IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 June 2022, accepted 21 July 2022, date of publication 1 August 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195338

RESEARCH ARTICLE

On the Evaluation, Management and Improvement of Data Quality in Streaming Time Series

MERITXELL GÓMEZ-OMELLA[1,2], BASILIO SIERRA[2], AND SUSANA FERREIRO[1]

[1]Tekniker, Basque Research and Technology Alliance (BRTA), 20600 Eibar, Spain
[2]Faculty of Informatics, University of the Basque Country (UPV/EHU), 20018 Donostia-San Sebastian, Spain

Corresponding author: Meritxell Gómez-Omella (meritxell.gomez@tekniker.es)

This work was supported in part by the SPRI-775 Basque Government through the ELKARTEK Program through the Project 3KIA under Grant KK-2020/00049.

ABSTRACT The Internet of Things (IoT) technologies plays a key role in the Fourth Industrial Revolution (Industry 4.0). This implies the digitisation of the industry and its services to improve productivity. To obtain the necessary information throughout the different processes, useful data streams are obtained to provide Artificial Intelligence and Big Data algorithms. However, strategic decision-making based on these algorithms may not be successful if they have been developed based on inadequate low-quality data. This research work proposes a set of metrics to measure Data Quality (DQ) in streaming time series, and implements and validates a set of techniques and tools that allow monitoring and improving the quality of the information. These techniques allow the early detection of problems that arise in relation to the quality of the data collected; and, in addition, they provide some mechanisms to solve these problems. Later, as part of the work, a use case related to industrial field is presented, where these techniques and tools have been deployed into a data management, monitoring and data analysis platform. This integration provides additional functionality to the platform, a Decision Support System (DSS) named *DQ-REMAIN (Data Quality REport MAnagement and ImprovemeNt)*, for decision-making regarding the quality of data obtained from streaming time series.

INDEX TERMS Data quality, streaming time series, decision support system.

I. INTRODUCTION

The *Internet of Things (IoT)* is a new evolution of the Internet that includes many applications in different domains such as transportation and logistics, healthcare, smart environments and personal and social interactions as explained in [1]).

Large amounts of data have been captured with the recent digitisation of the industry, which represents a link between the physical and cyber world [2]. The analysis of the large amount of available data from historical data bases is an important step in obtaining information in different fields. This type of information can be used for anomaly detection, diagnosis, and/or forecasting as shown in [3] and [4] to obtain knowledge about the behaviour or conditions of a system. Despite the many types of analysis that can be carried out, the common goal of any of these studies based on the Data Information-Knowledge-Wisdom (DIKW) model [5] is wisdom. Data are the basis of the DIKW pyramid (Figure 1). Therefore, Data Quality (DQ) is a crucial requirement for any data analysis.

Poor DQ has a negative effect on these activities. Therefore, the accuracy of the techniques and algorithms can decrease significantly incorrect or poor-quality data have been used as inputs. Consequently, the conclusions drawn by understanding the results may be incorrect. Some studies have revealed that poor data quality is responsible for millions of annual losses [6]. Data gathered from the global-scale deployment of smart-things are the base for making intelligent decisions and providing services in IoT applications. Low quality has a high impact ranging from increased

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang.

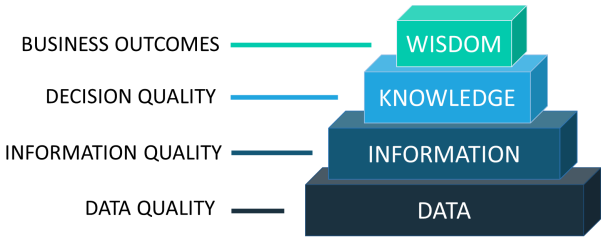


FIGURE 1. The basic structure of the Data-Information-Knowledge-Wisdom pyramid.

difficulty in setting strategies, derived from data analysis, to reduced customer satisfaction. [7]. The use of low-quality data leads to unsustainable decision and unsuccessful strategies and induces inefficient decision-making. The study of DQ is crucial for achieving user participation and acceptance of the IoT paradigm and services [2].

Organisations are aware of the importance of measuring the quality of data to identify errors and face losses. In addition, identifying these problems provides information on whether the data can be useful for their purposes. However, according to Gartner, nearly the 60 % of organisations do not measure the annual financial cost of poor-quality data. the annual spending on on-premises DQ tools remains high, with an average of $208000 and a median of $150000, and it prevents more pervasive adoption of tools [8]. Measuring and understanding DQ with the right tools is therefore necessary to improve outcomes and increase confidence in data-driven decisions [9].

Many definitions of the DQ can be found in the literature. Because of this variety, choosing suitable methods, that are advantageous for the DQ of a certain problem or in a particular context is a challenge [10]. These definitions usually refer to technical documents (standards) established by analysts or relevant organisations. In these cases, controlling hte DQ simply ensures that the data follow that standard. However, to assess the quality of the data (DQ), it is not only necessary to define the influencing aspects, but also to associate them with numerical scores. It is a scientific and statistical evaluation process that allows the calculation of a numerical data quality value for each problem or factor that influences DQ [11]. It can be considered a set of techniques or equations used to quantify and improve the quality of the data. The data set can be of low quality owing to a set of problems of various types and nature (e.g. loss of data, low accuracy, etc.). Therefore, it is necessary to have the required mechanisms to identify and quantify the problems that may exist in the data set with respect to its low quality, as well as having a set of corrective functions that allow handling each of those problems to improve the quality of the data. This is the focus of this work, the definition and implementation of some metrics to measure the quality of the data in streaming time series, and proposals to increase it.

The remainder of this paper is organised as follows. An extensive review of the methods proposed in the literature is presented in Section II. A mathematical representation to assess DQ in streaming time series is presented in Section III. Section IV describes the complete flow of the proposed methodology. The implementation of these metrics and functionalities resulted in an R package called `dqts`. Main functions available in `dqts` R package are explained in Section V. The methodology has been validated for different datasets (simulated, open-source and real) in Section VI. Finally, the design of the *DQ-REMAIN DSS (Data Quality REport MAnagement and Improvement Decision Support System)* for DQ analysis is presented in Section VII. The DSS has been described in functional blocks based on the use of `dqts` package and it provides an easy and intuitive solution for user interaction with DQ analysis for an industrial use case.

II. RELATED WORK

In this section, existing studies on definitions, approaches and implementations of DQ are analysed.

The need for the research community to define DQ was born at the end of the twentieth century when the quantity of data flows increased considerably with the digitisation of the industry. Initially, DQ focused on measuring dispersion using statistics when the data follow a known distribution. The confidence interval (CI) of the mean can be defined and precision can be evaluated according to the allowed variability [12].

Currently, DQ is not limited to traditional techniques based on the study of the standard deviation. There are other aspects to consider to achieve a high DQ. The main aim of most DQ publications in the late 20th century was to provide a formal definition of the term. Since then, the concept has been most often associated with the 'fitness for use' principle [13].

Therefore, it was wanted to delve into the aspects that define the concept of DQ, and several authors gave different sets of DQ dimensions such as *Accuracy*, *Timeliness*, *Interpretability* and *Accessability* [14], [15]. Data dimensions are attributes of the DQ that can, when measured correctly, indicate the overall quality level of the data. There are many possible dimensions of DQ depending on the context and nature of the data. These dimensions come from the issues specific to each field [16]–[18]. An overview of dimension definitions can be found in [19]. Furthermore, as the topic became more interesting, the quality of the data in specific sectors began to be defined and some authors focused their work on defining dimensions of quality for data received by sensors [20]. Essentially, the data collected by the sensors are streaming time series because the data are recorded together with the moment of time in which it was received.

Increased interest in the search for quality standards has led to the creation of DQ metrics. Metrics are formulas that allow quantification of different quality aspects within a dimension. Therefore, the most common approach to measure DQ is to define a set of metrics that provide numerical results to detect and correct data failures, and combine them to provide a numerical score of the overall DQ [21], [22]. The range of metrics available varies widely, because of the several definitions of the DQ concept depending on the context.

Nevertheless, one of the aspects in which the authors of research on DQ agree is the study of reliability [23]. It focuses on the use and trust of data. In other words, the study of metrics that can characterise the quality of the data received to provide an indicator of quality for future studies on the exploitation of these data.

The maximum knowledge of the domain and the problems presented is required for the correct definition of the metrics. It is common to find different definitions of metrics in relation to the problems presented by the data in different fields. Some authors place more emphasis on the diversity and volume of the data and the problem that the data change very quickly [23] whereas others consider missing values as the main quality problem. Furthermore, different definitions of DQ metrics can be found according to the origin of the data; for instance, [24] provided a specific approach when data come from sensors.

Then, the need arose to define complete methodologies for the study of DQ. This is divided into four activities: (1) state reconstruction, (2) DQ measurement or assessment, (3) data cleansing or improvement and (4) continuous data monitoring [25], [26]. The first phase of state reconstruction refers to the collection of contextual information, which is beyond the scope of this work. Although measurement and evaluation are concepts often treated at the same level, it is important to differentiate them, in terms of DQ. The term 'measure' describes the assignment of a numerical value or degree that allows quantification. Instead, assessment is the evaluation of the nature, ability, or quality of something and consists of analysing the results of measurements to draw a conclusion. Step (3) concerns the strategies to achieve the highest DQ. Finally, the techniques proposed for the periodic report and control of the temporal evolution are monitored.

The studies [27], [28] and [29] mathematically approximated some of the metrics proposed in the literature to quantify the quality of the data in a time series. However, the mathematical formulation of the mathematical formulation of the complete methodology to measure DQ in time series is still not available in the community of data analysts, who must manually adapt their own analysis methodology for DQ to the problem.

On the other hand, a work that isstill to be solved is building a personalised DQ management platform. Each data consumer has a unique vision of how ''good'' data should depend on their core business and needs [2]. The first step in the design of this platform is to define a general methodology for calculating the DQ score in a time series. Thus, a methodology that combines DQ measurement, assessment, improvement and monitoring is still unavailable in the literature.

