal firm's costs/revenues (Fee & Thomas, 2004; Hertzel et al., 2008; Mackelprang & Malhotra, 2015).

The effects of sampling biases and missing nodes/links are becoming increasingly central in the social network analysis literature due to their significant implications for model specification and inference; however, effective approaches to address these issues are still underdeveloped (Crane, 2018; Graham, 2020; Smith & Moody, 2013; Smith et al., 2017). Caution is needed in addressing potential disclosure biases through instrumental variables (Ketokivi & McIntosh, 2017; Lu & Shang, 2017), an approach that can be problematic in case of a correlation between the errors in calculated parameters and the other variables in the model (Bound et al., 2001; Schennach, 2020). Moreover, a theoretical justification for the exogeneity of the instrumental variable is often difficult in the analysis of supply chains (Hsieh & Lee, 2016; Lu & Shang, 2017).

The approach adopted in previous studies consists of identifying focal firms with limited data availability (e.g., Gualandris et al., 2021; Kim & Davis, 2016; Schwieterman et al., 2020). Thanks to the presence of a relevant number of relationships for which the monetary value is available, in Bloomberg, it is also possible to classify firm-level data coverage based on the share of known revenues/costs (Mackelprang & Malhotra, 2015). Furthermore, more precise coverage levels can be inferred by combining firms' COGS data with the average purchasing spend of the industry (Hoberg et al., 2017). Once identified, firms with limited data availability can be

- removed from the final sample. This can be done, for instance, for firms with less than five suppliers (Gualandris et al., 2021; Kim & Davis, 2016) and those with only unquantified relationships (Gualandris et al., 2021). Gualandris et al. (2021) also ascertained through a one-way analysis of variance that the values of the dependent variable did not significantly differ between excluded firms and the final sample. This approach could be used in studies in which there is a limited number of entities characterized by low coverage levels (e.g., number of available suppliers); or

- handled through control variables in the main analysis or in further robustness tests, as done by Wang et al. (2021) with respect to the number of first- and second-tier suppliers and by Kashiwagi et al. (2021) and Schwieterman et al. (2020) for the number of customers and suppliers. This alternative might be adopted in studies in which the number of affected entities is higher.

In general, a potential correlation between the coverage levels of the sample (e.g., percentage of known COGS/revenues and number of first-tier suppliers with unquantified second-tier relationships) and the dependent variables needs to be tested. If a significant correlation is found, the direction and extent of the bias can be assessed by running two separate regressions for subsamples split around the median coverage level. This approach was applied, for instance, by Gualandris et al. (2021), who found that their main results were downward biased (i.e., underestimated) by the presence of firms characterized by low data coverage.

Regarding the factors related to different disclosure propensity and monitoring efforts (i.e., firm/industry/country characteristics), some studies included specific controls. These may affect both the dependent variable under investigation and the predictors (Adhikary et al., 2020; Lu & Shang, 2017; Wang et al., 2021). Some scholars (e.g., Agarwal et al., 2017; Dai et al., 2021; Wang et al., 2021; Yamamoto et al., 2021) have also performed robustness tests to verify whether the main results hold when analyzing subsamples segmented by size, geography, or industry.

The application of the approaches presented above should always be critically challenged in the light of the research context (Bettis et al., 2014; Guide Jr & Ketokivi, 2015; Ketokivi & McIntosh, 2017). Researchers can also strengthen their results by adopting complementary approaches, such as combining quantitative analyses using supply chain databases with case studies or expert interviews (e.g., DeCampos et al., 2022). Moreover, as shown by Lam (2018), there is a strong tradition in SCM research of using indicators derived from focal firms' financial statements as proxies for supply chain characteristics. These could be used in robustness tests as alternative measures to those retrieved from supply chain databases. Finally, some articles test their hypotheses at different levels of analysis, for example, leveraging industry data from input–output tables (Bae et al., 2019; Gofman et al., 2020; Osadchiy et al., 2016, 2021). Whether these approaches can be applied depends on the research questions. Potential limitations due to data coverage should, at any rate, be fully reported.

SUMMARY GUIDELINES AND ILLUSTRATIVE EXAMPLE

Against the research opportunities provided by large-scale supply chain databases, researchers should weigh the relevant limitations of the data and be aware of how data are collected and processed by the ISPs (i.e., scope, classification, approximation, and estimation). The severity of the issues and the appropriateness of corrective approaches depend on the unit of analysis and the topics under investigation.

This section presents a general summary of the key steps to consider throughout the entire research process[4] (Figure 1). For the sake of clarity, these steps are also applied to an illustrative example based on a simplified version of the article by Elking et al. (2017) on the relationship between focal firm dependence on suppliers and financial performance (with the supplier base as the unit of analysis). This example has been selected because it is focused on a research topic that is known to the *Journal of Supply Chain Management*'s readership and allows a step-by-step explication of data handling procedures and their impact. To increase transparency, the example is also reported in full in Table S4.

Step 1: Evaluating the potential and drawbacks of the databases

The first step starts with a deep understanding of the databases' characteristics and methodological approaches. It is important to compare the pros and cons of the solutions provided by the different ISPs in terms of what information is made available (e.g., estimates on the value of the relationships, classifications, and keywords) and the efforts needed for data retrieval. In building the illustrative example, the authors selected Bloomberg SPLC, as it provides the largest number of buyer–supplier relationships for which the monetary value is available, which are needed to calculate dependence measures.

[4] These guidelines exclusively address the implications of using the three databases (i.e., Bloomberg, Mergent, and FactSet), excluding other, more general methodological issues that may arise when investigating supply chains, such as those outlined in Serpa and Krishnan (2018), Bray et al. (2019), and Mukandwal et al. (2020).

