Classification of AMI Residential Load Profiles in the Presence of Missing Data

Poppy R. Harvey, Bruce Stephen, *Senior Member IEEE* and Stuart Galloway

***Abstract*—Domestic energy usage patterns can be reduced to a series of classifications for power system analysis or operational purposes, generalizing household behavior into particular load profiles without noise induced variability. However, with AMI data transmissions over wireless networks becoming more commonplace data losses can inhibit classification negating the benefits to the operation of the power system as a whole. Here, an approach allowing incomplete load profiles to be classified while maintaining less than a 10% classification error with up to 20% of the data missing is presented.**

***Index Terms*- Load modeling, Power Systems, Advanced Metering Infrastructure**

I. Residential Load Classification

WITH Advanced Metering Infrastructure (AMI) being rolled out to residential networks in many network areas, the prospect of operating the electrical grid more efficiently at this level has been keenly received. Heterogeneity characterizes residential loads, making them a particular challenge for power system operation and analysis. Despite this, daily load profiles can be partitioned into finite sets permitting generalization [1, 2, 3], to yield important subpopulations within load groups, gaining a detailed characterization of energy consumption, for example, when aligning generation with demand on microgrids [4]. However, AMI systems, being primarily based on wireless communications technologies, face the risk of data loss, so this paper proposes modifications that enables an existing technique for classifying daily load profiles in the presence of inevitable missing data and demonstrates the applications for this technique for robust classification and variable horizon short term load forecasting.

II. Load Classification Using Mixture Models

As proposed in [2], a daily electrical load profile l on a given date t can be represented as a multivariate Gaussian distribution with mean μ and covariance Σ as

$$P(l_t;\mu,\Sigma)=\frac{1}{l_t}\frac{1}{2\pi|\Sigma|^{\frac{d}{2}}}\exp\frac{1}{2}(\ln l_t-\mu)^T\Sigma^{-1}(\ln l_t-\mu) \quad (1)$$

The time resolution is encoded as d variates within the distribution, which results in a d-dimensional mean vector μ representing expected load. To accommodate multiple behaviors which will embody themselves as modes in the empirical distribution of load at a given time of day, this representation can be embedded in a mixture model comprising a linear combination of M distributions as

$$P(l_t)=\sum_{m=1}^{M}P(m)P(l_t;\mu_m,\Sigma_m) \quad (2)$$

Akin to general clustering approaches [1, 3], for any given load profile, this model, once learned from data [2], will yield a class label c evaluated as the one distribution out of M most likely to have generated it

$$c_t=\max_c\frac{P(c)P(l_t;\mu_c,\Sigma_c)}{\sum_{m=1}^{M}P(m)P(l_t;\mu_m,\Sigma_m)} \quad (3)$$

While techniques such as Self Organising Maps and Fuzzy type classifiers [1, 6] can achieve a similar objective, they typically need a complete set of inputs, e.g. l_t, on which to base their classifications. In AMI systems points in l_t are captured at a number of regular time periods though the day; measured values will be real but *null* readings can result from hardware or communication failure. The set s_t contains the indices of the daily load profile l_t for which the load data is valid. Formally this can be expressed as,

$$s_t=\{a:1\le a\le A\,|\,l_{ta}\in\Re\}, \quad (4)$$

where A is the load profile advance resolution (A=48 half hourly values in the cases considered here) and l_{ta} is the a-th advance of the load profile. The dimensionality of a multivariate Gaussian distribution can be reduced by simply extracting a subset of its mean vector as

$$\mu'\subseteq\mu,\text{where }\mu=\{\mu_1,\mu_2,\ldots,\mu_A\} \quad (5)$$

applying this to the daily load profile data allows s_t to be used synonymously with μ', *i.e.*

$$\forall\mu'_i\,\exists i\in s_t \quad (6)$$

This labeling will therefore permit a classification to be made irrespective of the amount of data missing during the day.

III. Load Classification with Incomplete Data

Models for load classification were learned from held out AMI data [7] using the procedure described in [2] and benchmark data was created retrospectively to form incomplete data sets. In the first instance, load data were randomly censored similar to the effect that transient power outages or communication failures might have [6]. Sustained

Dr. B. Stephen is a Senior Research Fellow in the Advanced Electrical Systems Research Group, Institute of Energy and Environment, University of Strathclyde, Glasgow, G1 1RD (phone: +44 (0)141 444 7260, e-mail: bruce.stephen@strath.ac.uk)

gaps in AMI readings would likely be caused by longer power outages or device failure hence are not considered here. The effect of censoring groups of measurements to simulate latency of the GPRS network, or failure of short range wireless networks, both used in AMI systems [5], at particular times of the day is another potentially interesting case. Figure 1 compares the effect of null readings ('not a number' - NaNs) when they are randomly located throughout the data, demonstrating daily energy usage is most likely to be incorrectly classified if data is missing in the early hours of the morning. This may be attributable to the load profile classes being characterized by less variable off-peak advances [2, 3].

