Incomplete Data in Smart Grid: Treatment of Values in Electric Vehicle Charging Data

Contribution:

Mostafa Majipour, Peter Chu, and Rajit Gadh
Smart Grid Energy Research Center, UCLA
Los Angeles, California, USA

Hemanshu R. Pota
School of Engineering & Information Technology
The University of New South Wales
Canberra ACT2610, Australia

Acknowledgement
This material is based upon work supported by the Department of Energy under Award Number DE-OE0000192.

Incomplete Data in Smart Grid: Treatment of Missing Values in Electric Vehicle Charging Data

Mostafa Majidpour, Peter Chu, Rajit Gadh
Smart Grid Energy Research Center
UCLA
Los Angeles, California USA
mostafam@ucla.edu

Hemanshu R. Pota
School of Engineering & Information Technology
The University of NSW
Canberra ACT 2610 Australia

***Abstract*— In this paper, five imputation methods namely Constant (zero), Mean, Median, Maximum Likelihood, and Multiple Imputation methods have been applied to compensate for missing values in Electric Vehicle (EV) charging data. The outcome of each of these methods have been used as the input to a prediction algorithm to forecast the EV load in the next 24 hours at each individual outlet. The data is real world data at the outlet level from the UCLA campus parking lots. The Median imputation improved the prediction results. Given that in most missing value cases in our database, all values of that instance are missing, the multivariate imputation methods did not improve the results significantly compared to Median imputation approach.**

Keywords— Electric Vehicle; Imputation; Time Series; Forecasting.

I. INTRODUCTION

Smart cities are equipped with lots of sensors and meters to monitor and record different quantities throughout the day. This huge number of measurements calls for analysis suitable for Big Data. Among the issues of analyzing Big Data such as the computational complexity of the algorithms, and scaling the well behaved algorithms to a larger number of data points, there is the problem of missing values in data. Sometimes data is not reported due to sensor outage. While at other times, the value that is reported is far from the expected range, therefore rendering the reported value as invalid. In these scenarios, the reported value is not the actual one, and the value is considered to be a missing value. In this short paper, we are investigating the effect of the missing values on the prediction of the Electric Vehicle (EV) loads in the UCLA parking structures.

The virtue of prediction in the case of the EV loads in the UCLA campus is that each charging station has multiple outlets, and, in view of the upper limit on the available power, the charging is multiplexed among the outlets. Due to this multiplexing, the expected charging time for each outlet can vary; therefore, by predicting the available power at each outlet, the charging time for such multiplexed outlets can be computed.

II. IMPUTATION METHODS

The process of providing the best guess for a missing value is called imputation. In the case of time series, some of the imputation methods such as "case deletion" cannot be applied since they will change the relative order of events and make the time series lose its ordinal properties such as periodicity. On the other hand, multivariate imputation methods such as "Maximum Likelihood" is applicable to datasets with more than one variable observed at a time (multivariate). We will be using three imputation methods which are applicable to both multivariate and univariate time series and two imputation methods that are only applicable to multivariate time series. Our imputation implementations are briefly explained below. A more detailed discussion can be found in [1].

A. Constant Imputation

In this imputation, a constant value is substituted for missing values. While this number is arbitrary, we have chosen it to be zero in order to have a sparser time series. Thus, every time the power is missing because any of the voltage, current, or power factor is missing, we substitute it with zero.

B. Mean Imputation

This imputation is the extension of the Constant imputation where the constant number is the mean of the available data points. Thus, for missing values, we substitute the mean of the available values. The advantage of inserting the mean value is that the mean value of the data will not change.

C. Median Imputation

One disadvantage of the Mean imputation is that the imputed value (mean) might not be any of the observed values; for instance, in our case, the received power by the EV can be zero, the maximum power, or half of it, etc. depending on the number of EVs being charged at that specific station. Thus, because of the discrete nature of the values, the mean value might not exist as an observed value while median is always one of the observed values of the data.

D. Maximum Likelihood Imputation

The idea behind Maximum Likelihood (ML) is to impute the unobserved values by considering the other variables at that moment. For instance, at some moment, the voltage value might be missing but the current value might be available. Then, the missing value for voltage can be replaced by considering

This work has been sponsored in part by grant from the LADWP/DOE fund 20699 & 20686, Smart Grid Regional Demonstration Project.

corresponding voltage of the moments with similar currents. This idea is usually implemented through the EM algorithm.

E. Multiple Imputation

A disadvantage of the aforementioned methods is that they underestimate the error by adding more data points without adding more information. Consider the error in the average of data points (sample mean) for instance: by adding more data points, the sample mean error (which is inversely proportional with the number of data points) reduces, without adding any new information to the available dataset.

The Multiple imputation method is similar to ML with the difference that each missing value is imputed by adding an error term so that it estimates the actual values more accurately. This process iterates a few times (usually five) and the final imputed value is the average of these iterations [1].

