Hochkamp, Florian; Rabe, Markus

Conference Paper

Outlier detection in data mining: Exclusion of errors or loss of information?

Florian Hochkamp and Markus Rabe

Outlier Detection in Data Mining: Exclusion of Errors or Loss of Information?

HICL

Published in: Changing Tides
Wolfgang Kersten, Carlos Jahn, Thorsten Blecker and Christian M. Ringle (Eds.)
ISBN 978-3-756541-95-9, September 2022, epubli

Outlier Detection in Data Mining: Exclusion of Errors or Loss of Information?

Florian Hochkamp[1] *and Markus Rabe*[1]

1 – Technical University Dortmund

Purpose: *Our research emphasizes the importance of considering outliers in production logistics tasks. With a growing amount of data, we require data mining to cope with these tasks. We underline that the widespread exclusion of outliers in data pre-processing for data mining leads to a loss of information and that using outlier interpretation can be used to address the issue.*

Methodology: *The paper discusses the data pre-processing of data mining in production logistics problems. Methods of outlier interpretation are collected based on a literature review. In addition to the literature-based investigation, the work relies on a case study that illustrates the individual evaluation of outliers.*

Findings: *This work shows that outliers take a special focus on the information generation. Within data pre-processing, a distinction must be made between an outlier as a defect and an outlier as a special datum. This can be conducted by methods presented in the literature.*

Originality: *This paper adds to existing literature in the research field of insufficiently analyzed outlier interpretation and shows a need for research in data pre-processing of data mining.*

First received: 11. Mar 2022 *Revised: 22. Aug 2022* *Accepted: 25. Aug 2022*

1 Introduction

In 2020, the manufacturing sector in Germany continues to be characterized by the quality claim of the “Made in Germany” brand with approximately 23 % share of gross value added in Germany (Statista, 2021), which is complemented by logistics as one of the most important economic fields. Steady developments in production led first to mass production and later to customized mass production according to Chen, et al. (2015). This is accompanied by a volatile competitive environment of the manufacturing companies. An important unique selling proposition is product quality, which enables companies to win customers and meet their expectations (Jacobson and Aaker, 1987). In order to meet the requirements of customized mass production in a competitive environment, various developments are being used in the context of digitization (Kusiak, 2018). Buzzwords such as Industry 4.0, Big Data, intelligent production and communication systems, and business intelligence have been shaping trends in recent years. At the same time, these also condition the field of logistics through collaboration, globalization, and just-in-time production. All the mentioned trends imply information in companies about the heterogeneous production situations, which is indispensable for a modern production infrastructure (Pennekamp, et al., 2019). Addressing increasingly complex production systems (Alkan, et al., 2018) implies a high level of internal and external information exchange, so that concrete knowledge about production and logistics can be used for implementation, improvement, and quality assurance (Kersten, Blecker and Ringle, 2020). Without suitable analysis procedures, with increasing data volumes, contained information and contained knowledge are no longer tangible for analyses. This knowledge is to be interpreted as a valuable enterprise resource and, thus, requires special attention (Wenzel and Stolipin, 2017).

Increasingly, knowledge discovery methods in databases (KDD) are used for this purpose, with data mining (DM) being the most important process step. For many manufacturing companies it is unclear which methods to apply in DM. An implementation of every DM technology requires a data pre-processing matching the specific DM algorithm, which deals with data quality deficiencies and ensures the structuredness of the data. While the literature includes manifold sources for removing noise, handling missing values, and

detecting outliers, a detailed analysis is rarely sought for the latter. This is exacerbated in the KDD context, where outliers are mostly excluded from DM after detection, resulting in an exclusive definition of outliers as defects without more-detailed analysis. This creates the possibility of information loss and, thus, risks for the analysis in the KDD process. Especially for critical systems and products, such a defect can have serious consequences. The consideration of outliers is of particular interest for various application areas, such as credit card fraud, clinical trials, network security (Ben-Gal, 2005), but also fault diagnosis, detection of structural defects, time series analysis, or erroneous entries in databases (Hodge and Austin, 2004). Also, in the domains of production and logistics, time, security, and increasing costs are highly relevant for a detailed examination of outliers to improve analyses.

This paper points out the critical gap in the literature in the area of outlier investigation in DM for manufacturing and logistics. For formal derivation of the argument, data and information quality are separated by their definitions. Furthermore, the relevance of a differentiated consideration of outliers is discussed against the background of existing literature. Here, a discussion of data pre-processing methods for DM takes place, identifying possible reasons for information loss due to outlier exclusion. The research reinforces the use of outlier interpretation to consider outliers in production logistic issues.

The paper is structured as follows: In Section 2, the required terms are first put into context. Against the background of the DM literature, the terms data quality and information quality are separated. Furthermore, the pre-processing of data in DM is discussed in particular. In this context, different methods from the literature are compared and classified. Section 3 discusses technology support in the area of data analysis with a focus on dealing with outliers in the domain context. Section 4 presents a case study including the testing of a selected method described in the previous chapters. The thesis concludes in Section 5 with a summary and outlook.

2 Theoretical Background on Information in Outliers in Data Mining

In the following sections, the necessary basics for the work are discussed and differentiated from related research fields. First, the necessary terms are explained, and data quality is distinguished from information quality in particular. After a short summary of the usual procedure in the KDD process, data pre-processing methods of the DM are classified.

2.1 Data and Information Quality

Both information and knowledge are based on data (North, 2022). For data to be used, they must first be generated, collected, or measured, and then stored. The data are then available in different formats, structures, and qualities in companies. While the question of the correct format can usually be answered in a sufficiently trivial manner, the choice of the appropriate structure of the database is subject of debate. Even though relational databases are most common in companies (Saake, Sattler and Heuer, 2019) other database structures up to the polyglot persistence of a data network are possible (Khine and Wang, 2019). In particular, NOSQL databases such as graph databases are cited as a natural representation option in logistics contexts such as supply chains (Hunker, Scheidler and Rabe, 2020).