After a review of the existing studies, we can conclude that there are an extensive number of definitions of metrics and dimensions depending on the subsequent use of the data and the context in which they are being analysed. In addition, no methods have been found to calculate metrics that are not based on a reference data set provided by the user. The need for a practical methodology for the treatment of DQ has been identified, which focuses on measuring and treating its different aspects instead of providing abstract definitions. Regarding the R packages available for data quality assessment, as far as we know, we conclude that there are no specific R packages for the quality of time series data and although some authors mention the dimensions of data quality, definitions of most of them are lack. Packages can be found in CRAN that are focused on a specific use of data in different fields such as *dataquieR* [30] to calculate the quality of epidemiological research data or *RawHummus* [31] in metabolomics. In addition, the *daqapo* R package offers a DQ assessment for process-oriented data that allows the detection of violations in frequency, order and range, detect outliers and missing values, incorrect names and unique vales but it requires a preprocessing step consisting of creating a certain type of data set [32]. The available R packages to assist data quality in general data sets calculate statistical indicators such as the mean and standard deviation, report unique values, and evaluate the number of missing values. These are *StatMeasures* [33], *dlookr* [34], *skimr* [35] and *xplorerr* [36]. In addition, *dlookr* include outliers indicators and *xplorerr* provides an interactive application in Shiny to show these results with open data sets or evaluate the data provided by the user. The only correction function found was provided by *StatMeasures*. It is a simple imputation function that replaces missing values with the value that the user enters as an argument of the function. The remaining packages are excluded from this comparison because the available documentation is not updated, or it is not possible to access the functions to use them in R.

This work was motivated by the need to monitor quality metrics in time series to ensure high DQ over time through the possibility of correcting the problems identified [37]. In addition, the available tools do not address the implementation of quality metrics. There is a gap in the relationship between the theory of data quality and tools available for its exploitation. Furthermore, for the DQ tools generated, approximately half are domain specific and of those that provide automatic support, there are no definitions of the functions. Finally, for the best of our knowledge, there are no tools that allow the correction of the metrics with the worst scores, nor the interactive design of DSS for the management of the DQ of streaming time series.

III. DATA QUALITY METRICS IN TIME SERIES

The analysis of the quality of the data received for subsequent statistical modelling and prediction studies is an important preliminary step in any data analysis. Although it is a matter that resides in the characteristics of each data set and in the intention and objective of the subsequent study, a generalisation of this concept is desired. Exact guidelines cannot be provided, however, the data are expected to conform to established standards. This section describes in detail the concept of DQ when data have an ordered structure, that is, a particular definition for DQ in time series is given by mathematical formulation and some proposed solutions.

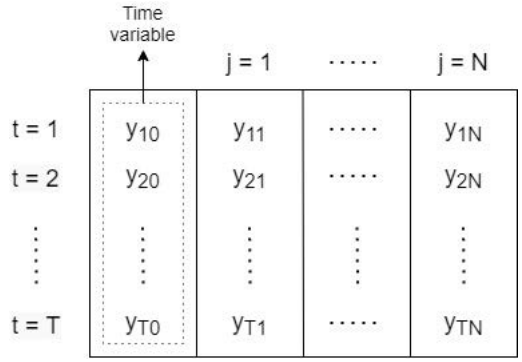


FIGURE 2. Schematic representation of the structure of a multivariate time series.

This type of data has special characteristics and should be treated in a special manner.

First, the necessary notation is introduced to understand the formulation given below. A time series is a collection of values obtained over time, often at regular time intervals. Therefore, consecutive observations can usually be recorded at equally spaced time intervals such as hourly, daily, weekly, monthly or yearly time intervals [38]. A time series is called univariate when it collects information about one characteristic and can be written as $Y = \{y_t\}$, where $t = 1, \ldots, T$ represents the elapsed time. A multivariate time series contains information from more than one characteristic, each of them having a univariate time series structure. In that case, the data set can be written as a table of dimension $T \times N$, where T is the number of observations, that is, the time elapsed, and N is the number of variables or univariate time series available. Often, a multivariate time series can be represented as a data table, as shown in Figure 2, where the first column is the time data and the remaining columns are the N variables to be analysed. The following methodology can be applied to multivariate time series data sets, considering each variable as a univariate time series. to simplify the notation, the time variable represented by $\{y_{10}, \ldots, y_{T0}\}$ is denoted by $\{t_1, \ldots, t_T\}$ from this point forward. The data set $Y = \{y_{tj}\}$, where y_{tj} is the value of variable j at time t, and $t \in \{1, \ldots, T\}$ and $j \in \{1, \ldots, N\}$, is denoted by Y in the remainder of this work, regardless of whether the time series is univariate ($N = 1$) or multivariate ($N > 1$).

The metrics are adapted to time series, taking the definitions provided by other authors summarised in Section II. Eleven metrics necessary to calculate the DQ in time series are presented below, classified into five dimensions. These metrics are functions that return values between 0 and 1 where 0 represents poor quality of the data and 1 represents the highest quality. Once the results of each quality metric have been calculated, a final DQ indicator can be provided by the arithmetic or weighted mean of the eleven metrics. The problem to be identified is described before each formula used to calculate the DQ scores. An optimal solution is proposed to deal with the poor quality identified by each of the metrics. This information is summarised in Table 1.

A. CONFORMITY

1) PROBLEM IDENTIFICATION

Conformity measures the amount of data stored in a standard format. Then, it determines the proportion of variables that are in the correct format and have the correct names. Conformity analysis is performed in two steps and both use a reference data set that can be provided or simulated before executing these metrics. The reference data set contains the same variables as the time series to be analysed, with the same names and in the same order. Each contains only one string element indicating the type of variable that is expected.

First, it is analysed that the names with which the variables have been labelled are correct. This is done by comparing the names with the names of the variables in the reference data set. It is then checked whether the format of the values contained in each of the variables is correct. The variables can be numerical, categorical or dates. This check is carried out by comparison with the reference data set.

Let $\{\mathrm{l}_1, \ldots, \mathrm{l}_N\}$ denote the names of the N variables available, and $\{\mathrm{l}_1^*, \ldots, \mathrm{l}_N^*\}$ denote the names of the variables in the reference data set. Similarly, $\{\mathrm{f}_1, \ldots, \mathrm{f}_N\}$ are the types of the N variables available and $\{\mathrm{f}_1^*, \ldots, \mathrm{f}_N^*\}$ are the correct formats of the same N variables in the same order in the reference data. The coincidence sets are defined as $\{\mathrm{c}_1^F, \ldots, \mathrm{c}_n^F\}$ and $\{\mathrm{c}_1^L, \ldots, \mathrm{c}_n^L\}$, where

$$\mathrm{c}_j^L = \begin{cases} 1, & \text{if } \mathrm{l}_j = \mathrm{l}_j^* \\ 0, & \text{if } \mathrm{l}_j \neq \mathrm{l}_j^* \end{cases} \quad \text{and} \quad \mathrm{c}_j^F = \begin{cases} 1, & \text{if } \mathrm{f}_j = \mathrm{f}_j^* \\ 0, & \text{if } \mathrm{f}_j \neq \mathrm{f}_j^* \end{cases} \tag{1}$$

So using that notation,

$$Names = \frac{\sum_{j=1}^{N} \mathrm{c}_j^L}{N} \quad \text{and} \quad Format = \frac{\sum_{j=1}^{N} \mathrm{c}_j^F}{N} \tag{2}$$

2) SOLUTION

There are two possible solutions when the *Conformity* value is low. The first is the transformation of the malformed variables into the desired format, taking those of the reference set as valid formats. The same is true for the names of variables. This process is not always possible because an external interpretation is sometimes necessary to convert an element from one type to another. Alternatively, the deletion of the conflicting variable from both the analysis and the reference data set is considered as a solution.

B. UNIQUENESS

1) PROBLEM IDENTIFICATION

The second dimension defined by some authors for the DQ is *Uniqueness*. The definition of this metric can vary according to the characteristics of the data set and objectives of the study. Each data set requires a different uniqueness in the captured variables. The uniqueness in the time variable is an important point regarding the DQ in time series because the repetition of the timestamps is not allowed. *Time Uniqueness* is the metric proposed to calculate the proportion of unique

values in the time variable and it is complementary to the duplicated timestamps in the data set. Let $\tau \in \mathbb{N}$ be the number of unique values of the time variable,

$$Time\ Uniqueness = \frac{\tau}{T}, \quad \tau \leq T \tag{3}$$

2) SOLUTION

Two solutions were proposed to address the low scores obtained in this metric. A straightforward method to increase the value of the *Time Uniqueness* score is to delete observations with duplicate values in the time variable. Another more complex method is the combination of repeated observations, for instance averaging. The second method is equivalent to the first one when the values of the other variables are repeated. An example of timestamp repeated three times is shown in Figure 3.

C. TIMELINESS

1) PROBLEM IDENTIFICATION

Time series data are typically saved at uniform time intervals. However, when data are received by sensors, there are usually small imbalances that should not be alarming, causing the time waits longer than what you want to allow. Therefore, the following metric calculates the proportion of observations received without a waiting time. *Timeliness* provides information on whether the data are available at the right time. Let Y_{t0} be the observations of the time variable where $t = 1, \ldots, T$. A set containing the difference between the time values was performed using

$$\mathcal{D} = \{\delta_1, \ldots, \delta_{T-1}\} \tag{4}$$

where $\delta_t = Y_{t+10} - Y_{t0}$ for $t = 1, \ldots, T-1$. Let δ_{max} be the maximum time difference allowed by two consecutive observations, *Timeliness* is computed as follows

$$Timeliness = \frac{\mathcal{D}^*}{T-1} \tag{5}$$

where $\mathcal{D}^* = \{\delta_t \in \mathcal{D} \mid \delta_t \leq \delta_{max}\}, \quad t = 1, \ldots, T-1.$

2) SOLUTION

Timeliness is the complementary value of the events of time that had been lost during the acquisition of time series data. The methods proposed to increase *Timeliness* value are based on the artificial generation of missing intermediate timestamps. Waiting times were examined and the necessary values were created for the time variable. Three methods are proposed to address the other variables. In the first method, no value is assigned to the rest of the variables, so the value of *Completeness* decreases after applying this method. The other two, take the average and median of the available data for each variable and use them to complete the observations.

D. COMPLETENESS

1) PROBLEM IDENTIFICATION

This concept refers to the degree of complete and present elements in the data set and it is one of the most important points of DQ. This value is the complementary of the degree of missing values, that are present in many studies and create uncertainty in research results. Given a multivariate time series in a data set Y, $M = \{m_{tj}\}$ is defined as the set of missing values where

$$m_{tj} = \begin{cases} 0, \text{ if } y_{tj} \text{ is missed} \\ 1, \text{ if } y_{tj} \text{ is known} \end{cases} \tag{6}$$

The first proposed metric has the same name as that of the dimension to be calculated. The score obtained by the following formula refers to the number of values present in the global view of the data set.

$$Completeness = \left(\frac{\sum_{t=1}^{T} \sum_{j=1}^{N} m_{tj}}{T \times N} \right) \tag{7}$$

Depending on the objective of each study, knowing whether the loss of data occurs at the same time for different variables may be of interest. Therefore, it is necessary to calculate the present values by observations. Similarly, in some cases it would be interesting to identify if missing data appear in the same variable over time. Thus, it is possible to identify faults in a specific characteristics of the set.

