Handling Bad or Missing Smart Meter Data through Advanced Data Imputation

Jouni Peppanen, Xiaochen Zhang, Santiago Grijalva
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA

Matthew J. Reno
Sandia National Laboratories
Albuquerque, NM, USA

***Abstract*— Smart meters and other the modern distribution measurement devices provide new and more data, but usually they are subject to longer delays and lower reliability than transmission system SCADA. Accurate and robust use of the modern distribution system measurements will be a cornerstone of the future advanced distribution management systems. This paper presents a novel and computationally efficient data processing method for imputing bad and missing load power measurements to create full power consumption data sets. The imputed data periods have a continuous profile with respect to the adjacent available measurements, which is a highly desirable feature for time-series (power flow) analyses. The method is shown to be superior in accuracy to a utility best practice approach. Our simulations use actual AMI data collected from 128 smart meters on the Georgia Tech campus.**

***Index Terms*— Data Handling, Data Preprocessing, Load Modeling, Power System Measurements, Smart Grids**

I. Introduction

In order to operate distribution systems under pervasive distributed energy resources (DERs), faster and more accurate monitoring, coordination and control are imperative [1]. The increasing DER installments are leading to the deployment of advanced distribution management systems (ADMS) [1], [2] to provide functions such as conservation voltage reduction (CVR), Volt/VAr optimization (VVO), and distribution state estimation [1]–[3]. The emerging data from smart meters and other sensors has the potential to provide information for the new operational needs [1], [4], [5]. However, compared to supervisory control and data acquisition (SCADA) measurements, modern distribution system measurements typically have lower reliability and longer delays. Accurate and robust use of all available measurements will be essential to manage ADMS functions with ubiquitous DERs [6].

Before storing the incoming measurement data to a database, the data must be preprocessed. Typically, the incoming (smart meter) measurement data preprocessing consists of data validation and data imputation [1]. The goal of the data validation process is to identify whether the data correctly represents the measured situation. Following the data validation, the data imputation process estimates values for the identified bad and missing measurements. This paper proposes a novel computationally efficient method for imputing missing and bad measurements in load power measurement data.

This paper has the following structure. Section II briefly introduces typical statistical data imputation methods and methods for estimating smart meter measurements. In section III, the proposed data imputation method is presented. Section IV demonstrates the proposed method with the Georgia Tech AMI data. Section V concludes the paper.

II. Load Power Data Imputation

A. Statistical Perspective on Data Imputation

Methods for handling missing data is a well-established area in statistics [7], [8]. The most common approach to handle missing data entries is to ignore them completely. The "ignoring methods", including list-wise deletion and pairwise deletion, are very easy to implement, but they reduce the amount of usable data and may lead to biased estimates in statistical analyses such as linear regression [7].

Full data sets can be generated by filling in the missing data periods with imputed data [8]. Common data imputation methods are categorized as single imputation (SI), multiple imputation (MI), and maximum likelihood estimation (MLE) [7], [8]. SI methods are the most commonly used approaches to fill in missing values. They fill in precisely one value for each missing one, as opposed to MI methods that generate multiple values for each missing entry to better reflect the uncertainty of the missing data. SI methods, such as replacing the missing values by the mean of available values or using linear regression to estimate the values, are simple to implement, but can lead to biased estimates of certain parameters in statistical modeling such as linear regression [7], [8]. Compared to SI methods, MI and MLE methods have better statistical properties, but require much more computational resources and data and thus, are not practical for imputing the bad and missing measurements in the Big Data provided by smart meters and DER sensors.

B. Load Power Data Imputation

A power industry best practice to impute bad/missing smart meter data is presented in [9]. Intervals shorter than two hours are typically imputed by applying linear interpolation to the surrounding data. For periods longer than two hours, the typical approach is to construct daily load profiles based on previously validated historical data of "like weekdays" and "like days". Holidays and other special cases are typically considered separately.

