Proceedings of the Hamburg International Conference of Logistics (HICL) – 33

Lorenz Kiebler, Nikolas Ulrich Moroff and Jens Jakob Jacobsen

Preliminary Analysis on Data Quality for ML Applications

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

Preliminary Analysis on Data Quality for ML Applications

Lorenz Kiebler[1], Nikolas Ulrich Moroff[1] and Jens Jakob Jacobsen[1]

1 – Fraunhofer Institute for Material Flow and Logistics

Purpose: *This publication investigates preliminary data quality analyses to estimate the efforts and expected results of the use of data sets for ML solutions already in the data understanding phase of an implementation. Knowledge about the necessary data cleaning efforts and result qualities allows potentials to be estimated early in the process.*

Methodology: *Through a literature research, characteristics of a time series as well as methods of data cleaning are analysed. Based on the results, a test environment is implemented in Python, enabling the evaluation of individual methods using sample data sets from the process industry and comparing them with different error analyses.*

Findings: *The publication describes a detailed overview of data cleaning procedures and addresses a first Indication of a connection between the final achievable forecast quality and the degree of error of the original data set. Insights into the influence of the choice of preprocessing method on the achievable quality of the AI-based forecast can be concluded.*

Originality: *Within the publication, the link between data characteristics in time series and preprocessing methods is established to draw conclusions in advance about the quality improvement to be expected from selected data cleaning methods and to provide decision support for the selection of the method.*

First received: 03. Mar 2022 Revised: 22. Aug 2022 Accepted: 25. Aug 2022

1 Motivation

Corporate supply chains and business units have changed significantly in recent decades due to increasing globalization and local constraints. Corporate networks are becoming more complex and at the same time need to become more responsive to meet growing and changing market demands. As a result, planning inventory, production capacity, and transportation is becoming increasingly important.

The focus of companies is shifting more and more from pure product competence to flexible customer service, speed, and on-time deliveries at the lowest possible cost (Wassermann, 2013). Consequently, the entire production process and the logistic periphery must react immediately to market fluctuations if they cannot be planned in advance with sufficient quality using advance planning methods (Erben and Romeike, 2003). In order to improve the quality of planning despite growing market demands, data-driven methods are increasingly used, especially in the areas of demand forecasting, production, pricing, and delivery (Dash, et al., 2019).

The level of integration and operational use of these data-driven technologies varies greatly by company and by industry. One industry with comparably low implementation levels of Artificial Intelligence (AI) or Big Data (BD) technologies is the process industry (Winter and Peters, 2019). Despite the strong suitability of the basic conditions in the process industry for AI use cases due to the already strong data collection by distributed control systems in the past (Ge, et al., 2017), there are large differences in implementation by sector and application (see Fornasiero, et al., 2021) and a broad and systematic strategy for implementation is not yet in place (Wostmann, et al., 2020).

For a uniform understanding, the term AI according to Kreutzer et al. is used within this article as a top-level term for Machine Learning and Deep Learning. Machine learning describes the first stage of knowledge generation with the aid of self-learning algorithms and is supplemented by the area of deep learning, which deals with more in-depth learning methods. Here, the algorithm works with larger data sets and can achieve better quality with decreasing result transparency due to higher complexity handling. As a related topic, the term Big Data within this article describes a mostly quantitatively large amount of data (used in the context of AI applications) that additionally contains qualitative information for processes or company operations (Kreutzer et al., 2019).

This paper focuses on the process industry as an example to analyse the reasons for the restrained use of new technologies and to identify possible solutions for individual problem areas. Findings and results in this paper with reference to the process industry originate in the research project AI-CUBE. In this European Union (EU) funded project, the focus is to help exploit and optimise the potential of AI and BD in the European process industry. To achieve this, the project AI-CUBE has the following specific objectives:

- Create a multi-dimensional AI and BD map (or "CUBE") that identifies available best practices and assesses the current state and level of penetration of AI and BD in different organizational processes
- Identify AI and BD solution white spaces that can be covered by adapting best practices from other process industries, and create an adaptation roadmap
- Define the data requirements and capabilities as well as research and innovation requirements for future AI and BD business cases emerging across process industry sectors.

This paper primarily addresses the topic of data requirements and explores preliminary analyses to estimate the effort and expected results of using a dataset for ML solutions already in the data understanding phase of an implementation process. In particular, the focus is on data quality, which according to a study in the AI-CUBE project is one of the main obstacles to the successful use of data-driven methods.

The research question to be addressed within the article is thus: Can a data set be tested for usability within an AI application with a preliminary analysis or can the effort required to prepare the data set be measured with the help of a key performance indicator? The goal here is to measure data set characteristics and to link them to the necessary data preparation effort. Thus, the aim is to develop a concrete translation scale of data characteristics and processing efforts, which is currently not available in sufficient quality. Due to the heterogeneity of the research field, the focus is limited to a data set from the process industry.

In Chapter 2, the term data quality is first introduced in the AI context and, building on this, the classic implementation processes of an AI application with the different phases of data preprocessing are presented. The focus of the paper here is on data cleaning. By means of a literature review, characteristics of data sets as well as methods of data cleaning are analysed (Chapter 3). Based on these results, a test environment in Python is implemented, which allows to evaluate individual methods using sample data sets and to compare them with different error analyses (Chapter 4).

The publication thus describes a detailed overview of possible data cleaning methods and addresses the relationship between the ultimately achievable forecast quality and the degree of error of the original data set. From this, conclusions can be drawn about the influence of the choice of preprocessing method on the achievable quality of the AI-based forecast.

2 Data Quality in AI implementations

The quality of the database is both a requirement and a challenge for AI implementations. As will be shown in the following, ensuring the necessary data quality is a major implementation barrier both in general and in the process industry and therefore also represents an important and demanding step in the adoption process.

2.1 Data quality as an implementation barrier (focus on the process industry)

For AI implementation projects as for every information technology (IT) or information system (IS) adaptation, there are a lot of different barriers and challenges that can hamper the implementation or even cause it to fail. A review on existing studies on implementation barriers for data driven solutions like artificial intelligence or big data technologies shows that the already identified and validated barriers can be structured in the three categories of *organizational and environment related barriers*, *technological and data related barriers* as well as *human related barriers* (Alsheibani, Cheung and Messom, 2019; Dasgupta and Wendler, 2019; Moktadir, et al., 2019; Eager, et al., 2020).

