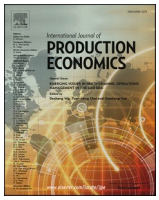
Int. J. Production Economics 256 (2023) 108708

Contents lists available at ScienceDirect

International Journal of Production Economics

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



A mixed methods approach to analyze and predict supply disruptions by combining causal inference and deep learning

Frank Bodendorf [a,*], Maximilian Sauter [b], Jörg Franke [a]

[a] Institute for Factory Automation and Production Systems, Friedrich-Alexander-University of Erlangen-Nuremberg (FAU), Egerlandstraße 7-9, 91058, Erlangen, Germany
[b] Friedrich-Alexander-University of Erlangen-Nürnberg (FAU), Schloßplatz 4, 91054, Erlangen, Germany

ARTICLE INFO

Keywords:
Supply disruption
Supply chain management
Risk management
Deep learning
Causal inference
Design science

ABSTRACT

In today's complex supply networks, disruptions caused by diverse events represent major unknown operational conditions and risk factors requiring research on supply chain risk management. Hereby, operations management (OM) researchers are traditionally focusing on defining analytical and mathematical risk mitigation strategies characterized by a variety of different constraints, such as capacities, timeliness, or relations, leading to limited transferability into practice. Given the complexity of modern supply chains we argue that a data-driven approach to supply chain risk management enhances the assessment and comparability of risk mitigation strategies. Thereby an approach that allows conclusions about causal relationships between supply chain interventions and potential outcomes is of particular importance. Inspired by deep learning as well as causal learning theory and following a design science research approach, we design an analytical model that can predict supply disruptions based on external and internal data and quantify causal effects on delivery reliability. The model is evaluated in a single case study using data from a large first tier automotive supplier. The results show that the model has a high predictive performance and can learn causalities from observed data and analyze interventions effecting supply disruptions. The identification of causal relationships offers the potential to identify lacking supplier relationships and consequently bundle supply chain risk management activities. Building on the empirical and analytical insights, we discuss implications for both theory and practice.

1. Introduction

In today's complex supply chain networks risks of disruptions due to organizational factors within and environmental factors outside of supply chains represent major challenges for both practitioners and researchers (Baghersad and Zobel, 2021). Therefore, international companies are investing significant effort to develop viable supply chain risk management strategies. Our systematic literature review identifies a dominant tendency to focus on risk mitigation scenarios, relying on mathematical and analytical research approaches. Theoretical limitations raise doubts about the real-world implications and validity of OM's normative mathematical models (Choi et al., 2016). In recent years, accompanied by the surge in "big data" and the increasing demand for business analytics, the importance of empirical research based on observational data in the field of supply chain risk management is experiencing a significant boost (Ho et al., 2017). However, reliable forecasts based on the analyses of large datasets represent a fundamental challenge. Building upon recent advances in deep learning receives great attention lately (Zhu et al., 2021). Especially in big datasets as well as non-linear feature relationships, the underlying flexible and complex architectures of layers and neurons promise superior performance capabilities compared to traditional machine learning (ML) models and even shallow neural networks. The significant impact of deep learning is not only limited to research areas of marketing and sales (Chui et al., 2018) but is on its way to becoming the standard for predictive analytics within operations research and management, benefiting from the increasing availability of processable big data in businesses (Kraus et al., 2020). On this way you must cope with the so-called black box problem (Gunning et al., 2019). Based on the deep networks' complex architectures and functionalities as well as the limited knowledge in managerial environments, ML and specifically deep learning models are considered opaque, limiting their explicability and consequently their practical relevance (Kraus et al., 2020; Bodendorf et al., 2022). To improve the limited interpretability for managers, this paper enhances the deep

* Corresponding author.
E-mail addresses: frank.bodendorf@fau.de (F. Bodendorf), max.sauter@fau.de (M. Sauter), joerg.franke@faps.fau.de (J. Franke).

https://doi.org/10.1016/j.ijpe.2022.108708
Received 12 September 2021; Received in revised form 18 October 2022; Accepted 31 October 2022
Available online 21 November 2022

learning approach by answering the "why" question, based on the theory of causal inference. The resulting causal effects and relationships between features provide a better decision support for management. To prove this, we define the following objectives as the main pillars of our research: As a theoretical foundation this paper first reviews OM literature focusing on supply disruptions and outlines the most significant theoretical constraints. Additionally, theories of causality as well as deep learning in OM are introduced (Section 2). We build upon the gained theoretical insights to design three key artifacts to analyze and forecast supply disruptions (Section 3 and 4). In a single embedded case study, we evaluate our artifacts providing empirical evidence (Section 5). The paper is concluded outlining research and management implications as well as proposing directions for future research (Section 6 and 7).

2. Related work

To our knowledge, this paper is the first approach to model and analyze supplier disruptions using deep learning and causal inference in OM research and therefore can be seen as an extension to supply chain risk management (SCRM). The enhancement relies on the analytical investigation of causal effects and relations, exploiting a large empirical data base. The main hurdles of comparable approaches in the past might have been on the one hand the lack in empirical research for theory building and verifying based on observable business data as well as on the other hand the lack in deep learning and causal inference capabilities. A basic understanding of causal inference theory is essential to predict supplier disruptions by examining causal relationships and effects. For such an approach we need to (1) know the causes and consequences of supply disruptions, (2) understand how supply disruption in OM research has been dealt with in the past, (3) acquire fundamental knowledge of causal reasoning as well as deep learning, and (4) provide an overview about their applications in OM. In the following sections, related work is reviewed to provide the theoretical basis for the subsequent model design and empirical investigation.

2.1. The significance of supply disruptions

A variety of incentives such as cost advantages, access to local workforce potential or presence in foreign markets drive companies to globally expand their supply chains. In all sectors of the economy, such globalization trends have been evident in the past, enabling companies to exploit regional, financial or organizational advantages in a stable environment. However, on the downside of global operations, the robustness of supply chains has suffered, reducing their transparency and making them more vulnerable to various types of disruptions (Tang, 2006).

The term disruption or supply chain disruption is not consistently used throughout SCRM literature (Bode and Wagner, 2015). In this paper, we refer to the definition of Craighead et al. (2007), who characterize supply chain disruptions as "unplanned and unanticipated events that disrupt the normal flow of goods and materials within a supply chain […] and, as consequence, expose firms within the supply chain to operational and financial risks.". Just to give examples for business relevance in 2000, a lightning strike and the resulting shut-down of Philips Semiconductor plant in New Mexico had led to a shortage of components for Sony Ericsson, provoking a loss of $400 million in potential revenue (Tomlin, 2006). In 2007, the market capitalization of Menu Foods Corp., a Canadian producer of pet food, dropped by 50% based on severe food quality issues, caused by an unannounced process change at the supplier side (Yang et al., 2009).

A deeper source-related differentiation of disruptions by Tang (2006) seems to be useful, which distinguishes between operational, inherent scenarios, caused by uncertain supply, demand or cost events, and external disruptions based on natural or man-made disasters as well as economic or political crises (Kleindorfer and Saad, 2005).

For affected businesses however, the consequence of all kinds of "unplanned and unanticipated events" (Dada et al., 2007) remains consistent: financial loss of varying extent. Hendricks and Singhai (2005a), for instance, report that companies experience 33% to 40% lower stock returns, relative to their industry benchmarks, in case they are suffering from supply chain disruptions.

Being motivated by the persistent occurrence of supply chain disruptions as well as their severe consequences, extensive SCRM research has been published in the field of OM, tackling the challenges of disruptions by focusing on empirical as well as analytical research methodologies.

2.2. Supply disruption related research in the OM discipline

This literature review is structured following Vom Brocke et al. (2009), covering all essential phases of a literature review process: (1) Definition of Extent, Thematic Delimitation, (2) Literature Search Process, (3) Literature Analysis, (4) Elaboration of Future Research Opportunities.

To investigate related work on supply disruptions, the interrelated research field of operations and supply chain management is looked at. Following Jourqual 3 journal ranking, qualitatively oriented OM journals which appeared after 2005 are investigated, such as the journal of *Operations Management*, the journal of *Operations Research*, the journal of *Decision Sciences*, the journal of *Manufacturing & Service Operations Management,* and the journal of *Production and Operations Management.*

Having chosen EBSCOhost, SCOPUS and Clarivate Analytics Web of Science as the three major databases for the subsequent keyword search, a search phrase mapping (see Fig. 1) is carried out in one database after the other, relying on a title, abstract and keyword search, combining the search terms one, two, and three (e.g., [Supplier OR Vendor] AND [Disruption OR Discont*] AND [Method OR System]). By collecting the articles of all three databases and excluding those, which are mentioned several times, an overall volume of 713 articles is gathered. Subsequently, a backward and forward search in alignment with Webster and Watson (2002) is conducted to realize a bibliometrics related investigation. With the support of this guideline, multiple analyses such as the *Most Relevant Sources* and the *Most Cited References* (both based on the h-Index) are executed.

The bibliometric results shown in Fig. 2a highlight the twenty most relevant sources, measured by the number of published papers contained in the final sample. As illustrated, the journal of *Operations Management* (75), the journal of *Manufacturing and Service Operations Management* (52), the journal of *Production and Operations Management* (46), *Omega – International Journal of Management Science* (35), the journal of *Decision Sciences* (27), and the journal of *Operations Research* (23) are the greatest contributors to the field of interest as far as the number of published articles is concerned. Based on the most relevant journals displayed in Fig. 2b as well as on the article's number of local citations within the database, the paper published by Tomlin (2006) is overshadowing the subsequent articles.

In a qualitative content analysis, the applied research methods, the applied data analysis methods, the covered topics, the specific limitations as well as the proposed future research questions of the above mentioned most cited references are summarized in Appendix A.

Analyzing the articles in detail, we see the dominance of intervention-based *analytical mathematical* research methods, which are applied in 50% of the qualitatively investigated references. In this paper the definition of Wacker (1998) describes analytical mathematical research as the development of new mathematical relationships to investigate the behavior of risk mitigation models under different circumstances. Hereby, the evaluation of the designed models is investigated using simulated data instead of real-world company data. The related articles mainly deal with the determination of the optimal number of suppliers and the allocated procurement volume under supply uncertainty as well as with the mathematical comparison of multiple

Keywords: Search algorithm					
1		**2**		**3**	
Supplier		Disruption		Method	Approach
Vendor		Discont*		System	Study
Contractor	**AND**	Disturbance	**AND**	Case	Design
Subcontractor		Interruption		Analy*	Framewok
Provider		Interference		Model*	Simulation
Furnisher				Strateg*	Assess*

Fig. 1. Search phrase mapping.

and single sourcing strategies to minimize the risk of supply chain disruptions overall (Tomlin, 2006; Babich et al., 2007; Dada et al., 2007). In summary, the collected articles contribute to the theoretical framework of SCRM, which targets the "identification, assessment, treatment, and monitoring of supply chain risks with the aid of the internal implementation of tools […] and of external coordination and collaboration with supply chain members […]" (Fan and Stevenson, 2007). For instance, Tomlin (2006) presents an analytical model based on a Markovian inventory system to study a "single-product setting in which a firm can source from two suppliers, one that is unreliable and another that is reliable but more expensive" and provides disruption-management strategies for risk-neutral and risk-averse companies. In addition to the *analytical mathematical* research, Kim et al. (2015) follow an *analytical conceptual* research approach to analyze four fundamental supply network structures regarding their resilience. The research effort is built upon the supply chain resilience theory, which focuses on methods and tools to enhance the "organizational capability to survive in a turbulent environment" (Chowdhury and Quaddus, 2017).

Besides the *analytical mathematical* or *analytical conceptual* research approaches, seven out of twenty examined articles focus on empirical contributions to SCRM theory (Kleindorfer and Saad, 2005; Bode and Wagner, 2015; Hendricks and Singhai, 2005a). Referring to the Scudder and Hill (1998) classification of empirical research and data analysis methods, it is evident that empirically extracting insights is the preferred course of action. For instance, the authors Hendricks and Singhai (2005a & 2005b) investigate correlations between supply chain disruptions and operating performance indicators using data provided by The Wall Street Journal and the Dow Jones News Services. In addition, Azadegan et al. (2020) use two empirical studies to elaborate whether business continuity programs limit the damage caused by supply chain disruptions and improve company financial performance.

Finally, a marginal number of articles carry out *Literature Reviews* (LR) (Snyder et al., 2016; Tang, 2006). Tang (2006) develops "a unified framework for classifying SCRM articles" through differentiating strategies regarding supply, product, demand, and information management. Ivanov and Dogui (2021) on the other hand reflect on the ability of operations research methods to cope with the ripple effect in specific supply chain risk management scenarios.

The results of the presented systematic literature review leads us to a supply disruption research map (see Appendix B). It consists of five main pillars in accordance with Tang (2006)'s framework for SCRM article classification. Each pillar is characterized by a short content description, an illustration of the research targets and methods as well as a list of key references. In addition to the main pillars differences between operational and disruption risks are outlined.

Finally, based on the outlined research efforts, their limitations as well as future research opportunities (Appendix A), we can draw three key conclusions underpinning our research problem, which can also be seen as gaps in the SCRM research:

1. Mostly, Supply Disruption is investigated in an analytical way, which is frequently associated with limitations and restrictions of real supplier-customer relationships. Tomlin (2006), for example, consideres equal lead times for all suppliers as well as a stochastic, but stationary demand in each calculated period. Furthermore, Yu et al. (2008) ground their model on the assumption that the investigated suppliers are characterized by unlimited capacities. Finally, Sawik (2011) limits his designed model to a single-period view with respect to supplier selection and contract awards. Therefore, real-world relevance is highly debatable. The dilemma between theory and practice in OM is further emphasized by Choi et al. (2016), who similarly derive the conclusion that highly theoretical assumptions hurt "the real impact of OM research" as well as arise "doubts on the validity".
2. The mitigation of operational risks (e.g., uncertain supply, demand, costs), which seem to be more tangible than disruption risks (e.g., earthquakes, hurricanes), dominate present research efforts, although as illustrated by Kleindorfer and Saad (2005) disruptive events such as earthquakes can lead to far more serious consequences for affected businesses.
3. A dominant number of articles aim at developing optimal supplier order allocation models, e. g., for uncertain supply volumes rather than designing methods to prematurely forecasting operational supply disruptions.

Building on the identified research opportunities, we address the research gaps in Supply Disruption by considering:

1. a mixed methods research approach targeting practical and theoretical relevance. The approach combines empirical real-world data gathering and their analytical investigation (Choi et al., 2016).
2. a data analytics perspective which strives to capitalize on the promising capabilities of deep learning to make data driven predictions. Thereby, the focus does not lie on classification algorithms long established in literature, such as Support Vector Machine, KNearest Neighbor, Random Forest, and Logistic Regression (Caiado et al., 2021), but on rather novel combinations of deep learning with methods suitable for reasoning and analyzing causality.

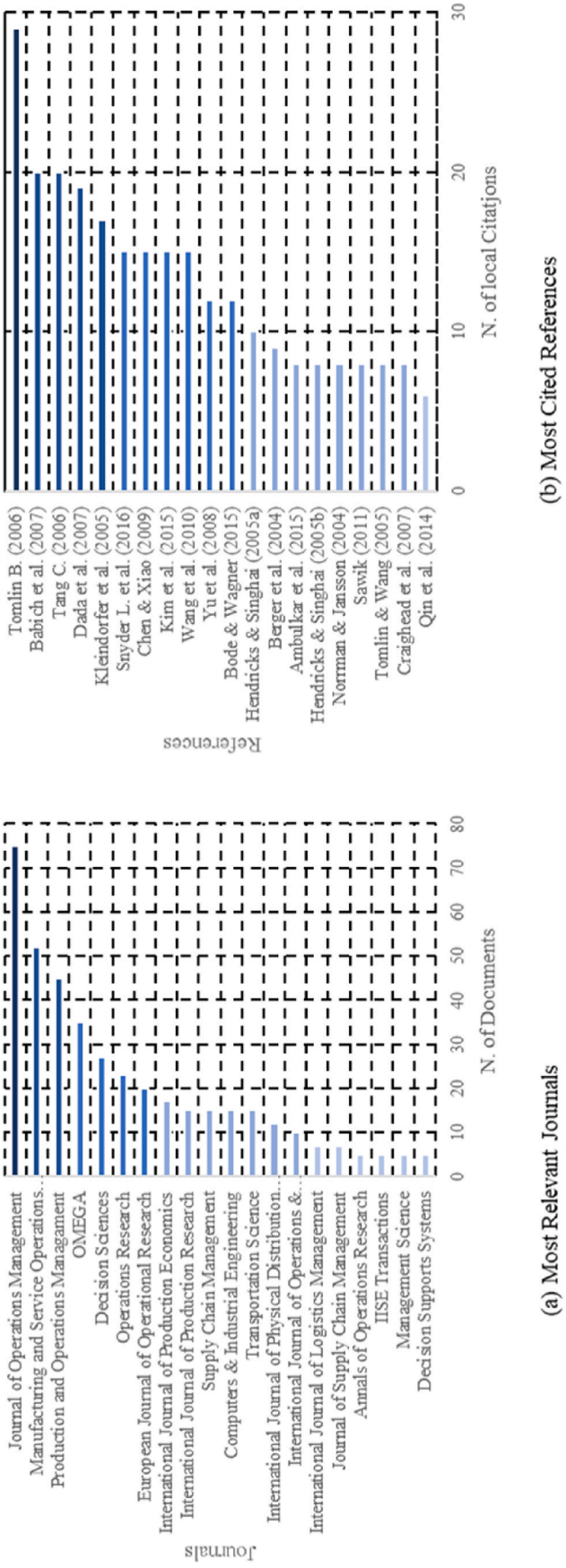


Fig. 2. Most relevant journals (a) and most cited references (b).

3. an observation and analysis of disruption risks. Therefore, external factors compared to internal factors receive a higher level of attention in our data selection process.

2.3. Causal inference fundamentals

To gain a fundamental understanding of causal inference theory, the following section introduces the Structural Causal Model Framework (SCMF), highlighting structural equations and concluding with an overview about causality-based learning mechanisms. For a more granular overview we refer to Appendix C.