Load power data imputation is related to (very) short-term load forecasting (STLF) that has been extensively studied in the literature [10]. However, STLF research typically focuses on forecasting the total system load, which is a fairly different

978-1-5090-5167-0/16/$31.00 ©2016 IEEE

problem compared to imputing missing/bad data of individual smart meters or other sensors that can have highly variable measurement profiles. Many STLF approaches also require additional data such as temperature, etc. Smart meter measurements can be used for constructing advanced customer type specific load profiles [11], [12] that can be efficiently applied for imputing bad/missing measurements. However, utilizing an average load profile for each customer segment clearly ignores any customer specific load behaviors and does not account for spatial load characteristics, such as a load that tends to be higher in certain distribution system area. These approaches also do not leverage the adjacent available measurements in data imputation. Although more sophisticated approaches for smart meter data imputation, such as [13], have been proposed, the methods tend to make unrealistic assumptions on the load data characteristics or be impractical to implement. In the future similarly to transmission system state estimation, bad data detection and estimation may be integrated into distribution system state estimation (DSSE). Utilizing AMI data for DSSE has been studied in, e.g., [14], [15]. However, since most utilities have no DSSE today, data imputation remains as a separate process.

III. Optimally Weighted Average Data Imputation Method

This section presents a computationally and data efficient optimally weighted average (OWA) load power data imputation method that is practical for offline and online applications. The method only requires the historical load power measurements from the smart meter (or other sensor). In particular, the method does not require measurement (e.g. customer) specific information or other explanatory variables such as weather. The proposed load data imputation scheme leverages two typical load data characteristics. First, the data tends to be rather continuous over a short time interval, meaning that short time intervals of missing/bad measurement samples have likely similar characteristics as the adjacent available data. Second, since the load data is strongly driven by human consumption patterns, the data tends to have similar characteristics over time periods with similar human activity. For example, the data characteristics of weekdays tend to be different to weekend days, mornings different to evenings, etc.

A. Linear Interpolation Imputation

There are several ways to estimate short intervals of missing samples from the adjacent available samples. Nearest-neighbor and interpolation are particularly commonly used approaches. In the nearest-neighbor approach, the missing samples are simply set equal to the closest available sample or an average of them. For slightly longer missing data periods, interpolation is preferred since it results in estimates that are continuous with the adjacent available measurements. The data imputation method proposed in section III.C. uses linear interpolation since it tends to have more consistent behavior for missing data with different characteristics compared to cubic or other more complicated interpolation methods [16].

Linear interpolation (LI) imputation estimates a missing value y_i from the closest preceding and succeeding available values y_h and y_j with

$$\hat{y}_i^{LI} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h), x_h < x_i < x_j. \quad (1)$$

LI imputation is simple, fast, and requires only two available samples to impute each missing data period. On the other hand, the accuracy of LI imputation typically decreases as the length of the missing data period increases.

B. Historical Average Imputation

LI imputation tends to perform poorly on long periods of missing data, and better estimates can be derived from representative periods of historical data. The simplest approach to impute missing values with historical data is to use the sample from the previous hour, day, or month. Using a single sample however, can result in highly variable estimates whose accuracy may strongly depend on the missing sample times. The data imputation method proposed in section III.C. utilizes historical average (HA) imputation method that estimates each missing sample y_i as an average of N_H representative historical samples $y_j, j \in \mathcal{H}, |\mathcal{H}| = N_H$

$$\hat{y}_i^{HA} = \frac{1}{N_H} \sum_{j \in \mathcal{H}} y_j. \quad (2)$$

To characterize the set $\mathcal{H}$, we define "weeknum" (WN)

$$WN = WD + \frac{HH}{24} + \frac{MM}{24 \times 60}, \quad (3)$$

as a function of the weekday $WD \in \{1,\ldots,7\}$ (1=Monday, …, 7=Sunday), hour of the day $HH \in \{1, \ldots, 24\}$, and minute of the hour $MM \in \{1, \ldots, 60\}$. Now, the set $\mathcal{H}$ is defined to consist of historical samples whose day of the year (DOY) and WN are within selected spans of the missing sample. In this paper, the DOY span of ± 8 days and the WN span of $\pm 1/24 + 1/(24 \times 60)$ (1 hour and 1 minute) were used. The DOY assures that the historical mean is calculated over samples with similar seasonal characteristics. The WN guarantees that the historical mean is calculated over samples with similar days of the week and times of the day. Holidays and other special days are handled separately or if sufficient data is not available for them, they are categorized as Sundays. This definition of $\mathcal{H}$ results in smooth historical average profiles for sequential missing samples. If "hard" time selection criteria, such as equal season, equal WD, and equal HH was used, the sequential imputed samples would have jumps when the season, weekday, hour, etc. change.

