Missing Data Handling for Meter Data Management System*

Ru-Sen Jeng[1], Chien-Yu Kuo[1], Yao-hua Ho[1], Ming-Feng Lee[2], Lin-Wen Tseng[2],
Chia-Lin Fu[2], Pei-Fang Liang[2], and Ling-Jyh Chen[3]
[1] Department of Computer Science and Information Engineering, National Taiwan Normal University
[2] Green Energy and Environment Research Lab, Industrial Technology Research Institute
[3] Institute of Information Science, Academia Sinica

ABSTRACT

We study the meter data management systems (MDMS) with a focus on missing data handling, and propose two approaches, called Lookback-N and Sandwich-N, based on the historical data. Using a realistic dataset, we demonstrate that Lookback-N is effective for online processing, and Sandwich-N outperforms the conventional offline approach, in terms of estimation accuracy, in all test cases. The proposed approaches are simple, effective, and show promise in handling missing data for emerging smart meter data management systems.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support (e.g., MIS)*

Keywords

Missing data, Smart meter, MDMS

1. INTRODUCTION

In this study, we tackle the meter data management systems (MDMS) with a focus on missing data handling. The issue is challenging because 1) it has to be accurate as it is the basis for electricity billing, planning, and provisioning; and 2) it has to be simple in computation to support large-scale deployments. As a consequence, conventional missing data handling methods are not applicable to smart meter data management because 1) they are based on sophisticated statistical models that are computationally expensive [3, 5]; and 2) they require additional knowledge of environment and social factors [1, 4] that are infeasible in real deployments.

In the following sections, we propose two approaches to deal with missing data for meter data management systems, namely, the *Lookback-N* and *Sandwich-N* schemes. Using a realistic dataset, we evaluate the proposed schemes against the baseline scheme, which is based on linear interpolation and has been widely used by electric power companies. The results demonstrate that 1) the *Lookback-N* scheme can support *online processing* while keeping estimation error acceptable; and 2) the *Sandwich-N* scheme can yield a smaller estimation error than the baseline scheme, especially when missing data takes place in a burst.

2. PROPOSED APPROACHES

We propose the *Lookback-N* scheme and the *Sandwich-N* scheme to deal with missing data for meter data management systems. Specifically, let M_t^i be the reading of the i-th meter on time t; and let δ_t^i be a status variable that is equal to 0 if M_t^i is missing and equal to 1 otherwise. Suppose the reading of the i-th meter is missing at time t_k (i.e., $\delta_{t_k}^i = 0$), and the last/next non-missing reading before/after $M_{t_k}^i$ is on time t_x and t_y respectively. Let $\widetilde{M}_{t_k}^i$ be the estimate of $M_{t_k}^i$, the conventional approach (i.e., the baseline scheme) implements the linear interpolation method to obtain $\widetilde{M}_{t_k}^i$, i.e., $\widetilde{M}_{t_k}^i = \frac{M_{t_x}^i + M_{t_y}^i}{2}$.

2.1 The Lookback-N Approach

The rationale of the *Lookback-N* scheme is that '*similar things come together*'. Let $f(i, t_u, j, t_v, n)$ be a matching function that returns 1 when the last n readings of the i-th meter since t_u are the same as the last n readings of the j-th meter since t_v; and it returns 0 otherwise. There are two cases in *Lookback-N* to estimate $M_{t_k}^i$:

1. When $t_x = t_k - 1$, it looks up the historical data to identify the set of the 'similar' meter readings that have the same values in its preceding N readings as $M_{t_k}^i$ by
$$R_{t_k}^i = [M_{t_u}^j | \forall j \forall t_u \in [t_k - L, t_k) : f(i, t_k, j, t_u, N) = 1], \quad (1)$$
where L is the *lookback* factor that determines the length of the historical data to look up. Then, it estimates $\widetilde{M}_{t_k}^i$ based on $R_{t_k}^i$ using *roulette-wheel selection* [2].
2. When $t_x < t_k - 1$, it estimates $\widetilde{M}_{t_x+1}^i$ using the procedure used in the above case, and then it uses the estimated value to estimate $\widetilde{M}_{t_v}^i$, for $t_v = t_x + 2, ..., t_k$.

2.2 The Sandwich-N Approach

The *Sandwich-N* scheme is similar to *Lookback-N*, except that it considers both the preceding and succeeding N readings of the missing one when identifying the set of 'similar' meter readings. We let $g(i, t_u, j, t_v, n)$ be a matching function that returns 1 when the preceding and succeeding n readings of the i-th meter since t_u are the same as those of the j-th meter since t_v, and it returns 0 otherwise.

When $t_x = t_k - 1$, the scheme identifies the 'similar' set using Eq. 2, and it uses *roulette-wheel selection* to estimate $\widetilde{M}_{t_k}^i$ based on $S_{t_k}^i$. When $t_x < t_k - 1$, it estimates $\widetilde{M}_{t_x+1}^i$ first, and then it uses the estimated value to estimate $\widetilde{M}_{t_v}^i$, for $t_v = t_x + 2, ..., t_k$, iteratively.

$$S_{t_k}^i = [M_{t_u}^j | \forall j \forall t_u \in [t_k - L, t_k) : g(i, t_k, j, t_u, N) = 1]. \quad (2)$$

*The authors wish to express their sincere gratitude to Bureau of Energy, Ministry of Economic Affairs, Taiwan, for financial support.

e-Energy'13, May 21–24, 2013, Berkeley, California, USA.
ACM 978-1-4503-2052-8/13/05.