III. Simulations and Analysis

A. Data

The imputation methods described above are applied to charging stations located on the UCLA campus. The data used in this paper were recorded from December 7, 2011 to October 16, 2013; however, not all outlets were in use on all days. Among charging outlets at UCLA, 15 outlets have charging data for more than 60 effective days (days with some nonzero charging); these outlets have been used in our implementation.

Data for each outlet is in the format that is called Station Records. Each station Record contains the hourly voltage, current, and power factor of the charging outlet. In order to find the time series of real power at each hour, we multiply the voltage, current, and power factor.

B. Prediction Algorithm

The applied prediction algorithm in this paper is Nearest Neighbor (NN). This algorithm has been found to have better accuracy with respect to Symmetric Mean Absolute Percentage Error (SMAPE). Based on the NN algorithm, each sample (training, test or validation) is composed of input and output pairs. In our application, the output is the energy consumption for the next 24 hours, $y(t) = \underline{E}(t)$, and the input is the concatenation of the consumption records for up to D prior days, $x(t) = \{\underline{E}(t-24), \underline{E}(t-48), \dots, \underline{E}(t-24D)\}$. This concatenation repeats for all days: if there are N days in the data set, there will be N-D+1 of these input-output pairs. The total number of data points is $n = 24N$. Now, in order to find an estimate for $y(t_{s^*})$ where $t_{s^*} \in t_s$ is an instance of test set indexes, first the distance between $x(t_{s^*})$ and all other $x(t_r)$ that belong to the training set is computed. Distance could be any norm of their difference; however, we have used the Time Weighted Dot Product (TWDP) distance here based on [3]. After determining the closest $x(t_r)$ to $x(t_{s^*})$, the corresponding $y(t_r)$ is generated as $y(t_{s^*})$. Fig. 1 illustrates the algorithm. More detailed discussions can be found in other references [1][2]-[3].

For this algorithm, we need to specify the optimum depth parameter (D) for each outlet; this value has been derived with a cross-validation method suitable for time series [4].

Nearest Neighbor Algorithm

Inputs: $x(t_r), y(t_r), x(t_{s^*})$
Output: $y(t_{s^*})$
1. **for** $j \in t_r$
2. $dist[j] = \|x(t_{s^*}) - x(j)\|$
3. $idx =$ index of smallest$(dist)$
4. $y(t_{s^*}) = y(idx)$

Figure 1. Nearest Neighbor Algorithm.

C. Results and analysis

The training set in our simulations was the first 90% of the data which makes the test set the last 10% of the data. We used five blocks in the cross validation procedure.

Table I shows the average SMAPE on test days for each imputation method and outlet. Among the constant imputation methods, Median imputation results in a lower SMAPE. Also, Multiple imputation has the lowest error; however, since in most missing value cases, all the variables are missing, its performance is not that different from Median imputation.

TABLE I. Average of Smape (%) on test days for each Outlet and Imputation Method

No	Constant (zero)	Mean	Median	Maximum Likelihood	Multiple
1	6.7	6.72	6.68	6.68	6.68
2	6.66	6.74	6.58	6.57	6.55
3	1.93	1.97	1.89	1.88	1.87
4	9.91	9.99	9.83	9.83	9.81
5	24.42	24.43	24.41	24.41	24.41
6	50.56	50.59	50.53	50.52	50.52
7	28.18	28.24	28.12	28.12	28.10
8	22.44	22.47	22.41	22.41	22.40
9	20.63	20.73	20.53	20.52	20.50
10	7.8	7.86	7.74	7.73	7.72
11	18.31	18.31	18.31	18.31	18.31
12	20.87	20.88	20.86	20.86	20.86
13	19.78	19.87	19.69	19.68	19.66
14	9.47	9.56	9.38	9.37	9.36
15	14.05	14.08	14.02	14.02	14.01
Mean	17.45	17.50	17.40	17.39	17.38

All simulations were run with RStudio Version 0.98.507 on an Intel Core i-7 CPU at 3.40 GHz with 16 GB RAM.

IV. Conclusion

Five imputation methods have been applied to replace the missing values of EV charging data time series. Median imputation seems the best among constant imputation methods. Due to lack of the cases where data is partially missing (e.g. voltage is available but not current), Multiple imputation was not that effective.

References

[1] Allison, P. D. (2001) Missing Data Thousand Oaks, CA.

[2] M. Majidpour, C. Qiu, C-Y. Chung, P. Chu, R. Gadh, H. Pota, "Fast Demand Forecast of Electric Vehicle Charging Stations for Cell Phone Application", in *Proc. IEEE/PES General Meeting, 27-31 July 2014, Washington, D.C.,USA*.

[3] M. Majidpour, C. Qiu, P. Chu, R. Gadh, H. Pota, "Modified Pattern Sequence-based Forecasting for Electric Vehicle Charging Stations", in *Proc. IEEE SmartGridComm, 3-6 Nov 2014, Venice, Italy,* in press.

[4] Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", *Journal of Stat. Software*.