IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE POWER & ENERGY SOCIETY SECTION

Received June 15, 2021, accepted June 24, 2021, date of publication July 5, 2021, date of current version July 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094671

Fuzzy Theory Based Quality Assessment of Multivariate Electrical Measurements of Smart Grids

SOUMYAJIT GANGOPADHYAY, (Graduate Student Member, IEEE), AND SARASIJ DAS, (Senior Member, IEEE)
Indian Institute of Science, Bengaluru, Karnataka 560012, India
Corresponding author: Sarasij Das (sarasij@iisc.ac.in)

ABSTRACT Electrical measurements of smart grids form the backbone of various data driven applications. Proper analysis of these measurements can result in improved system planning, operation, monitoring and protection. However, the efficacy of smart grid data mining is highly influenced by the quality of the data. Hence, quality assessment of smart grid data is essential prior to the usage of the data in various applications. A fuzzy assessment method is proposed in this paper for assessing quality of multivariate electrical measurements of smart grids. At first, relevant quality dimensions of smart grid data are identified. Then, based on certain desirable characteristics, novel membership functions are proposed for assessing the data quality with respect to each of the considered dimensions. The proposed membership functions are evaluated on the current and real power measurements obtained from the power flow analysis of the IEEE 14-bus system. In addition, the proposed method is also implemented on the voltage, current and the real power measurements obtained from the power flow analysis of an actual 34 node feeder located in Arizona. The impact of measurement noise is also investigated by polluting the original measurements with Gaussian noise. It is found that the quality of the noisy measurements worsens with the increase in variance of the added noise. The proposed method has also been validated on a database containing practical SCADA and PMU measurements of the Southern Regional Grid of India. It is found that the PMU datasets are relatively incomplete compared to the SCADA datasets. In addition, the obtained results indicate that PMU data are suitable for use in more number of applications compared to the SCADA datasets. Unlike the existing methods, the proposed method can be used for quantifying the quality of any smart grid dataset that contains electrical measurements of multiple power system variables including boolean variables such as circuit breaker status. Moreover, unlike the existing methods, the proposed method can measure the consistency among the measurements. In addition, the proposed method is found to be sensitive to the distribution of the bad measurements in a given database.

INDEX TERMS Data quality, multivariate data, smart grids.

I. INTRODUCTION

Smart grid data are increasingly being used in various data driven power system applications ranging from power theft detection [1], harmonic state estimation [2], forced oscillation source location [3] to transmission line parameters' estimation [4] and high impedance fault detection in distribution systems [5]. Electrical measurements of smart grids form the backbone of a majority of data driven applications. These measurements often contain important information about the state of the grid. As a result, proper analysis of these measurements can lead to better planning, operation, monitoring and protection of smart grids [6]. However, the outcome of smart grid data analytics is often influenced by the quality of smart grid data [7]–[9]. Quality of data can be affected by a number of factors. Some of the factors that can adversely affect the quality of electricity consumption data have been presented in [10]. Poor data quality may limit the usability of the data in various applications. For example, poor quality PMU data may inhibit forced oscillation source location in smart grids. Even distribution system state estimation can be hampered by poor data quality [11]. Moreover, mining of poor quality

The associate editor coordinating the review of this manuscript and approving it for publication was Elizete Maria Lourenco.

smart grid data can result in erroneous inferences. Actions taken based on incorrect inferences can adversely impact the grid safety and reliability. For example, the value of the transmission line parameters estimated using poor quality SCADA data can be significantly different from their true values. Usage of the erroneous estimates can adversely affect contingency analysis, short circuit analysis and state estimation. Hence, it is clear from the foregoing discussion that poor quality of smart grid data can act as a major roadblock in the path of advanced smart grid data analytics. Unless smart grid measurements are of sufficiently good quality, it may not be possible to extract useful information from such measurements. Hence it is necessary to assess the quality of the smart grid measurements. Moreover, data quality based smart grid data analytics are increasingly getting attention from the research community [12], [13]. Quality assessment of smart grid data can help to determine the usability of the data in various applications. Data quality assessment may also help the utilities in assessing their data management practices. For example, if the estimated quality of data is poor, then the utilities can take steps to improve the existing data management practices. Thus quality assessment of smart grid data is essential prior to the usage of the data in various applications.

In computer science and data mining literature, data quality has been defined as a measure of fitness of data for use in an application [14], [15]. While a particular dataset may be useful for a particular application, the same dataset may not be fit for use in another application. In other words, data quality is relative and the outcome of data quality assessment may depend on the context. Multiple dimensions have been defined in [16], [17] for assessing data quality which makes data quality estimation a multidimensional process. In addition, [17] has also proposed various metrics for measuring and improving data quality. However, majority of the metrics proposed in [17] are based on questionnaires and user surveys. In other words, [16], [17] do not provide any mathematical approach for assessing data quality. Moreover, the definition of the quality dimensions proposed in [17] may not be directly applicable for smart grid data due to the contextual nature of data quality. Literature [18] has proposed quality assessment of data based on the subjective perceptions of the data users. Reference [19] has shown (by testing several hypotheses on a set of users) that the characteristics of the decision maker (or data user) plays a significant role while assessing data quality in a context. In other words, different users may rate the same dataset differently. Quality assessment of linked data has been proposed in [20], [21]. In [20], a goal-question-metric driven approach has been used where simple ratios have been proposed for measuring the inherent quality characteristics of linked open datasets. Quality of a linked dataset is often influenced by the quality of the dataset to which it is linked. However, the quality of the linking between datasets has not been considered in [20]. In smart grids, the electrical measurements may be linked with each other through system equations. As a result, quality assessment of smart grid data may require estimation of the consistency among measurements which makes it difficult to apply the approach of [20] for smart grid data quality estimation. In [21], several dimensions have been proposed and defined for assessing quality of linked open data. However, most of the dimensions proposed in [21] overlap significantly with each other which makes it difficult to precisely quantify the individual dimensions. Big data quality assessment has been proposed in [22]. In [22], three data quality dimensions have been proposed and measured using simple ratios. Big datasets often suffer from data duplication and lack of interpretability of the data. However, [22] has not assessed the duplication and interpretability aspects of big data. In [23], seven data quality dimensions have been considered for assessing quality of big data. However, [23] has not provided mathematical formulation of the considered quality dimensions. Moreover, [23] has used only a subset of the data for estimating the quality of the entire dataset which may lead to flawed quality assessment in some situations. Reference [24] has proposed quality assessment for IoT data by considering the physical and the environmental characteristics of the sensors in addition to the quality dimensions used in [23]. However, [24] has not shed light on the evaluation of the sensor characteristics required for assessing data quality. In addition, [24] has not presented any mathematical formulation for measuring the proposed quality dimensions. A six step data quality assessment framework has been proposed in [25] for assessing quality of electronic medical record data of a group of patients. In [26], an automated framework has been developed for medical data curation that can detect outliers and missing values in a database containing records of various patients. Basically, the framework proposed in [26] essentially aims at controlling and improving the quality of a medical database by detecting outliers and incomplete data. However, the same framework may not be applicable to a database containing smart grid data. The automated framework in [26] detects outliers by using statistical techniques like z-score estimate, Grubb's test and interquartile range. These techniques consider a datapoint as an outlier if the datapoint is significantly different from majority of the data. However, in power system, data that are significantly different may represent data during system disturbances like faults and hence may not be outliers. Removal of such data from the database may lead to loss of crucial information about the system. Hence, the framework proposed in [26] may not be directly applicable for improving quality of smart grid data. Moreover, [25] and [26] do not provide any data quality metrics required for benchmarking acceptable levels of medical record data. Quality of building footprints data on OpenStreetMap has been assessed in [27] by comparing the data with the reference (or standard) building footprints data in German Authority Topographic–Cartographic Information System. Several quality indicators of volunteered geographic information (VGI) data have been proposed in [28]. For accuracy assessment of VGI data, VGI data are often compared against similar reference data (obtained from other sources).

However, in the area of power systems, such reference data are usually unavailable which makes it difficult to apply the quality assessment methods of [27], [28] for assessing smart grid data quality. Moreover, majority of the VGI data quality indicators (like population density, population age, income, social deprivation, contributors' behaviour and contributors' education) are not relevant for quality assessment of smart grid data. Mathematical estimation of data quality has been proposed in [29] for boolean datasets. However, electrical measurements of smart grids are not always boolean (for example, voltage or current measurements). As a result, the method proposed in [29] may not be applicable for assessing smart grid data quality. Literature [30] and [31] have proposed quality assessment of sensing data submitted by the participants participating in crowdsensing tasks. While [30] proposes employing multiple verifiers for evaluation of a single submitted task, the method proposed in [31] assesses the quality of the submitted task by estimating the effort spent by the participants in completing the assigned tasks. Literature [30] and [31] have mostly focused on the correctness of the task submitted by the participants. However, [30] and [31] have not considered the time taken by the participants to complete the task. Time taken by the participants may impact the quality of the submitted task. In other words, quality of data may depend on the availability of the data. Moreover, sensing data are completely different from multivariate time series data such as electrical measurements of smart grids. As a result, methods for quality assessment of sensing data may not be directly applicable to smart grid data. Thus, outside smart grid literature, there is a lack of data quality assessment techniques that can be used for assessing quality of smart grid data. Recently, quality assessment of smart grid data has been proposed in [32]. In [32], data quality has been assessed for a system variable (such as bus voltage or line current) independent of the other variables in a dataset. However, power system variables are usually correlated as they are governed by system equations. So, quality of one variable may also depend on quality of the other variables in a dataset. In other words, it is essential to measure the consistency among the measurements contained in a dataset. However, literature [32] has not considered inter-variable correlations while assessing data quality. Hence, the formulations presented in [32] may not be directly applicable for assessing quality of a time series of multivariate electrical measurements. So there is a need for a data quality assessment framework for assessing quality of a time series of multivariate electrical measurements of smart grids.

Literature [32] has focused on the quantification of data quality. Quantification of data quality may be necessary to understand (or measure) how good or how bad the quality of a dataset is. In [32], a number is generated to quantify a dataset's quality. However, it can be difficult to interpret the number that characterizes a dataset's quality. For example, suppose a dataset is found to have a quality of 0.8. This number (i.e. 0.8) may not be sufficiently comprehensible to a system operator (or any other user of the dataset). In other words, a single number (representing data quality) may not be helpful to draw conclusions about the quality of the dataset in hand. In order to have a better understanding of the quality of a dataset, a fuzzy assessment methodology can be useful in addressing the aforementioned problem. For example, usage of linguistic (or fuzzy) labels for assessing data quality (such as GOOD quality or BAD quality or VERY BAD quality) may be more comprehensible to a system operator (or any other user of the dataset). Hence, in this paper, a fuzzy assessment method is proposed for assessing quality of multivariate electrical measurements of smart grids. At first, relevant quality dimensions of smart grid data are identified. Then, novel membership functions are proposed for each quality dimension based on certain desirable characteristics. The proposed membership functions are used for measuring the degree of membership of a smart grid dataset in various fuzzy sets that represent various quality dimensions of smart grid data. Based on the value of membership functions, appropriate fuzzy labels are used for fuzzy assessment of each of the considered data quality dimensions. The proposed membership functions can be calculated on any smart grid dataset that contains electrical measurements of multiple power system variables. The proposed membership functions are evaluated on (i) the current and the real power measurements obtained from power flow analysis of the standard IEEE 14 bus system and (ii) on the voltage, current and the real power measurements obtained from power flow analysis of an actual 34 node feeder located in Arizona. The impact of measurement noise is also investigated by modifying the original measurements with zero mean Gaussian noise. The obtained results show that the quality of the original measurements is better than that of the noisy measurements. Moreover, quality of the noisy measurements worsens with the increase in variance of the added noise. The proposed membership functions are also used for assessing the quality of practical SCADA and PMU measurements. The obtained results show that the PMU datasets are relatively incomplete compared to the SCADA datasets. The obtained results also show that the PMU datasets have better usability (or suitability for use) across a wide variety of applications compared to the SCADA datasets. In summary, the main contributions of this paper are:

- This paper identifies and quantifies the dimensions that are relevant for assessment of smart grid data quality.
- Novel membership functions are proposed (based on certain desirable characteristics) for fuzzy assessment of quality of multivariate electrical data of smart grids.
- The proposed approach is validated on (i) the measurements obtained from the power flow analysis of the standard IEEE 14 bus system, (ii) the measurements obtained from the power flow analysis of an actual 34 node feeder located in Arizona and (iii) the practical PMU and SCADA measurements of the Southern Regional Grid of India.
- The impact of measurement noise on the performance of the proposed membership functions is investigated using various case studies.

