Moestue, Lars; Obermeier, Andreas; Widmann, Torben; Klier, Mathias

Assessing Completeness of IoT Data: A Novel Probabilistic Approach

Business & Information Systems Engineering

Provided in Cooperation with:
Springer Nature

Suggested Citation: Moestue, Lars; Obermeier, Andreas; Widmann, Torben; Klier, Mathias (2024) : Assessing Completeness of IoT Data: A Novel Probabilistic Approach, Business & Information Systems Engineering, ISSN 1867-0202, Springer Fachmedien Wiesbaden, Wiesbaden, Vol. 67, Iss. 6, pp. 797-814,
https://doi.org/10.1007/s12599-024-00889-0

This Version is available at:
https://hdl.handle.net/10419/333373

Bus Inf Syst Eng 67(6):797–814 (2025)
https://doi.org/10.1007/s12599-024-00889-0

RESEARCH PAPER

Assessing Completeness of IoT Data: A Novel Probabilistic Approach

Mathias Klier · Lars Moestue · Andreas Obermeier · Torben Widmann

Received: 26 September 2023 / Accepted: 9 July 2024 / Published online: 12 August 2024

Abstract The Internet of Things (IoT) is one of the driving forces behind Industry 4.0 and has the potential to improve the entire value chain, especially in the context of industrial manufacturing. However, results derived from IoT data are only viable if a high level of data quality is maintained. Thereby, completeness is especially critical, as incomplete data is one of the most common and costly data quality defects in the IoT context. Nevertheless, existing approaches for assessing the completeness of IoT data are limited in their applicability because they assume a known number of real-world entities or that the real-world entities appear in regular patterns. Thus, they cannot handle the uncertainty regarding the number of real-world entities typically present in the IoT context. Against this background, the paper proposes a novel, probability-based metric that addresses these issues and provides interpretable metric values representing the probability that an IoT database is complete. This probability is assessed based on the detection of outliers regarding the deviation between the estimated number of real-world entities and the number of digital entities. The evaluation with IoT data from a German car manufacturer demonstrates that the provided metric values are useful and informative and can discriminate well between complete and incomplete IoT data. The metric has the potential to reduce the cost, time, and effort associated with incomplete IoT data, providing tangible benefits in real-world applications.

Keywords Data quality · Data quality assessment · Completeness · Internet of Things · Probability-based metric

1 Introduction

The Internet of Things (IoT) is one of the driving forces behind Industry 4.0 (Okano 2017; Pivoto et al. 2021; Valderas et al. 2023) and has the potential to enable new business models, optimized supply chains, and new revenue streams (Kashyap 2022; Palmaccio et al. 2021; Steininger 2022; Vass et al. 2021). It is predicted to have enormous economic potential (Hamdan et al. 2022; Valderas et al. 2023) with an estimated impact of $11.1 trillion by 2025 (Edquist et al. 2021). Based on the vast amount of data that IoT devices produce (El-Hasnony et al. 2021; Gubbi et al. 2013; Valderas et al. 2023) and the high speed at which the data is created (Cai et al. 2017; Rahimi et al. 2018), big data from IoT can for example be used to automate production processes (Delsing et al. 2016), improve the efficiency of existing processes (Fatima et al. 2022), and support predictive maintenance (Compare et al. 2020). However, on the hand, all such applications require high-quality data to provide viable results (Jugulum 2016) and a lack of data quality can lead to incorrect decision-making and poor outcomes (Liu et al. 2020; Teh et al. 2020). On the other hand, including many sensors (Loebbecke and Boboschko 2020) and network connections

Accepted after two revisions by Natalia Kliewer.

M. Klier (✉) · L. Moestue · A. Obermeier · T. Widmann
Institute of Business Analytics, University of Ulm,
Helmholtzstraße 22, 89081 Ulm, Germany
e-mail: mathias.klier@uni-ulm.de

L. Moestue
e-mail: lars.moestue@uni-ulm.de

A. Obermeier
e-mail: andreas.obermeier@uni-ulm.de

T. Widmann
e-mail: torben.widmann@uni-ulm.de

(Powell et al. 2022), IoT systems are susceptible to a large number of potential sources of data quality problems (Teh et al. 2020). Therefore, it is essential to be able to assess the quality of IoT data (Mützel and Tafreschi 2021; Scheider et al. 2023) to identify and manage potential quality problems. However, this assessment of data quality poses a major challenge, as the large volume of data and the high velocity of data generation (Bansal et al. 2021; Fernandes and Wagh 2019) make a manual assessment impossible (Abbasi et al. 2016; Costantini et al. 2021; Evron et al. 2022; Karkouch et al. 2016).

One of the most frequently observed (Liu et al. 2020) and economically most costly (Côrte-Real et al. 2020; Liu et al. 2020) data quality defects in the IoT context is incomplete data. Therefore, in this paper, we focus on the data quality dimension completeness, which is defined "as the degree to which a given data collection includes data describing the corresponding set of real-world objects" (Batini et al. 2009, p. 7). In practice, it is not readily possible to assess the completeness of IoT data, i.e., whether all real-world entities under consideration are actually represented by respective data in the IS. In particular, a simple and straightforward assessment of completeness in terms of comparing the number of real-world entities with the number of digital entities (Pipino et al. 2002) is not feasible for IoT data due to the uncertainty regarding the typically unknown number of real-world entities that should be represented in the IS (Jugulum 2016). In the case of a German car manufacturer (cf. Sect. 5), for example, almost one million digital entities are generated every day for just one specific production process and the car manufacturer cannot say with certainty what the definite number of real-world entities (i.e., process runs) is daily. Thus, a metric for completeness of IoT data must deal with this uncertainty (Karkouch et al. 2016). However, existing approaches for assessing completeness assume that the number of real-world entities to be represented in the IS is known and neglect the associated uncertainty (cf. Sect. 3). To alleviate this drawback, we design and evaluate a novel probability-based metric to assess the completeness of IoT data considering the underlying uncertainty.

We argue that the principles and the knowledge base of probability theory provide well-founded methods for describing and analyzing such situations under uncertainty. Therefore, our completeness metric is grounded in probability theory and is based on an estimate of the number of real-world entities (e.g., using time series forecasting) to deal with the associated uncertainty. Furthermore, it assesses the probability that the IoT data is complete, by comparing the deviation of the estimated number of real-world entities from the number of entities actually stored in the IS. We demonstrate the practical applicability of our metric and evaluate its values, which represent probabilities, using the case of a German car manufacturer that stores IoT data from its production facilities. The results of two instantiations of our metric show that our metric can distinguish very well between complete and incomplete IoT data and thus support decision-making.

The remainder of the paper is organized as follows. In Sect. 2, we illustrate the problem context using the running example of an IoT system monitoring a manufacturing plant with numerous industrial robots. In Sect. 3, we provide an overview of prior works and outline the research gap to be addressed. In Sect. 4, we develop a novel probability-based metric for completeness of IoT data. We instantiate our metric in cooperation with a large German car manufacturer and evaluate the metric values in Sect. 5. Finally, we conclude with implications for theory and practice, reflect on limitations, and provide an outlook on further research in Sect. 6.

2 Problem Context

The IoT serves as a basis for countless applications in both personal and business contexts (Miles et al. 2018; Yang et al. 2022). In general, IoT refers to the interconnection of machines and devices of all types and sizes over the internet, enabling the creation of data that can provide analytical insights and support new operations (Nord et al. 2019; Valderas et al. 2023). The complex architecture of IoT systems with multiple devices and data sources (Byabazaire et al. 2020; Powell et al. 2022; Valderas et al. 2023), in combination with the large volume and high velocity at which the data is generated (Cai et al. 2017; Rahimi et al. 2018), makes the assessment of data quality in general, and completeness in particular, a major challenge. Indeed, assessing the completeness of IoT data requires an automated approach that takes into account the large number of potential sources of error as well as the high velocity and large volume of data.

To illustrate the challenges of assessing completeness in the IoT context, we introduce the example of a manufacturing plant, which will serve as a running example throughout the paper. In this manufacturing plant, several industrial robots repeatedly work on a particular process (e.g., applying adhesive seams). All robots are part of an IoT system and are equipped with multiple sensors that monitor each robot. We refer to each of the process runs (e.g., the application of one adhesive seam to a workpiece) as a real-world entity that arises from the respective process. For each real-world entity, the robots' sensors monitor and record various parameters such as timestamps, process parameters such as pressure, and environmental conditions such as temperature. The data for each real-world entity is sent to an IoT database upon completion,

where it is stored as a digital entity, i.e., the digital representation with the collected information regarding the corresponding real-world entity. However, due to issues such as sensor malfunctions, network problems, or storage errors, the digital entity may not be stored (i.e., not represented) in the IoT database. This would constitute a defect regarding the data quality dimension completeness. Incompleteness of IoT data, caused by the lack of representation of individual real-world entities in the IS, is illustrated in Fig. 1. Here, the real-world entities 2 and (n-1) – and thus the corresponding process runs – are not represented by digital entities in the IoT database. Such defects are referred to as relational completeness defects (Batini and Scannapieco 2006; Klein and Lehner 2009). In the following, we focus on relational completeness (Batini and Scannapieco 2006; Klein and Lehner 2009), as this form of completeness is particularly important and challenging (especially) in the IoT context (Liu et al. 2020). While assessing other forms of completeness such as tuple completeness (i.e., whether all attributes have a corresponding value or, for example, a specific sensor reading misses a value in the digital entity) is equally important (Batini and Scannapieco 2006), there already exist reliable approaches that can be easily applied to IoT data as well (cf. Sect. 3). Typically, these approaches inherently assume that relational completeness is fulfilled, which further motivates the need for a metric to assess relational completeness.

The assessment of relational completeness is fundamentally rooted in the comparison of the number of real-world entities (left side in Fig. 1) with the corresponding number of digital entities (right side in Fig. 1). However, the actual number of real-world entities is typically unknown and a large number of potential sources of error occurs. For example, in the case of the German car manufacturer that we will use to demonstrate and evaluate our approach (cf. Sect. 5) and the single production process *Bonding*, there are more than 250 individual robots equipped with multiple sensors, each of which generates almost one million digital entities and about 250 megabytes of data per day. At the same time, the actual number of process runs that should be represented in the IS is not known with certainty. Thus, an appropriate data quality metric must account for the underlying uncertainty regarding both the actual number of real-world entities and possible sources of error when assessing the completeness of IoT data in an automated way.

