Data-Driven Decision-Making with Incomplete Data: Optimisation of Product Ordering

Tetyana Hess[1,*], Uta Störl[1]

[1]FernUniversität in Hagen, Hagen, Germany

Abstract

The paper analyses the concept of applying *Data-Driven Decision Making* (DDDM) in the retail sector with a particular focus on order planning. The objective is to systematically review the current state of research, summarise existing findings, and identify research gaps. Special attention is given to the question of which methods already exist to enable well-founded decisions in areas such as sales forecasting and product ordering, even when data is incomplete. The quality of the underlying data is also considered, which includes both transactional data and the maintenance of master data.

The goal of this paper is to study the use of DDDM in situations where data is missing or faulty, with a practical focus on order planning. The main focus here is on developing a resilient forecasting system that ensures stable and transparent decisions for ordering goods, depending on stock levels, shelf life, and under conditions where the data is incomplete or missing.

To address issues related to data quality, two strategies are used. The first strategy is to use synthetic datasets that simulate real retail scenarios [1] and the second is to use the application of the Multiple Imputation by Chained Equations (MICE) method, a statistically sound technique for multivariate imputation of missing values [2].

This paper contributes to the practical operationalisation of data-driven decision-making (DDDM) by integrating modern imputation methods and simulation-based data modelling. The relevance of this work is determined by conducting an analysis of the existing literature that deals with and investigates data-driven decision-making in retail under conditions of incomplete or missing data.

Keywords

Data-Driven Decision Making, Incomplete Data, Data Quality, Syntetic Dataset, MICE

1. Introduction

Data-Driven Decision-Making (DDDM) refers to the systematic collection, analysis, examination, and interpretation of data, usually through the application of analytics or machine learning methods and techniques, to make informed decisions [3]. In retail, data-driven decision-making has gained significant importance, especially given its seasonality, the limited shelf life of products, complex logistical dependencies and routes, and the high level of competition.

Theoretically, other industries can also apply the basic methodological foundation of data-driven decision-making. In this work, however, the research is focused specifically on the retail sector, because one of the authors has significant practical experience in a leading German retail company.

To illustrate this practical dependency, consider a store manager who can physically inspect the shelves and immediately see what is missing and what needs to be ordered. An SCM manager responsible for ordering products for multiple stores and not physically present faces a bigger challenge because they have to rely on the data in the system. Sometimes there are discrepancies between the stock on the shelf, meaning in the store, and the stock in the database. It happens due to theft, unrecorded discounts, spoilage, or data entry errors. These differences between actual and system-recorded data highlight the problem of incomplete or inaccurate data.

This work is based on practical experience in retail data analysis, where effective sales forecasting and ordering decisions are crucial but complicated by the fact that the data is often incomplete, incorrect, or

36th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), September 29 - October 01 2025, Regensburg, Germany

*Corresponding author.

tetyana.hess@studium.fernuni-hagen.de (T. Hess); uta.stoerl@fernuni-hagen.de (U. Störl)

outdated [4, 5, 6]. This is especially true for perishable goods and stock values that are only updated periodically after inventory checks.

The aim of this paper is to provide a systematic overview of scientific methods for decision support under conditions of data uncertainty. The focus is on imputation methods, in particular *Multiple Imputation by Chained Equations* (MICE) [2], as well as the generation of synthetic datasets to simulate realistic retail scenarios [1].

This paper delivers a structured overview of data sources and forecasting models for ordering decisions and identifying practically relevant variables. It builds a foundation for developing a robust forecasting model, which will be validated in future work by using a Monte Carlo simulation.

The structure of this paper is organized as follows: *Background* introduces the basic concepts of data-driven decision-making. *State of the Art* provides a systematic review of the existing literature and practice-oriented solution approaches. *Research Questions* formulates key questions, which are answered in *Concept* through methodically sound concepts. *Conclusion and Future Work* closes with a critical reflection and an outlook on further research.

2. Background

The concept of *Data-Driven Decision-Making (DDDM)* means systematically making decisions based on the data we have in the system and have analyzed, including machine learning methods [3]. Especially in the retail sector, data-driven decision-making (DDDM) helps with managing demand, inventory, and pricing in such a dynamically evolving sector.