The occurrence and relevance of the different categories and individual barriers changes with the area of application and the industry or company specific environment. Through a survey of users and solution providers in the process industry as part of the AI Cube research project it was found that the most important and influential challenge here is the barrier of data quality in the context of technology and data-related barriers.

Building on this observation, factors and concepts enabling the successful adaptation of data-driven solutions can also be positively formulated from the observed barriers. In the literature these factors are summarized as AI readiness factors (Najdawi, 2020; Jöhnk, Weißert and Wyrtki, 2021) or success factors (Bole, et al., 2015; Hughes, Rana and Dwivedi, 2020). While some of the identified factors are well known for IT/IS implementations (e.g., top management support, financial budget), some factors are more technology specific. Regarding the process industry in Europe, the survey from the AI Cube research project shows that besides existing personnel and strategy related factors, the two data related enabling factors *data availability* and *data quality* are rated as most influential. Missing data quality is considered a critical challenge for data-based solutions because of its strong influence on the final performance and suitability (Jöhnk, Weißert and Wyrtki, 2021).

Recognizing the potential barriers and success factors can improve AI adoption and boost the overall development and adoption rates of AI and Machine Learning solutions (Alsheibani, Cheung and Messom, 2019). In the specific case of data quality, it is also important to assess the quality of the available database as early as possible, to be able to predict expenses that will arise later in the process. Both for the development of new solutions and for the transfer of existing solutions, the creation of the necessary data quality involves great effort.

2.2 Data quality in AI implementation process models

To achieve the necessary data quality, special focus must be placed on data preprocessing within the implementation process of an AI application. As schnell

eckenberg and Moroff (2021) have shown based on a literature review, most Machine Learning (ML) and AI developments and implementations follow iterative methodologies that are closely based on the Cross Industry Standard Process for Data Mining (CRISP-DM) or other knowledge discovery in data bases (KDD) process models such as SEMMA or the basic KDD process. The different process models, which have been developed into different forms in recent years (SEMMA, CRISP-DM etc.) as presented in Figure 1, are very similar in many phases and include many of the same central aspects and sequences. Universal and generic steps can also be extracted here, even if there are of course deviations and specifications to specific use cases (Kurgan and Musilek, 2006). Due to this strong use of KDD related process models for the management of AI and ML developments in research and in industrial applications, these processes are also considered in a focused manner within the scope of this work.

However, the data preprocessing phase can be found within each model, which shows the special importance of this phases. In addition, data preprocessing is also the most demanding phase, as it is here that the quality of the AI application is significantly determined.

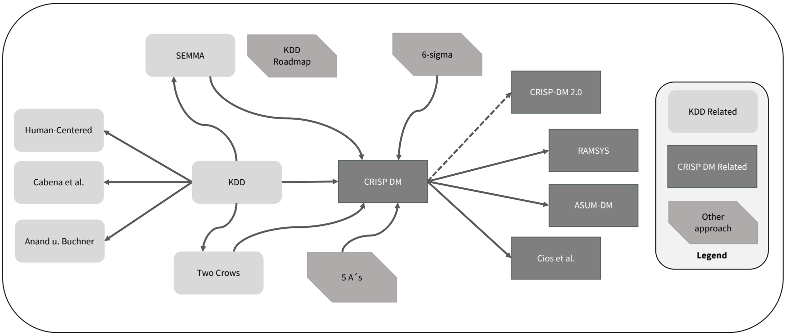


Figure 1: Development of KDD process models (Martinez-Plumed, et al., 2020; Plotnikova, Dumas and Milani, 2020)

Kurgan and Musilek found in a survey in 2006 that most of the effort within an AI project goes into data preparation as shown in Figure 2. The area of data preparation is further categorized into the sub-areas of data integration, data cleaning and data transformation and deals with the preparation of data and pursues the goal of generating usable data content from raw data. Here, techniques are used to minimize problems caused by data noise, inconsistent scales, or missing values (Data Cleaning and Transformation) (Alasadi and Bhaya, 2017).

To harmonize the concept of data quality and the closely related data preprocessing, the respective application must be considered (Jayawardene, Sadiq and Indulska, 2015). In general, two concepts can be followed when characterizing data quality: The *informational focus* (Data definition, data presentation and content data quality) and the *user-based focus* (Features of the dataset). In our publication, we restrict ourselves to the informational focus and especially on the aspect of *content-related data quality* and here we establish the connection to the data preparation process. We deliberately exclude the user-related approach, since a much broader focus of data quality is assessed here (e.g., accessibility, comprehensibility, or interpretability), and thus the transfer into an evaluable KPI system is a further challenge or research need.

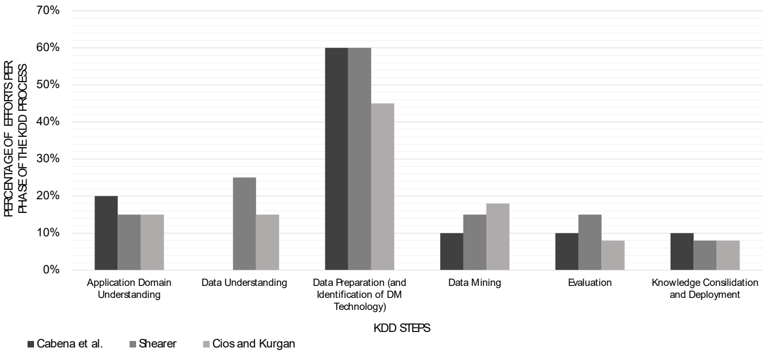


Figure 2: Relative effort spent on specific steps in the KDD process (Kurgan and Musilek, 2006)

The informational focus approach to data quality assessment focuses on assessing data definition, data set content quality, and data presentation. These three modules provide the basic framework for defining information-related data quality and were further detailed by English (1998). The first module focuses on the data collection framework. Only adequately specified data can be used to measure quality. The second and third module focus on content correctness in terms of uniqueness and completeness and with the availability of the data. Parameters for this part are, for example, the time of availability and compliance with the format.

In the further course of the work, we are particularly interested in module two, since the actual quality of the data set is checked here (freedom from errors, completeness, etc.). This is directly related to data cleaning, where existing data points with errors are corrected using different methods. The remaining phases of the data preparation (transformation, normalization, etc.) focus on the extraction of information by connecting different data sets or the transformation into processable formats for the algorithm. However, they do not revise the core information of the data set as is done in data cleaning, for this reason, data cleaning is usually performed before the rest of the data preparation steps (Brownlee, 2020).