- The proposed method is compared with some of the existing data quality assessment methods.

II. OVERVIEW OF FUZZY ASSESSMENT METHOD

Fuzzy assessment method has been used in a variety of power system applications including transient stability assessment of power systems [33]. In this method, linguistic labels (such as GOOD, VERY GOOD, HIGH, PARTIALLY HIGH, LOW etc.) are used for assessment of a variable. Usage of linguistic labels makes the assessment fuzzy. For example, linguistic labels (such as HOT, VERY HOT, COLD and VERY COLD) can be used for fuzzy assessment of weather at a particular place. Suppose an output (or dependent) variable depends on certain input (or independent) variables. So, assessment of the output variable cannot be made without assessment of the input variables. As a result, the first step in fuzzy assessment of an output variable is the fuzzy assessment of the input variables. This step is called fuzzification. In this step, a crisp input is assigned to various fuzzy sets with different degree of membership [34], [35]. The degree of membership of a crisp input in a particular fuzzy set is measured by the membership function associated with the set [34], [35]. In other words, every fuzzy set is characterized by a membership function that represents the degree of membership of a crisp input in that particular set. The output obtained from the fuzzification block basically represents fuzzy information about the input variables. Based on the fuzzy information about the input variables, appropriate linguistic labels can be used for fuzzy assessment of the input variables. This fuzzy information is then passed through a fuzzy rulebase. A fuzzy rulebase contains a set of fuzzy rules. Fuzzy rules are essentially IF-THEN statements that explicitly define the mapping between the input membership functions and the output membership functions [34], [35]. In other words, fuzzy rules help to generate the fuzzy information about the output variable from the fuzzy information about the input variables. Once the fuzzy information about the output variable is obtained, appropriate linguistic labels can be used for fuzzy assessment of the output variable. In some situations, a defuzzification block is also used after the fuzzy information about the output variable is obtained. The defuzzification block is used to generate a crisp value of the output variable that best represents the fuzzy information about the output variable [34], [35]. It is to be noted that unless a crisp estimation of the output variable is needed, the defuzzification block may not be required [34], [35]. Moreover, the fuzzy rules are useful only for generating fuzzy information about the output variable based on the fuzzy information about the input variables [34], [35]. Since fuzzy rules explicitly define the mapping between the input and the output variables, so fuzzy rules are required only for doing fuzzy assessment of output (or dependent) variable. In other words, fuzzy rules are not required for fuzzy assessment of the input (or independent) variables. However, the fuzzification block is necessary for fuzzy assessment of an input variable.

In this paper, eight different data quality dimensions (or indicators) are considered for assessing quality of a smart grid dataset. Membership functions are proposed for measuring the degree of membership of a smart grid dataset in various fuzzy sets that represent various dimensions of data quality. Based on the value of the membership functions, appropriate fuzzy labels are used for fuzzy assessment of each of the considered data quality dimensions. Thus, for a given smart grid dataset, the membership functions proposed in this paper essentially help in obtaining fuzzy information regarding each of the considered data quality dimensions. This fuzzy information is then used for fuzzy assessment of quality of a smart grid dataset.

III. PROPOSED FUZZY ASSESSMENT OF DATA QUALITY

In this paper, it is assumed that a smart grid dataset $\mathcal{D}$ has m rows where each row contains measurements of n power system variables (or attributes) at a particular timestamp. Hence, each row of $\mathcal{D}$ represents multivariate electrical data. The power system variables (whose measurements are contained in $\mathcal{D}$) may include voltage or frequency measurements of different buses of a network or current or power flow measurements in various transmission lines. In this section, quality of $\mathcal{D}$ is assessed by using a novel fuzzy assessment method. At first, it is assumed that the Universe of Discourse U consists of all possible multivariate electrical datasets of smart grids. In other words, every smart grid dataset is a member of U. Then, novel membership functions are proposed and defined over the entire Universe of Discourse U. For a given smart grid dataset $\mathcal{D} \in U$, the proposed membership functions are used for measuring the degree of membership of $\mathcal{D}$ in various fuzzy sets that denote various dimensions of smart grid data quality. Based on the degree of membership of $\mathcal{D}$ in the considered fuzzy sets, appropriate linguistic labels are used for fuzzy assessment of each of the considered data quality dimensions. The following data quality dimensions are considered in this paper for quality assessment of a given smart grid dataset $\mathcal{D}$.

- Completeness
- Accuracy
- Inter-variable Consistency
- Duplication
- Data Measurement Rate
- Amount of Data
- Interpretability
- Availability

Each of the aforementioned data quality dimensions is considered as an input (or independent) variable in this paper. It has been mentioned in section II that fuzzy assessment of input variables does not require a fuzzy rulebase or a defuzzification block. Hence neither fuzzy rules nor defuzzification is required for fuzzy assessment of each of the aforementioned data quality dimensions. In this paper, fuzzy assessment of each of the aforementioned data quality dimensions is proposed based on the value of the

membership functions. Detailed description and formulation of the membership functions are provided in the following subsections.

A. COMPLETENESS

A row of $\mathcal{D}$ can be called complete if measurement of each of the n attributes is contained in that row. As a result, $\mathcal{D}$ can be called complete if each of the m rows of $\mathcal{D}$ is complete. Now, each of the m rows of a given dataset $\mathcal{D}$ may not contain the measurement of all the n attributes. Measurements of some of the attributes may be missing in a given row. In this paper, the measurement of an attribute is considered missing if it is found empty or if the value reported against that attribute is NULL. Otherwise, the measurement of an attribute is considered available. Hence, a row having some missing attributes cannot be called complete. However, such a row can neither be called incomplete. This is because such a row may contain measurement of some other attributes. As a result, such a row is both partially complete and partially incomplete. Suppose x_k represents the number of attributes whose measurements are contained (or available) in row $k\,\forall k \in \{1, 2, \ldots, m\}$. Then $x_k \in \{0, 1, 2, \ldots, n\}$. In this section, membership functions are proposed to represent the degree of membership of a row of $\mathcal{D}$ in the set of complete and incomplete rows. Suppose the degree of membership of row k in the set of complete rows is denoted by $C(k)\ \forall k \in \{1, \ldots, m\}$. Then $C(k)$ should satisfy the following properties.

1) For row k, $C(k) = 1 \iff x_k = n\ \ \forall k \in \{1, 2, \ldots, m\}$
2) For row k, $C(k) = 0 \iff x_k = 0\ \ \forall k \in \{1, 2, \ldots, m\}$
3) Between two rows i and j (where $i \neq j$), $C(i) > C(j) \iff x_i > x_j$

Accordingly, $C(k)$ is given by (1) $\forall k \in \{1, 2, \ldots, m\}$.

$$C(k) = \left(\frac{x_k}{n}\right)^p \tag{1}$$

where $p \in \mathbb{Z}_{>0}$ and $\mathbb{Z}_{>0}$ is the set of positive integers. The degree of membership of row k in the set of incomplete rows is equal to $1-C(k)$. Thus the resulting fuzzy sets representing the set of complete and incomplete rows are $\{(k, C(k))\ \forall k \in \{1, \ldots, m\}\}$ and $\{(k, 1 - C(k))\ \ \forall k \in \{1, \ldots, m\}\}$ respectively.

The degree of membership of $\mathcal{D}$ in the set of complete datasets should be influenced by the degree of membership (in the set of complete rows) of each row of $\mathcal{D}$. Suppose r_c represents the number of rows of $\mathcal{D}$ having atleast one missing attribute each. Suppose there exist two datasets where each of the datasets has m rows and the same total (or sum) of the degree of membership of all the rows (in the set of complete rows). Then, the dataset having a lower value of r_c should be assigned a higher degree of membership (in the set of complete datasets) than the other. Suppose $C(\mathcal{D})$ and $IC(\mathcal{D})$ represent the degree of membership of $\mathcal{D}$ in the set of complete and incomplete datasets respectively. Then $C(\mathcal{D})$ and $IC(\mathcal{D})$ are given by (2) and (3) respectively.

$$C(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} C(k)\right\} - \frac{r_c}{m} + 1\right] \tag{2}$$

$$IC(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} (1 - C(k))\right\} + \frac{r_c}{m}\right] \tag{3}$$

The resulting fuzzy sets representing the set of complete and incomplete datasets are $\{(\mathcal{D}, C(\mathcal{D}))\ \ \forall \mathcal{D} \in U\}$ and $\{(\mathcal{D}, IC(\mathcal{D}))\ \ \forall \mathcal{D} \in U\}$ respectively. Both $C(\mathcal{D})$ and $IC(\mathcal{D})$ can be nonzero for a given dataset $\mathcal{D}$. In such situation, an appropriate linguistic label can be used to describe the completeness of $\mathcal{D}$ based on the values of $C(\mathcal{D})$ and $IC(\mathcal{D})$. For example, if $C(\mathcal{D}) \gg IC(\mathcal{D})$, then it can be said that $\mathcal{D}$ has HIGH completeness. Similarly, if $C(\mathcal{D})$ is slightly higher than $IC(\mathcal{D})$, then it can be said that $\mathcal{D}$ has SLIGHTLY HIGH completeness. On the other hand, if $C(\mathcal{D}) \ll IC(\mathcal{D})$, then it can be said that $\mathcal{D}$ has LOW completeness.

B. ACCURACY

A row of $\mathcal{D}$ can be called accurate if measurement of each of the available attributes (i.e. attributes for which measurement is available) of that row is accurate. Bad data detection algorithms can be used to determine whether the available measurements are accurate or not. It is to be noted that a row may not have the measurements of all the n attributes. However, a partially incomplete row can also be accurate if the measurement of each of the available attributes is accurate. In other words, unavailability of some of the measurements should not influence the degree of accuracy of a row. Accordingly, $\mathcal{D}$ can be called accurate if each of the m rows of $\mathcal{D}$ is accurate. It has been stated earlier that x_k represents the number of attributes whose measurements are available in row $k\ \forall k \in \{1, \ldots, m\}$. Now, each of the x_k measurements available in row k may not be accurate. In such a situation, row k can neither be called accurate nor be called inaccurate. In other words, row k is both partially accurate and partially inaccurate. Suppose y_k represents the number of accurate measurements of row $k\ \ \forall k \in \{1, 2, \ldots, m\}$. Then $y_k \in \{0, 1, 2, \ldots, x_k\}$. In this section, membership functions are proposed to represent the degree of membership of a row of $\mathcal{D}$ in the set of accurate and inaccurate rows. Suppose the degree of membership of row k in the set of accurate rows is denoted by $A(k)\ \forall k \in \{1, \ldots, m\}$. Then $A(k)$ should satisfy the following properties.

1) For row $k, A(k) = 1 \iff y_k = x_k\ \ \forall k \in \{1, 2, \ldots, m\}$
2) For row $k, A(k) = 0 \iff y_k = 0\ \ \forall k \in \{1, 2, \ldots, m\}$
3) Between two rows i and j (where $i \neq j$) having $x_i = x_j$, $A(i) > A(j) \iff y_i > y_j$
4) Between two rows i and j (where $i \neq j$) having $y_i = y_j$, $A(i) > A(j) \iff x_i < x_j$

Accordingly, $A(k)$ is given by (4) $\forall k \in \{1, 2, \ldots, m\}$.

$$A(k) = \left(\frac{y_k}{x_k}\right)^q \tag{4}$$

where $q \in \mathbb{Z}_{>0}$. The degree of membership of row k in the set of inaccurate rows is equal to $1 - A(k)$. Thus the resulting fuzzy sets representing the set of accurate and inaccurate rows are $\{(k, A(k)) \ \forall k \in \{1, \ldots, m\}\}$ and $\{(k, 1 - A(k)) \ \forall k \in \{1, \ldots, m\}\}$ respectively.