The overall target of studying causation is to understand the causal effect of some treatment A (manipulated variable) on some outcome Y (responsive variable) or in other words how and why causes influence their effects. Causation is thereby an enrichment of statistics, enabling the examination of problems, which cannot be communicated in the standard language of statistics and which pushes traditional statistical methods to their limits. In this context, a popular phenomenon is the so-called Simpson's Paradox (Simpson, 1951), which refers to the existence of data in which statistical associations based on an entire population are reversed in every subpopulation. In that regard, a full paradox occurs when each of the aggregated datasets illustrates a reversed effect compared to the disaggregated dataset. In case of a partial paradox, the reversal is not observable in every subpopulation of data. Overall, the Simpson's Paradox leads to the practical decision-making question which aggregation level is suitable to represent the results of interest (Shmueli and Yahav, 2017). At this point, analyzing causality and determining the causal structure or the so-called data generating process with its underlying probability models, provides a reliable background information for practitioners.

Because traditional statistical methods are limited in their ability to determine causal mechanisms in the data, additional tools based on causal inference theory have been developed that can illustrate and interpret causal relationships. Two questions, also referred to as causal inference questions, are hereby of central importance (Guo et al., 2020): By modifying which variables within the dataset can we change the status of the others? How would the values of the variables change, if we perform such manipulations, also referred to as interventions? Therefore, the fundamental task is the identification of causal relationships between two variables, which are mathematically described by causal models. In that respect, one of the most popular causal models is the Structural Causal Model Framework (SCMF), which is defined by Guo et al. (2020). It is also used in this paper.

Causal relationships can be depicted as directed acyclical graphs (DAGs). Based on the causal graph, causal effects, represented by the directed edges, can be specified by a set of equations called structural equations. These equations can therefore be seen as a causal interpretation of DAGs, which additionally allow statements about the distribution of a variable under experimental interventions (Guo et al., 2020; Drton and Maathius, 2017).

In a comprehensible example, displayed in Fig. 3, the three direct edges reflect the causal relations within our exemplary model, in which we assume binary variables for simplicity (t = treatment variable, y = outcome variable, x = confounder):

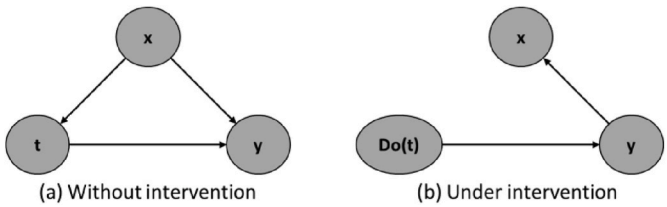


Fig. 3. SCMF without (a) and under (b) intervention.

- $t \rightarrow y$: Improving the infrastructure influences the volume of transported goods.
- $x \rightarrow y$: An economic crisis affects the volume of transported goods due to various described side factors.
- $x \rightarrow t$: An economic crisis impacts the overall state of the transportation infrastructure.

Based on the causal graph, a set of non-parametric structural equations can be used as the representation for the three causal effects $t \rightarrow y$, $x \rightarrow y$, and $x \rightarrow t$: $x = f_x(E_x)$, $t = f_t(x, E_t)$, $y = f_y(x, t, E_y)$. The terms E_x, E_y and E_t are hereby referred to as "noise", symbolizing any unknown or random impacts (unobserved variables), that may influence the relation between the endogenous variables x, y, and t. Learning causal effects, referred to as causal inference, is concerned with quantifying an expected change of a defined outcome variable y in case a modification of the treatment variable t is executed (Guo et al., 2020).

2.4. Artificial neural networks and deep learning

In this section, we provide insights into the functionality of artificial neural networks (ANN) and especially deep neural networks (DNN), referring to Goodfellow (2017), Kraus et al. (2020), and Géron (2019).

Being inspired by the architecture of biological neural networks, an ANN is a collection of units, called neurons, in which each neuron is further interconnected with a predefined number of neighbors, while being structured in consecutive layers. Based on the inputs received from other neurons, each artificial neuron generates certain output information through the weighted sum of inputs (with the help of the adjustable parameter v) and a characteristic activation function, displayed in Fig. 4a. The output is then passed to the subsequent layer and neurons (see Fig. 4b)

Multilayer ANNs are additionally referred to as deep neural networks, exemplary visualized in Fig. 4b. The term DNN is hereby not consistently used in literature. In this paper, we relate to the definition of Géron (2019): "When an ANN contains a deep (1 or more) stack of hidden layers, it is called a deep neural network." Consequently, especially for DNN, the optimization of hyperparameters such as the number of layers and the neurons per layer represents a challenging task and is still subject to active research. Kraus et al. (2020) hereby outline the opportunity to intentionally realize overfitting of the selected predictive model on the training data and then execute regularization methods to eliminate overadjustments.

2.5. Deep learning and causal inference in OM

Fig. 5 outlines the interlinking of a deep learning model with a causal interference approach and the respective structural equations.

However, although a connection is suggested in the displayed theoretical framework, both components are predominantly addressed separately in OM research. Kraus et al. (2020), for instance, emphasize the promising potential of DNN in relation to predictive analytics by comparing performance between traditional ML and deep embedded networks in a trisected case study. The suitability of DNN for use in the context of causal questions, however, is not addressed. The same applies also for Zhu et al. (2021), who implement deep as well as recurrent neural networks to improve the accuracy of demand forecasts affecting pharmaceutical supply chains, or Ketzenberg (2020), who targets the assessment of customer return behaviors based on neural network classifiers. In contrast, Shmueli and Yahav (2017) solely focus on the usage of classification and regression trees to study causal effects among different data aggregation levels without mentioning a potential implementation of DNN.

Consequently, we target to contribute to OM research by combining both research areas of deep and causality based learning in the field of examining and forecasting supply disruptions. In doing so, not only the superior forecast performance of deep learning models in comparison to shallow ANNs is presented, but additionally the benefits provided through the addition of causal inference methods are outlined. The architectures and peculiarities are described in detail in Section 5.

3. Research approach

We follow van Aken and Romme (2009) and Chandrasekaran et al. (2020) who both promote a design science research approach in operations management to investigate new theoretical and managerial insights by engaging with practice and solving complex field problems, hereby concluding in design artifacts, bridging the fundamental gaps between scholars and practitioners. The overall objective is to define and design theory ingrained artifacts that support the application of causality based deep learning methods for supply disruption analysis and prediction. The approach proposed by Peffers et al. (2007) can hereby be seen as a guideline for our further course of action.

First, the problem needs to be clarified. The key issues in OM are identified as (1) the questioned real-word relevance of analytical risk mitigation strategies due to their limitations and restriction of real supplier-customer relationships, (2) a significant imbalance between research targeting the mitigation of operational risks in contrast to

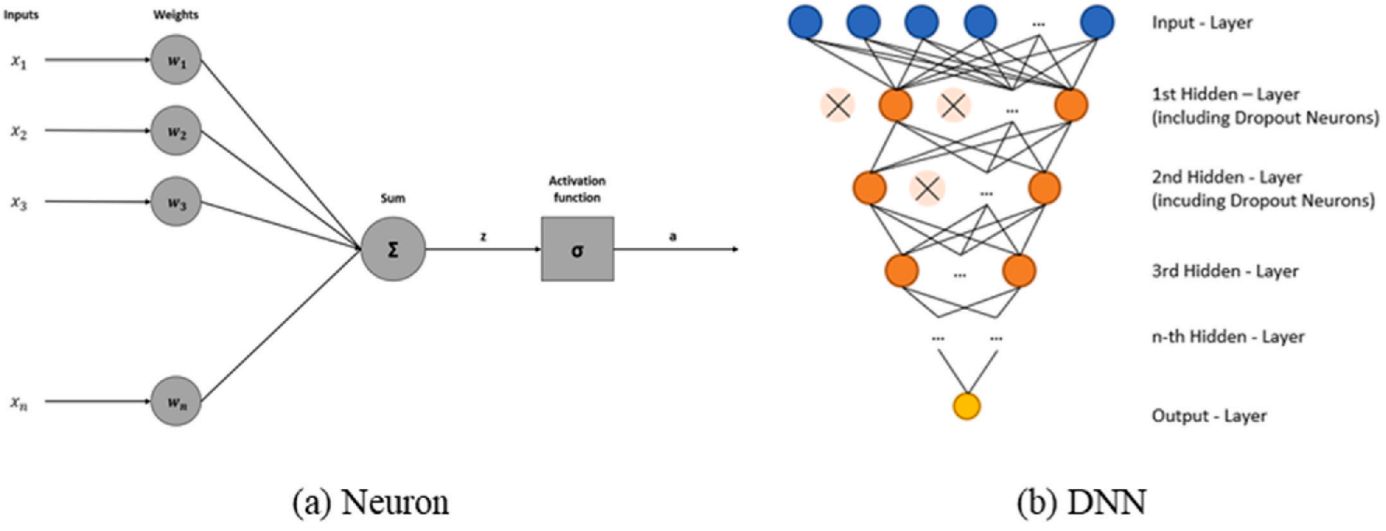


Fig. 4. Architecture of (a) artificial neuron and (b) exemplary perceptron.

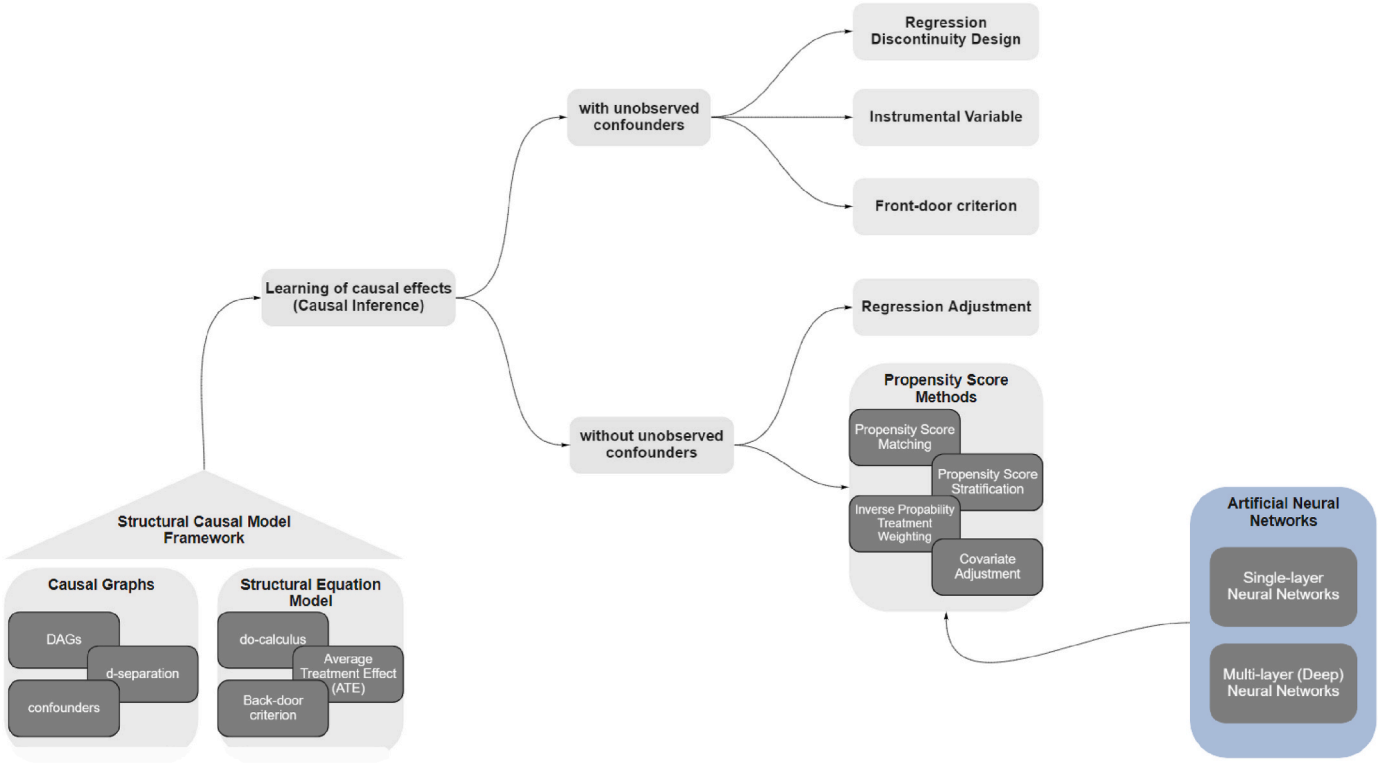


Fig. 5. Theoretical framework - causal inference and DNN

research aiming at minimizing disruption risks, and (3) the overall research direction to develop optimal order allocation models rather than focusing on supplier disruption forecasts. Afterwards we define qualitative and distinct targets to illustrate the requirements for a potential approach. Our first objective is to develop a mixed methods solution, which is inspired by theory and practice through combining empirical real-world data collection and their analytical penetration (Choi et al., 2016). Additionally, since the investigation relies on "Big Data", the second objective is to implement deep learning methods enriched by the benefits of causal inference. Finally, the third objective focuses on the observation of disruption risks to dismantle the imbalance between the analysis of observational and disruption risks.

At the design and development stage, we lack knowledge in (1) requirements and processes which are essential for a causality-related data mining phase, (2) a suitable selection of features to enable a best possible forecast of supply disruptions, and (3) an effective combination of the DNN and causal inference theory. Based on these requirements our artifacts are designed, which are displayed in Fig. 6 (step 3). Building on iteratively gained knowledge during the development of the above-mentioned artifacts, we enter the demonstration phase. The detailed explanation of the data selection process as well as its decisive requirements are of central importance. In addition, a prototype for the prediction of supply disruption, based on the identified suitable combination of deep learning and causal inference, is designed and reviewed. Finally, the prototype is trialed using fictitious datasets to eliminate any technical errors and secure its functionality.

Following the technical and theoretical review, we conduct an embedded case study in cooperation with a first-tier supplier operating in the automotive industry, to evaluate whether the causality based deep learning approach shows higher performance and better interpretability than single deep learning applications. Based on the results of the embedded case study, further potential for improvement is derived, which is discussed in the last section of this paper.

4. Designing for deep causal learning

In this section, we describe how our objectives for causality-based supply disruption forecasting lead to the design of artifacts. These artifacts include a causal data mining process, a feature selection step enabling the analysis of supply disruptions as well as a prototype combining DNN and causal inference.

4.1. A data mining process for causality-based data analytics

Causality-based data mining covers the process to detect relationships between variables and forecast the effect of interventions through the application of causal ML algorithms. Since the characteristics of causal data mining, which include causal discovery and causal inference, are not fully reflected in popular data mining techniques such as SEMMA or CRISP-DM (Truong, 2021), we propose a novel five step process to improve the capturing of causality in data.

Step 1. *Defining*: Based on gathered information about the industry, suppliers, third parties, components, geographic locations, and other influencing factors detailed in Artifact 2, objectives are identified and research questions are derived. Step 1 concludes with the complexity of the system and the expected flexibility including the definitions of possible variables and features within the model. Hereby, the concept of external validity or generalizability is key, allowing the reader to understand the clear rationale for the case study selection and its context (Gibbert and Ruigrok, 2010).

Step 2. *Sampling & Pre-processing*:

The second step is characterized by collecting data from reliable sources. Hereby the objective of constructing internal validity is central, e.g., by applying the strategy of triangulation of different sources of data (Gibbert and Ruigrok, 2010) as well as deviant-case analysis. The optimal number of instances and features depends on the application-specific dilemma between potential information loss and modelling efficiency. Subsequently, the collected dataset is examined

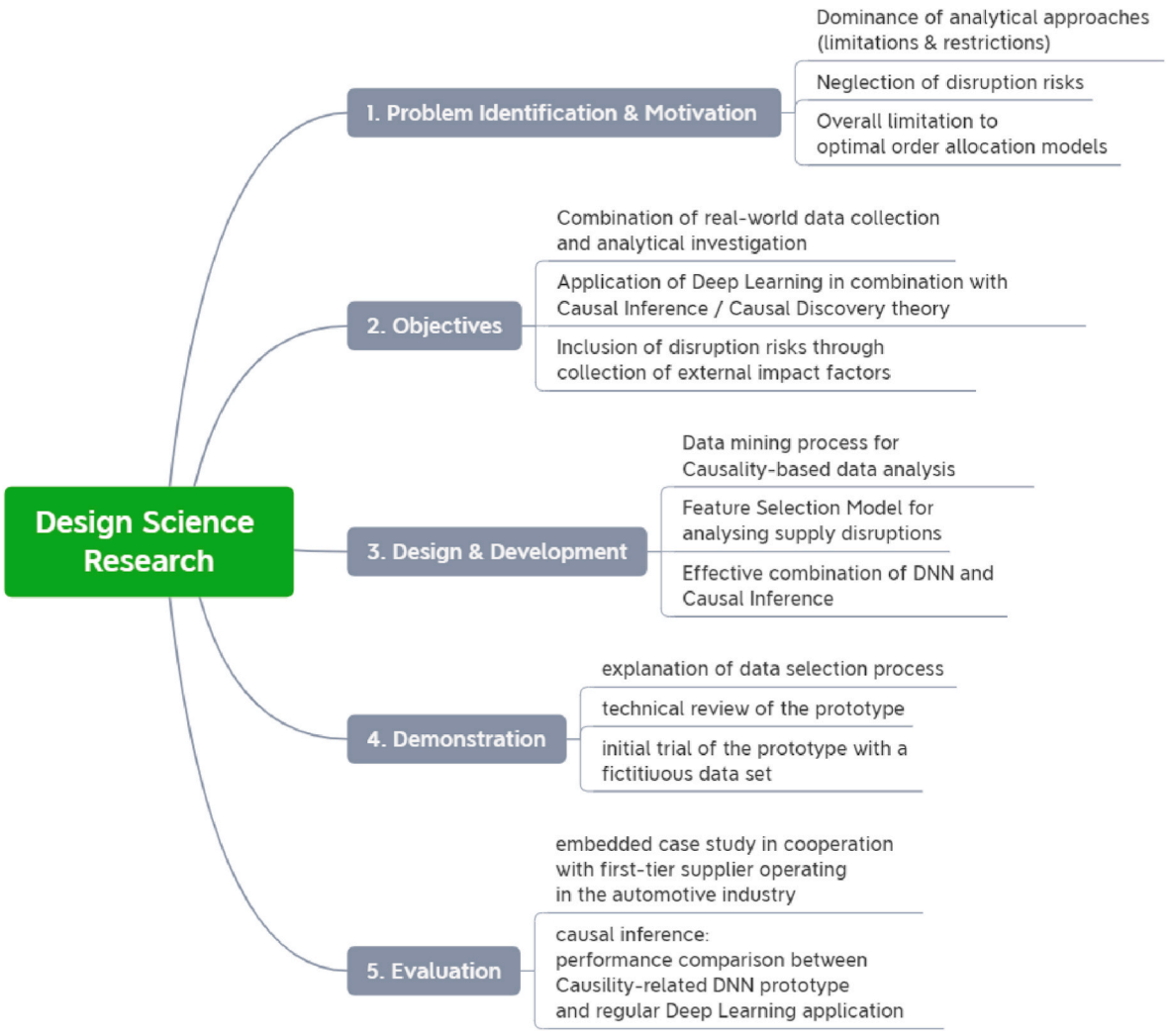


Fig. 6. Design science research approach following Peffers et al. (2007).

carefully to gain knowledge about the features' distribution, additionally descriptive characteristics as well as outliers and missing values. Since many machine learning algorithms achieve higher performance if the numerical input variables are subject to standard probability distributions, the next crucial step after data validation is discretization. Discretization is considered as a "date reduction mechanism since it diminishes data from a large domain of numeric values to a subset of categorical values" (Ramirez-Gallego et al., 2016). In that sense, the equal-width, equal-frequency, K-means or decision tree-based discretization are the most influential approaches (Gupta, 2019). Afterwards, the dataset is split into training and validation samples. An application of the k-fold cross-validation is conceivable.