The accuracy of the HA imputation depends on the characteristics of the data and requires clear historically repeating patterns. With these assumptions, on long missing data periods, HA imputation is expected to have a better average performance compared to LI imputation.

C. Optimally Weighted Average Imputation

Next, an optimally weighted average (OWA) imputation method is presented with the objective of leveraging the LI imputation accuracy for short missing data periods and the HA imputation accuracy for longer missing data periods. The OWA imputation estimates a missing data sample y_i as the weighted average of the LI imputed values $\hat{y}_i^{LI}$ and the HA imputed values $\hat{y}_i^{HA}$

978-1-5090-5167-0/16/$31.00 ©2016 IEEE

$$\hat{y}_i^{OWA} = w_i \hat{y}_i^{LI} + (1 - w_i)\hat{y}_i^{HA}. \quad (4)$$

The weight parameter w_i is set to exponentially decay with respect to $d_i > 0$, the (positive) distance (in samples) to the closest (preceding or succeeding) available sample

$$w_i = e^{-\alpha d_i}, \quad (5)$$

where α is a (positive) weight parameter. For small d_i (i.e. $w_i \approx 1$), the OWA imputed value $\hat{y}_i^{OWA}$ mainly depends on the LI imputed value $\hat{y}_i^{LI}$. For large d_i (i.e. $w_i \approx 0$), the OWA imputed value $\hat{y}_i^{OWA}$ depends mainly on the HA imputed value $\hat{y}_i^{HA}$. Figure 1 illustrates the weight function w_i dependence on α and d_i. For $\alpha > 2$, the HA imputed values are almost exclusively used for all but the first missing sample. Thus, it is reasonable to restrict $\alpha \in [0,2]$. The optimal value of α depends on the measurement data characteristics including the variability and the historical patterns of the data. The question remains about what value of α to select, so next, a method to optimize α is presented.

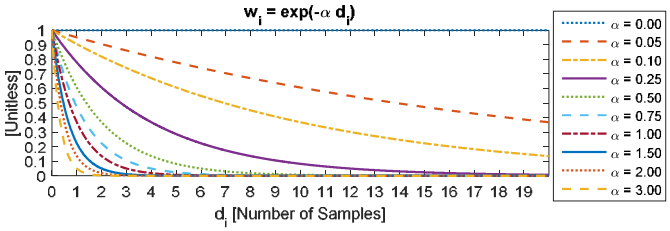

Figure 1. Optimally weighted average imputation weight function shape for different weight parameters and distance to the closest available sample

D. Optimal Weight Parameter for A Training Data Period

The optimal weight parameter (for a training data period) α_{opt} minimizes the error $F(\alpha)$ between the imputed samples and the training data samples

$$\alpha_{opt} = \underset{\alpha}{\operatorname{argmin}}\, F(\alpha) = \operatorname{argmin}_{\alpha} \sum_{i=1}^{N} F_i(\alpha). \quad (6)$$

With squared error, $F_i(\alpha)$ is given by

$$F_i(\alpha) = \left(\hat{y}_i^{OWA} - y_i^{true}\right)^2 = (e^{-\alpha d_i}\delta_i^{LH} + \delta_i^{HA})^2 \quad (7)$$

where $\delta_i^{LH} = \hat{y}_i^{LI} - \hat{y}_i^{HA}$ and $\delta_i^{HA} = \hat{y}_i^{HA} - y_i^{true}$. A necessary condition for an optimal solution α_{opt} is that the derivative vanishes $F'(\alpha) = 0$. Such so-called critical points can be found, e.g., with the Newton's method starting at initial value $\alpha = \alpha_0$ and iterating with

$$\alpha_{k+1} = \alpha_k - \frac{F\prime(\alpha_k)}{F\prime\prime(\alpha_k)} \quad (8)$$

until a selected convergence criteria is satisfied. The error function $F(\alpha)$ is nonconvex for any set of training samples y_i^{true} and imputed samples $\hat{y}_i^{LI}$ and $\hat{y}_i^{HA}$ that result in $F''(\alpha) > 0$. As a result, Newton's method may diverge for a poorly chosen α_0. In practice, good convergence is obtained by selecting a small (but positive) α_0 (e.g., $\alpha_0 = 0.001$).