The degree of membership of $\mathcal{D}$ in the set of accurate datasets should be influenced by the degree of membership (in the set of accurate rows) of each row of $\mathcal{D}$. Suppose r_a represents the number of rows of $\mathcal{D}$ having atleast one inaccurate attribute each. Suppose there exist two datasets where each of the datasets has m rows and the same total (or sum) of the degree of membership of all the rows (in the set of accurate rows). Then, the dataset having a lower value of r_a should be assigned a higher degree of membership (in the set of accurate datasets) than the other. Suppose $A(\mathcal{D})$ and $IA(\mathcal{D})$ represent the degree of membership of $\mathcal{D}$ in the set of accurate and inaccurate datasets respectively. Then $A(\mathcal{D})$ and $IA(\mathcal{D})$ are given by (5) and (6) respectively.

$$A(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} A(k)\right\} - \frac{r_a}{m} + 1\right] \tag{5}$$

$$IA(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} (1 - A(k))\right\} + \frac{r_a}{m}\right] \tag{6}$$

The resulting fuzzy sets representing the set of accurate and inaccurate datasets are $\{(\mathcal{D}, A(\mathcal{D})) \ \forall \mathcal{D} \in U\}$ and $\{(\mathcal{D}, IA(\mathcal{D})) \ \forall \mathcal{D} \in U\}$ respectively. Both $A(\mathcal{D})$ and $IA(\mathcal{D})$ can be nonzero for a given dataset $\mathcal{D}$. In such situation, an appropriate linguistic label can be used to describe the accuracy of $\mathcal{D}$ based on the values of $A(\mathcal{D})$ and $IA(\mathcal{D})$. For example, if $A(\mathcal{D}) \gg IA(\mathcal{D})$, then it can be said that $\mathcal{D}$ has HIGH accuracy. On the other hand, if $A(\mathcal{D}) \ll IA(\mathcal{D})$, then it can be said that $\mathcal{D}$ has LOW accuracy.

C. INTER-VARIABLE CONSISTENCY

The measurements (recorded at a given timestamp) corresponding to a set of power system variables (or attributes) can be called inter-consistent if they satisfy the equations governing the power system. For example, if a smart grid dataset contains the line current measurement of various transmission lines meeting at a node, then inter-variable consistency among these measurements can be assessed by using Kirchoff's Current Law (KCL). It is to be noted that checking consistency among the power system measurements is different from assessing accuracy of the measurements. Power system measurements can be inter-consistent even in the presence of bad (or inaccurate) measurements. Inter-variable consistency only captures whether the measurements of the attributes reported in $\mathcal{D}$ satisfy the system equations or not. Suppose $v_1, v_2, \ldots, v_n$ represent the n power system attributes whose measurements (at different timestamps) are contained in $\mathcal{D}$. These measurements can be called inter-consistent if they satisfy (7) at every timestamp.

$$\begin{aligned} f_1(v_1, v_2, \ldots, v_n) &= 0 \\ &\vdots \\ f_h(v_1, v_2, \ldots, v_n) &= 0 \end{aligned} \tag{7}$$

Depending on which power system attributes are contained in $\mathcal{D}$, the number of equations and the exact set of equations that can be validated may vary. Moreover, for certain network equations to be validated, the knowledge of the network topology and the network parameters (like line resistance, line inductance and shunt capacitance) may be necessary. Hence, such equations can be a part of (7) if the network topology or the network parameters are known beforehand. Suppose the measurements of row k of $\mathcal{D}$ produce the values $e_{k1}, \ldots, e_{kh}$ against the functions $f_1, \ldots, f_h$. Thus the measurements produce an absolute error $|e_{k1}|$ against f_1, $|e_{k2}|$ against f_2 and so on. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets having inter-consistent variables and non-interconsistent variables are denoted by $ICo(\mathcal{D})$ and $NICo(\mathcal{D})$ respectively. Then $ICo(\mathcal{D})$ should satisfy the following properties.

1) If the absolute error produced by measurements of row k against the function f_j increases, then $ICo(\mathcal{D})$ should decrease i.e. $\frac{dICo(\mathcal{D})}{d|e_{kj}|} < 0 \ \forall |e_{kj}| > 0, \forall k \in \{1, \ldots, m\}, \forall j \in \{1, \ldots, h\}$.
2) For a given set of equations to be validated, a small absolute error produced by the measurements of a row against a particular function may be intolerable. However, a large absolute error against another function may be tolerable. So, $ICo(\mathcal{D})$ should be influenced by the weighted sum of absolute errors produced by the measurements of each row.
3) $ICo(\mathcal{D})$ should asymptotically decrease to zero with the increase in the weighted sum of absolute errors produced by the measurements of a row. So $ICo(\mathcal{D}) \in (0, 1]$.

Accordingly, $ICo(\mathcal{D})$ and $NICo(\mathcal{D})$ are given by (8) and (9) respectively.

$$ICo(\mathcal{D}) = \exp\left(-\frac{1}{m}\sum_{k=1}^{m} w^T \cdot e_k\right) \tag{8}$$

$$NICo(\mathcal{D}) = 1 - ICo(\mathcal{D}) \tag{9}$$

where

$$e_k = \begin{bmatrix} |e_{k1}| & |e_{k2}| & \ldots & |e_{kh}| \end{bmatrix}^T \quad w = \begin{bmatrix} w_1 & w_2 & \ldots & w_h \end{bmatrix}^T$$

In (8), exp is the exponential function, e_k refers to the absolute error vector consisting of the absolute errors produced by measurements of row k against the functions $f_1, \ldots, f_h$ and w refers to the positive real valued weight vector where w_j represents the weight per unit absolute error produced by measurements of row k against the function $f_j \ \forall j \in \{1, \ldots, h\}$ and $\forall k \in \{1, \ldots, m\}$. The resulting fuzzy sets representing the set of datasets having inter-consistent variables and

non-interconsistent variables are $\{(\mathcal{D}, ICo(\mathcal{D}))\ \forall \mathcal{D} \in U\}$ and $\{(\mathcal{D}, NICo(\mathcal{D}))\ \forall \mathcal{D} \in U\}$ respectively. Both $ICo(\mathcal{D})$ and $NICo(\mathcal{D})$ can be nonzero for a given dataset D. In such situation, an appropriate linguistic label can be used to describe the inter-variable consistency of $\mathcal{D}$ based on the values of $ICo(\mathcal{D})$ and $NICo(\mathcal{D})$. For example, if $ICo(\mathcal{D}) \gg NICo(\mathcal{D})$, then it can be said that $\mathcal{D}$ has HIGH level of consistency among the measurements. On the other hand, if $ICo(\mathcal{D}) \ll NICo(\mathcal{D})$, then it can be said that $\mathcal{D}$ has LOW level of consistency among the measurements.

D. DUPLICATION

Duplication occurs when same data are stored more than once in a database. Duplication can occur in two ways.

1) *Duplication among attributes:* Occurs when one or more attributes appear more than once in a dataset. This implies that all the n attributes of $\mathcal{D}$ are not distinct.
2) *Duplication among data:* Occurs when one or more rows appear more than once in a dataset. This implies that measurements corresponding to a particular timestamp are reported more than once.

Presence of duplication may limit the usability of the data. Hence data quality can be adversely affected with increasing duplication. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets having duplicate attributes and non-duplicate attributes are denoted by $AD(\mathcal{D})$ and $NAD(\mathcal{D})$ respectively. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets having duplicate data and non-duplicate data are denoted by $DD(\mathcal{D})$ and $NDD(\mathcal{D})$ respectively. Then $AD(\mathcal{D})$, $NAD(\mathcal{D})$, $DD(\mathcal{D})$ and $NDD(\mathcal{D})$ are given by (10), (11), (12) and (13) respectively.

$$AD(\mathcal{D}) = \begin{cases} 0 & \text{if } n_r = 0 \\ \left(\frac{n_r}{n_d}\right)^a \left(\frac{\sum_{k=1}^{n_r} v_k}{n}\right)^b & \text{if } n_r \neq 0 \end{cases} \tag{10}$$

$$NAD(\mathcal{D}) = 1 - AD(\mathcal{D}) \tag{11}$$

$$DD(\mathcal{D}) = \begin{cases} 0 & \text{if } m_r = 0 \\ \left(\frac{m_r}{m_d}\right)^a \left(\frac{\sum_{k=1}^{m_r} w_k}{m}\right)^b & \text{if } m_r \neq 0 \end{cases} \tag{12}$$

$$NDD(\mathcal{D}) = 1 - DD(\mathcal{D}) \tag{13}$$

In (10), n_d is the number of distinct power system attributes contained in $\mathcal{D}$. Out of n_d attributes, the number of attributes appearing more than once is n_r. The parameter v_k is the number of appearances of the k^{th} repeating attribute and n is the total number of attributes given by $n = n_d - n_r + \sum_{k=1}^{n_r} v_k$. In (12), m_d is the number of distinct rows. Out of m_d rows, the number of rows appearing more than once is m_r. w_k is the number of appearances of the k^{th} repeating row and m is the total number of rows given by $m = m_d - m_r + \sum_{k=1}^{m_r} w_k$.

The parameters a and b in (10) and (12) are positive real numbers that can be adjusted to control the relative influence of the ratios on the respective functions. The value of a and b in (10) should neither be too high nor be too low. Since $0 \leq \frac{n_r}{n_d} \leq 1$ and $0 \leq \frac{\sum_{k=1}^{n_r} v_k}{n} \leq 1$, so a high value of a and b (i.e. $a \gg 1$ and $b \gg 1$) will cause $AD(\mathcal{D})$ to be negligibly small even when the amount of duplication in $\mathcal{D}$ is appreciable. Similarly, a low value of a and b (i.e. $a \ll 1$ and $b \ll 1$) will cause $AD(\mathcal{D})$ to be close to 1 (indicating a high duplication) even when the amount of duplication in $\mathcal{D}$ is small. Similar reasoning applies for (12). From (10), it can be seen that $AD(\mathcal{D}) = 1$ if there does not exist an attribute that appears only once. Then, $n_r = n_d$ and $\sum_{k=1}^{n_r} v_k = n$. Similarly from (12), it can be seen that $DD(\mathcal{D}) = 1$ if there does not exist a single row that appears only once. Then, $m_r = m_d$ and $\sum_{k=1}^{m_r} w_k = m$. Both $AD(\mathcal{D})$ and $NAD(\mathcal{D})$ can be nonzero for a smart grid dataset $\mathcal{D}$. Similarly, both $DD(\mathcal{D})$ and $NDD(\mathcal{D})$ can be nonzero for a smart grid dataset $\mathcal{D}$. In such situation, an appropriate linguistic label can be used to describe the duplication in $\mathcal{D}$ based on the values of $AD(\mathcal{D})$, $NAD(\mathcal{D})$, $DD(\mathcal{D})$ and $NDD(\mathcal{D})$. For example, if $AD(\mathcal{D}) \gg NAD(\mathcal{D})$, then it can be said that $\mathcal{D}$ has a HIGH level of attribute duplication. Similarly, if $DD(\mathcal{D}) \ll NDD(\mathcal{D})$, then it can be said that $\mathcal{D}$ has a LOW level of data duplication.

E. DATA MEASUREMENT RATE

Various data driven power system studies require data of different frequencies. For example, transient studies usually require high frequency data whereas a load flow study need not require too much data. Thus, a smart grid dataset (containing data sampled at a fixed rate (or frequency)) may not be suitable for every data driven power system application. The data sampling rates may also vary across measuring devices and sensors. Most applications require a minimum data sampling rate depending on the context. Suppose f_d be the threshold (or minimum) sampling frequency of data required for an application and f_a be the actual sampling frequency of the data contained in $\mathcal{D}$. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets that are suitable for use in a given application (with respect to data measurement rate) is denoted by $DR(\mathcal{D})$. Then $DR(\mathcal{D})$ should satisfy the following properties.