The subsets $T_{obs} \subseteq T$, $N_{var} \subseteq N$ are defined to calculate completeness by observations and variables, respectively.

$$T_{obs} = \left\{ t \in T \mid \sum_{j=1}^{N} m_{tj} = 0 \right\} \tag{8}$$

$$N_{var} = \left\{ j \in N \mid \sum_{t=1}^{T} m_{tj} = 0 \right\} \tag{9}$$

The *Completeness by observations* can be calculated as

$$Completeness_{obs} = \left(1 - \frac{|T_{obs}|}{T} \right) \tag{10}$$

and the *Completeness by variables* can be calculated as

$$Completeness_{var} = \left((1 - \frac{|N_{var}|}{N} \right) \tag{11}$$

2) SOLUTION

To increase the quality of the *Completeness* metric, different methods of handling missing values can be found in the literature. Methods of dealing with missing data can be classified into three groups: Case/Pairwise Deletion, Parameter Estimation and Imputation [39]. The first approach discards the observations that contain missing values. Usually if the amount of missing data is not relevant in the study, that is, there are very few values that are unknown, we choose to discard them. In this way, all statistical analysis is carried out without considering them and only the available data are proceeded. In the case of identifying any variable made up entirely of missing data, it will be removed from the data set. Subsequently, the dimensions of the data set are reduced and the statistical properties change [40]. The treatment of lost data is outside the scope of this work, and thus, possible

Timestamp	V1	V2	V3
03/08/2020 12:00	16	3	100
03/08/2020 13:00	12	1	85
03/08/2020 14:00	10	1	80
03/08/2020 14:00	20	1	90
03/08/2020 14:00	20	3	85
03/08/2020 15:00	17	2	90

Deletion

Timestamp	V1	V2	V3
03/08/2020 12:00	16	3	100
03/08/2020 13:00	12	1	85
03/08/2020 14:00	10	1	80
03/08/2020 15:00	17	2	90

Average combination

Timestamp	V1	V2	V3
03/08/2020 12:00	16	3	100
03/08/2020 13:00	12	1	85
03/08/2020 14:00	16.67	1.67	85
03/08/2020 15:00	17	2	90

FIGURE 3. Two solutions proposed when timestamps are repeated.

solutions are mentioned without detailing the methodologies. Imputation replaces missing values with estimations using various methods. Different imputation methods can be applied to treat data when the number of missing values or their effects on the study are high. The process of imputing the missing data consists of calculating an estimation based on the available data and replacing them in the set. There are simple techniques and others that are more sophisticated, but the effectiveness of each of the methods is not known because it will depend on the nature of the series.

The simplest method is imputation based on the average of all available data or the average of the maximum and minimum values. Similarly, the median can be used to estimate all missing values. Another possibility is to use interpolation between the last value and the next value available in each gap. Depending on the case, we can use interpolation of a different degree, although the most common is linear interpolation. On the other hand, Machine Learning techniques can be used to estimate missing values. In these cases, the values available prior to data loss are used to estimate the best model. The predictions made with that model will be the estimates of missing data in the series. Finally, a variant of the KNN algorithm can be used to find similar patterns in the time series. This method called KNPTS finds the most similar subseries in the available history and estimates the future values with combinations of the values that follow each of the selected subseries. [41].

E. ACCURACY

1) PROBLEM IDENTIFICATION

Four metrics are defined to measure the accuracy of the data and they should be interpreted according to their nature. These metrics provide information about the degree of reproducibility of measured values that may or may not be close to real world values. These metrics are calculated for each numerical variable in the data set. The final value is the average of the results obtained for each variable.

First, the *Range* metric quantifies the values within the lower and upper bands which can be provided or simulated before executing this metric. Therefore, measuring the *Range* value requires expert knowledge of the problem and ignorance can significantly vary the tolerance level and

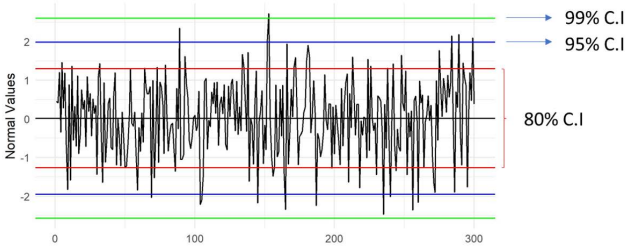


FIGURE 4. Example of simulated values following a Gaussian distribution and the three boundaries of the confidence intervals (CI) with confidence levels of 80%, 95% and 99%.

consequently the result of DQ. Let $y_{min} = (y^1_{min}, \ldots, y^N_{min})$ and $y_{max} = (y^1_{max}, \ldots, y^N_{max})$ be the sets of lower and upper values, respectively. The subset of Y contains values into these limits and is defined as follows:

$$Y^*_j = \{y_{tj} \in Y \mid y_{tj} \in [y^j_{min}, y^j_{max}]\}, \quad t = 1, \ldots, T \tag{12}$$

So using that notation,

$$\begin{aligned} Range &= \frac{1}{N}\sum_{j=1}^{M}(Range_j) \\ &= \frac{1}{N}\sum_{j=1}^{N}\left(\frac{Y^*_j}{T}\right) \end{aligned} \tag{13}$$

The remaining *Accuracy* metrics are *Consistency*, *Typicality* and *Moderation*. These three assume a Gaussian distribution in variables and 80%, 95% and 99% confidence intervals (CI) are built using the data available for each numerical variable. *Consistency* is the proportion of values in an interval with a confidence level of 80% and its corresponding z-score is 1.28. The set of consistent values for variable j is expressed as follows

$$Y^C_j = \{y_{tj} \in Y \mid y_{tj} \in (\bar{y}_j - 1.28s_j, \bar{y}_j + 1.28s_j)\} \tag{14}$$

where $\bar{y}_j$ is the average, and s_j is the standard deviation of the variable j calculated in a random selection of 30% of the points in the first part of the series.

So using that notation,

$$Consistency = \frac{1}{N}\sum_{j=1}^{M}(Consistency_j) = \frac{1}{N}\sum_{j=1}^{N}\left(\alpha_C - \frac{Y_j^C}{T}\right), \quad t = 1, \ldots, T \quad (15)$$

where, $\alpha_C = 0.8$ is the level of confidence in that interval. Similarly, *Typicality* is calculated using a $\alpha_T = 0.95$ confidence level and a z-score of 1.96 and *Moderation* is calculated using a confidence level of $\alpha_M = 0.99$ and a z-score of 2.58. The corresponding typical values set and moderated values set are named Y_j^T and Y_j^M, respectively.

2) SOLUTION

The proposed methods to obtain high *Accuracy* are different in the *Range* metric than in the rest of the normality metrics. Although the methods to solve Range can be applied in *Consistency*, *Typicality* and *Moderation*, it must be considered that certain manipulations in the data could cause the distribution to change and they could no longer be considered Gaussian. The first solution is the simplest and consists of removing the values that are outside the boundaries. The remaining are different imputation techniques, that is, the estimation of the values identified by other data with more appropriate values. All the methods explained above to increase the value of the *Completeness* metric can also be applied with data out of range or outside the normality bands. In addition, the band values can be used to calculate other types of estimates. On the one hand, the values can be imputed by the mean value between the minimum and maximum allowed, that is, the mean of the upper and lower limits. On the other hand, it is possible to impute the values that exceed the upper limit using the maximum allowed, and the values that are below the lower limit using the minimum value allowed. All these methods, as we have commented, affect the distribution of the data. Therefore, an additional method is proposed to try to increase the *Consistency*, *Typicality* and *Moderation* values. A random period of initial data is used to calculate the theoretical mean (μ) and standard deviation (σ) of the data distribution. Once these statistics are calculated, random data that follow a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ are generated. These values are used as estimates of the values outside of normality in each metric.

IV. IMPLEMENTATION OF DATA QUALITY METRICS

This section describes the complete flow from data acquisition to obtain the set with the desired quality. The process is illustrated in Figure 5.

The first step after accessing the data set is the identification of the Gaussian variables to assign weights to the *Consistency*, *Typicality* and *Moderation* metrics (A). Next, the existence of the Reference Data Set, Range Data Set, Maximum Time Difference and Unit of Time is checked (B). If they are unavailable, they are simulated. The third step is the computation of DQ metrics using the `DQ` function in the complete set or by moving windows (C). If all the metrics reach the highest quality, it is concluded that the analysis set is correct. If the quality is not 1, a list of the metrics is returned in the order in which they should be treated. Next, using the `deepDQ` function, the problems with the first metric in the list are analysed in depth (D). Finally, the decision is made to rectify the data set. In that case, using the `handleDQ` function the metric is corrected until it reaches a value of 1 (D) and the metrics are recomputed in step C. If data are not modified in relation to this metric, they are removed from the list and if there were more items left in the list, it returns to step D using the next metric in the list. After the complete inspection of the list is done, the final modified data set is available.

A. NORMALITY CHECK

The first step in the process of calculating DQ is the search for variables that are expected to be normal in the data set. It is decided which of the available numerical variables follows a Gaussian distribution and which do not by means of the Shapiro-Wilks test [42]. This test considers the normality of the data as the null hypothesis, and a p-value lower than 0.05 is considered significant, leading to the rejection of the null hypothesis and the assumption of normality in the distribution.

A 30% random sample is taken from the beginning of the series (if possible in the first third of the time series) from each numerical variable and the Shapiro-Wilks test is evaluated for every sample. If the test allows us to decide that none of the available variables follow a Gaussian distribution, then the metrics *Consistency*, *Typicality* and *Moderation* are not calculated and their weights are set to zero. The remaining weights corresponding to the other metrics are recomputed as identically distributed. Therefore, 1/8 weights are assigned to each of the remaining eight metrics. Otherwise, some of the variables in the data set are considered to follow a Gaussian distribution and the normality metrics will have a weight of 1/11, the same as the rest of the variables.

B. CALCULATION OF MISSING PARAMETERS

It is important that the user knows the data to be analysed well and provides as much information as possible to obtain accurate results. Thus, it is ensured that the values of the metrics are in accordance with the quality standards expected from the data. Providing wrong values of initial parameters may lead to errors in the conclusions of the data quality analysis, for instance, assuming an incorrect frequency of the data or time units. When initial information is lacking, the system is prepared to simulate these values in the most realistic way to avoid erroneous conclusions.

