Stream Data Cleaning for Dynamic Line Rating Application

Authors:

Hassan M. Nemati, A. Laso, M. Manana, Anita Sant'Anna, S?awomir Nowaczyk

Date Submitted: 2018-09-21

Keywords: data mining, stream data cleaning, dynamic line rating, smart grids

Abstract:

The maximum current that an overhead transmission line can continuously carry depends on external weather conditions, most commonly obtained from real-time streaming weather sensors. The accuracy of the sensor data is very important in order to avoid problems such as overheating. Furthermore, faulty sensor readings may cause operators to limit or even stop the energy production from renewable sources in radial networks. This paper presents a method for detecting and replacing sequences of consecutive faulty data originating from streaming weather sensors. The method is based on a combination of (a) a set of constraints obtained from derivatives in consecutive data, and (b) association rules that are automatically generated from historical data. In smart grids, a large amount of historical data from different weather stations are available but rarely used. In this work, we show that mining and analyzing this historical data provides valuable information that can be used for detecting and replacing faulty sensor readings. We compare the result of the proposed method against the exponentially weighted moving average and vector autoregression models. Experiments on data sets with real and synthetic errors demonstrate the good performance of the proposed method for monitoring weather sensors.

DOI of Published Version: https://doi.org/10.3390/en11082007 energies

Article

Stream Data Cleaning for Dynamic Line Rating Application

Hassan M. Nemati [1,†], A. Laso [2], M. Manana [2], Anita Sant'Anna [1] and Sławomir Nowaczyk [1,*]

[1] Center for Applied Intelligent Systems Research, Halmstad University, SE-30118 Halmstad, Sweden; hassan.nemati@hh.se (H.M.N.); anita.santanna@hh.se (A.S.)
[2] Department of Electrical and Energy Engineering, University of Cantabria, 39005 Santander, Spain; lasoal@unican.es (A.L.); mananam@unican.es (M.M.)
* Correspondence: slawomir.nowaczyk@hh.se; Tel.: +46-35-16-7930
† Current address: Center for Applied Intelligent Systems Research, Halmstad University, P.O. Box 823, 30118 Halmstad, Sweden.

Received: 16 July 2018; Accepted: 31 July 2018; Published: 2 August 2018

Abstract: The maximum current that an overhead transmission line can continuously carry depends on external weather conditions, most commonly obtained from real-time streaming weather sensors. The accuracy of the sensor data is very important in order to avoid problems such as overheating. Furthermore, faulty sensor readings may cause operators to limit or even stop the energy production from renewable sources in radial networks. This paper presents a method for detecting and replacing sequences of consecutive faulty data originating from streaming weather sensors. The method is based on a combination of (a) a set of constraints obtained from derivatives in consecutive data, and (b) association rules that are automatically generated from historical data. In smart grids, a large amount of historical data from different weather stations are available but rarely used. In this work, we show that mining and analyzing this historical data provides valuable information that can be used for detecting and replacing faulty sensor readings. We compare the result of the proposed method against the exponentially weighted moving average and vector autoregression models. Experiments on data sets with real and synthetic errors demonstrate the good performance of the proposed method for monitoring weather sensors.

Keywords: smart grids; dynamic line rating; stream data cleaning; data mining

1. Introduction

In smart grids, large renewable sources are integrated into the grid by using overhead transmission lines. As renewable sources have a variable production capacity, this often causes the transmission lines to operate close to their maximum current limit. Reaching the maximum limit increases the temperature in the conductor [1], and out of control high conductor temperatures can weaken the line and decrease the cable's elongation.

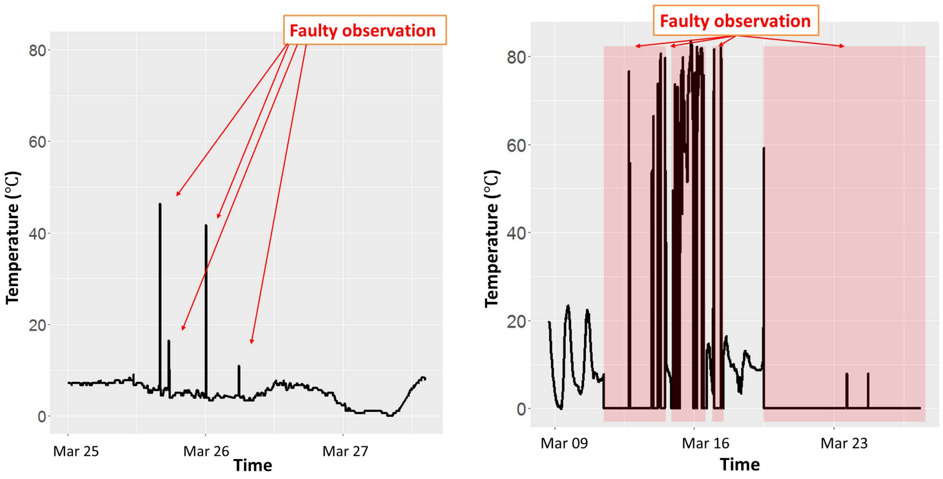
The thermal current limit (TCL) is defined as the maximum amount of electrical current that a cable's conductor can carry before deterioration [2]. In calculation of the static rating (SR) or static thermal current limit, the conductor is considered to be operating under *presumed* atmospheric conditions; in dynamic line rating (DLR), the conductor is considered to be operating under *real* atmospheric conditions.

DLR is a technique that allows the increase of the TCL in power transmission lines without damaging their conductor [3]. In DLR, the current transmission capability (also known as ampacity) of a line is calculated in real time using weather information. This dynamic management of ampacity considers the physical and electrical properties of power cables to estimate the maximum allowable conductor temperature for a particular set of weather parameters [4,5].

DLR is commonly used when renewable sources with variable production capacity (such as solar power stations, wind farms, tidal or wave power plants) have to be integrated into the grid [3,6,7]. The production of these sources is not constant and depends on environmental conditions, such as wind speed or solar radiation. Employing dynamic ampacity management allows power companies to temporarily, as needed, increase the lines' transmission capability over the limits defined by static design in order to reduce the curtailment of electricity supplied from renewable energy sources.

One of the main problems in implementing such dynamic management is the presence of erroneous or missing weather sensor data. A lack of data forces the operator to return to static ampacity management, and it may result in reducing or even disconnecting renewable energy production. Faulty data may deceive the operator into allowing the current to increase over the maximum temperature that the cable was designed for. This could lead to serious damage to the conductor. Being able to detect and correct this faulty data in real time will allow power companies to use the resources efficiently, without putting the smart grids' components at risk.

Figure 1 shows examples of temperature sensor readings with two types of faulty observations: *spike errors*, and a *sequence of consecutive errors*. Spike errors are the faulty sensor readings that happen in a short time, e.g., a single time stamp. A sequence of consecutive errors refers to continuous faulty sensor readings, e.g., over several hours.



(**a**) spike errors (**b**) sequence of consecutive errors

Figure 1. Examples of temperature sensor readings with faulty observations.