The usability of a data set is thus directly related to the success of data cleaning, since the correctness of the data is the basis for the success of an ML model. At the same time, besides the high potential, there is also the highest risk in data cleaning, since due to adjustments the data set can be processed too much and thus the core information can be lost. Therefore, to reduce the effort for data preprocessing or explicitly data cleaning, an estimation of whether a dataset can be successfully revised during data cleaning would be helpful to avoid costly and unsuccessful data reprocessing. When analysing the current scientific literature for thresholds of statistical quantities (number of missing values, number of outliers, etc.) that give you information about when a data set can be used for a machine learning model, the formulations remain only general and vague (Brownlee, 2020). The following article will therefore provide an indication of how effective data cleaning is at a given level of uncertainty, with the aid of statistical parameters. It is important here that the outliers considered in the further course are regarded as clear errors in the data series and not as information to be interpreted. Here it must be considered whether e.g., a 20% share of outliers must not already be interpreted as information pattern, so that no correction takes place.

Due to the high importance of data cleaning for AI implementations both in terms of effort and cost as well as in terms of result quality and the associated perceived usability of the developed solutions, the thresholds for evaluation support proposed here also have a major impact on industrial practice. Especially in the process industry, there are often problems with incomplete and error-prone data (Khaydarov, et al., 2020) despite the fundamentally very extensive data collection and evaluation already in place. Thus, this industry represents a good use case for considering the influence of data quality on ML implementations. As the results of the survey on existing barriers show, the hurdle of insufficient data quality and high data complexity is also well known in the industry, so that methods for assessment and process optimization can be integrated very well into currently existing implementation or technology transfer efforts here. As shown in the context of a self-examination of the European Process Industry by the A.SPIRE association, especially such methods for dealing with insufficient data quality in the industrial environment are not yet available and require a deeper consideration (Winter and Peters, 2019).

3 Methods for data cleaning in preprocessing and basic characteristics of time series

After discussing the challenge of data quality and the overall context of data cleaning in the data preprocessing process in the previous section, this section will focus on specific data cleaning methods that enable a dataset to be optimised for a Machine Learning application. For content limitation, only methods for numerical values or time series values that can be used in demand forecasting will be discussed in this paper. This is a special case in the field of data preprocessing since data points influence each other over time and are not independent of each other.

The method used in this chapter to conduct a literature review on different data cleaning methods is based on the methodology formulated by Randolph (Randolph, 2009). The steps presented there for developing a complete literature review and its goals are themselves based on Cooper's taxonomy (Cooper, 1988). The five phases of the research that were followed are: 1. Problem formulation, 2. Data collection 3. Data evaluation, 4. Analysis and interpretation, 5. Public presentation. The goal of the research is the analysis and synthesis of scientific findings. According to Cooper's taxonomy, therefore, the objective is to be found in the "Research Outcome". The problem to be addressed hereby is: What types of time series error types exist and what techniques are suitable for correcting them. The data is collected in the scientific databases Web of Science, Scopus and IEEEexplore, since these cover the topics Computer Science, Engineering and Economics. To ensure the relevance of the publications used, the data base is limited based on timeliness. Starting from 2016, there is a continuous increase in publications, therefore the year 2016 is set as the limit for the timeliness of the publications. The evaluation follows the research question and examines the publications for the types of errors in time series data mentioned in each case, procedures for dealing with the errors mentioned, and correlations between usability of the data and the degree of errors. In the following, these results will be summarized, and the further procedure is derived.

When errors in time series are mentioned, reference is usually made to the three categories of *noise*, *outliers*, and *missing values*. These errors make it difficult to use the data set in data driven applications and need to be handled in the phase of preprocessing. The extend of these errors in a data set can be determined and analysed directly with statistical quantities, so that a direct overview of the current state of the data set can be given. The understanding of the three error categories used in this paper and the statistical quantities for identifying the errors are described in the following:

- **Outliers**: Values within a data set that deviate too far from other observations within the same data set. Outliers can be determined by analysing the number of values that are outside the expected value. These numbers can still be legitimate data points, which makes identification more difficult (Rustum and Adeloye, 2007).
- **Noise:** Random or irregular fluctuation of a measured value caused by random errors overlaying the original true value (Lazzeri, 2020) that can be identified though a high standard deviation
- **Missing values:** Observations for which no value is stored. This can have various reasons, for example, values can be lost during transmission or there can be an error in the recording system. Identification of missing values is possible through the number of undefined or unrepresentable values within the time series (Rustum and Adeloye, 2007).

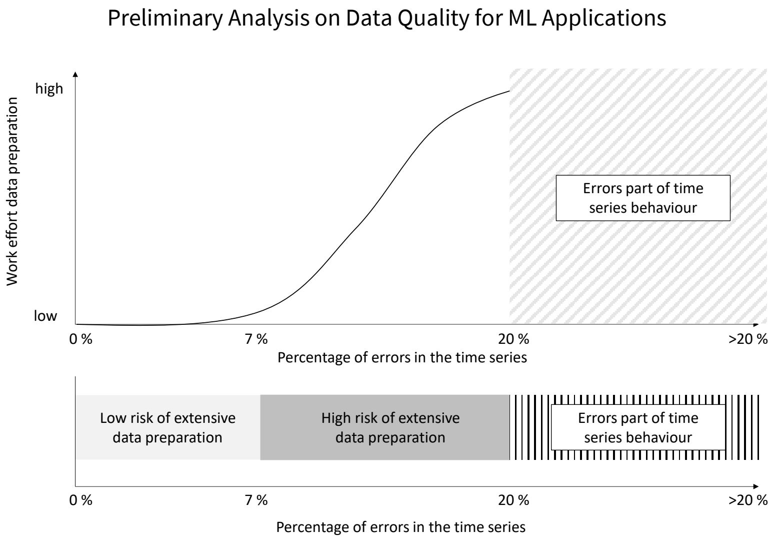


Figure 3: Presumed correlation of percentage of errors and work effort

One strategy for estimating the effort of data preprocessing that will be presented in this paper is thus to find a threshold value that evaluates the effectiveness of a data preparation (e.g., "up to 8% outliers can be in a data set to use it"). This pre-analysis can help in estimating the usability of the dataset after data cleaning. In Figure 3, this concept is graphically broken down and explained in detail based on the error types. In the field of time series analysis, the "Replace" strategy is mostly used to avoid generating gaps within a time series. Based on the literature review, the most frequently used methods identified for the replace approach for the different errors were identified (see Table 1). A detailed mathematical description of these individual methods will not be given in this paper, as the focus is less on the concrete mode of action and more on the results of the various methods relative to each other. More information on the described methods can be found in the literature of Garcia, Luengo and Herrera (2015).

