Addressing Data Veracity in Big Data Applications

Saima Aman†, Charalampos Chelmis‡, Viktor Prasanna‡
†Department of Computer Science ‡Department of Electrical Engineering
University of Southern California, Los Angeles, CA
Email: {saman, chelmis, prasanna}@usc.edu

***Abstract*—Big data applications such as in smart electric grids, transportation, and remote environment monitoring involve geographically dispersed sensors that periodically send back information to central nodes. In many cases, data from sensors is not available at central nodes at a frequency that is required for real-time modeling and decision-making. This may be due to physical limitations of the transmission networks, or due to consumers limiting frequent transmission of data from sensors located at their premises for security and privacy concerns. Such scenarios lead to partial data problem and raise the issue of *data veracity* in big data applications. We describe a novel solution to the problem of making short term predictions (up to a few hours ahead) in absence of real-time data from sensors in Smart Grid. A key implication of our work is that by using real-time data from only a small subset of *influential* sensors, we are able to make predictions for *all* sensors. We thus reduce the communication complexity involved in transmitting sensory data in Smart Grids. We use real-world electricity consumption data from smart meters to empirically demonstrate the usefulness of our method. Our dataset consists of data collected at 15-min intervals from 170 smart meters in the USC Microgrid for 7 years, totaling 41,697,600 data points.**

***Keywords*—data veracity, prediction model, smart grid**

I. Introduction

Low cost wireless sensors are increasingly being deployed in large numbers for tracking, monitoring, and control in emerging big data applications such as in smart electric grids, transportation, and remote medical and environment monitoring. Examples of such sensors include sensors for monitoring climate features such as temperature and green-house gas measurements [8]; smart meters for measuring energy consumption [15], [9]; and loop detectors installed under pavements for recording traffic [13]. Sensor based big data applications often encounter problems with respect to *availability* and *timeliness* of data [5], where only partial data from sensors is available in real-time, and complete high resolution data is available only after certain periods. This may be due to *physical limitations* of existing transmission networks, such as latency, bandwidth and high energy consumption [4], or due to consumers opting out of or limiting frequent transmission of information from sensors located at their premises for *security and privacy* concerns [10]. For example, fine-grained electricity consumption data collected through smart meters can be used to infer activities of the consumers and also indicate the presence or absence of dwellers in the consumer premises [11].

Forecasting models in such applications often face challenges due to their assumption of data availability in real time being invalidated. While volume, velocity, and variety characterize the qualitative aspect of big data, *veracity*, refers to its quantitative aspect. Without addressing veracity, big data solutions risk degradation in performance and inaccurate interpretation of generated insights. We describe our novel approach to address the problem of data veracity raised in the context of Smart Electricity Grids where high volume electricity consumption data is collected by smart meters at consumer premises and securely transmitted back to the electric utility over wireless or broadband networks to be used for forecasting [2]. Due to physical limitations of existing transmission networks, data from all smart meters is not readily available in real-time [3],[9]. Instead, data from a subset of meters is transmitted in real-time, whereas data from the rest of the meters is collected locally and transmitted in batches periodically (Fig. I).

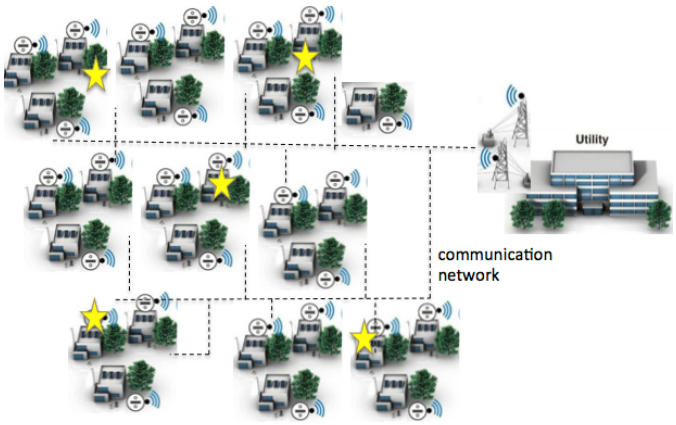


Fig. 1. In a smart grid network, data from a subset of meters (shown starred) is transmitted in real-time, whereas for the rest of the meters, it is collected locally and transmitted in batches periodically.