1) $DR(\mathcal{D}) = 1 \iff f_a \geq f_d$.
2) If f_a is more, then the suitability of $\mathcal{D}$ (for use in an application) cannot be less i.e. $\frac{dDR(\mathcal{D})}{df_a} \geq 0 \quad \forall f_a > 0$.
3) For $f_a \ll f_d$, a small increase in f_a cannot increase the suitability of $\mathcal{D}$ significantly. Similarly, if $f_a < f_d$ but very close to f_d, then the suitability of D cannot be significantly less than the case when $f_a = f_d$. So, both for i) $f_a \ll f_d$ and ii) $f_a < f_d$ but very close to f_d, $\frac{dDR(\mathcal{D})}{df_a} \ll 1$.

Accordingly, $DR(\mathcal{D})$ is given by (14).

$$DR(\mathcal{D}) = \begin{cases} \dfrac{\sigma\left(\frac{f_a}{f_d} - 0.5\right) - \sigma(-0.5)}{\sigma(0.5) - \sigma(-0.5)} & \text{if } 0 \leq \dfrac{f_a}{f_d} < 1 \\ 1 & \text{if } \dfrac{f_a}{f_d} \geq 1 \end{cases} \tag{14}$$

where $\sigma(x)$ in (14) is the sigmoid function given by $\sigma(x) = \frac{1}{1+\exp(-x)}$ $\forall x \in \mathbb{R}$ where $\mathbb{R}$ is the set of real numbers. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets that are not suitable for use in an application (with respect to data measurement rate) is given by $NDR(\mathcal{D})$. Then $NDR(\mathcal{D})$ is given by (15).

$$NDR(\mathcal{D}) = 1 - DR(\mathcal{D}) \tag{15}$$

F. AMOUNT OF DATA

Amount of data usually represents the volume of data needed for a given power system study. Amount of data recorded over a time interval depends on the measurement rate of a device. For example, suppose measurement rate is 40 milliseconds for PMU data and 1 minute for SCADA data. Then a PMU dataset containing 12000 seconds of data actually corresponds to 300, 000 rows of data whereas a SCADA dataset containing 12000 seconds of data corresponds to 200 rows of data only. Most power system studies require a minimum amount of data depending on the context. Suppose the degree of membership of $\mathcal{D}$ in the set of datasets that are suitable for use in a given application (with respect to amount of data) is denoted by $AoD(\mathcal{D})$. Then $AoD(\mathcal{D})$ should satisfy properties similar to the properties satisfied by $DR(\mathcal{D})$ with f_a replaced by d_a (the actual number of rows of $\mathcal{D}$ excluding the duplicate data samples) and f_d replaced by d_d (the threshold (or minimum) number of rows required for using $\mathcal{D}$ in an application). So, the function used in (14) can be used as $AoD(\mathcal{D})$ with $\frac{f_a}{f_d}$ replaced by $\frac{d_a}{d_d}$. Hence $AoD(\mathcal{D})$ is given by (16).

$$AoD(\mathcal{D}) = \begin{cases} \dfrac{\sigma\left(\frac{d_a}{d_d} - 0.5\right) - \sigma(-0.5)}{\sigma(0.5) - \sigma(-0.5)} & \text{if } 0 \leq \dfrac{d_a}{d_d} < 1 \\ 1 & \text{if } \dfrac{d_a}{d_d} \geq 1 \end{cases} \tag{16}$$

Suppose the degree of membership of $\mathcal{D}$ in the set of datasets that are not suitable for use in an application (with respect to amount of data) is given by $NAoD(\mathcal{D})$. Then $NAoD(\mathcal{D})$ is given by (17).

$$NAoD(\mathcal{D}) = 1 - AoD(\mathcal{D}) \tag{17}$$

G. INTERPRETABILITY

A row of $\mathcal{D}$ can be called interpretable if measurement of each of the available attributes (i.e. attributes for which measurement is available) of that row is interpretable. A measurement (or value) can be called interpretable if it does not contain any illogical characters that prevent the measurement from being comprehended. For example, a line current magnitude of $23@d4$ amperes is non-interpretable due to the presence of the characters '@' and 'd'. It is to be noted that a row may not have the measurements of all the n attributes. However, a partially incomplete row can also be interpretable if the measurement of each of the available attributes is interpretable. In other words, unavailability of some of the measurements should not influence the degree of interpretability of a row. Moreover, a measurement that is interpretable may not be accurate. On the other hand, an accurate measurement must always be interpretable. Accordingly, $\mathcal{D}$ can be called interpretable if each of the m rows of $\mathcal{D}$ is interpretable. Each of the x_k measurements available in row k may not be interpretable. In such a situation, row k can neither be called interpretable nor be called non-interpretable. In other words, row k is both partially interpretable and partially non-interpretable. Suppose z_k represents the number of interpretable measurements of row k $\forall k \in \{1, 2, \ldots, m\}$. Then $y_k \leq z_k \leq x_k$. In this section, membership functions are proposed to represent the degree of membership of a row of $\mathcal{D}$ in the set of interpretable and non-interpretable rows. Suppose the degree of membership of row k in the set of interpretable rows is denoted by $I(k)$ $\forall k \in \{1, \ldots, m\}$. Then $I(k)$ should satisfy the following properties.

1) For row k, $I(k) = 1 \iff z_k = x_k$ $\forall k \in \{1, 2, \ldots, m\}$
2) For row k, $I(k) = 0 \iff z_k = 0$ $\forall k \in \{1, 2, \ldots, m\}$
3) Between two rows i and j (where $i \neq j$) having $x_i = x_j$, $I(i) > I(j) \iff z_i > z_j$
4) Between two rows i and j (where $i \neq j$) having $z_i = z_j$, $I(i) > I(j) \iff x_i < x_j$

Accordingly, $I(k)$ is given by (18) $\forall k \in \{1, 2, \ldots, m\}$.

$$I(k) = \left(\frac{z_k}{x_k}\right)^s \tag{18}$$

where $s \in \mathbb{Z}_{>0}$. The degree of membership of row k in the set of non-interpretable rows is equal to $1 - I(k)$. Thus the resulting fuzzy sets representing the set of interpretable and non-interpretable rows are $\{(k, I(k))\ \forall k \in \{1, \ldots, m\}\}$ and $\{(k, 1 - I(k))\ \forall k \in \{1, \ldots, m\}\}$ respectively. From (4) and (18), it can be seen that if the value of q is chosen equal to the value of s, then $I(k) \geq A(k)$ $\forall k \in \{1, \ldots, m\}$.

The degree of membership of $\mathcal{D}$ in the set of interpretable datasets should be influenced by the degree of membership (in the set of interpretable rows) of each row of $\mathcal{D}$. Suppose r_i represents the number of rows of $\mathcal{D}$ having atleast one non-interpretable attribute each. Suppose there exist two datasets where each of the datasets has m rows and the same total (or sum) of the degree of membership of all the rows (in the set of interpretable rows). Then, the dataset having a lower value of r_i should be assigned a higher degree of membership (in the set of interpretable datasets) than the other. Suppose $I(\mathcal{D})$ and $NI(\mathcal{D})$ represent the degree of membership of $\mathcal{D}$ in the set of interpretable and non-interpretable datasets respectively. Then $I(\mathcal{D})$ and $NI(\mathcal{D})$ are given by (19) and (20) respectively.

$$I(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} I(k)\right\} - \frac{r_i}{m} + 1\right] \tag{19}$$

$$NI(\mathcal{D}) = \frac{1}{2}\left[\left\{\frac{1}{m}\sum_{k=1}^{m} (1 - I(k))\right\} + \frac{r_i}{m}\right] \tag{20}$$

The resulting fuzzy sets representing the set of interpretable and non-interpretable datasets are $\{(\mathcal{D}, I(\mathcal{D}))\ \forall \mathcal{D} \in U\}$ and $\{(\mathcal{D}, NI(\mathcal{D}))\ \forall \mathcal{D} \in U\}$ respectively. Both $I(\mathcal{D})$ and $NI(\mathcal{D})$ can be nonzero for a given dataset $\mathcal{D}$. In such situation, appropriate linguistic label can be used to describe the interpretability of $\mathcal{D}$. For example, if $I(\mathcal{D}) \gg NI(\mathcal{D})$, then it can be said that $\mathcal{D}$ has HIGH interpretability.

H. AVAILABILITY

Availability of a dataset usually represents the ease of accessing the dataset. Suppose a user wants to access dataset $\mathcal{D}$ for using it in an application. The user requests for $\mathcal{D}$ at time $t = 0$ on an online data platform. Suppose t_{dl} and t_{del} represent the deadline (i.e. the time within which the user needs access to $\mathcal{D}$) and the time of delivery of $\mathcal{D}$ to the user respectively. Suppose $Av(\mathcal{D})$ represents the degree of membership of $\mathcal{D}$ in the set of available datasets. Then $Av(\mathcal{D})$ should satisfy the following properties.

1) $Av(\mathcal{D}) = 1 \iff t_{\text{del}} \leq t_{\text{dl}}$
2) If $t_{\text{del}} > t_{\text{dl}}$, then $Av(\mathcal{D}) < 1$.
3) Suppose t_{del1} and t_{del2} be two delivery time instants such that $t_{\text{del2}} \geq t_{\text{del1}} \geq t_{\text{dl}}$. Then $Av(\mathcal{D})$ cannot be more when $t_{\text{del}} = t_{\text{del2}}$ compared to the case when $t_{\text{del}} = t_{\text{del1}}$ i.e. $\frac{dAv(\mathcal{D})}{dt_{\text{del}}} < 0\ \forall t_{\text{del}} > t_{\text{dl}}$. Moreover, $Av(\mathcal{D}) \to 0$ as $(t_{\text{del}} - t_{\text{dl}}) \to \infty$.

Accordingly, $Av(\mathcal{D})$ is proposed in (21).

$$Av(\mathcal{D}) = \begin{cases} 1 & \text{if } t_{del} \leq t_{\text{dl}} \\ \exp\left(-k\left(1 - \frac{t_{dl}}{t_{\text{del}}}\right)\right) & \text{if } t_{del} > t_{\text{dl}} \end{cases} \tag{21}$$

where $k \in \mathbb{Z}_{>0}$. Suppose the degree of membership of $\mathcal{D}$ in the set of unavailable datasets is given by $NAv(\mathcal{D})$. Then $NAv(\mathcal{D})$ is given by (22).

$$NAv(\mathcal{D}) = 1 - Av(\mathcal{D}) \tag{22}$$

Increasing k will significantly reduce the $Av(\mathcal{D})$ even if t_{del} is beyond t_{dl} by a small margin. A high value of k is recommended when even a small delay beyond the deadline is undesirable. For example, the utility usually uses consumption data of a block (typically consisting of 15 minutes) to estimate the consumption in the following block. This information (i.e. the predicted consumption in the next block) is then passed to the generating stations to control the generation for the next block. All these activities should be completed before the onset of the next block. As a result, the utility should get the consumption data of a block as quickly as possible so that all other activities can be completed on time. In such situation, even a small delay in receiving the consumption data of the current block is undesirable and can cost the utility a lot. From the foregoing discussion, it is seen that data availability is a very important factor for determining the data usability and the data quality. Unless data are available within the deadline, the usefulness of the data may reduce significantly.

IV. SUMMARY OF THE PROPOSED METHOD

The working of the proposed data quality assessment method is summarized in the form of a block diagram as shown in Fig. 1. For quality assessment of a given smart grid dataset $\mathcal{D}$, the dataset $\mathcal{D}$ is first fuzzified with the help of the membership functions proposed in section III. The proposed membership functions represent the degree of membership of $\mathcal{D}$ in various fuzzy sets where each fuzzy set essentially captures (or represents) a particular data quality dimension. Based on the value of the membership functions, appropriate fuzzy labels are then used for fuzzy assessment of each data quality dimension. In this paper, four fuzzy labels are considered for quality assessment namely HIGH, SLIGHTLY HIGH, SLIGHTLY LOW and LOW. The label HIGH is used when the value of a membership function (representing a particular data quality dimension) falls in the interval (0.75,1). For example, if $C(\mathcal{D}) > 0.75$, then it can be said that the dataset $\mathcal{D}$ has HIGH completeness. Similarly, the label SLIGHTLY HIGH is used for value of a membership function belonging to the interval (0.5, 0.75). The labels SLIGHTLY LOW and LOW are used for value of a membership function belonging to the interval (0.25, 0.5) and (0, 0.25) respectively. As shown in Fig. 1, a fuzzy assessment is made for each data quality dimension by assigning a particular fuzzy label based on the value of the membership functions associated with the dimensions. Usage of fuzzy (or linguistic) labels results in fuzzy assessment of each of the data quality dimensions.