1) REFERENCE DATA SET

The reference data set is made up of as many variables as the data set to be analysed has. Each of these variables must be

TABLE 1. Summary of problems and solutions for each of the proposed DQ metrics.

Dimension	Metric	Problem identification	Solutions
Conformity	Names	Wrong variable names	Copy names or formats from the reference data set
	Format	Different data formats	Delete variables with incorrect names or format
Uniqueness	Time Uniqueness	Repeated timestamps	Delete repeated observations Combine the repeated observations
Timeliness	Timeliness	Excessive waiting times between observations	Add missing observations
Accuracy	Range	Values out of range	Delete observations containing values out of range Imputation of values out of range
	Consistency	Values out of the 80% confidence interval	
	Typicality	Values out of the 95% confidence interval	
	Moderation	Values out of the 99% confidence interval	
Completeness	Global Completeness	There are missing values	Delete observations containing missing values Imputation of missing values
	Compleness by Observations	Some observations are lost	
	Completeness by Variables	Some variables are lost	

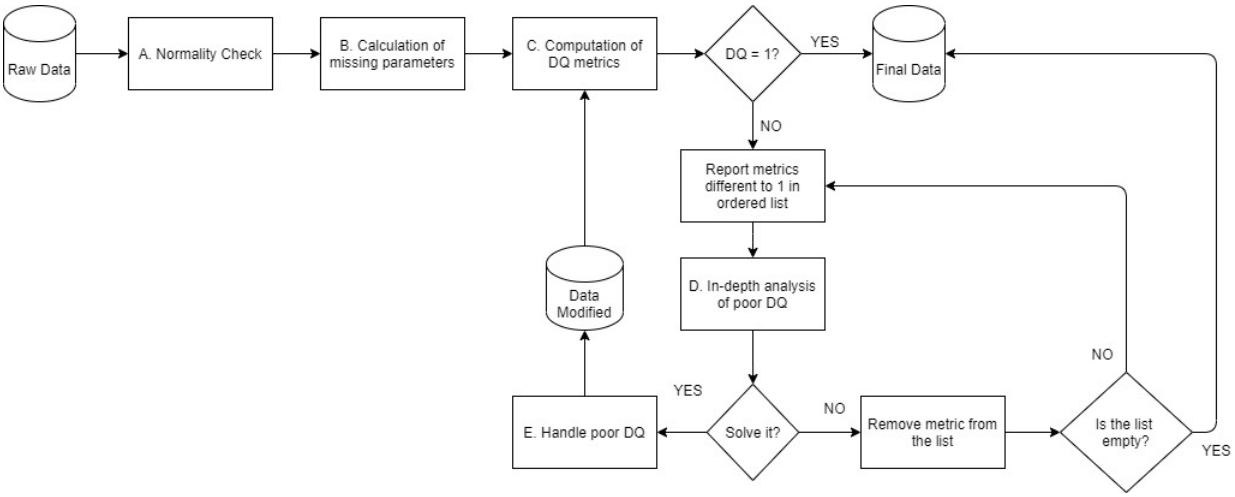


FIGURE 5. Flow for the detection, inspection and resolution of poor DQ problems.

correctly named. The content of each variable is the type of value expected from each variable.

If this reference set is not provided by the user when calculating the *Formats* and *Names* metrics, a set will be calculated by extracting the information from a random sample of 30% of the data located in the first third of the series.

2) RANGES DATA SET

The data set of ranges has the same number of variables as the data set to be analysed. The variables have the same name as the original data set and each contains two values. The first value of each variable contains the minimum value allowed in each case and the second is the maximum.

In the event that this information is not available at the time of computing the *Range* metric, a set of ranges will be calculated by extracting the minimum and maximum values of each variable in a random sample of 30 % of the available data located in the first third of the series.

3) MAXIMUM TIME DIFFERENCE AND UNITS

Observations are expected to be received at regular time intervals in time series data sets. For this reason, there is a maximum time difference allowed between the observations with their corresponding unit of time.

The value with the highest mode is extracted from a random sample of 30% of the available data located in the first third of the series if this information is not available at the time of calculating the *Timeliness* metric. In the same way, the unit of that time difference is extracted from that most common value in that random sample.

C. COMPUTATION OF DQ METRICS

This step can be performed in two ways depending on the interests of the user, as described in the following subsections. On the one hand, the value of the metrics can be obtained from the complete data set. On the other hand, there are three different ways to compute quality metrics by moving windows to inspect the evolution of DQ over time. This method of computing metrics is recommended when a large amount of data is available, because the information provided will be more accurate.

1) OVERALL DQ

The main objective in generalising the concept of DQ is to provide a general value that allows the comparison of the quality of different data sets. A value between 0 and 1 is obtained for each quality metric, in addition to a general value that is calculated from a combination of the remaining values.

2) MOVING WINDOWS

Calculation of the DQ metrics using moving windows provides information on the evolution of the metric values during the course of the temporal data. It is convenient that the windows are of the same size, that is, the scores are calculated with the same amount of data or that the windows start at the same point. This is forced so that the achieved values are comparable and bias is avoided. Three methods for calculating the quality of the data by moving windows are explained. On the one hand, it is decided whether the size of the window is fixed or changing. On the other hand, if the values of the series can belong to more than one window, that is, if the windows can overlap.

- Figure 6 shows the case in which the size of the windows is constant and does not overlap. In this option, the values to be analysed are completely different in each iteration of the DQ metrics.
- In the second case, the partition by windows of a constant size with overlap is shown. In this way, as it can be seen in Figure 7, in each iteration the first data of the window are deleted and new data are added at the end; however, some overlap is allowed in the windows. Therefore, some final values of one iteration are used to be the first values in the next computation of the DQ metrics.
- Finally, Figure 8 shows the case of a window of varying sizes and overlapping. In this case, the beginning of the interval remains fixed and more data are added to the window in each iteration.

Once the DQ value for each metric is known, they are ordered in a specific manner. This order places the perfect metrics at the bottom of a list and orders the rest so that their arrangement affects the rest as little as possible. In this way, if fixes are required for all metrics, the metrics relative to *Conformity* (*Names* and *Format*) will be dealt with first. These metrics could add or remove all variables and for that reason should be fixed first. The time metrics (*TimeUniqueness* and *Timeliness*) are then inspected. The reason for studying these metrics at this point is that the first could remove all observations from the data set and the second could add missing values to the time series. Next, the *Accuracy* metrics will be discussed, starting with *Range* and continuing with the three normality metrics (*Consistency*, *Typicality* and *Moderation*). Finally, the *Completeness* metrics, impute both the initial missing values and those that could have been added in previous steps.

Each time the data set is modified to raise quality in one of the dimensions, the quality metrics are recalculated. Thus, the list of metrics are modified. Once a list of metrics is obtained with values that are substantially good for the user, the new set can be saved and the modified data downloaded for possible analysis.

D. IN DEPTH ANALYSIS OF POOR DQ

The information obtained from an in-depth analysis of each metric is as follows

- *Names:* Name in the reference set and analysis set of variables with incorrect names and their positions in the original set.
- *Formats:* Formats in the reference set and analysis set of the malformed variables and their positions in the original set.
- *Time Uniqueness:* Repeated dates and frequencies they appear in the analysis data set.
- *Timeliness:* Instant in which a temporary wait longer than allowed begins and ends, in addition to the waiting time and the number of values that were expected to have been received in that period.
- *Range:* Name of the variables with out-of-range values accompanied by information about these values. This information can be in two forms: the value of the *Range* metric for each variable or the time positions in which these out-of-range values occur.
- *Consistency, Typicality and Moderation:* Name of the normal variables with more values outside the confidence intervals (CI) than allowed in each case. There are two options to show the information: the value of the corresponding metric in each of the variables or the time of all values outside the CI. Note that some of the values that will be shown outside the CI do not necessarily have to be incorrect because we assume that in every normal distribution, there will be a number of values outside the CI corresponding to the confidence level with which the interval has been built.
- *Completeness:* Name of variables with missing values with some useful information. Two options: The value of the *Completeness* metric by variables or the time in which data were lost for each of the variables.

E. HANDLE POOR DQ

One of the possible actions for the treatment of low data quality is the use of corrective functions for each of the metrics that do not reach the maximum quality score. These actions are mentioned in Section III. The options available for each DQ metric are as follows.

- *Names:* If any of the variables do not have the same tag or name as the variables given in the reference set, there are two options. First, defective variable are eliminated from the study set. The second option is the manual change of the name of the variable in which it fails, in the case that the expected name is known.
- *Formats:* There are several functions in the base R package that allow switching from one format to another. However, it must be borne in mind that this is not always possible. Therefore, apart from the typical format change options such as changing from character to numeric, it is possible to eliminate the variable from the study.
- *Time Uniqueness:* The first option available is to eliminate observations that contain repeated dates and leave only the first one. In the case of uncertainty about which of them provides more information, it is possible to

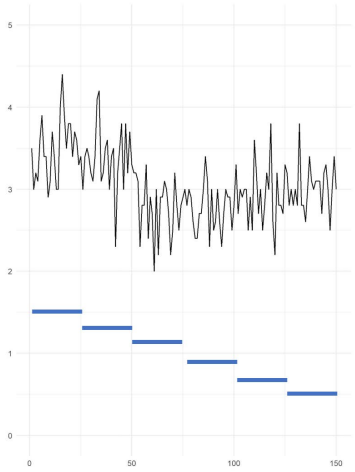


FIGURE 6. Constant windows without overlapping.

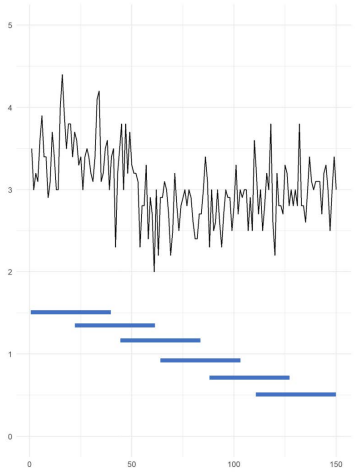


FIGURE 7. Constant overlapping windows.

combine the values of the rest of the variables by arithmetic mean, median, minimum or maximum.