Prior work in this area has focused on (i) reducing communication requirements by developing sampling or compression strategies, or methods to estimate missing real-time data [14], [6], and (ii) estimating missing real-time data by techniques such as regression-based interpolation [7]. We use a different approach where instead of trying to estimate the missing real-time data, we try to make predictions using partial real time data by learning dependencies among time series originating from different sensors.

II. Problem Formulation

Consider a large set of sensors $\mathcal{S} = \{s_1, ..., s_n\}$ deployed in a big data application producing time series output in form of ordered sequence of readings $\mathcal{T}_i = \{x_j^i\}, j = 1, ..., t$. Some of these sensors can send data back to a central node in real-time, while the rest of the sensors send back data periodically.

Our goal is to use this *partial data* to make predictions for *all* sensors for a given prediction horizon h. Given a set of sensor time series outputs $\{x_j^i\}, j = 1, ..., t, i = 1, ..., n$, short-term prediction is to estimate $\{x_j^i\}, j = t+1, ..., t+h, i = 1, ..., n$ for a horizon h, which is a few hours ahead.

We formulate the problem of short term prediction with partial data as follows: Given a set of sensors $\mathcal{S}$ with time series outputs $\{x_j^i\}, j = 1, ..., t, i = 1, ..., n$, make short-term predictions $\{x_j^i\}, j = t+1, ..., t+h, i = 1, ..., n$ for each sensor $s_i \in \mathcal{S}$, when readings $\{x_k^o\}, k = t-r+1, ..., t$ for $o \in \mathcal{O}$ are missing for a subset $\mathcal{O}$ of sensors, $\mathcal{O} \subset \mathcal{S}$.

III. Veracity Aware Short-term Prediction

We propose a two-stage solution for building a short term electricity consumption prediction model that needs to collect real-time data from only a small subset of smart meters selected on the basis of causal influence. First, we learn temporal correlations among historic time series data collected from smart meters. Then, we build our predictive model leveraging such discovered temporal dependencies.

Influence Discovery: We cast the problem of short term prediction with partial data as a regression problem. In regression, given data $(\mathbf{x}^i, y_i), i = 1, 2, ..., n$, the response y_i for the i^{th} observation is estimated in terms of p predictor variables, $\mathbf{x}^i = (x_{i1}, ..., x_{ip})$ by minimizing the residual squared error. For prediction with partial data, the predictors for s_i are sequences $\{\mathcal{P}_k\}_{k \neq i}$ of past values from *other* sensors, excluding the sensor's own past values, which are not available in real-time. We used *lasso* to identify sensors with strong *influence* on the given sensor s_i and leave out others. Lasso is known to improve variance and reduce overall prediction errors by shrinking or reducing to zero some coefficients [16].

Given n sensor outputs in form of time series $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n$, with readings at timestamps $\mathbf{t} = 1, ...T$, for each series $\mathbf{x}^i$, lasso gives a sparse solution for coefficients $\mathbf{w}$ by minimizing the sum of squared error and a constant times the L1-norm of the coefficients:

$$\mathbf{w} = \arg\min \sum_{t=l+1}^{T} \left\| x_t^i - \sum_{j=1}^{n} \mathbf{w}_{i,j}^T \mathcal{P}_t^j \right\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad (1)$$

where $\mathcal{P}_t^j$ is the sequence of past l readings, i.e., $\mathcal{P}_t^j = [x_{t-l}^j, ..., x_{t-1}^j]$, $\mathbf{w}_{i,j}$ is the j-th vector of coefficients $\mathbf{w}_i$ representing the dependency of series i on series j, and λ is a parameter which determines the sparseness of $\mathbf{w}_i$ and can be determined using cross-validation method.