V. RESULTS

In this section, the performance of the proposed membership functions is demonstrated by evaluating them on (i) the line current and the real power measurements obtained from the power flow analysis of the standard IEEE 14 bus system and (ii) the voltage, line current and the real power measurements obtained from the power flow analysis of an actual 34 node feeder located in Arizona. In addition, the measurements are also polluted with noise to test the sensitivity of the proposed membership functions. The membership functions are also used for comparing the quality of practical SCADA and PMU measurements of the Southern Regional Grid of India.

A. VALIDATING THE PROPOSED MEMBERSHIP FUNCTIONS ON IEEE 14 BUS SYSTEM

In this section, the proposed membership functions are evaluated on various datasets to assess their quality. At first, the power flow analysis of the standard IEEE 14 bus system (shown in Fig. 2) is carried out at two different time instants. It is assumed that at both the time instants, the system conditions (like the load at each bus, the number of transmission lines, the number of generators etc.) are identical. Hence the results obtained from the power flow analysis are identical at both the time instants. Then at every bus, the real power generation and the real power load are recorded (at each of the two time instants). As a result, 28 measurements are recorded for each time instant. Moreover, the sending end real

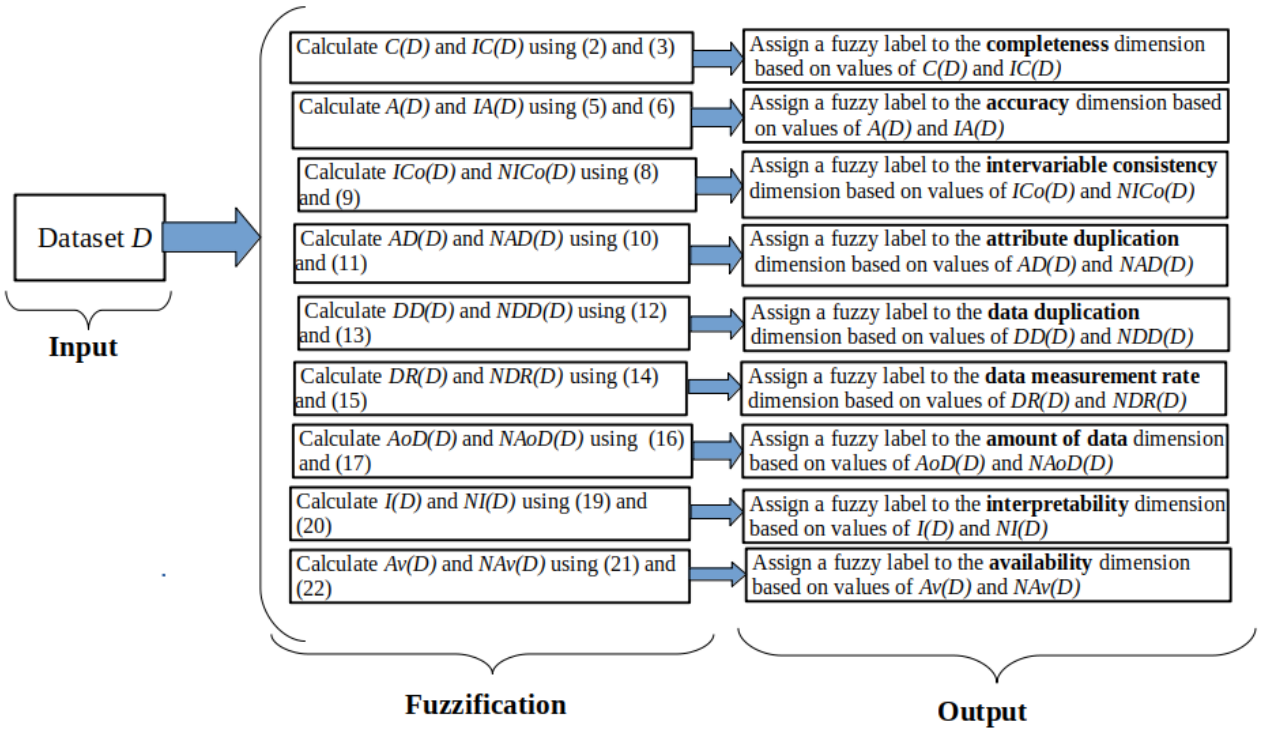

FIGURE 1. Working of the proposed data quality assessment method.

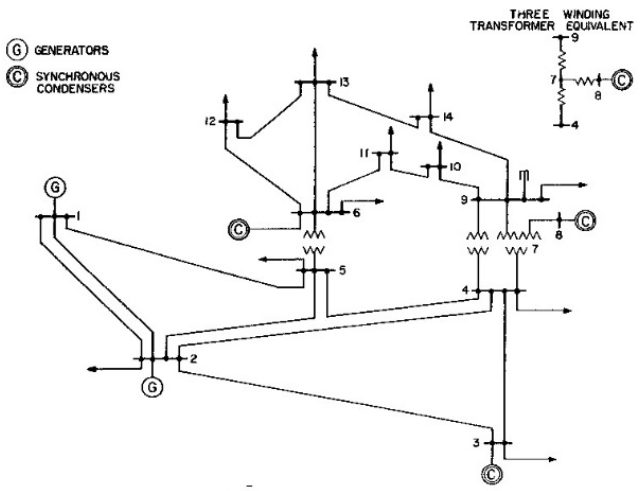

FIGURE 2. IEEE 14 bus test system [36].

power and the receiving end real power are also calculated (at each of the two time instants) for 20 different transmission lines in the system. As a result, a total of 40 measurements are recorded for each time instant. The line current is also calculated (at each of the two time instants) and recorded for each of the 20 transmission lines (that are considered for this study), resulting in 20 current measurements for each time instant. So, the total number of measurements recorded (at each time instant) for quality assessment is equal to 88. So $n = 88$. Moreover, measurements are recorded corresponding to two different time instants. So $m = 2$. So in this section, it is assumed that a dataset contains measurement of 88 power system variables where each variable is measured at two different time instants. Hence the dataset has 2 rows where each row has measurement of 88 power system variables. For validation of the proposed membership functions, the following cases are considered.

- Case 1: Each row of the dataset contains all the 88 measurements obtained from power flow analysis of the standard IEEE 14 bus system.
- Case 2: In this case, the 20 line current measurements are eliminated from the first row only. The remaining measurements are kept unchanged. As a result, the new dataset contains 68 measurements in first row and 88 measurements in the second row.
- Case 3: In this case, 10 line current measurements are eliminated from each of the two rows of the dataset. The remaining measurements are kept unchanged. As a result, the new dataset contains 78 measurements in each of the rows.
- Case 4: In this case, each of the 20 line current measurements of the first row is corrupted by adding zero mean unit variance Gaussian noise. It is to be noted that adding Gaussian noise does not make the measurements non-interpretable. However, the rest of the measurements are kept unchanged (compared to the measurements contained in dataset of Case 1). As a result, the new

dataset contains 88 measurements in each row. However, the line current measurements in the first row do not represent the original line current measurements (i.e. the line current measurements contained in first row of the dataset of Case 1).

- Case 5: In this case, 10 line current measurements are corrupted (by adding Gaussian noise) in both the rows. However, the rest of the measurements are kept unchanged (compared to the measurements contained in dataset of Case 1). As a result, the new dataset contains 88 measurements in each row. However, there are 10 line current measurements in each of the rows that differ from the original line current measurements (i.e. the line current measurements contained in dataset of Case 1).
- Case 6: In the first row, the real power generation data and the real power load data are modified at each bus. The rest of the measurements are kept unchanged (compared to the measurements contained in dataset of Case 1). Moreover, the data modification is carried out in a way that ensures that the algebraic sum of real power measurements is equal to zero at every bus.
- Case 7: In this case, each of the 20 line currents (i.e. the line current measurements contained in dataset of Case 1) is repeated in a way such that each line current measurement appears twice in each row of the new dataset. As a result, the total number of measurements in each row of the new dataset is 108.
- Case 8: In this case, certain illogical characters (like '$' and '@') are incorporated into the value of each of the 20 transmission line currents of the first row, making them non-interpretable. However, the remaining measurements are kept unchanged (compared to the measurements contained in dataset of Case 1).
- Case 9: In this case, certain illogical characters (like '$' and '@') are incorporated into the value of 10 transmission line currents in each of the two rows, making them non-interpretable. However, the remaining measurements are kept unchanged (compared to the measurements contained in dataset of Case 1).

For each of the aforementioned cases, the proposed membership functions are calculated. For calculating the membership functions, value of various parameters are required such as value of p in (1), value of q in (4), value of s in (18) etc. The choice of the parameters (required for calculating the membership functions) depends on the context. For example, suppose, for using a dataset in an application, it is required that the number of erroneous measurements in the dataset should be very small. So to assess the quality of dataset in such a situation, a high value of q should be chosen. This is because, if the value of $A(\mathcal{D})$ turns out to be high (i.e. close to 1) even after choosing a high value of q, then it can be concluded that the dataset has very few erroneous measurements. However, if a low value of q is used, then the value of $A(\mathcal{D})$ can still be significantly high even in the presence of a large number of erroneous measurements. As a result, the choice of the parameters is highly dependent on the context. In this paper, the parameter p (in (1)) is taken as 2. The parameter q (in (4)) is taken as 2. The parameter s (in (18)) is taken as 2. The parameters a and b (in (10) and (12)) are taken as 1 and 1 respectively. The weight vector w in (8) is chosen such that each coordinate is equal to 1. The obtained results are summarized in Table 1. It is seen from Table 1 that the value of the proposed membership functions differs among the nine cases that represent nine different datasets. This indicates that the nine datasets differ in their quality.

In Case 1, the dataset contains all the 88 measurements that are accurate and interpretable. There is no missing measurement. Moreover, the measurements also satisfy the system equations which ensures that the measurements are inter-consistent as well. A total of 34 equations is considered for checking consistency among the measurements. Out of the 34 equations, 14 equations are used to check whether the algebraic sum of the real power measurements (entering/leaving a particular bus) is zero or not at each of the 14 buses of the system. The remaining 20 equations are used to check whether the difference between the sending end and the receiving end power measurements equals the transmission loss or not in each of the 20 transmission lines. In Case 1, all the measurements are obtained from the power flow analysis carried out on the standard IEEE 14 bus system. Hence, all the measurements are accurate. These measurements also satisfy the 34 equations used for checking consistency among the measurements. Hence, the dataset in Case 1 has HIGH completeness, HIGH accuracy, HIGH interpretability, HIGH inter-variable consistency and LOW duplication.

In Case 2, the dataset does not have the 20 current measurements in the first row. However, the second row is complete. As a result, the dataset has SLIGHTLY HIGH completeness with value of r_c equal to 1. However, the available measurements are both accurate and interpretable. Moreover, the available measurements also satisfy the system equations. Out of 34 equations, 14 equations have been validated for the first row. The remaining 20 equations (that check whether the difference between sending and receiving end power is equal to the transmission loss or not) could not be validated for the first row due to the unavailability of the line current measurements in the first row. As a result, consistency of the available measurements of the first row is checked only with respect to 14 equations in this case. Since the available measurements satisfy the 14 equations, so the available measurements of the first row are inter-variable consistent. For the second row, all the 34 equations have been validated. Hence the dataset in case 2 has SLIGHTLY HIGH completeness, HIGH accuracy, HIGH interpretability, HIGH inter-variable consistency and LOW duplication.

In Case 3, the dataset does not contain 10 current measurements in both the first and the second row. Although the total number of missing entries is 20 in both Case 2 and Case 3, still the dataset in Case 3 is less complete than that of Case 2 as seen from Table 1. This is because, in this case, $r_c = 2$. However, the available measurements are both accurate and

TABLE 1. Quality assessment of line current and real power flow measurements of the standard IEEE 14 bus system.

