Improving data quality assessment of connected vehicles data with machine learning and statistical methods

Mulluken Wondie[1], Yang Li[1] and Julie Wall[1]
[1]School of Architecture, Computing and Engineering, University of East London, London, UK

Abstract – The connected vehicle is a fast-growing phenomenon enabling enterprises to generate new revenue streams, reduce costs and increase safety by utilizing the data collected. Quality data is a fundamental pre-requisite in the process of extracting the intended value. Therefore, developing data quality assessment methods is important. Classical data quality assessment methods are not good enough for connected vehicles data due to the complex nature those system, such as the spatio-temporal aspects. In this research, machine learning and statistical methods for data quality assessment are investigated. For this, two scenarios are selected. The first is to classify if a vehicle, which is not communicating (not sending data), is experiencing a real issue or simply properly parked and the power is off. If issues are detected earlier, measures can be taken to avoid delays or lose data, which can be formulated as timeliness and completeness of data quality dimensions. Using real life data, a new feature is constructed using DBSCAN and a logistic regression model is trained to identify real issues from false alarms, with a 0.76 F-score. The second scenario is to detect inaccurate fuel consumption (accuracy of data quality dimension). Using a public dataset, first a machine learning model is trained to predict fuel consumption. Then the difference between the actual and predicted value is calculated. A control chart is applied on the calculated difference and values which are out of the control are marked as inaccurate. This method can accurately detect 85% of inaccurate values correctly from the test dataset.

Keywords – Connected Vehicles, Data Quality, Machine learning, statistical methods, Data Quality dimensions, DBSCAN, Logistic Regression, Control Chart

1 INTRODUCTION

Connectivity is increasing around the world and its expansion to vehicles is no exception (Siegel, Erb and Sarma, 2017). The main principle of the connected vehicles (CV) ecosystem is wireless communication between a vehicle and another vehicle (V2V), a vehicle with infrastructure (V2I) or a vehicle to everything (V2X), which includes V2V, V2I and Vehicle to Person (V2P) (Mahmood, 2020). V2I refers to traffic signs or other fixed objects. The CV can, thus, be defined as the interconnectivity and collaboration of the various components within and outside of the vehicle including various devices , networks and infrastructure (Jadaan, Zeater and Abukhalil, 2017).

CVs create new data rich environments and are considered key enablers for many applications and services that will make roads safer, less congested, and more eco-friendly (Siegel, Erb and Sarma, 2017). This connectivity will be substantial to support different features and systems, such as adaptive routing, real-time navigation, and slow and near real-time infrastructure. Further examples include environmental sensing, advanced driver assistance systems, automated driving systems, mobility on demand, and mobility as a service (Jadaan, Zeater and Abukhalil, 2017), (Abdelkader, Elgazzar and Khamis, 2021).

Supported by faster development in different areas, such as connectivity, sensors and other computational capabilities, the CV is having an impact on society. If the data is garnered properly, its impact is wide in scope, including improving transport, helping tackle environmental change, economic development, as well as helping improve the design of the vehicle itself. However, there are challenges that have to be tackled for the full potential of CVs, such as dependable coverage, security threats, complex regulatory requirements, managing data, etc. (Siegel, Erb and Sarma, 2017).

This research aims to investigate one of those challenges i.e., data quality challenges, particularly data quality assessment, through the exploration of machine learning and statistical methods.

This research is organized as follows. In section 2, background information is presented and section 3 discusses the related literature. Section 4 presents the proposed solution and section 5 gives conclusions, some limitations and future work.

2 BACKGROUND

DAF Trucks N.V., one of the leading truck manufacturers, has implemented a CV system to collect information from trucks in real time for different purposes, such as predictive and preventive maintenance, economic activity analysis and warranty management, thereby increasing safety and decreasing costs. Good quality data is necessary to achieve all the objectives set. However, low data quality remains a challenge.
In this section, first a brief description of CVs is given. Next data quality issues are discussed and finally, impacts of data quality issues are discussed.

2.1 Connected Vehicles (CV)

Different phrases are used to refer to V2V such as Car-to-X (X-refers to everything), "Internet of Vehicles", Connected Vehicles, and "Talking Vehicles" (Mahmood, 2020). However, the CV is used by many practitioners and researchers.
The CV is made possible by various technologies supporting this paradigm, such as the Internet of Things (IoT) (Gubbi et al., 2013). IOT provides direct integration between the physical and digital worlds (Miraz et al., 2015), resulting in smart applications, such as smart transport. Generally, IoT exhibits a layered architecture, notably composed of acquisition layer (sensors), network or processing layer and application layers (Sethi and Sarangi, 2017). Another technology supporting CVs is cloud computing, for cheaper distributed processing (Qian et al., 2009). Dedicated Short-Range Communication (DSRC) protocols is another important enabling component to enable vehicles to communicate with each other and other road users or surrounding infrastructure directly, by providing high-speed, secure communication without involving cellular or other infrastructure. Artificial Intelligence (AI) also plays an important role by providing intelligent capability of systems (Jordan and Mitchell, 2015). Modern vehicles are also equipped with enabling technologies, such as a complex network of sensors forming a wide area network to collect thousands of signals within the vehicle and sensing the environment surrounding the vehicle (Siegel, Erb and Sarma, 2017).

2.2 Data Quality (DQ)

Data Quality (DQ) is a classical subject, with an extensive amount of research in the literature. DQ, also referred to as Information Quality (IQ), is considered as one of the success factor criteria for enterprises to thrive. Good data is a precondition for every action taken on the basis of data. For example, a machine learning solution used to discover new patterns and relationships from a given data set, essentially depends on good data quality, as bad data with too many outliers might transform patterns to be unrecognizable (Batini, Scannapieco and others, 2016).
No common DQ definition exists in the literature, rather, different definitions are used. These include "fitness for use" (Cai and Zhu, 2015), (Juran and Godfrey, 1999), "conformance to requirements" (Fox, Levitin and Redman, 1994), "degree of data fitness for a given purpose" (Gudivada, Apon and Ding, 2017), "how much data satisfies user expectations" (Sebastian-Coleman, 2012) and many more. From the literature, it is evident that there is agreement on the concept of "fitness-for-purpose", i.e. DQ depends on the consumer, the situation and the time (Cai and Zhu, 2015). Fitness-for-use, according to (Brimicombe, 2010), denotes the evaluated value of the outcomes of analysis applied for decision making. Fitness-for-use can be inferred from the reliability of the data if one topic is considered for decision making. On the other hand, if multiple themes and data sets are used, the outcome of the analysis can be evaluated from the blend of data reliability of each topic. Therefore, (Brimicombe, 2010) emphasizes that fitness-for-use is not always derived from analytical outcomes, rather depends on contexts.
From a general quality perspective, according to ISO 9000:2021, quality refers to the "degree to which a set of inherent characteristics fulfils a set of requirements" (International Organization for Standardization, 2021). In relation to this, it seems logical that many researchers agree on the notion which can be summarized as "DQ indicates to what extent or degree a specific data complies to the requirements of a certain use case". Even though DQ is commonly expressed as 'fitness-for-use' since the term was coined first (Juran and Godfrey, 1999), there is no exact or unique and universally accepted definition as DQ is a multifaceted subject. Therefore, the more practical and easiest way to describe DQ is by using DQ dimensions. This is evident as many studies use DQ dimensions to formulate DQ assessment methodologies (Karkouch et al., 2016). For example, (Gudivada, Apon and Ding, 2017) stress that the DQ definition "fit for business purpose" is broad and subjective and they suggest using DQ dimensions, such as accuracy, currency and

consistency for a concrete and pragmatic definition. The different DQ dimensions capture one or more facets or aspects of DQ, such as accuracy, completeness , timeliness and so on, which contribute towards the overall DQ measurement. The following section highlights the common DQ dimensions.