In the academic literature, data-driven decision-making is presented as a combination of human expert understanding and algorithmic machine learning support, making this approach integrated and allowing dynamic response to market change.

A suitable model for illustration is the DECAS framework by Elgendy et al. [3].

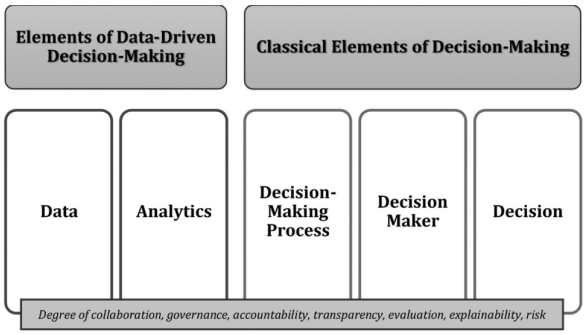


Figure 1: Data-Driven Decision-Making Elements | DECAS Model (Elgendy et al., 2022)[3]

This model differentiates between the main components: “Data-Driven” with data and analytics and “Decision-Making” with decision-making process, decision makers and the decision itself. The visualization (Figure 1) shows the importance of structured data preparation for well-founded decisions.

Figure 1 shows that data-driven decision-making describes a process where decisions are not only based on intuition but are systematically supported and validated by data.

3. State of the Art

Despite technological progress, there are many challenges in the practical implementation of DDDM. These challenges must be solved to fully use the benefits of Big Data and modern analytics tools. One of the biggest challenges is the **data quality and reliability of data**.

3.1. Data Quality as a Success Factor

A large number of studies confirm that data-driven decisions require high-quality data. McAfee and Brynjolfsson [7] demonstrate that data-driven companies are more innovative and efficient because they can respond more quickly to market changes. Janssen et al. [8] point out that low-quality data can lead to bias and uncertainty in analysis results, which can reduce the quality of decisions.

High-quality and reliable data are the foundation of data-driven decision-making. However, if data is incomplete or missing, companies risk making decisions based on incorrect or insufficient information, which can lead to incorrect decisions and serious consequences. According to Janssen et al. [8], **the quality of data is a key factor that strongly influences decision-making**, because low-quality data often causes bias and uncertainty in the results.

The quality of decisions in retail is mainly determined by the availability, consistency, and accuracy of the underlying data. In conclusion, data that is inconsistent, missing, or faulty represents a challenge in the implementation of forecasting models or data-driven decision systems.

3.2. Forecasting Models with Incomplete Data

A key challenge of the Forecasting Models is the integration of incomplete sales data into predictive models. Pedregal et al. [9] present the Tobit Exponential Smoothing method for censored time series, which allows robust forecasts even when the available data is limited. Other approaches use model-based classification frameworks to distinguish between zero values and classify demand into different types. This improves the accuracy of forecasts [10].

Basic Principle of Exponential Smoothing

Exponential Smoothing (ETS) is one of the most widely used forecasting methods in both practice and research. The method is based on the foundational work by Charles C. Holt in 1957 [11].

Holt describes an effective method for time series forecasting. This method is based on exponentially weighted moving averages. The idea of exponential smoothing is that both seasonal fluctuations and forecast trends are considered. At the same time, the requirements for using this algorithm do not involve huge volumes of data or large computational resources.

Therefore, among other things, in exponential smoothing, when we observe historical data - from today backward, for example - the coefficient applied over time will be larger for more recent data than for older and older observations.

Why do Zero Values happen?

In the article “Why do zeroes happen? "A model-based approach for demand classification”, Svetunkov et al. [10] explain that an understanding of what zero values mean can help you make better choices and make more accurate forecasting, especially in fields like retail and logistics. In other words, not all zero values mean the same. Some may be caused by:

- actual zero demand
- missing data
- technical errors

The authors suggest explicitly classifying and modeling these causes [10], instead of treating all zeros the same. The article introduces a two-step classification model:

- Identifying the origin of the zero value:
  - “true” zero demand (e.g., no purchase)
  - “false” zeros (e.g., incorrect or missing data)
- Classifying characteristics of the demand:
  - regular/intermittent
  - intermittent smooth
  - fractional/count

3.3. Imputation Techniques for Missing Values

Multiple Imputation by Chained Equations (MICE)

The method of *Multiple Imputation by Chained Equations (MICE)* has especially become widely used for handling missing or inaccurate data. This approach takes into account multivariate relationships between variables and has proven to be a dynamic and reliable tool, particularly when working with complex datasets, such as those found in medical or psychological research [12].