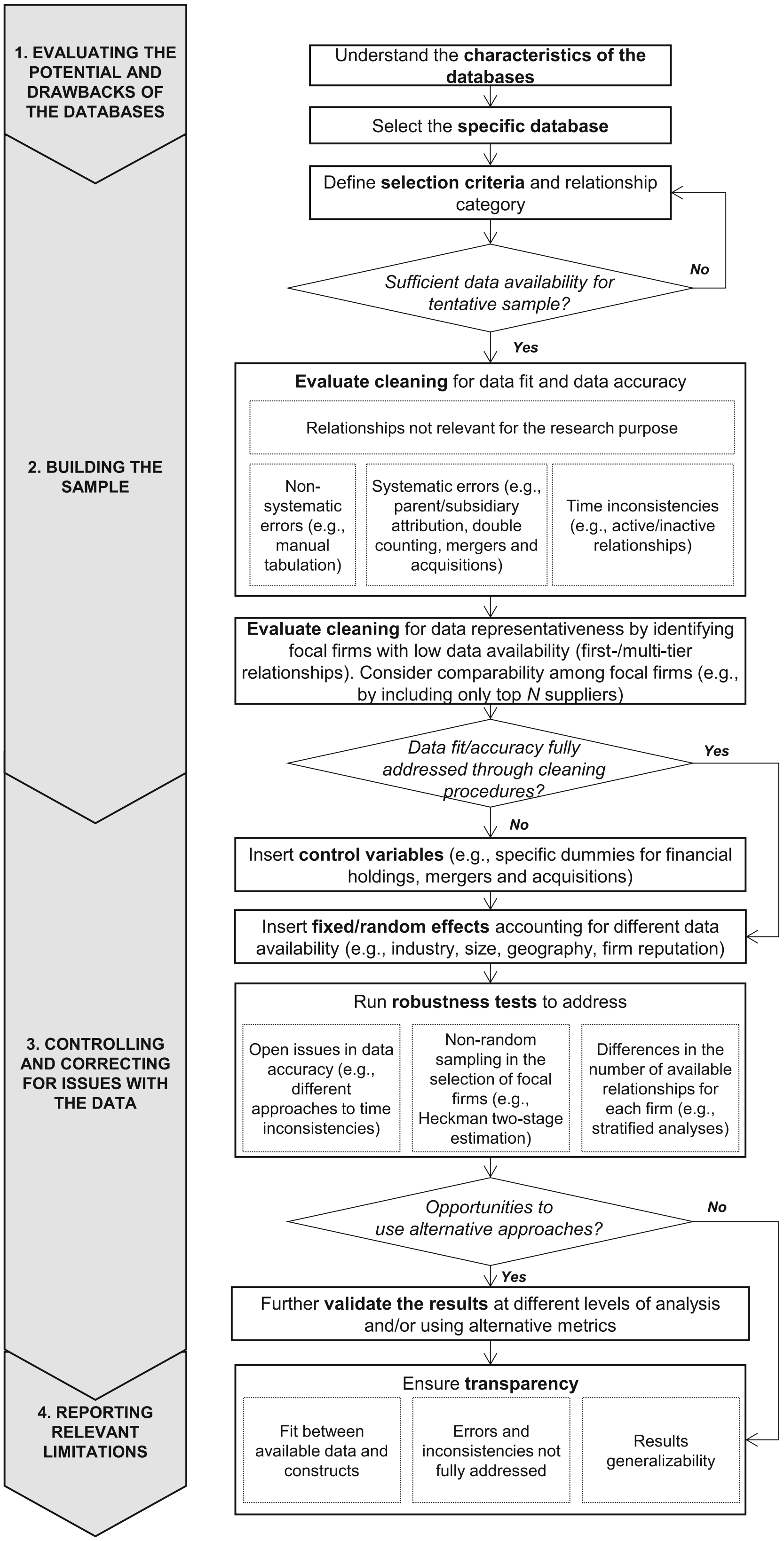
1. EVALUATING THE POTENTIAL AND DRAWBACKS OF THE DATABASES
2. BUILDING THE SAMPLE
3. CONTROLLING AND CORRECTING FOR ISSUES WITH THE DATA
4. REPORTING RELEVANT LIMITATIONS
Understand the characteristics of the databases
Select the specific database
Define selection criteria and relationship category
Sufficient data availability for tentative sample?
No
Yes
Evaluate cleaning for data fit and data accuracy
Relationships not relevant for the research purpose
Non-systematic errors (e.g., manual tabulation)
Systematic errors (e.g., parent/subsidiary attribution, double counting, mergers and acquisitions)
Time inconsistencies (e.g., active/inactive relationships)
Evaluate cleaning for data representativeness by identifying focal firms with low data availability (first-/multi-tier relationships). Consider comparability among focal firms (e.g., by including only top N suppliers)
Data fit/accuracy fully addressed through cleaning procedures?
Yes
No
Insert control variables (e.g., specific dummies for financial holdings, mergers and acquisitions)
Insert fixed/random effects accounting for different data availability (e.g., industry, size, geography, firm reputation)
Run robustness tests to address
Open issues in data accuracy (e.g., different approaches to time inconsistencies)
Non-random sampling in the selection of focal firms (e.g., Heckman two-stage estimation)
Differences in the number of available relationships for each firm (e.g., stratified analyses)
Opportunities to use alternative approaches?
No
Yes
Further validate the results at different levels of analysis and/or using alternative metrics
Ensure transparency
Fit between available data and constructs
Errors and inconsistencies not fully addressed
Results generalizability

FIGURE 1 Key steps

Step 2: Building the sample

Several limitations of the databases can already be addressed while building the sample. Selection criteria should, in fact, consider the level of data availability, which is higher for large public firms and firms based in large Western countries. For each focal firm, it is also important to select those buyer–supplier relationships that are relevant to the study. This can be done using the categories provided by the databases or based on industry classifications. In the illustrative example, only the top five firms by total assets in each three-digit NAICS manufacturing industry were sampled. COGS relationships were selected as the most stable and relevant in daily business activities (Adhikary et al., 2020; Elking et al., 2017; Gualandris et al., 2021). The resulting sample consisted of 92 firms with, on average, 50.85 COGS suppliers each. The average size of the supplier base was aligned with previous studies focusing on large firms (e.g., Chedid et al., 2021). An initial ordinary least squares regression was performed on the sample without the application of any corrective measures. The results showed only a very weak association between dependence on suppliers and firm performance ($\beta = -0.950$, $SE = 0.639$, $p = 0.1405$[5]).

As a next step, data cleaning procedures should address nonsystematic and systematic errors, including those related to time inconsistencies. Moreover, firms with low data availability should be identified and properly handled. The data cleaning procedures adopted in the illustrative example address the type of relationship (consistency between focal firms' and suppliers' industries), the presence of rolled-up relationships (i.e., relationships involving a private subsidiary, which are attributed to the ultimate public parent), contract manufacturers, mergers and acquisitions, and the stability of buyer–supplier relationships over time. Moreover, suppliers accounting for less than 0.01% of focal firms' COGS were removed to increase the comparability of firms with different coverage levels. As a result of these steps, the sample contained 87 (focal) firms, with an average of 39.39 COGS suppliers each.