3 Related Work

Completeness is one of the most important data quality dimensions to assess, especially given the high costs that incomplete data causes (Miao et al. 2022; Zhang et al. 2019). This is particularly true in the context of the IoT. Studies show that maintaining a high level of completeness

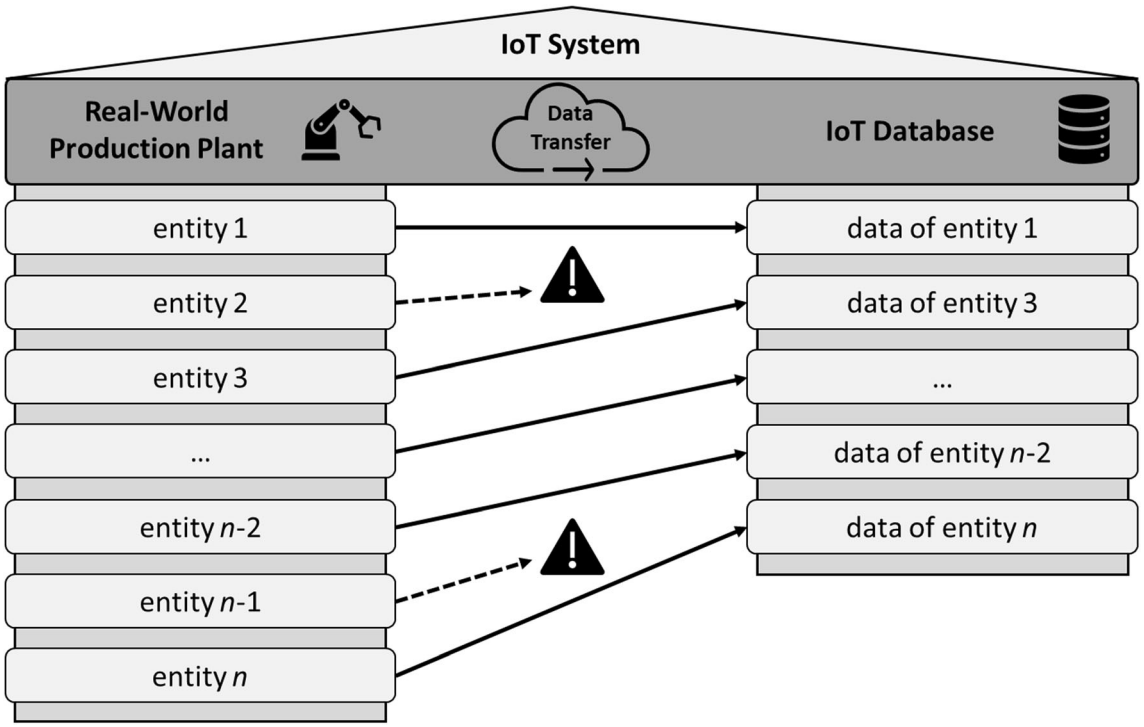

Fig. 1 Schematic representation of the running example

of IoT data is positively correlated with an organization's competitive advantage (Côrte-Real et al. 2020; Ge et al. 2018; Liu et al. 2020; Scheider et al. 2023). In the following, we first take a broader view of the data quality dimension completeness, focusing on concepts and ideas that are potentially applicable to IoT data. We then provide an overview of existing metrics for assessing the completeness of IoT data.

Regarding the data quality dimension completeness, the literature provides different perspectives. Schema completeness is defined as the degree to which a database schema represents the entities and their associated characteristics in the real world (Pipino et al. 2002). For instance, if the attribute 'last name' is absent in a customer database (e.g., there exists no column comprising the customers' last names), it would be deemed lacking schema completeness. Further, at the data level, the literature can be divided into two strands (Batini and Scannapieco 2006): works that assume a closed world and works that assume an open world. Under the closed-world assumption, it is assumed that all real-world entities are represented by digital entities (using a predefined schema with specified attributes) in the corresponding database and that the schema contains all necessary columns (Batini and Scannapieco 2006). In this case, completeness defects can only occur at the level of single entities if individual digital entities do not contain values for all attributes for a corresponding real-world entity (e.g., a database containing information about all 50 states in the U.S. is missing the value for the attribute *population* for one state). In this line, Cykana et al. (1996) propose to assess the so-called tuple completeness for each digital entity as the percentage of the number of attribute values actually available in the database relative to the number of specified attributes in the database. Based on this approach, for example, Wang et al. (2001) and Lee et al. (2002) developed similar metrics to measure tuple completeness. However, several authors point out that tuple completeness addresses only one relevant aspect of completeness, even when assuming a closed world (Batini and Scannapieco 2006; Laranjeiro et al. 2015; Pipino et al. 2002). They argue for considering further completeness metrics, such as column completeness, defined as the fraction of digital entities that actually contain values for specific columns, i.e., attributes (e.g., the fraction of states in the database that contain a value for the attribute *population*). However, the closed-world assumption often does not hold in practice – especially in the context of IoT systems. As pointed out in our running example, it is common for real-world entities (e.g., process runs of applying adhesive seams) not to be represented by corresponding digital entities in the IoT database, e.g., due to device malfunctions, network errors, or storage capacities (Bansal et al. 2021; Liu et al. 2020; Powell et al. 2022). Here, the closed-world assumption is not fulfilled but has to be replaced by the open-world assumption, where it is possible that (individual) real-world entities are not represented by digital entities. In this context, both Batini and Scannapieco (2006), as well as Pipino et al. (2002), propose the concept of relational completeness as the percentage of real-world entities that are actually represented by digital entities in the corresponding database. Although this metric mitigates the closed-world assumption, it still suffers from a major caveat in that it cannot cope with uncertainty about the actual number of real-world entities but assumes that this number is known. While this assumption may be realistic in simple cases such as a database containing information on only 45 states of the U.S., it is rather unrealistic for typical IoT applications (cf. Sect. 2). Nevertheless, by addressing the core of the completeness problem in the case of IoT data, namely real-world entities that are not represented by digital entities, these works can serve as a promising starting point for investigating the assessment of completeness of IoT data.

Indeed, most of the contributions to assessing the completeness of IoT data focus on relational completeness. Assuming that the number of real-world entities is known, many of them apply the relational completeness metric of Batini and Scannapieco (2006) as well as Pipino et al. (2002) in specific IoT contexts such as RFID chips in the logistics industry (van der Togt et al. 2011), contextual information processes (Anagnostopoulos and Kolomvatsos 2016), process mining systems (Janssenswillen and Depaire 2019), object tracking systems (Bardaki et al. 2010), or metrological weather measuring stations (Sicari et al. 2016, 2018). Overall, however, since in the IoT context the actual number of real-world entities is typically unknown and thus associated with uncertainty (Bansal et al. 2021; Karkouch et al. 2016; Liu et al. 2020), the methodical applicability of these metrics and approaches is rather limited and only given for selected (special) cases.

A first approach to mitigate the assumption that the actual number of real-world entities is known with certainty was presented by Biswas et al. (2006) for the case of smart home applications. They assume that real-world entities (e.g., sensor readings) are generated in regular time intervals (e.g., one reading per minute), allowing a calculation of the expected number of entities per time interval. Thereby, completeness is defined as the ratio of the number of available digital entities to the number of expected entities. This idea has been applied by various authors in a broad context of IoT devices, general wireless sensor networks (Cheng et al. 2018; Klein and Lehner 2009; Liu et al. 2014), and in the context of logistics chains (Ahmed et al. 2021). Although these approaches do not require the number of real-world entities to be known a priori, they are limited in their practical application, especially in the

context of IoT data, because they assume that entities are generated at regular and uniform intervals. However, this means that they cannot be used, for example, for sensors that respond to environmental factors such as irregular temperature changes, or, as in the running example, for sensors that monitor many robots performing different processes with varying run times (Saravanamohan et al. 2021). Therefore, these approaches are limited in their practical applicability because they cannot be used in many areas where IoT data is used.

Overall, promising methods exist for assessing data completeness, both with and without the closed-world assumption. In the context of IoT data, the closed-world assumption is not realistic. Thus, metrics for relational completeness that address an open world are needed. Building on foundational work for (classical) databases, several completeness metrics have been proposed for IoT data. However, all of these approaches are hampered in their applicability by either relying on a known definite number of real-world entities or by assuming that real-world entities are generated in regular and uniform time intervals. In fact, in the IoT context, both of these assumptions are usually not met due to the uncertainty involved. To address this research gap, in the following, we propose a novel probability-based metric that explicitly accounts for the underlying uncertainty when assessing the completeness of IoT data.

4 Development of the Probability-Based Metric for Completeness

In this section, we first outline the general setting and the basic idea of our approach. Based on this, we design our novel, probability-based completeness metric for IoT data. To reduce the manual effort when applying the metric, we provide an extension of the metric that can deal with a limited amount of quality-assured IoT data by leveraging expected value calculus.

4.1 General Setting and Basic Idea

In IoT databases, real-world entities may not be represented by digital entities for a variety of reasons, such as device malfunctions, network errors, or storage capacities (Liu et al. 2020). Therefore, assessing the completeness of IoT data (i.e., whether all real-world entities arising from the corresponding IoT processes are represented by respective digital entities in the IoT database) is crucial, as discussed in our problem context (cf. Sect. 2). In IoT systems, data is constantly and continuously created, sent, and stored over time. Thus, to account for this temporal dimension, we assess the completeness of IoT databases with respect to concrete time steps. Consequently, we consider an IoT database $D = D_1 \cup D_2 \cup \ldots \cup D_N$ as a composition of several disjoint subsets D_i, each containing the respective digital entities of the corresponding IoT processes for a given time step $i \in \{1, \ldots, N\}$. These time steps can be, for example, an hour, a day, or a specific production shift. During each time step, several real-world entities arise, such as process runs performed by industrial robots (cf. running example in Sect. 2). Each of these real-world entities $e^r_{i,k}$ with $k \in \{1, \ldots, n^r_i\}$ arising in time step i should be represented by a respective digital entity $e^d_{i,k}$ in the subset D_i of the IoT database. Hence, the number of real-world entities n^r_i arising in time step i should be equal to the number of respective digital entities $n^d_i = |D_i|$ for each time step i. However, for a variety of reasons, some real-world entities may not be represented by respective digital entities in the IoT database. This results in a difference between the number of real-world entities and the number of respective digital entities $\varepsilon^{DQ}_i = n^r_i - n^d_i$, which represents a data quality defect (i.e., incompleteness). Therefore, the IoT database D is complete with respect to time step i if and only if $\varepsilon^{DQ}_i = 0$ (Klein and Lehner 2009). On this basis, the completeness of IoT data in a database D for time step i is defined as follows:

$$COMP(D,i) = \begin{cases} true, & if\, \varepsilon^{DQ}_i = n^r_i - n^d_i = 0 \\ false, & else \end{cases} \quad (1)$$