Figure 2 shows an example of a training data period of 50 samples with the true known values and the values estimated with HA, LI, WA ($\alpha = 0.10$), and OWA imputation. Clearly, for such a long time period, the LI imputation accuracy suffers. Better imputation accuracy is achieved with a linear combination of LI and HA imputation (WA) and best accuracy is obtained with the optimal weight parameter (OWA).

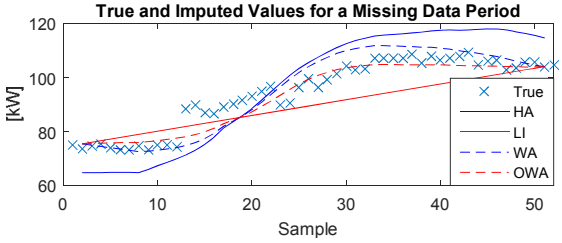

Figure 2. An example of a training data period with the true known values and the HA, LI, WA ($\alpha = 0.10$), OWA ($\alpha_{opt} = 0.040$) imputed values

E. Globally Optimal Weight Parameter

The optimal weight parameter α_{opt} depends on the characteristics and the length of the missing period. Thus, different α_{opt} values are obtained using different training data period characteristics and lengths. The distribution of α_{opt} can be estimated by optimizing α over a set of training data periods with randomly selected lengths and locations. The missing data period lengths can be sampled from known distribution of missing data period lengths (if available). The globally optimal α can be estimated from the mean (or median) of the obtained α_{opt} sample distribution.

Algorithm 1 lists the process of estimating the weight parameter α_{opt} for a meter. The optimal weight parameter α_{opt} of a meter is optimized only once and is stored in the MDMS. Afterwards, missing data is estimated with (4) using the optimized α_{opt}. The results shown in section IV indicate that good estimate of α_{opt} distribution can be obtained with $N_{period} = 100$ for typical missing data period lengths. If dealing with a large number of meters, Algorithm 1 can be executed for a subset of the meters and the mean (or median) of the resulting α_{opt} distribution can be utilized for all meters.

Algorithm 1: OWA Weight Parameter Optimization

1. Randomly choose the first samples of the training data periods for N_{period} training data periods and $N_{lengths}$ training data period lengths.
2. Construct an array of timestamps of all the samples needed for imputing the training data samples with HA and LI imputation.
3. Fetch the samples with the timestamps from the MDMS.
4. Repeat **1. – 3.** for periods with (true) bad/missing samples.

FOR $N_{lengths}$ training data period lengths

5. For each sample of each training data period, impute the values $\hat{y}_i^{HA}$ and $\hat{y}_i^{LI}$ and calculate δ_i^{LH}, δ_i^{HA}, and d_i.
6. Use (6)-(8) to find α_{opt} that minimizes $F(\alpha)$ over all missing data periods. Store α_{opt}.

ENDFOR

7. Choose the globally optimal α, e.g., as the mean (or median) of the distribution of α_{opt} values.

978-1-5090-5167-0/16/$31.00 ©2016 IEEE

IV. Imputation on Georgia Tech AMI Data

Georgia Tech owns and maintains its electricity distribution system serving more than 200 campus buildings. The measurements from the approximately 400 revenue-grade smart meters in the buildings are recorded and aggregated into a database every 15 minutes [17]. Next, the OWA data imputation method is shown for smart meter measurements from the Georgia Tech distribution system.

A. Detailed Analysis for A Georgia Tech Smart Meter

The OWA data imputation was first analyzed with the active power measurements of one of the Georgia Tech smart meters. The analyzed smart meter is located in a building that is mainly dedicated for classroom and office purposes. As a result, the building energy consumption has a clear historical pattern driven by the classroom and office activity as illustrated for a two-week period in 2013 in Figure 3.