2.2.1 DQ Dimensions

There is no consensus on the list of DQ dimensions. According to (Cichy and Rass, 2019), there is a relatively high variation in data quality dimensions considered per framework (depending on the usage and subject). There exist many different definitions and prioritizations for the different DQ dimensions, such as those given in (Loshin, 2010), (Aziz, Saman and Jusoh, 2012), (Strong, Lee and Wang, 1997) and (Liaw et al., 2011) where conceptual DQ frameworks are identified that differ mostly in the types of dimensions and their categorization. Many of them are similar in their approach in that their definitions of the dimensions tend to be descriptive and subjective (Loshin, 2010), (Pipino, Lee and Wang, 2002), often using adjectives for which the semantics are overlapping or fuzzy. (Batini et al., 2015) provided a detailed study and literature review of the different data quality dimensions and variations. However, the most acknowledged conceptual DQ framework proposed by the research community is, probably the one from Wang and Strong (Strong, Lee and Wang, 1997). They defined information as a product (Wang, 1998) and as every product has a quality criteria so has information. They conducted a series of surveys and studies, which identified, sorted and ranked DQ aspects into 179 granular dimensions according to their importance to consumers. The output of their work is a hierarchical framework that organizes the identified DQ dimensions into four categories – intrinsic, contextual, representational and accessibility – and fifteen major DQ dimensions. DQ dimensions can correspond to the extension of data, which means that they refer to data values, or to the intension of data, i.e. to the metadata (Eppler, 2006).

In this work, we are only focusing on DQ dimensions that refer to the extension of data which (Eppler, 2006) states is most relevant to real-life applications. A short selection of the most widely used DQ dimensions with corresponding definitions is given in Table 2.1 below.

Table 2.1: Selected DQ dimensions

DQ Dimension	Definition
Accuracy	Sometimes referred to as correctness, it indicates the extent to which data correctly represents the true value; or the extent to which data is correct, reliable and certified (Cichy and Rass, 2019).
Completeness	The extent to which an entity has values for all expected data elements or the degree to which data fulfills the expected breadth, depth and scope for the purpose in context (Wang and Strong, 1996).
Consistency	The degree to which data is free from contradiction or the absence of difference, when comparing two or more representations of a thing against a definition (Group and others, 2013).
Timeliness	Also referred to as currency, it represents the extent to which data is up-to-date or the degree to which data represents reality from the required point in time (Group and others, 2013).
Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition (Group and others, 2013).

DQ is well studied and accuracy, completeness, consistency and timeliness are frequently discussed dimensions (Batini et al., 2009). Accuracy is the proximity of value x to value $x`$, which is the right representation of the real world i.e. x (Batini et al., 2009). (Strong, Lee and Wang, 1997) express accuracy in terms of "how correct, reliable and certified a value is". As (Batini et al., 2009) stated, data is considered to be accurate when its stored value matches with the real world.

Completeness is the degree to which data are good enough in terms of breadth, depth and scope for specific objective (Batini et al., 2009).

Validity is the extent to which data satisfies a given business rule, predefined standard or domain. Contrary to accuracy and completeness, it does not compare values to the real world. Timeliness is a measure of the time difference between refresh/update time and availability for usage of data. Data should be in the correct state at the time of usage, otherwise it is assumed to be outdated (Batini et al., 2009).

Dimensions can be rated as more or less important, depending on the data requirements and the domain. For example, one application could require the underlying data mainly to be correct while it would not require all data to be complete. Another application might strive for completeness while accepting a reduction in consistency (Group and others, 2013). Slightly different DQ dimensions are used in smart systems. For example; (Juddoo et al., 2018) states consistency, accuracy, completeness, and timeliness are major challenges. (Cai and Zhu, 2015) stresses completeness, accuracy and timeliness as major problems and (Olufowobi et al., 2016) adds provenance, which refers to the documentation of the

chronology (timeline) of data ownership and any data transformations or modifications applied to a data set. (Hazen et al., 2014) has elaborated on the DQ problems in CVs, especially on their spatio-temporal aspects.

As indicated in Section 2.1, smart connected systems are composed of multiple layers, such as acquisition using sensors, processing, and utilization (Perez-Castillo *et al.*, 2018), (Olufowobi et al., 2016), (Juddoo et al., 2018). This increases complexity such as scale of deployment, shortage of resources, network, sensor, situation, destruction, fail-dirty, security vulnerability, data stream processing and so on (Perez-Castillo *et al.*, 2018), in addition to generating big data, the various components are susceptible to malfunctions. Moreover, machine generated data lacks appropriate metadata and big data is known for its loose structure and missing values (Gudivada, Apon and Ding, 2017), (Juddoo et al., 2018). Errors such as offset, continuous varying or drifting, crashed or jammed, trimming error, outlier, noise and so on are common (Perez-Castillo *et al.*, 2018), (Megler, Tufte and Maier, 2016). Therefore, terms like "trustworthiness", "confidence", "credibility" etc., are also commonly used in DQ discussions of smart systems.

In summary, it seems that in IoT applications, which include CVs, the DQ dimensions remain more or less the same. However, some data elements such as timeliness, accuracy and completeness become more critical than others (Farooqi, Khattak and Imran, 2018).

2.3 Data Quality impact

In order to give proper emphasis to DQ, it is worth studying its impact on enterprises. A problem with DQ poses problems in organizations in many aspects. (Cichy and Rass, 2019) emphasize that since decision making highly depends on accessible data, good data quality plays a critical role for businesses to succeed. An inadequate level of DQ will have far-reaching impacts, such as poor decision-making, increased costs, poor operational performance, and nonconformity to regulations (Strong, Lee and Wang, 1997) , (Floridi, 2013).

The impact of DQ has been explored by many researchers. For example, (Spruit, Linden and others, 2019) have made an extensive study on enterprises and identified 11 DQ impacts. However, the classification given by (Loshin, 2011) provides a good summary by categorizing DQ impacts into 1. Financial, 2. Confidence and Satisfaction, 3. Productivity, and 4. Risk and Compliance.

To give some financial figures, for example, according to the Data Warehousing Institute (TDWI), the DQ issue costs US organizations 700 billion dollars yearly (Gudivada, Apon and Ding, 2017). A recent Gartner study states that poor DQ causes businesses to lose 15 million dollar on average annually (Spruit, Linden and others, 2019). In 2016, IBM research estimated that poor DQ costs U.S. companies more than 3 Trillion dollars per year (Spruit, Linden and others, 2019). (Redman, 2017) states that the cost of poor DQ approximately ranges from 15% to 25% of the revenue of most companies. It is evident from these figures that DQ is costing companies significantly in terms of money.

2.3.1 Data Quality impact in Connected systems

The DQ problem in IoT-based applications, including CVs, worsens as the nature and volume of data generated from connected objects is complex. (Keller et al., 2017) refers to telematics/CVs data as opportunity data. Data is generated massively but the underlying logic is not transparent, and the nature of error and statistical characteristics are not clear. Even though there is no study quantifying specifically the DQ impact on connected systems, it is obvious that the impact will be higher due to the complex nature of connected systems as many components are involved.

It is impossible to materialize the potential benefits of IoT if data is unreliable (Davenport and Redman, 2015). In the same study, it is described that DQ issues are higher in IoT-based smart systems, as these devices introduce more errors on top of the common human errors familiar in traditional systems. (Fekade et al., 2018) listed some of the causes that lead to poor DQ, including connection problems, interference from the environment, or sensor problems. This may lead to missing or inconsistent data, which causes financial loss, lack of customer satisfaction, loss of productivity as well as not complying with regulatory requirements.

Another very important characteristic of IoT connected systems is time. This is due to the fact that change in data fast and the freshness of data is very short these days, which in turn suitable technology. (Cai and Zhu, 2015). In this regard, a preventative alert system implemented based on a streaming CV will not be able to serve the intended purpose if data is delayed or missing, which may lead to increased costs or even to loss of life

3 RELATED WORKS

3.1 DQ Assessment and Improvement

The objective of DQ Assessment is to identify erroneous data elements and estimate their impact on various data-driven business processes so that a mitigating strategy can be devised. Therefore, before using data for any application, it should be assessed for fitness-of-use. This is because, as stated in (Brimicombe, 2010), it is highly unlikely to achieve 100% data accuracy as errors and uncertainty are inevitable, and further deterioration can even be caused when two or more data sets are combined, which will have an impact on the quality of the output from the data. Therefore, it is important to know the level of DQ and means of mitigating DQ issues. For this, a DQ assessment strategy should be developed.

For a long time, DQ has been considered as a multidimensional concept and as a result its assessment is viewed as a complex process with multifaceted challenges (Cichy and Rass, 2019). Besides, DQ is understood as a multidisciplinary problem spanning subjects such as computing, quality control, human factors and statistics (Cichy and Rass, 2019). Therefore, many researchers approached DQ assessment from this multi-faceted perspective. For example, (Eppler, 2006) states that the majority of DQ assessment frameworks are built for a particular domain and only a few are general enough to be applied to any domain.