- *Timeliness:* The low-value solution functions create the missing dates in the time variable. Once these values are generated, there are several options for completing the corresponding observations in the remaining variables. The four general options are using the minimum, the maximum, the arithmetic mean and the median. If none of these values adapts to the behaviour of the series, there is the option of not filling those gaps and leaving a null value in them. In that case, the *Completeness* metric changes its value, because artificial missing values are being introduced.
- *Range:* There are four simple methods that replace the values out of range by the average, the median, the minimum or the maximum value of that variable. Another possibility is to substitute the values by the mean of the last value in the range received and the next value. In addition, the KNPTS algorithm can be used to estimate out of range values. Finally, there are two methods related to the values introduced in the Ranges Data Set. On the one hand, a function that replaces the values out of range with the average of the minimum and maximum allowed values for each variable. On the other hand, the values that exceed the values allowed are replaced by the maximum and the values that fell short are replaced by the minimum value allowed. Those options are implemented using the variables.
- *Consistency, Typicality and Moderation:* The options available are the same as in the *Range* case but it is not recommended to solve problems related to normality distribution because it can undermine the properties of the time series with a Gaussian distribution. Alternatively, the mean (μ) and standard deviation (σ) of the data in the first part of the time series are calculated and used to simulate random data that follow a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. The last option is to do nothing with variables with values out of normality but to consider that the problem with their distribution exists.
- *Completeness:* There are four simple methods that replace the missing values with the average, the median, the minimum or the maximum of the available values in each variable. Another option is to substitute the missing values with the mean of the last value in the range received and the next value. Finally, the KNPTS algorithm can be used to estimate the missing values.

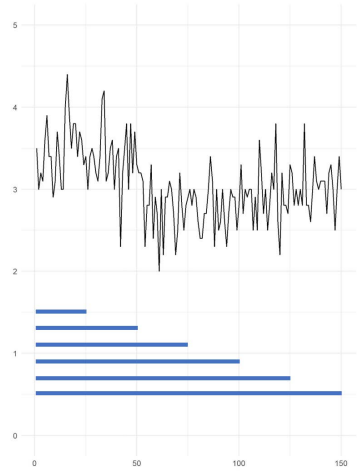


FIGURE 8. Non-constant windows with overlapping.

V. THE DQTS R PACKAGE

The work explained in the previous sections was implemented using the statistical software R. In R, the fundamental unit of shareable code is the package. A package bundles code, data, documentation, and tests together, and is easy to share with others. The `dqts` R package is available in GitHub in the following link for use by the entire R community.[1]

[1] https://github.com/MeritxellGomez/dqts-R-package

The R library is made up of four main functions explained below.

- The `DQ` function performs the three steps described in Section IV. First, it checks the normality of the variables and the availability of the necessary parameters in the input arguments. If unavailable, they are estimated. Finally, the values of the DQ metrics are calculated and this function allows two ways to do so, the overall one and by windows in three different ways.
- The `deepDQ` function takes the data, the name of the metric to inspect, and the parameters required for that metric as inputs. This function returns precise information regarding the data failures in the selected metric.
- The `handleDQ` function estimates solutions to faults found in the data for the metric introduced as an argument. It returns the data set with the necessary changes for that metric to achieve the highest quality score.
- The `plotDQ` function allows visualisation of the quality of the data. The output of the `DQ` function is introduced as an argument. In the case that quality has been calculated in the complete data set, the `plotDQ` function shows a bar graph where each bar indicates the numerical value of each of the metrics with magnitudes between 0 and 1. On the other hand, if the quality of the data has been computed by windows, a scatterplot is displayed with as many lines as metrics have been calculated and the time evolution of the metrics is shown.

VI. VALIDATION IN CASE STUDIES

This section presents an evaluation of the metrics developed by applying them to three data sets. To demonstrate the effectiveness of the quality metrics for different scenarios, it was decided to test the proposed method on a simulated data set created by the authors, on an open data set and on a real data set.

A. DETECTION AND SOLUTION IN SIMULATED DATA

1) DATA DESCRIPTION

In the first experiment, a multivariate time series was created with minute data and information on five random variables collected over one day. Three of these variables (G1, G2, and G3) follow a Gaussian distribution, another follows a chi-square distribution (V1) and the last was a binomial variable (V2) that takes values 0 or 1. In addition, the temporal variable (timestamp) collects minute values from "2021-01-01 00:00:00 UTC" to "2021-01-01 23:59:00 UTC". In summary, six variables were available for the analysis of this data set.

Problems related to each of the DQ dimensions defined in table 1 were simulated. The table 2 contains detailed information regarding the simulations to be detected and fixed throughout this section.

2) DQ ANALYSIS

The multivariate time series after simulating quality errors was displayed in Figure 9. Next, the process of Figure 5 starts to examine and correct the quality of the data.

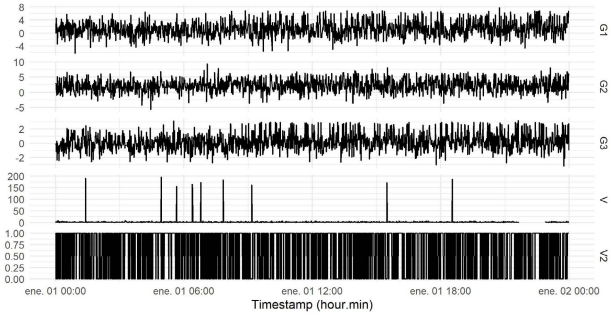


FIGURE 9. Visualisation of data distribution of all 5 variables of the simulated data set.

First, normality was evaluated and it was established that the variables G1, G2, and G3 follow a normal distribution, with p-values of 0.87, 0.15 and 0.59 in the Shapiro-Wilk test, respectively. This statement was possible because the three p-values were greater than 0.05, the null hypothesis of normality in their distributions can be considered as valid. For that reason, the weights for the *Consistency*, *Typicality*, and *Moderation* metrics were set the same as the other metrics. In this case, the combination of the metrics was balanced; therefore, the weights were all 1/11. Regarding the input parameters, the type of each variable was known, so the reference set was given as input in the DQ function. On the other hand, some fictitious ranges were established that were also entered as inputs in the function. Finally, we know the data acquisition frequency; therefore, the maximum value of time difference was set at 1 min.

In the first evaluation of the quality metrics in the complete set, the results of the metrics were obtained, as listed in Table 3. The order in which the results appear indicates the order in which problems should be addressed.

It can be seen that the fault related to *Names* was in the variable that was in position number 5, using an in-depth inspection of the problem. This variable was called V1 in the reference set and is now called V. As for *Formats*, V2 should be an integer, according to the reference set, but it was received as a character. These two problems were solved by renaming the variable V by V1 and changing the character values of V2 to integers.

Following the established order, the next metric that falls short of the highest quality score was *TimeUniqueness*. One timestamp was found more than once as shown in Table 4. This problem was solved by eliminating the observations that were repeated.

The *Timeliness* was discussed below. An in-depth inspection of the problems related to this metric indicates that temporary waits are due to a stop between 2021-01-01 03:15:00 and 2021-01-01 03:20:00. Four observations were excluded. The missing observations were generated using the `handleDQ` function, and the remaining values were imputed by the median of each variable.

None of the three normal variables (G1, G2, or G3) achieved the highest score for *Consistency*, *Typicality*

TABLE 2. Problems forced to appear in simulated data.

Metric	Variable	Description
Names	V1	renamed to "V"
Format	V2	Character type instead of numeric
Time Uniqueness	timestamp	"2021-01-01 00:09:00 UTC" is repeated 4 times
Timeliness	timestamp	time gap between "2021-01-01 03:15:00 UTC" and "2021-01-01 03:20:00 UTC"
Range	V1	10 values out of range in random position
Consistency	G1	10% of the correct values move outside the 80% of CI
Typicality	G2	10% of the correct values move outside the 95% of CI
Moderation	G3	10% of the correct values move outside the 99% of CI
Completeness	V1	5 % of NA in random position

TABLE 3. DQ metrics in overall evaluation.

Metric	Value	Interpretation
Names	0.8333	83.33% of variables with correct names
Format	0.8333	83.33% of variables with correct formats
Time Uniqueness	0.9979	99.79% of unique values in timestamp
Timeliness	0.9972	99.72% of correct waiting times between consecutive timestamps
Range	0.9988	99.88% of values within ranges
Consistency	0.9492	94.92% of values following Gaussian distribution according to CI with 80% of confidence level
Typicality	0.9556	95.56% of values following Gaussian distribution according to CI with 95% of confidence level
Moderation	0.9791	97.91% of values following Gaussian distribution according to CI with 99% of confidence level
Completeness	0.9915	99.15% of data received have a value
Completeness Observations	1	100% of observations have at least one value
Completeness Variables	1	100% of variables have at least one value

TABLE 4. Deep inspection of time uniqueness quality.

Repeated Date	Frequency
2021-01-01 00:09:00	4

TABLE 5. Deep inspection of timeliness quality.

Loss Start Date	Loss End Date	Waiting times	Missing values
2021-01-01 03:15:00	2021-01-01 03:20:00	5 mins	4

and *Moderation*. Note that these three metrics are related. If *Moderation* lowers its quality value, *Typicality* and *Consistency* also worsen. Suppose that the analyst needs those variables in the future. Therefore, they can not be deleted, and the best option for dealing with low quality is to simulate the values using the mean and standard deviation. The means of the first third of G1, G2 and G3 were 0.91, 1.91 and 0 and their standard deviations were 2.07, 1.97 and 1.04, respectively. The rest of the time series was imputed by random values following a $\mathcal{N}(0.91, 2.07^2)$, $\mathcal{N}(1.91, 1.97^2)$ and $\mathcal{N}(0, 1.04^2)$ distributions.

Finally, *Completeness* was inspected in depth, and it can be seen that the metric was 1 for all the variables except for V1 which was 0.9493. The KNPTS method was used to impute the missing values in V1.

Once the DQ analysis has been completed and the conflicts of the different variables resolved, the total quality of the final set was 1.

B. DETECTION AND SOLUTION IN OPEN DATA

1) DATA DESCRIPTION

The airline passenger data set contains monthly data of airline passengers in thousands from January 1949 to December 1960 [43]. AirPassenger is one of the most well-known open data sets available in R. It is a univariate time series with 144 monthly values that does not follow a Gaussian distribution.

This section takes advantage of the availability of data to compare the precision of time series predictive models when they were trained with quality data and when they were trained with data that lack quality.

An ARIMA model was trained using the univariate time series of the original data for the first 11 years to predict the data for the previous year. The ARIMA model that best fits the data was an ARIMA(1,1,0)(0,1,0)[12] model. This model was chosen because of the clear trend and seasonality presented in the time series. The forecast for the next year of data was shown in Figure 10. The graph also shows the confidence bands of the intervals with a confidence levels of 80% and 95%. The RMSE was used to validate the precision of this forecast. We obtained an RMSE of 23.93. The results were used as benchmarks.