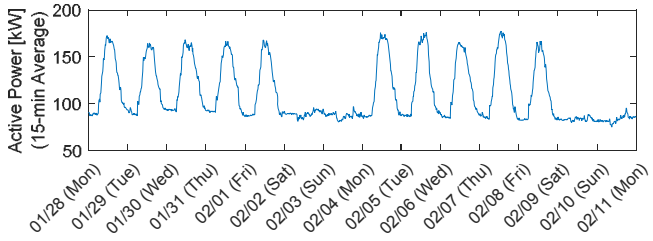

Figure 3. The 15-min average active power measurements for the analyzed Georgia Tech building from January 28, 2013 through February 11, 2013

First, α_{opt} of the meter was searched with Algorithm 1 using $N_{length} = 29$ training data period lengths varying from 3 (45 minutes) to 100 (25 hours) each with $N_{period} = 100$ randomly chosen period locations. For each period length, α_{opt} was solved to minimize the imputation error over the period locations. This resulted in 29 α_{opt} values. To get a better estimate of the α_{opt} distribution, this process was repeated 100 times resulting in a 100×29 array of α_{opt} values. The overall average α_{opt} was 0.1387. The distribution of the 100 α_{opt} values for each of the 29 training period lengths is visualized in Figure 4. For missing data period lengths above 8 (2 hours), the majority of the α_{opt} values are between 0 and 0.4. For short missing data period lengths, smaller α_{opt} is preferred effectively putting more emphasis on the linear interpolation.

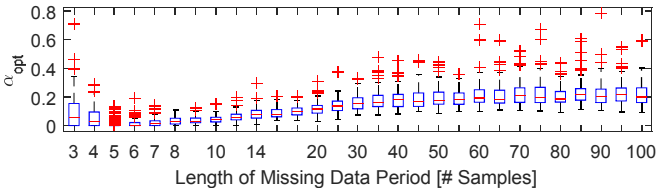

Figure 4. The boxplots of 100 α_{opt} values for each of the 29 different missing data period lengths for the analyzed Georgia Tech smart meter

Next, the sample mean $\alpha_{opt} = 0.1387$ was utilized to compare the performance of the OWA imputation to HA, LI, and industry best practice (BP) imputations. As discussed in Section II, BP imputation uses LI imputed values for missing data periods shorter than 2 hours and an average of three previous days for missing data periods above 2 hours [9]. The validation was done for $N_{length} = 29$ missing data period lengths (same as used for α_{opt} training) each with 50 randomly chosen period locations. The period locations were chosen independent of the period locations used for α_{opt} training. For each period length, a mean absolute percentage error ($MAPE = \frac{1}{N}\sum_{i=1}^{N_{sample}} \left|\hat{y}_i^{imputed} - y_i^{true}\right|$) was calculated over the N_{sample} samples of the 50 missing data periods. This resulted in 29 MAPE values. To obtain a more stable estimate of the MAPE distribution, the process was repeated 100 times resulting in a 100×29 array of MAPE values. Figure 5 illustrates the distribution of the 100 MAPE values for each 29 validation data period lengths. On average, OWA outperforms HA, LI, and BP imputations for all missing data period lengths. Compared to HA and LI imputation, the advantage of OWA imputation is greater for short and long periods, respectively. For periods under 2 hours, OWA operates fairly similarly to BP but for periods over 2 hours, OWA outperforms BP imputation. Only average MAPE reduction can be expected since no imputation method is guaranteed to be effective for all missing data period lengths and characteristics.

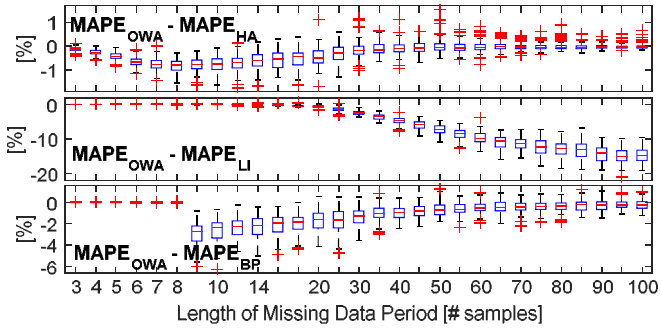

Figure 5. The boxplots of 100 MAPE differences between OWA imputation method and HA, LI, and BP imputation methods for each of the 29 different missing data period lengths for the analyzed Georgia Tech smart meter