Usually, the assessment process is based on DQ dimensions, tools to describe the concept of DQ and researchers suggest to use DQ dimensions in DQ assessment endeavour (Juddoo et al., 2018). DQ dimensions are features of data that may provide the overall fitness level if measured properly and it is recommended to begin the DQ assessment effort by clearly listing relevant DQ dimensions (Cichy and Rass, 2019). (Cichy and Rass, 2019) also stress that DQ assessment varies based on different factors, such as the context, nature and type of data, DQ dimensions that are deemed to be relevant for the chosen objective. To this end, there has been much research in DQ, which is evolving to respond to changes such as various data types and innovations (Karkouch et al., 2016), shifting from monolithic to networked systems (Batini et al., 2009) as these changes introduce complexities.

3.2 Machine Learning Approaches

The best way to prove that a piece of information is correct is to compare with some trusted source, that is a source which is consistently correct. Such a source may not exist or at least may not be readily available (Maydanchik, 2007). To illustrate this with an example, assume that an organization needs to verify the ages of its 4,000 staff members. This could be done by checking necessary documents showing birthdates. But if such a document does not exist, it may be necessary to call and ask each and every individual, which is difficult and costly to apply on a large scale. This issue is compounded in the IoT era where a big volume of data is generated, multiple layers are involved and the data is of various types, including timestamps, device data, location data and others (Kim et al., 2019). Therefore, traditional DQ assessment methodologies may not work as a proper DQ assessment method for smart connected systems. Realizing this fact, some researchers have employed machine learning methodologies for DQ assessment. In this section, a review of DQ assessment methodologies that applied machine learning is presented.

Most of the machine learning methods for DQ assessment focused on sensors and anomaly detection, which is equivalent to the validity DQ dimension (Vasta et al., 2017), (Barnes and Hu, 2013), (Diop et al., 2017). For example, (Wang et al., 2017) explored deep learning to detect sensor drift, which helped to automatically calibrate sensors and thus improving data quality. One comprehensive work that covers multiple DQ issues by combining traditional methods together with advanced machine learning methods is found in the works of Shrivastava et al., (2019). In this work, a data quality advisor is developed that helps with the assessment and improvement of DQ. The DQ advisor is supported by a visual inspection functionality, so that experts can have a judgement on the proposed DQ improvement by the DQ advisor. Different modules are included in the framework. The general validator module ranging from simple null, uniqueness and duplicate checks to statistical functions, such as correlation and summary statistics enable to assess relatively easier DQ issues. The AI and time series modules help to identify relatively complex DQ issues such as anomaly detection.

Another work, which aimed to improve Intelligent Transport System (ITS) data, employs unsupervised machine learning techniques (Megler, Tufte and Maier, 2016). Intended for outlier detection, the method adopted a framework consisting of k-means

clustering at the core. Using this method, each observation is assigned to a certain cluster and distance is calculated to the cluster. This framework helps to detect two possible anomalies, i.e, anomalous records within each cluster and an anomaly from the whole cluster.

Advanced methods such as deep learning is investigated in the works of (Dai, Yoshigoe and Parsley, 2018). The framework developed combines deep learning with statistical quality control to detect outliers that can be removed from the data set, to improve big data DQ. First a deep learning model is trained and using the trained model, prediction is made on the data. Then the statistical method is used to identify suspicious values. This helps to visually detect outliers. Even though this method was not applied on connected data directly, it is an important step to show the application of machine learning and advanced statistical methods for big data DQ assessment.

Random forest regression is also used to assess the accuracy of IoT data (Farooqi, Khattak and Imran, 2018). In this work, historical weather data is used to train the random forest regression model. Using the results obtained from this model, rules to detect inaccurate data are established. The proposed assessment approach for operationalization is that after predicting the new value, if it does not comply with the rules established, then the data is considered as inaccurate and the recommendation is to remove this specific data point.

From the literature, it is evident that various machine learning methods are explored for the different DQ issues in smart connected systems. While there are efforts to incorporate advanced techniques such as machine learning to DQ assessment of connected products, it is still not enough and most of the existing works do not have a holistic approach. Moreover, most of the studies have focused on outlier detection, which is equivalent to the validity DQ dimension. Many DQ dimensions are not studied enough. Specifically, timeliness and accuracy are not tackled with enough depth and breadth.

4 PROPOSED SOLUTION

In this research, machine learning and statistical methods have been proposed to enhance CVs DQ assessment. This approach is proposed to replace classical DQ assessment methods, rather to fill the gaps that classical DQ assessment methods cannot fulfil and provide an improved DQ assessment method. The following two scenarios, mapping to three DQ dimensions, are selected to demonstrate DQ assessment.

1. Detecting missing data or delays (Completeness/Timeliness)

   In CVs, space and time are important elements. Connectivity may be affected by different factors such as telecommunication, network issues, cabling issues, an accident and so on. In such a situation, the vehicle may not be sending data. This will introduce a delay or missing data.

2. Detect inaccurate values (Accuracy)

   CVs are composed of different components. Inaccuracies can be caused by any of these, such as sensor drift, processing errors and so on.

4.1 Data set

The datasets used in model training and testing for the CVs DQ assessment in this research was originally real-life data from DAF Trucks N.V. connected system. This real data includes a wide range of vehicle data collected from Controller Area Network (CAN) bus using the onboarding connected device including vehicle health data (fault codes, dash-lights), driver behaviour (such as braking), economic data (such as fuel), environment data (such as ambient temperature and road type), GPS information and so on. To enable reproducibility and privacy, a public dataset of anonymized data is also used.

1. Anonymized data

The first dataset is anonymized real life telematics data. Anonymization is applied to avoid privacy concerns. This data set includes selected trip level telematics data elements and the following approach is followed:

- Real life data is collected containing the required features.
- Privacy sensitive data is identified. In this step, Vehicleid together with GPS location (latitude and longitude) are identified as privacy sensitive data.
- Replace sensitive data with anonymized data. To avoid privacy concerns, the vehicleid is replaced with sequence numbers.

2. Public data

The second dataset is a publicly available dataset of public transportation buses data collected using sensors about fuel consumption and contextual condition, (Rosameo, 20221)

4.2 Selected scenarios

For each selected scenario, machine learning models are trained following machine learning processes including data pre-processing, feature selection, modelling and evaluation. In this section, the proposed solution for each scenario is presented.

1. Detecting missing data or delay (Completeness/Timeliness)

For a CV to garner all the benefits, reliable data flow is important. But there are situations where information is delayed or missing. A vehicle may not be sending information for two reasons:

1. Because it is properly parked and the main switch is off, which is a normal behaviour, or
2. Due to technical reasons, which causes delays or missing data

If a vehicle is not sending data due to technical reasons, one of the following will happen as a consequence:

1. Information/data will be received delayed if/when the issue is fixed, or
2. Information/data will be wiped out and lost

Both have a negative impact on the business operation by affecting the quality of the data. As such, the problem can be formulated as a DQ problem, specifically, the timeliness and completeness DQ dimensions, amongst others (Juddoo and George, 2018).

If the issue is detected in time, different measures can be taken, including rebooting the system remotely or manually, or even replacing the system if other measures are not successful. However, acting on all vehicles that are not sending data is costly or even impossible. Therefore, it is important to identify the real problematic vehicles from parked vehicles.

In order to monitor the DQ problem, a DQ assessment report can be developed (Gitzel, Turring and Maczey, 2015). However, classical DQ assessment methods are not good enough for CVs, where the spatio-temporal aspect is important. In this research, machine learning methods are investigated to identify normally parked vehicles from vehicles not sending data due to real technical issues. This method consists of two machine learning components.

- An unsupervised machine learning method is used to create a new feature. parking location
- A supervised learning method, specifically binary classification, is applied using the newly constructed feature together with other features.

The data used in this section is the anonymized real-life data containing the data elements described in Table 4.1 below.

Table 4.1: Dataset description

Field name	Example value	Description
Vehicleid	1	Vehicle identifier
EVENTID	4	Reason why the data was generated
EVTDATETIME	44237.57955	The datetime when the data was generated
GPSLATITUDE	47.75903702	The GPS latitude when the data was generated
GPSLONGITUDE	9.889554024	The GPS longitude when the data was generated
VDIST	148812965	The cumulative vehicle mileage when the data was collected
RECEIVEDDATETIME	44237.60455	The datetime when the generated data was received in the back end
COUNTRY	Germany	Country where the vehicle was driving when data was collected

In addition to these, a number of features are derived and given in Table 4.2 below.