It is important to know why the data is missing or incomplete. In statistics, there are three main mechanisms of missing value: MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random) [13].

The MCAR mechanism means that the data is missing completely at random. MAR implies that the probability of missingness depends on observed variables and not on the missing values themselves. MNAR means that data are missing for reasons directly related to the missing values.

To determine which missingness mechanism is present, specific statistical tests are used. Little's test [14] will be used for MCAR. For MAR and MNAR, pattern analysis or logistic regression models are often applied.

MICE is particularly useful under the Missing at Random (MAR)[1]mechanism. The MICE method works on the principle of an iterative chain of regressions. In this case for each variable with missing values, a separate prediction model is built using other known variables. The process is then repeated several times until the results become consistent. Depending on the type of data, different models are used:

- linear regression (for numerical data),
- logistic regression (for binary data),
- ordinal logistic regression (for ordinal data).

Thanks to this flexibility, the MICE method is considered a powerful tool for working with datasets that contain missing values.

Synthetic Datasets

In addition to the MICE method described above, other approaches enable data imputation. One of them, which is gaining increasing importance, is *synthetic datasets*. They are particularly relevant when working with confidential data. The goal of such synthetic data is to create an artificial dataset that preserves the statistical properties of the real data but contains no confidential information.

What approaches exist? Three will be highlighted in this work:

- **Generative Adversarial Networks (GANs)** [16] – There are two neural networks in a GAN: the generator, which aims to create realistic data, and the discriminator, which tries to tell the difference between real data and data that was developed [17]. This adversarial learning process creates synthetic data that is very similar to the original data.
- **Variational Autoencoders (VAEs)** [18] use variational inference and the structure of autoencoders to make new data. This is especially useful when the original data's distribution is hard to understand or unknown.
- **Monte Carlo-based Methods** – This method is one of the most commonly used tools of stochastic modeling that is also employed to generate synthetic datasets without using any personally identifiable information. An interesting study in this context is "Nested Stochastic Valuation of Large Variable Annuity Portfolios: Monte Carlo Simulation and Synthetic Datasets" by Guojun Gan and Emiliano A. Valdez (2018) [19]. A synthetic dataset is generated to realistically

[1]**Missing at Random (MAR)**[15]: For any individual, the probability that the value of a variable X_j is missing does not depend on the (unobserved) value of X_j itself, but only on the observed values of the other variables. Formally, this means:

$$P(X_j = \text{mis} \mid X_1, \ldots, X_p) =$$

$$= P(X_j = \text{mis} \mid X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p).$$

simulate complex financial processes, such as the assessment of important insurance portfolios, without using sensitive data. The data is generated through Monte Carlo simulations combined with a nested stochastic simulation framework. These models show how markets change over time [19].

A key advantage of synthetic data is its compliance with data protection [20, 21]. This allows simulation and data analysis without exposing confidential information – an important factor in data-driven fields such as retail.

Actual research suggests combining MICE and synthetic data generation. MICE is used to impute real data gaps, while synthetic data improves diversity and robustness in forecasting models [21]. In retail this hybrid strategy is useful.

4. Research Questions

The quality of data-based forecasts in the retail sector plays a big role in decision-making and strongly depends on the completeness, consistency, and relevance of the underlying data.

The authors present the following research questions:

- **RQ1.** Which methods are suitable for systematically identifying data gaps and anomalies, and how can data quality be assessed beyond traditional metrics?
- **RQ2.** Which methods are appropriate for analyzing and improving the semantic, syntactic, and pragmatic quality of master data?
- **RQ3.** What criteria determine whether data is suitable for accurate forecasting?
- **RQ4.** Which imputation techniques are particularly suitable for use in the retail sector?