The stored data cannot be used directly by the viewer. According to North (2022), data become information when meaning is attached to them and they enable action. This meaning can be assigned by formal description criteria (Piro and Gebauer, 2011) or by the observer himself. In the context of production logistics, information includes, for example, details of processes, intended uses, and decision support. The interlinking of different information denotes North (2022) as knowledge.

Defects are possible in data and information. The International Organization for Standardization (2010) defines a defect as a result-altering problem, a failure to meet requirements, and a designation for an error. Making a connection of the defect definition to data and information inevitably leads to data quality and information

quality, with quality as the degree to which requirements are met (International Organization for Standardization, 2015). Data quality and information quality are often used interchangeably in the literature (Gebauer and Windheuser, 2021). Gebauer and Windheuser (2021) define data quality as the suitability of the dataset to fulfill quality characteristics and meet specified requirements. Accordingly, data quality serves conceptually as a classification of the problems arising during generation, collection, measurement, storage, and merging. Thus, high data quality is equivalent to few relevant errors in the dataset. The separation of data and information quality was investigated by preliminary work at the department IT in Production and Logistics (ITPL) and is possible via the data and information concept. The data quality evaluates the mapping quality between the real world as well as the representation by the data and the information quality evaluates the suitability of the data to fulfill a certain purpose (Mengering, 2021).

Even generated data, e.g., through data farming (Brandstein and Horne, 1998), may contain data quality deficiencies. Measured and collected real data are mostly burdened with data quality deficiencies, e.g., due to faulty data collection measures, missing data, or definition inconsistencies (García, Luengo and Herrera, 2015). In the case of real production data in particular, their heterogeneity is also reflected within the data quality. The data quality is, thus, in the context of the structure and the format of the data, but it is also dependent on manifold influencing factors on the level of data storage and analysis (Oliveira, Rodrigues and Henriques, 2005). Collected data are subject to external factors at various levels, such as environmental phenomena, strategic changes, or changes in machine behavior. Information about the external factors is necessary to be able to quantify the partially influencing factors. This also requires a consideration of the different levels of aggregation within the available production and logistics data. Various works in the literature propose different dimensions to evaluate the data quality. Internationally used sources in this respect (Miller, 1996; Redman, 1996; Wang and Strong, 1996; English, 1999) have already been supplemented by German-language ones in preliminary work at the ITPL (Müller, 2000; Rohweder, et al., 2011) as well as examined in the supply chain context (Türkmenoglu, 2021). These were synthesized into the following dimensions: format, consistency, accuracy, completeness, comprehensibility, lack of redundancy, trustworthiness, accessibility, security, timely and accrual-based

posting, response time, relevance, and timeliness. Before using data to generate information, data quality can be used as an assessment, accordingly. Similarly, the information quality must be quantified before the information is used.

According to the definition of information, information quality must evaluate the suitability of a piece of information for use. This definition of information quality is consistent with many sources in the literature, which describe information quality specifically in terms of suitability, which is evaluated intrinsically or externally, of a piece of information for use (Stvilia, et al., 2007). This evaluation is supported by Lee, et al. (2002) in the following 15 dimensions for the measurement of the information quality: accessibility, appropriate amount, believability, completeness, concise representation, consistent representation, ease of operation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, and understandability. These differ significantly according to the application domain, e.g., in the context of Big Data the appropriate amount is questionable. The occurrence of low data and information quality is not synonymous with information loss, but there is a correlation. At the same time, the underlying data quality to which the information refers to is relevant for the information quality.

While information quality is directly related to data quality, studies on the use of information need to consider data and information quality separately. For example, poor data quality does not exclusively lead to poor information quality, even if there is a correlation. The inherent information of each dataset can be used with caution even when data quality is poor. Already Parsons (1996) discusses the handling of imperfect information as a consequence of basing information on data that are real-world and uncertain. Accordingly, in addition to considering data quality as a whole, individual datasets must be examined for their information content and placed in the context of relevant data quality deficiencies. However, an unpredictable data situation is not synonymous with disruptions, errors, or poor planning. There may be information in the data that has not yet been allocated or that cannot yet be estimated. When evaluating the suitability of a datum for generating information, deficiencies like noise, missing values, and outliers exist apart from database-specific data quality. Both noise, which can be seen as a corruption of real data (Dong, Chan and Xu, 2007) and as an

unintentional obstacle for the analyst (Chandola, Banerjee and Kumar, 2009), as well as missing values, which do not allow use as data, are negatively associated in the context of data quality in the literature. In contrast, outliers occupy a separate role and are sharply distinguishable from noise (Chandola, Banerjee and Kumar, 2009).

Outliers are data with a sufficiently big difference from expectation, suggesting a deviant mechanism of origin (Hawkins, 1980). For outliers, there is no fixed basis of occurrence (Barnett, 1978) which can be, e.g., a measurement error or an undetected external influence in the data. Aggarwal (2017) adds the terms abnormality, discordant, deviation, and anomaly used in the literature for outliers. Furthermore, Aggarwal names inliers, which, unlike outliers, do not deviate from the expected data model. Wainer (1976) lists the designation of fringeliers, which are to be categorized between the outliers and inliers and for which no direct classification as outlier or inlier is possible. An overview of outliers, fringeliers and inliers is shown in Figure 1.