Under the closed-world assumption, completeness as expressed in Eq. (1) would be trivial to determine since both n^r_i as well as n^d_i, and hence their difference ε^{DQ}_i would be known with certainty. However, as discussed in Sect. 3, in the IoT context this assumption is usually not met since the number of real-world entities n^r_i and hence ε^{DQ}_i are typically not known with certainty or not known at all. Therefore, the determination of completeness must be based on an estimate of the number of real-world entities, denoted by $\hat{n}_i$. Consequently, our approach is based on comparing this estimated number of real-world entities with the number of respective digital entities. More precisely, we observe the deviation $\Delta_i = \hat{n}_i - n^d_i$ of these two values. Comparing the number of digital entities with the estimated number of real-world entities avoids the problem of an unknown true number of real-world entities but at the same time introduces uncertainty. More precisely, while a non-zero difference ε^{DQ}_i (based on the true number of real-world entities) is certainly associated with incompleteness, a non-zero deviation Δ_i (based on the estimated number of real-world entities) could also be attributed to an unavoidable estimation error, since the estimate may differ from the true value to some extent. Such situations under uncertainty can be described and analyzed using well-

founded methods based on the principles and the knowledge base of probability theory. Moreover, defining the values of a data quality metric in terms of a probability has several advantages (Heinrich and Klier 2015): They have a concrete unit of measurement, are interval scaled, and can be included in the calculation of expected values. Thus, in developing our metric, we aim at a metric that is based on probability theory and that provides an indication in terms of a probability. More precisely, the values of our metric represent the probability, that for a given time step i all real-world entities $e^r_{i,k}$ are represented by respective digital entities $e^d_{i,k}$ in the corresponding IoT database, i.e., that no malfunctions or errors occurred that caused some entities not to be represented. On this basis, our probability-based metric $Q_{COMP}(D,i)$ for the completeness of IoT data in database D for time step i is defined as follows:

$$Q_{COMP}(D,i) = P(COMP(D,i) = true) \tag{2}$$

This probabilistic approach accounts for the uncertainty in the required estimate of the unknown number of real-world entities and the associated unknown estimation errors. To assess the probability from Eq. (2), we further investigate the deviation Δ_i between the estimated number of real-world entities and the number of digital entities in time step i. Without completeness defects, i.e., when all real-world entities are represented by respective digital entities in the IoT database, the deviations for all time steps can be traced back to the estimation error alone. Thus, if multiple deviations are observed from time steps without completeness defects, they are identically and independently distributed following the distribution of the estimation error. If a deviation is excessively large with respect to this distribution, a completeness defect (i.e., missing digital entities) should have occurred that caused the deviation not to be determined by the estimation error alone but to be amplified by a missing number of digital entities. Accordingly, the affected deviation in question cannot be explained based on the distribution of the deviations without completeness defects and thus represents an outlier with respect to this distribution. Therefore, the probability that the deviation for a given time step does not represent an outlier (with respect to the deviations from time steps without quality defects) corresponds to the probability that the IoT data is complete. This probability, which can be determined using statistical outlier tests, represents the value of our metric for the completeness of IoT data. In the next section, we present our approach for measuring the completeness of IoT data in detail.

4.2 Design of the Basic Model

Following the basic idea, our approach is divided into two distinct phases, as shown in Fig. 2. The objective of the first phase is to calibrate the metric, which includes two key aspects. First, it aims to establish a robust estimation model capable of accurately estimating the number of real-world entities for individual time steps. Second, it seeks to assess the distribution of the resulting estimation errors associated with this estimation model based on time steps without completeness defects. In the second phase, once calibrated, the metric can be used to calculate the metric values by determining the probability that a deviation between the number of estimated real-world entities and the number of respective digital entities does not constitute an outlier with respect to the distribution of estimation errors derived in the first phase of our approach.

In the IoT context, assessing completeness ultimately boils down to comparing the number of digital entities to the number of real-world entities. However, the exact number of real-world entities is unknown. Thus, in the first phase of our approach, the *calibration of the metric*, it is first necessary to *derive an estimation model for the number of real-world entities*. Many well-established methods use historical data for this purpose, including time series forecasting techniques like ARIMA(X), TBATS, and decomposition models (Perone 2022; Shaub 2020). Furthermore, regression models can be used that incorporate data from other databases (e.g., the number of parts produced) or the expertise of domain experts can be used. The selected estimation method can then be instantiated (e.g., by determining its model parameters on historical data), to produce reliable and accurate estimates of the number of real-world entities. However, even reliable and accurate estimates may have small deviations from the true value and are therefore subject to uncertainty. Mathematically, the estimated number of real-world entities for time step i, denoted as $\hat{n}_i = n^r_i + \varepsilon^{\hat{n}}_i$, represents the sum of the true number of real-world entities n^r_i and an identically and independently distributed estimation error $\varepsilon^{\hat{n}}_i \in \mathbb{R}$, as shown in Fig. 3. To account for this uncertain estimation error, which typically follows a normal distribution (Taylor and Letham 2018), we *assess the estimation error for the number of real-world entities during a quality-assured period* by determining a sample of estimation errors. For this purpose, we perform estimations for a set of quality-assured time steps I^q, i.e., time steps with no data quality defects Here, we can directly observe the estimation errors as the deviation from the quality-assured numbers of digital entities. The sample of identically and independently distributed estimation errors is referred to as the reference values R. Both the estimation model for real-world entities

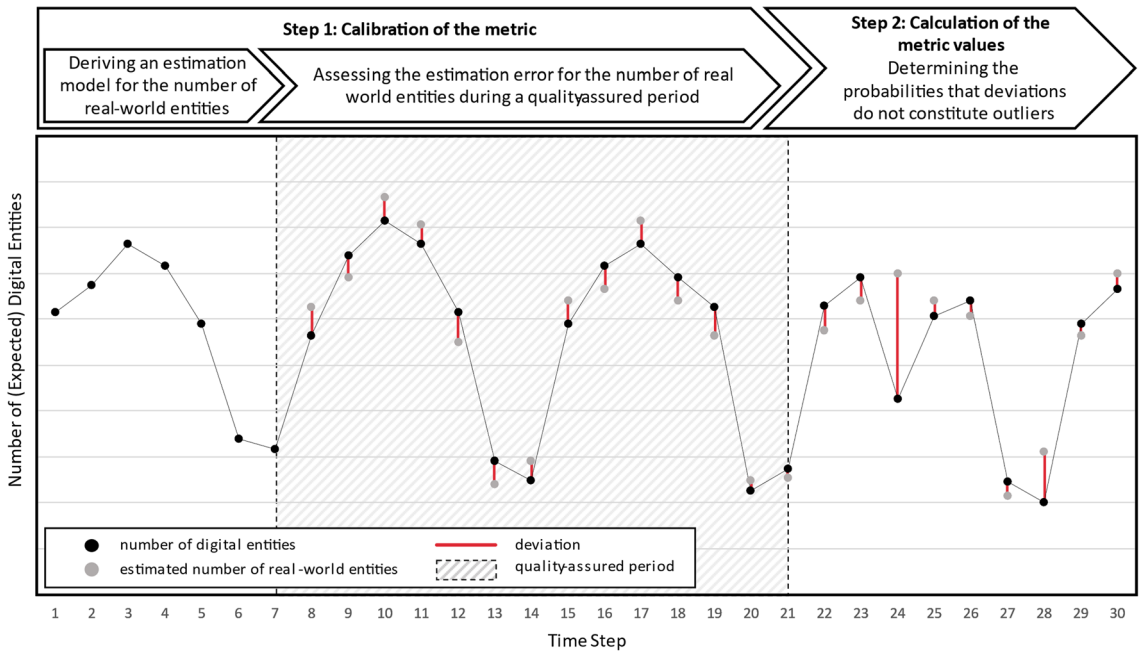

Fig. 2 Illustration of the two phases of our approach with an exemplary time series indicating the number of digital entities and the estimated number of real-world entities

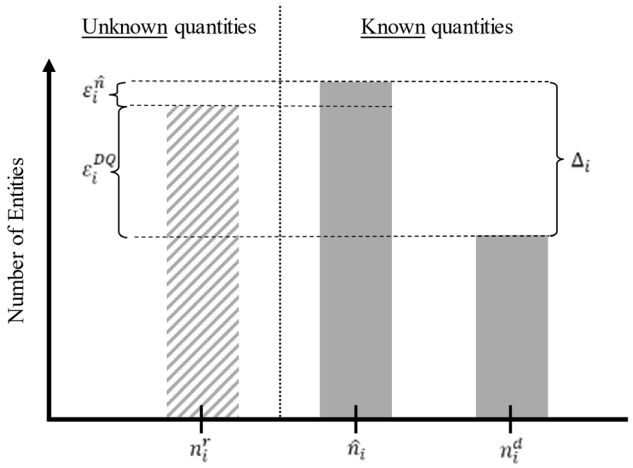

Fig. 3 Illustration of the unknown number of real-world entities n_i^r, its estimate $\hat{n}_i$, the number of digital entities n_i^d in the IoT database as well as the error terms $\varepsilon_i^{\hat{n}}$, ε_i^{DQ} and the observable deviation Δ_i for a time step i

and the reference values are required to calibrate the metric in the first phase. In the next phase, the metric can be used to obtain reliable metric values for the completeness of IoT data.

In the second phase of our approach, the *calculation of the metric values*, the metric can be applied to determine the metric values in terms of the probability that the IoT data is complete for a given time step. This is equivalent to *determining the probabilities that deviations between the estimated number of real-world entities and the number of respective digital entities do not constitute outliers*. In this vein, we assess whether the associated deviation between the estimated number of real-world entities and the number of respective digital entities is in line with the distribution of the reference values. Thereby, the more likely this deviation constitutes an outlier compared to the reference values, the more likely it is that some real-world entities are not represented by respective digital entities (i.e., the data quality error ε_i^{DQ} is greater than zero). For a rigorous assessment of completeness, the term *outlier* must be assessed objectively and on a mathematical basis. The statistical field of outlier detection offers a wide range of well-founded methods for this purpose. In the following, we will derive the assessment of the metric values in a mathematically sound and rigorous way.