Real-world data is usually incomplete, noisy, and inconsistent [8]. Employing faulty data can lead to incorrect decisions and unreliable analysis. Therefore, data cleaning (data cleansing) has been a key area in data analysis and machine learning.

Most of the existing faulty data detection techniques are designed for offline applications where the entire data set is available. Moreover, the methods which are proposed for online data cleaning are mainly effective for repairing spike errors [9].

Smoothing techniques are widely used for online data cleaning to eliminate noisy data [9]. Moving average (MA) [10], weighted moving average (WMA) [11], exponentially weighted moving average (EWMA) [11], and sliding window bottom-up (SWAB) [12] are examples of smoothing methods. MA smooths time series data by computing the unweighted mean of the last k points. In WMA, the data at different timestamps are given different weights. EWMA assigns exponentially decreasing weights over time. SWAB uses linear interpolation or linear regression to find the approximating model for the data.

The detection and replacement of faulty data can also be done by autoregressive methods, such as AR [13–15], ARX [13,16], or ARIMA [13,17]. ARIMA consists of an autoregressive process and a moving average process. A data point is considered faulty if its prediction is significantly different from the observation.

Constraint-based replacement is also one of the techniques used for detecting and cleaning faulty data [18–20]. These methods are based on defining some constraints that the data should satisfy. In [20], a method for stream data cleaning based on speed constraints was proposed. According to this method, the derivative of the signal (change in consecutive values over time difference) should be bounded.

The above-mentioned methods are commonly used for a univariate time series analysis, i.e., the models are based on considering only one variable at a time. In many applications, such as environmental monitoring, several variables are continuously collected. In these applications, the univariate methods are limited by their inability to capture and model important dynamic interrelationships between variables of interest.

Multidimensional models take into account the correlation and dependencies between different variables to improve data replacement. A multivariate EWMA control chart [21] is a technique commonly used to simultaneously monitor several correlated variables. However, this control chart is based on only the most recent observation. Another example of multidimensional models is the vector autoregression (VAR) model [22,23], which is a generalization of the univariate autoregressive model for forecasting a collection of variables, i.e., a vector of time series. VAR is used to capture only the *linear interdependencies* among multiple variables.

The detection and replacement of faulty data only based on the change in a recent sensor reading is a useful method when dealing with spike errors. However, in the case of sequences of faulty data (e.g., several hours), this method fails [9]. In this case, the replaced values for the faulty data gradually increases/decreases and eventually causes a big offset error after several inputs.

In smart grids, a large amount of historical data related to measurement readings are available but rarely used. In this work, we propose a method for exploiting historical data to detect and replace sequences of consecutive faulty observations originating from streaming weather sensors. The proposed data-driven method does not require any assumption about the underlying population from which the data are obtained. It is a combination of: (a) a set of constraints in derivatives of sensor data (locally), and (b) a set of association rules automatically generated from historical data (globally). Therefore, it is not only based on the recent sensor readings. To generate association rules, 3 years of historical data from weather sensors of a power station in the north of Spain were used.

To evaluate the proposed method, experiments on data sets with real and synthetic errors were performed. From Figure 1b, we can observe that there are two types of consecutive errors: (1) sequence of erroneous samples with an offset increase, and (2) zero-value samples. In this work, we considered the faulty data of type 1 and, in generating synthetic faults, we created a sequence of samples with an offset error.

The rest of the paper is structured as follows: Section 2 describes the calculation of ampacity and the proposed method; Section 3 demonstrates the results; Section 4 is devoted to the discussion; and Section 5 concludes the paper.

2. Materials and Methods

2.1. Weather Information and Ampacity

The ampacity limit is computed using both the physical characteristics of the cables and the weather conditions surrounding the conductor. According to CIGRE TB601 [5], the maximum current that can be transmitted by a cable is computed considering the steady-state heat balance equation that is shown in Equation (1).

CIGRE TB601 steady-state heat balance

$$P_J + P_M + P_S + P_I = P_C + P_R + P_W, \tag{1}$$

where P_J is the Joule heating, P_M is the magnetic heating, P_S is the solar heating, P_I is the corona heating, P_C is the convective cooling, P_R is the radiative cooling, and P_W is the evaporative cooling.

The heat balance equation according to IEEE 738 [4] does not consider magnetic heating, corona heating, and evaporative cooling, because their impact is usually insignificant compared to the other terms. The simplified IEEE 738 equation for non-steady-state (transient) heat balance includes the total heat capacity of the conductor mC_p.

IEEE 738 transient heat balance

$$P_J + P_S + P_I = P_C + P_R + mC_p\frac{dT_c}{dt} \tag{2}$$

In addition, in steady-state conditions, $\frac{dT_c}{dt} = 0$, and, therefore, the current rating I_{DR} can be computed by the following.

IEEE 738 steady-state heat balance

$$I_{DR} = \sqrt{\frac{P_C + P_R - P_I - P_S}{R(T_{c,avg})}} \tag{3}$$

The computation of both CIGRE TB601 and IEEE 738 is based on the weather conditions surrounding the overhead transmission lines. Figure 2 illustrates the basic architecture of a weather measurement system for ampacity computation.

The weather stations include temperature, humidity, wind, radiation, rain, and atmospheric pressure sensors. However, prototype versions of the stations usually lack rain and pressure sensors, as they are not used for the calculation of ampacity, and so it is not present in long-term data. It is interesting to point out the need to use ultrasonic wind sensors, as recommended by CIGRE TB299 [24]. The reason is the high precision required at very low wind speeds, as the change from natural convection cooling and forced convection cooling happens when wind speed is lower than 1 m/s [4].

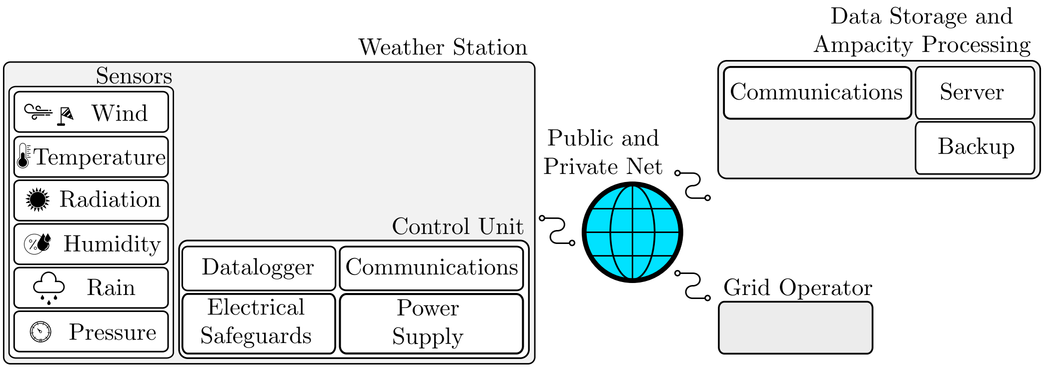


Figure 2. Architecture of ampacity management system.

Usually, the weather sensors are installed in substations to facilitate the access to the hardware if needed. However, sometimes it is required to install weather stations at midpoint locations along the power lines. Past experiences show that locating the sensors along the lines can cause problems, such as loss of signal, SIM card deterioration, and sensor power failure. An example of this type of installation can be seen in Figure 3.