Table 1: Methods for Replace-Strategy per error category

Error category	Methods for replacing erroneous data points	
Missing Values	Mean and Median Imputation Forward/ Backward Imputation Linear Interpolation	Moving Average Interpolation und Rolling Median Interpolation Weighted Moving Average Interpolation
Outliers	Mean und Median Imputation Moving Average und Rolling Median Imputation Linear Interpolation	Weighted Moving Average Imputation Nearest-Neighbour Regression Forward/ Backward Imputation
Noise	Moving Average Imputation k-Nearest neighbour Regression	Weighted Moving Average Imputation

Apart from these error types and methods for improving the data quality of the original time series, broader strategies like filtering of different parts of the time series or differentiation and integration can be used depending on the basic characteristics of the time series before the actual data cleaning: Time series are composed of different characteristics, whose superposition results in the pattern of the time series (see Figure 4). The characteristics are *trend*, *seasonality*, and *structural breaks*. The effectiveness of the individual strategies depending on the timeline characteristics is examined below. (Montgomery, 2015; Lazzeri, 2020; Mukhiya, 2020)

After identifying and presenting the most common types of errors and existing methods for correcting these errors as well as basic characteristics of time series, the following chapter shows how the effects of these methods on different time series were collected and analysed.

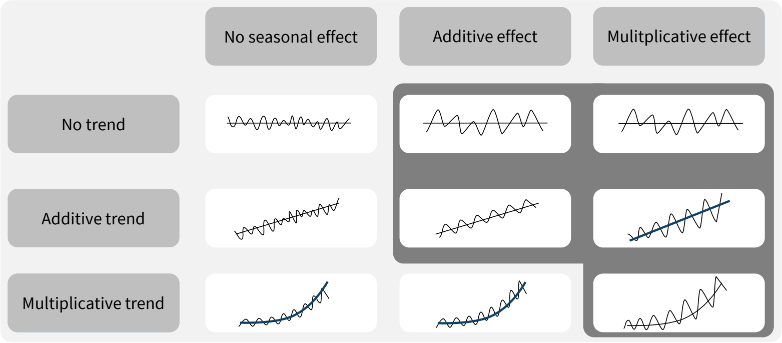


Figure 4: Demand Patterns resulting from trend and seasonal components

4 Experiments and prognosis results

Due to the fact, that the demand forecasting is one of the most important research subjects in supply chain management (Bertolini, et al., 2021), the optimization of the data basis and the prediction of potentially achievable forecast qualities are important when AI capabilities are to be implemented in the industry. As time series provide the basis for forecasting (Koller, 2014), it is important to determine the influence of their different factors and characteristics on the accuracy of the forecast and to explore strategies to further improve the results. This Improved forecast accuracy corresponds directly to more effective planning and better availability of inventory (Kuhn and Hellingrath, 2002; Crone, 2010).

4.1 Experimental plan

To determine correlations between the mentioned error types, cleaning methods and forecast accuracy, as well as between the characteristics of the time series and forecast accuracy, the following model is established: A selection of sufficiently large timeseries of demands of individual products from the chemical sector in the process industry with different combinations of characteristics (marked grey in the following illustration) were chosen as data basis.

The products selected here each represent specific product groups with comparable demand patterns. The individual error types analysed further were artificially added to each of these time series. In this way a new data basis is generated from the original time series with different error types of multistage occurrences (phase 1) that can be cleaned and used for an ML based prognosis. After performing this experimental design to determine the influence of error and data cleaning on the forecast performance, different approaches are formulated to eliminate the influence of the basic time series characteristics on the forecast (phase 2). While phase 1 will show the improvement of the data quality and the corresponding prognosis quality through data cleaning methods, phase 2 will show the influence of the characteristic-depending cleaning strategies.

The artificially implemented errors in the data sets are processed with the listed data cleaning methods. Outlier and missing values are replaced by imputation methods and noise is removed using smoothing methods. In the subsequent step, the cleaned data basis is used for forecasting with a selected ML model. A random forest regressor is used as the Machine Leaning model, which offers the advantage of low training time while maintaining high forecasting accuracy (Darapaneni, et al., 2019; Du Ni, Xiao and Lim, 2020; Lu, et al., 2021). At the same time, it is one of the most widely used ML methods in time series forecasting and therefore represents a method that is particularly representative in decision trees. The results worked out here in the paper are not transferable 1:1 to other procedures, since they require other conditions in the processing of the data basis particularly with fundamentally different model architectures (e.g., vector-based). Parameter tuning of the random forest is performed via a random grid search in which a suitable parameter setting is determined for the given data set. Depending on the results in both phases for the different error rates, general threshold values can be estimated up to which recognizable error rates in the original data a sufficient data quality for subsequent AI forecasts can be achieved by data cleaning. Both the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) were calculated as error values. The RMSE indicates how much the forecast deviates on average from the actual values and the MAPE allows a good interpretation, or serves as an indicator for the relationship between the forecast and the actual observation. It is clearly interpretable and at the same time dimensionless. In the following the general calculation formulas of the error values are shown.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\hat{y}_i} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

In the formulas, $\hat{y}_i$ stands for the actual value of the demand, y_i for the predicted value, and n for the number of predicted values.

The experiments are performed based on a single-factor model. The different data cleaning methods form the individual parameters of the respective experiment plan. The number of errors of one type in the time series set represents the level intervals of the experiment and the resulting prediction accuracy maps the effect to be investigated. (Siebertz et al. 2017, S. 1–7) In this way, the effect on forecast quality can be determined for each error type and data cleaning method. Not considered in this experimental design is the interaction from different error types within the same time series.

In both phases of the experiment described in Figure 5, twenty variations of manipulation were integrated into the original time series for each error type (missing values, noise, outliers). The error rates were increased step by step from 0% (original data set) up to 20%. For example, this means that in the manipulated time series for 20% missing values, 20% of the data points are missing. Per phase, this leads to an experimental plan of three times twenty run-throughs per original time series. While this approach evaluates the effectiveness of the data cleaning methods in phase 1 and allows initial error thresholds to be determined for the successful use of the methods, phase 2 focuses on the further improvement of preprocessing through the upstream manipulation of the data series using characteristic-dependent strategies and its influence on the identified thresholds.