Step 3: Controlling and correcting for issues with the data

The main analysis and robustness tests should be designed to control for and correct issues with the data. In addition to the data cleaning steps described above, a dummy variable for firms with fewer than three suppliers was included. The (negative) association between firm dependence on suppliers and ROA became stronger ($\beta = -1.377, SE = 0.631, p = 0.0319$). Other control variables were then introduced in the model to account for firm, industry, and country specificities, and this further strengthened the results; the coefficient size increased, whereas the standard error decreased ($\beta = -1.607$, $SE = 0.493$, $p = 0.0017$).

Consistent with Lu and Shang (2017), the Heckman two-stage estimation approach was applied to check whether there was a selection bias in the sample of focal firms ($\beta = -1.554$, $SE = 0.507$, $p = 0.0030$). Further robustness tests were performed by restricting the analyses to European, North American, and Japanese firms only (i.e., contexts characterized by high coverage) and removing focal firms for which the monetary value of known COGS relationships accounted for less than 1% of the total COGS. In both cases, the association between firm dependence on suppliers and ROA was stronger than in the main analysis ($\beta = -2.028$, $SE = 0.513$, $p = 0.0002$; $\beta = -2.049$, $SE = 0.577$, $p = 0.0008$). In line with Gualandris et al. (2021), a potential correlation between firm-level data availability and the dependent variable was also verified.

Step 4: Reporting relevant limitations

Finally, studies based on supply chain databases must openly and transparently report their relevant limitations. In particular, researchers should reflect upon the fit between the available data and the investigated supply chain constructs, explicitly mention any errors or inconsistencies that have not been fully addressed, and discuss any limitations related to the generalizability of the results.

CONCLUSIONS

SCM research is increasingly relying on secondary data (Ellram & Tate, 2016; Miller et al., 2021). However, the growing use of supply chain databases in academic studies has not been matched by an open discussion about their shortcomings and corrective approaches to deal with them. This article presents the main features of the three most widely used supply chain databases (i.e., Bloomberg, Mergent, and FactSet) and provides an overview of their use in SCM research. It then outlines a set of potential challenges related to data fit, accuracy, and representativeness and highlights potential solutions to address

[5] As recommended by the American Statistical Association, cutoff levels of statistical significance are not used. Therefore, this study reports only exact *p*-values and interprets them appropriately in the text (for more details in this regard, see Wasserstein & Lazar, 2016).

them. Finally, the article illustrates a sequence of key steps that can guide researchers in using these databases and provides an illustrative example of their use.

This effort leads to four closing considerations for scholars using secondary data in SCM. First, it is important to invest time in understanding why the databases were developed, how the data are collected, and how the original sources are classified and estimations are made. This results in a research design that is more consistent with the available data. Second, researchers should assess how data can be matched with the supply chain constructs under investigation. For example, supply chain databases should not be used to identify multitier material flows as it is difficult to ensure that a supplier of a focal firm's first-tier supplier is also a second-tier supplier of the focal firm. Third, the factual credibility of the data should not be taken for granted; any inconsistency must be checked with the original source and properly handled. Even more caution is needed in longitudinal analyses; researchers should consider any change over time of coverage and monitoring efforts that might influence what information is included and how it is presented. Fourth, the limitations of the samples in terms of representativeness must be acknowledged and addressed through proper statistical procedures. These limitations might not be fully overcome, but knowledge of the research context can support a better interpretation of the results.

In general, supply chain databases open unprecedented opportunities. The data they contain are indeed rich and extensive and allow to overcome some relevant issues of primary research, such as perceptual, acquiescence, habituation, and informant biases. However, researchers should also be aware of their limitations, which must be transparently reported and explicitly addressed. In this regard, this article echoes previous calls (e.g., Miller et al., 2021; Roth et al., 2015) for greater transparency on the characteristics of the samples and the approaches adopted for data preparation and analysis. Researchers using supply chain databases could also share the lists of focal firms included in the analyses, the data cleaning steps and scripts (e.g., Python, Stata, or R codes), or—subject to the authorization of ISPs—even the analyzed datasets. The disclosure of this information can indeed facilitate the comparability and replicability of studies and, ultimately, ensure the advancement of academic knowledge.

Future methodological research could then extend the approaches and methods presented in this article, including methods that rely on text mining, machine learning, and artificial intelligence. Moreover, despite the difficulties in executing them, validation studies could provide additional insights into the potential presence and extent of systematic biases in the databases.

ACKNOWLEDGMENTS

The authors are very grateful for the insightful guidance provided by the whole editorial team including the four anonymous reviewers. The authors also thank Jury Gualandris (Ivey Business School, Canada), Ruggero Bellio (University of Udine, Italy), Ilaria Giannoccaro (Polytechnic University of Bari, Italy), and Albachiara Boffelli (University of Bergamo, Italy) for their advice and support on the preliminary versions of the manuscript. This work was supported by the Regione Autonoma Friuli Venezia Giulia within the specific program 89/2019 – Fondo Sociale Europeo 2014/2020. The authors thank the Department of Innovation, Research University and Museums of the Autonomous Province of Bozen/Bolzano for covering the Open Access publication costs.