5. Concept

The quality, completeness, and consistency of the underlying data have a significant impact on data-driven decisions in retail. For this reason, the focus of this chapter is on the methodological framework for answering the defined research questions. Additionally, ideas and perspectives are discussed regarding which methodological approaches can be followed and which research areas should be explored in future work.

RQ1. Which methods are suitable for systematically identifying data gaps and anomalies, and how can data quality be assessed beyond traditional metrics?

To systematically **identify data gaps and anomalies**, classical methods of Exploratory Data Analysis (EDA) [22] are applied. These help to detect structural patterns, outliers, and distributions both visually and statistically. Different visualization techniques such as boxplots, heatmaps, and time series plots support intuitive detection of missing values, extreme values, and unusual distribution patterns.

In addition, specialized anomaly detection methods [23, 24] are used to identify unusual, implausible, or faulty data points automatically. There are different statistical methods (e.g., Z-scores), machine learning algorithms (e.g., Isolation Forest), and time series-based approaches (e.g., seasonal decomposition).

The combination of exploratory analysis and algorithmic anomaly detection allows for robust data validation, covering both obvious and subtle anomalies. This makes quality control more effective and flexible in fields like retail.

Beyond traditional metrics, a **context-based assessment of data quality** is recommended to reflect practical retail requirements. In addition to the well-known 15 dimensions of Wang and Strong (1997) [25], which include aspects such as accuracy, timeliness, and completeness, the analysis should be extended to include domain-specific quality indicators.

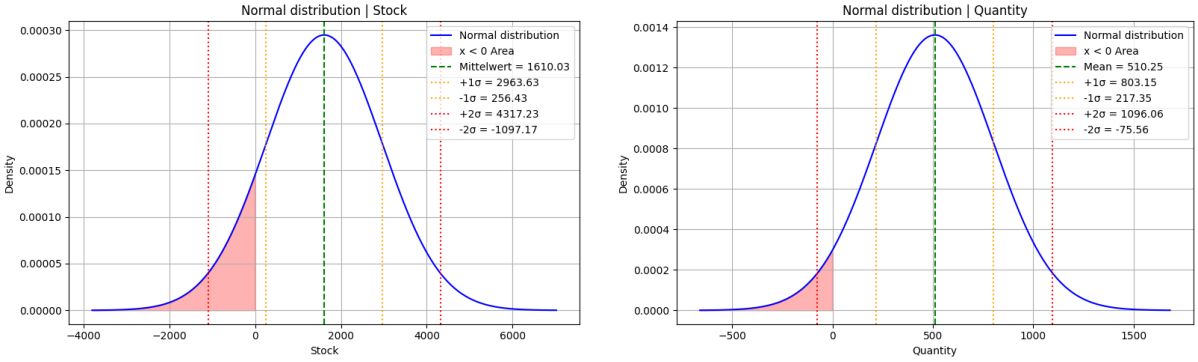
Therefore, the concept also integrates retail-specific Key Performance Indicators (KPIs), such as stock accuracy and sales volume.

A key element of the concept is the use of exploratory data analysis (EDA), particularly univariate and bivariate methods. The study begins with the assumption of a normal distribution, because it is a simple and understandable way to identify outliers using sigma intervals (e.g., ±1σ, ±2σ, ±3σ). KPIs such as sales volume, order quantity, and stock are checked to determine whether they lie within these intervals. Values outside these intervals are marked as potential outliers that require further review.

At the same time, the assumption of normality must be critically questioned. In retail, sales and stock data often show right-skewed distributions, for example due to the Pareto effect (20% of products generate 80% of sales). As a result the methodological study also includes testing alternative statistical models, such as log-normal distributions, to ensure meaningful outlier detection.

To illustrate this, a sample dataset is used with KPIs such as daily sales, order quantity, and stock. This dataset with inserted faulty values was manually created for demonstration purposes to simulate typical retail data problems.

In the first step, an univariate analysis is performed (Figure 2a). The stock data shows a mean of approximately 1610 units under the assumption of a normal distribution. Sigma intervals show the separation of normal values and outliers. Stock values below zero are implausible and immediately flagged as erroneous. Stock values below zero are clearly implausible and immediately flagged as erroneous. The values beyond the ±2σ range are marked as potential outliers for further inspection.