Next, *Completeness* and *Range* problems were forced into the training set for the first 11 years of data. In the first case, missing values were simulated in 12 values corresponding to 1955. In addition, an extremely high value (1500) was added in May 1957.

2) DQ ANALYSIS

The DQ flow in Figure 5 was executed in the set in which we simulated the problems. First, the variable that measures passengers does not follow a Gaussian distribution; therefore, the *Consistency*, *Typicality* and *Moderation* metrics will not be computed. In this case, the minimum value of the range that is allowed for the number of passengers is zero, because a negative number is meaningless. The maximum value was simulated to be 1000. The reference data set was randomly calculated from the first third part of the time series.

TABLE 6. DQ metrics and RMSE achieved in forecasting using original and simulated data.

Metric	Original Data	Simulated Data
Names	1	1
Formats	1	1
Time Uniqueness	1	1
Timeliness	1	1
Range	1	0.99
Completeness	1	0.95
DataQuality	1	0.99
RMSE	23.93	90.68

In the first computation of the DQ of the original data set, all metrics achieved the maximum score and therefore the total value of the quality of this set was 1. Therefore, the AirPassengers set does not originally present any problems with DQ. The quality metrics in the simulated set assign a value of 0.95 to *Completeness* and a value of *0.99* to Range. These two values indicate that there were two problems with DQ. The total value of the quality of this time series was 0.99, with null weights assigned to the three metrics related to normality.

The effect of poor data quality can be seen when the search process for the best time series model was repeated with data from the same time period. The seasonality was no longer captured and the best model was ARIMA (0,1,1). Those issues were reflected in the precision of future estimates. The ARIMA model cannot accurately predict the next values as shown in Figure 11. Furthermore, the confidence interval bands widened, highlighting the uncertainty associated with new predicted values. The RMSE obtained in this forecast was 90.68, which represents an increase of 378% compared to the forecast made with the original data that had full quality. This result suggests that if the quality of the data is not controlled, erroneous results can be obtained with algorithms that worked correctly.

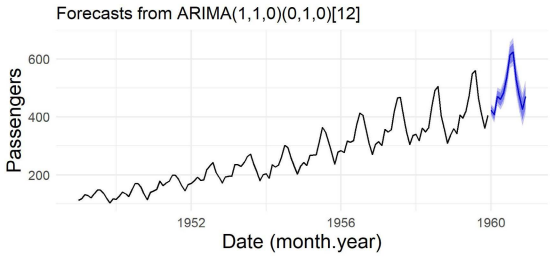


FIGURE 10. One year ahead forecasting training ARIMA model with original data.

Table 6 reflects the quality problems that appear in the new set and the impact that a decrease in DQ has on an increase in the prediction error.

The `handleDQ` function was used to solve problems with *Range* and *Completeness*. On the one hand, the correction of the value that was out of range was made using the mean method. On the other hand, the KNPTS method was used for

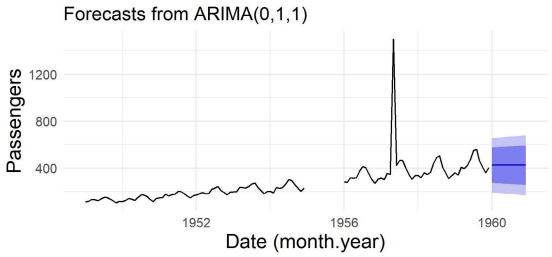


FIGURE 11. One year ahead forecasting training ARIMA model with simulated data with 0.95 in Completeness and 0.99 in Range.

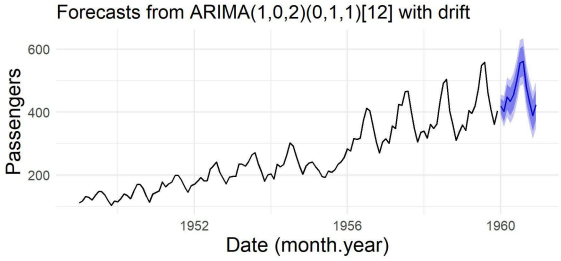


FIGURE 12. One day ahead forecasting using fixed data with highest quality.

the correction of Completeness. After these two corrections, the total data quality was achieved. The model that best fits these new training data was ARIMA (1,0,2)(0,1,1)[12]. The forecast data for the following year were shown in Figure 12. The improvement was remarkable. The confidence interval bands were adjusted to the predicted data and the RMSE error obtained for this forecast was 30.70. This value represents a 66% error reduction compared to the forecast made with the erroneous data. The difference between the RMSE obtained with the corrected data and original data was 6.77. In other words, the RMSE increases only 28%.

This simulation study shows, on the one hand, the importance of having quality data at the time of starting statistical analyses that give rise to predictions of future values. On the other hand, the advantages that the package of quality metrics presents to detect and solve quality problems quickly and effectively.

C. DETECTION AND SOLUTION IN REAL DATA

1) DATA DESCRIPTION

Monitoring of electrical power systems in industry has been on the rise in recent years. Forecasting electricity demand using Time Series techniques and Machine Learning is one of the tasks on the rise in the study of the production and consumption of electricity. The study of electricity consumption data began with the task of capturing data from high-frequency meters. These data were expected to be cyclical because of the activation behaviour of certain devices. However, they were not expected to follow a normal

TABLE 7. Metrics that not achieve the highest value of quality in electricity consumption dataset.

Metric	Lowest Value	Highest Value
TimeUniqueness	0.9583	1
Timeliness	0.9600	1
Range	0.6667	1

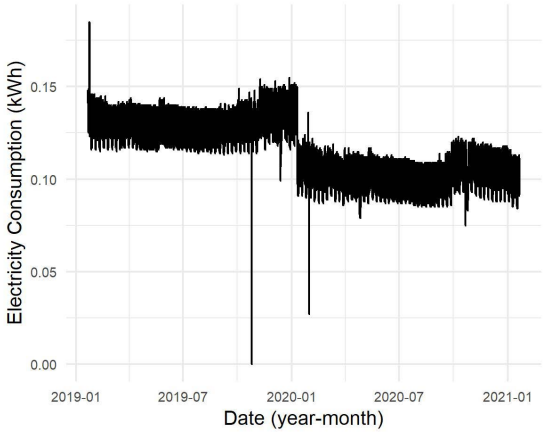


FIGURE 13. Distribution of the electricity consumption

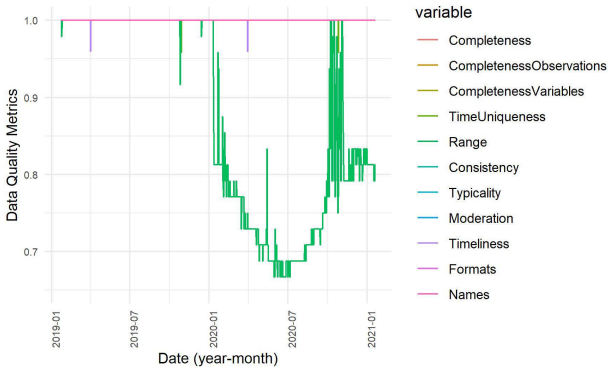


FIGURE 14. Evolution of the DQ metrics.

TABLE 8. Deep inspection of time uniqueness quality.

Repeated Date	Frequency
2019-07-27 02:00:00	2
2020-10-25 02:00:00	2

TABLE 9. Deep inspection of timeliness quality.

Loss Start Date	Loss End Date	Waiting times	Missing values
2019-03-31 01:00:00	2019-03-31 03:00:00	2 hours	1
2020-03-29 01:00:00	2020-03-29 03:00:00	2 hours	1

distribution; therefore, the normality metrics in this case were weighted with a null weight.

In this case, data were taken from a meter of a sentry box in which an electric pump is installed that supplies water to nearby farms. The small construction site is located in the facilities of the Tekniker technology centre, in Eibar, Spain. The data set corresponds to a univariate time series capturing hourly electricity consumption data measured in kWh. A total of two years of data were captured from 21 January, 2019 at 00:00 a.m. to 21 January, 2021 at 11:00 p.m. Figure 13 shows the distribution of the data over time.

2) DQ ANALYSIS

The flow in Figure 5 was executed to compute the DQ in the time series of the electricity consumption data. Figure 13 shows the distribution of the data. In this case, the allowed ranges for the electricity consumption variable were not known, but the function that automatically generates them was used, taking a random sample of 30% of the data located in the first third of the series. Figure 14 shows the evolution of the quality metrics computed using constant 24-hour windows without overlap.

The value of the DQ in each moving window ranged from 0.958 to 1. The reason why the quality was not 1 in the entire set was that some metrics have detected quality problems at different points in the series. The system returns a list like that in Table 7 in which it can be seen the metrics that fail in the order in which they should be treated.

The `deepDQ` function was used in this step to check the metrics that did not reach the highest quality score. Table 8 lists the output of the *Time Uniqueness* inspection. Low values of that metric are due to repetitions on two different dates. The first on 2019-07-27 02:00:00 and the second on 2020-10-25 02:00:00. Both with a frequency of 2. The Deletion method was used to solve this problem.

Next, the *Timeliness* metric was inspected in depth and the information collected in Table 9 was obtained. Two waiting times of 2 hours were identified at two different points in the series. The first between the dates 2019-03-31 01:00:00 and 2019-03-31 03:00:00 and the second between the dates 2020-03-29 01:00:00 and 2020-03-29 03: 00:00. Both involve a loss of one observation. Those two observations were generated. Note that this arrangement introduces two missing values in the series and causes the *Completeness* to decrease from 1 to 0.999.

Finally, *Range* was analysed and it was seen that in the first half of the series there were values out of range on the dates that were collected in Table 10. In the second half of the time series, none of the windows achieved the highest score in *Range*. This was due to a change in the data trend. We identified by the low value of Range for all iterations that the boundaries calculated in the first section of the series could no longer be applied. In that case, if the metrics were calculated daily after data acquisition, after collecting a significant number of followed *Range* values different from 1, we could make the decision to recompute the maximum and minimum values allowed. In the calculation we would use the consumption values for the year 2020, which is when a change in the distribution of the data was identified. The values identified

TABLE 10. Values out of range in 2019.

Date	Consumption (kWh)
2019-01-23 17:00:00	0.185
2020-10-25 01:00:00	0.027
2019-10-25 02:00:00	0.0
2019-10-25 03:00:00	0.0
2019-10-25 04:00:00	0.0
2019-10-25 05:00:00	0.004
2019-12-13 09:00:00	0.099

TABLE 11. Values out of range in 2020.

Date	Consumption (kWh)
2020-01-29 09:00:00	0.136
2020-01-30 14:00:00	0.027

out of range in the second half of the time series using the new boundaries were listed in Table 11.