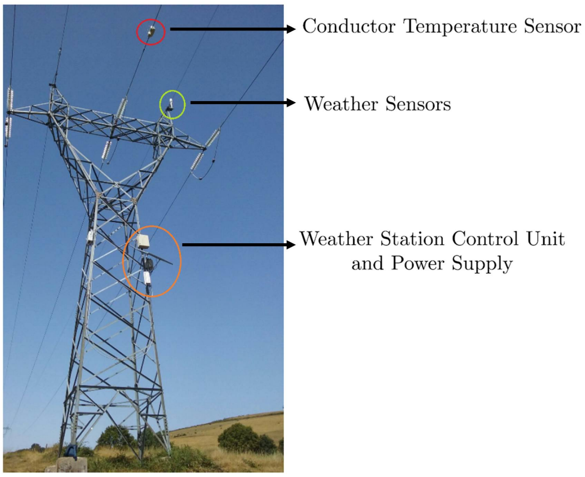

Conductor Temperature Sensor
Weather Sensors
Weather Station Control Unit and Power Supply

Figure 3. An example of an installation of a weather station on a transmission tower.

The accuracy of the ampacity calculation depends on the accuracy of the weather sensor readings. Figure 4 illustrates the relative connection between the accuracy of input data and ampacity calculation. In this figure, the uncertainty of the ampacity computation process is shown by the gray area, which depends on the weather parameter that is faulty or unavailable.

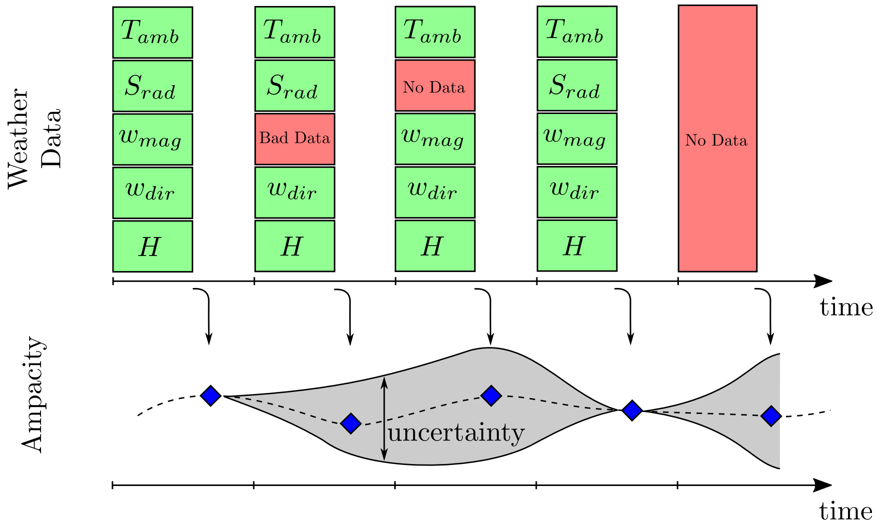


Figure 4. The general relation between the accuracy of input data and ampacity calculation.

In this work, we had access to the data of only one substation in the north of Spain. For this station, we assumed that sensors fail independently of each other. Furthermore, problems such as communication issues or power loss, which lead to many or all data corruption at the same time, were not considered.

2.2. Approach

Figure 5 illustrates the system architecture of our proposed method. The method requires historical environmental conditions from the location where the weather sensors are installed. The historical data are used for two purposes: (a) calculating the derivative in every two consecutive data inputs, and (b) generating a set of rules to estimate the correlation between the sensor readings under different weather conditions. Furthermore, to combine the results of (a) and (b), we accounted for the confidence of the rules.

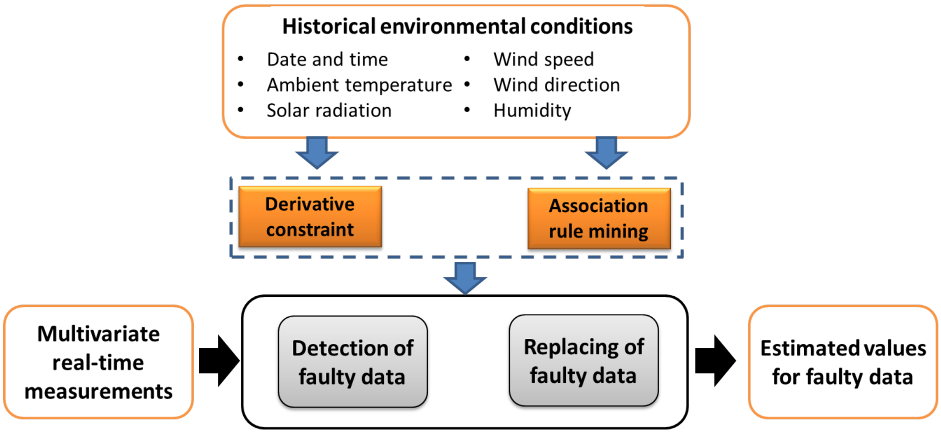


Figure 5. System architecture of the proposed method.

2.2.1. Signal Derivative Constraints

To identify which data points are faulty, we noticed that the derivative is often bounded—so-called speed constraints [18,20]. Given that weather variables are slow changing, the difference between two consecutive samples should not be very large.

Consider the historical data $\mathbf{Z}$ with columns $j \in [1, ..., k]$ representing the numerical attributes of observations $i \in [1, ..., n]$. Each $z_{i,j}$ has a timestamp t_i. For any consecutive observations $z_{i-1,j}$, $z_{i,j}$, the derivative is defined as the absolute ratio of the change in the value z over time t as $S_i^k = \left| \frac{z_{i,j} - z_{i-1,j}}{t_i - t_{i-1}} \right|$ for $i \in [2, ..., n]$ [20].

A *derivative constraint* for attribute k (δ_k) is defined as the maximal change in the derivative within all of its consecutive observations $\delta_k = max \left\{ S_2^k, S_3^k, ..., S_n^k \right\}$. The derivative constraint was calculated separately for each attribute $j \in [1, 2, ..., k]$, *i.e.*, $\Delta = \{\delta_1, \delta_2, ..., \delta_k\}$.

2.2.2. Association Rule Mining

Usually, association rule mining is applied to discrete data [25]. Therefore, the historical data need to be transformed into discrete values before generating the rules. In data discretization, the numerical attributes are replaced by interval labels (e.g., 0–3) or conceptual labels (e.g., *Spring*). There are several techniques for data discretization, such as using experts' knowledge, binning (equal-width or equal-frequency), histogram analysis, clustering, decision tree, and correlation analysis [8].

Figure 6 illustrates part of the data set before and after discretization. Here, we used the same notation as before for representing the historical data $\mathbf{Z}$, which is a two-dimensional real-valued matrix with size $n \times k$. After the data discretization, matrix $\mathbf{Z}$ was transformed into a two-dimensional Boolean matrix $\bar{\mathbf{Z}}$ with size $n \times m$, where $m > k$. A value of 1 in the Boolean matrix $\bar{\mathbf{Z}}$ indicates the presence of a feature (item) in an observation, and a value of 0 indicates the absence of the feature.