B. Results for 128 Georgia Tech Smart Meters

Next, Algorithm 1 was used to search the α_{opt} for 128 Georgia Tech smart meters. For each meter, Algorithm 1 was executed with $N_{length} = 29$ training data period lengths (same as in Section IV.B.) and $N_{period} = 100$ randomly chosen missing data periods. This resulted in a 128×29 array of α_{opt} values. The average α_{opt} (over all meters and all training data period lengths) was 0.1081. Figure 6 visualizes the distribution of α_{opt} values for different training data period lengths and meters. As shown in the top plot of Figure 6, median α_{opt} seems to be relatively independent of the training data period lengths except for very short period lengths less than 8 (2 hours) for which α_{opt} seems to be slightly higher. The bottom plot of Figure 6 indicates that α_{opt} takes similar values for most meters but that there are also meters for which $\alpha_{opt} = 0$ or $\alpha_{opt} = 2$ for many training data period lengths. For these meters, better imputation accuracy can be achieved by solely using LI imputation or HA imputation, respectively.

Next, the overall average $\alpha_{opt} = 0.1081$ was utilized to compare the performance of the OWA imputation to the HA,

978-1-5090-5167-0/16/$31.00 ©2016 IEEE

LI, and industry best practice (BP) imputations. The validation was done for the same $N_{length} = 29$ missing data period lengths each with 100 period locations that were chosen randomly and independent of the period locations used for α_{opt} training. The distribution of MAPE differences between the OWA and the HA, LI, and BP imputations are illustrated in Figure 7. Ignoring outliers ($\geq 10\%$ and $\leq -10\%$), the average (over all missing data period lengths and meters) MAPE reductions and the respective 95% confidence intervals of the OWA approach compared to the HA, LI, and BP imputation methods were (-0.8070±0.0189)%, -(0.9831±0.0381)%, and, (-1.8592±0.0520)%, respectively.

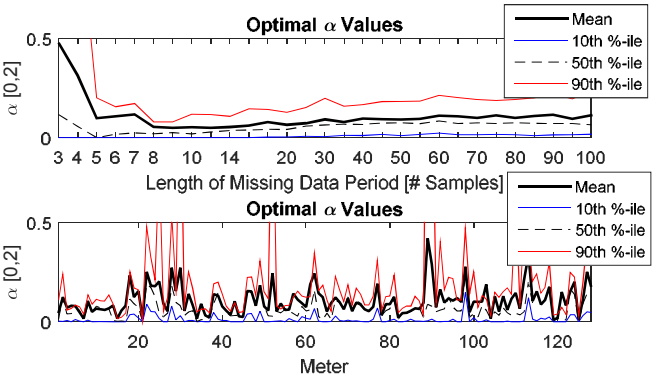

Figure 6. The percentiles of α_{opt} distribution for different training data period lengths (top) and for the analyzed 128 Georgia Tech smart meters (bottom)

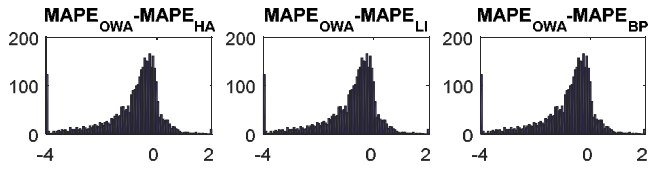

Figure 7. The histograms of MAPE differences between OWA imputation method and HA, LI, and BP imputation methods

The level of MAPE reduction varies among meters but compared to HA, LI, and BP, OWA reduces the average (over the 29 missing data period lengths) MAPE values for 93.0%, 79.0%, and 93.0% of all the meters, respectively. Figure 8 visualizes the MAPE reductions for different missing data period lengths. Compared to HA, LI, and BP, OWA achieves smaller average MAPE values for medium, long, and long period lengths, respectively.