In Summary, ten original time series were manipulated with twenty proportion variations per error type, leading to 600 time series per phase that are each cleaned with 3 types of data cleaning methods (constant, linear, non-linear) with 2 specific methods each. Combined, in each phase, 3600 prognoses were calculated.

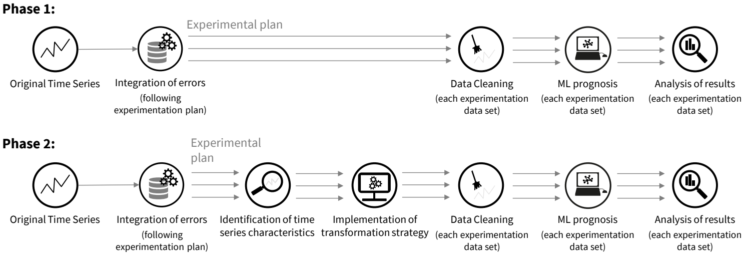

Phase 1:
Experimental plan
Original Time Series
Integration of errors (following experimentation plan)
Data Cleaning (each experimentation data set)
ML prognosis (each experimentation data set)
Analysis of results (each experimentation data set)
Phase 2:
Experimental plan
Original Time Series
Integration of errors (following experimentation plan)
Identification of time series characteristics
Implementation of transformation strategy
Data Cleaning (each experimentation data set)
ML prognosis (each experimentation data set)
Analysis of results (each experimentation data set)

Figure 5: Phases of the experiment

4.2 Results from the different prognoses and phases

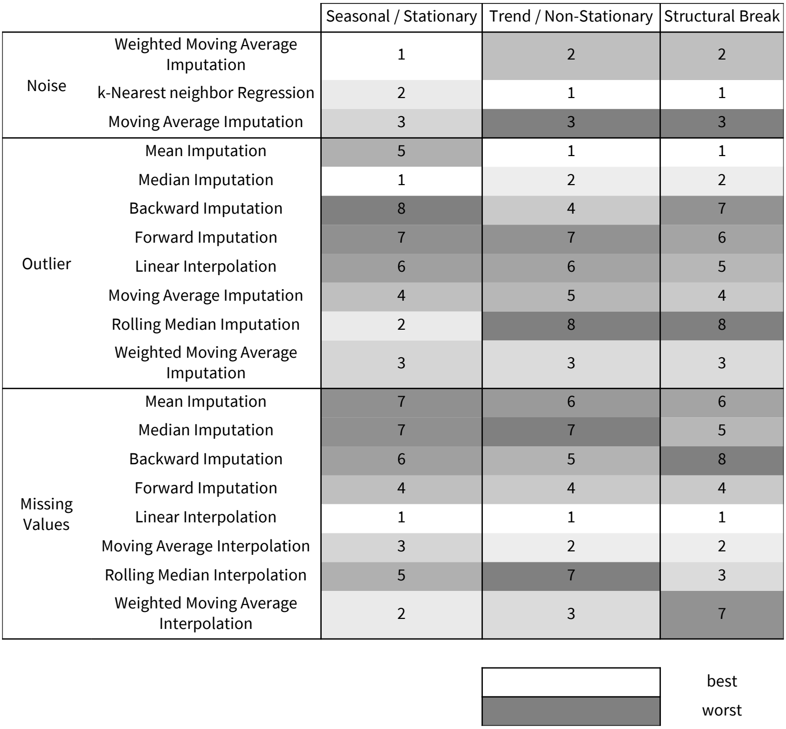
In this chapter, results of the observations from both experimental phases are presented and underlying explanations and interrelationships are identified.

Results of Phase 1: Data Cleaning without characteristic-depending manipulation

Regarding missing values, a clear trend can be observed, that the forecast quality decreases as the number of missing values increases. This behaviour is independent of the time series character and was observed in every test series. In the area of outliers, a similar behaviour emerges, the higher the number of outliers, the poorer the forecast quality. However, an interesting phenomenon is that trended time series can be forecasted with constant quality if the outliers are in the direction of the future trend development. As a last error type, an artificial noise was integrated into the data set. Here, as with the missing values, there was a decreasing forecast quality with an increasing proportion of noise in the data set.

Looking at the achieved improvements in forecast quality through the various data cleaning methods, especially the linear *interpolation* and the *moving average* could achieve good results (across time series). When using the *Weighted Moving Average*, one can see a clear difference between the forecast quality for seasonal time series (second best method) and time series with structural break (worst method). Thus, a preliminary analysis of the time series is essential here. In the following Figure 6 the used data cleaning methods are ranked by quality in the range of the considered error type (1 best result and maximum value worst result).

In addition to the evaluation of the data cleaning methods on the test data sets, limit values were also determined to estimate the effort of a data preparation with the help of a preliminary analysis. The limit values were connected to the time series characteristics (Structural Break, Trend, Seasonal) in the analysis to enable a first general statement.



		Seasonal / Stationary	Trend / Non-Stationary	Structural Break
Noise	Weighted Moving Average Imputation	1	2	2
	k-Nearest neighbor Regression	2	1	1
	Moving Average Imputation	3	3	3
Outlier	Mean Imputation	5	1	1
	Median Imputation	1	2	2
	Backward Imputation	8	4	7
	Forward Imputation	7	7	6
	Linear Interpolation	6	6	5
	Moving Average Imputation	4	5	4
	Rolling Median Imputation	2	8	8
	Weighted Moving Average Imputation	3	3	3
Missing Values	Mean Imputation	7	6	6
	Median Imputation	7	7	5
	Backward Imputation	6	5	8
	Forward Imputation	4	4	4
	Linear Interpolation	1	1	1
	Moving Average Interpolation	3	2	2
	Rolling Median Interpolation	5	7	3
	Weighted Moving Average Interpolation	2	3	7

□ best
■ worst

Figure 6: Achieved forecast improvement by data cleaning method

The goal of the analysis was to identify the so-called *elbow area* in the time series, which describes a percentage of errors in the data set. From the *elbow point*, the data set can no longer be corrected using the data cleaning methods to obtain a similar result as with the original data (loss of forecast performance). The following Figure 7 shows a representative result diagram from the experimental plan for the exemplary determination of the *elbow area*. The highlighted area indicates the percentage of error after which the forecast quality decreases significantly.