We define the set of items of $\bar{\mathbf{Z}}$ as $I = \{I_1, I_2, ..., I_m\}$. Each observation $\bar{z}_i$ in the data set may or may not contain a specific item, e.g., $\bar{z}_1 = \{I_1, I_2, I_5\}$ means that observation $\bar{z}_1$ only contains items I_1, I_2, and I_5.

The objective of mining association rules is to find the most frequently occurring combinations of items in a data set [26]. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and A, B are disjoint itemsets, i.e., $A \cap B = \emptyset$. In this case, the itemset $A = \{a_1, a_2, ...\}$ is the prior and the itemset $B = \{b_1, b_2, ...\}$ is the posterior of the rule.

We define X_A as $X_A = \{\bar{z}_i \in \bar{\mathbf{Z}}$ that contains items $A\}$. In this case, the *support* of itemset A, represented by $S(A)$, is the ratio of the cardinality of X_A over the cardinality of the data set $\bar{\mathbf{Z}}$ (number of observations) [26]:

$$S(A) = \frac{|X_A|}{|\bar{\mathbf{Z}}|} = P(X_A). \tag{4}$$

The support of a rule, denoted as $S(A \Rightarrow B)$, is the percentage of observations in the data set that contain both A and B:

$$S(A \Longrightarrow B) = \frac{|X_{A \cup B}|}{|\bar{\mathbf{Z}}|} = \frac{|X_A \cap X_B|}{|\bar{\mathbf{Z}}|} = P(X_A, X_B). \tag{5}$$

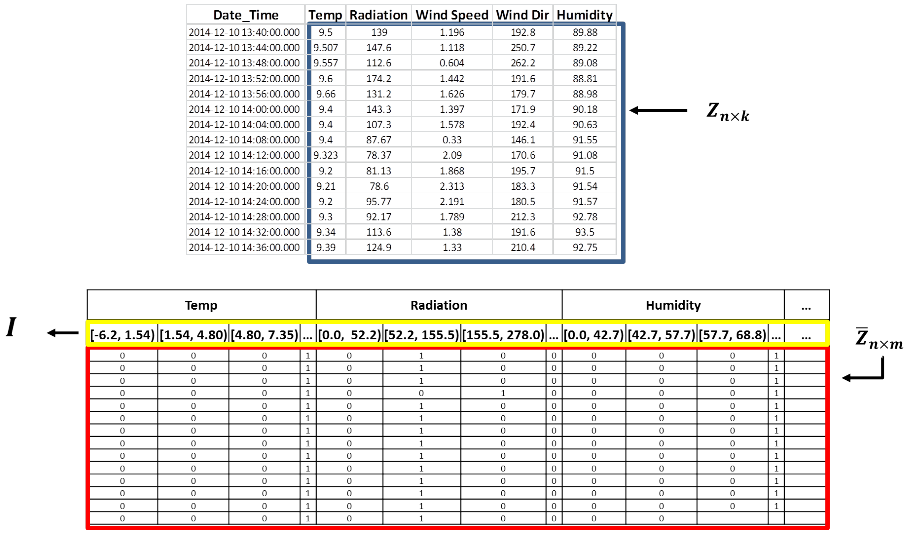
The *confidence* of an association rule is the percentage of examples containing A that also contain B; or, in other words, a fraction that shows how frequently B occurs among all the observations containing A. The confidence value indicates how reliable the rule is.

$$C(A \Rightarrow B) = \frac{S(A \Rightarrow B)}{S(A)} = P(X_B|X_A) \tag{6}$$

The *lift* of an association rule is the ratio of the confidence of the rule to the frequency of observations containing B. It is a value between 0 and infinity that measures the deviation of a rule from statistical independence:

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{S(B)} = \frac{P(X_A, X_B)}{P(X_A)P(X_B)}. \tag{7}$$

A lift value smaller than 1 indicates negative correlation, equal to 1 indicates no correlation, and greater than 1 indicates a positive correlation between features A and B among all the observations.



Date_Time	Temp	Radiation	Wind Speed	Wind Dir	Humidity
2014-12-10 13:40:00.000	9.5	139	1.196	192.8	89.88
2014-12-10 13:44:00.000	9.507	147.6	1.118	250.7	89.22
2014-12-10 13:48:00.000	9.557	112.6	0.604	262.2	89.08
2014-12-10 13:52:00.000	9.6	174.2	1.442	191.6	88.81
2014-12-10 13:56:00.000	9.66	131.2	1.626	179.7	88.98
2014-12-10 14:00:00.000	9.4	143.3	1.397	171.9	90.18
2014-12-10 14:04:00.000	9.4	107.3	1.578	192.4	90.63
2014-12-10 14:08:00.000	9.4	87.67	0.33	146.1	91.55
2014-12-10 14:12:00.000	9.323	78.37	2.09	170.6	91.08
2014-12-10 14:16:00.000	9.2	81.13	1.868	195.7	91.5
2014-12-10 14:20:00.000	9.21	78.6	2.313	183.3	91.54
2014-12-10 14:24:00.000	9.2	95.77	2.191	180.5	91.57
2014-12-10 14:28:00.000	9.3	92.17	1.789	212.3	92.78
2014-12-10 14:32:00.000	9.34	113.6	1.38	191.6	93.5
2014-12-10 14:36:00.000	9.39	124.9	1.33	210.4	92.75

Temp				Radiation				Humidity				...
[-6.2, 1.54)	[1.54, 4.80)	[4.80, 7.35)	...	[0.0, 52.2)	[52.2, 155.5)	[155.5, 278.0)	...	[0.0, 42.7)	[42.7, 57.7)	[57.7, 68.8)	...	...
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	0	1	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0	0	1	
0	0	0	1	0	1	0	0	0	0			

Figure 6. Part of the data set before (**Z**) and after discretization ($\bar{\mathbf{Z}}$).

2.2.3. Detection of Faulty Data

We used both the derivative constraints and the association rules for detecting faulty data. For the derivative constraint, only the last two observations are considered, i.e., each attribute in the new observation O_t^j is evaluated "locally" against the value of the same attribute from the previous observation. If, for any feature (e.g., feature j where $j \in [1, ..., k]$), the difference between the two values exceeds the constraint (δ_j), the current observation will be labeled as faulty. Furthermore, each attribute in the new observation is also compared against the association rules generated from all the historical data (globally). If the current observation from streaming data stays out of the intervals which are identified by the rules, it will be labeled as a faulty observation.

2.2.4. Replacing of Faulty Data

When a new observation of one attribute is labeled as faulty, the "estimation" of the correct value needs to be performed. First, we define $C = \{c_1, c_2, ...\}$ to be the combination of all items except the faulty attribute. In the list of association rules, we search for the rules which have C as prior, i.e., $X_C = \{\bar{z}_i \in \bar{\mathbf{Z}}$ that contains items $C\}$. The rule with the highest confidence specifies the intervals for the faulty observation.