Our approach focuses on the deviation $\Delta_i = \hat{n}_i - n_i^d$ between the estimate $\hat{n}_i$ and the number of digital entities n_i^d that can be directly observed for each time step i. Combining the definitions of the data quality error $\varepsilon_i^{DQ} = n_i^r - n_i^d$ and the estimated number of real-world entities $\hat{n}_i = n_i^r + \varepsilon_i^{\hat{n}}$, it becomes evident that the observable deviation Δ_i can also be calculated as the sum of the two existing error terms, $\Delta_i = \varepsilon_i^{\hat{n}} + \varepsilon_i^{DQ}$ (cf. Figure 3). This transformation underlines the relevance of the determined reference values $R = \{\Delta_j | j \in I^q\}$ as they provide the distribution of the expected deviation of our estimate for the number of real-world entities in the absence of data quality errors, i.e., $\varepsilon_i^{DQ} = 0$. To assess whether the deviation Δ_i is

in line with the distribution of estimation errors, and thus can be explained by the general estimation error $\varepsilon_i^{\hat{n}}$ alone, we can use the reference values from the first phase of our approach. If Δ_i constitutes an outlier with respect to the distribution of the reference values, this must be due to the presence of an additional error term $\varepsilon_i^{DQ} > 0$. This additional error term, which is caused by a completeness defect resulting in real-world entities not being represented by respective digital entities, amplifies the observed deviation and thus the distribution obtained from quality-assured data cannot explain the larger magnitude of the deviation. Thus, there is a congruence of outliers to the top within the deviations and the incompleteness of IoT data resulting from completeness defects. This congruence can be used to assess the completeness of IoT data by identifying outliers within the observed deviations between the expected number of real-world entities and the number of respective digital entities. Therefore, the values of our metric can be assessed by estimating the probability that the deviation observed in the time step in question represents no outlier to the top with respect to the reference values. Thus, the approach of measuring the completeness of IoT data while considering the uncertainty of possible data quality defects as well as the uncertainty regarding the actual number of real-world entities leads to the task of identifying outliers to the top within a time series composed of multiple deviations. By denoting the set of outliers with respect to the reference values as$O$, our metric (i.e., the probability that the IoT data in the database D is complete in time step$i$) is defined as

$$P(\Delta_i \notin O) = 1 - P(\Delta_i \in O) = P(COMP(D, i) = true). \tag{3}$$

Statistics and the hypothesis testing-based branch of outlier detection provide a wealth of sound methods to help estimate this probability (Chandola and Kumar 2009; Hodge and Austin 2004). The p-value, a well-known concept in hypothesis testing, can be used to provide a mathematically sound indication of whether outliers are present for given time steps (Hodge and Austin 2004). Indeed, given the null hypothesis that there is no outlier to the top in the data under consideration, the corresponding p-value represents the highest level of significance α at which this null hypothesis cannot be rejected. Applied to our context, the probability that the deviation of interest Δ_i is not an outlier with respect to the error-distribution that the reference values R follow can be assessed by means of the p-value p_i of the hypothesis test based on the null hypothesis that the deviation of interest Δ_i is indeed no outlier to the top with respect to all reference values.

There are many well-established methods in the field of outlier detection, most of which are based on statistical hypothesis testing (Chandola and Kumar 2009; Hodge and Austin 2004). The Grubbs test (Grubbs 1969; Stefansky 1972; Thompson 1935) is one of the most widely used methods, mainly because it is reliable, robust, computationally inexpensive (Urvoy and Autrusseau 2014), and does not require manually adjusted parameters (Hodge and Austin 2004). Based on these advantages, we apply the one-sided Grubbs test with the null hypothesis that there is no outlier at the top at time step i and use the corresponding p-value to determine $P(\Delta_i \notin O)$. In general, the test statistic G of the one-sided Grubbs test is determined as the difference between the maximum and the mean of the observed values – in our case, the observed reference values – divided by their standard deviation. This test statistic is compared to a critical value $Z_{\alpha,n}$ based on the number of values n and the significance level α. As indicated previously, $P(\Delta_i \notin O)$ can then be identified with the p-value p_i of the Grubbs test whether the deviation Δ_i constitutes an outlier with respect to all reference values R. The Grubbs test statistic G_i^a is given by the difference between the deviation of interest Δ_i and the mean $\overline{\Delta}_{i,R}$ of all reference values R as well as Δ_i, divided by the standard deviation $\sigma_{i,R}$ of all these deviations:

$$G_i^a = \frac{\Delta_i - \overline{\Delta}_{i,R}}{\sigma_{i,R}} \tag{4}$$

The null hypothesis that Δ_i is not an outlier to the top with respect to the reference values is rejected at the significance level α_i if G_i^a is greater than the critical value Z_{α_i,n_i}. This critical value Z_{α_i,n_i}, depending on the number of values n_i (all reference values as well as the deviation of interest) and the significance level α_i, is defined by

$$Z_{\alpha_i,n_i} = \frac{n_i - 1}{\sqrt{n_i}} \sqrt{\frac{t^2_{\frac{\alpha_i}{n_i},n_i-2}}{n_i - 2 + t^2_{\frac{\alpha_i}{n_i},n_i-2}}}. \tag{5}$$

Here, $t^2_{\frac{\alpha_i}{n_i},n_i-2}$ represents the upper critical value for a t-distribution with $n_i - 2$ degrees of freedom at the level α_i/n_i. The p-value can also be thought of as the largest significance level α_i at which the null hypothesis cannot be rejected (i.e., the test statistic G_i^a is less than the critical value Z_{α_i,n_i}). Thus, as the significance level α_i increases, the critical value Z_{α_i,n_i} decreases (Grubbs and Beck 1972) and thus maximizing α_i is equivalent to minimizing the critical value Z_{α_i,n_i}. Consequently, the p-value p_i can be determined as the solution to the optimization problem of minimizing the critical value Z_{α_i,n_i} as a function of the significance level α_i under the condition that the null hypothesis cannot be rejected (i.e., the Grubbs test statistic G_i^a is less than the critical value Z_{α_i,n_i}). Thus, the optimization problem for assessing the p-value p_i provides the value of our metric for

the completeness of IoT data for the time step i representing the probability that the number of digital entities n_i^d is equal to the number of real-world entitiesn_i^r:

$$Q_{COMP}(D,i) = P(\Delta_i \notin O) = p_i = \underset{\alpha_i}{\text{argmin}}\left(Z_{\alpha_i,n_i} | G_i^a \leq Z_{\alpha_i,n_i}\right) \quad (6)$$

In conclusion, the developed metric for completeness of IoT data provides the probability that all real-world entities are represented by respective digital entities in this IoT database in a given time step. The calculation of this probability is based on the Grubbs outlier test. Thereby, the deviation between a reliable and accurate estimate of the number of real-world entities and the number of respective digital entities in the IoT database is compared to reference values for the deviations previously obtained using quality-assured data. Then, the Grubbs outlier test determines the probability for a given time step that the associated deviation is not an outlier (compared to the reference values), meaning that no digital entities are missing in the IoT database.

4.3 Extension of the Basic Model

Defining metric values as probabilities has many advantages such as the ease of interpretation and the possibility to calculate expected values. In this section, we present an extension of our basic model based on the calculation of expected values in order to reduce the initial (manual) effort in calibrating the metric. Indeed, in the first phase of our approach, a set of quality-assured time steps is needed to initially calibrate our metric in order to assess the completeness of IoT data. On the one hand, as many reference values as possible should be used to ensure meaningful results of the Grubbs test (Hodge and Austin 2004) and the metric. On the other hand, the effort to provide quality-assured reference values limits their number. To address this issue, we propose the following extension of our probability-based completeness metric which allows further extending the reference values without the need to obtain additional quality-assured data.

In the case of a completeness defect during time step i, the observed deviation Δ_i is unexpectedly large and would thus deteriorate the sample of reference values if included directly in the initial sample. This possible degradation by the biased deviation, however, is reflected in the determined probability p_i (representing the value of our metric, cf. Equation (6)). Therefore, we use the metric value p_i as a weight to correct the deviation (possibly degrading the sample of reference values) by basing it on the corrected number of digital entities n_i^{corr}. Thereby, n_i^{corr} represents the weighted average of the estimated number of real-world entities $\hat{n}_i$ and the number of digital entities n_i^d in the IoT database.

$$n_i^{corr} = p_i \cdot n_i^d + (1 - p_i) \cdot \hat{n}_i. \quad (7)$$

If a deviation is an outlier caused by a completeness defect, the metric value p_i is small, which causes the corrected number of digital entities n_i^{corr} being closer to the estimated number of real-world entities $\hat{n}_i$ (since the number of digital entities cannot be trusted due to the completeness defect). On the other hand, if the deviation is not an outlier, the determined metric value p_i approaches a value of 1, and the factor $(1 - p_i)$ becomes small. Then, the corrected number of digital entities remains very close to the actual number of digital entities n_i^d present in the IoT database. Consequently, the reference value that expands the initial sample of reference values is calculated by the difference $\Delta_i = \hat{n}_i - n_i^{corr}$. This allows the sample of reference values to be expanded at each time step to continuously improve the determination of the probability that the IoT data is complete, without requiring additional manual effort.

5 Demonstration and Evaluation

In this section, we demonstrate and evaluate our probability-based metric for the completeness of IoT data. First, we discuss the selected case of a German car manufacturer that uses an IoT system to monitor its production facilities. Then, we describe the instantiation and application of our metric for the real-world case. Finally, we conclude with a presentation and evaluation of the results.

5.1 Case Selection and Dataset

To demonstrate and evaluate our approach, the metric is applied to the IoT data of a German car manufacturer. The IoT data contains information about the manufacturing process of *bonding*, which involves the application of various adhesive seams to car components. Among the diverse application areas of IoT, the automotive sector in particular has seen remarkable progress (Ghosh et al. 2022). In addition to enabling smart vehicles, autonomous driving, and efficient supply chain management (Krasniqi and Hajrizi 2016; Rahim et al. 2021), the automotive sector has the opportunity to enhance production processes (Liu et al. 2012; Rahim et al. 2021) by leveraging big data to reduce costs and production downtime (Liu et al. 2012; Siddhartha et al. 2021). In this line, the bonding process serves as a prime example of a production process in which industrial robots connected by IoT perform different adhesive seams on different parts of a car (Ray and Rao

2019). The process is characterized by a high degree of automation and speed, with industrial robots applying up to several thousand adhesive seams per day (Banea et al. 2018; Banea and Da Silva 2009). Such processes are irreplaceable not only in the automotive body manufacturing process (Valášek and Müller 2015), but also in other manufacturing industries (He et al. 2020; Zhong et al. 2017), and generate a significant amount of IoT data (Banea et al. 2018; Banea and Da Silva 2009). Thus, our case of the German car manufacturer and its IoT data in the context of the manufacturing process *bonding* seems particularly suitable and relevant for an automated assessment of the completeness of IoT data.