Table 4.2: Derived data elements for dataset 1

Field name	Calculation
Distancedone	VDIST at end of trip minus VDIST at start of trip
Missingdistance	The gap in mileage of the previous trip and next trip
Time_diff_prev_trip	The time gap in between previous trip and next trip
Delay_hrs	The delay in hours from the time the trip was made (end of trip) and the information is received to the back end
Nocomm [Yes/No]	If a trip has a delay of 13 hours or more or it has a missing trip, then it is marked as non communicating
Number_of_missing	How many times has the vehicle showed a missing trip in the past
Nr_buffer	How many times has the vehicle been delayed in the past

In this dataset, **nocomm [Yes/No]** is the target variable for the classification algorithm.

i. Unsupervised learning

During data exploration, it becomes apparent that the places where the vehicles are parked has an influence. To investigate this, dealer locations were incorporated. Most of the vehicles stopped at dealer locations were normally parked and came back with no issue. However, dealer locations are only a few of the places where vehicles normally park. Vehicles park at more places such as on customer bases (customer locations) and other parking locations. But these locations are not available. Since this feature is necessary, a clustering algorithm is employed to construct parking locations from the dataset. The process is described as follows.
From the dataset, using the GPS, the positions where vehicles are parking can be determined by comparing the GPS point where it stopped and where it started its next trip (duration between time t2 and t3) as depicted in Figure 4.1 below.

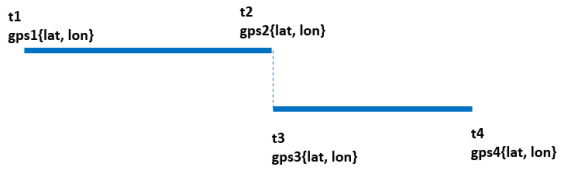

Figure 4.1: Trip Sequence

When these positions are the same and the time duration is long enough (8 hours is used to exclude rest periods), it is considered that the vehicle has parked at this position. To identify these locations, the GPS points from the data set are extracted and a density-based spatial clustering of applications with noise (DBSCAN) is trained. DBSCAN is one of the most widely used clustering algorithms, a density-based nonparametric algorithm that forms sub-groups of closely located points together and marks the less densely points as outliers. DBSCAN is selected for the following reasons (Ester *et al.*, 1996), (Khan *et al.*, 2014).

1. It performs well for oddly shaped data such as clusters with irregular shape
2. There is no need to specify the number of clusters apriori
3. Works well for data of different distributions,i.e., it can discover different shapes and sizes

The DBSCAN algorithm uses two parameters (Ester et al., 1996):

- minPts: The smallest number of items to form a group to qualify to be dense enough.
- eps (ε): specifies how close items should be to each other to be considered a part of a group

In this process, the centroids of each cluster are taken as a parking location. This set is stored to construct a new feature for supervised learning.
The following steps are performed to determine parking locations.

1. GPS locations where trucks stopped for 8 hours or more in the same location in the past were collected as shown in Figure 4.1.
2. DBSCAN is trained on the anonymized dataset with (eps = 0.4, minPts = 10), which is an assumption of parking locations with ~400m radius and either 10 vehicles have been parked in the past or one vehicle was parked 10 times in the past on the same location. To determine these parameters, previous domain knowledge was used and DBSCAN was trained.
3. Centroids of each cluster are retrieved.
4. The centroid points as centres of parking locations are stored

Fig 4.2 shows the result of the DBSCAN. 13,000 parking locations are identified and stored using this approach. Fig 4.3 shows one example of an identified parking location.

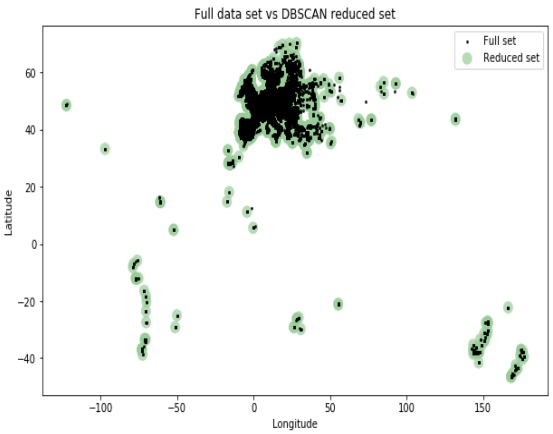

Figure 4.2: DBSCAN result plot for parking points

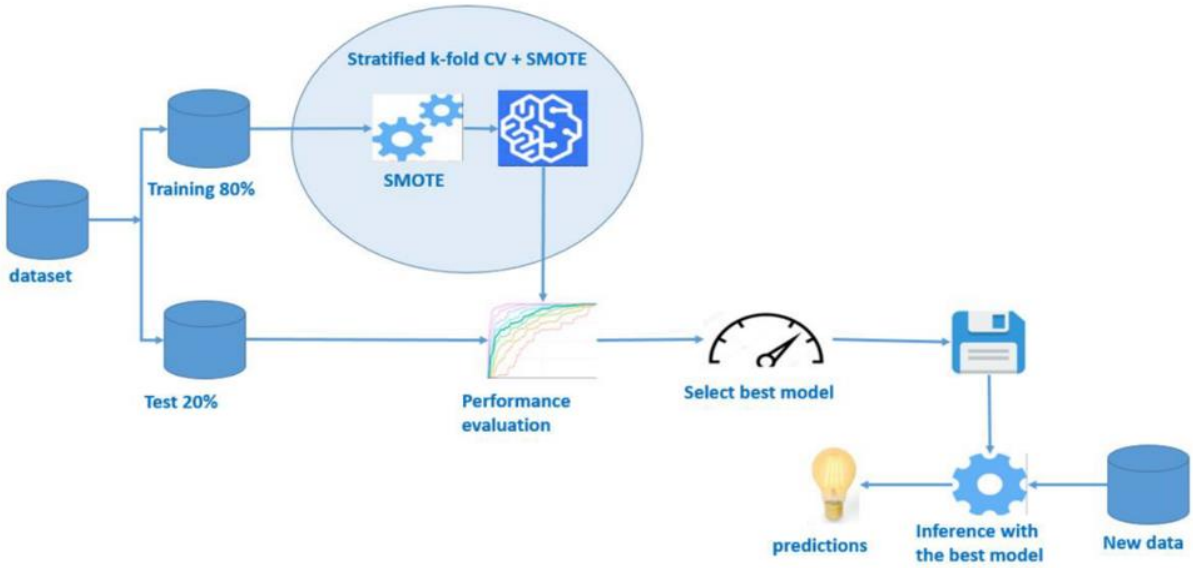

Figure 4.4: Architecture of supervised learning

Figure 4.3: Example of identified parking locations

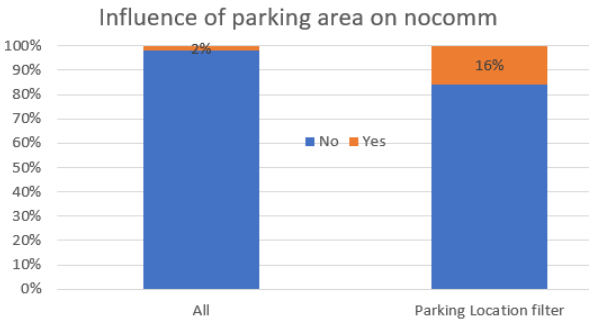

Figure 4.5: Influence of parking location on non-communication

ii. Supervised Learning

To determine if a vehicle is communicating or not, a supervised machine learning algorithm is trained using the architecture given in Figure 4.4. First the dataset is enriched with the distance to the nearest parking location when the CV last communicated (as explained in Section 4.2). Without this feature, the distribution of nocomm is 84% (No) to 16% (Yes). Adding this feature and filtering <=400m (which means near one of the identified parking locations), the distribution becomes 98% (No) to 2% (Yes), which means only 2% of vehicles whose last position was in the identified parking location are actually non-communicating.

Another important data element that has an influence, is the event observed just before the CV stopped sending data, as shown in Figure 4.6, where it the nocomm proportion is higher.