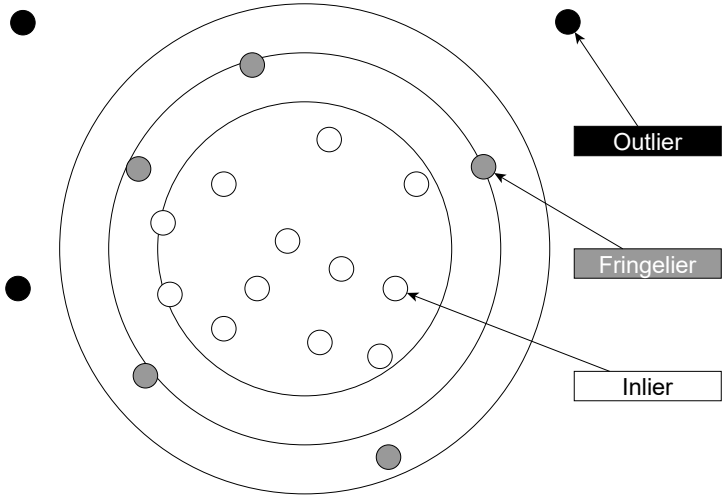

Figure 1: Concept illustration of the outlier (black), fringelier (grey) and inlier (white)

Already Beckman and Cook (1983) give reasons for dealing with outliers: special interest within the research question, detection of special alternative phenomena, accumulation of outliers, and influence of outliers. The consideration of outliers is, however, likewise in KDD processes and statistical procedures a part of the data pre-processing, even if outliers should not represent the actual investigation goal.

It is to be expected that information can also be extracted from data declared as outliers. In some areas, such as quality assurance or risk analysis, sometimes explicitly the outlier data themselves as well as findings about the outliers represent the relevant information or have a significant influence on it. For this reason, outlier detection and interpretation, also explanation or description, receive special attention here.

2.2 Data and Error Processing in Data Mining

KDD is concerned with extracting useful information and knowledge from large amounts of digital data (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Diverse process models with different focuses developed in KDD are presented in the literature, such as the model of Fayyad, Piatetsky-Shapiro and Smyth (1996), the Cross Industry Standard Process for Data Mining (CRISP-DM) (Wirth and Hipp, 2000) or the Sample Explore Modify Model Assess (SEMMA) of the SAS Institute (Azevedo and Santos, 2008). The KDD process, according to the overlaps of the models, is based at least on the research question, data selection, data pre-processing, DM, and post-processing of the data mining result (Scheidler and Rabe, 2021). Figure 2 gives an overview concerning the classification of the phase results of the KDD process according to Fayyad, Piatetsky-Shapiro and Smyth (1996) on the knowledge staircase (North, 2022).

While the research question and the evaluation mostly provide the contextual reference to the use case, the DM is to be considered as the central aspect in the KDD process and to be understood as a collective term for knowledge extraction procedures. DM is already used in the context of production for various applications, such as quality assurance and improvement (Köksal, Batmaz and Testik, 2011), but also maintenance and special production processes (Harding, Shahbaz and Kusiak, 2006). The phase results of DM are patterns discovered in the data. The pattern term is polysemous in definition and representability in the KDD context.

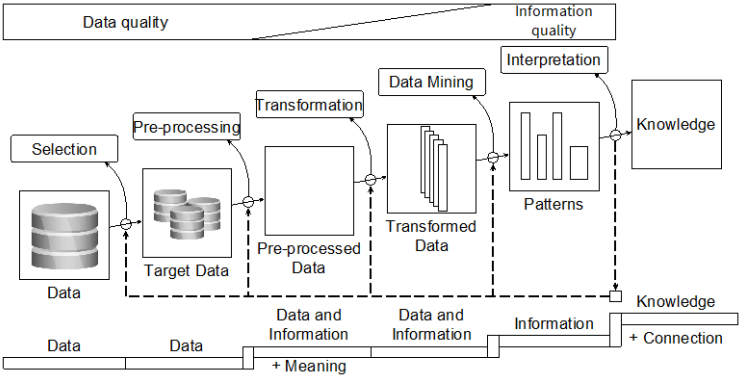

Figure 2: Data, information, and knowledge in the KDD process based on Fayyad, Piatetsky-Shapiro and Smyth (1996) and North (2022)

Nevertheless, there is the possibility to compare patterns (Geng and Hamilton, 2006) and to examine them for their interestingness (Silberschatz and Tuzhilin, 1995). To ensure the analysis quality of the KDD process, data pre-processing is essential, as it compensates for data quality deficiencies and is a mandatory perquisite for the application of data mining methods. In most KDD processes, data preparation consists of data integration, data cleansing, data normalization, filling missing data, identifying noise, data transformation, and reducing data (García, Luengo and Herrera, 2015). In this paper, a special focus is given to the KDD model of Fayyad, Piatetsky-Shapiro and Smyth (1996).

Data integration addresses the merging of different data sources and the handling of the resulting sources of error, such as the different formatting of the weight column in different databases as weight in grams or kilograms. However, the model of Fayyad, Piatetsky-Shapiro and Smyth (1996) assumes an already integrated database.

Data cleansing, data normalization, filling missing data, and identifying noise is addressed in the KDD model of Fayyad, Piatetsky-Shapiro and Smyth (1996). In data cleansing, data errors are corrected, which can include entry errors, data transmission errors, and errors in the data processing system. For example, an entry *Dortmund* in the column *Postal code* is cleaned. Data normalization ensures that data inappropriate to the

DM algorithm are converted to a different form so that new attributes with appropriate values can be generated and used for analysis. A treatment of missing values is essential for the use of most DM algorithms, insofar as they are not robust. Common practice is the exclusion of the respective dataset with missing values, but also an estimation of the missing values via dependencies and similarities to other values. When identifying noise, the imperfect data must be cleaned of corruptions. In particular, noise hinders the calculation of sharp boundaries, e.g., for clusters. At the same time, however, it also hinders other analyses. Solutions in the DM context provide robust learners, a partial exclusion of noise, or a filter to eliminate noise.

The data transformation aggregates raw data values to adapt the value ranges or distributions according to the requirements of the underlying DM algorithm. Both the data transformation and the subsequent data reduction are assigned to the transformation step and not to the data pre-processing step in the KDD model of Fayyad, Piatetsky-Shapiro and Smyth (1996). The goal of data reduction is to address the curse of dimensionality, avoiding the unnecessary processing of too much data. In this process, data are cleverly excluded so that the DM process produces the same or a nearly identical result (García, Luengo and Herrera, 2015).