To demonstrate and evaluate our probability-based metric for completeness, we selected IoT data from the German car manufacturer's bonding process. To allow for a rigorous evaluation, including thorough manual labeling despite the time and effort involved, we limited our scope to 22 industrial robots and the period from January 1st to May 25th, 2021. Each industrial robot executes different bonding programs over time, applying a varying number of adhesive seams to different vehicle components, depending on factors such as the type of adhesive being applied and the specific vehicle model being manufactured. Successfully stored digital entities include technical information about the bonding process, such as temperature and maximum pressure, as well as organizational information such as the program executed, time stamp, and duration of execution. To determine the completeness of the IoT data, we assessed each industrial robot individually. As time steps, we chose a so-called production day with a duration of 24 h starting at 6 a.m. and including early, late, and night shifts in their entirety, as commonly employed by the car manufacturer when calculating KPIs. During the considered time period, over four million digital entities were stored in the IoT database in 1,349 time steps of our evaluation, each representing the number of digital entities for one industrial robot during one production day. Of these 1,349 time steps, we used 616 to calibrate our metric (28 production days for each of the 22 robots). Thereby, the first seven production days are used to derive an estimation model (i.e., to train the model) and the subsequent 21 production days are used to assess the estimation error. This left a total of 733 time steps remaining to calculate and evaluate the values of our metric for completeness. To rigorously evaluate the values of our metric, we worked extensively with technical experts from the car manufacturer to laboriously derive the actual numbers of real-world entities including using another database containing information on the number of specific vehicles produced, and manually investigating and reconciling the number of adhesive seams for each vehicle model. Since this information was not readily available and required significant manual effort to obtain, this evaluation approach was only possible for the limited sample size of 22 robots. Among the resulting 733 time steps, a total of 697 (i.e., 95.1%) were manually labeled as complete while the remaining 36 (i.e., 4.9%) showed missing entities. The tremendous effort required to derive the actual number of real-world entities for evaluation purposes emphasizes the need to develop a metric for completeness of IoT data capable of accounting for the underlying uncertainty about the number of real-world entities.

5.2 Instantiation and Demonstration of the Practical Applicability of the Metric

To determine the probability that the IoT data is complete for given time steps, we instantiated both our basic model with a fixed sample of reference values as well as our extended model with a stepwise expansion of the sample of reference values. For this purpose, we built on the German car manufacturer's Databricks platform, which gave us access to a part of the Azure cloud's data lake, where all production data is stored. For both models, we used Meta's time series forecasting tool *Prophet* to estimate the number of real-world entities. *Prophet's* estimates are based on the generalized additive regression model (Hastie and Tibshirani 1987; Taylor and Letham 2018) and are known for their efficacy and accuracy in accounting for trends, seasonality, and holidays, while also incorporating regressor values if available (e.g., related additional information from another database) in the estimation model (Taylor and Letham 2018). The *Prophet* model consists of a trend component, a seasonal component, an event component that allows for the integration of regressors, and the residual error term. It is capable of representing long-term linear or logistic trends as well as trend changes, capturing periodic effects over multiple periods, and allowing the inclusion of additional variables to be taken into account, such as holidays or planned production breaks (Taylor and Letham 2018). As a result, it is particularly suitable for forecasting in complex and dynamic environments, excels at automating tasks such as trend detection, and is capable of providing accurate predictions even with limited historical data while striking a balance between automation and intuitive customizability (Jha and Pande 2021; Ning et al 2022; Taylor and Letham 2018). *Prophet's* technical details and advantages align with our observations when comparing several candidate time series models on the German car manufacturer's data, as it stood out as the best-performing model as measured by Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Specifically, we evaluated the performance of *Prophet* against classical models such as (Seasonal) Autoregressive Integrated Moving Average with eXogenous regressors

((S)ARIMAX) and TBATS, as well as modern approaches such as DeepAR, an architecture developed by Amazon based on recurrent neural networks and Light Gradient Boosting Machine (LightGBM), developed by Microsoft, which uses gradient-boosting decision trees. Table 1 summarizes the results of our comparison on data from the German car manufacturer. Due to the reasons described above and the superior performance in our comparison, we chose the *Prophet* model for our instantiation.

For the first phase of our approach, the calibration of the metric, we used a total of 28 time steps for each robot. Of these 28 time steps, seven were used to derive the estimation model, i.e., to train the *Prophet* time series model. These trained models are then iteratively applied to the remaining 21 time steps of the first phase to estimate the corresponding number of real-world entities and thus assess the distribution of the estimation error. With the resulting 21 reference values for each time series, we ensured a sufficiently large sample to apply the Grubbs test.

In the second phase of our approach, starting from the 29th day of each time series, we calculate the metric values for the remaining time steps for each industrial robot. To accomplish this, we estimate the number of real-world entities for each time step and industrial robot using the *Prophet* time series model derived in the first phase of our approach. Then, the deviation between the estimated number of real-world entities and the corresponding number of digital entities is calculated. Finally, we use the Grubbs test based on the distribution of reference values obtained in the previous phase to calculate the metric value based on each deviation. In the extended model, the sample of reference values is expanded after each time step (cf. Sect. 4.3) by using each determined probability to correct the deviation (possibly biased by a completeness problem) by calculating the corrected number of digital entities using Eq. (7).

After instantiating the metric, we automatically calculated metric values for all 733 time steps. Figure 4 shows the distribution of the relative frequency of the metric values in terms of probabilities in ten bins. For both models, our approach mainly assigned either very low or very high metric values for most time steps. This distribution pattern of the estimated probabilities is beneficial because it provides the basis for a clear and comprehensible classification. Furthermore, these results show a striking similarity between the depicted distributions and the actual proportions of incomplete (4.9%) and complete (95.1%) time steps present within the dataset.

Table 1 Performance comparison based on MAE and RSME across different candidate models for time series forecasting using data from the car manufacturer (ranked by MAE)

	MAE	RMSE
Prophet	**624.6**	**1123.8**
LightGBM	724.7	1233.7
TBATS	796.7	1172.6
DeepAR	865.7	1387.2
ARIMAX	907.2	1620.5
SARIMAX	959.6	1786.9

Best performance in bold

5.3 Evaluation of the Metric Values

We evaluate our approach with respect to two different aspects. First, we examine the metric values in terms of probabilities regarding their reliability and discriminative power. Second, we evaluate the performance of our approach considering the classification into complete and incomplete time steps. Thereby, we also compare our results with those of the commonly used Six Sigma approach.

Reliability refers to the agreement between estimated probabilities and actually observed relative frequencies (Murphy and Winkler 1977). In our context, reliability means that the determined completeness probabilities should correspond to the observed relative frequencies of complete time steps. Reliability is often evaluated using the reliability curve, which plots the estimated probabilities against the observed relative frequencies. In addition, reliability can be assessed quantitatively using the reliability score, which is defined as the mean squared deviation from the diagonal weighted by the number of pairs of data in each bin (Murphy 1973). The left part of Fig. 5 shows the reliability curve for both the basic and the extended model. With the reliability curves closely following the diagonal and reliability scores of 0.0026% (basic model) and 0.0021% (extended model), the results show that our approach provides reliable results for both versions. Considering the imbalance in the completeness assessment of IoT data (i.e., many more time steps with complete IoT data rather than time steps with incomplete IoT data), it is crucial that the determined probabilities also have high discriminative power. Thus, we assessed the discriminatory power in terms of the area under the curve (AUC) under the receiver operating characteristic (ROC) curve, which is a commonly used choice to evaluate the discrimination of a probability-based metric (Hanley and McNeil 1982; Hosmer et al. 2013). To get the ROC curve, the classification threshold is varied and the corresponding true positive rate is plotted against the false positive rate. The ROC curves for both models are shown in the right part of Fig. 5. With the ROC curves closely aligning with perfect discrimination and ROC AUCs of 92.62% (basic model) and 97.64% (extended model), the discrimination is deemed

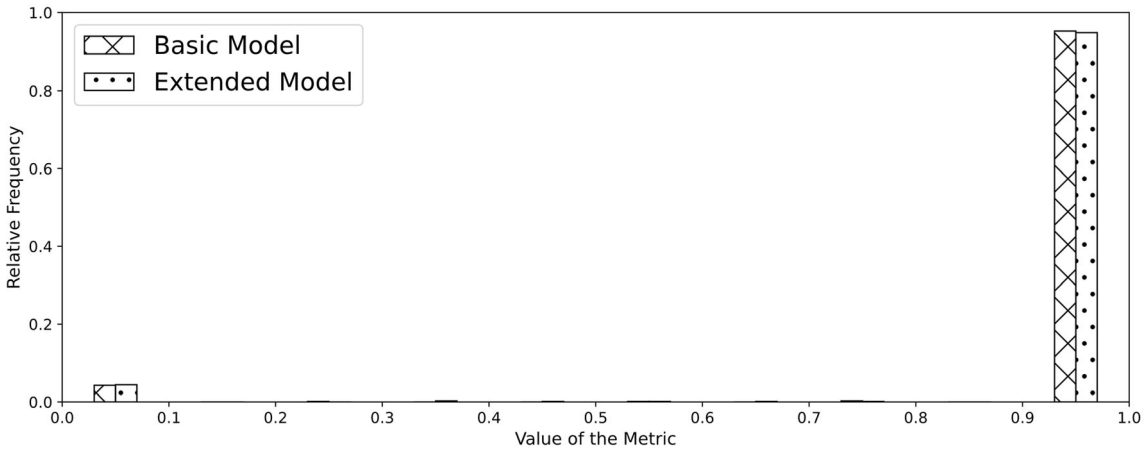

Fig. 4 Relative frequencies of the metric values in form of probabilities

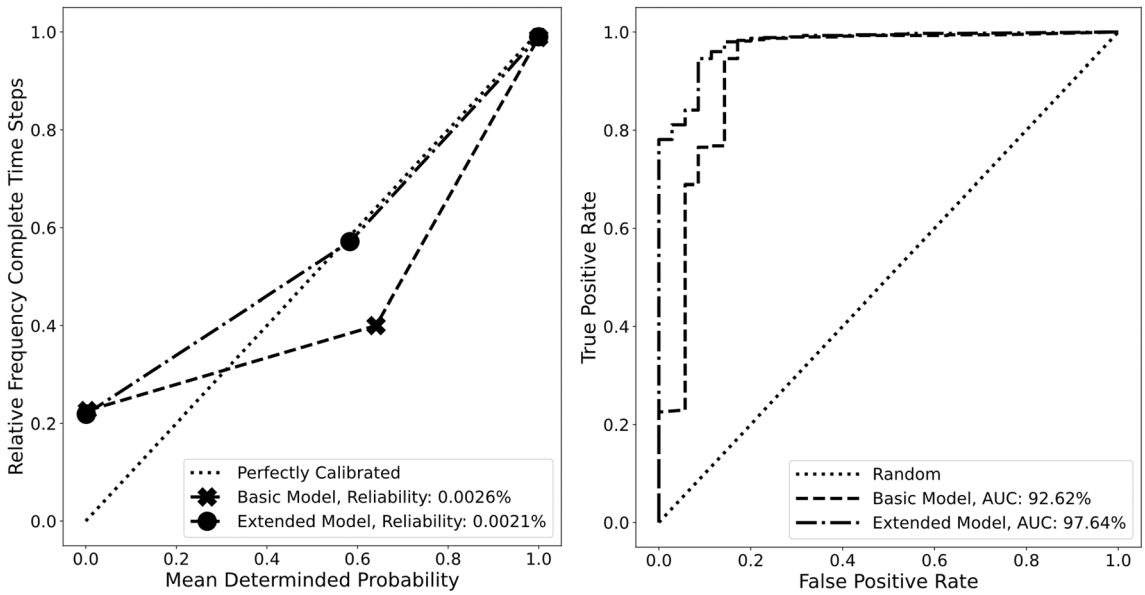

Fig. 5 Reliability curves (left) and ROC curves (right) of the basic and extended version of our metric

outstanding (Hosmer et al. 2013). Overall, these results support that the probabilities provided by our approach are reliable and have high a discriminative power.