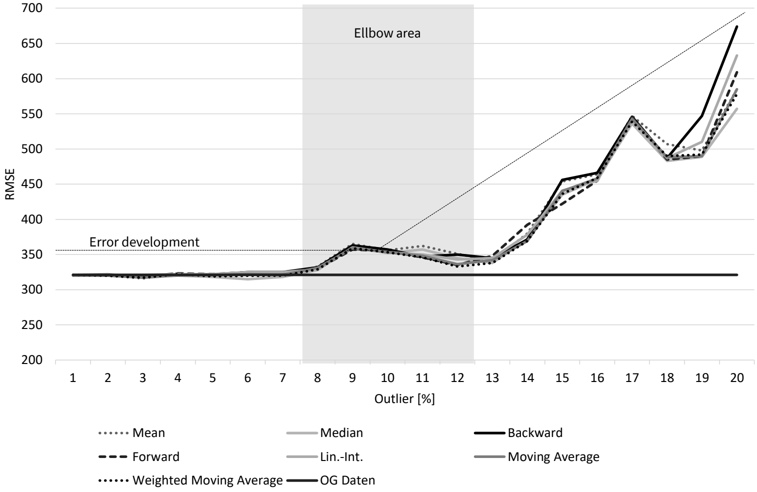


Figure 7: Identification of the elbow area

In the range of the error type "missing values", an elbow range can be identified at approx. 6-8%. As soon as the number of missing values exceeds 8%, the forecast quality decreases significantly. The elbow range is to be defined the same over all timeseries types and thus in the case of the test scenarios independent of the timeseries characteristics. In the analysis of outliers, the range was identified at 5-7% for seasonal (stationary) time series and at 3-6% for trending (non-stationary) time series.

So here we can summarize that a small number of outliers have a positive influence for the forecast results (result from the previous analysis), but as soon as the error percentage is above 6%, the forecast quality reduces again. In the area of outliers, only a valid calculation of the elbow range for seasonal products could be performed (7-12%), since the remaining time series did not show a common pattern. In general, the time series specific characteristics must be considered for the limit values, so that the determined values (shown in Figure 8) cannot be used as limit values, but as orientation values. This individuality will be used in phase 2 to further improve the forecast quality despite high error rates.

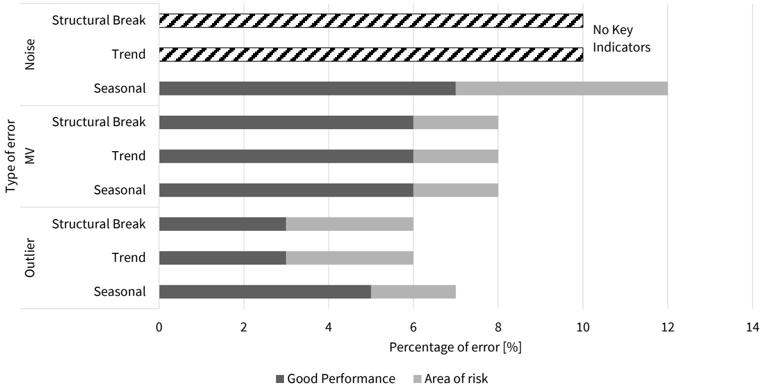


Figure 8: Limit values of the elbow areas by time series specific characteristics

Results from phase 2 (Usage of characteristic-depending strategies)

To account for the individuality of the time series, three strategies depending on the time series were used to perform preprocessing of the data before applying data cleaning procedures. The data basis is thus already prepared in advance.

(A) Trend or Seasonal: Derive, perform forecast, integrate

(B) High Noise: apply median filter,

(C) Structural Break: split data set, use actual data to train and test

In the following Table 2, the potentials are briefly shown by a time series specific preparation (always compared to the initial accuracy without data cleaning and other preprocessing strategies). The evaluation shows the percentage development of the forecast quality compared to a non-preprocessed data set. The positive development due to the integration of contextual knowledge can be seen well, especially when considering the time series with structural break, since here the time series has experienced a break in content and the sections must therefore be considered independently.

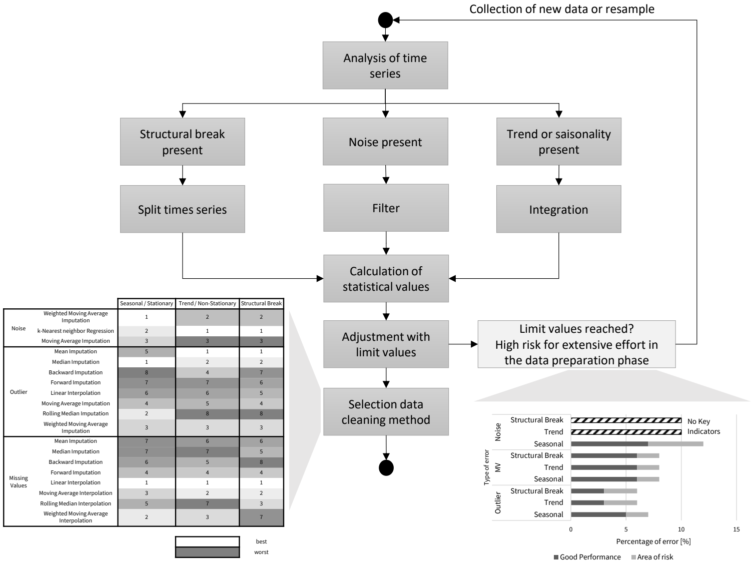
Table 2: Potentials of the different strategies for prognosis accuracy

Characteristic-depending strategy	Development Baseline accuracy	Development Accuracy after data Cleaning
A	+1%	+1-10% (depending on the error type)
B	~0%	+10%
C	+25%	+30%

Based on the identified potentials, a flowchart is drawn up in the following section (see Figure 9) that provides a strategy for analysing data sets in advance. In addition, the procedure offers the possibility to make an estimation based on the calculated critical values whether there is a high risk for an effort-intensive preprocessing.

4.3 Recommendations for the use of data cleaning methods and preprocessing efforts

The process of preliminary analysis using the flowchart presented in Figure 9 starts with the general analysis (visual or analytical) of the time series. Based on the identified characteristics (noise, trend, season, or structural break), a strategy for preprocessing is proposed (split time series, filter, integration). After preprocessing, the error types of the time series are evaluated (e.g., number of missing values) and compared with the limit values from the previous analyses. If the critical values are exceeded, it can be assumed that the effort in data preprocessing will be much more intensive. Based on the results, the original data set can be adjusted again (collection or addition of data points), or one or more data cleaning methods can be selected for further processing. In the following figure, the process is described again graphically with the help of a flowchart.