The treatment of outliers different from exclusion in data pre-processing and the overall KDD process of Fayyad, Piatetsky-Shapiro and Smyth (1996), the CRISP-DM, and SEMMA, is not provided for. This leads to an incomplete knowledge discovery and consequently to a lower analysis quality as well as non-consideration of the formation mechanisms of unexpected data. Changes to the patterns, which were extracted under exclusion of the outliers, can influence the result of the knowledge discovery and the information contained therein can be lost. This is especially relevant in application fields with required high analysis accuracy, such as medicine, and high-performance applications, or with focus on data deviations, such as the considered outliers in quality control. For these reasons, usage-intended detection of outliers is an important, but not explicitly listed, step in data pre-processing. The literature on outlier detection, i.e., declaring data as outliers, has been extensively studied from general procedures to domain-specific algorithms. A good overview about the general topic of outliers can be found in Chandola, Banerjee and Kumar (2009) and Aggarwal (2017).

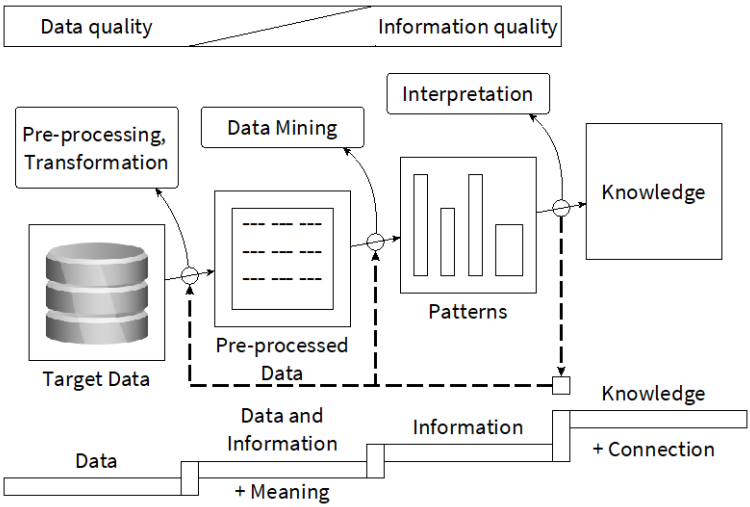

Figure 3: Data, information, and knowledge in the aggregated KDD process based on Fayyad, Piatetsky-Shapiro and Smyth (1996) and North (2022)

In the model of Fayyad, Piatetsky-Shapiro and Smyth (1996) the selection step is predominantly determined by the research question and a subset is formed from the existing database. The database formed by this process was reduced by a large number of records, which changes the subsequent analysis result. For the result of the KDD process, it is irrelevant whether by the selection the relevant data are extracted from a database or exclusively relevant data are collected in a database without selection. For this reason, the selection step must be evaluated separately from the rest of the KDD process. The step of transformation, which includes data reduction and data projection, is integrated into data pre-processing in the context of this work. The background is the missing transformation of data, information, and knowledge on the knowledge staircase. The underlying object of consideration, i.e., data and information, is not changed, but the set of the respective data is reduced as well as formatted for the data mining step. According to the statements of Fayyad, Piatetsky-Shapiro and Smyth (1996) and Han, Pei and Kamber (2012), no or almost no change of the result takes place in the

transformation step, which is why the merging has no effects on a global level. These adaptions result in the presented reduced KDD model in Figure 3.

3 Information Loss Prevention in Data Mining Pre-processing

In the following section, the technology-supported implementation of DM is discussed and outlier interpretation is considered. Here, conventional methods are distinguished from technology-supported methods postulated in the literature against the background of large datasets.

3.1 Technology Support for Data Pre-Processing of Data Mining in Production Logistics

In production logistics, more and more use is being made of technology-supported analysis methods such as DM. As data volumes in companies continue to grow, consideration of data exceeds the manual manageability and evaluation of datasets. Production and logistics data, most of which are stored in relational databases, are not only being more frequently collected by sensors, but the level of detail and the scope of the data are also increasing. Thus, the increase in the network size of supply chains also leads to more complex subordinate processes, which in turn leads to larger databases (Scheidler, 2017). New parameters add to the previous considerations and dimensionality of the data. This increases the complexity of the DM, but especially also that of the data pre-processing, in data reduction, data cleansing, and the consideration of outliers.

Even for frequently performed analyses, subject matter experts are needed who are familiar with DM methods and can at the same time classify the issue under consideration from a technical point of view. The prevailing shortage of subject matter experts also necessitates extensive technical support for pre-processing and analysis execution. In this way, subtasks can be further automated or reduced in complexity.

3.2 Method of Outlier Detection and Outlier Interpretation

Within the KDD process, outliers are addressed within the data pre-processing and as a result of the DM as explained in Section 2.2. Here, it is to be distinguished whether the outliers represent the analysis result of the DM or a research question is examined, in which outliers were detected as a secondary result in data pre-processing.

In both cases, different detection methods are used, and their selection depends on different factors, such as the types of data, the amount of data, the knowledge about former outliers in the dataset, and the interpretability of the detected outliers (Aggarwal, 2017). After applying the detection procedure, procedure-dependent results are presented to the analyst for his interpretation. The classification is complicated by possible false positives, i.e., inliers that are classified as outliers, and false negatives, i.e., outliers that are classified as inliers. Both types of incorrect classification occur more frequently in the region of fringeliers.

Conventional detection algorithms list outliers and inliers. Here, the analyst lacks contextual information that facilitates interpretation. Regarding the background of the research question, the detected outliers are compared to the existing knowledge about former outliers in the dataset. Application-specific rationales are also reviewed, such as detected outliers before a machine failure occurred. The consequences and causes of the outliers can be classified by the precise technical examination and used in subsequent analyses of similar data. However, the effort required to interpret the respective outliers represents a significant disadvantage. The individual examination of each outlier quickly exceeds the time frame and, thus, also leads to increasing analysis costs. The ever-increasing data volumes justify the expectation that there will also be more outliers in the data. At the same time, there are few analysts available due to the shortage of skilled labor, and this work is quite expensive. For these reasons, only specially selected outliers can be considered in detail or the interpretation of the outliers must be simplified.