In some applications, it is necessary to perform a binary classification into time steps with complete and incomplete IoT data. In the following, we evaluate the performance with respect to such a classification using the common performance measures accuracy, precision, recall, and F1-measure (i.e., the harmonic mean of precision and recall). To determine the values for these performance measures, we use the probabilities estimated in the versions of our approach and classify metric values by assigning them to the most likely class, i.e., we use the probability 0.5 as a natural classification threshold. As a benchmark to compare the performance, we used a method based on the idea of the *Six Sigma* management system. The method originated as a general quality management approach and emphasizes the statistical aspect. The name is derived from the statistical Six Sigma principle, which states that more than 95% of all observations of random variables (under certain weak conditions) lie in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ (where μ is the mean and σ the standard deviation of the observations) (Pukelsheim 1994). In quality management, this describes the goal of having more than 95% of all outputs within a specified tolerance range. This means that the principle can also be used for outlier detection. The idea is that outputs outside the tolerance range can be considered outliers. Mathematically, this means that values outside the interval $[\mu - 3\sigma, \mu + 3\sigma]$ are defined as outliers. Here, we determine both the mean and the standard deviation ('sigma') of the reference values of each time series and use a fixed decision boundary of three sigma

above the mean to determine whether the IoT data is complete or incomplete for each time step. In the IoT context, Six Sigma has a major advantage in quality management: The high availability of data means that production processes can be constantly monitored and controlled (Rahimi et al. 2018; Valderas et al. 2023). As a result, faulty processes can be quickly identified and improved (Tissir et al. 2023). For similar reasons, the Six Sigma method is also frequently used for outlier detection in the IoT context (Huang et al. 2019; Kale et al. 2022) and manufacturing (Lee and Lee 2022; Pugna et al. 2016), among others (Čampulová et al. 2017). It is expected to provide convincing results, as it has shown good performance in similar settings (Čampulová et al. 2017; Han and Lee 2002; Huang et al. 2019; Kale et al. 2022; Lee and Lee 2022; Pugna et al. 2016). In addition, it is easy to implement and simple to use. Therefore, for our use case, it seems to be a particularly suitable baseline to assess the performance of our metric. To this end, we followed existing implementations from the literature that have already applied this method in the IoT context (cf. e.g., Huang et al. 2019). Specifically, we examined which of the differences in observed values for each robot exceeded the respective deviation of 3σ and classified them as outliers and thus as completeness defects.

On the given dataset, our approach provides very promising results for both the basic as well as the extended model, especially – as expected – for the majority class, i.e., the *complete* class. However, due to the given imbalance of complete and incomplete time steps, the crucial performance metrics refer to the minority class, i.e., the *incomplete* class, as this class is more difficult to predict and generally of higher interest (Sun et al. 2007, 2015; Yin et al. 2013). For all performance measures considered, the two instantiations of our approach outperform the Six Sigma method (cf., Table 2).

The basic model is able to identify 71,43% of the time steps with incomplete data (recall), while the extended model is even able to identify 77.14%. Moreover, our approach is 78.13% correct with the basic model and 77.14% correct with the extended model (precision) when identifying time steps with incomplete IoT data. This significantly outperforms the Six Sigma method, which only achieves a precision of 48.08%. Consequently, the F1-measure provides good results for both of our models with 74.63% (basic model) and 77.14% (extended model), especially in comparison to the Six Sigma method with 57.47%. Overall, the extended model shows superior performance in assessing the completeness of IoT data compared to the basic model, once again highlighting the benefits of a larger sample size and the ability to calculate weighted averages based on the metric values in terms of probabilities. Both instantiations result in very high accuracies of over 97% clearly outperforming the Six Sigma method frequently used in comparable settings (Čampulová et al. 2017; Han and Lee 2002). In conclusion, these results confirm that the values of our proposed metric – for both the basic and the extended model – are reliable and can discriminate very well between complete and incomplete IoT data.

Convinced by the ease of implementation and favorable evaluation, the German car manufacturer adopted our metric to assess IoT data completeness for all bonding processes across all production sites and other techniques such as welding. In addition, dashboards were established, displaying real-time completeness defects for individual robots. In the final discussions, the data scientists of the German car manufacturer were very satisfied with respect to the metric, pointing out that for the first time it is now possible to systematically and comprehensively monitor the completeness of their IoT data. This also enables effective monitoring and optimization of production processes for increased efficiency.

6 Discussion and Conclusion

In this section, we discuss theoretical and practical implications as well as limitations of our work. Finally, we conclude with a brief summary.

6.1 Theoretical Contributions

In this paper, we designed and evaluated a novel probability-based metric for (relational) completeness of IoT data. Our contribution to research is twofold. First, unlike

Table 2 Performance measures for classification into complete and incomplete IoT data using both instantiations of our approach as well as the Six Sigma method

		Accuracy (%)	Recall (%)	Precision (%)	F1-measure (%)
Basic model	Incomplete	97.68	71.43	**78.13**	74.63
	Complete		99.00	98.57	98.79
Extended model	Incomplete	**97.82**	**77.14**	77.14	**77.14**
	Complete		98.85	98.85	98.85
Six sigma method	Incomplete	94.95	71.43	48.08	57.47
	Complete		96.13	98.53	97.32

Best performance in bold

existing metrics, which assume that the exact number of real-world entities is known exactly or that the real-world entities appear in a regular pattern, our approach prevents the closed-world assumption. Indeed, the closed-world assumption is often not satisfied in practical applications since the number of real-world entities is typically unknown. This uncertainty about the definite number of real-world entities arises from the high number of potential sources of error combined with the high velocity and large volume of IoT data (Bansal et al. 2021; Fernandes and Wagh 2019). Thus, our metric provides a new perspective for assessing the completeness of IoT data, which enables a much broader range of applications, by relying only on an estimation method and the easily obtainable number of digital entities. Second, unlike existing approaches, our proposed metric accounts for uncertainty by leveraging probability theory and provides an indication rather than a binary score. Such binary scores cannot reflect the actual uncertainty associated with the completeness of IoT data and do not provide a measure of the degree of confidence with respect to the classification into complete and incomplete data. Thus, they express a certainty that does not exist. In contrast, our metric values are interval-scaled and can be unambiguously interpreted as probabilities that adequately reflect the uncertainty associated with the occurrence of completeness defects. Furthermore, the metric values can be integrated into expected value calculations in a methodically well-founded manner. This also allows for extensions as presented in Sect. 4.3, which can be especially important for practical applications by supporting decision-making with higher quality results even in the case of limited data availability (Liu et al. 2020; Teh et al. 2020).

6.2 Practical Implications

Next to the theoretical contributions, our findings also point to practical implications. First, our evaluation supports the proof-of-concept of our metric, as it shows that the obtained metric values exhibit very good discrimination between incomplete and complete IoT data, especially when compared to the widely used Six Sigma method. This finding is well in line with previous research that demonstrates the potential and advantages of probability-based approaches to assess (other dimensions of) data quality (Heinrich and Klier 2015; Klier et al. 2021). Therefore, our metric can improve data-driven decision-making by allowing data quality and its uncertainty to be represented and incorporated into decision-making, thereby enhancing the quality of decisions made and instilling confidence in their outcomes. Overall, this leads to better decisions and has the potential to reduce the cost, time, and effort associated with incomplete IoT data, providing tangible benefits in real-world applications (Côrte-Real et al. 2020; Liu et al. 2020). Second, to realize its potential, the approach must be economically feasible. Specifically, this means that the expected benefits of applying the metric must exceed the costs. There are different types of benefits and costs associated with data quality initiatives and metrics, all of which typically depend on the context of application (cf. e.g., Batini and Scannapieco (2016) for an overview). In the concrete setting used to demonstrate our metric, the German car manufacturer was faced with a situation where data-driven projects (e.g., regarding predictive maintenance) failed due to poor data quality and, in particular, incomplete data. Based on our metric, the completeness of data can now be actively measured and managed to avoid the delay or even cancellation of such projects, which leads to economic benefits. These benefits outweigh the costs of applying our metric. More concretely, little effort of only nine person-days was necessary to instantiate and evaluate our metric after its development. Moreover, the metric can be adapted to other IoT-supported manufacturing processes with minimal effort. In this vein, the German car manufacturer has implemented the metric for further manufacturing processes, such as stud welding, and established data quality dashboards that provide real-time information on completeness for all robots involved in production. In summary, our metric can help to unlock the potential of data, and the low effort required to apply it makes it economically feasible in many contexts.

6.3 Limitations and Future Work

Despite its merits, our work also has limitations that can be a starting point for future research. In our demonstration and evaluation, we focused on a single case of IoT data regarding the bonding process of a German car manufacturer. Further research could investigate the generalizability of the approach to other cases, such as other manufacturing processes or IoT data in other contexts. Moreover, our approach provides metric values in the form of probabilities for the completeness of IoT data for predefined time steps (e.g., a production day). However, it does not provide additional insight into the specific real-world entities that are missing and the underlying reasons for the completeness defects. Future research could for example investigate the use of our approach in combination with predictive maintenance to detect potential completeness problems in advance and thus avoid them in the first place. Furthermore, while completeness is one of the most important data quality dimensions for IoT data (Côrte-Real et al. 2020; Liu et al. 2020; Miao et al. 2022), our approach does not consider other potential data quality defects such as inaccurate data. Future research could incorporate our

approach into a more holistic IoT data quality framework that combines methods for accuracy assessment (Omar et al. 2020; Tkachenko et al. 2020) and other data quality dimensions (Bai et al. 2018; Heinrich and Hristova 2016).