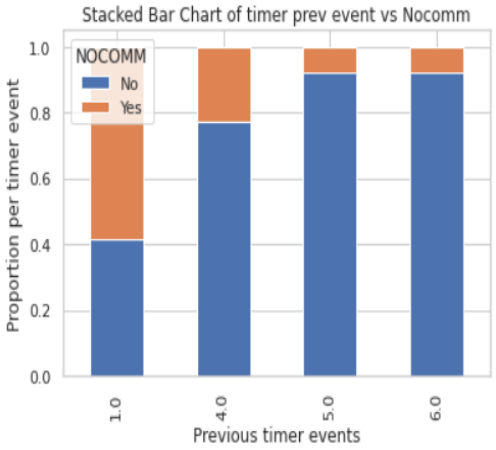

Figure 4.6: Influence of event on non-communication

The dataset was split, 80% for training and 20% for testing. Then a logistic regression model with stratified k-fold (k=10) cross validation (Delen, Walker and Kadam, 2005) is trained. Logistic regression is chosen for interpretability. In the data exploration phase, it became clear that the data is highly imbalanced (see Fig 4.5). Different methods have been investigated to handle the class imbalance and the synthetic minority oversampling technique (SMOTE) (Kotsiantis *et al.*, 2006) resulted in the best outcome. Table 4.3 gives the different configurations and the corresponding outcome of the model training and testing. Due to the class imbalance, the F-Measure is used as the performance evaluation (He and Garcia, 2009).
The initial model resulted in 0.61. Then SMOTE was introduced and the F-score went up to 0.73. Finally, grid search was introduced to choose an appropriate weight parameter (Li, 2013) and the F-Measure went up to 0.76.

Table 4.3: Selected results

Configuration	**F-score**
Logistic regression	0.61
Logistic regression + SMOTE	0.73
Logistic regression + SMOTE + Grid Search	0.76

Finally, the best model is stored and used to predict new non-communicating vehicles. The output gives the *vehicleid* together with the probability that the vehicle has a problem (missing or delayed) when it communicates back, see Table 4.4. The output is used to filter vehicles that are highly likely to be facing an issue so that action can be taken to avoid further delays or missing data.

Vehicleid	**predicted (probability)**
9748	0.30852841
49749	0.962966256
58830	0.975912128
59205	0.972544542
65956	0.990541118
67443	0.275451636
69754	0.275451493

Table 4.4: Sample prediction output

2. Detect inaccurate values

CV data is collected using multiple sensors and passes through different points. Due to sensor issues or other problems, inaccurate values may be reported for end users. While validity can be easily assessed by checking the values against the accepted range of values, accuracy is difficult to assess using classical DQ assessment methods. Therefore, in this section, machine learning and statistical methods are investigated to assess if a specific data is accurate or not. For the experiment, the publicly available dataset is used and the fuel consumption data element is selected as a target for the regression task.
The proposed solution consists of two main modules.

1. Machine learning module, used to predict the value of the fuel consumption
2. Statistical quality module, used to apply a quality control on top of the difference between the actual value and the predicted value from the machine learning module.

The architecture of the proposed solution is given in Figure 4.7 below. The dataset used for this scenario is the public dataset described in Section 4.1, containing the data elements in Table 4.5.

Table 4.5: Data elements of the public dataset

Field name	**Example value**	**Description**
Date-time	43480.25762	Trip datetime
VehicleID	0	Identifier of the vehicle
avg_slope	0.009036145	Average slope of the path
mass	19.614	Mass in ton of the vehicle including passengers
aircond_ptime	0	Percentage of travel time with air conditioning on
stop_ptime	0.12244898	Percentage of the travel time with the vehicle stopped and with the engine on
brake_usage	0.367346939	Percentage of the travel time with the brake and with the engine on
Accel	0.617674419	Percentage of the travel time with thge accelerator pedal pressed
fuel_per_km	0.75	Fuel used in the trip

Using these features and fuel consumption, the following five machine learning models were investigated and evaluated:

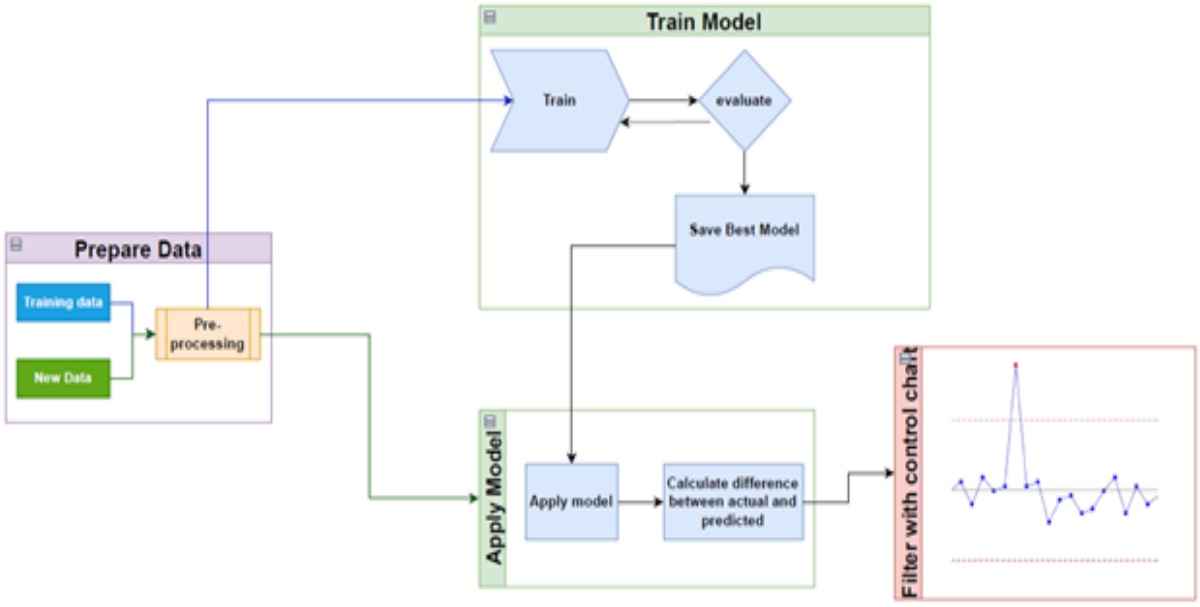

Figure 4.7: Architecture of ml and statistical method for accuracy detection

- Linear Regression (LR)
- RandomForest Regressor (RF),
- GradientBoosting Regressor (GBR)
- Elastic Net
- Lasso Regressor

GBR is selected as it resulted in the best result as given in Table 4.6.

Table 4.6: Results of different algorithms

Model	**MAE**	**MSE**	**RMSE**	**R2**	**RMSLE**	**MAPE**	**TT (Sec)**
RF	0.0602	0.0060	0.0778	0.6785	0.0521	0.1338	3.0290
LR	0.1104	0.0186	0.1364	0.0107	0.0931	0.2817	0.2340
GBR	0.0674	0.0072	0.0849	0.6171	0.0570	0.1524	1.2740
Eastic Net	0.1112	0.0188	0.1372	-0.0002	0.0936	0.2838	0.0210
Lasso Regressor	0.1112	0.0188	0.1372	-0.0002	0.0936	0.2838	0.0210

3. Statistical module

This is motivated by the statistical control process, the basis of which is the central limit theorem (Benneyan, 1998). Process control employs a control chart which is used to depict process change over time. A control chart always consists of (Nelson, 1985):-

- a central line for the average (CL),
- an upper line for the upper control limit (UCL),

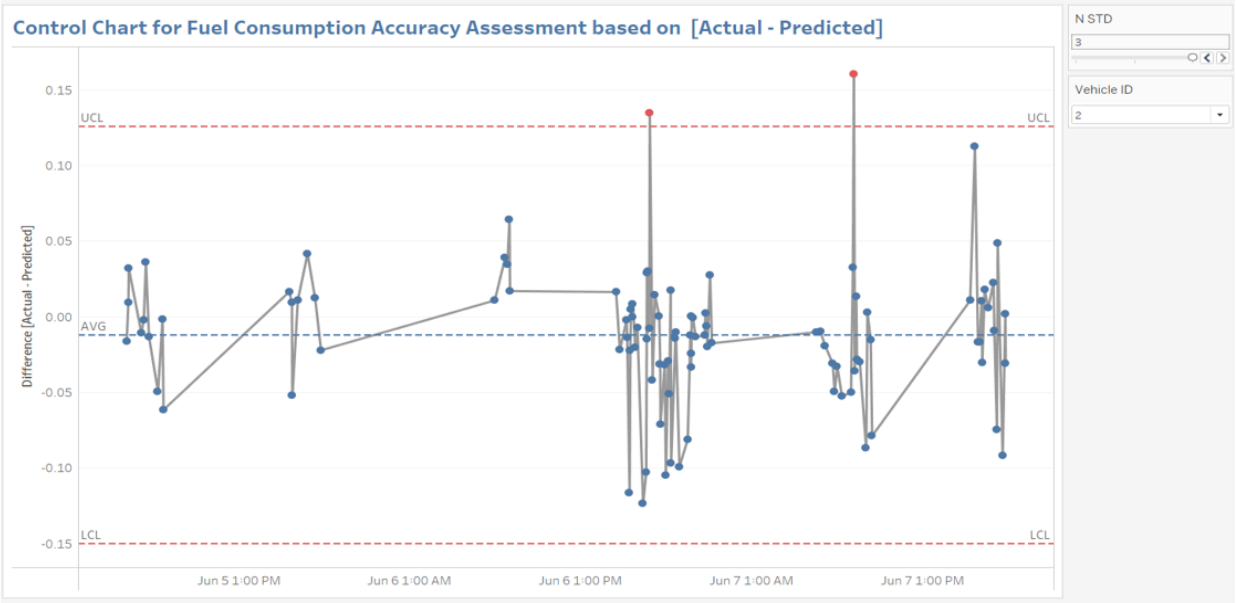

Figure 4.9: Control chart showing to detect inaccurate values

- a lower line for the lower control limit (LCL).