By using outlier scores in detection algorithms, data are ranked according to their outlier tendency or according to the distinctness of the outliers, without considering the context. Based on the ranking, the most relevant outliers should be estimated. However, when forming the ranking, it cannot be determined which outliers are relevant for the

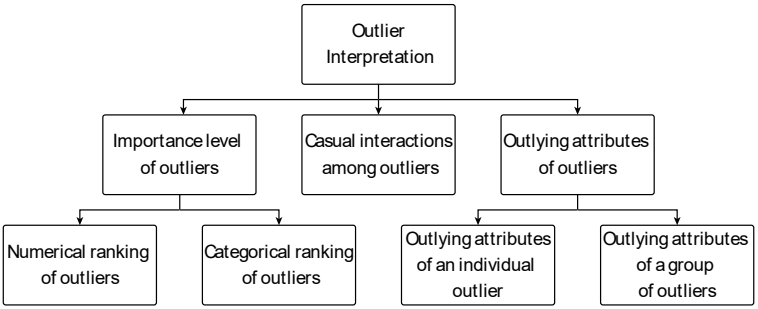

Figure 4: Categories of outlier interpretation according to Panjei, et al. (2022)

underlying research question. Particularly conspicuous outliers, such as decimal shifts resulting from entry errors, may overshadow those of technical interest. The incomplete examination of outliers also leaves the problem of undetected influences on the dataset and the question of exclusion or inclusion in analyses unresolved (cf. Section 2.2).

Outlier interpretation methods address the above problems. Panjei, et al. (2022) provide the only current overview of outlier interpretation methods and Xu, et al. (2021) a comparison of different outlier interpretation methods with a focus on algorithmic effectiveness. Panjei, et al. (2022) postulate three categories of interpretation methods: the importance level of outliers, causal interactions between outliers, and outlying attributes of outliers. An overview is shown in Figure 4.

The sources provided by Panjei, et al. (2022) particularly focus on the *interpretation of the outlying attributes of an individual outlier*. Both *numerical ranking of outliers* and *outlying attributes of a group of outliers* are addressed in only a few sources, *causal interaction among outliers* in two sources and *categorical ranking* in only one source. Panjei, et al. (2022) evaluate the given outlier interpretation aspects derived from the explanatory classification of Molnar (2019). At the same time, outliers contain information (cf. Section 2.1) and information is provided to the analyst by methods of outlier interpretation. Both are, thus, dependent on information quality. A comparison of the information quality criteria according to Lee, et al. (2002) and the aspects of Panjei, et al. (2022) is shown in Table 1.

Table 1: Comparison of criteria of outlier interpretation aspects and information quality criteria. *Added by the authors of this paper.

Outlier interpretation aspects according to Panjei, et al. (2022)	**Information quality criteria according to Lee, et al. (2002)**
Contrastive	-
Selected	Appropriate amount Ease of operation
Focus on the abnormal	-
Social	Interpretability
Truthful	Accessibility Believability Free-of-error Objectivity Reputation Security Timeliness Understandability
Consistent with prior beliefs of the explainee	-
General and probable	Completeness
Understandable for the explainee*	Concise representation Consistent representation

By comparison of the outlier interpretation aspects and the information quality criteria an overlap can be identified. This paper derives the outlier interpretation quality criteria

from the overlap and divides it into three categories: *the effect on people*, the *research question*, and the *mapping of reality*.

The dimensions named *selected*, *social,* and consis*tent with prior beliefs of the explainee* can be mapped to the category *effect on people*. This category is in clear contrast to the information quality criteria since the information itself and not the *effect on people* is the object of consideration. Nevertheless, *appropriate amount* and *ease of operation* are related to *selected* by ensuring the relevance of the information. *Interpretability* and *social* can be recognized as *connected*, which represent the possibility of the interpretation kind of the respective viewer. No information quality criterion verifies that the information is *consistent with prior beliefs of the explainee*. The *effect on people* represents a central component of information loss prevention. A possible misinterpretation of given information is prohibited by methods of technology support only by a representation type adapted to the viewer.

Contrastive and *focus on the abnormal* are assigned to the second category *research question*. Both find, by the attempted assurance of the information quality in the associated criteria, no agreement to a deviation-centered view. The intersection with the third category is also found here. In the context of the predefined question, the expressions are to be associated with a trivial validation.

The third category of *mapping of reality* includes *truthful* as well as *general and probable*. *General and probable* implies a result-centered view of the completeness of given information, which is expressed by *completeness* in the information quality criteria. *Truthful* expresses the correct representation of reality on data as well as on information level. Therefore, most of the information quality criteria can be assigned to *truthful*. The expressions of this category are to be associated with a kind of verification.

In the study published by Panjei, et al. (2022), presentation types, e.g., *concise presentation* and *consistent presentation,* find no consideration as a derived criteria. However, references exist, e.g., to “lookout” of Gupta, et al. (2019), which deal intensively with information visualization for outlier interpretation. Accordingly, the consideration of Panjei, et al. (2022) must be extended by *understandable for the explainee*, which can be assigned to the category *effect on people*. As a direct result of the comparison of the

work of Panjei, et al. (2022) and Lee, et al. (2002) in Table 1, an overview of the derived categories and dimensions of outlier interpretation quality is presented in Table 2.

In reference back to the direct use of outlier interpretation in the context of the DM, the literature confirms existing relevant information in outliers. This is already to be considered in the data pre-processing step of the DM, which is why an inclusion of suitable outliers provides the analyst with additional information in the data pre-processing and in the DM itself. The analyst's review of the information content is ensured by the *effect on humans* and *understandable for the explainee* in particular.

In summary, in the case of technology support, information loss within outliers in DM can occur at three different levels. At the data level, outliers may not be detected or may be incorrectly excluded. At the information level, information may be placed in the wrong context or declared unimportant, creating incorrect information or excluding correct information. Lastly, at the human interaction level, communication problems can lead to incorrect evaluation of the information.