6.4 Conclusion

Assessing the completeness of IoT data in an automated way is an important issue in both research and practice. In this paper, we propose a probability-based approach for this task. It aims to determine the probability that an IoT database is complete for a given time step based on the detection of outliers regarding the deviation between the estimated number of real-world entities and the number of digital entities. Existing approaches are limited in their ability to accurately identify completeness defects in IoT data since they either assume that the definite number of real-world entities is known exactly or that the real-world entities appear in regular patterns. In fact, existing approaches cannot cope with the uncertainty arising from the high number of potential sources of error combined with the large volume and high velocity of IoT data. Our proposed probability-based metric for completeness of IoT data addresses these issues and yields interpretable metric values representing the probability that an IoT database is complete for a given time step. We demonstrate the practical applicability of the metric and evaluate its values based on a real-world case of a large German car manufacturer. The results show that the provided metric values are useful and informative and can well discriminate between complete and incomplete IoT data. The positive evaluation, along with the practical applicability of our metric resulted in the car manufacturer introducing the metric to assess the completeness of all bonding processes across all production facilities.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abbasi A, Sarker S, Chiang R (2016) Big data research in information systems: toward an inclusive research agenda. J Assoc Inf Syst 17:1–32

Ahmed M, Taconet C, Ould M, Chabridon S, Bouzeghoub A (2021) IoT data qualification for a logistic chain traceability smart contract. Sensors 21:2239. https://doi.org/10.3390/s21062239

Anagnostopoulos C, Kolomvatsos K (2016) A delay-resilient and quality-aware mechanism over incomplete contextual data streams. Inf Sci 355:90–109. https://doi.org/10.1016/j.ins.2016.03.020

Bai L, Meredith R, Burstein F (2018) A data quality framework, method and tools for managing data quality in a health care setting: an action case study. J Decis Syst 27:144–154. https://doi.org/10.1080/12460125.2018.1460161

Banea MD, Da Silva LFM (2009) Adhesively bonded joints in composite materials: an overview. Proc Inst Mech Eng, Part l: J Mater: Des Appl 223:1–18. https://doi.org/10.1243/14644207JMDA219

Banea MD, Rosioara M, Carbas R, Da Silva L (2018) Multi-material adhesive joints for automotive industry. Compos B Eng 151:71–77. https://doi.org/10.1016/j.compositesb.2018.06.009

Bansal M, Chana I, Clarke S (2021) A survey on IoT big data. ACM Comput Surv 53:1–59. https://doi.org/10.1145/3419634

Bardaki C, Kourouthanassis P, Pramatari K, Doukidis GI (2010) Modeling the information quality of object tracking systems. In: MCIS proceedings, Tel Aviv. https://aisel.aisnet.org/mcis2010/10

Batini C, Scannapieco M (2006) Data quality: data-centric systems and applications. Springer, Heidelberg

Batini C, Scannapieco M (2016) Data and information quality: dimensions, principles and technique. Springer, Cham

Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Comput Surv 41:1–52. https://doi.org/10.1145/1541880.1541883

Biswas J, Naumann F, Qiu Q (2006) Assessing the completeness of sensor data. In: proceedings of the 11th international conference on database systems for advanced applications, pp 717–732. Singapore. https://doi.org/10.1007/11733836_50

Byabazaire J, O'Hare G, Delaney D (2020) Using trust as a measure to derive data quality in data shared IoT deployments. In: 29th international conference on computer communications and networks, Honolulu, pp 1–9. https://doi.org/10.1109/ICCCN49398.2020.9209633

Cai H, Xu B, Jiang L, Vasilakos AV (2017) IoT-based big data storage systems in cloud computing: perspectives and challenges. IEEE Internet Things J 4:75–87. https://doi.org/10.1109/JIOT.2016.2619369

Čampulová M, Veselík P, Michálek J (2017) Control chart and Six Sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM 10. Atmospheric Pollut Res 8:700–708. https://doi.org/10.1016/j.apr.2017.01.004

Chandola V, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41:1–58. https://doi.org/10.1145/1541880.1541882

Cheng H, Feng D, Shi X, Chen C (2018) Data quality analysis and cleaning strategy for wireless sensor networks. EURASIP J Wireless Commun Netw 2018:1–11. https://doi.org/10.1186/s13638-018-1069-6

Compare M, Baraldi P, Zio E (2020) Challenges to IoT-enabled predictive maintenance for Industry 4.0. IEEE Internet Things J 7:4585–4597. https://doi.org/10.1109/JIOT.2019.2957029

Côrte-Real N, Ruivo P, Oliveira T (2020) Leveraging internet of things and big data analytics initiatives in European and

American firms: is data quality a way to extract business value? Inf & Manag 57:103141. https://doi.org/10.1016/j.im.2019.01.003

Costantini F, Galvan F, de Stefani MA, Battiato S (2021) Assessing information quality in IoT forensics: theoretical framework and model implementation. J Appl Logics 8:2373–2406

Cykana P, Paul A, Stern M (1996) DoD guidelines on data quality management. In: Proceedings of the 1st international Conference on Information Quality, Cambridge, pp 154–171

de Vass T, Shee H, Miah S (2021) IoT in supply chain management: opportunities and challenges for businesses in early Industry 4.0 context. Oper Supply Chain Manag Int J 14:148–161

Delsing J, Eliasson J, van Deventer J, Derhamy H, Varga P (2016) Enabling IoT automation using local clouds. In: 2016 IEEE 3rd World Forum on Internet of Things, Reston, pp 502–507. https://doi.org/10.1109/WF-IoT.2016.7845474

Edquist H, Goodridge P, Haskel J (2021) The Internet of Things and economic growth in a panel of countries. Econ Innov New Technol 30:262–283. https://doi.org/10.1080/10438599.2019.1695941

El-Hasnony IM, Mostafa RR, Elhoseny M, Barakat SI (2021) Leveraging mist and fog for big data analytics in IoT environment. Trans Emerg Telecommun Technol 32:e4057. https://doi.org/10.1002/ett.4057

Evron Y, Soffer P, Zamansky A (2022) Model-based analysis of data inaccuracy awareness in business processes. Bus Inf Syst Eng 64:183–200. https://doi.org/10.1007/s12599-021-00709-9

Fatima Z, Tanveer MH, Waseemullah, Zardari S, Naz LF, Khadim H, Ahmed N, Tahir M (2022) Production plant and warehouse automation with IoT and Industry 50. Appl Sci 12:2053. https://doi.org/10.3390/app12042053

Fernandes NA, Wagh R (2019) Quality assurance in big data analytics: an IoT perspective. Telfor J 11:114–118. https://doi.org/10.5937/telfor1902114A

Ge M, Bangui H, Buhnova B (2018) Big data for Internet of Things: a survey. Future Gener Comput Syst 87:601–614. https://doi.org/10.1016/j.future.2018.04.053

Ghosh RK, Banerjee A, Aich P, Basu D, Ghosh U (2022) Intelligent IoT for automotive Industry 4.0: challenges, opportunities, and future trends. In: Ghosh U et al (eds) Intelligent Internet of Things for healthcare and industry. Springer, Cham, pp 327–352

Grubbs FE (1969) Procedures for detecting outlying observations in samples. Technometrics 11:1–21. https://doi.org/10.2307/1266761

Grubbs FE, Beck G (1972) Extension of sample sizes and percentage points for significance tests of outlying observations. Technometrics 14:847–854. https://doi.org/10.2307/1267134

Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst 29:1645–1660. https://doi.org/10.1016/j.future.2013.01.010

Hamdan A, Alareeni B, Hamdan R, Dahlan MA (2022) Incorporation of artificial intelligence, Big Data, and Internet of Things (IoT): an insight into the technological implementations in business success. J Decis Syst. https://doi.org/10.1080/12460125.2022.2143618

Han C, Lee Y-H (2002) Intelligent integrated plant operation system for Six Sigma. Ann Rev Control 26:27–43. https://doi.org/10.1016/S1367-5788(02)80008-6

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hastie T, Tibshirani R (1987) Generalized additive models: some applications. J Am Stat Assoc 82:371–386. https://doi.org/10.2307/2289439

He G, Dang Y, Zhou L, Dai Y, Que Y, Ji X (2020) Architecture model proposal of innovative intelligent manufacturing in the chemical industry based on multi-scale integration and key technologies. Comput Chem Eng 141:106967. https://doi.org/10.1016/j.compchemeng.2020.106967

Heinrich B, Hristova D (2016) A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. J Decis Syst 25:16–41. https://doi.org/10.1080/12460125.2015.1080494

Heinrich B, Klier M (2015) Metric-based data quality assessment — developing and evaluating a probability-based currency metric. Decis Support Syst 72:82–96. https://doi.org/10.1016/j.dss.2015.02.009

Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22:85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. Wiley, Hoboken

Huang R, Chen Z, Liu Z, Song S, Wang J (2019) TsOutlier: explaining outliers with uniform profiles over IoT data. In: 2019 ieee international conference on big data, Los Angeles, pp 2024–2027. https://doi.org/10.1109/BigData47090.2019.9006232

Janssenswillen G, Depaire B (2019) Towards confirmatory process discovery: making assertions about the underlying system. Bus Inf Syst Eng 61:713–728. https://doi.org/10.1007/s12599-018-0567-8

Jha BK, Pande S (2021) Time series forecasting model for supermarket sales using FB-Prophet. In: proceedings of the 5th international conference on computing methodologies and communication, Erode, pp 547–554. https://doi.org/10.1109/ICCMC51019.2021.9418033

Jugulum R (2016) Importance of data quality for analytics. In: Sampaio P, Saraiva P (eds) Quality in the 21st century: perspectives from ASQ Feigenbaum Medal winners. Springer Nature, Cham, pp 23–31. https://doi.org/10.1007/978-3-319-21332-3_2

Kale V, Katke C, Dayane S, Thakar P (2022) Challenges of introducing Lean Six Sigma, IoT in Industry 4.0, and supply chain management: a review. In: Reddy ANR et al (eds) Intelligent manufacturing and energy sustainability. Springer, Singapore, pp 303–315. https://doi.org/10.1007/978-981-16-6482-3_31

Karkouch A, Mousannif H, Al Moatassime H, Noel T (2016) Data quality in internet of things: a state-of-the-art survey. J Netw Comput Appl 73:57–81. https://doi.org/10.1016/j.jnca.2016.08.002

Kashyap R (2022) The internet of value and Internet of Things. In: Vadgama N, Xu J, Tasca P (eds) Enabling the internet of value: how blockchain connects Global businesses. Springer, Cham, pp 147–156. https://doi.org/10.1007/978-3-030-78184-2_13