ORCID

Giovanna Culot https://orcid.org/0000-0003-1402-1622
Matteo Podrecca https://orcid.org/0000-0003-1130-8759
Guido Nassimbeni https://orcid.org/0000-0002-1532-2980
Guido Orzes https://orcid.org/0000-0001-7740-3885
Marco Sartor https://orcid.org/0000-0001-9286-1382

REFERENCES

Adhikary, A., Sharma, A., Diatha, K. S., & Jayaram, J. (2020). Impact of buyer-supplier network complexity on firms' greenhouse gas (GHG) emissions: An empirical investigation. *International Journal of Production Economics*, *230*, 107864. https://doi.org/10.1016/j.ijpe.2020.107864

Agarwal, A., Leung, A. C. M., Konana, P., & Kumar, A. (2017). Cosearch attention and stock return predictability in supply chains. *Information Systems Research*, *28*(2), 265–288. https://doi.org/10.1287/isre.2016.0656

Agca, S., Babich, V., Birge, J. R., & Wu, J. (2022). Credit shock propagation along supply chains: Evidence from the CDS market. *Management Science*, *68*(9), 6506–6538. https://doi.org/10.1287/mnsc.2021.4174

Ali, A., Klasa, S., & Yeung, E. (2008). The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *The Review of Financial Studies*, *22*(10), 3839–3871. https://doi.org/10.1093/rfs/hhn103

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Bae, J. W., Elkamhi, R., & Simutin, M. (2019). The best of both worlds: Accessing emerging economies via developed markets. *The Journal of Finance*, *74*(5), 2579–2617. https://doi.org/10.1111/jofi.12817

Bansal, P., Gualandris, J., & Kim, N. (2020). Theorizing supply chains with qualitative big data and topic modeling. *Journal of Supply Chain Management*, *56*(2), 7–18. https://doi.org/10.1111/jscm.12224

Banz, R. W., & Breen, W. J. (1986). Sample-dependent results using accounting and market data: Some evidence. *The Journal of*

Finance, *41*(4), 779–793. https://doi.org/10.1111/j.1540-6261.1986.tb04548.x

Barnes, B. G., Harp, N. L., & Oler, D. (2014). Evaluating the SDC mergers and acquisitions database. *Financial Review*, *49*(4), 793–822. https://doi.org/10.1111/fire.12057

Basole, R. C. (2009). Visualization of interfirm relations in a converging mobile ecosystem. *Journal of Information Technology*, *24*(2), 144–159. https://doi.org/10.1057/jit.2008.34

Basole, R. C. (2016). Topological analysis and visualization of interfirm collaboration networks in the electronics industry. *Decision Support Systems*, *83*, 22–31. https://doi.org/10.1016/j.dss.2015.12.005

Basole, R. C., & Bellamy, M. A. (2014). Visual analysis of supply network risks: Insights from the electronics industry. *Decision Support Systems*, *67*, 109–120. https://doi.org/10.1016/j.dss.2014.08.008

Basole, R. C., Bellamy, M. A., & Park, H. (2017). Visualization of innovation in global supply chain networks. *Decision Sciences*, *48*(2), 288–306. https://doi.org/10.1111/deci.12213

Basole, R. C., Bellamy, M. A., Park, H., & Putrevu, J. (2016). Computational analysis and visualization of global supply network risks. *IEEE Transactions on Industrial Informatics*, *12*(3), 1206–1213. https://doi.org/10.1109/TII.2016.2549268

Basole, R. C., Ghosh, S., & Hora, M. S. (2018). Supply network structure and firm performance: Evidence from the electronics industry. *IEEE Transactions on Engineering Management*, *65*(1), 141–154. https://doi.org/10.1109/TEM.2017.2758319

Basole, R. C., & Karla, J. (2011). On the evolution of mobile platform ecosystem structure and strategy. *Business & Information Systems Engineering*, *3*(5), 313–322. https://doi.org/10.1007/s12599-011-0174-4

Basole, R. C., & Karla, J. (2012). Value transformation in the mobile service ecosystem: A study of app store emergence and growth. *Service Science*, *4*(1), 24–41. https://doi.org/10.1287/serv.1120.0004

Bellamy, M. A., Dhanorkar, S., & Subramanian, R. (2020). Administrative environmental innovations, supply network structure, and environmental disclosure. *Journal of Operations Management*, *66*(7–8), 895–932. https://doi.org/10.1002/joom.1114

Bellamy, M. A., Ghosh, S., & Hora, M. (2014). The influence of supply network structure on firm innovation. *Journal of Operations Management*, *32*(6), 357–373. https://doi.org/10.1016/j.jom.2014.06.004

Ben-Jebara, M., & Modi, S. (2021). Product personalization and firm performance: An empirical analysis of the pharmaceutical industry. *Journal of Operations Management*, *67*(1), 82–104. https://doi.org/10.1002/joom.1109

Bettis, R., Gambardella, A., Helfat, C., & Mitchell, W. (2014). Quantitative empirical analysis in strategic management. *Strategic Management Journal*, *35*, 949–953. https://doi.org/10.1002/smj.2278

Bloomberg, L. P. (2011). *Supply chain relationships (SPLC) database*. Bloomberg.

Bloomberg LP (2019a). How to analyse coronavirus-related risks. Retrieved from https://www.bloomberg.com/professional/blog/researching-supply-chain-exposures-how-to-analyse-coronavirus-related-risks/ [Accessed 20.09.21]

Bloomberg, L. P. (2019b). *Estimating supply chain relationships: The Bloomberg way*. Bloomberg.

Borgatti, S. P., & Li, X. (2009). On social network analysis in a supply chain context. *Journal of Supply Chain Management*, *45*(2), 5–22. https://doi.org/10.1111/j.1745-493X.2009.03166.x

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. Heckman & E. Leamer (Eds.). *Handbook of econometrics*. Elsevier. https://doi.org/10.1016/S1573-4412(01)05012-7

Boyer, K. K., & Swink, M. L. (2008). Empirical elephants—Why multiple methods are essential to quality research in operations and supply chain management. *Journal of Operations Management*, *26*, 337–348. https://doi.org/10.1016/j.jom.2008.03.001

Bray, R. L., Serpa, J. C., & Colak, A. (2019). Supply chain proximity and product quality. *Management Science*, *65*(9), 4079–4099. https://doi.org/10.1287/mnsc.2018.3161

Brintrup, A., Wang, Y., & Tiwari, A. (2017). Supply networks as complex systems: A network-science-based characterization. *IEEE Systems Journal*, *11*(4), 2170–2181. https://doi.org/10.1109/JSYST.2015.2425137

Calantone, R. J., & Vickery, S. K. (2010). Introduction to the special topic forum: Using archival and secondary data sources in supply chain management research. *Journal of Supply Chain Management*, *46*(4), 3–11. https://doi.org/10.1111/j.1745-493X.2010.03202.x

Cao, S. S., Fang, V. W., & Lei, L. G. (2021). Negative peer disclosure. *Journal of Financial Economics*, *140*(3), 815–837. https://doi.org/10.1016/j.jfineco.2021.02.007

Carnovale, S., & Yeniyurt, S. (2014). The role of ego networks in manufacturing joint venture formations. *Journal of Supply Chain Management*, *50*(2), 1–17. https://doi.org/10.1111/jscm.12015