These lines are determined from historical data.

A point outside of the LCL and UCL is considered to be out of the control signal, as shown in Figure 4.8.

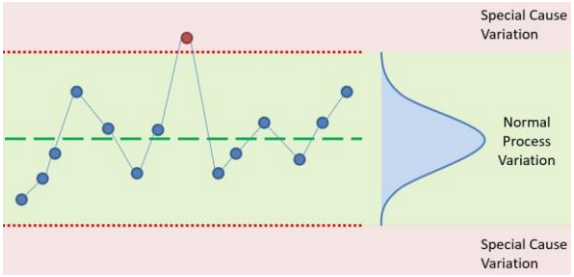

Figure 4.8: Control Chart

In this research, this method is used to determine if the difference between the predicted value and the actual value is within normal variation or not.

A. Evaluation

To evaluate the proposed solution, 10% of the dataset is kept aside. The dataset has 6,660 samples in total. Then for 10% this set (which is 666), the actual value is replaced with systematically calculated inaccurate values. The inaccurate values are introduced as follows. As mentioned earlier, the experiment is performed using real life data. The outcome shows that the inaccurate values were from 10% to 80%. Therefore, new inaccurate values are introduced with the same proportion for the 10% of the data kept. Accordingly, eight different error rates are used with equal proportion. For the first 12.5% of 666, i.e., 83, the actual value is replaced with a value of 10% higher, for the next 12.5%, the actual values is replaced with 20% higher and so on.

Then the trained model is applied to find the predicted values. Next, the difference to the corresponding actual values (i.e. the systematically introduced error values) is calculated. Lastly, UCL and LCL values are calculated for each record using the calculated difference, and values beyond the UCL are identified as inaccurate. To visualize this, a control chart is developed with an option to filter for a specific vehicle and required standard deviation (sigma). Table 4.7 shows the introduced inaccurate values for a single vehicle and Figure 4.9 gives the corresponding control chart detecting two of the incorrect values accurately when 3 sigma is selected.

Table 4.7: Example of error values for validation

Date-time	fuel_per_km	predicted_fuel_per_km	actual_value
6/7/2019 8:09	0.667651403	0.506997609	0.517651
6/7/2019 16:39	0.748088235	0.635558882	0.658088235
6/6/2019 17:49	0.603460722	0.468844247	0.473460722

The outcome shows that this method has detected 85% of the inaccurate values correctly. The result is visualized in Figure 4.9.

5 DISCUSSION AND CONCLUSIONS

This study aims to provide an improved DQ assessment for CVs and demonstrates that machine learning and statistical methods can augment classical DQ assessment methods. This has been demonstrated using two selected scenarios. The first scenario tries to see if a CV not sending data is facing a real issue or it is parked, and the power is intentionally switched off. If it is due to problems of software or connectivity, there will be either a delay in data, some missing information or both. The problem is formulated as a DQ problem by mapping the delay to timeliness and missing to completeness DQ dimensions. For this scenario, anonymized real life CV data is used. The available relevant features are considered and several derived features (see table 4.2) are constructed. One of the contributions of this study is the demonstration of how important feature engineering is to machine learning. To this end, DBSCAN was used to locate parking locations. Using this this real-life dataset, a new feature was constructed to determine if the vehicle was at a proper parking location or not during its last communication. This feature proved to be a strong predictor for the supervised learning approach.

On this real-life dataset, a logistic regression model was trained to classify non-communicating vehicles as those having a real problem or those which are normally parked. Imbalanced data treatment, specifically SMOTE, and stratified k-fold cross validation was applied. Further optimization was performed using Grid Search. Since the dataset is imbalanced, the F-score wass used as a performance measure, with final model results of an F-score of 0.76.

This solution has been deployed in production and helps to prematurely detect vehicles that are having issues with a corresponding likelihood so that preventive action can be taken to solve the problem. there by avoiding further delays and missing data.

The second scenario demonstrates a combined application of machine learning and statistical quality control to detect inaccurate values. For this experiment a public dataset was collected, based on the fuel consumption of transport buses. The main goal here is to determine if the actual fuel consumption reported is accurate or not. To do that, first a machine learning model was trained to predict the fuel consumption from historical values. Then the difference between the actual values and the predicted values is calculated. Using this difference, a statistical control chart was depicted and values beyond the upper control limit are marked as inaccurate values. To evaluate this approach, systematic erroneous values were introduced and the method accurately detected 85% of the erroneous values accurately.

Both scenarios are difficult to assess using traditional DQ assessment methods and advanced methods like machine learning and statistical methods can help to achieve improved DQ assessment.

Limitation and future studies

One of the limitations of this study was finding relevant public data. Since the investigated topics require ground truth verifications, it is important to have real and verifiable data. However, due to privacy and other factors, it becomes challenging to find such a dataset. One option to mitigate this challenge is to generate synthetic data. Therefore, a future study can be done to generate representative data synthetically, removing the privacy concern using machine learning techniques. Besides, the publicly available dataset used for this experiment does not contain all relevant features. Therefore, more features can be added and investigated to improve prediction models.

Future research will also investigate other more advanced algorithms to achieve a better prediction outcome, such as deep learning.

REFERENCES

Abdelkader, G., Elgazzar, K. and Khamis, A. (2021) 'Connected vehicles: technology review, state of the art, challenges and opportunities', *Sensors*. Multidisciplinary Digital Publishing Institute, 21(22), p. 7712.

Aziz, A. A., Saman, M. Y. M. and Jusoh, J. A. (2012) 'Data investigation: Issues of data quality and implementing base analysis technique to evaluate quality of data in heterogeneous databases', *Journal of Theoretical and Applied Information Technology*, 45(1), pp. 360–372. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84874528600&partnerID=40&md5=c0a93b1f761e14519748e0717d2f0282.

Barnes, B. B. and Hu, C. (2013) 'A Hybrid Cloud Detection Algorithm to Improve MODIS Sea Surface Temperature Data Quality and Coverage Over the Eastern Gulf of Mexico', *IEEE Transactions on Geoscience and Remote Sensing*, 51(6), pp. 3273–3285. doi: 10.1109/TGRS.2012.2223217.

Batini, C. *et al.* (2009) 'Methodologies for data quality assessment and improvement', *ACM computing surveys (CSUR)*. ACM, 41(3), p. 16.

Batini, C. *et al.* (2015) 'From data quality to big data quality', *Journal of Database Management*, 26(1), pp. 60–82. doi: 10.4018/JDM.2015010103.

Batini, C., Scannapieco, M. and others (2016) 'Data and information quality', *Cham, Switzerland: Springer International Publishing*. Springer.

Benneyan, J. C. (1998) 'Use and interpretation of statistical quality control charts', *International Journal for Quality in Health Care*. JSTOR, 10(1), pp. 69–73.

Brimicombe, A. (2010) 'GIS, environmental modeling and engineering'. CRC Press/Taylor & Francis Group.

Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal*. Ubiquity Press, 14.

Cichy, C. and Rass, S. (2019) 'An overview of data quality frameworks', *IEEE Access*. IEEE, 7, pp. 24634–24648.

Dai, W., Yoshigoe, K. and Parsley, W. (2018) 'Improving Data Quality Through Deep Learning and Statistical Models', in *Information Technology-New Generations*. Springer, pp. 515–522.

Davenport, T. and Redman, T. (2015) 'Build data quality into the internet of things', *Wall Str. J.*

Delen, D., Walker, G. and Kadam, A. (2005) 'Predicting breast cancer survivability: a comparison of three data mining methods', *Artificial intelligence in medicine*. Elsevier, 34(2), pp. 113–127.