Step 3. *Learning & Evaluating*:

Step 3 starts with a supervised learning process. The network generated by this allows practitioners to evaluate the causal effect of each input on the defined target variable and to make predictions about its outcome. After the supervised learning process, an evaluation of the causal network performance in terms of accuracy and predictive power is required to guarantee the model's validity and reliability. Criteria such as the confusion matrix, ROC chart and AUC score, classification accuracy, logarithmic loss, F1 score or GINI coefficient are applicable to gain insights into the network's overall performance (Hao and Fan, 2019). In case the resulting insights are not satisfactory, further post-hoc analyses need to be conducted to achieve the desired performance improvements. In this context, Truong (2021) proposes to gradually remove characteristics that have a small effect on the target variable prediction. Also worth mentioning is the approach of Van Vliet and Salmelin (2020) who focus on fine-tuning a learned model using domain information beyond the underlying dataset and implement an appropriately applicable framework.

Step 4. *Concluding*: In the last step, the most influential attributes need to be highlighted in conjunction with their individual scope of effects on the target variable. Subsequent sensitivity analysis additionally shows how a variation of one variable affects the defined target feature, relying on different methods of transformation such as uniform distribution or permutation. Finally, causal discovery and causal inference analysis can be used to support managers in their decision-making. Hereby, Truong (2021) emphasizes the importance of different what-if scenarios.

Reviewing the causal data mining process, certain similarities to the CRISP-DM as well as the SEMMA approach are apparent in terms of sampling, pre-processing, modeling, and evaluation. However, the four-step process mentioned here, with its focus on decision support as well as management implications, has a higher suitability for causality inspired data mining projects compared to traditional CRISP-DM and SEMMA. Consequently, we apply the proposed causal data mining process in the subsequent embedded case study.

4.2. Feature selection for analyzing supply disruptions

Since supplier selection processes are designed to avoid potential

supply disruptions, a variety of supplier selection criteria can be used to forecast future disruptions. Due to its good transparency and structure, Chan et al. (2008)'s hierarchy for supplier selection is vertically extended by more recent supplier selection research efforts by Rajesh and Ravi (2015) and Mina et al. (2021) and serves as a basis for feature selection. Our feature selection step is consequently built upon four main pillars, namely the *quality* and *service* aspects, the supplier's *background* information and *external risk factors* such as geographical, social, or economic environment (see Appendix D). In the following, each main criterion is discussed in detail and further sub-criteria are presented. Each criterion is defined in accordance with related research in supplier selection as well as discussions in internal cross-functional expert workshops.

4.2.1. Service

The demand-oriented availability of goods or services is influenced to a large extent by supplier-specific performance characteristics (Chan et al., 2008). We therefore define the following list of attributes:

- *Delivery reliability*: A continuous flow of material along the entire supply chain depends on the schedule and quantity adherence of the suppliers. Therefore, attributes describing the historical delivery reliability of suppliers provide insights into possibilities of future disruptions.
- *Information sharing:* As illustrated by Yang et al. (2009), information transfer running in parallel to the material flow can cause disruptions in the supply chain and affect the profitability of all parties involved. Therefore, features enabling an assessment of the interorganizational communication between supplier and customer as well as public information sharing of suppliers are considered.
- *Flexibility and responsiveness*: The ability of the supplier to adapt to fluctuations of customer demands, price structures, and order frequencies is essential in selecting suitable suppliers, but additionally in forecasting of supply disruptions especially in highly volatile markets and industries.

4.2.2. Quality

Quality criteria not met by suppliers bear the risk to disrupt the material flow within supply chains. Consequently, attributes determining the quality of suppliers are of central importance. Following Chan et al. (2008), Chen et al. (2005), Vilko and Hallikas (2012) and Ho et al. (2015) these attributes are decisive:

- *Product reliability*: Besides the schedule and quantity adherence mentioned within the prime criterion "Service", the quality of goods related to a certain supplier as well as the quality-specific complaints are further key aspects, enabling conclusions about interferences as far as the supply is concerned.
- *Quality assurance and process capability:* Quality deviations within the production process on the supplier-side as well as the implemented quality assurance process are further indicators towards the future development of quality fidelity. Moreover, missing or short warranty periods of suppliers at the expense of customers can be a sign of a lack in quality awareness and thus a potential risk for supply interruptions

4.2.3. Supplier background information

We follow Chan et al. (2008), Vilko and Hallikas (2012), and Ho et al. (2015) who emphasize the importance of evaluating the technological, financial, and infrastructural characteristics of suppliers:

- *Technological Capability:* The ability of the supplier to keep up with the fast pace of product and service innovations is crucial to ensure an appropriate supply level even across multiple product life cycles. Features in relation to the technological alignment of the supplier, such as its research and development commitments, are important as they may indicate upcoming qualitative or functional challenges.
- *Financial Status*: The supplier's financial status is crucial for its current and future operational room for maneuver. Therefore, data attributes that indicate the financial performance and stability of the supplier are deemed important.
- *Facility and Infrastructure*: The comparison between the planned delivery quantities in future periods and the available production capacity at the supplier-side, which may indicate future supply bottlenecks, are considered.

4.2.4. External risk factors

Supply chain operations can be affected by exogenous factors, also referred to as external disruption risks (Tang, 2006). In alignment with the frameworks of Chan et al. (2008), Klibi et al. (2010), Pettit et al. (2013), Thun and Hoenig (2011), Vilko and Hallikas (2012), and Ho et al. (2015), we specify the heterogenous nature of external impacts as:

- *Geographical Location*: Based on the geographic location of the suppliers and the corresponding climatic conditions, estimates regarding the probability of natural disasters are made possible. As mentioned in Section 1, these external climatic impacts bear a significant threat towards a firm's profitability. Furthermore, spatial distances between suppliers and customers are derivable, allowing conclusions to be drawn about the geographical complexity of supply chains.
- *Political Stability and Infrastructure:* The political system of the supplier country, its stability, and government intervention in the areas of trade and customs are key features that can indicate disruptive effects on existing buyer-supplier relationships. Characteristics to be examined here are the political stability index and the government effectiveness index. Additionally, we consider attributes measuring the development status of the country's associated public physical infrastructure, such as the ratings towards quality of roads, railways, ports or airports as relevant (Neven, 2021).
- *Economic and Social Development*: In addition to the political stability of the country the economic development of the country shapes the monetary status of the supplier and influences its financial room for maneuver. Consequently, regional or even global economic recessions have the potential to disrupt the operations of individual businesses and, beyond that, entire supply chains. Therefore, features indicating domestic or global economic trends, such as the economic growth rate, the inflation rate as well as the prevailing exchange rate, are of central importance. Closely related to the country's economic development is its social well-being. Negative economic and financial developments are often accompanied by waves of layoffs and labor disputes, which limit the operational performance of affected businesses. Therefore, characteristics such as the unemployment rate, the poverty rate or the human development index should not be neglected when monitoring social framework conditions.
- *Terrorism and Crime Rate:* By the example of the 9/11 incident, it becomes apparent that terrorism is one of the greatest enemies for global business processes and internationally operating supply chains. Manufacturers locate their geographic supplier base in regions characterized by low crime rates and terrorism to minimize the impact on planned delivery schedules and routes (Chan et al., 2008). Therefore, features expressing the crime or terrorism rate, such as the security threats index (Neven, 2021), are included.

4.3. An effective combination of DNN and causal inference

Causal inference theory offers multiple possibilities to utilize the beneficial operation of ANNs as well as DNNs. This section combines both approaches by the implementation of a DNN for propensity score assessment (Austin, 2011; Westreich et al., 2010; Cannas and Arpino, 2019; Luo et al., 2020).

To estimate the effects of treatment, exposures, and interventions on outcomes, randomized controlled trials (RCTs) are seen as the silver bullet due to their avoidance of confounding through random treatment allocation (Lane et al., 2012; Austin, 2011; Kim et al., 2016; Beal and Kupzyk, 2014). However, causal observational studies in complex real-world environments do not satisfy the characteristics of randomization as non-randomized groups systematically differ from each other based on the number of covariates (Lane et al., 2012; Rosenbaum and Rubin, 1983). Consequently, the outcome difference between the treated and controlled groups may not reflect the true treatment effect of a RCT (Kim et al., 2016). Being challenged with the handling of confounding biases associated with observational data, propensity score methods experience an increasing prominence (see Appendix C). Here, the propensity score matching (PSM) takes center stage due to its effectiveness in removing covariate imbalance (Kim et al., 2016).

Focusing on the estimation of treatment effects through the PSM related generation of matched sets between untreated and treated subjects, a four-step process following Lane et al. (2012) is considered:

1. *Covariate Determination and Propensity Score Calculation:* First, all theoretically relevant covariates likely to forecast group membership need to be clarified, whereby the number of considered covariates is not subject to restrictions (Lane et al., 2012). Then, propensity scores, or probabilities of group membership, are calculated across all participants, using a research dominant logistic regression method.
2. *Propensity Score Matching:* Following the calculation of the propensity scores, participants are matched between groups to control for covariates. Hereby, the processes of greedy and optimal matching need to be distinguished. As far as greedy matching is concerned, nearest-neighbor matching algorithms equipped with thresholds regarding the maximum allowable distance in probabilities are applied (Austin, 2011). In contrast, optimal matching processes tend to match both participants based on minimizing the total absolute distance between control and treatment propensity scores (Beal and Kupzyk, 2014).
3. *Balance Assessment:* Once the matching process has been completed, the effectiveness needs to be evaluated by investigating the balance within the newly matched sample. Hereby Rosenbaum and Rubin (1983) propose to stratify treatment and control groups across quintiles of the propensity score to evaluate the interaction effects as well as the statistical significance of group differences. Further examination of the standardized mean differences between groups on covariates provides additional information on effect size and strength.
4. *Estimation of Treatment Effects:* Once a sufficiently matched sample has been created, the PSM allows an estimation of the treatment effect, the average treatment effect (ATE), and the average treatment effect for those treated (ATT) by directly comparing results between treated and untreated subjects (Austin, 2011; Lane et al., 2012).

Having detailed the determination of treatment effects with the PSM approach, it is necessary to clarify how we aim to ensemble the functionality of ANN and DNN and PSM. In literature, logistic regression is used for the score estimation due to its attractive mathematical constraint of defining probabilities in the range of 0 to 1 as well as its ability to easily converge on parameter estimates. However, following the argumentation of Westreich et al. (2010) this way of estimating propensity scores with its key assumptions of linearity to the logit appears to be naïve as it may result in a poor model fit leading to residual confounding in propensity score analysis and biased estimates of treatment effects. At this point, neural networks provide the option to replace logistic regression. In comparison, the DNNs offer a variety of advantages. First, being challenged with the analysis of high-dimensional data, characterized by many covariates, DNNs outperform traditional logistic regression in various research applications. Second, any smooth polynomial function is approximable by a DNN, regardless of polynomial order or interaction terms, due to adaptive design and hyperparameter modification (Westreich et al., 2010). In addition, inspired by Setoguchi et al. (2008), neural network approaches to propensity score estimation should result in less bias than comparable logistics regression approaches, especially within nonlinear settings, and benefit from a superior generalization performance overall (Zhang et al., 2021). Finally, Cannas and Arpino (2019) further emphasize the estimation robustness of neural networks, particularly in cases where logistic regression approaches fail to guarantee any sufficient matching balance.

5. Embedded case study

In the final phase of our design science research approach we carry out an embedded case study in collaboration with a first-tier supplier in the automotive industry. We have access to internal and external data coming from private as well as public sources covering various subunits of supplier-customer relationships. Given the characteristic three-layer architecture of understanding the underlying case, conceptualizing a realistic model and explaining the lessons learned as proposed by Scholz and Tietje (2022), the embedded case study is particularly relevant for examining scenarios in which boundaries and causal relationships between treatment and target variable characteristics are not clear. To demonstrate the benefits of deep learning as well as its combination with causal inference in data analytics, we guide through the case study in five subsequent phases, being aligned with the data mining process outlined in Section 4.1.

5.1. Problem and target definition

The first-tier supplier in the embedded case study, consecutively referred to as company A, offers complex engine-related products and systems to a variety of different globally operating automotive original equipment manufacturers (OEMs). Customers, warehouses, production plants as well as suppliers are interconnected through multiple cross-continental supply chains monitored by company A. Being frequently challenged by the risk of upstream supply disruptions as well as their subsequent downstream consequences, available external and internal data allows to identify the underlying causal relationships and derive key recommendations for improvement. In cooperation with Company A all upstream supplier-customer relationships for a specific production site are examined.

5.2. Data sampling, exploration, and pre-processing

The embedded case study draws upon a proprietary dataset containing information from three different sources: the supplier's enterprise resource planning system, the Global Economy platform providing economic data on foreign countries (Neven, 2021) as well as Visual Crossing's climatic data (Visual Crossing, 2021). The dataset is collected from January 2017 to December 2020, including a sample of 23.000 observations. In reference to the main pillars of the feature selection step (see Section 4.2) a total of 54 variables are included in the dataset (Appendix E depicts each feature considered as well as its corresponding data type). Looking at the external disruption risks, the largest part of the variables belongs to the classes *supplier background information* and *external risk factors*. After eliminating missing and duplicated instances the overall number of observations is reduced to 21.942, still providing a sample size in accordance with the recommendation of Figueroa et al. (2012) to include about 100 observations per variable (5.400 in this case).

In alignment with the six defined tasks of purchasing by Wannenwetsch (2010) – *right product, right time, right quality, right quantity, right location, right costs* - which can be regarded as a key performance indicator for supply interruptions, we supplement the data set with the additional target variable "Delivery Reliability", which indicates a

supply disruption if the overall purchasing activity fails to achieve a defined performance level of 99%. In this context, the target variable returns the statement FALSE. Throughout the dataset, there is a class imbalance between the two states of the target variable "Delivery Reliability" (True = no disturbance (35%) / FALSE = disturbance (65%)), which poses a significant challenge to the predictive performance of the neural networks and consequently requires additional methods to address the class imbalance (Buda et al., 2018).

Having sampled and defined all relevant features within the dataset, we focus on the process of data exploration and visualization. Special attention is given to the frequency distribution of the data and the presence of outliers due to their impact on modeling accuracy (Khamis et al., 2005).

Appendix F illustrates the frequency distributions of all features in the dataset. There are outliers for each characteristic, highlighted in red, indicating a non-normal distribution of the data. Especially in connection with the application of shallow ANN, such outlier scenarios are proven to negatively affect its modeling accuracy and performance (Khamis et al., 2005). However, since each outlier is related to a specific characteristic of a certain supplier and we target to implement a deep ANN with a higher degree of anomaly robustness, any loss of information resulting from the removal of statistical outliers must be prevented (Cox, 2017). Conversely, any extraction of anomalies would hinder the subsequent process of analyzing causal effects and relationships. E.g., looking at the distribution of the feature "Distance" in detail (marked with red frame), it turns out that most data is allocated in a range of 0–1000 km, suggesting that the selected production site of company A mainly has relationships with intracontinental suppliers and consequently pursues a strategy to reduce the spatial complexity of its supply chains. However, as the outliers show, sourcing relationships with distant suppliers cannot be avoided. Detailed analysis of outliers and frequency distributions is followed by preprocessing of the data, which deals primarily with multicollinearity testing, exhaustive feature selection, feature coding, scaling, and preparation of the training and test data sets. Here, we start with transforming the existing categorical variables "Country", "City", "Product Group", and "Shipping Mode" to numerical attributes using the One-Hot encoding approach. Then, the entire data set is split into a training and a test data set, which are aligned in a ratio of 80:20. Since deep learning algorithms prefer processing normally distributed data (Géron, 2019), both training and test datasets undergo a standardization process.

5.3. Predictive supply disruption analytics with DNN

After data preprocessing, the task is forecasting supply disruptions measured by the defined target variable "Delivery Reliability" using DNNs. Within this procedure, we outline the performance advantage of deep ANN by comparing to a default single-layer ANN. However, before any performance characteristic can be evaluated, a closer look at the architecture of both ANNs is required.