Carter, C. R., Rogers, D. S., & Choi, T. Y. (2015). Toward the theory of the supply chain. *Journal of Supply Chain Management*, *51*(2), 89–97. https://doi.org/10.1111/jscm.12073

Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639–2657. https://doi.org/10.1002/smj.2475

Chae, S., Yan, T., & Yang, Y. (2020). Supplier innovation value from a buyer–supplier structural equivalence view: Evidence from the PACE awards in the automotive industry. *Journal of Operations Management*, *66*(7–8), 820–838. https://doi.org/10.1002/joom.1063

Chedid, F., Kocabasoglu-Hillmer, C., & Ries, J. M. (2021). The interaction between supply networks and internal networks: Performance implications. *International Journal of Operations & Production Management*, *41*(6), 860–881. https://doi.org/10.1108/IJOPM-10-2020-0710

Chellappa, R. K., & Saraf, N. (2010). Alliances, rivalry, and firm performance in enterprise systems software markets: A social network approach. *Information Systems Research*, *21*(4), 849–871. https://doi.org/10.1287/isre.1090.0278

Chen, H. A., Karim, K., & Tao, A. (2021). The effect of suppliers' corporate social responsibility concerns on customers' stock price crash risk. *Advances in Accounting*, *52*, 100516. https://doi.org/10.1016/j.adiac.2021.100516

Chen, K., Li, Y., & Linderman, K. (2022). Supply network resilience learning: An exploratory data analytics study. *Decision Sciences*, *53*(1), 8–27.

Chen, L., Zhang, G., & Zhang, W. (2016). Return predictability in the corporate bond market along the supply chain. *Journal of Financial Markets*, *29*, 66–86. https://doi.org/10.1016/j.finmar.2016.03.005

Choi, T. Y., & Hong, Y. (2002). Unveiling the structure of supply networks: Case studies in Honda, Acura, and DaimlerChrysler. *Journal of Operations Management*, *20*(5), 469–493. https://doi.org/10.1016/S0272-6963(02)00025-6

Cook, R. G., Campbell, D. K., & Kelly, C. (2012). An issue of trust: Are commercial databases really reliable? *Journal of Business & Finance Librarianship*, *17*(4), 300–312. https://doi.org/10.1080/08963568.2012.712501

Cowton, C. J. (1998). The use of secondary data in business ethics research. *Journal of Business Ethics*, *17*(4), 423–434. https://doi.org/10.1023/A:1005730825103

Craighead, C. W., Ketchen, D. J., Dunn, K. S., & Hult, G. T. M. (2011). Addressing common method variance: Guidelines for survey research on information technology, operations, and supply chain management. *IEEE Transactions on Engineering Management*, *58*(3), 578–588. https://doi.org/10.1109/TEM.2011.2136437

Crane, H. (2018). *Probabilistic foundations of statistical network analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315209661

Dai, R., Liang, H., & Ng, L. (2021). Socially responsible corporate customers. *Journal of Financial Economics*, *142*(2), 598–626. https://doi.org/10.1016/j.jfineco.2020.01.003

DeCampos, H. A., Rosales, C. R., & Narayanan, S. (2022). Supply chain horizontal complexity and the moderating impact of inventory turns: A study of the automotive component industry. *International Journal of Production Economics*, *245*, 108377. https://doi.org/10.1016/j.ijpe.2021.108377

Dhanorkar, S., Kim, Y., & Linderman, K. (2019). An empirical investigation of transaction dynamics in online surplus networks: A complex adaptive system perspective. *Journal of Operations Management*, *65*(2), 160–189. https://doi.org/10.1002/joom.1006

Ding, L., Lam, H. K., Cheng, T. C. E., & Zhou, H. (2021). The contagion and competitive effects across national borders: Evidence from the 2016 Kumamoto earthquakes. *International Journal of Production Economics*, *235*, 108115. https://doi.org/10.1016/j.ijpe.2021.108115

Ding, W., Levine, R., Lin, C., & Xie, W. (2021). Corporate immunity to the COVID-19 pandemic. *Journal of Financial Economics*, *141*(2), 802–830. https://doi.org/10.1016/j.jfineco.2021.03.005

Dong, Y., Skowronski, K., Song, S., Venkataraman, S. & Zou, F. (2020). Supply base innovation and firm financial performance. *Journal of Operations Management*, *66*(7–8), 768–796, 768, https://doi.org/10.1002/joom.1107

Elking, I., Paraskevas, J.-P., Grimm, C., Corsi, T., & Steven, A. (2017). Financial dependence, lean inventory strategy, and firm performance. *Journal of Supply Chain Management*, *53*(2), 22–38. https://doi.org/10.1111/jscm.12136

Ellinger, A. E., Adams, F. G., Franke, G. R., Herrin, G. D., deCoster, T. E., & Filips, K. E. (2020). A triadic longitudinal assessment of multiple supply chain participants' performance and the extended enterprise concept. *International Journal of Physical Distribution & Logistics Management*, *50*(7–8), 745–767. https://doi.org/10.1108/IJPDLM-07-2019-0209

Ellram, L. M., & Tate, W. L. (2016). The use of secondary data in purchasing and supply management (P/SM) research. *Journal of Purchasing and Supply Management*, *22*(4), 250–254. https://doi.org/10.1016/j.pursup.2016.08.005

Elton, E. J., Gruber, M. J., & Blake, C. R. (2001). A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases. *The Journal of Finance*, *56*(6), 2415–2430. https://doi.org/10.1111/0022-1082.00410

FactSet (2013). FactSet acquires Revere Data. Retrieved from https://www.globenewswire.com/news-release/2013/09/05/571631/10047372/en/FactSet-Acquires-Revere-Data.html. [Accessed 23.03.22].

FactSet. (2021a). Supply chain relationships data and methodology guide. Retrieved from https://open.factset.com/api/public/media/download/products/documents/9723b7f3-64a1-430c-9391-0017efa51d45?_gl=1%2A1d5d4u8%2A_ga%2AMzMzNzkzODkuMTY0NDQ5OTUxMA..%2A_ga_2Q3PTT96M8%2AMTY0Njg5NzE4NC4xMy4wLjE2NDY4OTcxODcuMA. [Accessed 14.03.22].