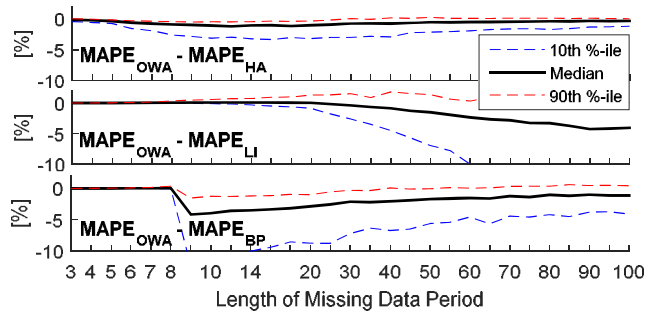

Figure 8. The percentiles of MAPE differences between OWA imputation method and HA, LI, and BP imputation methods for different missing data period lengths for the 128 analyzed Georgia Tech smart meters

V. Conclusions

While modern distribution system measurement sources such as AMI and DER sensors provide more data, they are typically subject to longer delays and have lower reliability than transmission system SCADA. This paper presents a novel load power data imputation method, which could be beneficial to support advanced DMS functions. The imputed data periods have a continuous profile with respect to the adjacent available measurements, which is a highly desirable feature for time-series (power flow) analyses. The method outperforms conventional linear interpolation, historical average, and an industry best practice imputation approaches in imputing Georgia Tech AMI measurements. The weight parameter of the developed imputation method is trained offline after which the method is computationally and data efficient making the method suitable for both online and offline settings.

References

[1] J. Fan and S. Borlase, "The evolution of distribution," *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 63–68, Mar. 2009.

[2] "VOICES of Experience - Insights into Advanced Distribution Management Systems," Feb. 2015.

[3] T. Taylor and H. Kazemzadeh, "Integrated SCADA/DMS/OMS: Increasing Distribution Operations Efficiency," *Electric Energy Online*, Apr-2009.

[4] J. S. John, "Can Microinverters Stabilize Hawaii's Shaky Grid? : Greentech Media," 02-Feb-2015. [Online]. Available: http://www.greentechmedia.com/articles/read/enphase-to-help-hawaii-ride-its-solar-energy-wave.

[5] R. F. Arritt and R. C. Dugan, "Distribution System Analysis and the Future Smart Grid," *IEEE Trans. Ind. Appl.*, vol. 47, no. 6, 2011.

[6] "Report on Use of Distribution State Estimation Results for Distribution Network Automation Functions, D2.2.1," BU, EF, Fraunhofer IWES, INDRA, Korona, FP7 - 248135, Apr. 2011.

[7] P. Allison, "Missing Data," 2001.

[8] X.-H. Zhou, *Applied missing data analysis in the health sciences*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014.

[9] "Uniform Business Practices for Unbundled Electricity Metering - Volume 2," Edision Electric Institute, Dec. 2000.

[10] T. Hong, "Short Term Electric Load Forecasting," North Carolina State University, 2010.

[11] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Jarventausta, "Enhanced Load Profiling for Residential Network Customers," *IEEE Trans. Power Deliv.*, vol. 29, no. 1, Feb. 2014.

[12] B. Stephen and S. J. Galloway, "Domestic Load Characterization Through Smart Meter Advance Stratification," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1571–1572, Sep. 2012.

[13] G. Mateos and G. B. Giannakis, "Load Curve Data Cleansing and Imputation Via Sparsity and Low Rank," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2347–2355, Dec. 2013.

[14] E. Manitsas, R. Singh, B. Pal, and G. Strbac, "Modelling of pseudo-measurements for distribution system state estimation," in *IET-CIRED SmartGrids for Distribution*, Frankfurt, Germany, 2008.

[15] X. Feng, F. Yang, and W. Peterson, "A practical multi-phase distribution state estimation solution incorporating smart meter and sensor data," in *IEEE Power and Energy Society General Meeting*, San Diego, CA, 2012.

[16] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmos. Environ.*, vol. 38, no. 18, Jun. 2004.

[17] J. Peppanen, M. J. Reno, M. Thakkar, S. Grijalva, and R. G. Harley, "Leveraging AMI Data for Distribution System Model Calibration and Situational Awareness," *IEEE Trans. Smart Grid*, 2015.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

978-1-5090-5167-0/16/$31.00 ©2016 IEEE