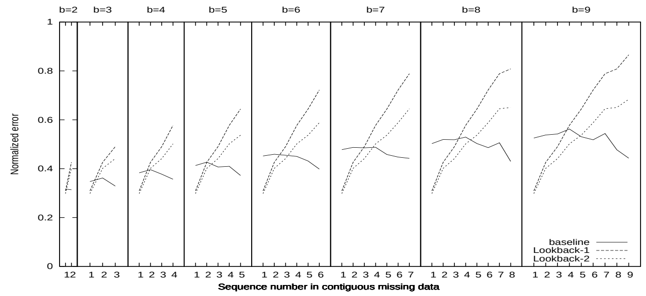

Figure 1: Comparison of the *Lookback-N* scheme and the baseline scheme under contiguous missing data

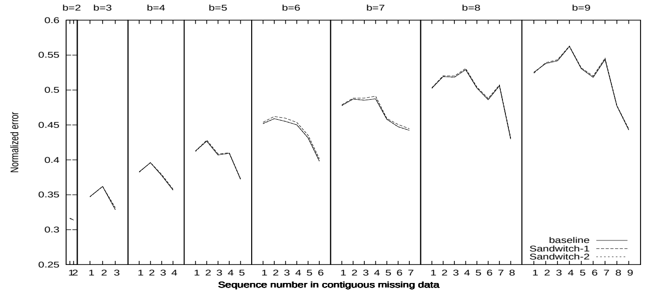

Figure 2: Comparison of the *Sandwich-N* scheme and the baseline scheme under contiguous missing data

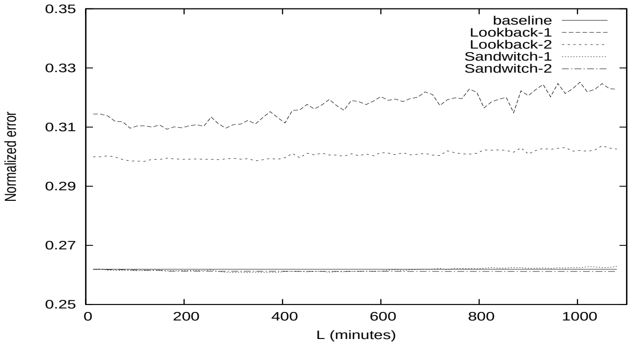

Figure 3: Comparison of the different schemes under different lengths of historical data to look up (i.e., L)

3. EVALUATION

We evaluate the proposed schemes using a real dataset from the Pilot Smart Meter Deployment in Taiwan. The dataset comprises eleven household electricity meters, from 2011/6/30 to 2011/9/30, in the same building in Taipei, Taiwan; and the sample rate of each meter is one sample every 15 minutes. There are no missing data in the dataset, and we suppose the energy consumption behavior of each household is unchanged in the measurement period.

We randomly generate different lengths of contiguous missing data in the dataset, and observe the normalized error (i.e., the estimation error over the original value) achieved by each scheme. Figures 1 and 2 show the average results based on 10,000 runs for each setting, and there are four observations:

1. The error achieved by *Lookback-N* is greater than that by the other schemes, but it remains affordable for *online processing* in meter data management systems.

2. When *Lookback-N* is used, the error accumulates with the sequence number of missing data in a burst, and the error is smaller than that of the baseline scheme when estimating the first few missing data in a burst.

3. When *Sandwich-N* or the baseline scheme is used, the error increases with the sequence number of the contiguous missing data.

4. *Sandwich-N* is comparable to the baseline scheme when dealing with contiguous missing data. The error achieved by the two schemes is greater when estimating the middle few missing data, and it is smaller when estimating the last few missing data in a burst.

Figure 3 shows the normalized error achieved by different schemes under different lengths of historical data to look up (i.e., L). We observe that the two *Lookback-N* curves (N = 1 and 2) have greater normalized errors, and the error increases with L. The reason is that, when N is small or when L is large, *Lookback-N* is more likely to yield an estimate based on wrong energy consumption behavior models (e.g., it may estimates the missing data on working hours based on the non-working hours behavior model), thereby leading to a greater estimation error.

Moreover, we also find that the estimation error is smaller when a large N is used in *Sandwich-N*. As shown in Figure 3, the normalized error of *Sandwich-2* is consistently lower than *Sandwich-1* and the baseline scheme. However, we note that the larger the value of N used in *Sandwich-N*, the fewer 'similar' smart meters might be included in $S^i_{t_k}$, leading to bias in estimation. Moreover, the larger the value of L is, the higher computational complexity is resulted for both *Lookback-N* and *Sandwich-N*.

4. CONCLUSION

In this paper, we propose two missing data handling schemes, namely, *Lookback-N* and *Sandwich-N*, for emerging smart electricity meter systems. Using a realistic dataset, we show that *Sandwich-N* can achieve the lowest estimation error in all settings, and *Lookback-N* is effective for online missing data handling, as well as estimating the first few missing data in a burst. Work on considering additional factors when identifying 'similar' meter data is ongoing, and we plan to report the results in the near future.

5. REFERENCES

[1] J.-Y. Fan and J. D. McDonald. A Real-Time Implementation of Short-Term Load Forecasting for Distribution Power System. *IEEE Transactions on Power Systems*, 9(2):988–994, May 1994.

[2] D. E. Goldberg. *Genetic algorithms in search, optimisation, and machine learning*. Addison Wesley Longman, Inc., 1989.

[3] N. J. Horton and K. P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, February 2007.

[4] D. Matheson, C. Jing, and F. Monforte. Meter Data Management for the Electricity Market. In *International Conference on Probabilistic Methods Applied to Power Systems*, 2004.

[5] T. D. Pigott. A Review of Methods for Missing Data. *Educational Research and Evaluation*, 7(4):353–383, 2001.