(a) Stock

(b) Sales Quantity

Figure 2: Normal Distribution

A similar procedure is applied to sales volumes (Figure 2b). Again, assuming a normal distribution, clear upper and lower boundaries can be established.

By combining classical dimensions, domain-specific KPIs, and well-considered analytical techniques, a more nuanced evaluation of data quality is achieved.

RQ2. Which methods are appropriate for analyzing and improving the semantic, syntactic, and pragmatic quality of master data?

This paper presents a practical approach for analyzing and improving master data quality, based on classical methods and exploratory data analysis (EDA).

While rule-based checks, similarity analysis, and ontologies are established methods [4, 26], the integration of anomaly detection and EDA can reveal semantic and pragmatic quality issues at an early stage.

- **Syntactic quality**
  In addition to standard checks (e.g., format and value ranges), EDA methods can reveal frequent deviations from syntactic standards. For example, length analysis may identify invalid IDs, or analysis of BBD (best-before date) clusters may reveal unrealistic values (e.g., very high or low years) or placeholders such as “01.01.1900”.
- **Semantic quality**
  Beyond classical ontology use, EDA can help identify semantic inconsistencies. For example,

grouped bar plots or heatmaps can be used to check whether product assignments to categories or product IDs are consistent (e.g., revenue distribution by product group). Another example is whether specific product groups have unusually short or long shelf lives compared to industry averages.

- **Pragmatic quality**
  This is especially important because it helps identify inconsistencies in master data.. For example, comparing sales velocity with the remaining shelf life may show that products with high BBD values but low sales are either slow-moving goods or contain incorrect shelf-life data.

Extending the analysis of these three aspects of data quality (semantic, pragmatic, and syntactic) with EDA provides greater flexibility and helps identify not only syntactic errors but also semantic and pragmatic inconsistencies. Especially in the retail sector, where master data plays a major role in analysing a wide product range, this offers valuable opportunities to ensure high data quality.

RQ3. What criteria determine whether data is suitable for accurate forecasting?

For accurate forecasting, it is essential to check the data for suitability, that is, for quality. Firstly, the data completeness is very important, because missing values can distort time patterns. It's also crucial that historical data goes back long enough to accurately identify recurring cycles and trends.

Another key criterion is autocorrelation, which measures the dependency of observations over time. In this context, autocorrelation should be analysed per product group. It is important to distinguish between short-term autocorrelation and seasonal autocorrelation, as both imply different modelling approaches [27]. Stable and significant autocorrelation may indicate high predictability, while strongly fluctuating patterns suggest irregular behaviour.

Variance stability is also important, as the heteroskedastic time series makes the estimation of forecast intervals more difficult [27].

Additionally, the distribution of the data matters for forecasts—special attention should be paid to asymmetry and sensitivity to outliers.

In conclusion, the data should be relevant and have appropriate granularity. For example, in the context of forecasting and sales prediction, the data should be available at least on a daily basis.

RQ4. Which imputation techniques are particularly suitable for use in the retail sector?

The study "Inventory record inaccuracy in grocery retailing: Impact of promotions and product perishability, and targeted effect of audits" [6] shows that accurate inventory records in retail can increase sales by up to 11%. Why is the stock data not always accurate? This is because stock levels are typically a calculated value and inventory checks are only done once a month or even less frequently. As a result, incorrect inventory data can lead to wrong or missed product orders. Errors in delivery recording may also cause inflated inventory levels. Accurate inventory is essential for sales forecasting, since forecasts are based on historical values. If products are not available on the shelf due to incorrect stock data, they are also missing from sales data.

As described, data quality in retail is a critical success factor [8]. Faulty or incomplete data can cause major inefficiencies across stock and sales processes. A frequently neglected aspect is the proper handling of missing data and distortions, which directly affect stock and sales forecasts. This section therefore, focuses on the suitability of imputation methods, especially the Monte Carlo [16] method, and illustrates them with practical examples from retail.