For the multi-layer DNN, we suggest a sequential neural network architecture consisting of one input and one output layer as well as five stacked hidden neural layers in-between. Each layer is associated to its neighbors in a fully connected fashion. Taking the requirements of the subsequently executed PSM into account, the layer's activation function is set to the sigmoid function, ensuring an output value between 0 and 1, even if the training process is relatively slow in comparison to the ReLu activation function (Krizhevsky et al., 2012). Hidden dense and dropout layers alternate each other, leading to an overall number of three dense layers and two dropout layers. The number of neurons per layer is decreasing towards the output layer consisting of a single neuron. A total of 15,800 parameters are adjustable during the training phase of the DNN. To counteract overfitting, the architecture includes dropout layers that randomly set inputs to zero with a certain frequency at each step during training and weight optimization of the model. The neurons "dropped out" do not contribute to forward propagation and do not participate in the backpropagation phase. Thus, because neurons of subsequent layers do not rely on the presence of other neurons, increased learning stability is enforced (Krizhevsky et al., 2012). In addition, regularization techniques are integrated into the DNN's hidden dense layers by applying penalties on the layer's kernel, bias, and output. These penalties are then added to the loss function being optimized, further reducing the tendency of the model to overfit. Finally, a k-fold cross-validation approach is implemented to furthermore avoid overfitting.

The default single-layer ANN, in contrast, consists of a fully connected architecture of one input layer, one hidden layer and one output layer. The sigmoid activation function is also chosen. Dropout layers are not used, resulting in a total number of 108 trainable parameters. However, to avoid overfitting and ensure a certain degree of comparability, regularization and the k-fold cross-validation approach is implemented. Fig. 7 visualizes both neural network architectures in direct comparison. Following Kraus et al. (2020) and Géron (2019), the optimal set of hyperparameters is identified based on an exploration of a predefined parameter space using the RandomizedSearchCV method. The tuned set of hyperparameters includes the number of hidden layers, the number of neurons per hidden layer, the dropout rate, the learning rate for all layers as well as the relevant batch size and epochs. Having identified the optimal set of hyperparameters, both tuned models are fitted using an adaptive learning rate optimization algorithm specified as Adam (Géron, 2019). The optimal batch size is set to 500 examples. In several training and adaptation cycles, the kernel's weight regulation is set to use the sum of absolute and squared weights, while the bias and output regulations refer only to the sum of squared weights. In this regard, a weight decay of 0.0001 has turned out to be advantageous.

For performance evaluation of the default ANN as well as the DNN, a variety of different classification metrics is used. To assess the impact of the previously outlined class imbalance this experiment compares the DNN applied to the originally unbalanced data set, the DNN applied to an undersampled dataset (Nanni et al., 2015) as well as the DNN applied to a synthetically oversampled data set using the SMOTE process (Chawla et al., 2002). Table 1 evaluates the prediction results with exemplary covariates depicted in Table 2.

Overall, the SMOTE approach delivers the highest performance. Besides the observation of the above illustrated performance characteristics, we monitor the model's classification accuracy and loss development (see Fig. 8). Looking at the accuracy and loss curves of the DNN, a stable development comes with a small absolute loss and accuracy difference between training and validation dataset, with a marginal overfitting tendency. The default ANN however is characterized by an unstable loss and accuracy development accompanied by significant absolute differences between training and SIMP validation data, indicating overfitting. Focusing on a comparison of absolute loss and accuracy differences between both models at 40th epoch, the DNN's accuracy difference between training and validation data set is 0.004, whereas the loss difference is 0.007. The ANN's accuracy difference however is 0.011, whereas the loss difference is 0.081.

Overall, taking the performance comparison of Fig. 8 and Table 1 into account, the conclusion can be drawn, that our DNN model is able to successfully improve the predictive accuracy over default ANN architectures, thereby allowing for a more precise identification of supply disruptions and reduced delivery reliabilities.

5.4. Method application

This section intends to shed light into black-box DNN models by calculating propensity scores and estimating treatment effects. As mentioned in Section 4.3, we follow the introduced four step process.

The first step for estimating and matching propensity scores is to select multiple covariates likely to impact group membership, meaning covariates that may support forecasting the likelihood of deliveries being affected by disruptions. The selection of covariates is already

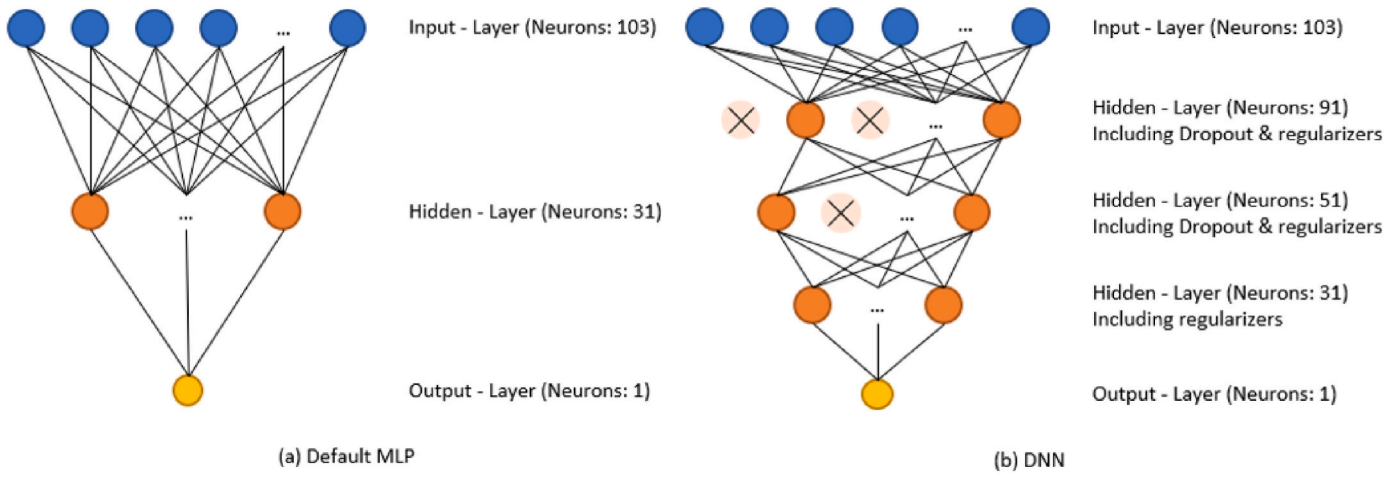


Fig. 7. Architectures of Default MLP and DNN.

Table 1
Comparison of prediction performance - Default MLP vs. DNN.

Model	Free Parameters	Performance					
		Accuracy	Loss	AUC	Category	Recall	F1 Score
Default ANN	104	0.82	0.51	0.86	True	0.89	0.88
					False	0.62	0.72
DNN (unbalanced)	15,800	0.85	0.14	0.88	True	0.91	0.9
					False	0.72	0.8
DNN (undersampling)	15,800	0.86	0.12	0.9	True	0.94	0.9
					False	0.71	0.8
DNN (SMOTE)	15,800	**0.86**	**0.12**	**0.91**	True	**0.97**	**0.9**
					False	**0.75**	**0.8**

Table 2
Pre-matched exemplary covariates.

	Distance <200 km (n = 8.504)		Distance >200 km (n = 13.438)		Total Cohort (n = 21.492)	
Exemplary Covariates	Mean	SD	Mean	SD	Mean	SD
Euler Hermes Rating	4.00	0.00	3.31	0.95	3.58	0.82
GDP per Capita	46,363.54	1116.23	46,100.51	4228.60	46,327.66	3369.72
Globalization index	88.80	0.06	85.21	1.15	88.74	0.89
Industrial production (%. Dev. p. a.)	−2.28	6.09	−1.36	5.84	−1.72	5.96

considered within the feature selection step (see Section 4.2). Afterwards, the collected features are used to estimate the probability of the treatment. By compiling and fitting of the DNN, propensity scores including their logit transformation are calculated, which are subsequently used in the matching process. To calculate the ATE of all features, binary treatment scenarios are defined for all attributes, requiring a repetitive execution of the PSM method. However, to provide an example as well as to prove the quality of matching, we subsequently outline the PSM application for the below illustrated binary treatment scenario:

- T = 0: Supplier's distance < 200 km (control group)
- T = 1: Supplier's distance ≥ 200 km (treatment group)

Table 2 depicts a comparison of exemplarily selected baseline features between the respective treatment and control group. The portion of deliveries coming from suppliers with a distance greater than 200 km outweighs the portion of the counterfactual group considering an overall total of 21,492 deliveries.

Following the neural network's application as a substitute for logistic regression, the propensity scores for all observations are calculated, which conceptualize the observation's probability of being treated as a function of measured baseline covariates (Jacovidis, 2017). Austin (2011) proposes to consider different sets of covariate variables in the propensity model, including "all measured baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (e.g., potential confounders), and all covariates that affect both treatment assignment and the outcome (e.g., true confounder)". In consequence, all features integrated in the original dataset except the treatment are utilized within the DNN fitting process. Once the neural network is fitted with the treatment group symbolizing the outcome, subjects with similar propensity scores are required to be matched to enable the determination of treatment effects. To be able to evaluate the pre-matched propensity score overlap between treatment and control group, Fig. 10a visualizes the pre-matched propensity score distribution of both treatment groups. It is obvious that the propensity scores are distributed differently in each treatment group, indicating that treatment and control group include sourcing relations with generally different baseline features. However, as required by the propensity score matching method, a common support region is visible that maps a sufficiently common area of treatment and control groups under the distribution of propensity scores to allow generalizability of treatment effects. In this example the common support region can range from a propensity score of t = 0.32 to a propensity score of t = 0.86. Instances

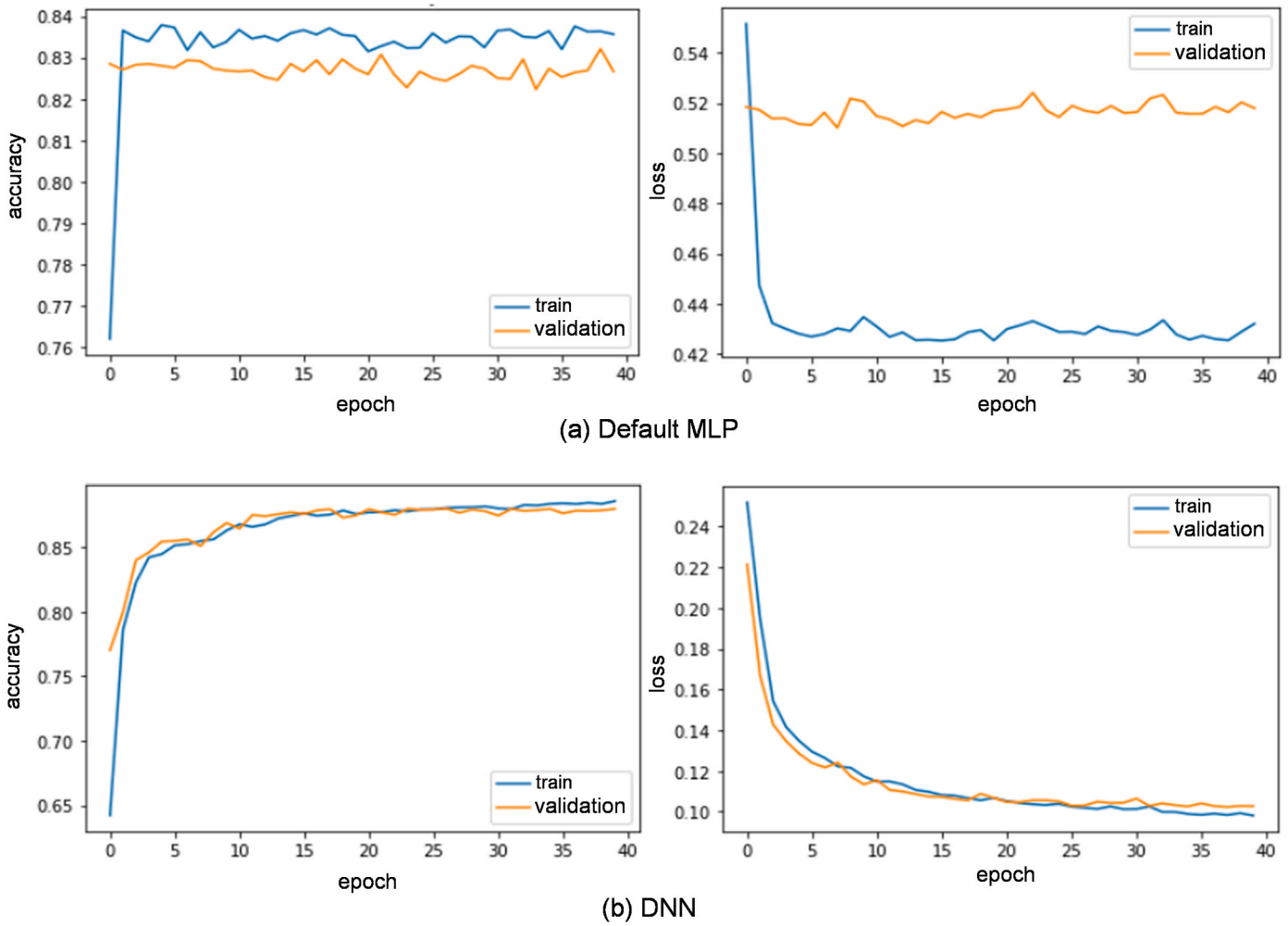


Fig. 8. Model accuracy and loss development of (a) default MLP and (b) DNN.

outside this range are excluded from the matching process and consequently contribute to bias in the estimation of treatment effects (Lane et al., 2012).

After the determination and evaluation of propensity scores, we enter the phase of matching on these scores. A method applicable in this context is the nearest neighbor matching procedure, which sequentially matches observations in the treatment group to units in the control group based on their estimated propensity scores (Abadie and Imbens, 2016). Hereby, a distinction must be made between the 1-to-1 nearest neighbor matching, generating a relation between one treatment and one control instance, and the k-to-1 matching for any positive integer k (Staffa and Zurakowski, 2018), enabling relations to multiple control instances. In our embedded case study, we perform the second nearest neighbor matching approach to limit the overall loss of data points and consequently the bias. Following the definition of the matching procedure, three key decisions are required to gain well-matched treatment and control pairs.

The first decision is to match with or without replacement. When matching without replacement is selected, each control group member can only be matched to one treatment group member. When matching with replacement is selected, an untreated control group member can be matched with multiple treatment group members. Matching with replacement may cause a violation of the independence requirement of observations, representing one of the fundamental assumptions of causal inference, and its practical relevance is debatable (Jacovidis, 2017). We focus on matching without replacement, while solving the issue of a potential sequence dependency through random data sampling.

The second decision to be made is the choice between greedy and optimal matching. As far as the greedy matching process is concerned, the allocation of untreated control group members to members of the treatment group is done sequentially through the list of treated instances, whereby the matched control subject is removed from any further consideration. In contrast, the optimal matching method based on the Relax-IV algorithm continuously creates, breaks, and rearranges matches to minimize the overall sum of match distances (NCSS, 2021). Therefore, due to its ability to minimize the overall difference in propensity scores within pairs, we select the optimal matching approach as the preferred method.

Finally, the third decision is related to the caliper width size. Here, the caliper width is the maximum difference between propensity scores allowed for two potentially matched observations (Staffa and Zurakowski, 2018). Following the recommendation of Stuart (2010) concerning a reasonable width to minimize bias, a caliper of 0.25 times the standard deviation of the propensity score's logit is implemented, where the logit of a given propensity score is equal to the natural logarithm $\ln(t/[1-t])$ with variable t representing the propensity score. As a result, 2187 well-matched pairs of treatment (supplier distance $\geq$ 200 km) and control group (supplier distance $<$ 200 km) are identified in the dataset. Once the matching process is complete, a balance assessment is required between the two treatment groups with respect to their baseline characteristics, which provides information on matching quality. Since the sample size may significantly impact p-values comparing the baseline factors between treatment and control group (Staffa and Zurakowski, 2018), we follow the suggestion of Rubin (2001) to use the standardized

mean difference for balance evaluation, which should be close to zero (d < 0.20) in case of an effective matching. Based on the comparison of each covariate characteristics in treatment and control group, an absolute standardized mean difference less than 0.20 indicates a negligible difference between both groups for this respected covariate and consequently a sufficient matching.

As illustrated in Fig. 9, the absolute standardized mean difference *d* for several exemplarily selected features before and after the implemented matching process is displayed. This shows both a reduction in the mean difference and an undercutting of the proposed threshold value. Based on the characteristics of the absolute standardized mean difference, the matching procedure in our embedded case study results in a sufficiently matched data sample suitable for inferring treatment effects.

Overall, in reference to the above illustrated performance characteristics as well as the improved propensity score overlap highlighted in Fig. 10, a sufficiently matched dataset of treatment and control group is created, forming the fundament for the subsequent and final phase of estimating treatment effects.

Any difference found within the matched dataset is consequently more reflective of the RCT's true treatment effect due to the reduction of the confounding bias (Staffa and Zurakowski, 2018). Based on the classification approach introduced by Rosenbaum and Rubin (1983), we split the estimated propensity scores in n groups, for which the average outcomes for treatment and control groups are determined. Following the estimation of treatment effects for each propensity score group, the ATT can be determined. We rely on the causal inference assumption of independence, additionally referred to as ignorability assumption, indicating that treatment status is independent of potential outcomes (Rosenbaum and Rubin, 1983; Keele, 2015). Based on this assumption of Rosenbaum and Rubin (1983), we consequently deduce the average treatment effect for the entire matched sample.

5.5. Presentation and interpretation of results

In the embedded case study, 43 binary treatment scenarios are analyzed. Appendix G provides a tabular overview about all investigated scenarios, the average outcome within treatment and control group, the calculated ATE as well as interpretations. We subsequently highlight and interpret the most influential results for each feature group. Fig. 11 depicts their associated ATEs.

5.5.1. Supplier-specific background

In this feature group, we need to focus on three key attributes: the "Shipping Mode" describing the mode of transport between supplier and customer, the "Product Group" defining the procured object, as well as the "Euler Hermes Rating" evaluating the supplier's financial market position, operating performance, and capital adequacy developments (Serra, 2020). The defined treatment scenario for each feature is listed in Appendix G.