FactSet. (2021b). Standard datafeed user guide: Supply chain relationships V1.4.2. Retrieved from https://open.factset.com/api/public//media/download/products/documents/badec2f3-623f-415e-9277-21fcf2e06d9f?_gl=1*msjum7*_ga*MzMzNzkzODkuMTY0NDQ5OTUxMA..*_ga_2Q3PTT96M8*MTY0Njg5NzE4NC4xMy4wLjE2NDY4OTcxODcuMA. [Accessed 14.03.22].

FactSet. (2021c). Supply chain report. Retrieved from https://my.apps.factset.com/oa/pages/17447 [Accessed 14.03.22].

Falcone, E. C., Fugate, B. S., & Dobrzykowski, D. D. (2022). Supply chain plasticity during a global disruption: Effects of CEO and supply chain networks on operational repurposing. *Journal of Business Logistics*, *43*(1), 116–139. https://doi.org/10.1111/jbl.12291

Fee, C. E., & Thomas, S. (2004). Sources of gains in horizontal mergers: Evidence from customer, supplier, and rival firms. *Journal of Financial Economics*, *74*(3), 423–460. https://doi.org/10.1016/j.jfineco.2003.10.002

Gofman, M., Segal, G., & Wu, Y. (2020). Production networks and stock returns: The role of vertical creative destruction. *The Review of Financial Studies*, *33*(12), 5856–5905. https://doi.org/10.1093/rfs/hhaa034

Gofman, M., & Wu, Y. (2022). Trade credit and profitability in production networks. *Journal of Financial Economics*, *143*(1), 593–618. https://doi.org/10.1016/j.jfineco.2021.05.054

Graham, B. S. (2020). Network data. In S. N. Durlauf, L. P. Hansen, J. Heckman, & R. L. Matzkin (Eds.). *Handbook of econometrics*. Elsevier. https://doi.org/10.1016/bs.hoe.2020.05.001

Gualandris, J., Longoni, A., Luzzini, D., & Pagell, M. (2021). The association between supply chain structure and transparency: A large-scale empirical study. *Journal of Operations Management*, *67*(7), 803–827. https://doi.org/10.1002/joom.1150

Guide Jr, V. D. R., & Ketokivi, M. (2015). Notes from the editors: Redefining some methodological criteria for the journal. *Journal of Operations Management*, *37*, v-viii. https://doi.org/10.1016/S0272-6963(15)00056-X

Håkansson, H., Kraus, K., & Lind, J. (2010). *Accounting in networks*. Routledge. https://doi.org/10.4324/9780203854310

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, *47*(1), 153–161. https://doi.org/10.2307/1912352

Hertzel, M. G., Li, Z., Officer, M. S., & Rodgers, K. J. (2008). Inter-firm linkages and the wealth effects of financial distress along the supply chain. *Journal of Financial Economics*, *87*(2), 374–387. https://doi.org/10.1016/j.jfineco.2007.01.005

Hoberg, K., Protopappa-Sieke, M., & Steinker, S. (2017). How do financial constraints and financing costs affect inventories? An empirical supply chain perspective. *International Journal of Physical Distribution & Logistics Management*, *47*(6), 516–535. https://doi.org/10.1108/IJPDLM-05-2016-0142

Hofer, C., Barker, J., & Eroglu, C. (2021). Interorganizational imitation in supply chain relationships: The case of inventory leanness. *International Journal of Production Economics*, *236*, 108134. https://doi.org/10.1016/j.ijpe.2021.108134

Houston, M. B. (2004). Assessing the validity of secondary data proxies for marketing constructs. *Journal of Business Research*, *57*(2), 154–161. https://doi.org/10.1016/S0148-2963(01)00299-5

Hsieh, C. S., & Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, *31*(2), 301–319. https://doi.org/10.1002/jae.2426

Jacobs, B. W., & Singhal, V. R. (2020). Shareholder value effects of the Volkswagen emissions scandal on the automotive ecosystem. *Production and Operations Management*, *29*(10), 2230–2251. https://doi.org/10.1111/poms.13228

Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. *The Journal of Financial and Quantitative Analysis*, *31*(3), 309–335. https://doi.org/10.2307/2331394

Kashiwagi, Y., Todo, Y., & Matous, P. (2021). Propagation of economic shocks through global supply chains—Evidence from Hurricane Sandy. *Review of International Economics*, *29*(5), 1186–1220. https://doi.org/10.1111/roie.12541

Keil, J. (2017). The trouble with approximating industry concentration from Compustat. *Journal of Corporate Finance*, *45*, 467–479. https://doi.org/10.1016/j.jcorpfin.2017.05.019

Ketokivi, M., & McIntosh, C. N. (2017). Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations. *Journal of Operations Management*, *52*, 1–14. https://doi.org/10.1016/j.jom.2017.05.001

Kim, H. Y., & Davis, G. F. (2016). Challenges for global supply chain sustainability: Evidence from conflict minerals reports. *Academy of Management Journal*, *59*(6), 1896–1916. https://doi.org/10.5465/amj.2015.0770

Kim, J., Verdi, R. S., & Yost, B. P. (2020). Do firms strategically internalize disclosure spillovers? Evidence from cash-financed M&As. *Journal of Accounting Research*, *58*(5), 1249–1297. https://doi.org/10.1111/1475-679X.12337

Kim, Y. H., & Swink, M. (2021). Contingency role of a supplier's operational efficiency in the customer relationship–performance links. *International Journal of Operations & Production Management*, *41*(8), 1379–1403. https://doi.org/10.1108/IJOPM-06-2020-0382

Kumar, A., Cantor, D. E., & Grimm, C. M. (2019). The impact of a supplier's environmental management concerns on a buyer's environmental reputation: The moderating role of relationship criticality and firm size. *Transportation Research Part E*, *122*, 448–462. https://doi.org/10.1016/j.tre.2019.01.001

Kumar, A., Steven, A., & Paraskevas, J.-P. (2020). Impact of buyer-supplier TMT misalignment on environmental performance. *International Journal of Operations & Production Management*, *40*(11), 1695–1721. https://doi.org/10.1108/IJOPM-01-2020-0046

Lam, H. K. (2018). Doing good across organizational boundaries: Sustainable supply chain practices and firms' financial risk. *International Journal of Operations & Production Management*, *38*(12), 2389–2412. https://doi.org/10.1108/IJOPM-02-2018-0056