A key advantage of Monte Carlo imputation is that missing values are not replaced by deterministic statistics such as mean values but by stochastic sampling from estimated probability distributions. This is especially relevant in retail, where both sales quantity and inventory data often show strong skewness and outliers.

Choosing the correct underlying distribution is critical. One central criterion is based on the **Pareto Principle**, which states that a small number of products in many categories generate the majority of

the revenue. For such heterogeneous data sets, assuming a normal distribution is not sufficient. Instead, specialized distributions are needed to reflect the real structure of the data.

For items with highly uneven sales patterns – especially slow-moving non-food products or promotional items with rare sales peaks – the **Pareto distribution**[2] is a suitable choice.

It allows modeling of rare but high sales values and addresses the phenomenon of "few products, high sales share." Careful attention must be paid to the choice of the minimum value x_m and the shape parameter α.

For regularly sold products, such as fast-moving consumer goods or fresh products with recurring demand, the **log-normal distribution**[3] is more appropriate.

Since sales are always positive and often right-skewed, the log-normal distribution is well-suited. Its logarithmic values follow a normal distribution. The parameters μ and σ can be estimated from the log-transformed data, providing a realistic representation of mean and variance while preserving skewness in the original data.

Finally, the Monte Carlo imputation offers a well-founded method to fill data gaps in the retail sector while considering the sales structure. Selecting the correct distribution is very important to avoid bias in forecasting models and to support commercially reasonable decisions.

6. Conclusion and Future Work

This paper is based on the scientifically supported assumption that data-driven decision-making in retail – especially in forecasting and sales prediction – is strongly influenced by the quality and completeness of the underlying data [8]. This assumption is confirmed by both academic literature and practical experience in the retail industry. The aim of this exploratory study was to identify and discuss initial approaches for addressing data quality challenges. The focus was on methods for identifying data gaps and anomalies, as well as on techniques for analysing and improving data quality.

Within the scope of this work, exploratory data analysis, synthetic datasets, and Monte Carlo simulations were considered as potential tools to improve the quality of the available data and to close missing values realistically.

Future research should further develop the methods discussed here and systematically compare them with established approaches from the current scientific literature. Particular focus should be given to the selection and evaluation of suitable imputation methods in order to assess their effectiveness and applicability in the specific context of retail. Moreover, it would be of particular interest to validate the developed methods using real data from the retail sector, thereby formulating well-founded practical recommendations for the application of data-driven decision-making based on empirically verified evidence.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4 in order to: Grammar and spelling check.

[2] A continuous random variable X is called *Pareto distributed* $\mathrm{Par}(k, x_{\min})$ with parameters $k > 0$ and $x_{\min} > 0$ if it has the probability density function

$$f(x) = \begin{cases} \frac{\alpha\, x_{\min}^{\alpha}}{x^{\alpha+1}} & \text{for } x \geq x_{\min}, \\ 0 & \text{for } x < x_{\min} \end{cases}$$

[3] A continuous random variable X is called *log-normally distributed* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if it has the probability density function

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$

References

[1] Y. Xia, C. Wang, J. Mabry, G. Cheng, Advancing retail data science: Comprehensive evaluation of synthetic data, CoRR abs/2406.13130 (2024). arXiv:2406.13130.

[2] S. van Buuren, Flexible imputation of missing data, CRC press (2018) 8–10, 120–129.

[3] N. Elgendy, A. Elragal, T. Päivärinta, Decas: A modern data-driven decision theory for big data and analytics, Journal of Decision Systems 31 (2022) 337–373. URL: https://doi.org/10.1080/12460125.2021.1894674.

[4] J. Becker, M. Matzner, O. Mueller, A. Winkelman, Towards a semantic data quality management using ontologies to assess master data quality in retailing, AIS eLibrary (2008) 1–11. URL: https://aisel.aisnet.org/amcis2008/129.

[5] A. Winkelman, D. B. und Christian Janiesch, J. Becker, Improving the quality of article master data: Specification of an integrated master data platform for promotions in retail, AIS eLibrary (2008). URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1138&context=ecis2008.

[6] Y. Rekik, R. Oliva, C. Glock, A. Syntetos, Inventory record inaccuracy in grocery retailing: Impact of promotions and product perishability, and targeted effect of audits (2025). arXiv:2506.05357.