VII. DECISION SUPPORT SYSTEM FOR DATA QUALITY MANAGEMENT AND IMPROVEMENT

The work presented in this article contributes an important advance toward the development and integration of Decision Support System (DSS) to improve the quality management of streaming time series data.

A system that supports the decision-making process to improve the quality of streaming time series data has been designed and implemented based on the methodology and the `dqts` R package. The system provides functionalities and mechanisms to assess and handle problems regarding the quality of the data. In addition, it provides support for the end user and helps to select the most appropriate alternatives, regardless of the characteristics of the end user (that does not require an analyst profile to use the module).

The *DQ-REMAIN (Data Quality REport MAnagement and ImprovemeNt)* system, improves the data management, focuses on the DQ improvement, and consequently, the final exploitation through the KDD (Knowledge Discovery in Databases) process. The findings will be more accurate and will improve the competitiveness and capabilities in areas where these types of systems are integrated. Manufacturing and energy sectors, for example, could be potential users and beneficiaries of this type of systems.

The functionality of the system is described by 4 functional blocks that can be seen in Figure 15 and are described below.

- Database connectivity: it provides a connection to the data source where the information is located or acquired (data streaming). It provides access to the data, represented as "Raw data".
- Engine: it calls the functions provided by the `dqts` module to generate the results for the data quality assessment.
- `dqts` module: it provides the functionalities associated with DQ assessment.
- User Interface: it informs throughout the visualisation about the status of the DQ based on the representation over time of the value obtained for the "Metrics" and also for the "Quality index". In addition, it facilitates interaction between the module and the end user to solve quality problems.

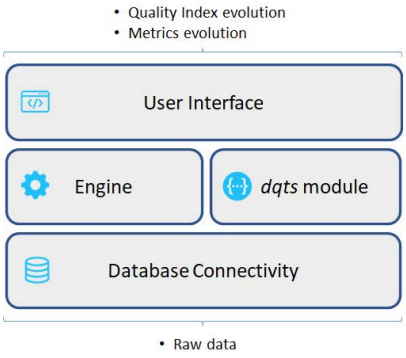


FIGURE 15. Functional blocks of DSS.

The system provides a new functionality that, integrated into existing data management, monitoring and analysis platforms, allows monitoring and correcting the quality of the data over time. The quality of the data improves significantly and data do not lose value for further exploitation.

Figure 16 presents the architecture of the solution for deploying the *DQ-REMAIN* system in a manufacturing information management platform. The platform is responsible for acquiring and managing all information, such as condition monitoring data along with process data, from its assets (test benches, machines, industrial robots, etc.). The platform acquires the necessary information and stores it in different databases for later exploitation using advanced analytical techniques. This enables the user to generate new insights, obtain an asset health assessment, and be able to perform predictive or condition-based maintenance. However, this generation of new knowledge depends on both the amount and quality of data received. The deployment of the *DQ-REMAIN* system within the platform, periodically computes the quality of the data received and assists in solving data quality issues.

The *DQ-REMAIN* system provides a graphical interface through which the status of the quality of the data monitored from the different assets can be visualised. Figure 17 shows an example of the computation of the quality of data obtained from a test bench. After detecting that the quality of the data has decreased over a period of time, the list of metrics that need to be improved in order of priority is shown by the *DQ-REMAIN* system. In the example of Figure 18, there are some metrics that cause a decrease in the quality index value: *TimeUniqueness*, *Timeliness* and *Completeness*. The system offers alternatives to correct the problems that arise depending on the metric being affected. Users can then select different methods to solve the low quality of each metric. In this example, the system offers some alternatives, a list of the algorithms available to solve repetitions in time variable, create missing dates and impute missing values. The end user

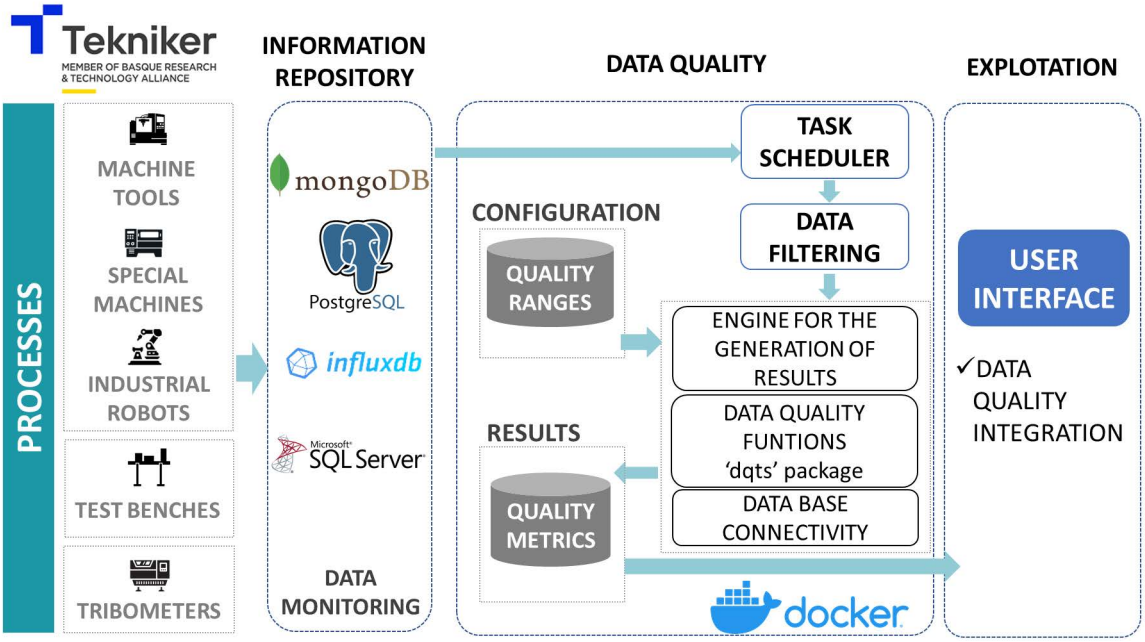


FIGURE 16. DQ-REMAIN integration into Manufacturing platform.

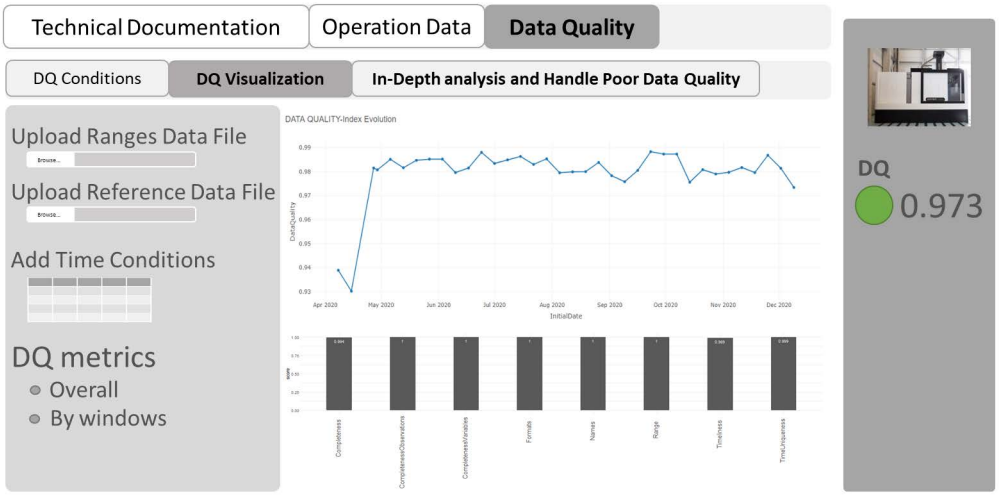


FIGURE 17. Visualisation of the computation of DQ metrics of data from a test bench.

can use the algorithm that best suits the situation and recalculates metrics. If the solution proposed by the system is good enough and the metrics improve, then it is possible to save new data in the database for future exploitation guaranteeing their highest quality.

There are many scenarios from industrial applications of IoT where the ingestion of data through sensors presents certain difficulties that affect the quality of the collected data and this type of decision support system is exploitable. Furthermore, as it is composed of functional blocks (library packages), the deployment and distribution of software for its integration into third-party solutions is carried out quickly and easily. Only ad-hoc development will be necessary to connect with new data repositories and to integrate the visual exploitation in another type of interface.

At present, the DSS is being integrated and will be used in the facilities of Tekniker (a Technology Research Centre, www.tekniker.es) to monitor and correct quality problems in certain operating machines (mainly test benches and machining centres). The results of this improvement in the quality of the data will be exploited in the near future to make use of this good quality data, and implement a health assessment of the state of the machines. This is one more step towards Smart Facturing Systems for Industry 4.0.

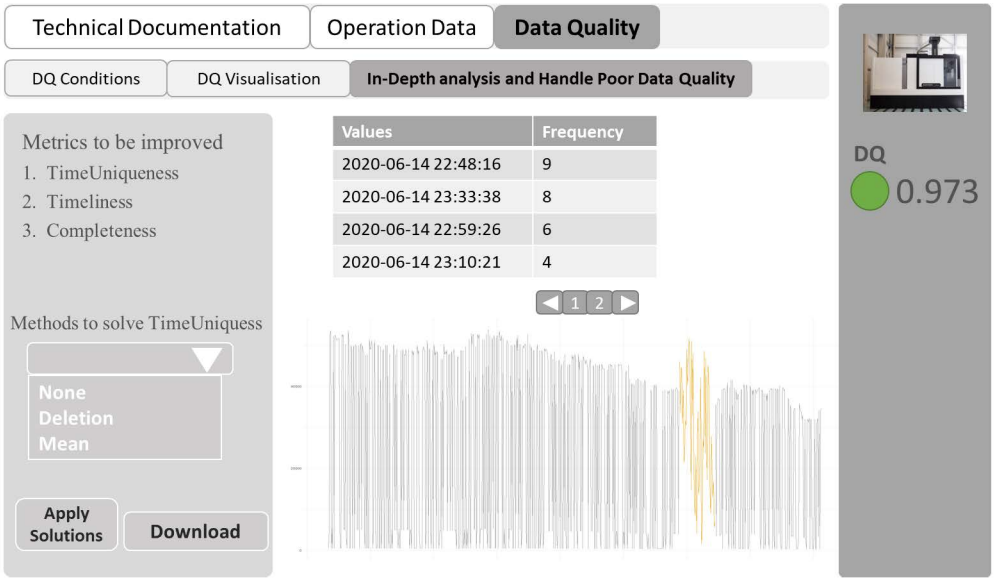


FIGURE 18. Interactive dashboard in which to choose low quality resolution methods and inspect the results.