		Seasonal / Stationary	Trend / Non-Stationary	Structural Break
Noise	Weighted Moving Average Imputation	1	2	2
	k-Nearest neighbor Regression	2	1	1
	Moving Average Imputation	3	3	3
Outlier	Mean Imputation	5	1	1
	Median Imputation	1	2	2
	Backward Imputation	8	4	7
	Forward Imputation	7	7	6
	Linear Interpolation	6	6	5
	Moving Average Imputation	4	5	4
	Rolling Median Imputation	2	8	8
	Weighted Moving Average Imputation	3	3	3
Missing Values	Mean Imputation	7	6	6
	Median Imputation	7	7	5
	Backward Imputation	6	5	8
	Forward Imputation	4	4	4
	Linear Interpolation	1	1	1
	Moving Average Interpolation	3	2	2
	Rolling Median Interpolation	5	7	3
	Weighted Moving Average Interpolation	2	3	7

□ best
■ worst

Figure 9: Flowchart for strategy for analysing data sets in advance

It should be added that the flow chart should be used as an orientation and does not claim to be complete in the selection of preliminary analyses and data cleaning methods. It is only intended to provide a systematic procedure for estimating the effort required for the data preparation phase.

5 Conclusion

The analyses presented in this paper provide recommendations for dealing with the existing errors and characteristics of time series so that they can be cleaned optimally in the preprocessing phase. While the identified data cleaning methods for individual error types (noise, missing values, outliers) improve the subsequent ML-based prediction in the presence of errors with great reliability, the characteristic-based strategies improve the forecast results noticeably only for certain characteristics. While especially time series with structural breaks can be preprocessed quite effectively for subsequent data cleaning by using strategies beforehand, the influences on the results are not so reliably attainable in the case of trends, seasonality, or high noises. Nevertheless, positive influences have also been observed for these time series characteristics. Especially the combination of characteristic-depending strategies like splitting, filtering, or integration with the classical methods of data cleaning influence the prediction results well.

By means of the flowchart shown in Figure 9, a procedure model has been developed which provides an assessment of the suitability of time series for forecasting with little effort through limit values per error type (Figure 8) and, in the event of a positive outcome, provides indications of the next steps by evaluating the effectiveness data cleaning methods in Figure 6.

It should be emphasised, however, that the limit values and suitability descriptions presented were determined based on the underlying time series originating from the process industry, so that possible structural properties or errors may have been included here. Thus, they are only first indicators of these limits, which must be verified in the future by other data sets from other application areas or industries. Nevertheless, the initial evaluations on this data set show the described good indicators for the determination of the limit values. By conducting comparable experiments in other application areas or industrial environments in further research work, it will be possible to identify both the reproducibility of the results and any environment-related influencing factors.

In particular, the transferability to other algorithms or compositions of the original data represents a further research need for the future, as the effectiveness and the statements were only assessed for the random forest approach used initially. Through a growing body of experience regarding the behaviour and influence of data cleaning procedures on forecast quality under different environmental and technological influences, further patterns can be identified here which could be related to potential underlying causalities.

However, the initial presentation of these limit values shows that there is the possibility to make later efforts more economically evaluable. Since data quality is a significant barrier for AI implementation projects generally and explicitly in the process industry (see Chapter 2.1) and the creation of suitable data quality takes up a large part of the effort in the implementation process (see Chapter 2.2), both the presentation of solution strategies for the creation of sufficient data quality and the indication of the expected effort and the expected results are important building blocks of an implementation guideline. By integrating this proactive assessment of the potential outcome into the implementation process before the effort is deployed, avoidable costs can be circumvented, and the success of the implementation project can be increased.

The AI Cube research project will develop such guidelines for the process industry based on the experiences and maturity levels of the different sectors and building on existing enabling factors, strategies for overcoming barriers, potential impacts and business models for AI and Big Data solutions. The solutions and effort estimations around data quality presented in this paper provide important parts for these guidelines. Knowledge of the best possible solution to existing barriers such as poor data quality simplifies implementation and makes it easier to plan in terms of capacities, skills, and budgets. This plannability is further supported by the upstream success estimation through the limit value consideration of the error rate. Due to the fundamentally widespread data availability in the process industry through the already existing process control solutions, the further influencing parameters can be particularly well investigated experimentally in this industry in further research. By at least initially fixing the industry focus and thus a potential influencing factor on the correlations further influences can be identified and investigated. However, the objective of the test series presented here does not only influence the process industry from which the data for the test execution originate. Rather, the relationship between the pre-estimation of the forecast quality improvement, the data cleaning methods to be used and the limit values presented here has a generic significance for the preprocessing step of AI implementation processes.

Acknowledgements

This paper is part of the research in the AI-CUBE Project funded from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement N°958402 (*https://www.ai-cube.eu/*).

References

Alasadi, S. A. and Bhaya, W. S., 2017. Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), pp. 4102–4107.

Alsheibani, S., Cheung, Y. and Messom, C., 2019. Factors Inhibiting the Adoption of Artificial Intelligence at organizational-level: A Preliminary Investigation. In: M. Santana, and R. Montealegre. AMCIS 2019 Proceedings. Americas Conference on Information Systems 2019. Cancun, Mexico, 15/08/19.

Bertolini, M., Mezzogori, D., Neroni, M. and Zammori, F., 2021. Machine Learning for industrial applications: A comprehensive literature review. Expert Systems with Applications, 175, p. 114820. http://dx.doi.org/10.1016/j.eswa.2021.114820.

Bole, U., Popovič, A., Žabkar, J., Papa, G. and Jaklič, J., 2015. A case analysis of embryonic data mining success. International Journal of Information Management, 35(2), pp. 253–259. http://dx.doi.org/10.1016/j.ijinfomgt.2014.12.001.

Brownlee, J., 2020. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python: Machine Learning Mastery.

Cooper, H. M., 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, [e-journal] 1(1), pp. 104–126. http://dx.doi.org/10.1007/BF03177550.

Crone, S. F., 2010. Neuronale Netze zur Prognose und Disposition im Handel. Wiesbaden: Gabler.

Darapaneni, N., Muthuraj, S., Prabakar, K. and Sridhar, M., 2019. Demand and Revenue Forecasting through Machine Learning. In: 2019 International Conference on Communication and Signal Processing (ICCSP). Chennai, India, 04.04.2019 - 06.04.2019. Piscataway, NJ: IEEE, pp. 328–331.