Then, we search for the corresponding observations in the historical data $\mathbf{Z}$ that contain C in $\bar{\mathbf{Z}}$. Let us call this $Y_C = \{z_i \in \mathbf{Z}$ that corresponds to C in $\bar{\mathbf{Z}}\}$. The average of derivatives in Y_C for the faulty attribute j specifies the changes between previous and current observations.

For example, assume the attribute temperature is faulty, and the following is the generated rule with the highest confidence:

$$\{SolarRadiation = [766.3, 1275.5], WindSpeed = [0.932, 1.619), WindDirection = D2,$$
$$Humidity = [8.9, 42.7), Time = [11:00 - 14:59], Month = Aug\} =>$$
$$\{Temp = [24.99, 36.30]\}$$

According to this rule, the correct value for the temperature sensor would be within the interval $[24.99, 36.30]$. In this example, the terms before the arrow sign "⇒" belong to the itemset C. In the historical data $\mathbf{Z}$, all the observations that are within the intervals of C correspond to Y_C.

The faulty observation O_t^j will be replaced by $\hat{O}_t^j$ using the following formula.

$$\hat{O}_t^j = O_{t-1}^j + mean(S_{i \in Y_C}^k) \tag{8}$$

In power companies, if there is a missing value or faulty data input from weather sensors, the experts manually replace them by looking into previous observations (for example, the last 10 days). Formula (8) automatically applies the same concept while using the historical data for 3 years, searching for similar conditions, and considering all the attributes at the same time.

This estimated value of $\hat{O}_t^j$ *should* be within the intervals captured from the rules. However, in some cases, the estimated value is not within the limits. In these cases, the confidence of the rule is considered. The confidence value indicates how reliable a rule is and is used as a weight to modify the limits by using the following equation.

$$N\_\hat{O}_t^j = \begin{cases} \hat{O}_t^j - (\left|\hat{O}_t^j - I_{max}\right| \times C_{rule}), & \text{if } \hat{O}_t^j > I_{max} \\ \hat{O}_t^j + (\left|\hat{O}_t^j - I_{min}\right| \times C_{rule}), & \text{if } \hat{O}_t^j < I_{min} \end{cases} \tag{9}$$

In this equation, I_{max} is the maximum of the rule interval, I_{min} is the minimum of the rule interval, C_{rule} is the confidence of the rule, and $N\_\hat{O}_t^j$ is the new estimation of the faulty observation.

In the next section, we demonstrate that, by using the confidence value as a weight to modify the limits, we can improve the accuracy of replacing faulty data.

3. Results

The available dataset contains 337,035 weather sensor readings from 10 December 2014 until 16 August 2017 for one weather station. The data set does not contain any faulty data; however, there are some missing values in the data. From this data set, 310,000 observations were used for generating association rules and calculating speed constraints. The last part of the data (test set) was used for evaluation. This test data set contains multivariate *real-time* measurements.

The methodology was developed using R programming language and RStudio as the graphical front-ends. The R code was run on a PC which was configured with a 2.50 GHz Intel (R) Core (TM) i5-4300U CPU and 8 GB memory.

Based on 310, 000 observations, the maximum change constraints Δ (within 1 min) for the attributes solar radiation, ambient temperature, humidity, wind speed, and wind direction are $\delta_{S_{rad}} = 50$, $\delta_{T_{amb}} = 0.25$, $\delta_H = 1.5$, $\delta_{W_{mag}} = 0.75$, and $\delta_{W_{dir}} = 45$ (see Figure 7). In the calculation of the derivative constraints, only the absolute value of the changes were considered important.

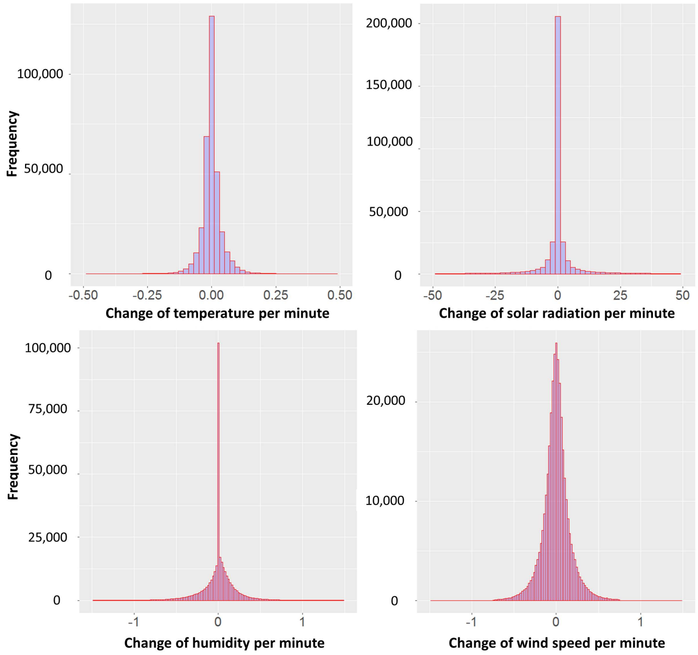


Figure 7. Histogram plot of the change per minute for the attributes solar radiation, temperature, humidity, and wind speed for all the historical data. The absolute value of the changes were used for estimating the constraints.

In order to generate association rules, first we discretized the values in the training data set. We used two methods for data discretization: (1) consulting with experts, and (2) using k-means clustering [8]. Using k-means, the numerical attributes ambient temperature, solar radiation, humidity, and wind speed were clustered into 10, 7, 7, and 7 categories, respectively. The number of clusters were chosen based on the quality measures of the final estimations.

In addition to the available attributes, we added *Hour* and *Month* based on the timestamps in the data set. The *Hour* corresponds to the time of the day, which was also discretized into equal-width intervals. The *Month* corresponds to the month number.

We refer to our proposed association rule-based method as:

- $A - rule(EC)$, when using **E**xperts' knowledge for data discretization and the **C**onfidence is used as a weight to modify the limits based on Equation (9).
- $A - rule(E)$, when using **E**xperts' knowledge for data discretization and confidence is not considered.
- $A - rule(K)$, when using **K**-means for data discretization and confidence is not considered.

After data discretization, we applied association rule mining using an "apriori" function implemented in the R `arules` library. The thresholds for confidence and lift were set to 60% and 1, respectively. For each attribute as posterior (e.g., ambient temperature), the rules were generated separately. Accordingly, 1235 rules were generated for predicting the ambient temperature, and 329 rules for predicting wind speed. Some of these rules are presented in Table 1.

Table 1. Example of the generated association rules with confidence greater than 60% and lift greater than 1 for ambient temperature and wind speed as posterior.