Cases	Dimensions									
	Completeness		Accuracy		Inter-variable Consistency		Duplication		Interpretability	
	Our Method	[32]	Our Method	[32]	Our Method	[32]	Our Method	[32]	Our Method	[32]
Case 1	$C = 1$ $IC = 0$	1	$A = 1$ $IA = 0$	1	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 2	$C = 0.6492$ $IC = 0.3508$	0.8863	$A = 1$ $IA = 0$	1	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 3	$C = 0.3928$ $IC = 0.6072$	0.8863	$A = 1$ $IA = 0$	1	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 4	$C = 1$ $IC = 0$	1	$A = 0.6492$ $IA = 0.3508$	0.8863	$ICo = 0.67$ $NICo = 0.33$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 5	$C = 1$ $IC = 0$	1	$A = 0.3928$ $IA = 0.6072$	0.8863	$ICo = 0.36$ $NICo = 0.64$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 6	$C = 1$ $IC = 0$	1	$A = 0.6162$ $IA = 0.3838$	0.8409	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 7	$C = 1$ $IC = 0$	1	$A = 1$ $IA = 0$	1	$ICo = 1$ $NICo = 0$	–	$AD = 0.0841$ $NAD = 0.9159$ $DD = 0$ $NDD = 1$	–	$I = 1$ $NI = 0$	1
Case 8	$C = 1$ $IC = 0$	1	$A = 0.6492$ $IA = 0.3508$	0.8863	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 0.6492$ $NI = 0.3508$	0.8863
Case 9	$C = 1$ $IC = 0$	1	$A = 0.3928$ $IA = 0.6072$	0.8863	$ICo = 1$ $NICo = 0$	–	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	–	$I = 0.3928$ $NI = 0.6072$	0.8863

interpretable. In this case, out of 34 equations, 24 equations have been validated for the measurements of both the first and the second row. The remaining 10 equations could not be validated due to the unavailability of 10 line current measurements. The available measurements satisfy the 24 equations. So the available measurements are also inter-variable consistent. So, in this case, the dataset has SLIGHTLY LOW completeness, HIGH accuracy, HIGH interpretability, HIGH inter-variable consistency and LOW duplication.

In Case 4, all the measurements are available. However, the line current measurements of the first row are modified with noise. As a result, the current measurements of the first row are erroneous and inaccurate. Hence the dataset has SLIGHTLY HIGH accuracy in this case. Moreover, the inaccurate current measurements of the first row do not satisfy the 20 transmission line based system equations (that check whether the difference between sending and receiving end power is equal to the transmission loss or not). So, the measurements of the first row are less consistent among each other. However, the second row contains inter-consistent measurements. So, in this case, the dataset has HIGH completeness, SLIGHTLY HIGH accuracy, SLIGHTLY HIGH inter-variable consistency, HIGH interpretability and LOW duplication.

In Case 5, all the measurements are available. However, 10 line current measurements of both first and second row are modified with noise which makes them inaccurate. Although the total number of inaccurate entries is 20 in both Case 4 and Case 5, still the dataset in Case 5 is less accurate than that of Case 4 as seen from Table 1. This is because, the value of r_a is 2 in this case whereas the value of r_a was 1 in Case 4. Moreover, in each row, the 10 inaccurate current measurements do not satisfy 10 transmission line based system equations (that check whether the difference between sending

and receiving end power of a transmission line is equal to the transmission loss or not). As a result, the measurements in each of the rows are less consistent among each other. So, in this case, the dataset has HIGH completeness, SLIGHTLY LOW accuracy, SLIGHTLY LOW inter-variable consistency, HIGH interpretability and LOW duplication.

In Case 6, the real power generation data and the real power load data are modified for each bus in the first row. Hence, in the first row, the power generation data and the power demand data for each bus are inaccurate. However, the data are modified in a way that ensures that the algebraic sum of the real power measurements is zero at each bus. As a result, the measurements of the first row continue to remain inter-consistent. On the other hand, the measurements of the second row are accurate and consistent among each other. Since measurements in both the rows are consistent, so the dataset has HIGH inter-variable consistency as shown in Table 1. However, the dataset has SLIGHTLY HIGH accuracy (with $r_a = 1$) due to the presence of erroneous measurements in the first row. Moreover, the dataset has HIGH completeness, HIGH interpretability and LOW duplication.

In Case 7, the measurement of each of the line currents appears twice in each row of the dataset. As a result, the dataset exhibits attribute duplication. However, none of the measurements is missing which makes the dataset complete. Moreover, all the measurements are accurate and interpretable. So, in this case, the dataset has HIGH completeness, HIGH accuracy, HIGH inter-variable consistency, HIGH interpretability and LOW duplication.

In Case 8, presence of illogical characters makes the current measurements of the first row non-interpretable. So the usefulness of the current measurements of the first row is reduced. However, the measurements of the second row are interpretable. As a result, the dataset has SLIGHTLY HIGH interpretability (with $r_i = 1$). Moreover, the non-interpretable current measurements of the first row cannot be accurate. Hence the dataset has poor quality in terms of accuracy as well. However, the dataset has good quality in terms of inter-variable consistency. This is because, inter-variable consistency is checked only among the measurements that are interpretable. In other words, presence of non-interpretable measurements does not affect the quality of the dataset in terms of inter-variable consistency. So, in this case, the dataset has HIGH completeness, SLIGHTLY HIGH accuracy, HIGH inter-variable consistency, SLIGHTLY HIGH interpretability and LOW duplication.

In Case 9, illogical characters are incorporated into the measurements of 10 line currents (in both the first and the second row) which makes such measurements non-interpretable. In this case, the usefulness of the current measurements of both the rows is reduced. As a result, the dataset has SLIGHTLY LOW interpretability in this case (with $r_i = 2$) compared to the dataset in Case 8. Moreover, the non-interpretable current measurements are also considered as inaccurate. Hence the dataset has SLIGHTLY LOW accuracy as well. Moreover, in this case, the dataset is less accurate compared to the dataset in Case 8. This is because the inaccurate and non-interpretable measurements are distributed in both the rows of the dataset. This case shows that it is possible for a dataset to have SLIGHTLY LOW accuracy and SLIGHTLY LOW interpretability but HIGH level of inter-variable consistency. In addition, the dataset has HIGH completeness and LOW duplication.

From the foregoing discussion, it is seen that a single quality dimension may not be sufficient for assessing the overall quality of a dataset. While a dataset may have good quality along a given quality dimension, the same dataset may have bad quality along another dimension. For example, a dataset may have HIGH accuracy but LOW completeness. This shall imply that most of the measurements is missing in the dataset. However, the available measurements are accurate. Similarly, a dataset may have HIGH inter-variable consistency but LOW accuracy. Hence, it is clear from the foregoing discussion that assessment of each of the quality dimensions is necessary for assessing the overall quality of a dataset. Moreover, fuzzy assessment of each of the data quality dimensions helps in better understanding of a dataset's quality.

B. VALIDATING THE PROPOSED MEMBERSHIP FUNCTIONS ON AN ACTUAL 34 NODE FEEDER

In this section, the proposed membership functions are evaluated on a dataset containing voltage, current and real power measurements obtained from power flow analysis of an actual 34 node feeder located in Arizona [37], with a nominal voltage of 24.9kV. The feeder is characterized by long and lightly loaded, two in-line regulators, an in-line transformer for short 4.16 kV section, unbalanced loading, and shunt capacitors as shown in Fig. 3. More details about the feeder can be found in [37]. For the purpose of quality assessment, the following measurements are considered.

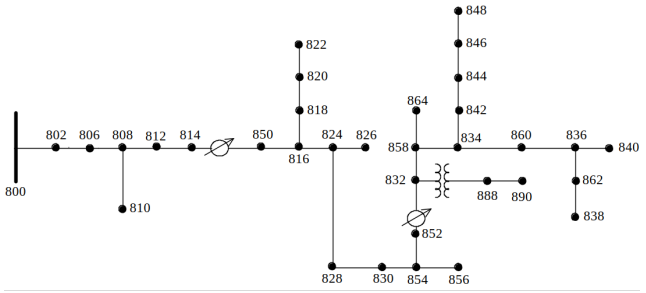

FIGURE 3. A real 34 node feeder located in Arizona [37].

- Voltage measurements (i.e. voltage magnitude and angle) of each of the 3 phases are recorded at node 802, 806, 808, 812, 814. So a total of 30 measurements are recorded.
- Line current measurements (i.e. magnitude and angle) are recorded for each of the 3 phases (or lines) connecting nodes 802 and 806, 806 and 808, 808 and 812, 812 and 814. So a total of 24 measurements are recorded.

- Transmission loss measurements are also recorded for each of the 3 lines (or phases) connecting nodes 802 and 806, 806 and 808, 808 and 812, 812 and 814. So a total of 12 measurements are recorded.

Thus a total of 66 measurements is considered in this study. Quality assessment of these measurements is performed under the following cases.

- Case 1: Zero mean zero variance Gaussian noise is added to each of the 66 measurements independently. In other words, no noise is added to the measurements.
- Case 2: Zero mean unit variance Gaussian noise is added to each of the measurements independently, except the line current measurements. In other words, the magnitude and angle of the line current measurements are not polluted with noise.
- Case 3: Zero mean Gaussian noise with variance equal to 10 is added to each of the measurements independently, except the line current measurements. In other words, the magnitude and angle of the line current measurements are not polluted with noise.
- Case 4: The voltage measurements are removed from the dataset in order to simulate a situation where a dataset contains missing measurements. However, the current measurements and the transmission loss measurements are kept unchanged. In other words, the current measurements and the transmission loss measurements in this case are same as that in Case 1.
- Case 5: The transmission loss measurements are repeated in a way such that each transmission loss measurement appears twice in the dataset. This case simulates a situation where a particular measurement appears more than once in a database. However, each of the current measurements and each of the voltage measurements appears only once.
- Case 6: Illogical characters (such as '$' and '@') are incorporated into each of the voltage measurements, making them non-interpretable. This case simulates a situation where the measurements in a database are corrupted and hence non-interpretable. However, the current measurements and the loss measurements are kept unchanged. In other words, the current measurements and the loss measurements in this case are same as that in Case 1.

For each of the aforementioned cases, the proposed membership functions are calculated in order to assess the quality of the measurements. A total of 24 equations is considered for checking the consistency among the measurements. Out of the 24 equations, 12 equations are used to check whether the difference between voltage of two consecutive nodes equals the line current scaled by the line impedance. This is done for each of the 3 phases (or lines) connecting two consecutive nodes. The remaining 12 equations are used to check whether the transmission loss in each phase (or line) equals the product of square of line current and line resistance. This is also done for each of the 3 phases (or lines) connecting two consecutive nodes. The value of the parameters p, q, s, a and b are taken as 2, 2, 2, 1 and 1 respectively. The weight vector w is chosen such that each coordinate is equal to 1. The obtained results are summarized in Table 2.

In Case 1, no noise was added to the measurements. Hence each of the measurements is accurate and interpretable. Moreover, the measurements satisfy the system equations. Hence $ICo = 1$. In addition, there is no duplication since $AD = 0$ and $DD = 0$. Moreover $C = 1$ indicates the absence of missing measurements.

In Case 2, zero mean unit variance Gaussian noise is added to each of the measurements (except the current measurements) independently. Addition of noise makes the measurements inaccurate. So, the value of A decreases with the addition of noise. Since noise is added independently to each of the measurements (except the current measurements), so the noisy voltage measurements and the noisy power measurements do not satisfy the system equations. So, the value of ICo also reduces with the addition of independent noise. However, addition of noise does not affect the completeness, duplication and the interpretability aspects of the data.

In Case 3, zero mean Gaussian noise with variance equal to 10 is added to each of the measurements (except the current measurements) independently. It can be seen from Table 2 that the value of ICo decreases with the increase in the variance of the added Gaussian noise. When the variance is increased by a factor of 10 (from Case 2 to Case 3), the value of ICo decreases by 71.5%. This is because, with increase in variance, the measurements fail to satisfy the system equations by a larger margin. Hence ICo decreases significantly with increase in variance of the added noise. However, it can be seen from Table 2 that increasing the variance of the added noise does not affect the value of A. This is because, the value of A depends on whether the measurements are erroneous or not. A measurement is erroneous if it contains noise irrespective of its variance. Hence, the value of A is unaffected by the variance of the added noise.