Influence Model (IM): We work on day long windows of data to ensure stationarity within each window, implying that the dependence on the preceding values does not change with time. We discover influence and train our model on a previous similar day. We consider two cases of similarity: *previous week*, which is the same day of the week in preceding week and *previous day*, which is the day preceding the given day. Our model is formally described below:

1) Split the readings for each smart meter into a set of daily series $\{\mathcal{D}_j^i\}_{i=1,...,n, j=1,...,q}$.
2) Define dependency matrix $\mathcal{M}_|$ for each day using the weight vectors $\mathbf{w}_i$ in equation 1 as its rows. Set the diagonal of the dependency matrix to zero, i.e., $\mathcal{M}[i,i] = 0$, for each i.
3) Define a regression tree for each smart meter that uses predictors from all meters with non-zero coefficients in the dependency matrix learned from a similar day, i.e., predictors are taken from $\{\mathcal{D}^k\}, \forall k : \mathcal{M}_{sim}[i,k] \neq 0$.

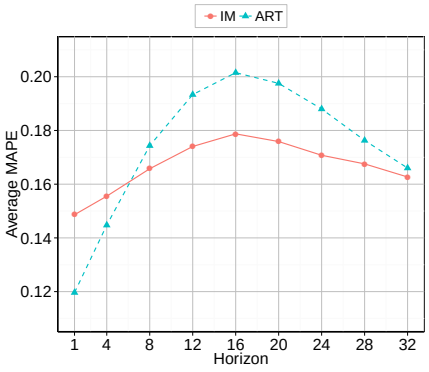


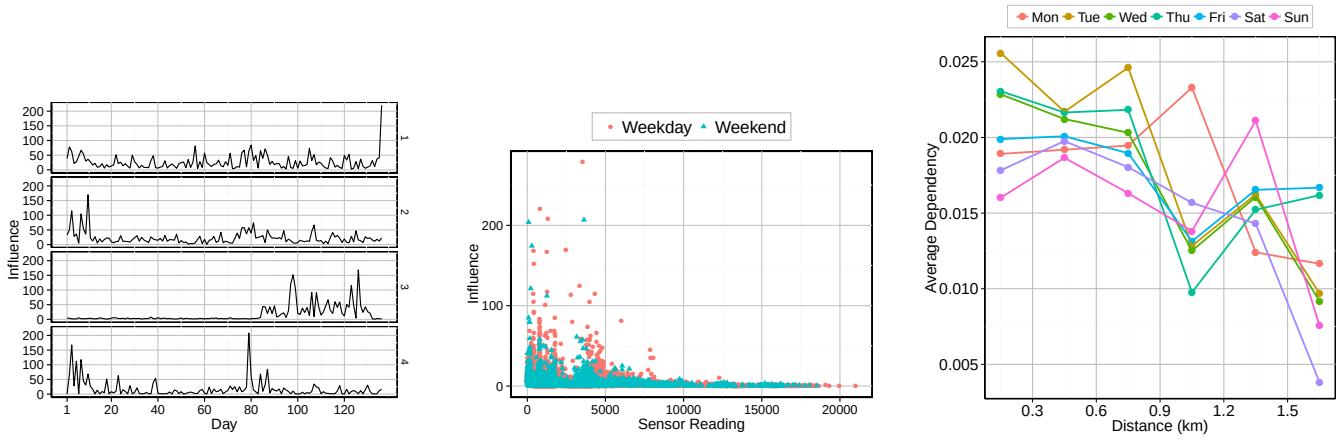
Fig. 3. Prediction performance of influence model with respect to ART.

A key benefit of this model is that we are able to make predictions for a sensor in absence of its own past values by using past values of its influential sensors.

IV. Experimental Evaluation

We evaluate the feasibility and accuracy of our proposed method in a real-life cyber-physical system, while at the same time quantifying scalability savings. We use electricity consumption data collected by smart meters installed in the USC campus microgrid [15] in Los Angeles, and weather data taken from NOAA's USC campus station. For evaluation, we use the Auto-Regressive Tree (ART) as the baseline. ART uses recent values as features in a regression tree model and has been shown to offer high predictive accuracy on a large range of datasets [12]. We implement a specialized $ART(p, h)$ model that uses recent p values of a variable for making h interval ahead prediction. ART was a natural choice for baseline comparison, as our proposed model is also based on regression tree. However, it is to be noted that while ART model uses a variable's own recent observations, our model only uses other variables' observations to make predictions. We used Mean Absolute Percentage Error (MAPE $= \frac{1}{n}\sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{x_i}$) as the evaluation metric as it is scale-independent [1], and allows comparison across different ranges.