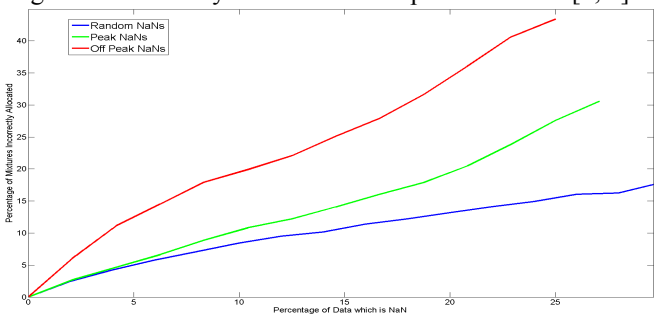

Figure 1: Comparison of the number of incorrectly allocated labels as the percentage of nulls in the data increases.

Null readings that are randomly dispersed throughout the data have the least effect on the classification process.

TABLE I

CORRECT CLASSIFICATION OF ENERGY USAGE FOR NULL READINGS OCCURRING AT PEAK TIMES

Null Reading Occurrence	% of Load Profile Data Missing	№ Erroneous Allocations	% of Allocations which are Erroneous
7-7.30pm	4.16	117	4.66
7-8.30pm	8.33	231	9.20
7-9.30pm	12.50	399	15.89
6-9.30pm	16.66	494	19.67
5-9.30pm	20.83	618	24.61
4-9.30pm	25.00	666	26.52
4-10.30pm	29.16	799	31.82

TABLE II

CORRECT CLASSIFICATION OF ENERGY USAGE WHEN NULL READINGS OCCUR AT OFF-PEAK TIMES

Null Reading Occurrence	% of Missing Data Points	№ Erroneous Allocations	% Erroneous Allocations
3-3.30am	4.16	108	4.30
3-4.30am	8.33	213	8.48
2-4.30am	12.50	344	13.70
2-5.30am	16.66	536	21.35
1-5.30am	20.83	729	29.03
12-5.30am	25.00	937	37.32
12-6.30am	29.16	1161	46.24

Tables I and II indicate that it is off-peak times that influence the distinguishing features of a load profile – lower classification errors result from more than 20% of the load profile missing at peak times than for off-peak.

IV. Partial Day Forecasting

Another motive to accommodate partial observations is the need to classify the current day at different time horizons, for example to formulate a robust demand response schedule for the following day [4].

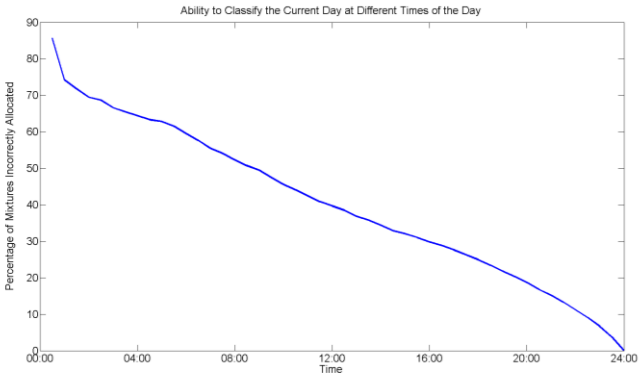

Figure 2: Number of incorrectly allocated mixtures against time of day.

At midday in figure 2, with 50% of readings unknown, 60% of class allocations were correct. In order to confidently predict the days' allocation, around 90% of the time this cannot be made until 10pm but this still buys a valuable window in which to carry out processing for day ahead operational activities.

V. Conclusions

Future electricity grids place an increasing dependence on data and communications, particularly in understanding load on distribution networks. The approach presented allows data with small gaps to still be reliably used to inform models that classify load behaviors, accommodating null or delayed readings inevitable with wireless data collection. Energy usage in a given premises on a given day continues to be assigned to the correct sub-profile class with almost 90% accuracy even with several hours of data missing. Where this addition will be most useful in practice is in residential demand response or storage schemes that deal with highly dynamic load behaviour across a small number of customers on relatively short time scales with the added challenge of reliance on public cellular or wireless networks.

VI. References

[1] Seem, J. E. "Pattern recognition algorithm for determining days of the week with similar energy consumption profiles," Energy Build., vol. 37, no. 2, pp. 127–139, 2005.

[2] Stephen, B., Mutanen, A., Galloway, S., Burt, G. & Jarventausta, P. "Advanced Load Profiling for Residential Customers", IEEE Trans. Power Delivery, vol. 29, no. 1, pp. 88-96, February 2014.

[3] Kwac, J., Flora, J. & Rajagopal, R. "Household Energy Consumption Segmentation Using Hourly Data", IEEE Trans. Smart Grid, vol. 5, pp. 420-430, 2014.

[4] Palensky, P., & Dietrich, D. "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," IEEE Trans. Industrial Informatics, vol.7, no.3, pp.381-388, Aug. 2011

[5] Khalifa, T., Naik, K. & Nayak, A., "A Survey of Communication Protocols for Automatic Meter Reading Applications," IEEE Comm. Surveys & Tutorials, vol.13, no.2, pp.168-182, 2nd Quarter 2011

[6] Zhang, Y. & Årvidsson, A. "Understanding the characteristics of cellular data traffic." ACM SIGCOMM Computer Comm. Review 42.4 (2012): 461-466.

[7] CER. (2014, Oct. 17). Smart Metering Trial Data Publication. Data from the Commission for Energy Regulation. (2013, Sep. 20). [Online]. Available:http://www.ucd.ie/issda/data/commissionforenergyregulationcer/