In Case 4, the voltage measurements were removed from the dataset. In other words, the voltage measurements were missing. As a result, the value of C has reduced indicating the presence of missing measurements. However, the available measurements are accurate and hence they satisfy the system equations. It is to be noted that in this case, out of 24 equations, only 12 equations have been validated. The remaining 12 equations (that check whether the difference between voltage of two consecutive nodes equals the line current scaled by the line impedance or not) could not be validated due to the absence of the voltage measurements. As a result, consistency is checked only with respect to 12 equations in this case. Since the available measurements satisfy the 12 equations, so the measurements are inter-variable consistent. In addition, $I = 1$ indicates that the measurements are interpretable.

In Case 5, each of the transmission loss measurements appears twice in the dataset. This indicates that the dataset suffers from attribute duplication. Hence the value of AD is nonzero. However, none of the measurements is missing.

TABLE 2. Quality assessment of voltage, current and real power measurements of a real 34 node feeder located in Arizona.

Cases	Dimensions				
	Completeness	Accuracy	Inter-variable Consistency	Duplication	Interpretability
Case 1	$C = 1$ $IC = 0$	$A = 1$ $IA = 0$	$ICo = 1$ $NICo = 0$	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	$I = 1$ $NI = 0$
Case 2	$C = 1$ $IC = 0$	$A = 0.0661$ $IA = 0.9339$	$ICo = 0.431$ $NICo = 0.569$	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	$I = 1$ $NI = 0$
Case 3	$C = 1$ $IC = 0$	$A = 0.0661$ $IA = 0.9339$	$ICo = 0.123$ $NICo = 0.877$	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	$I = 1$ $NI = 0$
Case 4	$C = 0.1487$ $IC = 0.8513$	$A = 1$ $IA = 0$	$ICo = 1$ $NICo = 0$	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	$I = 1$ $NI = 0$
Case 5	$C = 1$ $IC = 0$	$A = 1$ $IA = 0$	$ICo = 1$ $NICo = 0$	$AD = 0.0559$ $NAD = 0.9441$ $DD = 0$ $NDD = 1$	$I = 1$ $NI = 0$
Case 6	$C = 1$ $IC = 0$	$A = 0.1487$ $IA = 0.8513$	$ICo = 1$ $NICo = 0$	$AD = 0$ $NAD = 1$ $DD = 0$ $NDD = 1$	$I = 0.1487$ $NI = 0.8513$

In addition, the measurements are accurate, interpretable and hence they are inter-variable consistent.

In Case 6, illogical characters are incorporated into the voltage measurements which affects the interpretability of the dataset. Hence the value of *I* reduces. Non-interpretable measurements are also considered as inaccurate. Hence, the value of *A* also decreases in the presence of non-interpretable measurements. It has been mentioned in section V-A that presence of non-interpretable measurements does not affect the quality of a dataset in terms of inter-variable consistency. This is because consistency is checked among the interpretable measurements only. Hence the value of *ICo* is unaffected in the presence of non-interpretable measurements.

From the value of the membership functions shown in Table 2, the following inferences can be drawn.

- In Case 1, the dataset has HIGH completeness, HIGH accuracy, HIGH inter-variable consistency, LOW duplication and HIGH interpretability.
- In Case 2, the dataset has HIGH completeness, LOW accuracy, SLIGHTLY LOW inter-variable consistency, LOW duplication and HIGH interpretability.
- In Case 3, the dataset has HIGH completeness, LOW accuracy, LOW inter-variable consistency, LOW duplication and HIGH interpretability.
- In Case 4, the dataset has LOW completeness, HIGH accuracy, HIGH inter-variable consistency, LOW duplication and HIGH interpretability.
- In Case 5, the dataset has HIGH completeness, HIGH accuracy, HIGH inter-variable consistency, LOW duplication and HIGH interpretability.
- In Case 6, the dataset has HIGH completeness, LOW accuracy, HIGH inter-variable consistency, LOW duplication and LOW interpretability.

Real world measurements are often noisy because of the measurement noise associated with the measuring devices and the sensors. In addition, measurements often get stuck because of communication errors which leads to missing of various measurements from the database. Sometimes, multiple appearance of the same data are also observed in a smart grid database. Measurements may even get corrupted by illogical characters which can make them non-interpretable. As a result, a smart grid data quality assessment method should be capable of distinguishing between noisy and noise-free measurements. From Table 2, it can be concluded that the proposed data quality assessment method can effectively discriminate between datasets containing noisy and noise-free measurements. The proposed method can also detect missing or duplicate or non-interpretable entries. Hence, the applicability of the proposed method on real systems is justified.

C. COMPARISON WITH EXISTING METHODS

Table 3 shows the results obtained by comparing some of the existing data quality assessment methods. It is seen from Table 3 that none of the existing methods can measure the inter-variable consistency dimension. Inter-variable consistency is one of the most important quality indicators of smart grid measurements. This is because smart grid measurements are usually connected through the system equations. In other words, measurements of different power system variables

TABLE 3. Comparison of data quality assessment methods.

Data quality assessment methods	Completeness	Accuracy	Inter-variable Consistency	Duplication	Interpretability
[20]	✓	✓	✗	✗	✗
[21]	✗	✗	✗	✗	✗
[22]	✓	✓	✗	✗	✗
[24]	✗	✗	✗	✗	✗
[28]	✗	✗	✗	✗	✗
[29]	✗	✗	✗	✗	✗
[32]	✓	✓	✗	✗	✓
Proposed method	✓	✓	✓	✓	✓

are not independent and are governed by system equations. Hence, checking the consistency among the measurements is necessary for quality assessment of smart grid measurements. However, none of the existing methods can measure the consistency among the measurements of different power system variables. Similarly, duplication is one of the most relevant dimensions for smart grid data quality assessment. This is because presence of duplicate measurements may reduce the usefulness of a smart grid dataset, resulting in poor quality of the same. Hence, estimation of the level of duplication is necessary for assessing quality of a smart grid dataset. However, none of the existing methods can detect data duplication. Similarly, interpretability is also an important dimension in the context of data quality assessment. This is because presence of non-interpretable measurements in a dataset may adversely affect its usability in various applications. However, except [32], none of the existing methods can quantify the interpretability of the smart grid measurements. Although [21] and [24] have introduced various dimensions for assessing data quality, they have not provided any mathematical formulation for assessing the data quality dimensions. In other words, [21] and [24] have provided a theoretical description of several quality dimensions instead of quantifying them mathematically. Hence [21] and [24] cannot be used for mathematical assessment of the data quality dimensions considered in this paper. Literature [28] has proposed various quality dimensions for assessing the quality of VGI data. However, those dimensions are completely different from the quality dimensions considered in this paper. Basically the quality dimensions proposed in [28] are relevant only in the context of VGI data quality assessment. In other words, those dimensions are not applicable for quality assessment of smart grid data. In addition, [28] has not provided any metric for quantification of the quality dimensions considered in this paper. As a result, the method proposed in [28] cannot be used for assessing the data quality dimensions considered in this paper. Literature [29] has proposed quality assessment techniques only for boolean datasets (i.e. datasets containing binary variables). However, majority of the power system variables are not binary variables (for example, bus voltage, line current, power flow etc.). Hence the method proposed in [29] cannot be used for assessing the quality of smart grid datasets. Literature [32] has provided metrics for estimating some of the quality dimensions such as accuracy, completeness and interpretability. However, from Table 1 and Table 3, it is seen the method proposed in [32] fails to assess the quality of a dataset in terms of inter-variable consistency and duplication. Moreover, it is seen from Table 1 that if the approach proposed in [32] is used for quality assessment, then the completeness score remains unchanged in Case 2 and Case 3. Similarly, the accuracy score remains unchanged in Case 4 and Case 5 while the interpretability score remains unchanged in Case 8 and Case 9. This shows that the method proposed in [32] is not sensitive to the distribution of the missing measurements or the inaccurate measurements or the non-interpretable measurements. In other words, the method proposed in [32] fails to distinguish between datasets containing the same number of missing or inaccurate or non-interpretable measurements that are distributed in varying number of rows. Hence the existing data quality assessment methods have several limitations which make them unsuitable for quality assessment of smart grid data.

The approach proposed in this paper overcomes the limitations of the existing data quality assessment methods. Basically, unlike the existing methods, the proposed approach can quantify the data quality dimensions. As seen from Table 3, the approach proposed in this paper can be used for quantifying inter-variable consistency and duplication in addition to accuracy, completeness and interpretability. Moreover, as seen from Table 1, the approach adopted in this paper is sensitive to the distribution of the missing measurements or the inaccurate measurements or the non-interpretable measurements. In other words, the proposed approach can effectively discriminate between datasets containing the same number of missing or inaccurate or non-interpretable measurements that are distributed in varying number of rows. Hence, based on the foregoing discussion, it can be concluded that the proposed quality assessment approach offers a significant improvement over the existing data quality assessment methods.

D. ASSESSING QUALITY OF SCADA AND PMU DATA

The proposed membership functions are calculated on a database provided by the Power System Operation Corporation Ltd. (POSOCO), India. The database had several datasets comprising SCADA and PMU measurements of the Southern Regional Grid of India. SCADA datasets contained measurements at an interval of 1 minute while PMU datsets contained measurements at an interval of 40 milliseconds.

TABLE 4. Comparison of quality of SCADA and PMU data.

Dimensions	SCADA Datasets			PMU Datasets	
	Bus voltage data	Frequency data	Real power flow data	Data of 10.09.2013	Data of 28.05.2013
Completeness	$C = 0.933$	$C = 0.372$	$C = 0.933$	$C = 0.0193$	$C = 0.0807$
	$IC = 0.067$	$IC = 0.628$	$IC = 0.067$	$IC = 0.9807$	$IC = 0.9193$
Duplication	$AD = 0$	$AD = 0$	$AD = 0$	$AD = 0$	$AD = 0$
	$NAD = 1$	$NAD = 1$	$NAD = 1$	$NAD = 1$	$NAD = 1$
	$DD = 8.93 \times 10^{-7}$	$DD = 8.93 \times 10^{-7}$	$DD = 8.93 \times 10^{-7}$	$DD = 0$	$DD = 5.46 \times 10^{-4}$
	$NDD = 0.9999$	$NDD = 0.9999$	$NDD = 0.9999$	$NDD = 1$	$NDD = 0.99945$
Interpretability	$I = 1$	$I = 1$	$I = 1$	$I = 1$	$I = 1$
	$NI = 0$	$NI = 0$	$NI = 0$	$NI = 0$	$NI = 0$
Data Measurement Rate (Case 1: ($f_d = 5$ Hz))	$DR = 0.00343$	$DR = 0.00343$	$DR = 0.00343$	$DR = 1$	$DR = 1$
	$NDR = 0.99657$	$NDR = 0.99657$	$NDR = 0.99657$	$NDR = 0$	$NDR = 0$
Data Measurement Rate (Case 2: ($f_d = 1$ Hz))	$DR = 0.0162$	$DR = 0.0162$	$DR = 0.0162$	$DR = 1$	$DR = 1$
	$NDR = 0.9838$	$NDR = 0.9838$	$NDR = 0.9838$	$NDR = 0$	$NDR = 0$
Data Measurement Rate (Case 3: ($f_d = 0.033$ Hz))	$DR = 0.5$	$DR = 0.5$	$DR = 0.5$	$DR = 1$	$DR = 1$
	$NDR = 0.5$	$NDR = 0.5$	$NDR = 0.5$	$NDR = 0$	$NDR = 0$
Data Measurement Rate (Case 4: ($f_d = 0.0166$ Hz))	$DR = 1$	$DR = 1$	$DR = 1$	$DR = 1$	$DR = 1$
	$NDR = 0$	$NDR = 0$	$NDR = 0$	$NDR = 0$	$NDR = 0$
Data Measurement Rate (Case 5: ($f_d = 0.00833$ Hz))	$DR = 1$	$DR = 1$	$DR = 1$	$DR = 1$	$DR = 1$
	$NDR = 0$	$NDR = 0$	$NDR = 0$	$NDR = 0$	$NDR = 0$