Overall, based on the observed causal effects, three key conclusions can be drawn:

- "Shipping Mode": With respect to company A, shipments using trucks as the main mode of transport are characterized by a delivery reliability decrease of 38.9% in comparison to the respective sea shipments.
- "Product Group": Raw Material suppliers show a superior delivery performance (increase of 55.2%) in comparison to vendors offering operating supplies, measurement equipment, and spare parts.
- "Euler Hermes Rating": A positive association between delivery performance and financial well-being is observable, benefitting relations to financially well-equipped suppliers. Business relationships with suppliers with Euler Hermes Ratings less than 4 are to be avoided, since they are accompanied by a delivery reliability decrease of 22.5%.

5.5.2. External risk factors

As mentioned in Section 2.2, the investigation of supply disruptions caused by external events or circumstances is one central target of this paper. In consequence, a variety of attributes belong to the feature group of external risk factors. Starting with the subunit "Geographical Location", several attributes must be investigated, providing information about the supplier's country, its distance to the customer, the geographical risk of natural calamities as well as the climatic conditions at the point of shipment (temperature, precipitation, wind speed).

The most influential results within the subunit "Geographical Location" can be seen as follows:

- "Country": Comparing suppliers based on their origin, the strong performance of foreign suppliers is apparent. Overall, deliveries by domestic suppliers are affected by an average reliability decrease of 39.7%.
- "Distance": Suppliers with a distance of less than 200 km to Company A's plant are characterized by poorer delivery performance (−35.8%) compared to subcontractors located further away.
- "World Risk Index": Using the world risk index to quantify the risk of catastrophes resulting from extreme natural events, procurement relationships with suppliers in countries with an index of over 3 are associated on average with a 34.6% increased risk of supply disruptions.

The subsection "Political Stability & Infrastructure" focuses on the causal effects of political stability, globalization, government effectiveness, and infrastructure quality at the point of shipment on delivery reliability coming to the following conclusions:

- "Infrastructure Quality": Focusing on the state of the infrastructure at the point of shipment, the quality of roads turns out to have the greatest influence on the overall delivery performance in comparison to the other infrastructure related indicators (Neven, 2021). Less developed road networks (index <5) decrease the delivery reliability on average by 36.8%.
- "Government Effectiveness Index": Suppliers in countries with less effective government structures, such as those characterized by poor policy formulation, implementation, monitoring, and public or civil services (Neven, 2021), suffer a 30.1% drop in delivery reliability.

Besides the subunit of "Political Stability and Infrastructure", the causal effects of features characterizing the economic development of the supplier's country on the target variable "Delivery Reliability" are worth analyzing. Therefore, the PSM method is performed for several treatment variables, e.g., for the gross domestic product and trade balance, forecasts regarding investment and inflation developments as well

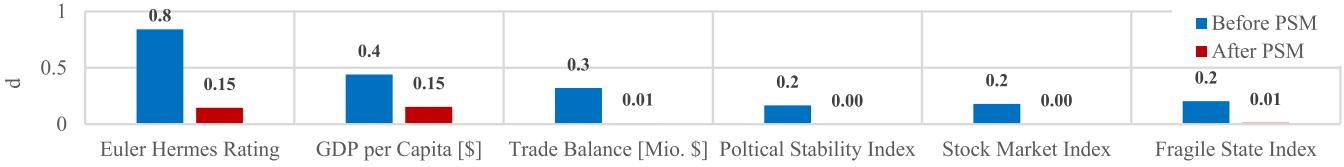


Fig. 9. Balance Assessment using Absolute Standardized Mean Difference d.

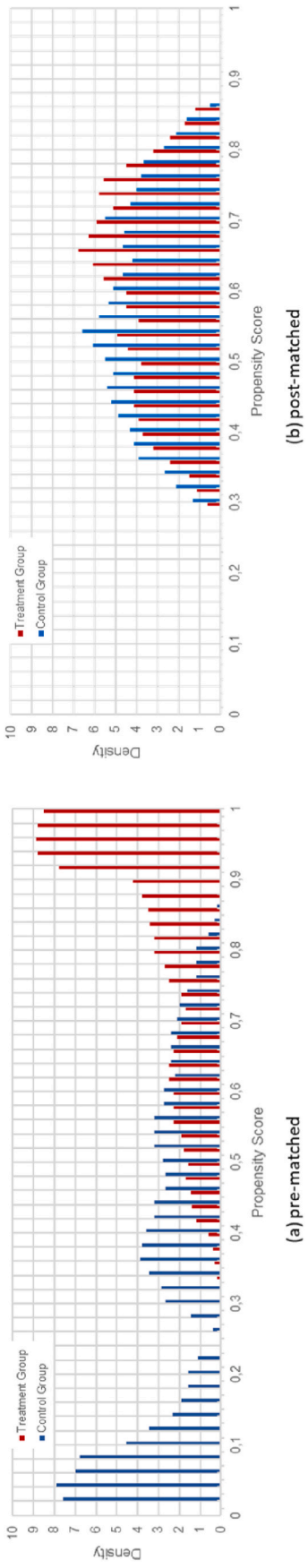


Fig. 10. Pre-matched (a) and post-matched (b) Propensity score distribution.

as the overall state of economic globalization.

5.5.3. Conclusions based on calculated ATEs

- "GDP per Capita": The supplier's delivery performance is affected by the GDP per capita of the respective state. Below a threshold of 45.000$, a delivery reliability decrease of 37.9% can be expected.
- "Inflation Forecast": The inflation forecasts of the respective supplier countries, which point to currency devaluations, and the average earnings difference directly influence delivery performance. For forecasts greater than 1%, delivery reliabilities of affected suppliers on average decrease by 51.6%.
- "Economic Globalization Index": The globalization of the national economy benefits the delivery performance of local suppliers. From an index value of 87 on, indicating low physical trade and capital transfer constraints, an average improvement in delivery reliability of 36.5% can be expected.

Focusing on the fourth subunit indicating the social development in the supplier's country, the poverty ratio, the unemployment rate as well as the human development index, quantifying the resident standard of living, are considered. In this context, the following causal relationship should be highlighted:

- "Human Development Index": For suppliers located in countries with a standard of living of less than 0.9 (measured by the human development index), there is on average a 53.9% increased risk of supply disruptions. Consequently, procurement relationships with such suppliers are to be avoided or comprehensively monitored.

Finally, we look at the causal effects of characteristics that can be assigned to the subunit "Terrorism & Crime Rate". Here, three key attributes are investigated: the "Control of Corruption" index, indicating the extent of corruption as well as public power abuse within the supplier's country, the "Fragile State Index" evaluating the country's vulnerability especially in terms of conflict situations and violent outbreaks, and the "External Interventions Index", quantifying the impact of external actors on the country's operations. Based on the observed causal effects, we can draw the following conclusions:

- "Control of Corruption Index": A lack of corruption prevention measures has a negative impact on the delivery reliability of local suppliers. Below an index threshold of 1.5, delivery reliability drops by an average of 34.9%. With regard to external interventions, however, no decisive causal effects on the target variable can be observed.
- "Fragile State Index": Suppliers located in countries with a high degree of fragility, characterized by complaints about violations of human rights, the occurrence of outbreaks of violence, and the effectiveness of the security apparatus (Neven, 2021), show poorer delivery performance. Above an index of 25, an average decrease in reliability of 46.3% must be expected.

Having presented the most essential causal effects within each feature group, we close this section by analyzing the overall supervised network structure. In Fig. 12 the solid lines represent the direct causal effects of features to the target variable "Delivery Reliability" for which the treatment effects have been calculated, whereas the dotted lines illustrate relations between features. The supervised network architecture is constructed from the dataset using the BNAN learning method.

Overall, the existence of several confounders influencing both treatment and outcome, such as the features "Country", "Fragile State Index", "Human Development Index", "GDP per Capita" or "Globalization Index", is observable. For Company A, it is therefore recommended to prioritize the analysis and interpretation of their treatment effects mainly because these confounding variables do not only map

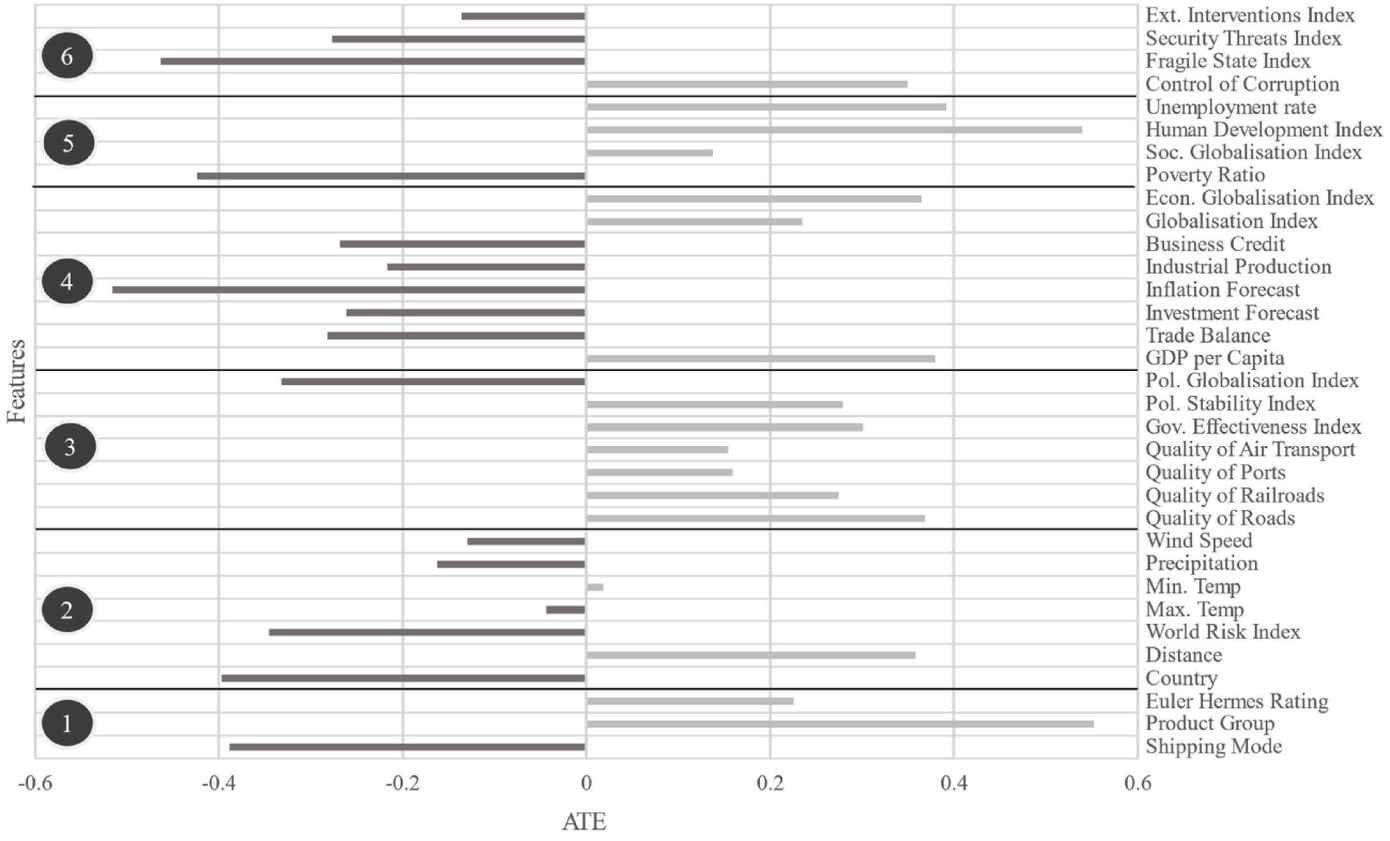


Fig. 11. ATEs for different feature groups.

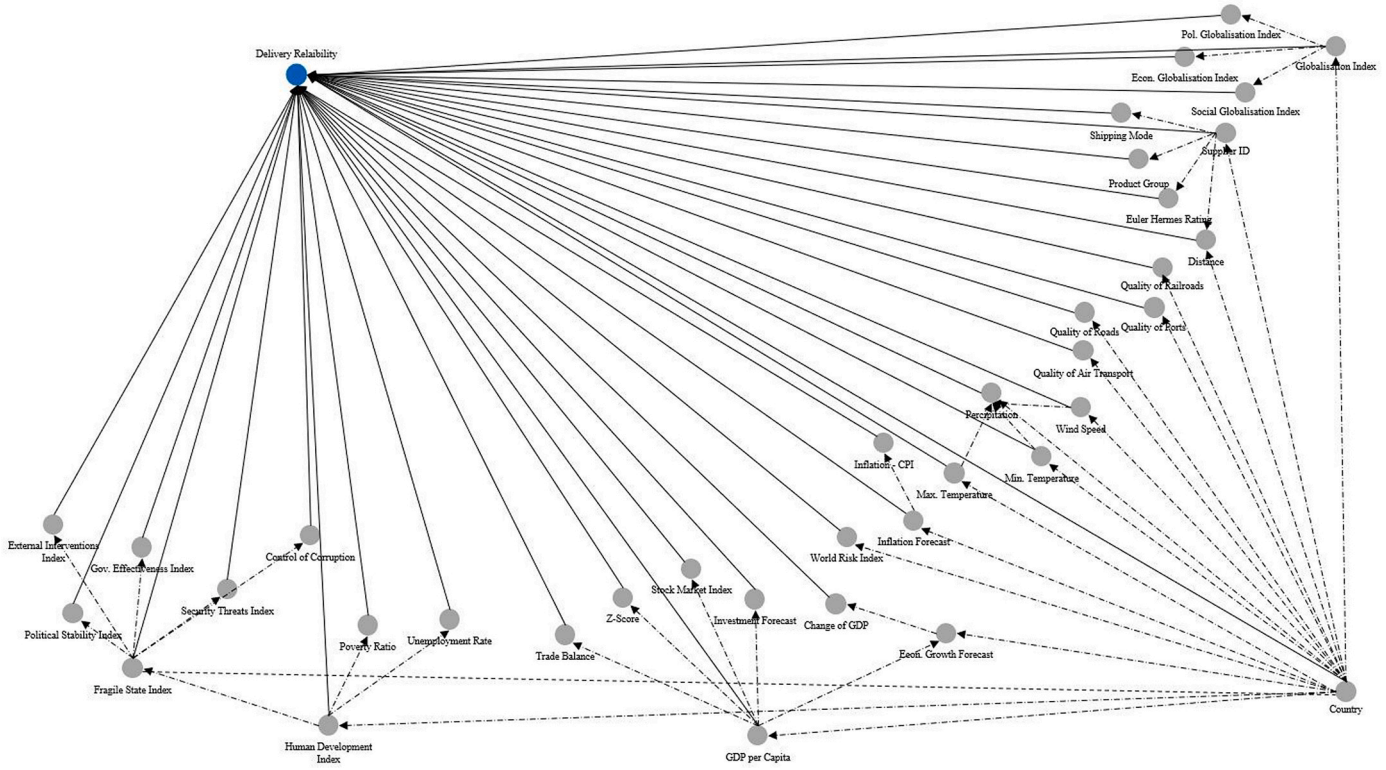


Fig. 12. Supervised network structure.

relationships to the target variable "Delivery Reliability". Specifically, this means that with respect to the confounder "Fragile State Index," the analysis of its treatment effect on the target variable as well as the derivation of risk mitigation measures must be prioritized over the analysis of the characteristics "Political Stability Index", "Government Effectiveness Index," or "Security Threats Index", since the characteristic "Fragile State Index" additionally influences their treatment scenarios on the target variable.

6. Discussion

The mixed methods approach, combining deep learning and causal inference, provides the opportunity to gain new insights from "Big Data" to analyze supply disruptions. The following sections critically reflect and discuss theoretical and managerial implications as well as future research opportunities.

6.1. Theoretical implications

Based on the findings, implications for research can be derived that contribute to and extend the current state of supply chain risk management theory. To systematically outline each contribution, we follow the five-step process by Manuj and Metzner (2008) for global supply chain risk management focusing on *Risk Identification*, *Risk Assessment & Evaluation*, *Selection of Risk Management Strategies*, *Implementation* and *Mitigation*.

Concerning *Risk Identification*, frameworks that classify supply chain risks into risk clusters are widely used. Christopher (2011) defines a general categorization into process, control, demand, supply, and environmental risks. As illustrated in Fig. 13, this research contributes to *risk identification* theory by increasing the level of detail of these frameworks from general categories (e.g., supply risks) to specific risk factors (e.g., supplier financial credibility) and providing evidence of causal relationships and impacts. Referring to the *Risk Assessment & Evaluation* strategies within supply chain risk management theory, the introduced 2 × 2 risk probability and impact matrix by Bhattacharya et al. (2009) is very useful, classifying risks from "low probability – low impact" to "high probability – high impact" designations. The presented causal deep learning model enables a more granular data-based categorization of risks, especially in terms of their impact. More specifically, the calculated ATEs enable a quantitative assessment of supply chain risks evoked by different feature groups. In selecting risk management strategies, this paper makes a threefold contribution to existing research efforts:

1. *Supply Risk Management:* Research efforts in procurement risk management have been focusing on performance and risk comparison of upstream sourcing strategies, mainly conducted in a mathematical and analytical manner. A recent example is the work of Li et al. (2022) who analyze sourcing decisions in competing multitier supply chains under basic conditions with practical relevance as well as generalizability not being proved. Using real-world data, our research model provides a remedy to the prevailing analytical studies, where risky single- or dual-sourcing relationships can be specifically identified. Strategic implications can be derived without having to make theoretical assumptions.
2. *Resilience Management*: In research on supply chain resilience preventive measures as well as rapid responses in case of disruptions take a central role. Looking at previously developed disruption management processes, such as the four-stage framework of Bode and Macdonald (2017), the stages of disruption detection and diagnosis are found to be the limiting factors for success in disruption preparation, recovery, and business continuity. Using the validity of causal relationships and monitored network structures, our research provides the ability to simulate potential disruption events embedded in what-if scenarios, revealing supply chain vulnerabilities.
3. *Supply Chain Complexity Management:* Research in the field of complexity management of supply chains aims at analyzing the relation between their complexity and performance characteristics such as responsiveness or transaction costs. Worth mentioning is the study of Ates et al. (2022) who perform a data-base meta-analysis to outline the effect of supply chain complexity on a firm's operational, innovation, and financial performance. Gaining insights into the causal relationships and effects, for example, between supplier location (geographic complexity) and delivery performance, and predicting possible outcomes can again facilitate the development of tailored risk mitigation strategies.