Table 2: Derived outlier interpretation quality categories and dimensions

Effect on people	**Research question**	**Mapping of reality**
Selected	Contrastive	Truthful
Social	Focus on the abnormal	General and probable
Consistent with prior beliefs of the explainee		
Understandable for the explainee		

4 Exemplary Investigation of a Technology-Supported Outlier Interpretation Method in Production

In the domains of manufacturing and logistics, to the best of the authors' knowledge, technology support through outlier interpretation has not been significantly studied so far. Only Xing, et al. (2015), with an investigation of cabs in a regional traffic model, as well as Zhang, Diao and Meliou (2017), who use synthetically generated supply chain data from an airline, are in the domain of logistics. In accordance with the focus on outlying attributes of an individual outlier in the literature, this paper examines the category as an example. The case study will be based on the COIN outlier interpretation method provided by Liu, Shin and Hu (2018) and a production dataset. The COIN method builds context-based outlier scores for relevance evaluation of individual outliers.

The copper wire production line dataset, published on Kaggle by Oscar (2020), contains 16 days of recorded disturbance data from a production line. The application of the COIN method to the dataset was complemented by increasing the iteration steps to ensure convergence. In Figure 5, the calculated contextual outlier score of the dataset by the COIN method is illustrated.

According to the study conducted by Liu, Shin and Hu (2018), the detected outliers with a higher outlier score are more likely to be true outliers than those with a low value. At the same time, they describe the outliers with high outlier score as more technically interesting. For example, for rows 93, 94, 102, 103, 110, 111, 121, and 126, the low outlier scores of the values suggest the defective machine 8, on whose basis a cluster of outliers was generated. Line 72 has an outlier score of 10.5 in connection with machine 8 and at the same time marks the point in time after which machine 8 exclusively produced outliers. Thus, exactly this outlier contains information of particular interest and is assigned a high outlier score.

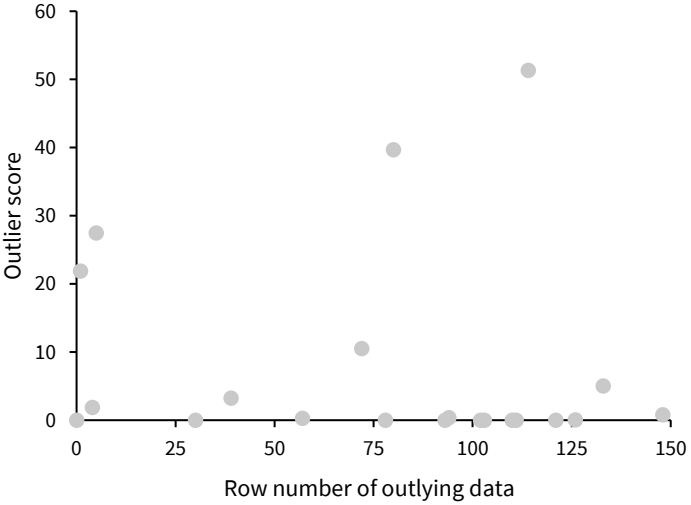

Figure 5: Outlier score of the copper wire production line dataset computed with the COIN-Method

The outliers in Figure 5 with an outlier score higher than 20 exhibit clear dependencies on multiple clusters and are, thus, also highlighted objects of study. Liu, Shin and Hu (2018) postulate for this case that they must be true outliers, but do not make any statement about included information. It must be recommended to consult experts of the production plant for these outliers.

Based on the sources mentioned above and the exemplary study presented here, it can nevertheless be shown that methods of outlier interpretation can in principle also be applied in the domains of production and logistics.

5 Summary and outlook

This paper discusses the possibilities of information loss prevention in DM data pre-processing in the domain of production logistics. In this context, relevant definitions of data and information quality were gathered, and methods of outlier interpretation were pointed out as well as classified. At the same time, the relevance of technology support in the given field was highlighted and placed in the context of information loss issues. In particular, three levels of possible information loss sources in outlier interpretation were highlighted: the data level, the information level, and the human interaction level. In addition to the literature-based argumentation on outlier interpretation methods, the exemplary case study also shows a possible evaluation of the interestingness for information stored in outliers. Due to the incomplete literature base on outlier interpretation in the domains of manufacturing and logistics as well as the DM context, a comprehensive classification of this work is difficult. Also, the exemplary domain suitability study needs to be extended by the case study with close subject matter expert contact based on various outlier interpretation methods.

In subsequent research the research field of outlier interpretation requires a framework concept for application in the KDD process. Here, the use of DM methods in outlier interpretation may result in a multi-phase implementation of a KDD process. A linkage of the KDD process with methods of outlier interpretation could generate improved analysis results by inclusion of all suitable information.

References

Aggarwal, C. C., 2017. An Introduction to Outlier Analysis. In: C. C. Aggarwal, ed. 2017. *Outlier Analysis.* 2nd ed. Cham: Springer, pp. 1–34.

Alkan, B., Vera, D. A., Ahmad, M., Ahmad, B. and Harrison, R., 2018. Complexity in Manufacturing Systems and its Measures: A Literature Review. *European Journal of Industrial Engineering,* 12(1), pp. 116–150.

Azevedo, A. and Santos, M. F., 2008. KDD, SEMMA and CRISP-DM: A Parallel Overview. *IADIS European Conference on Data Mining,* Amsterdam, The Netherlands, July 22-27, pp. 182–185.

Barnett, V., 1978. The Study of Outliers: Purpose and Model. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* [e-journal] 27(3), pp. 242–250. http://dx.doi.org/10.2307/2347159.

Beckman, R. J. and Cook, R. D., 1983. Outlier …… s. *Technometrics,* [e-journal] 25(2), pp. 119–149. http://dx.doi.org/10.1080/00401706.1983.10487840.

Ben-Gal, I., 2005. Outlier Detection. In: O. Maimon and L. Rokach, eds. 2005. *Data Mining and Knowledge Discovery Handbook.* 2nd ed. New York, NY: Springer, pp. 131–146.