Diop, M. *et al.* (2017) 'A methodology for prior management of temporal data quality in a data mining process', in *2017 Intelligent Systems and Computer Vision (ISCV)*, pp. 1–8. doi: 10.1109/ISACV.2017.8054906.

Eppler, M. J. (2006) *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science \& Business Media.

Ester, M. *et al.* (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise.', in *kdd*, pp. 226–231.

Farooqi, M. M., Khattak, H. A. and Imran, M. (2018) 'Data quality techniques in the internet of things: Random forest regression', in *2018 14th International Conference on Emerging Technologies (ICET)*, pp. 1–4.

Fekade, B. *et al.* (2018) 'Probabilistic recovery of incomplete sensed data in IoT', *IEEE Internet of Things Journal*. IEEE, 5(4), pp. 2282–2292.

Floridi, L. (2013) 'Information quality', *Philosophy & Technology*. Springer, 26(1), pp. 1–6.

for Standardization, I. O. (2021) *Quality Management Systems--Fundamentals and Vocabulary*. International Organization for Standardization.

Fox, C., Levitin, A. and Redman, T. (1994) 'The notion of data and its quality dimensions', *Information processing & management*. Elsevier, 30(1), pp. 9–19.

Group, D. Q. D. W. and others (2013) 'The Six Dimensions of EHDI Data Quality Assessment'. DAMA UK.

Gubbi, J. *et al.* (2013) 'Internet of Things (IoT): A vision, architectural elements, and future directions', *Future generation computer systems*. Elsevier, 29(7), pp. 1645–1660.

Gudivada, V., Apon, A. and Ding, J. (2017) 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations', *International Journal on Advances in Software*, 10(1), pp. 1–20.

Hazen, B. T. *et al.* (2014) 'Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications', *International Journal of Production Economics*. Elsevier, 154, pp. 72–80.

He, H. and Garcia, E. A. (2009) 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering*. Ieee, 21(9), pp. 1263–1284.

Jadaan, K., Zeater, S. and Abukhalil, Y. (2017) 'Connected vehicles: an innovative transport technology', *Procedia Engineering*. Elsevier, 187, pp. 641–648.

Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*. American Association for the Advancement of Science, 349(6245), pp. 255–260.

Juddoo, S. *et al.* (2018) 'Data governance in the health industry: investigating data quality dimensions within a big data context', *Applied System Innovation*. Multidisciplinary Digital Publishing Institute, 1(4), p. 43.

Juran, J. and Godfrey, A. B. (1999) 'Quality handbook', *Republished McGraw-Hill*, 173(8), pp. 34–51.

Karkouch, A. *et al.* (2016) 'Data quality in internet of things: A state-of-the-art survey', *Journal of Network and Computer Applications*. Elsevier, 73, pp. 57–81.

Keller, S. *et al.* (2017) 'The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches', *Annual Review of Statistics and Its Application*. Annual Reviews, 4, pp. 85–108.

Khan, K. *et al.* (2014) 'DBSCAN: Past, present and future', in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238.

Kim, S. *et al.* (2019) 'Extending data quality management for smart connected product operations', *IEEE Access*. IEEE, 7, pp. 144663–144678.

Kotsiantis, S. *et al.* (2006) 'Handling imbalanced datasets: A review', *GESTS international transactions on computer science and engineering*, 30(1), pp. 25–36.

Li, J. (2013) 'Logistic regression', *Course Notes. URL http://sites. stat. psu. edu/\~{} jiali/course/stat597e/notes2/logit. pdf*.

Liaw, S.-T. *et al.* (2011) 'Data quality and fitness for purpose of routinely collected data--a general practice case study from an electronic Practice-Based Research Network (ePBRN)', in *AMIA Annual Symposium Proceedings*, p. 785.

Loshin, D. (2010) *The practitioner's guide to data quality improvement*. Elsevier.

Loshin, D. (2011) 'Evaluating the business impacts of poor data quality', *Information Quality Journal*.

Mahmood, Z. (2020) 'Connected vehicles in the IoV: Concepts, technologies and architectures', in *Connected vehicles in the internet of things*. Springer, pp. 3–18.

Maydanchik, A. (2007) *Data quality assessment*.

Technics publications.

Megler, V. M., Tufte, K. and Maier, D. (2016) 'Improving data quality in intelligent transportation systems', *arXiv preprint arXiv:1602.03100*.

Miraz, M. H. *et al.* (2015) 'A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)', in *2015 Internet Technologies and Applications (ITA)*, pp. 219–224.

Nelson, L. S. (1985) 'Interpreting Shewhart X control charts', *Journal of Quality Technology*. Taylor \& Francis, 17(2), pp. 114–116.

Olufowobi, H. *et al.* (2016) 'Data provenance model for Internet of Things (IoT) systems', in *International Conference on Service-Oriented Computing*, pp. 85–91.

Perez-Castillo, R. *et al.* (2018) 'Data Quality Best Practices in IoT Environments', in *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pp. 272–275.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data quality assessment', *Communications of the ACM*. ACM, 45(4), pp. 211–218.

Qian, L. *et al.* (2009) 'Cloud computing: An overview', in *IEEE international conference on cloud computing*, pp. 626–631.

Redman, T. C. (2017) 'Seizing opportunity in data quality', *MIT Sloan Management Review. https://sloanreview. mit. edu/article/seizing-opportunity-in-data-quality. November*, 29.

Sebastian-Coleman, L. (2012) *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes.

Sethi, P. and Sarangi, S. R. (2017) 'Internet of things: architectures, protocols, and applications', *Journal of Electrical and Computer Engineering*. Hindawi, 2017.

Siegel, J. E., Erb, D. C. and Sarma, S. E. (2017) 'A survey of the connected vehicle landscape—Architectures, enabling technologies, applications, and development areas', *IEEE Transactions on Intelligent Transportation Systems*. IEEE, 19(8), pp. 2391–2406.

Spruit, M., Linden, V. van der and others (2019) 'BIDQI: The business impacts of data quality interdependencies model', *Technical Report Series*. UU BETA ICS Departement Informatica, (UU-CS-2019-001).

Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997) 'Data quality in context', *Communications of the ACM*. ACM, 40(5), pp. 103–110.

Vasta, R. *et al.* (2017) 'Outlier Detection for Sensor Systems (ODSS): A MATLAB Macro for Evaluating Microphone Sensor Data Quality.', *Sensors (14248220)*, 17(10), pp. 1–14. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=125917379&site=ehost-live.

Wang, R. Y. (1998) 'A product perspective on total data quality management', *Communications of the ACM*. Association for Computing Machinery, Inc., 41(2), pp. 58–66.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond accuracy: What data quality means to data consumers', *Journal of management information systems*. Taylor & Francis, 12(4), pp. 5–33.

Wang, Yuzhi *et al.* (2017) 'A deep learning approach for blind drift calibration of sensor networks', *IEEE Sensors Journal*. IEEE, 17(13), pp. 4158–4171.

Moore, R., Lopes, J. (1999). Paper templates. In *TEMPLATE'06, 1st International Conference on Template Production*. SCITEPRESS.

Smith, J. (1998). *The book*, The publishing company. London, 2nd edition.

Abdelkader, G., Elgazzar, K. and Khamis, A. (2021) 'Connected vehicles: technology review, state of the art, challenges and opportunities', *Sensors*. Multidisciplinary Digital Publishing Institute, 21(22), p. 7712.

Aziz, A. A., Saman, M. Y. M. and Jusoh, J. A. (2012) 'Data investigation: Issues of data quality and implementing base analysis technique to evaluate quality of data in heterogeneous databases', *Journal of Theoretical and Applied Information Technology*, 45(1), pp. 360–372. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84874528600&partnerID=40&md5=c0a93b1f761e14519748e0717d2f0282.

Barnes, B. B. and Hu, C. (2013) 'A Hybrid Cloud Detection Algorithm to Improve MODIS Sea Surface Temperature Data Quality and Coverage Over the Eastern Gulf of Mexico', *IEEE Transactions on Geoscience and Remote Sensing*, 51(6), pp. 3273–3285. doi: 10.1109/TGRS.2012.2223217.

Batini, C. *et al.* (2009) 'Methodologies for data quality assessment and improvement', *ACM computing surveys (CSUR)*. ACM, 41(3), p. 16.