The proposed membership functions are evaluated on three different SCADA datasets containing measurements of June 2016. One of the datasets had voltage measurements of 20 different buses. The other two datasets contained frequency measurements of the 20 buses and real power flow measurements in 20 different transmission lines. The membership functions are also calculated on two PMU datasets containing measurements of 10th September, 2013 and 28th May, 2013. The PMU datasets had frequency measurements and voltage measurements for 5 different buses. The obtained results are summarized in Table 4. From Table 4, it is seen that the degree of membership of the SCADA datasets in the set of complete datasets is higher than that of the PMU datasets. This implies that the SCADA datasets have better quality in terms of completeness. However, among the SCADA datasets, the voltage and the real power flow datasets have better completeness than the frequency dataset. No attribute is reported more than once in any of the considered datasets. As a result, the considered datasets do not suffer from attribute duplication. However, a small amount of data duplication is observed in both SCADA and PMU datasets. This is because the SCADA measurements corresponding to 12 a.m. were reported twice for each of the days in the month. Similarly, PMU measurements corresponding to an interval of 9 seconds were also reported twice in the dataset of 28th May, 2013. Neither the SCADA nor the PMU datasets contained illogical characters. As a result, each dataset has HIGH interpretability. For assessing the suitability of the datasets for use in an application with respect to data measurement rate, five different cases are considered as shown in Table 4. Each of the five cases represents a specific application with a particular value of f_d. It can be seen from Table 4 that the PMU datasets have $DR = 1$ in each of the five cases. This implies that the PMU datasets are suitable for use in each of the five applications. This is because sampling frequency of the considered PMU datasets is equal to $(1/40 \text{ milliseconds}) = 25$ Hz. However, in each of the five cases, $f_d < 25$ Hz. Hence it can be concluded that PMU datasets have HIGH suitability for use in various applications with respect to data measurement rate. SCADA datasets, on the other hand, have very low value of DR (i.e. $DR < 1$), especially in the first 3 cases. This implies that the SCADA datasets are not completely suitable for use in the first three applications. This is because, sampling frequency of the considered SCADA datasets is equal to $(1/1 \text{ minute}) = 0.0166$ Hz. As a result, in the first three cases, the value of f_d is higher than the actual sampling frequency of the SCADA datasets. Hence, the SCADA datasets have lower suitability for use in the first three applications. In other words, the SCADA datasets have LOW suitability in the first two cases, SLIGHTLY LOW suitability in the third case and HIGH suitability in the fourth and the fifth case. From the foregoing discussion, it can be concluded that PMU data are suitable for use in more number of applications compared to the SCADA datasets. The network topology and the network parameters were not available for the investigated SCADA

and PMU data. Due to the unavailability of the network parameters and the network topology, it is not possible to compare the quality of the SCADA and the PMU datasets on the basis of their accuracy and inter-variable consistency. Hence, results on accuracy and inter-variable consistency are not shown in Table 4.

VI. CONCLUSION

A novel fuzzy assessment method is proposed in this paper for assessing the quality of multivariate electrical measurements of smart grids. Novel membership functions are proposed for measuring the degree of membership of a smart grid dataset in various fuzzy sets that represent various dimensions of smart grid data quality. Based on the value of the membership functions, appropriate fuzzy labels are used for fuzzy assessment of data quality. The proposed membership functions are used to assess the quality of current and power measurements obtained from power flow analysis of the standard IEEE 14 bus system. The proposed method is also used to assess the quality of voltage, current and power measurements obtained from power flow analysis of an actual 34 node feeder located in Arizona. In addition, the measurements are also modified with Gaussian noise to test the applicability of the proposed method for quality assessment of real world measurements. The obtained results show that when the variance increases by a factor of 10, the consistency among the measurements decreases by 71.5%. The obtained results also show that the proposed method can detect the presence of non-interpretable or missing measurements along with data duplication. Unlike the existing methods, the proposed membership functions are found to be sensitive to the distribution of missing, inaccurate and non-interpretable measurements in a given dataset. The proposed method is also tested on practical SCADA and PMU measurements. It is found that PMU datasets are relatively incomplete compared to SCADA datasets. In addition, the obtained results indicate the presence of duplicate data in both SCADA and PMU datasets. Moreover, the obtained results show that the PMU datasets have better usability (or suitability for use) in a wide variety of applications compared to the SCADA datasets. The proposed membership functions can be used for developing fuzzy inference systems in various data quality driven smart grid applications.

ACKNOWLEDGEMENT

The authors would like to thank the Power System Operation Corporation Ltd., for providing the SCADA and PMU measurements of the Southern Regional Grid of India.

REFERENCES

[1] Y. Gao, B. Foggo, and N. Yu, "A physically inspired data-driven model for electricity theft detection with smart meter data," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5076–5088, Sep. 2019.

[2] W. Zhou, O. Ardakanian, H.-T. Zhang, and Y. Yuan, "Bayesian learning-based harmonic state estimation in distribution systems with smart meter and DPMU data," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 832–845, Jan. 2020.

[3] S. Chevalier, P. Vorobev, and K. Turitsyn, "A Bayesian approach to forced oscillation source location given uncertain generator parameters," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1641–1649, Mar. 2019.

[4] G. L. Kusic and D. L. Garrison, "Measurement of transmission line parameters from SCADA data," in *Proc. IEEE PES Power Syst. Conf. Expo.*, Oct. 2004, pp. 440–445.

[5] S. Chakraborty and S. Das, "Application of smart meters in high impedance fault detection on distribution systems," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3465–3473, May 2019.

[6] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *Proc. IREP Symp. Bulk Power Syst. Dyn. Control-IX Optim., Secur. Control Emerg. Power Grid*, Aug. 2013, pp. 1–9.

[7] S. Das and P. S. N. Rao, "Understanding power system behavior through mining archived operational data," *Int. J. Emerg. Electr. Power Syst.*, vol. 10, no. 1, pp. 1–19, Apr. 2009.

[8] S. Das and P. S. N. Rao, "Principal component analysis based compression scheme for power system steady state operational data," in *Proc. Innov. Smart Grid Technol.-India (ISGT India)*, Dec. 2011, pp. 95–100.

[9] S. Das and P. S. N. Rao, "Arithmetic coding based lossless compression schemes for power system steady state operational data," *Int. J. Electr. Power Energy Syst.*, vol. 43, no. 1, pp. 47–53, Dec. 2012.

[10] W. Chen, K. Zhou, S. Yang, and C. Wu, "Data quality of electricity consumption data in a smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 75, pp. 98–105, Aug. 2017.

[11] Y. Guo, Y. Zhang, and A. K. Srivastava, "Data-quality aware state estimation in three-phase unbalanced active distribution system," in *Proc. IEEE Ind. Appl. Soc. Annu. Meeting (IAS)*, Sep. 2018, pp. 1–7.

[12] M. Allalouf, G. Gershinsky, L. Lewin-Eytan, and J. Naor, "Smart grid network optimization: Data-quality-aware volume reduction," *IEEE Syst. J.*, vol. 8, no. 2, pp. 450–460, Jun. 2014.

[13] J.-W. Kang, L. Xie, and D.-H. Choi, "Impact of data quality in home energy management system on distribution system state estimation," *IEEE Access*, vol. 6, pp. 11024–11037, 2018.

[14] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, Mar. 1996.

[15] G. K. Tayi and D. P. Ballou, "Examining data quality," *Commun. ACM*, vol. 41, no. 2, pp. 54–57, Feb. 1998.

[16] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[17] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[18] C. Cappiello, C. Francalanci, and B. Pernici, "Data quality assessment from the user's perspective," in *Proc. Int. Workshop Inf. Qual. Inf. Syst.*, 2004, pp. 68–73.

[19] S. Watts, G. Shankaranarayanan, and A. Even, "Data quality assessment in context: A cognitive perspective," *Decis. Support Syst.*, vol. 48, no. 1, pp. 202–211, Dec. 2009.

[20] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jeremic, "A metrics-driven approach for quality assessment of linked open data," *J. Theor. Appl. Electron. commerce Res.*, vol. 9, no. 2, pp. 64–79, 2014.

[21] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, and P. Hitzler, "Quality assessment methodologies for linked open data," *Semantic Web J.*, vol. 1, no. 1, pp. 1–5, 2013.

[22] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big data quality: A quality dimensions evaluation," in *Proc. Int. IEEE Conf. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People, Smart World Congr. (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, Jul. 2016, pp. 759–765.

[23] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Future Gener. Comput. Syst.*, vol. 89, pp. 548–562, Dec. 2018.

[24] E. A.-M. Al-Masri and Y. Bai, "Invited paper: A service-oriented approach for assessing the quality of data for the Internet of Things," in *Proc. IEEE Int. Conf. Service-Oriented Syst. Eng. (SOSE)*, Apr. 2019, pp. 9–97.

[25] A. P. Reimer, A. Milinovich, and E. A. Madigan, "Data quality assessment framework to assess electronic medical record data for use in research," *Int. J. Med. Informat.*, vol. 90, pp. 40–47, Jun. 2016.

[26] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Comput. Biol. Med.*, vol. 107, pp. 270–283, Apr. 2019.

[27] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on OpenStreetMap," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 4, pp. 700–719, Apr. 2014.
[28] C. Fonte, V. Antoniou, L. Bastin, J. Estima, J. J. Arsanjani, J. L. Bayas, L. See, and R. Vatseva, "Assessing VGI data quality," Sep. 2017, pp. 137–163, doi: 10.5334/bbf.g.
[29] N. Raviv, S. Jain, and J. Bruck, "What is the value of data? On mathematical methods for data quality estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2825–2830.
[30] J. An, J. Cheng, X. Gui, W. Zhang, D. Liang, R. Gui, L. Jiang, and D. Liao, "A lightweight blockchain-based model for data quality assessment in crowdsensing," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 84–97, Feb. 2020.
[31] D. Peng, F. Wu, and G. Chen, "Data quality guided incentive mechanism design for crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 17, no. 2, pp. 307–319, Feb. 2018.
[32] A. Radhakrishnan and S. Das, "Quality assessment of smart grid data," in *Proc. 20th Nat. Power Syst. Conf. (NPSC)*, Dec. 2018, pp. 1–6.
[33] T. Liu, Y. Liu, J. Liu, Y. Yang, G. A. Taylor, and Z. Huang, "Multi-indicator inference scheme for fuzzy assessment of power system transient stability," *CSEE J. Power Energy Syst.*, vol. 2, no. 3, pp. 1–9, 2016.
[34] J. M. Mendel, "Fuzzy logic systems for engineering: A tutorial," *Proc. IEEE*, vol. 83, no. 3, pp. 345–377, Mar. 1995.
[35] M. R. H. M. Adnan, A. Sarkheyli, A. M. Zain, and H. Haron, "Fuzzy logic for modeling machining process: A review," *Artif. Intell. Rev.*, vol. 43, no. 3, pp. 345–379, Mar. 2015.
[36] *Power Systems Test Case Archive*. Accessed: Jun. 10, 2021. [Online]. Available: https://labs.ece.uw.edu/pstca/pf14/pg_tca14bus.htm
[37] *Resources*. Accessed: Jun. 10, 2021. [Online]. Available: https://site.ieee.org/pes-testfeeders/resources/

SOUMYAJIT GANGOPADHYAY (Graduate Student Member, IEEE) received the B.Tech. degree in electrical and electronics engineering from the National Institute of Technology Calicut, Kozhikode, India, in 2017. He is currently pursuing the M.Tech. (research) degree with the Department of Electrical Engineering, Indian Institute of Science, Bengaluru, India. His research interests include smart grid data analytics, distribution system optimization, and consumer behavior modeling along with applications of machine learning and deep learning algorithms in electrical power systems.

SARASIJ DAS (Senior Member, IEEE) received the Ph.D. degree from the University of Western Ontario, London, ON, Canada, in 2014. Previously, he was with Global Technology Centre, Schneider Electric India Private Ltd., Bengaluru, India; and Power Research and Development Consultants Private Ltd., Bengaluru. He is currently working as an Assistant Professor with the Department of Electrical Engineering, Indian Institute of Science, Bengaluru. His research interests include smart grid data analytics, and renewable energy along with analysis, protection, and monitoring of electrical power systems.