Prior	Posterior	Confidence	Posterior Attribute
{Radiation = [766.3, 1275.5], WindSpeed = [0.932, 1.619), WindDirection = D2, Humidity = [8.9, 42.7), Hour = T2, Month = 08}	{Temp = [24.99, 36.30]}	1.00	temperature
{Radiation = [587.4, 766.3), WindSpeed = [5.195, 10.340], WindDirection = D1, Humidity = [42.7, 57.7), Hour = T3, Month = 07}	{Temp = [20.49, 24.99)}	0.95	temperature
{Radiation = [0.0, 52.2), WindSpeed = [0.932, 1.619), WindDirection = D1, Humidity = [95.4, 100.0], Hour = T3, Month = 02}	{Temp = [−6.20, 1.54)}	0.67	temperature
{Radiation = [0.0, 52.2), WindSpeed = [3.153, 4.047), WindDirection = D2, Humidity = [95.4, 100.0], Hour = T2, Month = 11}	{Temp = [1.54, 4.80)}	0.92	temperature
{Radiation = [422.7, 587.4), WindSpeed = [4.047, 5.195), WindDirection = D1, Humidity = [87.3, 95.4), Hour = T2, Month = 07}	{Temp = [17.18, 20.49)}	0.96	temperature
...	...	...	...
{Temp = [12.04, 14.47), Radiation = [0.0, 52.2), WindDirection = D2, Humidity = [68.8, 78.4), Hour = T2, Month = 12}	{WindSpeed = [5.195, 10.340]}	0.79	wind speed
{Temp = [17.18, 20.49), Radiation = [766.3, 1275.5], WindDirection = D1, Humidity = [87.3, 95.4), Hour = T3, Month = 06}	{WindSpeed = [4.047, 5.195)}	0.74	wind speed
{Temp = [9.69, 12.04), Radiation = [0.0, 52.2), WindDirection = D1, Humidity = [42.7, 57.7), Hour = T5, Month = 12}	{WindSpeed = [3.153, 4.047)}	0.74	wind speed
{Temp = [12.04, 14.47), Radiation = [0.0, 52.2), WindDirection = D1, Humidity = [57.7, 68.8), Hour = T5, Month = 10}	{WindSpeed = [0.932, 1.619)}	1.00	wind speed
{Temp = [−6.20, 1.54), Radiation = [0.0, 52.2), WindDirection = D2, Humidity = [95.4, 100.0], Hour = T1, Month = 10}	{WindSpeed = [0.023, 0.932)}	1.00	wind speed
...	...	...	...

The real faulty samples shown in Figure 1b, which were collected from another power station, were used to visually evaluate the performance of the *A-rule(EC)* method. These data contained several faulty sensor readings corresponding to several hours. In these data, only the temperature sensor was faulty, and the other sensors were correct. The results of applying our method are shown in Figure 8. Since we did not have access to the real temperature, we used historical data from a nearby weather API station as the ground truth. According to the figure, the estimation of faulty observations is very close to the ones we captured from the API.

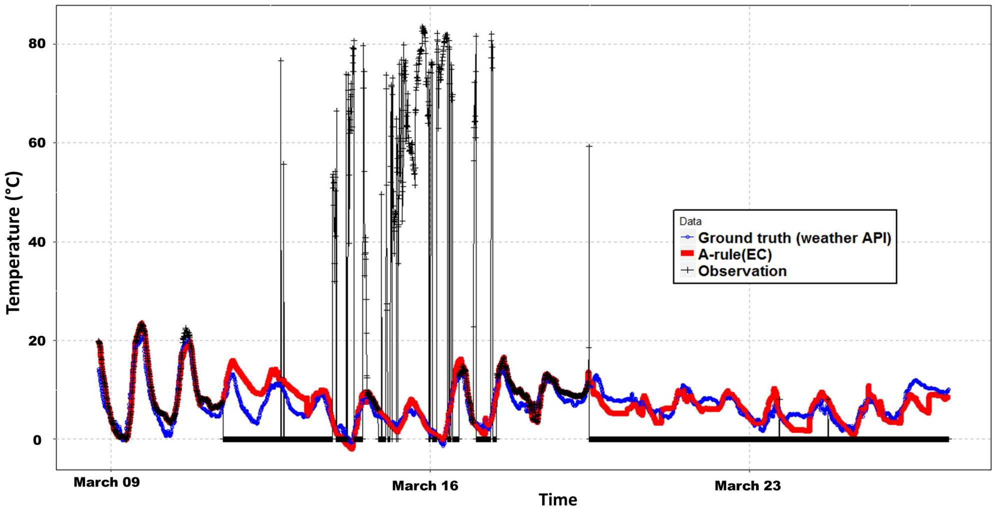


Figure 8. Detection and correction of real faulty data from Figure 1b.

4. Discussion

Unfortunately, we did not have access to a sufficient number of real faulty sensor readings to evaluate our method with the real data. Therefore, we generated synthetic errors and evaluated the performance of the proposed method using artificial faulty observations. To generate synthetic faults, a sequence of samples from the test set was selected and an offset error was added to them.

Figures 9 and 10 show four examples with synthetic errors. In these figures, in addition to repairing the faulty data using our proposed *A-rule(EC)* method, we also repaired the data by utilizing EWMA and VAR techniques. The parameters for a VAR model were calculated using the “VAR” function in the R `vars` library.

According to Figures 9 and 10, the EWMA technique could not correctly replace the faulty observations. After a few faulty samples, the estimation based on the EWMA became very close to the faulty observations. The VAR method was better than EWMA in the replacement of faulty data, but the estimation was still very far from the actual values. On the other hand, our proposed association rule method outperformed the other two methods.

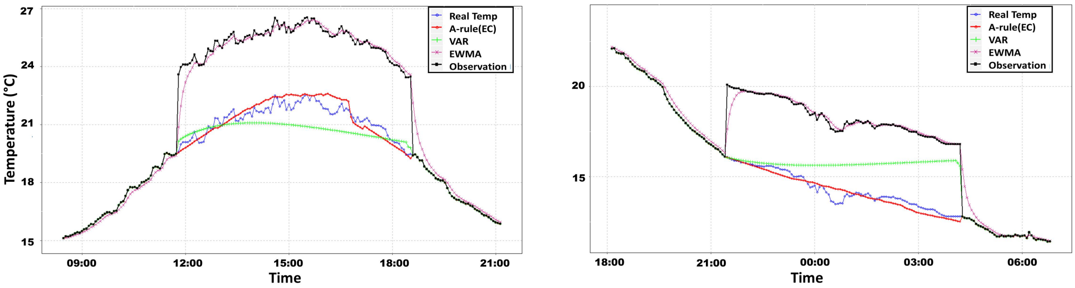


Figure 9. Detection and correction of synthetic faulty data from a temperature sensor.

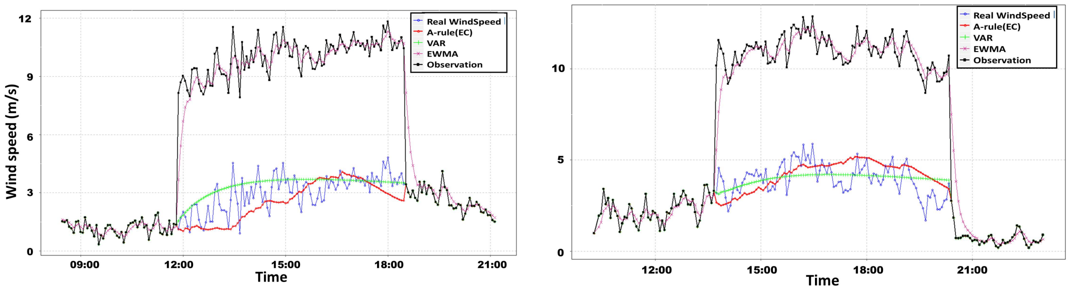


Figure 10. Detection and correction of synthetic faulty data from a wind speed sensor.