Batini, C. *et al.* (2015) 'From data quality to big data quality', *Journal of Database Management*, 26(1), pp. 60–82. doi: 10.4018/JDM.2015010103.

Batini, C., Scannapieco, M. and others (2016) 'Data and information quality', *Cham, Switzerland: Springer International Publishing*. Springer.

Benneyan, J. C. (1998) 'Use and interpretation of statistical quality control charts', *International Journal for Quality in Health Care*. JSTOR, 10(1), pp. 69–73.

Brimicombe, A. (2010) 'GIS, environmental modeling and engineering'. CRC Press/Taylor & Francis Group.

Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal*. Ubiquity Press, 14.

Cichy, C. and Rass, S. (2019) 'An overview of data quality frameworks', *IEEE Access*. IEEE, 7, pp. 24634–24648.

Dai, W., Yoshigoe, K. and Parsley, W. (2018) 'Improving Data Quality Through Deep Learning and Statistical Models', in *Information Technology-New Generations*. Springer, pp. 515–522.

Davenport, T. and Redman, T. (2015) 'Build data quality into the internet of things', *Wall Str. J.*

Delen, D., Walker, G. and Kadam, A. (2005) 'Predicting breast cancer survivability: a comparison of three data mining methods', *Artificial intelligence in medicine*. Elsevier, 34(2), pp. 113–127.

Diop, M. *et al.* (2017) 'A methodology for prior management of temporal data quality in a data mining process', in *2017 Intelligent Systems and Computer Vision*

(ISCV), pp. 1–8. doi: 10.1109/ISACV.2017.8054906.

Eppler, M. J. (2006) *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science \& Business Media.

Ester, M. *et al.* (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise.', in *kdd*, pp. 226–231.

Farooqi, M. M., Khattak, H. A. and Imran, M. (2018) 'Data quality techniques in the internet of things: Random forest regression', in *2018 14th International Conference on Emerging Technologies (ICET)*, pp. 1–4.

Fekade, B. *et al.* (2018) 'Probabilistic recovery of incomplete sensed data in IoT', *IEEE Internet of Things Journal*. IEEE, 5(4), pp. 2282–2292.

Floridi, L. (2013) 'Information quality', *Philosophy & Technology*. Springer, 26(1), pp. 1–6.

for Standardization, I. O. (2021) *Quality Management Systems--Fundamentals and Vocabulary*. International Organization for Standardization.

Fox, C., Levitin, A. and Redman, T. (1994) 'The notion of data and its quality dimensions', *Information processing & management*. Elsevier, 30(1), pp. 9–19.

Group, D. Q. D. W. and others (2013) 'The Six Dimensions of EHDI Data Quality Assessment'. DAMA UK.

Gubbi, J. *et al.* (2013) 'Internet of Things (IoT): A vision, architectural elements, and future directions', *Future generation computer systems*. Elsevier, 29(7), pp. 1645–1660.

Gudivada, V., Apon, A. and Ding, J. (2017) 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations', *International Journal on Advances in Software*, 10(1), pp. 1–20.

Hazen, B. T. *et al.* (2014) 'Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications', *International Journal of Production Economics*. Elsevier, 154, pp. 72–80.

He, H. and Garcia, E. A. (2009) 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering*. Ieee, 21(9), pp. 1263–1284.

Jadaan, K., Zeater, S. and Abukhalil, Y. (2017) 'Connected vehicles: an innovative transport technology', *Procedia Engineering*. Elsevier, 187, pp. 641–648.

Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*. American Association for the Advancement of Science, 349(6245), pp. 255–260.

Juddoo, S. *et al.* (2018) 'Data governance in the health industry: investigating data quality dimensions within a big data context', *Applied System Innovation*. Multidisciplinary Digital Publishing Institute, 1(4), p. 43.

Juran, J. and Godfrey, A. B. (1999) 'Quality handbook', *Republished McGraw-Hill*, 173(8), pp. 34–51.

Karkouch, A. *et al.* (2016) 'Data quality in internet of things: A state-of-the-art survey', *Journal of Network and Computer Applications*. Elsevier, 73, pp. 57–81.

Keller, S. *et al.* (2017) 'The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches', *Annual Review of Statistics and Its Application*. Annual Reviews, 4, pp. 85–108.

Khan, K. *et al.* (2014) 'DBSCAN: Past, present and future', in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238.

Kim, S. *et al.* (2019) 'Extending data quality management for smart connected product operations', *IEEE Access*. IEEE, 7, pp. 144663–144678.

Kotsiantis, S. *et al.* (2006) 'Handling imbalanced datasets: A review', *GESTS international transactions on computer science and engineering*, 30(1), pp. 25–36.

Li, J. (2013) 'Logistic regression', *Course Notes. URL http://sites. stat. psu. edu/\~{} jiali/course/stat597e/notes2/logit. pdf*.

Liaw, S.-T. *et al.* (2011) 'Data quality and fitness for purpose of routinely collected data--a general practice case study from an electronic Practice-Based Research Network (ePBRN)', in *AMIA Annual Symposium Proceedings*, p. 785.

Loshin, D. (2010) *The practitioner's guide to data quality improvement*. Elsevier.

Loshin, D. (2011) 'Evaluating the business impacts of poor data quality', *Information Quality Journal*.

Mahmood, Z. (2020) 'Connected vehicles in the IoV: Concepts, technologies and architectures', in *Connected vehicles in the internet of things*. Springer, pp. 3–18.

Maydanchik, A. (2007) *Data quality assessment*. Technics publications.

Megler, V. M., Tufte, K. and Maier, D. (2016) 'Improving data quality in intelligent transportation systems', *arXiv preprint arXiv:1602.03100*.

Miraz, M. H. *et al.* (2015) 'A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)', in *2015 Internet Technologies and Applications (ITA)*, pp. 219–224.

Nelson, L. S. (1985) 'Interpreting Shewhart X control charts', *Journal of Quality Technology*. Taylor \& Francis, 17(2), pp. 114–116.

Olufowobi, H. *et al.* (2016) 'Data provenance model for Internet of Things (IoT) systems', in *International Conference on Service-Oriented Computing*, pp. 85–91.

Perez-Castillo, R. *et al.* (2018) 'Data Quality Best Practices in IoT Environments', in *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pp. 272–275.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data quality assessment', *Communications of the ACM*. ACM, 45(4), pp. 211–218.

Qian, L. *et al.* (2009) 'Cloud computing: An overview', in *IEEE international conference on cloud computing*, pp. 626–631.

Redman, T. C. (2017) 'Seizing opportunity in data quality', *MIT Sloan Management Review. https://sloanreview. mit. edu/article/seizing-opportunity-in-data-quality. November*, 29.

Sebastian-Coleman, L. (2012) *Measuring data quality for ongoing improvement: a data quality assessment*

framework. Newnes.

Sethi, P. and Sarangi, S. R. (2017) 'Internet of things: architectures, protocols, and applications', *Journal of Electrical and Computer Engineering*. Hindawi, 2017.

Siegel, J. E., Erb, D. C. and Sarma, S. E. (2017) 'A survey of the connected vehicle landscape—Architectures, enabling technologies, applications, and development areas', *IEEE Transactions on Intelligent Transportation Systems*. IEEE, 19(8), pp. 2391–2406.

Spruit, M., Linden, V. van der and others (2019) 'BIDQI: The business impacts of data quality interdependencies model', *Technical Report Series*. UU BETA ICS Departement Informatica, (UU-CS-2019-001).

Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997) 'Data quality in context', *Communications of the ACM*. ACM, 40(5), pp. 103–110.

Vasta, R. *et al.* (2017) 'Outlier Detection for Sensor Systems (ODSS): A MATLAB Macro for Evaluating Microphone Sensor Data Quality.', *Sensors (14248220)*, 17(10), pp. 1–14. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=125917379&site=ehost-live.

Wang, R. Y. (1998) 'A product perspective on total data quality management', *Communications of the ACM*. Association for Computing Machinery, Inc., 41(2), pp. 58–66.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond accuracy: What data quality means to data consumers', *Journal of management information systems*. Taylor & Francis, 12(4), pp. 5–33.

Wang, Yuzhi *et al.* (2017) 'A deep learning approach for blind drift calibration of sensor networks', *IEEE Sensors Journal*. IEEE, 17(13), pp. 4158–4171.

Rosameo, (2021) '*GitHub - rosameo/Sensors-Data-about-Fuel-Consumption-in-Buses: This repository provides data collected by sensors about fuel consumption and contextual conditions of buses of public transportation*'