Li, Y., Chen, K., Collignon, S., & Ivanov, D. (2021). Ripple effect in the supply chain network: Forward and backward disruption propagation, network health and firm vulnerability. *European Journal of Operational Research*, *291*(3), 1117–1131. https://doi.org/10.1016/j.ejor.2020.09.053

Li, Y., & Zobel, C. W. (2020). Exploring supply chain network resilience in the presence of the ripple effect. *International Journal of Production Economics*, *228*, 107693. https://doi.org/10.1016/j.ijpe.2020.107693

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*, 107529. https://doi.org/10.1016/j.ijpe.2019.107529

Liu, G. (2020). Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. *Journal of Business & Finance Librarianship*, *25*(3–4), 1–55. https://doi.org/10.1080/08963568.2020.1847555

Lo, C. K., Wiengarten, F., Humphreys, P., Yeung, A. C., & Cheng, T. C. E. (2013). The impact of contextual factors on the efficacy of ISO 9000 adoption. *Journal of Operations Management*, *31*(5), 229–235. https://doi.org/10.1016/j.jom.2013.04.002

Lu, G., & Shang, G. (2017). Impact of supply base structural complexity on financial performance: Roles of visible and not-so-visible characteristics. *Journal of Operations Management*, *53*, 23–44. https://doi.org/10.1016/j.jom.2017.10.001

Mackelprang, A. W., & Malhotra, M. K. (2015). The impact of bullwhip on supply chains: Performance pathways, control mechanisms, and managerial levers. *Journal of Operations Management*, *36*, 15–32. https://doi.org/10.1016/j.jom.2015.02.003

Mergent. (2014). Mergent, Inc. announces the Integration of Horizon with Mergent Online. Retrieved from https://www.prweb.com/releases/2014/06/prweb11955932.htm [Accessed 04.03.22].

Mergent. (2021). About Mergent Online. Retrieved from https://www.mergentonline.com/noticescm.php?pagetype=about [Accessed 03.03.21].

Miller, J., Davis-Sramek, B., Fugate, B. S., Pagell, M., & Flynn, B. B. (2021). Editorial commentary: Addressing confusion in the diffusion of archival data research. *Journal of Supply Chain Management*, *57*(3), 130–146. https://doi.org/10.1111/jscm.12236

Mizgier, K. J., Pasia, J. M., & Talluri, S. (2017). Multiobjective capital allocation for supplier development under risk. *International Journal of Production Research*, *55*(18), 5243–5258. https://doi.org/10.1080/00207543.2017.1302618

Modi, S. B., & Cantor, D. E. (2021). How coopetition influences environmental performance: Role of financial slack, leverage, and leanness. *Production and Operations Management*, *30*(7), 2046–2068. https://doi.org/10.1111/poms.13344

Mukandwal, P. S., Cantor, D. E., Grimm, C. M., Elking, I., & Hofer, C. (2020). Do firms spend more on suppliers that have environmental expertise? An empirical study of U.S. manufacturers' procurement spend. *Journal of Business Logistics*, *41*(2), 129–148. https://doi.org/10.1111/jbl.12248

Orenstein, P. (2021). The changing landscape of supply chain networks: An empirical analysis of topological structure. *INFOR: Information Systems and Operational Research*, *59*(1), 53–73.

Osadchiy, N., Gaur, V., & Seshadri, S. (2016). Systematic risk in supply chain networks. *Management Science*, *62*(6), 1755–1777. https://doi.org/10.1287/mnsc.2015.2187

Osadchiy, N., Schmidt, W., & Wu, J. (2021). The bullwhip effect in supply networks. *Management Science*, *67*(10), 6153–6173. https://doi.org/10.1287/mnsc.2020.3824

Park, H., Bellamy, M. A., & Basole, R. C. (2018). Structural anatomy and evolution of supply chain alliance networks: A multimethod approach. *Journal of Operations Management*, *63*, 79–96. https://doi.org/10.1016/j.jom.2018.09.001

Park, Y. W., Blackhurst, J., Paul, C., & Scheibe, K. P. (2022). An analysis of the ripple effect for disruptions occurring in circular flows of a supply chain network. *International Journal of Production Research*, *60*(15), 4693–4711. https://doi.org/10.1080/00207543.2021.1934745

Piraveenan, M., Jing, H., Matous, P., & Todo, Y. (2020). Topology of international supply chain networks: A case study using FactSet Revere datasets. *IEEE Access*, *8*, 154540–154559. https://doi.org/10.1109/ACCESS.2020.3015910

Piraveenan, M., Senanayake, U., Matous, P., & Todo, Y. (2019). Assortativity and mixing patterns in international supply chain networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *29*(2), 023124. https://doi.org/10.1063/1.5082015

Potter, A., & Paulraj, A. (2020). Building supplier innovation triads: The effects of leadership relationships and alliance partner diversity. *International Journal of Operations & Production Management*, *40*(2), 144–172. https://doi.org/10.1108/IJOPM-07-2018-0418

Roth, A., Gray, J., Shockley, J., & Weng, H. H. R. (2015). The use of secondary source data for measuring performance in operations management research. *Available at SSRN*, 2271202.

Schennach, S. M. (2020). Mismeasured and unobserved variables. In S. N. Durlauf, L. P. Hansen, J. Heckman, & R. L. Matzkin (Eds.). *Handbook of econometrics*. Elsevier. https://doi.org/10.1016/bs.hoe.2020.07.001

Schwab, D. P. (2013). *Research methods for organizational studies*. Psychology Press. https://doi.org/10.4324/9781410611284

Schwieterman, M. A., Goldsby, T. J., & Croxton, K. L. (2018). Customer and supplier portfolios: Can credit risks be managed through supply chain relationships? *Journal of Business Logistics*, *39*(2), 123–137. https://doi.org/10.1111/jbl.12179

Schwieterman, M. A., Miller, J., Kremeyer, A. M., & Croxton, K. L. (2020). Do supply chain exemplars have more or less dependent suppliers? *Journal of Business Logistics*, *41*(2), 149–173. https://doi.org/10.1111/jbl.12249

Securities and Exchange Commission – SEC. (2018). Disclosure update and simplification. Retrieved from https://www.sec.gov/rules/final/2018/33-10532.pdf [Accessed 11.02.21].