In addition to the examples presented in Figures 9 and 10, we picked several other samples. To evaluate our proposed method, we selected 100 series of 200 observations from the test set. The samples between 70 and 170 in the series were changed by adding an offset to create a sequence of 100 consecutive faulty observations. The offset error for the temperature sensor was set to 4 degrees Celsius. These 100 observations correspond to 400 minutes. Then, these data sets were examined with our association rule-based method and the VAR and EWMA methods.

For all 100 data sets, the average of the difference between the real observations and corrected values was calculated as the estimation error. Figure 11 shows the results of replacing temperature data in all 100 data sets. Table 2 presents the mean value with 95% confidence interval (CI) and minimum and maximum error when the number of faulty samples is 25, 50, 75, and 100.

According to the results presented in Figure 11 and Table 2, in all the methods, the estimation error was rising as the number of faulty samples increased. The EWMA could not correctly replace the faulty data and the error reached the maximum value 4 (corresponding to the 4-degree Celsius offset) after a few observations. The VAR method worked well if the number of faulty samples was small, e.g., less than 10 samples. However, for more than 25 faulty samples, the error was higher than the association rule-based methods. Within all three rule-based methods, the error in *A-rule(EC)* —when we were using the confidence of the rules and applying experts' knowledge for data discretization—was the lowest. Furthermore, the 95% CI in estimating the correct temperature was lower than for all other methods. This shows that the *A-rule(EC)* method is also more robust in detecting and replacing faulty data compared to the other methods.

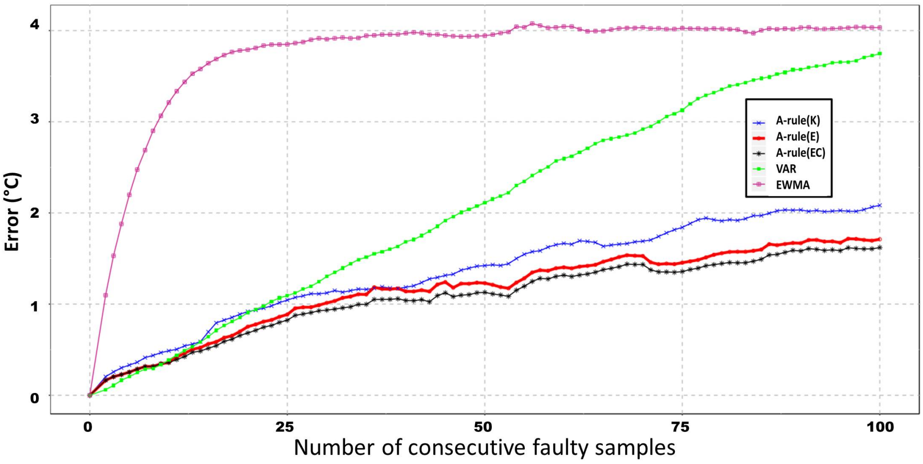


Figure 11. The average of the difference between the real observations and corrected values for 100 data sets for cleaning temperature data.

Table 2. Error in cleaning temperature data with a different number of faulty samples for 100 data sets.

	Method	**Mean Error**	**95% CI (Lower)**	**95% CI (Upper)**	**Min Error**	**Max Error**
25 faulty samples	*A-rule(K)*	1.073	0.800	1.346	0.001	9.809
	A-rule(E)	0.954	0.737	1.171	0.005	7.848
	A-rule(EC)	0.876	0.705	1.048	0.005	6.270
	VAR	1.121	0.939	1.304	0.009	4.850
	EWMA	3.863	3.773	3.954	2.485	5.139
50 faulty samples	*A-rule(K)*	1.431	1.152	1.710	0.018	9.087
	A-rule(E)	1.211	0.952	1.470	0.001	6.534
	A-rule(EC)	1.112	0.892	1.332	0.001	6.142
	VAR	2.152	1.789	2.515	0.024	9.409
	EWMA	3.956	3.871	4.041	2.144	5.081
75 faulty samples	*A-rule(K)*	1.884	1.564	2.204	0.019	8.976
	A-rule(E)	1.469	1.167	1.772	0.004	7.882
	A-rule(EC)	1.379	1.092	1.667	0.010	7.861
	VAR	3.197	2.679	3.714	0.012	12.060
	EWMA	4.022	3.929	4.115	2.521	5.370
100 faulty samples	*A-rule(K)*	2.091	1.727	2.455	0.010	10.593
	A-rule(E)	1.690	1.358	2.021	0.080	10.127
	A-rule(EC)	1.600	1.282	1.919	0.051	10.090
	VAR	3.780	3.194	4.367	0.040	13.437
	EWMA	4.018	3.940	4.097	2.933	5.304

The gray colored rows correspond to the lowest mean in the estimation error.

In this work, we only present the results for cleaning two attributes, ambient temperature and wind speed, but the methodology can be applied to other attributes as well.

In general, for most of the examples, our method outperformed EWMA and VAR. There are some situations wherein our method failed in replacing faults. Figure 12 demonstrates an example where the *A-rule(EC)* failed to correctly replace the faulty samples, especially after several observations (after time 21:00 in the figure). Further analysis showed that these situations were happening because of the lack of the rules generated from historical data. Since we were considering the rules with confidence greater than 60%, not all the conditions were included. One way to improve this is to use a larger historical data set and generate the rules based on that.

Moreover, the proposed method performed better when the location of the collected historical data was the same place where the method was going to be used. We considered that the environmental changes are very relevant to the geographical position, and when the method was "trained" with a data set with completely different environmental conditions from the test data, it failed. However, for stations not very far from each other (such as the data for the station in Figure 8), the method showed a very good performance.

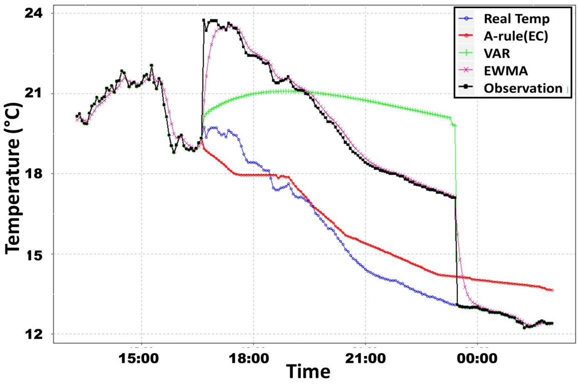


Figure 12. Detection and correction of faulty data from temperature sensor, when *A-rule(EC)* does not perform well.