Klein A, Lehner W (2009) Representing data quality in sensor data streaming environments. J Data Inf Qual 1:1–28. https://doi.org/10.1145/1577840.1577845

Klier M, Moestue L, Obermeier A, Widmann T (2021) Event-driven assessment of currency of wiki articles: a novel probability-based metric. In: ICIS 2021 Proceedings, Austin. https://aisel.aisnet.org/icis2021/data_analytics/data_analytics/14

Krasniqi X, Hajrizi E (2016) Use of IoT technology to drive the automotive industry from connected to full autonomous vehicles. IFAC-PapersOnLine 49:269–274. https://doi.org/10.1016/j.ifacol.2016.11.078

Laranjeiro N, Soydemir SN, Bernardino J (2015) A survey on data quality: classifying poor data. In: proceedings of the IEEE 21st pacific rim international symposium on dependable computing, pp. 179–188. https://doi.org/10.1109/PRDC.2015.41

Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inf Manag 40:133–146. https://doi.org/10.1016/s0378-7206(02)00043-5

Lee J, Lee I (2022) Exploratory data analysis of manufacturing data. In: 13th international conference on information and communication technology convergence, Jeju Island, pp 1797–1799. https://doi.org/10.1109/ICTC55196.2022.9952974

Liu C, Nitschke P, Williams SP, Zowghi D (2020) Data quality and the Internet of Things. Comput 102:573–599. https://doi.org/10.1007/s00607-019-00746-z

Liu T, Yuan R, Chang H (2012) Research on the Internet of Things in the automotive industry. In: 2012 international conference on management of e-commerce and e-government, Bejing, pp 230–233. https://doi.org/10.1109/ICMeCG.2012.80

Liu R, Wang G, Wang WH, Korn F (2014) iCoDA: interactive and exploratory data completeness analysis. In: proceedings of the 30th international conference on data engineering, Chicago, pp 1226–1229. https://doi.org/10.1109/ICDE.2014.6816747

Loebbecke C, Boboschko I (2020) Reflecting upon sensor-based data collection to improve decision making. J Decis Syst 29:18–31. https://doi.org/10.1080/12460125.2020.1776926

Miao X, Gao Y, Chen L, Peng H, Yin J, Li Q (2022) Towards query pricing on incomplete data. IEEE Trans Knowl Data Eng 34:4024–4036. https://doi.org/10.1109/TKDE.2020.3026031

Miles A, Zaslavsky A, Browne C (2018) IoT-based decision support system for monitoring and mitigating atmospheric pollution in smart cities. J Decis Syst 27:56–67. https://doi.org/10.1080/12460125.2018.1468696

Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorol 12:595–600. https://doi.org/10.1175/1520-0450(1973)012%3c0595:ANVPOT%3e2.0.CO;2

Murphy AH, Winkler RL (1977) Reliability of subjective probability forecasts of precipitation and temperature. J Royal Stat Soc Ser C Appl Stat 26:41–47. https://doi.org/10.2307/2346866

Mützel MM, Tafreschi O (2021) Data-centric risk management for business processes. In: HICCS proceedings, Weilea, pp 5728–5737

Ning Y, Kazemi H, Tahmasebi P (2022) A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. Comput Geosci 164:105126. https://doi.org/10.1016/j.cageo.2022.105126

Nord JH, Koohang A, Paliszkiewicz J (2019) The Internet of Things: review and theoretical framework. Expert Syst Appl 133:97–108. https://doi.org/10.1016/j.eswa.2019.05.014

Okano MT (2017) IOT and Industry 4.0: the industrial new revolution. In: proceedings of the ICMIS, Istanbul, pp 75–82

Omar N, Zen H, Nicole N, Waluyo W (2020) Accuracy and reliability of data in IoT system for smart agriculture. Int J Integr Eng 12:105–116. https://doi.org/10.30880/ijie.2020.12.06.013

Palmaccio M, Dicuonzo G, Belyaeva ZS (2021) The Internet of Things and corporate business models: a systematic literature review. J Bus Res 131:610–618. https://doi.org/10.1016/j.jbusres.2020.09.069

Perone G (2022) Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. Eur J Health Econ 23:917–940. https://doi.org/10.1007/s10198-021-01347-4

Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45:211–218. https://doi.org/10.1145/505248.506010

Pivoto DG, de Almeida LF, Da Rosa RR, Rodrigues JJ, Lugli AB, Alberti AM (2021) Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: a literature review. J Manuf Syst 58:176–192. https://doi.org/10.1016/j.jmsy.2020.11.017

Powell W, Foth M, Cao S, Natanelov V (2022) Garbage in garbage out: the precarious link between IoT and blockchain in food supply chains. J Ind Inf Integr 25:100261. https://doi.org/10.1016/j.jii.2021.100261

Pugna A, Negrea R, Miclea S (2016) Using Six Sigma methodology to improve the assembly process in an automotive company. Procedia - Soc Behav Sci 221:308–316. https://doi.org/10.1016/j.sbspro.2016.05.120

Pukelsheim F (1994) The three sigma rule. Am Stat 48:88–91. https://doi.org/10.1080/00031305.1994.10476030

Rahim MA, Rahman MA, Rahman MM, Asyhari AT, Bhuiyan MZA, Ramasamy D (2021) Evolution of IoT-enabled connectivity and applications in automotive industry: a review. Veh Commun 27:100285. https://doi.org/10.1016/j.vehcom.2020.100285

Rahimi H, Zibaeenejad A, Safavi AA (2018) A novel IoT architecture based on 5G-IoT and next generation technologies. In: 9th annual information technology, electronics and mobile communication conference, Vancouver, pp 81–88. https://doi.org/10.1109/IEMCON.2018.8614777

Ray P, Rao YV (2019) A review of Industry 4.0 applications through SMART technologies by studying examples from the automobile industry. Adv Innov Res 16:80–89

Saravanamohan M, Aswini D, Thanish GS (2021) Role of IOT in the development of Industry 4.0 and robot technology – a state of the art. In: 2021 international conference on advancements in electrical, electronics, Communication, computing and automation. Coimbatore. https://doi.org/10.1109/ICAECA52838.2021.9675634

Scheider S, Lauf F, Möller F, Otto B (2023) A reference system architecture with data sovereignty for human-centric data ecosystems. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-023-00816-9

Shaub D (2020) Fast and accurate yearly time series forecasting with forecast combinations. Int J Forecast 36:116–120. https://doi.org/10.1016/j.ijforecast.2019.03.032

Sicari S, Rizzardi A, Miorandi D, Cappiello C, Coen-Porisini A (2016) A secure and quality-aware prototypical architecture for the Internet of Things. Inf Syst 58:43–55. https://doi.org/10.1016/j.is.2016.02.003

Sicari S, Rizzardi A, Cappiello C, Miorandi D, Coen-Porisini A (2018) Toward data governance in the Internet of Things. In: Yager RR, Espada Jordán P (eds) New advances in the Internet of Things. Springer, Cham, pp 59–74

Siddhartha B, Chavan AP, HD GK, Subramanya KN (2021) IoT enabled real-time availability and condition monitoring of CNC machines. In: 2020 IEEE international conference on internet of things and intelligence system, Bali, pp 78–84. https://doi.org/10.1109/IoTaIS50849.2021.9359698

Stefansky W (1972) Rejecting outliers in factorial designs. Technometrics 14:469–479. https://doi.org/10.2307/1267436

Steininger DM (2022) Interview with Frank Petry on "Digital entrepreneurship: opportunities, challenges, and impacts." Bus Inf Syst Eng 64:111–114. https://doi.org/10.1007/s12599-021-00738-4

Sun Y, Kamel MS, Wong AK, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 40:3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

Sun Z, Song Q, Zhu X, Sun H, Xu B, Zhou Y (2015) A novel ensemble method for classifying imbalanced data. Pattern Recognit 48:1623–1637. https://doi.org/10.1016/j.patcog.2014.11.014

Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72:37–45. https://doi.org/10.1080/00031305.2017.1380080

Teh HY, Kempa-Liehr AW, Wang KI-K (2020) Sensor data quality: a systematic review. J Big Data 7:1–49. https://doi.org/10.1186/s40537-020-0285-1

Thompson WR (1935) On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. Ann Math Stat 6:214–219. https://doi.org/10.1214/aoms/1177732567

Tissir S, Cherrafi A, Chiarini A, Elfezazi S, Bag S (2023) Lean Six Sigma and Industry 4.0 combination: scoping review and perspectives. Total Qual Manag Bus Excell 34(3–4):261–290. https://doi.org/10.1080/14783363.2022.2043740

Tkachenko R, Izonin I, Kryvinska N, Dronyuk I, Zub K (2020) An approach towards increasing prediction accuracy for the recovery of missing IoT data based on the GRNN-SGTM ensemble. Sensors 20:2625. https://doi.org/10.3390/s20092625

Urvoy M, Autrusseau F (2014) Application of Grubbs' Test for outliers to the detection of watermarks. In: Proceedings of the 2nd ACM workshop on Information hiding and multimedia security, pp 49–60. https://doi.org/10.1145/2600918.2600931

Valášek P, Müller M (2015) Properties of adhesives used for connecting in automotive industry. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, Salzburg 63:463–470. https://doi.org/10.11118/actaun201563020463

Valderas P, Torres V, Serral E (2023) Towards an interdisciplinary development of IoT-enhanced business processes. Bus Inf Syst Eng 65:25–48. https://doi.org/10.1007/s12599-022-00770-y

van der Togt R, Bakker PJM, Jaspers MWM (2011) A framework for performance and data quality assessment of radio frequency identification (RFID) systems in health care settings. J Biomed Inform 44:372–383. https://doi.org/10.1016/j.jbi.2010.12.004

Wang YR, Ziad M, Lee YW (2001) Data quality. Kluwer international series on advances in database systems. Kluwer Academic, Boston, p 23

Yang Y, Wang H, Jiang R, Guo X, Cheng J, Chen Y (2022) A review of IoT-enabled mobile healthcare: technologies, challenges, and future trends. IEEE Internet Things J 9:9478–9502. https://doi.org/10.1109/JIOT.2022.3144400

Yin L, Ge Y, Xiao K, Wang X, Quan X (2013) Feature selection for high-dimensional imbalanced data. Neurocomput 105:3–11. https://doi.org/10.1016/j.neucom.2012.04.039

Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems. Bus Inf Syst Eng 61:575–593. https://doi.org/10.1007/s12599-019-00608-0

Zhong RY, Xu X, Klotz E, Newman ST (2017) Intelligent manufacturing in the context of Industry 4.0: a review. Eng 3:616–630. https://doi.org/10.1016/J.ENG.2017.05.015