Brandstein, A. and Horne, G., 1998. Data Farming: A Meta-Technique for Research in the 21st Century. In: Hoffmann, F., G. and G. Home, eds. 1998. *Maneuver Warfare Science 1998.* Quantico, Virginia: Marine Corps Combat Development Command Publication, pp. 93–99.

Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly Detection. *ACM Computing Surveys,* [e-journal] 41(3), pp. 1–58. http://dx.doi.org/10.1145/1541880.1541882.

Chen, D., Heyer, S., Ibbotson, S., Salonitis, K., Steingrímsson, J. G. and Thiede, S., 2015. Direct Digital Manufacturing: Definition, Evolution, and Sustainability Implications. *Journal of Cleaner Production,* [e-journal] 107, pp. 615–625. http://dx.doi.org/10.1016/j.jclepro.2015.05.009.

Dong, Y., Chan, R. H. and Xu, S., 2007. A Detection Statistic for Random-Valued Impulse Noise. *IEEE Transactions on Image Processing,* [e-journal] 16(4), pp. 1112–1120. http://dx.doi.org/10.1109/tip.2006.891348.

English, L. P., 1999. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits.* New York: Wiley.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* [e-journal] 17(3), pp. 37–54. http://dx.doi.org/10.1609/aimag.v17i3.1230.

García, S., Luengo, J. and Herrera, F., 2015. *Data Preprocessing in Data Mining.* Cham: Springer International.

Gebauer, M. and Windheuser, U., 2021. Strukturierte Datenanalyse, Profiling und Geschäftsregeln. In: K. Hildebrand, M. Gebauer and M. Mielke, eds. 2021. *Daten- und Informationsqualität. Die Grundlage der Digitalisierung.* 5th ed. Wiesbaden: Springer Vieweg, pp. 87–100.

Geng, L. and Hamilton, H. J., 2006. Interestingness Measures for Data Mining. *ACM Computing Surveys,* [e-journal] 38(3), pp. 1–32. http://dx.doi.org/10.1145/1132960.1132963.

Gupta, N., Eswaran, D., Shah, N., Akoglu, L. and Faloutsos, C., 2019. Beyond Outlier Detection: For Pictorial Explanation. In: M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley and G. Ifrim. *Machine Learning and Knowledge Discovery in Databases. Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Cham: Springer, pp. 122–138.

Han, J., Pei, J. and Kamber, M., 2012. *Data Mining: Concepts and Techniques.* 3rd ed. Amsterdam, Boston: Elsevier/Morgan Kaufmann.

Harding, J. A., Shahbaz, M. and Kusiak, A., 2006. Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering,* [e-journal] 128(4), pp. 969–976. http://dx.doi.org/10.1115/1.2194554.

Hawkins, D. M., 1980. *Identification of Outliers.* Dordrecht, The Netherlands: Springer.

Hodge, V. and Austin, J., 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review,* [e-journal] 22(2), pp. 85–126. http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9.

Hunker, J., Scheidler, A. A. and Rabe, M., 2020. A Systematic Classification of Database Solutions for Data Mining to Support Tasks in Supply Chains. In: 2020. *Hamburg International Conference of Logistics.* Hamburg, Germany, September 24-25, pp. 395–425.

International Organization for Standardization, 2010. *ISO/IEC/IEEE 24765-2010(E). Systems and Software Engineering – Vocabulary:* [online]. <https://cse.msu.edu/~cse435/Handouts/Standards/IEEE24765.pdf> [Accessed 13 May 2022].

International Organization for Standardization, 2015. *DIN EN ISO 9000:2015-11. Quality Management Systems – Fundamentals and Vocabulary (ISO 9000:2015); German and English Version EN ISO 9000:2015.* Berlin: Beuth Verlag.

Jacobson, R. and Aaker, D. A., 1987. The Strategic Role of Product Quality. *Journal of Marketing,* [e-journal] 51(4), pp. 31–44. http://dx.doi.org/10.1177/002224298705100404.

Kersten, W., Blecker, T. and Ringle, C. M., eds., 2020. *Data Science and Innovation in Supply Chain Management: How Data Transforms the Value Chain.* Berlin: epubli.

Khine, P. P. and Wang, Z., 2019. A Review of Polyglot Persistence in the Big Data World. *Information,* [e-journal] 10(4), p. 141–141. http://dx.doi.org/10.3390/info10040141.

Köksal, G., Batmaz, İ. and Testik, M. C., 2011. A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry. *Expert Systems with Applications,* [e-journal] 38(10), pp. 13448–13467. http://dx.doi.org/10.1016/j.eswa.2011.04.063.

Kusiak, A., 2018. Smart Manufacturing. *International Journal of Production Research,* [e-journal] 56(1-2), pp. 508–517. http://dx.doi.org/10.1080/00207543.2017.1351644.

Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y., 2002. AIMQ: A Methodology for Information Quality Assessment. *Information & Management,* [e-journal] 40(2), pp. 133–146. http://dx.doi.org/10.1016/s0378-7206(02)00043-5.

Liu, N., Shin, D. and Hu, X., 2018. *Contextual Outlier Interpretation.* Ithaca, NY: arXiv.

Mengering, B., 2021. *Erstellung eines Datenqualitätskonzeptes im Kontext der Eigenschaften von Big Data.* Master thesis. ITPL, TU Dortmund University.

Miller, H., 1996. The Multiple Dimenstions of Information Quality. *Information Systems Management,* [e-journal] 13(2), pp. 79–82. http://dx.doi.org/10.1080/10580539608906992.

Molnar, C., 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable.* Morisville, North Carolina: Lulu.

Müller, J., 2000. Funktionale und inhaltliche Aspekte der Transformation operativer Daten für analyseorientierte Informationssysteme. In: 2000. *Transformation operativer Daten zur Nutzung im Data Warehouse:* Deutscher Universitätsverlag, Wiesbaden, pp. 143–205.