Serpa, J. C., & Krishnan, H. (2018). The impact of supply chains on firm-level productivity. *Management Science*, *64*(2), 511–532. https://doi.org/10.1287/mnsc.2016.2632

Shao, B. M. B., Shia, Z. M., Choi, T. Y., & Cha, S. (2018). A data-analytics approach to identifying hidden critical suppliers in supply networks: Development of nexus supplier index. *Decision Support Systems*, *114*, 37–48. https://doi.org/10.1016/j.dss.2018.08.008

Sharma, A., Kumar, V., Yan, J., Borah, S. B., & Adhikary, A. (2019). Understanding the structural characteristics of a firm's whole buyer–supplier network and its impact on international business performance. *Journal of International Business Studies*, *50*(3), 365–392. https://doi.org/10.1057/s41267-019-00215-x

Sharma, A., Pathak, S., Borah, S. B., & Adhikary, A. (2020). Is it too complex? The curious case of supply network complexity and focal firm innovation. *Journal of Operations Management*, *66*(7–8), 839–865. https://doi.org/10.1002/joom.1067

Simchi-Levi, D., Schmidt, W., Wei, Y., Zhang, P. Y., Combs, K., Ge, Y., Gusikhin, O., Sanders, M., & Zhang, D. (2015). Identifying risks and mitigating disruptions in the automotive supply chain. *Interfaces*, *45*(5), 375–390. https://doi.org/10.1287/inte.2015.0804

Smith, J. A., & Moody, J. (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, *35*(4), 652–668. https://doi.org/10.1016/j.socnet.2013.09.003

Smith, J. A., Moody, J., & Morgan, J. H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, *48*, 78–99. https://doi.org/10.1016/j.socnet.2016.04.005

Son, B. G., Chae, S., & Kocabasoglu-Hillmer, C. (2021). Catastrophic supply chain disruptions and supply network changes: A study of the 2011 Japanese earthquake. *International Journal of Operations & Production Management*, *41*(6), 781–804. https://doi.org/10.1108/IJOPM-09-2020-0614

Steven, A. B., & Britto, R. A. (2016). Emerging market presence, inventory, and product recall linkages. *Journal of Operations Management*, *46*, 55–68. https://doi.org/10.1016/j.jom.2016.07.003

Steven, A. B., Dong, Y., & Corsi, T. (2014). Global sourcing and quality recalls: An empirical study of outsourcing-supplier concentration-product recall linkages. *Journal of Operations Management*, *32*, 241–253. https://doi.org/10.1016/j.jom.2014.04.003

Stewart, D. W., & Kamins, M. A. (1993). *Secondary research: Information sources and methods*. Sage. https://doi.org/10.4135/9781412985802

Sugrue, D., Martin, A., & Adriaens, P. (2021). Applied financial metrics to measure interdependencies in a waterway infrastructure system. *Journal of Infrastructure Systems*, *27*(1), 05020010. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000588

Varadarajan, P. R., Jayachandran, S., & White, J. C. (2001). Strategic interdependence in organizations: Deconglomeration and marketing strategy. *Journal of Marketing*, *65*(1), 15–28. https://doi.org/10.1509/jmkg.65.1.15.18129

Wang, Y., Li, J., & Anupindi, R. (2015). Risky suppliers or risky supply chains? An empirical analysis of sub-tier supply network structure on firm performance in the high-tech sector. *Ross School of Business Working Paper Series*, *no. 1297*. https://doi.org/10.2139/ssrn.2705654

Wang, Y., Li, J., Wu, D., & Anupindi, R. (2021). When ignorance is not bliss: An empirical analysis of subtier supply network structure on firm risk. *Management Science*, *67*(4), 2029–2048. https://doi.org/10.1287/mnsc.2020.3645

Wani, D., Dong, Y., Malhotra, M. K., & Xu, K. (2021). Emerging market penetration and emissions performance. *Decision Sciences*, *52*(1), 283–324. https://doi.org/10.1111/deci.12436

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *Management Information System Quarterly*, *26*(2), 13–23.

Wetzel, P., & Hofmann, E. (2019). Supply chain finance, financial constraints and corporate performance: An explorative network analysis and future research agenda. *International Journal of Production Economics*, *216*, 364–383. https://doi.org/10.1016/j.ijpe.2019.07.001

Wiedmer, R., & Griffis, S. E. (2021). Structural characteristics of complex supply chain networks. *Journal of Business Logistics*, *42*(2), 264–290. https://doi.org/10.1111/jbl.12283

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Wu, J., Zhang, Z., & Zhou, S. X. (2022). Credit rating prediction through supply chains: A machine learning approach. *Production and Operations Management*, *31*(4), 1613–1629. https://doi.org/10.1111/poms.13634

Xiong, Y., Lam, H. K., Hu, Q., Yee, R. W., & Blome, C. (2021). The financial impacts of environmental violations on supply chains: Evidence from an emerging market. *Transportation Research Part E*, *151*, 102345. https://doi.org/10.1016/j.tre.2021.102345

Xiong, Y., Lam, H. K., Kumar, A., Ngai, E. W., Xiu, C., & Wang, X. (2021). The mitigating role of blockchain-enabled supply chains during the COVID-19 pandemic. *International Journal of Operations & Production Management*, *41*(9), 1495–1521. https://doi.org/10.1108/IJOPM-12-2020-0901

Yamamoto, R., Kawadai, N., & Miyahara, H. (2021). Momentum information propagation through global supply chain networks. *The Journal of Portfolio Management*, *47*(8), 197–211. https://doi.org/10.3905/jpm.2021.1.264

Zhao, K., Zuo, Z., & Blackhurst, J. V. (2019). Modelling supply chain adaptation for disruptions: An empirically grounded complex adaptive systems approach. *Journal of Operations Management*, *65*(2), 190–212. https://doi.org/10.1002/joom.1009

Zhong, W., Ma, Z., Tong, T. W., Zhang, Y., & Xie, L. (2021). Customer concentration, executive attention, and firm search behavior. *Academy of Management Journal*, *64*(5), 1625–1647. https://doi.org/10.5465/amj.2017.0468

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Culot, G., Podrecca, M., Nassimbeni, G., Orzes, G., & Sartor, M. (2023). Using supply chain databases in academic research: A methodological critique. *Journal of Supply Chain Management*, *59*(1), 3–25. https://doi.org/10.1111/jscm.12294