VIII. CONCLUSION

In a world in which data plays a leading role in the evolution and fruitfulness of industries and companies, ensuring the quality of the data captured is a problem of great importance. Low-quality data greatly influences future results and conclusions reached from them. In this study, the need to define and develop a set of methodologies, procedures, and practical tools for the treatment of data quality focused on measuring and treating different aspects in time series data, the usual structure of data captured by IoT devices, has been identified. The technological novelty of this study addresses this gap through:

- The mathematical definitions of the metrics to quantify the quality of the data, beyond the existing theoretical formulation and the definition of the DQ indicator as a combination of them.
- Definition of the set of corrective functions, enhanced alternatives, to improve some aspects of quality (e.g., data imputation for missing values).
- Definition and representation of a complete methodology that describes the optimal workflow for a global process of detection, inspection, and resolution of potential problems regarding the quality of data in a time series.
- The implementation and packaging of all functions for their final use and exploitation.

This work has also been tested and validated in three data sets from different sources with different purposes to demonstrate the capacity of DQ metrics, compare the results from forecasting models model in both cases, address the quality of the data and without having done so and finally, to show the possibility of implementing the data quality treatment flow in a real case of data handled by an industrial sector.

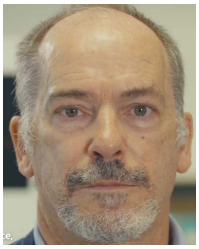
Improving the DQ is an essential task that must be addressed early, before, and during the process of transferring and storing data for its final exploitation. Doing this from the origin avoids many problems in the later phases. From the point of view of data analysis and the generation of detection and prediction models, having good quality data means drawing more precise and realistic conclusions. The models were developed from algorithms that learn based on the data provided. The better the data are, the more representative the models will be with respect to reality. The proposed research work presents two mechanisms to analyse and improve data quality in the early stages, once the data are collected and stored before the development of the model. First, the `dqts` package which implements all the required functionalities. Second, the decision support system, called *DQ-REMAIN*, which facilitates the management of the methodology proposed in this work for handling of the quality of data based on the `dqts` package.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
[2] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in Internet of Things: A state-of-the-art survey," *J. Netw. Comput. Appl.*, vol. 73, pp. 57–81, Sep. 2016.
[3] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification* (Integrated Series in Information Systems). New York, NY, USA: Springer, 2016.
[4] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017.
[5] M. H. Frické, "Data-information-knowledge-wisdom (DIKW) pyramid, framework, continuum," in *Encyclopedia Big Data*. Cham, Switzerland: Springer, 2018.
[6] S. Moore, "How to create a business case for data quality improvement," Gartner Inc., 2017. [Online]. Available: https://www.gartner.com/en/documents/3872887
[7] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Commun. ACM*, vol. 41, no. 2, pp. 79–82, Feb. 1998.
[8] S. Moore, "How to stop data quality unfermining your business," Gartner Inc., 2018.

[9] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, "Requirements for data quality metrics," *J. Data Inf. Qual.*, vol. 9, no. 2, pp. 1–32, Jan. 2018.
[10] C. Cichy and S. Rass, "An overview of data quality frameworks," *IEEE Access*, vol. 7, pp. 24634–24648, 2019.
[11] C. G. Jacobs, "Challenges to the quality of data-quality measures," *Food Chem.*, vol. 113, no. 3, pp. 754–758, Apr. 2009.
[12] W. H. Woodall, "The statistical design of quality control charts," *Statistician*, vol. 34, no. 2, p. 155, 1985.
[13] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.
[14] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996.
[15] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Berlin, Germany: Springer, 2006.
[16] P. Hodson, "Review: Data quality for the information age," *Comput. Bull.*, vol. 39, no. 6, p. 31, Dec. 1997.
[17] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.
[18] L. Orfanidis, P. D. Bamidis, and B. Eaglestone, "Data quality issues in electronic health records: An adaptation framework for the Greek health system," *Health Informat. J.*, vol. 10, no. 1, pp. 23–36, Mar. 2004.
[19] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A survey on data quality: Classifying poor data," in *Proc. IEEE 21st Pacific Rim Int. Symp. Dependable Comput. (PRDC)*, Nov. 2015, pp. 179–188.
[20] A. Klein and W. Lehner, "Representing data quality in sensor data streaming environments," *J. Data Inf. Qual.*, vol. 1, no. 2, pp. 1–28, Sep. 2009.
[21] M. Bovee, R. P. Srivastava, and B. Mak, "A conceptual framework and belief-function approach to assessing overall information quality," *Int. J. Intell. Syst.*, vol. 18, no. 1, pp. 51–74, Jan. 2003.
[22] I. El Alaoui, Y. Gahi, and R. Messoussi, "Big data quality metrics for sentiment analysis approaches," in *Proc. Int. Conf. Big Data Eng.*, Jun. 2019, pp. 36–43.
[23] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, p. 2, May 2015.
[24] C. C. G. Rodríguez and S. Servigne, "Managing sensor data uncertainty," *Int. J. Agricult. Environ. Inf. Syst.*, vol. 4, no. 1, pp. 35–54, Jan. 2013.
[25] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surveys*, vol. 41, no. 3, pp. 1–52, Jul. 2009.
[26] A. Maydanchik, *Data Quality Assessment* (Data quality for practitioners). Basking Ridge, NJ, USA: Technics Publications, 2007.
[27] R. Gitzel, S. Turring, and S. Maczey, "A data quality dashboard for reliability data," in *Proc. IEEE 17th Conf. Bus. Informat.*, Jul. 2015, pp. 90–97.
[28] R. Gitzel, "Data quality in time series data: An experience report," in *Proc. CEUR Workshop*, 2016, pp. 41–49.
[29] R. Gitzel, S. Subbiah, and C. Ganz, "A data quality dashboard for CMMS data," in *Proc. 7th Int. Conf. Oper. Res. Enterprise Syst.*, 2018, pp. 170–177.
[30] A. Richter, C. Schmidt, M. Krüger, and S. Struckmann, "DataquieR: Assessment of data quality in epidemiological research," *J. Open Source Softw.*, vol. 6, no. 61, p. 3093, May 2021.
[31] Y. Dong, Y. Kazachkova, M. Gou, L. Morgan, T. Wachsman, E. Gazit, and R. I. D. Birkler, "RawHummus: An R Shiny app for automated raw data quality control in metabolomics," *Bioinformatics*, vol. 38, no. 7, pp. 2072–2074, Mar. 2022.
[32] N. Martin, *daqapo: Data Quality Assessment For Process-Oriented Data*, R package version 0.3.1, 2020. [Online]. Available: https://cran.r-project.org/web/packages/daqapo/daqapo.pdf
[33] A. Jain, *StatMeasures: Easy Data Manipulation, Data Quality and Statistical Checks*, R package version 1.0, 2015.
[34] C. Ryu, *dlookr: Tools for Data Diagnosis, Exploration, Transformation*, R package version 0.5.4, 2021.
[35] E. Waring, M. Quinn, A. McNamara, E. A. De La Rubia, H. Zhu, and S. Ellis, *skimr: Compact and Flexible Summaries of Data*, R package version 2.1.3, 2021.
[36] A. Hebbali, *xplorerr: Tools for Interactive Data Exploration*, R package version 0.1.2, 2021.
[37] L. Ehrlinger, E. Rusz, and W. Wöß, "A survey of data quality measurement and monitoring tools," *Frontiers Big Data*, vol. 5, p. 28, Mar. 2022.
[38] R. Adhikari and R. Agrawal, "An introductory study on time series modeling and forecasting," Ph.D. thesis, Lambert Academic, Germany, 2013.
[39] J. Luengo, S. García, and F. Herrera, "On choice best imputation methods for missing values considering three groups classification methods," *Knowledge Information Systems*, vol. 32, no. 1, pp. 77–108, 2012.
[40] R. K. Elissavet, H. Spyros, and T. Ioannis, "Missing data in time series and imputation methods," M.Sc. dissertation, Univ. Aegean, Samos, Greece, Feb. 2017, p. 65.
[41] M. Gomez-Omella, I. Esnaola-Gonzalez, and S. Ferreiro, "Short-term forecasting methodology for energy demand in residential buildings and the impact of the COVID-19 pandemic on forecasts," in *Artificial Intelligence XXXVII* (Lecture Notes in Computer Science), M. Bramer and R. Ellis, Eds., vol. 12498. Cham, Switzerland: Springer, 2020, pp. 227–240.
[42] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, nos. 3–4, p. 591, Dec. 1965.
[43] M. Geurts, G. E. P. Box, and G. M. Jenkins, "Time series analysis: Forecasting and control," *J. Marketing Res.*, vol. 14, no. 2, p. 269, 1977.

MERITXELL GÓMEZ-OMELLA received the B.Sc. degree in mathematics from the Universitat Autònoma de Barcelona (UAB), in 2018, and the M.Sc. degree in modeling and mathematical, statistical and computing research from the University of the Basque Country (UPV/EHU), in 2019, with a master's thesis related to data quality and missing data imputation in collaboration with the Research Center Tekniker, where she is currently pursuing the Ph.D. degree in computer science engineering about analysis, modeling and forecasting of time series data in industrial environments. She is also a Data Scientist of the Intelligent Systems Unit, TEKNIKER, and she has been working in different national and European projects.



BASILIO SIERRA is currently a Full Professor with the Computer Sciences and Artificial Intelligence Department, University of the Basque Country (UPV/EHU). He is the Co-Director of the Robotics and Autonomous Systems Group, RSAIT. He is also a Researcher in the fields of robotics and machine learning, where he is working on the use of different paradigms to improve robot's behaviors. He works as well in multidisciplinary applications of machine learning paradigms, in agriculture, natural language processing, medicine, and so forth. He has published more than 50 journal articles, and several book chapters and conference papers.

SUSANA FERREIRO received the degree in computing science from the Technical School, University of the Basque Country, Spain, in 2005, and the Ph.D. degree *(cum laude)* in the area of probabilistic models for artificial intelligence and data mining, in 2012. Since 2005, she has been working with TEKNIKER in national and European projects in the application of the artificial intelligence for diagnosis and prognosis applied to different industrial sectors. She is also the Head of Predictive Analytics Research Line and a Senior Data Scientist of the Intelligent Systems Unit, TEKNIKER. She has several publications in conferences, journals with impact index and has collaborated in many books as well as in the supervision of theses. Her research interests include the technologies related to predictive analytics for data processing and analysis, including the descriptive and prescriptive analytics.