North, K., 2021. Die Wissenstreppe. In: K. North, ed. 2022. *Wissensorientierte Unternehmensführung.* Wiesbaden: Springer Fachmedien, pp. 33–69.

Oliveira, P., Rodrigues, F. and Henriques, P., 2005. A Formal Definition of Data Quality Problems. In: 2005. *Proceedings of the Tenth International Conference on Information Quality.* Boston, Massachusetts, November. Cambridge, Massachusetts: Massachusetts Institute of Technology (MIT), pp. 13–26.

Oscar, 2020. *Copper Wire Production Line Dataset: Data from a Real Copper Wire Production Line for Root Cause Analysis Purposes.* [online] Available at: <https://www.kaggle.com/datasets/osroru/copper-wire-production-line-dataset> [Accessed 1 May 2022].

Panjei, E., Le Gruenwald, Leal, E., Nguyen, C. and Silvia, S., 2022. A Survey on Outlier Explanations. *The VLDB Journal : Very Large Data Bases : A Publication of the VLDB Endowment,* pp. 1–32. http://dx.doi.org/10.1007/s00778-021-00721-1.

Parsons, S., 1996. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. *IEEE Transactions on Knowledge and Data Engineering,* [e-journal] 8(3), pp. 353–372. http://dx.doi.org/10.1109/69.506705.

Pennekamp, J., Glebke, R., Henze, M., Meisen, T., Quix, C., Hai, R., Gleim, L., Niemietz, P., Rudack, M., Knape, S., Epple, A., Trauth, D., Vroomen, U., Bergs, T., Brecher, C., Buhrig-Polaczek, A., Jarke, M. and Wehrle, K., 2019. Towards an Infrastructure Enabling the Internet of Production. In: 2019. *IEEE International Conference on Industrial Cyber Physical Systems (ICPS).* Taipei, Taiwan, May 6–10, pp. 31–37.

Piro, A. and Gebauer, M., 2011. Definition von Datenarten zur konsistenten Kommunikation im Unternehmen. In: K. Hildebrand, M. Gebauer, H. Hinrichs and M. Mielke, eds. 2011. *Daten- und Informationsqualität:* Vieweg+Teubner, pp. 143–156.

Redman, T. C., 1996. *Data Quality for the Information Age.* Boston, London: Artech House.

Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A. and Schmid, J., 2011. Informationsqualität – Definitionen, Dimensionen und Begriffe. In: K. Hildebrand, ed. 2011. *Daten- und Informationsqualität. Auf dem Weg zur Information Excellence.* 2nd ed. Wiesbaden: Vieweg + Teubner, pp. 25–45.

Saake, G., Sattler, K.-U. and Heuer, A., 2019. *Datenbanken: Implementierungstechniken.* 4th ed. Frechen: MITP.

Scheidler, A. A., 2017. *Methode zur Erschließung von Wissen aus Datenmustern in Supply-Chain-Datenbanken.* Göttingen: Cuvillier.

Scheidler, A. A. and Rabe, M., 2021. Integral Verification and Validation for Knowledge Discovery Procedure Models. *International Journal of Business Intelligence and Data Mining,* [e-journal] 18(1), pp. 73–87. http://dx.doi.org/10.1504/IJBIDM.2021.111744.

Silberschatz, A. and Tuzhilin, A., 1995. On Subjective Measures of Interestingness in Knowledge Discovery. In: 1995. *Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining.* Montreal, CA, August. Menlo Park, CA: AAAI Press, pp. 275–281.

Statista, 2021. *Anteil der Wirtschaftszweige an der Bruttowertschöpfung in Deutschland im Jahr 2020*. [online] Available at: <https://de.statista.com/statistik/daten/studie/163422/umfrage/umsatz-im-verarbeitenden-gewerbe-in-deutschland-2009/> [Accessed 13 May 2022].

Stvilia, B., Gasser, L., Twidale, M. B. and Smith, L. C., 2007. A Framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology,* [e-journal] 58(12), pp. 1720–1733. http://dx.doi.org/10.1002/asi.20652.

Türkmenoglu, B., 2021. *Systematische Untersuchung der Datenqualität und -strukturen in der Supply-Chain.* Bachelor thesis. ITPL, TU Dortmund University.

Wainer, H., 1976. Robust Statistics: A Survey and Some Prescriptions. *Journal of Educational Statistics,* [e-journal] 1(4), pp. 285–312. http://dx.doi.org/10.3102/10769986001004285.

Wang, R. Y. and Strong, D. M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* [e-journal] 12(4), pp. 5–33. http://dx.doi.org/10.1080/07421222.1996.11518099.

Wenzel, S. and Stolipin, J., 2017. Nachnutzung von Wissen in Simulationsstudien. In: S. Wenzel and T. Peter, eds. 2017. *Simulation in Produktion und Logistik 2017.* Kassel: kassel university press, pp. 209–218.

Wirth, R. and Hipp, J., 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. In: 2000. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining.* Manchester, UK, 1–13 April. London, UK: Springer, pp. 29–39.

Xing, L., Wang, W., Xue, G., Yu, H., Chi, X. and Dai, W., 2015. Discovering Traffic Outlier Causal Relationship Based on Anomalous DAG. In: Y. Tan, Y. Shi, F. Buarque, A. Gelbukh, S. Das and A. Engelbrecht, eds. 2015. *Advances in Swarm and Computational Intelligence.* Cham: Springer International, pp. 71–80.

Xu, H., Wang, Y., Jian, S., Huang, Z., Wang, Y., Liu, N. and Li, F., 2021. Beyond Outlier Detection: Outlier Interpretation by Attention-Guided Triplet Deviation Network. In: 2021. *Proceedings of the International World Wide Web Conference,* April 19-23. New York, NY, pp. 1328–1339.

Zhang, H., Diao, Y. and Meliou, A., 2017. *Exstream: Explaining Anomalies in Event Stream Monitoring. [e-book].* <https://par.nsf.gov/servlets/purl/10033440> [Accessed 13 May 2022].