The main drawback of this proposed method is that we are assuming only one sensor reading is faulty at a time and other sensors are correct. This might be problematic when there is a communication problem and all the data are missing. To continue this work, the authors are considering adopting the information from neighbor stations to detect and replace faulty observations.

In smart grids, intelligent sensors distributed throughout the grid allow for continuous collection of valuable data. In addition, large amounts of information related to historical faults, repairs, reported alarms, and so on, are recorded in different ways. This information can be used for many purposes, including data cleaning, fault detection, failure prediction, and load forecasting. However, most power electricity companies do not utilize this data fully and they often do not realize the full benefits of doing so. In this paper, we demonstrate that applying data-driven methods on the historical data provides valuable information that can be used for detecting and replacing faulty sensor readings.

5. Conclusions

In this paper, we propose a method for detecting and replacing sequences of consecutive faulty data originating from streaming weather sensors. To detect faulty data, a combination of both the derivative constraints and the association rules was used. Replacing the faulty data was done based on automatically generated association rules from the historical data. Furthermore, when replacing faulty data, we took into account the confidence of the rules. This means that instead of using the rules as crisp inferences, the rules are weighted by their confidence.

In order to evaluate the method, experiments on real data with real and synthetic errors were performed. The results show that the proposed A-rule-based method outperforms the commonly used methods, such as EWMA and VAR, especially when having a sequence of consecutive faulty weather sensor readings. Furthermore, among all three rule-based methods, the error in *A-rule(EC)*—when using the confidence of the rules and applying experts' knowledge for data discretization—is the lowest.

Author Contributions: The individual contributions for this research are described in the following. H.M.N.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing—Original Draft Preparation, Writing—Review & Editing, Visualization, Project Administration. A.L.: Conceptualization, Data Curation, Resources. M.M.: Conceptualization, Resources, Writing—Review & Editing. A.S. and S.N.: Supervision, Validation, Review.

Funding: This research was partially funded by Spanish Government under Spanish R+D initiative with reference ENE2013-42720-R and RETOS RTC-2015-3795-3.

Acknowledgments: This work was partially supported by Swedish Knowledge Foundation. The authors would also like to acknowledge Viesgo, the distribution system operator in northwest Spain, and Halmstad Energi och Miljö, the distribution system operator in the south of Sweden.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TCL	Thermal current limit
SR	Static Rating
DLR	Dynamic Line Rating
MA	Moving Average
WMA	Weighted Moving Average
EWMA	Exponentially Weighted Moving Average
SWAB	Sliding Window bottom-up
AR	Autoregressive
VAR	Vector Autoregression

References

1. Heckenbergerová, J.; Hosek, J. Dynamic thermal rating of power transmission lines and renewable resources. In Proceedings of the 11th International Conference on Environment and Electrical Engineering, Venice, Italy, 18–25 May 2012.
2. European Network of Transmission System Operators for Electricity(entsoe). *Dynamic Line Rating for Overhead Lines—V6;* Technical Report; European Network of Transmission System Operators for Electricity(entsoe): Brussels, Belgium, 2011.
3. Foss, S.D.; Maraio, R.A. Dynamic line rating in the operating environment. *IEEE Trans. Power Del.* **1990**, *5*, 1095–1105. [CrossRef]
4. IEEE Standard Association. *IEEE Standard for Calculating the Current-Temperature of Bare Overhead Conductors*; IEEE Standard Association: Piscataway, NJ, USA, 2013.
5. International Council on Large Electric Systems, CIGRE. *Guide for Thermal Rating Calculations of Overhead Lines*; Technical Brochure 601; CIGRE: Paris, France, 2014.
6. Morozovska, K.; Hilber, P. Study of the Monitoring Systems for Dynamic Line Rating. *Energy Proced.* **2017**, *105*, 2557–2562. [CrossRef]
7. Wallnerström, C.J.; Hilber, P.; Söderström, P.; Saers, R.; Hansson, O. Potential of dynamic rating in Sweden. In Proceedings of 2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Durham, UK, 7–10 July 2014; pp. 1–6.
8. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Elsevier: New York, NY, USA, 2011.
9. Zhang, A.; Song, S.; Wang, J.; Yu, P.S. Time series data cleaning: From anomaly detection to anomaly repairing. *Proc. VLDB Endow.* **2017**, *10*, 1046–1057. [CrossRef]
10. Brillinger, D.R. *Time Series: Data Analysis and Theory*; SIAM: Philadelphia, PA, USA, 2001.
11. Gardner, E.S., Jr. Exponential smoothing: The state of the art—Part II. *Int. J. Forecast.* **2006**, *22*, 637–666. [CrossRef]
12. Keogh, E.; Chu, S.; Hart, D.; Pazzani, M. An online algorithm for segmenting time series. In Proceedings of the 2001 IEEE International Conference on Data Mining Data Mining, San Jose, CA, USA, 29 November–2 December 2001; pp. 289–296.
13. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
14. Hill, D.J.; Minsker, B.S. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environ. Modell. Softw.* **2010**, *25*, 1014–1022. [CrossRef]
15. Yamanishi, K.; Takeuchi, J.I. A unifying framework for detecting outliers and change points from non-stationary time series data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–25 July 2002; pp. 676–681.
16. Park, G.; Rutherford, A.C.; Sohn, H.; Farrar, C.R. An outlier analysis framework for impedance-based structural health monitoring. *J. Sound.Vib.* **2005**, *286*, 229–250. [CrossRef]
17. Otto, M.C.; Bell, W.R. Two issues in time series outlier detection using indicator variables. In *Proceedings of the Business and Economic Statistics Section*; American Statistical Association (ASA): Alexandria, VA, USA, 1990; pp. 182–187.
18. Bohannon, P.; Fan, W.; Flaster, M.; Rastogi, R. A cost-based model and effective heuristic for repairing constraints by value modification. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 143–154.
19. Chu, X.; Ilyas, I.F.; Papotti, P. Holistic data cleaning: Putting violations into context. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, Australia, 8–11 April 2013; pp. 458–469.
20. Song, S.; Zhang, A.; Wang, J.; Yu, P.S. Screen: Stream data cleaning under speed constraints. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Australia, 31 May–4 June 2015; pp. 827–841.
21. Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving average control chart. *Technometrics* **1992**, *34*, 46–53. [CrossRef]
22. Zivot, E.; Wang, J. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*; Springer: New York, NY, USA, 2006; pp. 385–429.

23. Peña, D.; Tiao, G.C.; Tsay, R.S. *A Course in Time Series Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 322.
24. International Council on Large Electric Systems, CIGRE. *Guide for Selection of Weather Parameters for Bare Overhead Conductor Ratings;* Technical Brochure 299; CIGRE: Paris, France, 2006.
25. Zhang, C.; Zhang, S. *Association Rule Mining: Models and Algorithms*; Springer: New York, NY, USA, 2002.
26. Nemati, M.H.; Sant'Anna, A.; Nowaczyk, S. Bayesian Network representation of meaningful patterns in electricity distribution grids. In Proceedings of the 2016 IEEE International Energy Conference (ENERGYCON), Leuven, Belgium, 4–8 April 2016; pp. 1–6.