In addition to our contribution to supply chain risk management theory, we also promote analytical, data-driven research in OM. Traditionally, strategic-level models have been used to derive generalized results of how supply chains should operate (Misic and Perakis, 2020). However, using DNN architecture and processes helps building analytical capabilities in organizations, enabling high performance analysis. Moreover, with our presented data mining process, we enable the development of multi-criteria concepts for data-driven decision making (Ho et al., 2010), improving the transparency and interpretability for decision makers, capitalizing on causal inference and discovery theory.

Finally, in addition to the theoretical concept map illustrated in Fig. 13, we provide a remedy for the lack of real-world data insights emphasized by Barrat et al. (2011), when case studies are conducted. Based on the developed data mining and feature selection procedures and accompanied by the structured embedded case study, we introduce a methodological protocol that provides sufficient details of the research design, data collection, and data analysis, increasing the tractability of the subsequently derived inferences.

6.2. Managerial implications

While optimizing the design of global supply chains and networks, supply chain managers face two fundamental challenges. On the one hand, many supply chain design principles that have become popular over the last thirty years prioritize the responsiveness and efficiency of supply chain operations above the minimization of disruption risks (Barrat et al., 2011). On the other hand, the digital transformation and the associated increasing availability of "Big Data" impacts the way of managerial decision making, requiring new tools and models applicable to compress data to practically relevant insights. The main message of this study for practice and management can be summarized as:

- *Exploit the power of predictive analytics:* Predictive analytics supports the extraction of information from large data sets. Achieving forecast accuracies of 90% and more, deep learning models offer the possibility to practitioners to recognize potential supply disruptions within a certain lead time, whereby the scope for defining and preparing compensatory measures can be significantly increased.
- *Identify and manage high-risk sourcing relationships*: In today's supply chain management and mitigation processes established in globally operating firms, risk identification as well as risk assessment take on a central role. Identifying causal relationships between supplier characteristics and the expectable delivery performance and

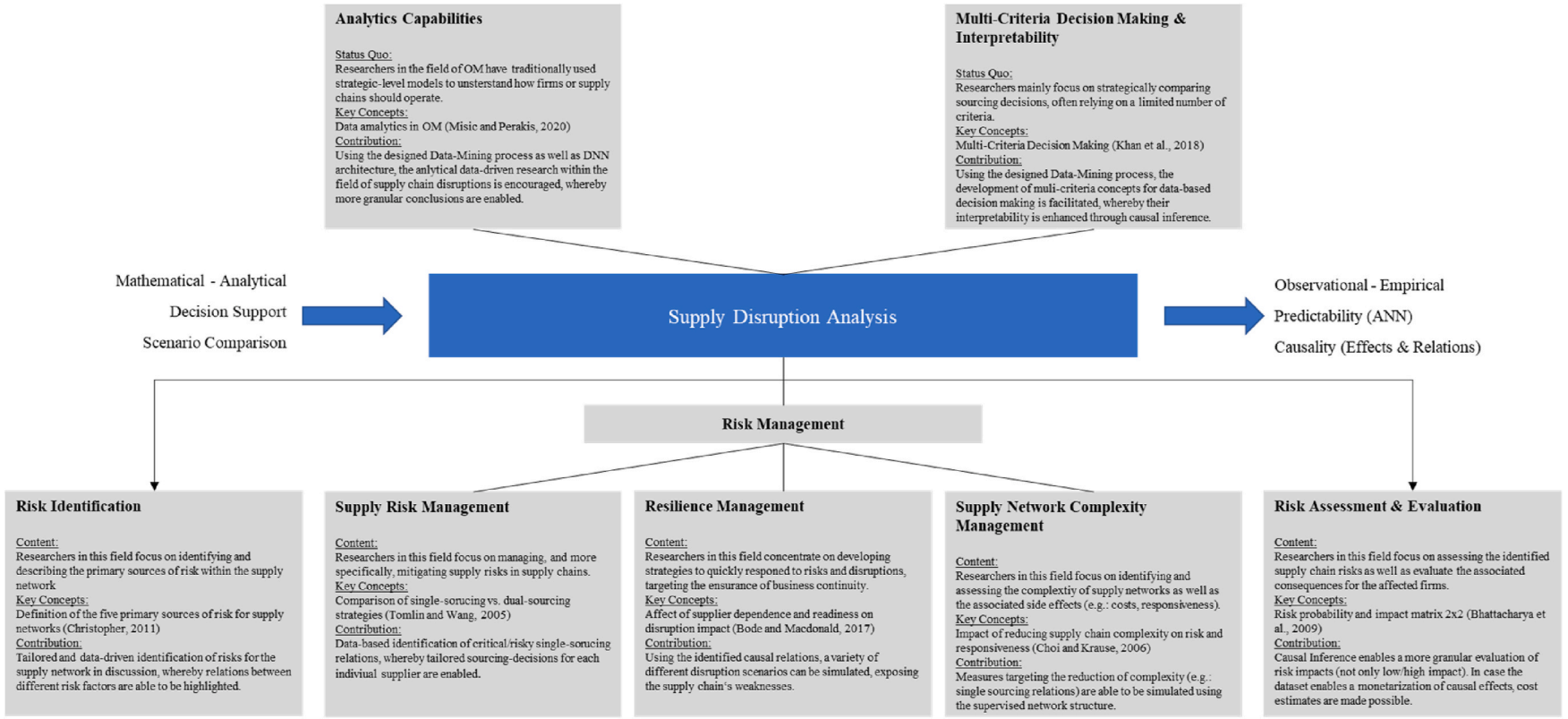


Fig. 13. Theory concept map for supply disruption analysis.

evaluating their effects using causal inference offers the potential to highlight questionable supply relationships and consequently concentrate activities. To illustrate this by an example, we refer to the detected causal relation between the supplier's financial credibility (Euler Hermes Rating) and its delivery performance. Using the gained insight from our embedded case study, company A is now able to *identify* risky suppliers as well as *assess* the risk extent. Having identified these crucial suppliers, risk management strategies such as hedging can be applied in a targeted and selective manner leading to a more efficient application of intracompany resources.

The identification of causal relationships between supplier characteristics and the expected delivery performance as well as the evaluation of their effect by means of causal inference offer the potential to identify lacking supplier relationships and consequently bundle supply chain risk management related activities.

However, data is key to the predictive analysis of supply disruptions. Therefore, business departments are challenged to establish a data-driven and data-mediated culture that improves intra-enterprise data sharing across different interfaces (Dremel et al., 2017). Managers mostly focus on the use of internal company data while many business analytics systems move towards the utilization of external information. As this study demonstrates, the great importance of external indicators like country-specific climate and financial conditions, the inclusion of publicly available international databases must additionally become established in the company's internal specialist departments.

6.3. Limitations and future research avenues

Analyzing supply disruptions using deep learning and causality-based learning poses challenges that should be addressed in future work.

First, the feature selection step results from extensive literature reviews and practical findings. However, for similar studies, a different set of features might be useful depending on the operational circumstances of the utility network under study. Especially for the downstream of supply chains, this framework needs further extensions. Future research should incrementally extend this existing framework, defining a standardized collection of features necessary for analyzing the causal risk management of supply chains.

Second, the effective combination of deep and causal inferences relies on some assumptions. It should be emphasized that the tuning of the DNN hyperparameters in this study is performed using the RandomizedSearchCV approach. For similar studies, we encourage the use of other hyperparameter tuning approaches, for example, the use of optimization libraries (Géron, 2019) to design neural networks that are well suited for the data analysis task at hand. In this context, it could be interesting to compare different hyperparameter optimization methods in order to analyze their impact on the model's prediction.

In addition, we highlight the capabilities of propensity score matching to gain crucial insights into the dataset. However, the PSM method is subject to certain limitations that need to be tackled in future research. First, propensity scores are only as unbiased as the features included in the dataset. Omitting important predictors could lead to biased results caused by unmeasured confounders. On the other hand, excluding unmatched individual cases that could form groups of significant size depending on the treatment scenario limits generalizability and reduces power (Kim et al., 2016; Beal and Kupzyk, 2014). Future research should therefore aim not only at applying and comparing different propensity score-based methods, such as stratification based on propensity scores, inverse probability weighting of treatment or the covariate adjustment approach, but also at applying propensity score-independent methods, such as the regression discontinuity design or the instrumental variable method.

Third, the data mining process for causality-based data analysis is applied only in the context of a single embedded case study. Although VossTsikriktsisFröhlich (2002) and Stuart et al. (2002) argue that single case studies are better suited to evaluate observations or methods in depth, for theory building and verification we suggest applying the defined data mining processes in further case studies to enable more robust and reliable conclusions about its performance and applicability in causality-based contexts (Eisenhardt and Graebner, 2007). There, its strengths and weaknesses in comparison to the already existing data mining approaches CRISP-DM as well as SEMMA should be investigated.

Furthermore, special attention should be paid to the subdomain of causal discovery learning due to its ability to vividly visualize complex causal relationships between features. Since the embedded case study is conducted in the automotive industry, where supply chains are characterized by a high degree of complexity due to component diversity as well as individual procurement characteristics (e.g., JIT delivery commitments), the application of the defined data mining process in other industries should be evaluated to prove its cross-industry validity.

7. Conclusion

Overall, this study illustrates the power and validity of a mixed methods and analytical modeling case study research in operations management (Choi et al., 2016) and more specifically in the context of supply disruption analysis. As a bottom line, the introduced business analytics model which incorporates both deep learning and causal inference techniques, excels in predictive performance and finds associations between supplier characteristics and supply disruption frequency, based on internal and external real-world enterprise data. In terms of practicality and relevance, the uncovered causal relationships of supply disruptions consequently assist managers and practitioners in prioritizing and executing operational and strategic decisions to optimize supply chains.

Declaration of competing interest

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

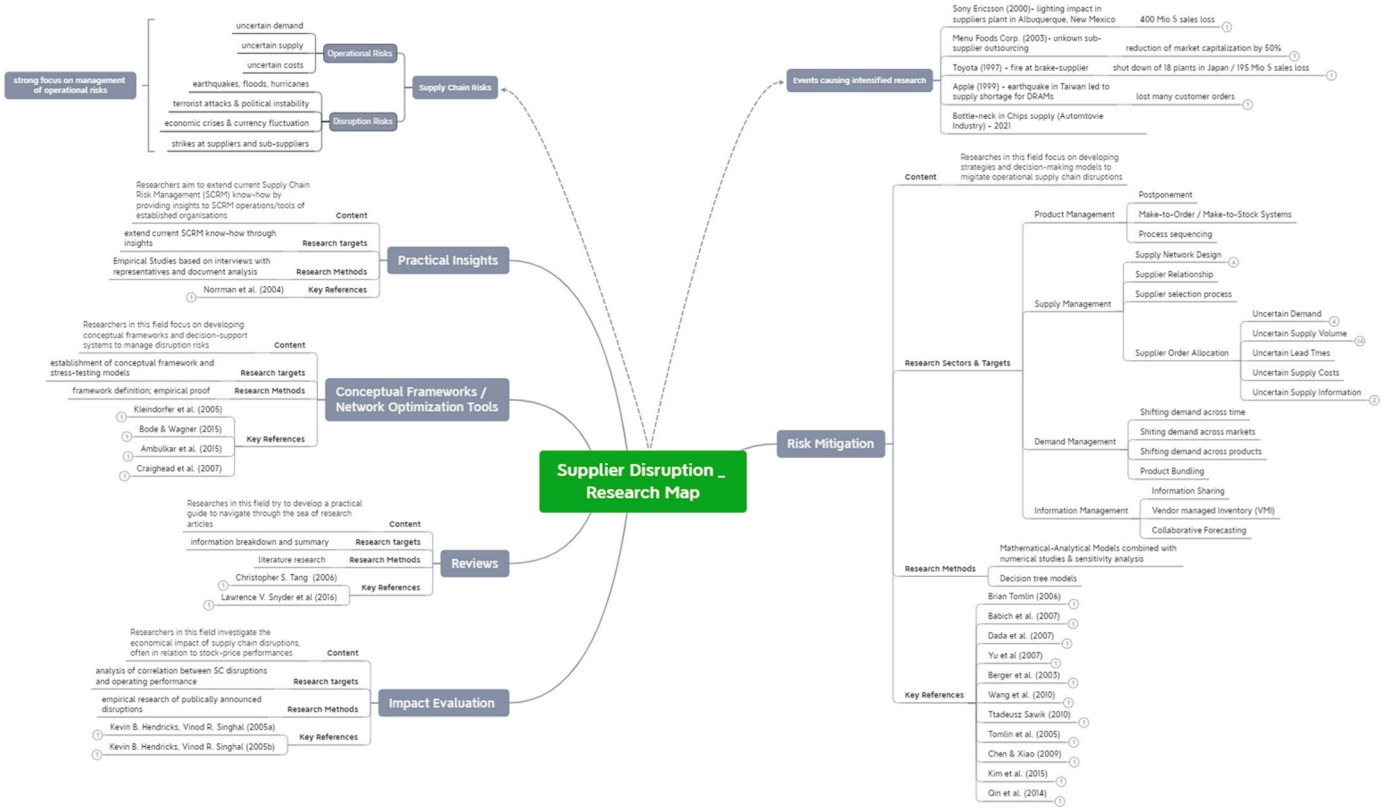
Data availability

Research data is available in the manuscript and online appendix

Appendix A. Characteristics of Most Cited References

#	Research Method	Data Analysis Method	Covered Topics	Future Research Questions	Limitations	Author
1	Analytical mathematical	–	Optimal sourcing strategy between reliable & unreliable supplier	Quantity flexible contracts vs. demand uncertainty	restricted to a specific customer-supplier scenario with limited practical transferability	Tomlin (2006)
2	Analytical mathematical	–	Influence of supplier default correlations on wholesale prices	Consideration of fluctuating lead times	restricted to a specific customer-supplier scenario with limited practical transferability	Babich et al. (2007)
3	Literature review		Review on perspectives in supply chain risk management	Non-stationary demand and supply process	–	Tang C. (2006)
4	Analytical mathematical	–	Formulation of newsvendor decision model (reliable/unreliable)	Pay-per-amount payment models	resticted analytical decision model focusing solely on operational risks	Dada et al. (2007)
5	Empirical database	Regression/ Correlation	Conceptual framework reflecting risk assessment & mitigation	Development of supply chain disruption indicators	limited number of risk factors considered for decisions on mitigation strategies	Kleindorfer and Saad (2005)
6	Literature review		Review of OR/MS models on supply chain disruptions	Combination of pro- and reactive mitigation strategies	–	Snyder L. et al. (2016)
7	Analytical mathematical	–	Development of two Supply Chain coordination models in the context of demand disruption	Consideration of one dominant retailer, acting as a monopolist	limited number of suppliers considered in decision making scenario	Chen and Xiao (2009)
8	Analytical conceptual		Comparison of four fundamental supply network structures to help understand supply disruptions	Equal probability of failure for every node and arc	–	Kim et al. (2015)
9	Analytical mathematical	–	Comparison of mutliple sourcing vs. single sourcing including supplier improvement	Supplier competition benefits of dual sourcing	focused solely on mitigation of operational risks such as supply uncertainties	Wang et al. (2010)
10	Analytical mathematical	–	Evaluation of supply disruption impacts on the choice between single and dual sourcing methods	Consideration of suppliers with limited capacity	focused solely on mitigation of operational risks such as supply uncertainties	Yu et al. (2008)
11	Empirical database	Means testing	Investigation of supply chain characteristics increasing the frequency of supply disruptions	Focus on upstream (supply-side) supply chain disruptions	stable supply chain structure assumed to simplify simulations	Bode and Wagner (2015)
12	Empirical database	Means testing	Investigation of long-term stock price effects caused by supply chain disruptions	Develop supply chain responsiveness indicators	limitations with respect to empirical sample group	Hendricks and Singhai (2005a)
13	Analytical mathematical	–	Determination of the optimal number of suppliers in the presence of risk through decision-tree models	Focus on “super effects”, affecting all global suppliers	limited number of risk factors considered for decisions on mitigation strategies	Berger et al. (2004)
14	Empirical survey	Path analysis	Investigation of factors contributing of a firm’s resilience to supply chain disruptions	–	cross-sectional data used limiting conclusions in relation to causality	Ambulkar et al. (2015)
15	Empirical database	Means testing	Investigation of association between supply chain glitches and operating performance	Development of methologies to predict SC glitches	methods to forecast operational & disruption risks not considered	Hendricks and Singhai (2005b)
16	Empirical case study	None	Analysis of executed supply chain risk management strategies in well-kown operation	–	analysis restricted to one specific enterprise	Norrman and Jansson (2004)
17	Analytical mathematical	–	Optimal selection of supply portfolio in the context of local and global disruptions	Consideration of multi-period supplier selection and order allocation	lack of practical evidence	Sawik (2011)
18	Analytical mathematical	–	Comparison of single-source dedicated strategies and single-source flexible strategies	Consideration of economies of scale and coordination costs	order allocation models based on significant restrictions of reality	Tomlin and Wang (2005)
19	Empirical survey	None	Ability of business continuity programs to help limit the damage caused by SC disruptions	Expansion to further empirical studies	no strong evidence for causailty	Azadegan et al. (2020)
20	Literature review	–	Ability of OR methods for coping with the ripplie effect in supply chain risk management scenarios	Evaluation of intertwined networks and structural dynamics	AI methods neglected	Ivanov and Dogui (2021)

Appendix B. Supplier Disruption Research Map



Appendix C. Overview on Causal Inference Terminology and Theory

Terminology	Alternatives	Explanation
causality	causal relationship, causation	causal relationship between variables
causal effects		strength of causal relationships
instance	unit, sample, example, individual	independent unit of the population
features	covariates, observables, pre-treatment variables	Variables describing instances
learning causal effects	forward causal inference, forward causal reasoning	identification and estimation of causal effects
learning causal relationships	causal discovery, causal learning, causal search	inferring causal graphs from data
causal graphs	causal diagram	a graph with variables as nodes and causality as edges
confounder	confounding variable	a variable causally influencing both treatment and outcome

Background Knowledge

To enable a deep dive into the structural causal model Framework (SCMF) related theory, we need to align the necessary notations and terminologies used in this paper. Lowercase letters like x are used to highlight random variables, whereas bold lowercase letters like $\mathbf{x}$ denote vectors and bold uppercase letters like $\mathbf{X}$ symbolize matrices. x_i thereby signifies features for the ith instance. In relation to causal graphs, calligraphic uppercase letters can either represent a set of nodes $\mathcal{V}$ or a set of edges $\mathcal{E}$. The ancestors of a node x in a graph $G = (\mathcal{V}, \mathcal{E})$ with $x \in \mathcal{V}$ is denoted by $An_G(x)$. Parents are accordingly illustrated by $Pa_G(x)$. The letter t is used to denote the treatment variable, whereas the outcome is represented by the letter y. In that respect, y_i^1 signifies the outcome when the instance i is treated with $t_i = 1$. The treatment effect, meaning a change in the outcome variable for different levels of treatment, is expressed by the letter τ. P_x signifies the probability distribution of x. The notation $x \perp\!\!\!\perp y$ represents the independence of x and y, whereas $x \perp\!\!\!\perp y|z$ denotes the conditional independence of x and y given z. In this context, the conditional expectation of y given x is expressed by the term $E(y|x)$. In addition to the above presented notations, listed terminologies in Appendix D will support the explanation of SCMF and causality-based learning mechanism.