Dasgupta, A. and Wendler, S., 2019. AI Adoption Strategies. CTGA Working Papers. Centre for Technology & Global Affairs at Oxford University.

Dash, S., Shakyawar, S. K., Sharma, M. and Kaushik, S., 2019. Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6(1). http://dx.doi.org/10.1186/s40537-019-0217-0.

Du Ni, Xiao, Z. and Lim, M. K., 2020. A systematic review of the research trends of machine learning in supply chain management. International Journal of Machine Learning and Cybernetics, 11(7), pp. 1463–1482. http://dx.doi.org/10.1007/s13042-019-01050-0.

Eager, J., Whittle, M., Smit, Jan, Cacciaguerra, Giorgio and Lale-Demoz, E., 2020. Opportunities of Artificial Intelligence. Policy Department for Economic, Scientific and Quality of Life Policies.

English, L., 1998. Data quality: Meeting customer needs. Pitney Bowes white paper.

Erben, R. F. and Romeike, F., 2003. Komplexität als Ursache steigender Risiken in Industrie und Handel. In: 2003. Erfolgsfaktor Risiko-Management: Springer, pp. 43–63.

Fornasiero, R., Nettleton, D. F., Kiebler, L., Martinez de Yuso, A. and Marco, C. E. de, 2021. AI and BD in Process Industry: A Literature Review with an Operational Perspective. In: A. Dolgui, A. Bernard, D. Lemoine, G. von Cieminski, and D. Romero, eds. 2021. Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. Cham: Springer International Publishing, pp. 576–585.

Garcia, S., Luengo, J. and Herrera, F., 2015. Data preprocessing in data mining: Springer.

Ge, Z., Song, Z., Ding, S. X. and Huang, B., 2017. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access, 5, pp. 20590–20616. http://dx.doi.org/10.1109/ACCESS.2017.2756872.

Hughes, D. L., Rana, N. P. and Dwivedi, Y. K., 2020. Elucidation of IS project success factors: an interpretive structural modelling approach. Annals of Operations Research, 285(1-2), pp. 35–66. http://dx.doi.org/10.1007/s10479-019-03146-w.

Jayawardene, V., Sadiq, S. and Indulska, M., 2015. An analysis of data quality dimensions.

Jöhnk, J., Weißert, M. and Wyrtki, K., 2021. Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors. Business & Information Systems Engineering, 63(1), pp. 5–20. http://dx.doi.org/10.1007/s12599-020-00676-7.

Khaydarov, V., Heinze, S., Graube, M., Knupfer, A., Knespel, M., Merkelbach, S. and Urbas, L., 2020. From stirring to mixing: artificial intelligence in the process industry. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). Vienna, Austria, 9/8/2020 - 9/11/2020. Piscataway, NJ: IEEE, pp. 967–974.

Koller, W., 2014. Prognose makroökonomischer Zeitreihen: Ein Vergleich linearer Modelle mit neuronalen Netzen. Teilw. zugl.: Wien, Wirtschaftsuniv., Diss., 2012. Frankfurt am Main: PL Acad. Research.

Kreutzer, Ralf T., and Marie Sirrenberg. Künstliche Intelligenz verstehen. Wiesbaden: Springer Fachmedien Wiesbaden, 2019.

Kuhn, A. and Hellingrath, B., 2002. Supply Chain Management. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kurgan, L. A. and Musilek, P., 2006. A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, 21(1), pp. 1–24.

Lazzeri, F., 2020. Machine Learning for Time Series Forecasting with Python®: Wiley.

Lu, H., Ma, X., Ma, M. and Zhu, S., 2021. Energy price prediction using data-driven models: A decade review. Computer Science Review, 39, pp. 3–42. http://dx.doi.org/10.1016/j.cosrev.2020.100356.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramirez Quintana, M. J. and Flach, P. A., 2020. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, p. 1–1. http://dx.doi.org/10.1109/TKDE.2019.2962680.

Moktadir, M. A., Ali, S. M., Paul, S. K. and Shukla, N., 2019. Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh. Computers & Industrial Engineering, 128, pp. 1063–1075. http://dx.doi.org/10.1016/j.cie.2018.04.013.

Montgomery, D. C., 2015. Introduction to Time Series Analysis and Forecasting. 2nd ed. Somerset: Wiley.

Mukhiya, S. K., 2020. Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data. Birmingham: Packt Publishing Limited.

Najdawi, A., 2020. Assessing AI Readiness Across Organizations: The Case of UAE. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Kharagpur, India, 01.07.2020 - 03.07.2020: IEEE, pp. 1–5.

Plotnikova, V., Dumas, M. and Milani, F., 2020. Adaptations of data mining methodologies: a systematic literature review. PeerJ. Computer science, 6, e267. http://dx.doi.org/10.7717/peerj-cs.267.

Randolph, J., 2009. A Guide to Writing the Dissertation Literature Review. Practical Assessment, Research, and Evaluation, 14(13). http://dx.doi.org/10.7275/b0az-8t74.

Rustum, R. and Adeloye, A. J., 2007. Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map. Journal of Environmental Engineering, 133(9), pp. 909–916.

Schreckenberg, F. and Moroff, N. U., 2021. Developing a maturity-based workflow for the implementation of ML-applications using the example of a demand forecast. Procedia Manufacturing, 54, pp. 31–38. http://dx.doi.org/10.1016/j.promfg.2021.07.006.

Siebertz, Karl; van Bebber, David; Hochkirchen, Thomas (2017): Statistische Versuchsplanung. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wassermann, O., 2013. Das intelligente Unternehmen: Mit der Wassermann Supply Chain Idee den globalen Wettbewerb gewinnen: Springer-Verlag.

Winter, M. and Peters, H., 2019. Artificial intelligence in EU process industry: A view from the SPIRE cPPP.

Wostmann, R., Schlunder, P., Temme, F., Klinkenberg, R., Kimberger, J., Spichtinger, A., Goldhacker, M. and Deuse, J., 2020. Conception of a Reference Architecture for Machine Learning in the Process Industry. In: X. Wu. 2020 IEEE International Conference on Big Data. Dec 10-Dec 13, 2020, virtual event: proceedings. Atlanta, GA, USA, 12/10/2020 - 12/13/2020. Piscataway, NJ: IEEE, pp. 1726–1735.