We use the *influence model (IM)* for making predictions for different prediction horizons up to 8 hours ahead (Fig. IV). The baseline ART model performs well up to 6 intervals (1.5 hour) due to very-short-term prediction horizon, where electricity consumption is not expected to drastically change from its previous 4 values. Beyond that, for ART, recent values used as predictors at the time of prediction become increasingly ineffective for longer horizons, when IM's use of more recent real-time values of other sensors become more



(a) (b) (c)

Fig. 2. Variance of Influence/dependency with (a) time, (b) size, and (c) distance: higher values observed for weekdays than for weekends.

useful predictors, and it consistently outperforms the baseline. Compared to ART, IM is able to reduce MAPE by up to 10% (Fig. IV). Thus, *more recent real-time values of other sensors actually become more useful predictors than a sensor's own relatively older values*. This is an important result and main advantage of the influence model. Thus, when real-time recent values are not available for a sensor, it uses recent real-time values of other sensors that were identified by learning dependencies among sensors on a similar day in past.

V. Conclusion

We address the issue of *veracity* in big data applications that arises when real time data from all sensors is not available at central nodes due to network limitations, or when limited by consumers for security and privacy reasons. Standard models for short term predictions are either unable to predict or perform poorly when trying to predict with *partial data*. We introduce a novel *influence* based model to make predictions in absence of real-time data from majority of sensors using real-time data from only a few influential sensors. We show that our influence model outperforms baseline approach by up to 10% for 2 to 8 hours ahead prediction despite lack of sensors' own real-time data. This model provide a simple and interpretable solution that is generalizable to big data applications in several domains.

Acknowledgment

This material is based upon work supported by the United States Department of Energy under Award Number DEOE0000192, and the Los Angeles Department of Water and Power (LA DWP). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, the LA DWP, nor any of their employees.

References

[1] S. Aman, Y. Simmhan, and V. Prasanna. Holistic measures for evaluating prediction models in smart grids. *IEEE Transactions in Knowledge and Data Engineering (To appear)*, 2014.

[2] N. Balac, T. Sipes, N. Wolter, K. Nunes, R. S. Sinkovits, and H. Karimabadi. Large scale predictive analytics for real-time energy management. In *IEEE International Conference on Big Data*, 2013.

[3] F. Bouhafs, M. Mackay, and M. Merabti. Links to the future: communication requirements and challenges in the smart grid. *IEEE Power and Energy Magazine*, 10(1), 2012.

[4] A. Ciancio and A. Ortega. A distributed wavelet compression algorithm for wireless multihop sensor networks using lifting. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 2005.

[5] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey. Addressing big data issues in scientific data infrastructure. In *International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2013.

[6] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *International conference on Very large data bases (VLDB 2004)*, 2004.

[7] D. M. Kreindler and C. J. Lumsden. The eects of the irregular sample and missing data in time series analysis. *Nonlinear dynamics, psychology, and life sciences*, 10(2), 2006.

[8] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*. ACM, 2009.

[9] A. Marascu, P. Pompey, E. Bouillet, O. Verscheure, M. Wurst, M. Grund, and P. Cudre-Mauroux. Mistral: An architecture for low-latency analytics on massive time series. In *IEEE International Conference on Big Data*, 2013.

[10] P. McDaniel and S. McLaughlin. Security and privacy challenges in the smart grid. *IEEE Security and Privacy*, 7, 2009.

[11] E. McKenna, I. Richardson, and M. Thomson. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41(C), 2012.

[12] C. Meek, D. M. Chickering, and D. Heckerman. Autoregressive tree models for time-series analysis. In *2nd International SIAM Conference on Data Mining (SDM)*. SIAM, 2002.

[13] B. Pan, U. Demiryurek, and C. Shahabi. Utilizing real-world transportation data for accurate traffic prediction. In *IEEE International Conference on Data Mining (ICDM'12)*, 2012.

[14] M. A. Razzaque, C. Bleakley, and S. Dobson. Compression in wireless sensor networks: A survey and comparative evaluation. *ACM Transactions on Sensor Networks*, 10(1), 2013.

[15] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna. Cloud-based software platform for data-driven smart grid management. *IEEE Computing in Science and Engineering*, 2013.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 1996.