Structural Causal Model Framework

SCMF builds on two pillars: causal graphs and structural equations. Based on the definition of Pearl et al. (2016), graphical models in the context of the SCMF consist of a collection of nodes Υ and edges E, targeting the representation of stochastic dependencies among a quantity of random variables (Drton and Maathius, 2017). The model's random variables, including treatment, outcome, observed and unobserved variables, are illustrated by the set of nodes, whereas the set of edges visualize specific relationships between those connected variables. In case of a connection of two variables by an edge, these variables are called *adjacent*. In Fig. 2b x_1 and x_2 are adjacent, whereas x_1 and x_3 do not satisfy the therefore necessary requirement. Edges in a graph can further be characterized by whether they are *directed* or *undirected*. Referring to Fig. 2a (Pearl et al., 2016), all edges within the graph are undirected as they do not display a single arrowhead on their endpoint marks. Consequently, the overall graph can be labelled as an *undirected graph* or skeleton of a graph.

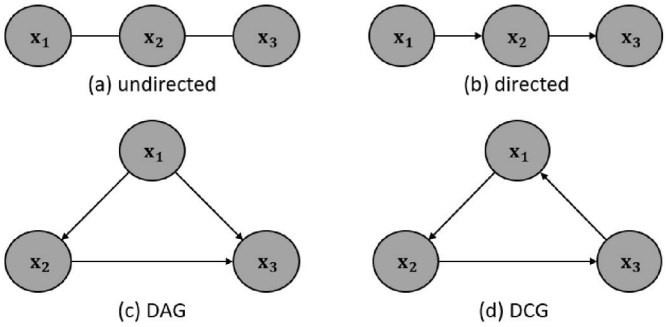


Fig. 2. Directed and undirected graphs

In contrast, a graph where all edges possess arrowheads is referred to as a *directed graph* (Roy,2021).

For *directed graphs*, a differentiation between Directed Acyclical Graphs (DAG) and Directed Cyclical Graphs (DCG) needs to be made (see Fig. 2c and d). In case a causal graph displays no undirected paths and no cyclical connections, the requirements of a DAG are satisfied. DCGs in contrast contain directed paths from one node back to itself, which lead to cyclical connections.

Depending on the application, the meaning of an edge between two variables may differ to a certain extent. In this paper, we refer to the definition of Pearl (2009), where an edge connecting x_1 and x_2, $x_1 \rightarrow x_2$, symbolizes a causal effect of x_1 on x_2. Therefore, a causal graph represents a special class of Bayesian networks and consequently enables the application of the conditional independence criteria (Guo et al., 2020). To understand the theory of conditional independence in connection with causal graphs, we briefly outline the concept of d-separation (Pearl, 1986) based on the definition of *blocked paths* in the following.

In general, conditioning on a set of nodes $S \subset \Upsilon$ will cause a path p to be blocked, if at least one of the nodes in p is blocked. Thereby, the process of conditioning can be referred to as knowing the value of a certain variable and blocking means that the flow of information or the dependency between the variables connected by a path is stopped. In the causal chain and in the causal fork, displayed in Fig. 2 (Guo et al., 2020), both variables x_1 and x_2 marginally depend on each other.

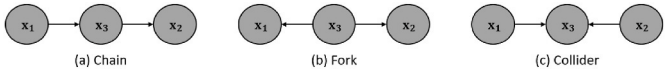


Fig. 3. Typical DAGs for conditional independence

In the chain x_1 has a causal effect on x_2 through its influence on x_3, whereas x_3 is the common cause of x_1 and x_2 in a fork. Overall, in both graphical models x_1 is associated with x_2 but a causation is not existent. The process of conditioning on x_3 in this case will lead to an independence between x_1 and x_2, in other words blocking the *path* and therefore the information flow between x_1 and x_2: x_3 *d-separates* x_1 from x_2, which can be mathematically displayed by $x_1 \perp\!\!\!\perp x_2 | x_3$. In contrast to chains and forks, colliders already block paths without the necessity of conditioning on a specific variable. Consequently, the variables x_1 and x_2 are (marginally) independent without the process of conditioning on x_3. However, if the common effect x_3 or *descendants* of x_3 are conditioned on, the path between x_1 and x_2 becomes unblocked leading to their mutual dependence, which can be represented by $x_1 \perp\!\!\!\perp x_2, x_1 \not\perp\!\!\!\perp x_2 | x_3$.

Overall, the following conclusion regarding *d-separation*, which plays a crucial role in explaining causal concepts, can be drawn. A *path* is *d-separated* or *blocked* by a set of nodes S_x, if:

- the path contains a chain $x_1 \rightarrow x_3 \rightarrow x_2$ or a fork $x_1 \leftarrow x_3 \rightarrow x_2$ such that the middle node x_3 is in S_x or,
- the path contains a collider $x_1 \rightarrow x_3 \leftarrow x_2$ such that the middle node x_3 as well as its descendants is not in S_x

The concept of d-separation connected with conditional independencies is highly correlated with the causal Markovian condition: every node and therefore every variable in a directed graph is independent of its nondescendents, conditioning on its parents. Hereby it is necessary to address the subject of confounding bias. Confounding can be defined as the distortion of an association between two variables (a treatment and its outcome) by a third often unmeasured variable, also referred to as confounder or latent variables, which leads to an over- or under-estimation of the observed association.

Structural Equation Models

In relation to the causal graph under intervention (manuscript Fig. 2b), the interventional distribution or post-intervention distribution $P(y|do(t^{'}))$ characterizes the transport volume increase distribution y, in case the infrastructure t was set to the value $t^{'}$ and all edges towards t were eliminated by intervention. In consequence of the denoted intervention, the set of structural equations associated with the causal graph in manuscript's Fig. 2b can be written as: $x = f_x(E_x), t = t^{'}, y = f_y(x,t^{'},E_y), P(y|do(t^{'})) = f_y(x,t^{'},E_y)$. Using the language of SCMF, we have the possibility to translate the problem of calculating causal effects into inquiries about the interventional distribution $P(y|do(t^{'}))$ with $t^{'}$ different values, assuming equal causal relations for each instance. Based on this assumption, SCMFs additionally enable the definition of an Average Treatment Effect (ATE), which can be defined as following in this paper:

$$ATE(t_+, t_-) = E[y|do(t_+)] - E[y|do(t_-)], t_+ > t_-$$

With:

ATE Average Treatment Effect
E conditional expectation of y given $do(t_+)$ or $do(t_-)$
y outcome variable (e. g., weight gain)
t_+, t_- subjects eat t_+ or forego t_- eating chocolate

Causal Inference

Finally, we briefly review an excerpt of algorithms applicable to learn causal effects, which are still an area of active research within the OM and ML community (Misic and Perakis, 2020).

Learning causal effects, referred to as *Causal Inference*, is concerned with quantifying an expected change of a defined outcome variable y in case a modification of the treatment variable t is executed (Guo et al., 2020). Depending on the research problem, an investigation of the causal effect for the whole population, for different subpopulations defined by same feature values or for unknown subpopulations is targeted. The most common treatment effect is the introduced Average Treatment Effect (ATE). However, if the observed population consists of multiple heterogeneous groups or subpopulations, the ATE can be misleading, causing the so-called spurious effect by including the impact of confounders on the outcome besides the effect of treatment (Yao et al., 2020). Consequently, the average should be defined for each individual and homogeneous group, leading to the Conditional Average Treatment Effect (CATE), which maps features defining subpopulations to its estimated ATE.

Causality-based learning techniques targeting the investigation of causal effects can be categorized into approaches with and without unobserved confounders. When formulating the assumption that all confounders are among the observed variables, conditioning, or adjustment on a set of variables S_x blocking all back-door paths is sufficient to solve the problem of learning causal effects. According to Guo et al. (2020), two different adjustment approaches are applicable in the context of learning causal effects: (1) regression adjustment and (2) propensity score methods.

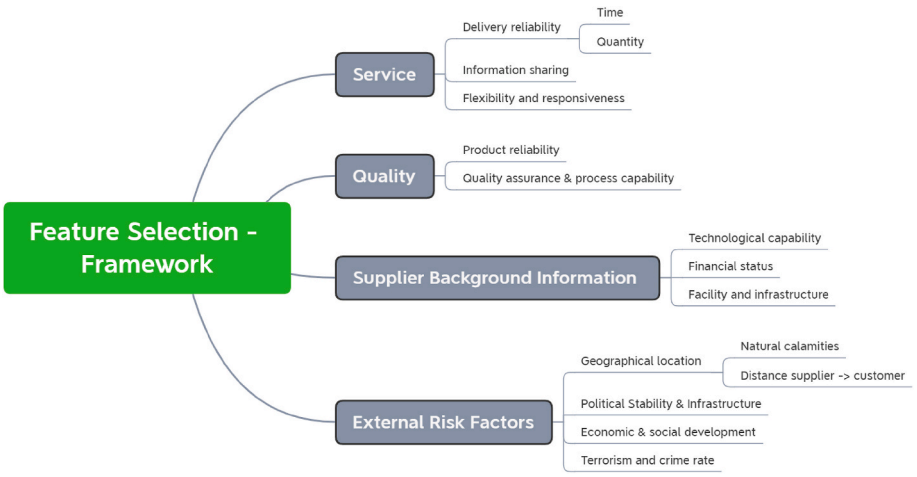
Starting with the regression adjustment method, researchers fit a function to estimate the probability distribution of y adjusting on certain confounders x and the treatment variable t. Practically, two types of regression adjustment methods need to be distinguished. On the one hand, single functions are applied to rate the probability distribution $P(y|x,t)$. On the other hand, it is additionally feasible to use two separate functions for the estimation of two potential outcomes e. g. $P(y|x,t=0)$ and $P(y|x,t=1)$, forcing the calculation of the ATE afterwards.

Continuing with the second adjustment approach, the propensity score represents a balancing score displaying the "probability of treatment assignment conditional on observed baseline covariates […]" (Austin, 2011). By dividing instances into strata and treating each stratum as a randomized control trial (Morgan and Winship, 2007), the effects of confounding on the causal effect between treatment and outcome are eliminated as well as potential upcoming sparse data problems. Building on the distribution, the ATE is estimated for each individual stratum. According to Austin (2011), four types of propensity score methods can be classified: (1) *propensity score matching*, (2) *propensity score stratification*, (3) *inverse probability treatment weighting* and (4) *covariate adjustment*.

Propensity Score Matching (PSM) matches a treated instance to a set of treated instances with similar propensity scores, mainly using the Greedy One-to-One as well as the Nearest Neighbor matching approaches (Guo et al., 2020). In that regard, p*ropensity score stratification* represents a natural extension of the PSM methods by stratifying instances into mutually exclusive subsets based on their defined propensity scores. The third method of *inverse probability treatment weighting* (IPTW) is introduced to solve the issue of data sparseness in feature subsets. Its intuition is to create a synthetic randomized control trial by weighting the instances with their inverse probability and consequently realizing an independency between covariate distribution and assigned treatment. The fourth and final approach of *covariate adjustment* regresses the outcome variable on the treatment variable as well as the estimated propensity score (Austin, 2011).

Having discussed multiple learning methods in the section above, it is essential to notice that for many real-world problems of learning causal effects, the fundamental assumption of no unobserved confounders is not satisfied and applicable. Therefore, further learning methods such as the *instrumental variable* (IV) method, the *front-door criterion* as well as the regression discontinuity design need to be considered.

Appendix D. Feature Selection Framework - Supply Disruptions



Appendix E. Feature Overview

Feature	Data Type
Supplier (short)	int64
Country	object
City	object
Product group	object
Day	int64
Month	int64
Year	int64
Timeliness	float64
Quantity loyalty	float64
Shipping mode	object
Euler hermes rating	int64
Distance	float64
Change of GDP	float64
GDP per capita	float64
Inflation - consumer price index	float64
Unemployment rate	float64
Trade balance	float64
Economic growth forecast	float64
Investment forecast	float64
Inflation forecast	float64
Government effectiveness index	float64
Control of corruption	float64
Political stability index	float64
Quality of roads	float64
Quality of railroad infrastructure	float64
Quality of port infrastructure	float64
Quality of air transport infrastructure	float64
Banking system Z-score	float64
Poverty ratio (<5.50 USD)	float64
Globalization index	float64
Economic globalization index	float64
Political globalization index	float64
Social globalization index	float64
Fragile state index	float64
Security threats index	float64
Economic decline index	float64
External interventions index	float64
Human development index	float64
Quarterly econimic growth	float64
Industrial production (annual change)	float64
Purchasing managers index	float64
Investment (% of GDP)	float64
Inflation (monthly)	float64
Business credit	float64
Exchange rate (USD)	float64
Unemployment rate (quarterly)	float64
Stock market index	float64

(*continued on next page*)

(*continued*)

World risk index	float64
Max. temperature	float64
Min. temperature	float64
Precipitation	float64
Snow	int64
Wind speed	float64
Conditions	int64
Delivery reliability	bool

Appendix F. Boxplots

Appendix G. – Management Implications through ATE

Instance	Treatment	$E[Y_0]$	$E[Y_1]$	ATE	Management ImplicationsImplications
1.	Country = Germany	0.748	0.351	**−0.397**	higher delivery reliability of foreign suppliers
2.	Shipping Mode = Truck	0.623	0.234	**−0.389**	Truck deliveries require extensive monitoring
3.	Distance >200 km	0.315	0.673	**0.358**	Reliability of close suppliers (<200 km) improvable
4.	Product Group = Raw Material	0.213	0.765	**0.552**	Reliability of non-raw material suppliers improvable
5.	Euler Hermes Rating >3	0.278	0.503	**0.225**	Financial well-being benefits delivery performance
6.	Change of GDP >1.3%	0.389	0.538	**0.149**	Domestic economic power requires to be monitored
7.	GDP per Capita >45.000 $	0.209	0.588	**0.379**	Domestic economic power requires to be monitored
8.	Trade Balance >230 [Bio $]	0.498	0.216	**−0.282**	Export orientation compromised delivery reliability
9.	Economic Growth Forecast >0%	0.744	0.225	**−0.519**	Lacking delivery performance of developing states
10.	Investment Forecast >21%	0.483	0.222	**−0.261**	Lacking delivery performance of developing states
11.	Inflation Forecast >1%	0.739	0.223	**−0.516**	Domestic financial framework as major influencing factor
12.	Industrial Production >1.7	0.457	0.24	**−0.217**	industrial production growth indicates underdevelopment
13.	Purchasing Managers Index >52	0.361	0.356	**−0.005**	no significant expressive power
14.	Investment >21.55% of GDP	0.311	0.578	**0.267**	Lacking delivery performance of developing states
15.	Business Credit >1040	0.435	0.167	**−0.268**	Avoid relations to suppliers with exuberant business credits
16.	Exchange Rate (USD) > 0.9	0.241	0.453	**0.212**	Weak local currencies benefits trade and reliability
17.	Stock Market Index >105	0.516	0.295	**−0.221**	Domestically strong stock markets compromise reliability
18.	Z-score < 26	0.727	0.342	**−0.385**	Buffer of domestic banking systems requires monitoring
19.	Quality of Roads >5	0.216	0.584	**0.368**	Domestic road networks require detailed investigations
20.	Quality of Railroads >5	0.213	0.487	**0.274**	Domestic rail networks require detailed investigations
21.	Quality of Ports >5	0.238	0.397	**0.159**	Domestic port networks require detailed investigations
22.	Quality of Air Transport >5	0.231	0.385	**0.154**	Domestic airport networks require detailed investigations
23.	Inflation - CPI >1.5%	0.593	0.207	**−0.386**	Domestic financial framework as major influencing factor
24.	GEI >1.6	0.204	0.505	**0.301**	Political independency necessary to monitored
25.	Control of Corruption >1.5	0.136	0.485	**0.349**	Political independency necessary to monitored
26.	Political Stability Index >0.6	0.208	0.487	**0.279**	Strong political environment required
27.	Poverty Ratio (<5.5 $) > 3	0.678	0.254	**−0.424**	Domestic living standard as additional pivotal indicator
28.	Globalization Index >90	0.243	0.478	**0.235**	Non-globalized states required to be avoided
29.	Econ. Globalization Index >87	0.226	0.591	**0.365**	Economically globalized mindset benefits reliability
30.	Pol. Globalization Index >87	0.676	0.344	**−0.332**	Politically globalized structure compromises reliability
31.	Soc. Globalization Index >87	0.256	0.394	**0.138**	Socially globalized mindset benefits reliability
32.	Fragile State Index >32	0.585	0.122	**−0.463**	Relations to suppliers in fragile states require to be avoided
33.	Security Threats Index >2.5	0.479	0.202	**−0.277**	Avoid relations to suppliers in potentially terrorized states
34.	Econ. Decline Index >1.8	0.591	0.216	**−0.375**	Domestic economic framework as major influencing factor
35.	Ext. Interventions Index >0.75	0.381	0.245	**−0.136**	Avoid relations to suppliers in potentially manipulated states
36.	Human Development Index >0.9	0.216	0.755	**0.539**	Standard of living necessary to be monitored
37.	Unemployment Rate <3.5	0.205	0.596	**0.391**	Increasing unemployment rate as indicator of disruptions
38.	World Risk Index >3	0.495	0.149	**−0.346**	Avoid relations to suppliers in natural hazardous regions
39.	Max Temperature >15 °C	0.361	0.317	**−0.044**	no significant expressive power
40.	Min Temperature <6 °C	0.341	0.359	**0.018**	no significant expressive power
41.	Precipitation >1.6 mm	0.496	0.334	**−0.162**	Domestic climatic conditions compromise reliability
42.	Wind Speed >20.7 km/h	0.364	0.234	**−0.13**	Domestic climatic conditions compromise reliability

References

Abadie, A., Imbens, G.W., 2016. Matching on the estimated propensity score. Econometrics 84 (2), 781–807.