[7] A. McAfee, E. Brynjolfsson, Big data: The management revolution, Harvard Business Review (2012). URL: https://hbr.org/2012/10/big-data-the-management-revolution.

[8] M. Janssen, H. van der Voort, A. Wahyudi, Factors influencing big data decision-making quality, Journal of Business Research (2017) 338–345. URL: https://doi.org/10.1016/j.jbusres.2016.08.007.

[9] D. J. Pedregal, J. R. Trapero, E. Holgado, Tobit exponential smoothing, towards an enhanced demand planning in the presence of censored data (2024). arXiv:2407.17920.

[10] I. Svetunkov, A. Sroginis, Why do zeroes happen? A model-based approach for demand classification, CoRR (2025). arXiv:2504.05894.

[11] C. C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, O.N.R. Research Memorandum (1957) 1–11.

[12] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, National Library of Medicine (2011) 40–49. URL: https://pubmed.ncbi.nlm.nih.gov/21499542/.

[13] I. Jansen, N. Hens, G. Molenberghs, M. Aerts, G. Verbeke, M. G. Kenward, The nature of sensitivity in monotone missing not at random models, Comput. Stat. Data Anal. 50 (2006) 830–858. doi:10.1016/J.CSDA.2004.10.009.

[14] R. J. A. Little, Inference about means from incomplete multivariate data, Biometrika (1976) 593–604. URL: https://academic.oup.com/biomet/article/63/3/593/270937.

[15] E. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, M. Cubiles-de-la-Vega, Missing value imputation on missing completely at random data using multilayer perceptrons, Neural Networks 24 (2011) 121–129. URL: https://doi.org/10.1016/j.neunet.2010.09.008.

[16] X. Qin, H. Shi, X. Dong, S. Zhang, L. Yuan, Improved generative adversarial imputation networks for missing data, Appl. Intell. 54 (2024) 11068–11082. URL: https://doi.org/10.1007/s10489-024-05814-2.

[17] J. Li, X. Wang, Y. Lin, A. Sinha, M. P. Wellman, Generating realistic stock market order streams, AAAI Press, 2020, pp. 727–734. URL: https://doi.org/10.1609/aaai.v34i01.5415.

[18] B. Roskams-Hieter, J. Wells, S. Wade, Leveraging variational autoencoders for multiple data imputation, in: Machine Learning and Knowledge Discovery in Databases: ECML PKDD, Springer, 2023, pp. 491–506. URL: https://doi.org/10.1007/978-3-031-43412-9_29.

[19] G. Gan, E. A. Valdez, Nested stochastic valuation of large variable annuity portfolios: Monte carlo simulation and synthetic datasets, Data 3 (2018) 31. URL: https://doi.org/10.3390/data3030031.

[20] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gans, NeurIPS (2019). URL: https://dl.acm.org/doi/10.5555/3454287.3454946.

[21] S. Mohapatra, J. Zong, F. Kerschbaum, X. He, Differentially private data generation with missing data, Proc. VLDB Endow. 17 (2024) 2022–2035.

[22] W. Polasek, Explorative daten-analyse: Eda; einführung in die deskriptive statistik, Springer-Verlag (1988) 3–28.

[23] G. M. Tavares, V. G. T. da Costa, V. E. Martins, P. Ceravolo, S. B. Jr., Leveraging anomaly detection in business process with data stream mining, Braz. J. Inf. Syst. 12 (2019) 54–75.
[24] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (2009) 15:1–15:58.
[25] D. M. Strong, Y. W. Lee, R. Y. Wang, Data quality in context, Commun. ACM 40 (1997) 103–110.
[26] M. Melkonian, C. Juigné, O. Dameron, G. Rabut, E. Becker, Towards a reproducible interactome: semantic-based detection of redundancies to unify protein interaction databases, Bioinform. 38 (2022) 1685–1691.
[27] R. J. Hyndman, G. Athanasopoulos, Forecasting: Principles and practice | 2.8 autocorrelation and 3.4 evaluating forecast accuracy (2018). URL: https://www.otexts.com/fpp3/acf.html.