Ambulkar, S., Blackhurst, J., Grawe, S., 2015. Firm's resilience to supply chain disruptions: Scale development and empirical examination. J. Oper. Manag. 33, 111–122.

Ates, M., et al., 2022. Order from chaos: a meta-analysis of supply chain complexity and firm performance. J. Supply Chain Manag. 58 (1), 3–30.

Austin, P.C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav. Res. 46 (3), 399–424.

Azadegan, A., et al., 2020. Supply chain disruptions and business continuity: an empirical assessment. Decis. Sci. J. 51 (1), 38–73.

Babich, V., Burnetas, A.N., Ritchken, P.H., 2007. Competition and diversification effects in supply chains with suppler default risks. Manuf. Serv. Oper. Manag. 9 (2), 123–146.

Baghersad, M., Zobel, C.W., 2021. Assessing the extended impacts of supply chain disruptions on firms: an empirical study. Int. J. Prod. Econ. 231, 107862.

Barrat, M., Choi, T.Y., Li, M., 2011. Qualitative Case Studies in Operations Management: Trends, Research Outcomes.

Beal, S.J., Kupzyk, K., 2014. An introduction to propensity scores: what, when, and how. J. Early Adolesc. 34 (1), 66–92.

Berger, P.D., Gerstenfeld, A., Zeng, A.Z., 2004. How many suppliers are best? A decision-analysis approach. Omega 32 (1), 9–15.

Bhattacharya, A., Geraghty, J., Young, P., 2009. On the analytical framework of resilient supply-chain network assessing excursion events. In: 2009 Third Asia International Conference on Modelling & Simulation. IEEE Computer Society, pp. 392–397.

Bode, C., Macdonald, J.R., 2017. Stages of supply chain disruption response: direct, constraining, and mediating factor for impact mitigation. Decis. Sci. J. 48 (5), 836–874.

Bode, C., Wagner, S.M., 2015. Structural drivers of upstream supply chain complexity and the frequency of supply chain disruptions. J. Oper. Manag. 36 (1), 215–228.

Bodendorf, F., Xie, Q., Merkl, P., Franke, J., 2022. A multi-perspective approach to support collaborative cost management in supplier-buyer dyads. Int. J. Prod. Econ. 245, 108380.

Buda, M., Maki, A., Mazurowksi, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network. 106 (10), 249–259.

Caiado, R., Scavarda, L., Gaviao, L., Ivson, P., Nascimento, D., Garza-Reyes, J., 2021. A fuzzy rule–based industry 4.0 maturity model for operations and supply chain management. Int. J. Prod. Econ. 231 (1).

Cannas, M., Arpino, B., 2019. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. Biom. J. 61 (4), 1049–1072.

Chan, F., Kumar, N., Tiwari, M., 2008. Global supplier selection: a fuzzy-AHP approach. Int. J. Prod. Res. 46 (4), 3825–3857.

Chandrasekaran, A., de Treville, S., Browning, T., 2020. Intervention-based research (IBR)—What, where, and how to use it in operations management. J. Oper. Manag. 66 (4), 370–378.

Chawla, N.V., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.

Chen, K., Xiao, T., 2009. Demand disruption and coordination of the supply chain with a dominant retailer. Eur. J. Oper. Res. 197 (1), 225–234.

Chen, C., Lin, C., Huang, S., 2005. A fuzzy approach for supplier evaluation and selection in supply chain management. Int. J. Prod. Econ. 102, 289–301.

Choi, T., Cheng, T.C.E., Zhao, X., 2016. Mutli–Methodological research in operations management. Prod. Oper. Manag. 25 (3), 379–389.

Chowdhury, M.M.H., Quaddus, M., 2017. Supply chain resilience: Conceptualization and scale development using dynamic capability theory. Int. J. Prod. Econ. 188, 185–204.

Christopher, M., 2011. Logistics and Supply Chain Management, fourth ed. Financial Times Series/Pearson Education Limited, Edinburgh Gate, UK.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Nel, P., Malhotra, S., 2018. Notes from the AI Frontier: Applications and Value of Deep Learning. McKinsey & Company. Available at: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning#part3.
Cox, V., 2017. Translating Statistics to Make Decisions. A Guide for the Non-Statistician. Apress, London, UK.
Craighead, C.W., Blackhurst, J., Rungtusanatham, M., Handfield, R., 2007. The severity of supply chain disruptions: design characteristics and mitigation capabilities. Decis. Sci. J. 38 (1), 131–156.
Crossing, Visual, 2021. Forecast & historical weather data. VisualCrossing. Available at: https://www.visualcrossing.com/weather/weather-data-services.
Dada, M., Petruzzi, N.C., Schwarz, L.B., 2007. A newsvendor's procurement problem when suppliers are unreliable. Manuf. Serv. Oper. Manag. 9 (1), 9–32.
Dremel, C., Herterich, M., Wulf, J., 2017. How AUDI AG estabished big data analytics in its digital transformation. MIS Q. Exec. 16 (2), 81–100.
Drton, M., Maathius, M.H., 2017. Structure learning in graphical modelling. Ann. Rev. Stat. Its Appl. 4, 365–393.
Eisenhardt, K.M., Graebner, M.E., 2007. Theory building from cases: opportunities and challenges. Acad. Manag. J. 50 (1), 25–32.
Fan, Y., Stevenson, M., 2007. A review of supply chain risk management: definition, theory, and research agenda. Int. J. Phys. Distrib. Logist. Manag. 48 (3), 205–230.
Figueroa, R.L., Zeng-Treitler, Q., Kandular, S., Ngo, L., 2012. Predicting sample size required for classification performance. BMC Med. Inf. Decis. Making 12 (8), 1–10.
Géron, A., 2019. HandsOn Machine Learning with ScikitLearn, Keras & TensorFlow. O'Reilly Media, Sebastopol, CA.
Gibbert, M., Ruigrok, W., 2010. The "what''and "how''of case study rigor: three strategies based on published work. Organ. Res. Methods 13 (4), 710–737.
Goodfellow, I., 2017. Deep Learning. MIT Press, Cambridge, MA.
Gunning, D., Sefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G., 2019. XAI—explainable artificial intelligence. Sci. Robotics 4 (37).
Guo, R., Cheng, L., Li, J., Hahn, P., Liu, H., 2020. A Survey of learning causality with data: problems and methods. ACM Comput. Surv. 53 (4).
Gupta, R., 2019. An introduction for discretization techniques for data scientists. Feature engineering: 4 discretization techniques to learn. Medium – Towards Data Sci. URL: https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2.
Hao, W., Fan, J., 2019. Review on evaluation criteria of machine learning based on big data. In: Journal Of Physics: Conference Series - 4th International Seminar On Computer Technology, Mechanical And Electrical Engineering. Chengdu, 13 December to 15 December 2019.
Hendricks, K.B., Singhai, V.R., 2005a. An empirical analysis of the effect of supply chain disruptions on Long–Run stock price performance and equity risk of the firm. Prod. Oper. Manag. 14 (1), 35–52.
Hendricks, K.B., Singhai, V.R., 2005b. Association between supply chain glitches and operating performance. Manag. Sci. 51 (5), 695–711.
Ho, W., Xu, Y., Dey, P., 2010. Multi-criteria decision making approaches for supplier evaluation and selection: a literature review. Eur. J. Oper. Res. 202 (1), 16–24.
Ho, W., et al., 2015. Supply chain risk management: a literature review. Int. J. Prod. Res. 53 (16), 5031–5069.
Ho, T., Lim, N., Reza, S., Xia, X., 2017. Causal inference models in operations management. Manuf. Serv. Oper. Manag. 19 (4), 1–17.
Ivanov, D., Dogui, A., 2021. OR-methods for coping with the ripple effect in supply chains during COVID-19 pandemic:Managerial insights and research implications. Int. J. Prod. Econ. 232, 107921.
Jacovidis, J.N., 2017. Evaluating the Performance of Propensity Score Matching Methods: A Simulation Study. Available at: https://commons.lib.jmu.edu/diss201019/149/.
Keele, L., 2015. The statistics of causal inference: a view from political methodology. Polit. Anal. 23 (3), 313–335.
Ketzenberg, M.E., 2020. Assessing customer return behaviours through data analytics. J. Oper. Manag. 66 (6), 662645.
Khamis, A., Ismail, Z., Haron, K., Mohammed, A., 2005. The effects of outliers data on neural network performance. J. Appl. Sci. 5 (8), 1394–1398.
Kim, Y., Chen, Y., Linderman, K., 2015. Supply network disruption and resilience: a network structural perspective. J. Oper. Manag. 3334 (1), 43–59.
Kim, D.H., Pieper, C., Ahmed, A., Colón-Emeric, C., 2016. Use and interpretation of propensity scores in aging research: a guide for clinical research. J. Am. Geriatr. Soc. 64 (10), 2065–2073.
Kleindorfer, P.R., Saad, G.H., 2005. Managing disruption risks in supply chains. Prod. Oper. Manag. 14 (1), 53–68.
Klibi, W., Martel, A., Guitouni, A., 2010. The design of robust value-creating supply chain networks: a critical review. Eur. J. Oper. Res. 203, 283–293.
Kraus, M., Feuerriegel, S., Otzekin, A., 2020. Deep Learning in business analytics and operations research: models, applications and managerial implications. Eur. J. Oper. Res. 281 (3), 638–641.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25 (2), 1–9.
Lane, F.C., To, Y., Shelley, K., Robin, K., 2012. An illustrative example of propensity score matching with educational research. Career Tech. Educ. Res. 37 (3), 187–212.
Li, Yi, et al., 2022. Managing disruption risk in competing multitier supply chains. Int. Trans. Oper. Res.
Luo, Y., Peng, J., Ma, J., 2020. When causal inference meets deep learning. Nat. Mach. Intell. 2 (8), 426–427.
Manuj, I., Metzner, J.T., 2008. Global supply chain risk management. J. Bus. Logist. 29 (1), 133–155.
Mina, Hassan, et al., 2021. Transition towards circular supplier selection in petrochemical industry: a hybrid approach to achieve sustainable development goals. J. Clean. Prod. 286.
Misic, V.V., Perakis, G., 2020. Data analytics in operations management: a review. Manuf. Serv. Oper. Manag. 22 (1), 158–169.
Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference. Cambridge University Press, Cambridge, UK.
Nanni, L., Tuunanen, T., Rothenberger, M., Chatterjee, S., 2015. Coupling different methods for overcoming the class imbalance problem. Neurocomputing 158, 48–61.
NCSS, 2021. Data matching – optimal and greedy. NCSS. Available at: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Data_Matching-Optimal_and_Greedy.pdf.
Neven, V., 2021. The Global Economy.com: business and economic data for 200 countries. Global Economy. Available at: https://www.theglobaleconomy.com/.
Norrman, A., Jansson, U., 2004. Ericsson's proactive supply chain risk management approach after a serious sub-supplier accident. Int. J. Phys. Distrib. Logist. Manag.
Pearl, J., 1986. Fusion, propagation, and structuring in belief networks. Artif. Intell. 29 (3), 241–288.
Pearl, J., 2009. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.
Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal Inference in Statistics: A Primer. John Wiley & Sons Ltd., Chichester, UK.
Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S., 2007. A design science research methodology for information systems research. J. Manag. Inf. Syst. 24 (3), 45–77.
Pettit, T., Croxton, K., Fiksel, J., 2013. Ensuring supply chain resilience: development and implementation of an assessment tool. J. Bus. Logist. 34 (1), 46–76.
Rajesh, R., Ravi, V., 2015. Supplier selection in resilient supply chains: a grey relational analysis approach. J. Clean. Prod. 86, 343–359.
Ramirez-Gallego, S., Garcia, S., Mourino-Tain, H., Martinez-Rego, D., Bolon-Cando, V., Alonso-Betanzos, A., Benitez, J., Herrera, F.-, 2016. Data discretization: taxanomy and big data challenge. WIREs Data Mining Knowledge Discover. 6 (1), 5–21.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrics 70 (1), 41–55.
Rubin, D.B., 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv. Outcome Res. Methodol. 2, 169–188.
Sawik, T., 2011. Selection of supply portfolio under disruption risks. Omega – Int. J. Manag. Sci. 41 (2), 194–208.
Scholz, R.W., Tietje, O., 2022. Embedded Case Study Methods: Integrating Quantitative and Qualitative Knowledge. Sage Publications, Inc., Thousand Oaks, USA.
Scudder, G.D., Hill, C.A., 1998. A review and classification of empirical research in operations management. J. Oper. Manag. 16 (1), 361–385.
Serra, B., 2020. MOODY's Investors Service: Euler Hermes SA. Available at: https://www.eulerhermes.de/content/dam/onemarketing/ehndbx/eulerhermes_de/dokumente/moodys-rating-euler-hermes-sa.pdf.
Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., Cook, E., 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol. Drug Saf. 17 (6), 546–555.
Shmueli, G., Yahav, I., 2017. Tackling Simpson's paradox with trees. Prod. Oper. Manag. 27 (4), 696–716.
Simpson, E.H., 1951. The interpretation of interaction in contingency tables. J. Roy. Stat. Soc. 13 (2), 238–241.
Snyder, L.V., Atan, Z., Peng, P., Rong, Y., Schmitt, A., Sinsoysal, B., 2016. OR/MS models for supply chain disruptions: a review. IIE Trans. 48 (2), 89–109.
Staffa, S.J., Zurakowski, D., 2018. Five steps to successfully implement and evaluate propensity score matching in clinical research studies. Int. Anesthesia Res. Soc. 127 (4), 1066–1073.
Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Stat. Sci. 25 (1), 1–21.
Stuart, I., McCutcheon, D., Handfield, R., McLachlin, R., Samson, D., 2002. Effective case research in operations management: a process perspective. J. Oper. Manag. 20 (5), 419–433.
Tang, C.S., 2006. Perspectives in supply chain risk management. Int. J. Prod. Econ. 103 (2), 451–488.
Thun, J., Hoenig, D., 2011. An empirical analysis of supply chain risk management in the German automotive industry. Int. J. Prod. Econ. 131, 242–249.
Tomlin, B., 2006. Mitigation and contingency strategies for managing supply chain disruption risks. Manag. Sci. 52 (5), 639–657.
Tomlin, B., Wang, Y., 2005. On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. Manuf. Serv. Oper. Manag. 7 (1), 37–57.
Truong, D., 2021. Using causal machine learning for predicting the risk of flight delays in air transportation. J. Air Transport. Manag. 91.
Van Aken, R., Romme, G., 2009. Reinventing the future: adding design science to the repertoire of organization and management studies. Organ. Manag. J. 6, 5–12.
Van Vliet, M., Salmelin, R., 2020. Post-hc modification of linear models: combining machine learning with domain information to make solid inference from noisy data. Neuroimage 204 (1), 116221.
Vilko, Jyri, Hallikas, J., 2012. Risk assessment in multimodal supply chains. Int. J. Prod. Econ. 140, 586–595.
Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., 2009. Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: 17th European Conference On Information Systems. Verona, 8 June to 10 June 2009.
Voss, C., Tsikriktsis, N., Fröhlich, M., 2002. Case research in operations management. Int. J. Oper. Prod. Manag. 22 (2), 195–219.

Wacker, J.G., 1998. A definition of theory: research guidelines for different theorybuilding research methods in operations management. J. Oper. Manag. 16 (4), 361–385.
Wang, Y., Gilland, W., Tomlin, B., 2010. Mitigating supply risk: Dual sourcing or process improvement? Manuf. Serv. Oper. Manag. 12 (3), 489–510.
Wannenwetsch, H., 2010. Integrierte Materialwirtschaft und Logistik: Beschaffung, Logistik, Materialwirtschaft und Produktion. Springer, Heidelberg, Germany.
Webster, J., Watson, R.T., 2002. Analyzing the past to prepare for the future. Manag. Inform. Syst. Quart. 26 (2), xiii–xxi.
Westreich, D., Lessler, J., Funk, M., 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol. 63 (8), 826–833.
Yang, Z., Aydin, G., Babich, V., Beil, D., 2009. Supply disruptions, asymmetric information, and a backup production option. Manuf. Serv. Oper. Manag. 55 (2), 192–209.
Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A., 2020. A Survey on causal inference. ACM Trans. Knowl. Discov. Data 15 (5), 1–46.
Yu, H., Zeng, A.Z., Zhao, L., 2008. Single or dual sourcing: decisionmaking in the presence of supply chain disruption risks. Omega – Int. J. Manag. Sci. 37 (4), 788–800.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.
Zhu, X., Ninh, A., Zhao, H., Liu, Z., 2021. CrossSeries demand forecasting using machine learning: evidence in the pharmaceutical industry. Prod. Oper. Manag.

References

Roy, J.A., 2021. A Crash Course in Causality: Inferring Causal Effects from Observational Data. Coursera. Available at: https://www.coursera.org/learn/crashcourseincausality/home/info.