Computing Electricity Consumption Profiles from Household Smart Meter Data

Omid Ardakanian
University of Waterloo
Waterloo, Ontario, Canada
oardakan@uwaterloo.ca

Negar Koochakzadeh[*]
Oracle
Vancouver, BC, Canada
negar@koochakzadeh.net

Rayman Preet Singh
University of Waterloo
Waterloo, Ontario, Canada
rmmathar@uwaterloo.ca

Lukasz Golab
University of Waterloo
Waterloo, Ontario, Canada
lgolab@uwaterloo.ca

S. Keshav
University of Waterloo
Waterloo, Ontario, Canada
keshav@uwaterloo.ca

ABSTRACT

In this paper, we investigate a critical problem in smart meter data mining: computing electricity consumption profiles. We present a simple, interpretable and practical profiling framework for residential consumers, which accounts for variations in electricity consumption at different times of day and at different external temperatures. Our approach is to isolate the effect of external temperature on electricity consumption and apply a time-series autoregressive model to the remaining signal. The proposed profiles may be used for making personalized energy-saving recommendations, detecting outliers, and generating very large realistic data sets for testing the scalability of smart meter data management systems. Using predictive power as a metric for the accuracy of consumption profiles, we show, using a real data set of 1000 homes, that our approach results in improved root-mean-squared prediction error compared to existing approaches.

1. INTRODUCTION

Smart electricity meters are rapidly replacing conventional meters in many parts of the world. Smart metering systems offer many operational advantages for energy utilities and policy makers, including

- enabling automated collection of fine-grained (typically half-hourly or hourly) consumption readings, thereby eliminating the need for utilities to send out estimated bills or to dispatch personnel to customer premises and manually read the meters,
- enabling dynamic pricing schemes that depend on the time-of-day in order to reduce demand for electricity during peak times.

However, exploiting smart metering systems to their fullest also requires mining the vast amounts of collected consumption data to obtain insights into grid operations and consumer behaviour [2, 4, 7, 12, 16, 21].

In this paper, we address the problem of computing electricity consumption profiles from smart meter data, with a focus on residential customers. The residential sector contributes a significant fraction to the total electricity demand (30 percent in Canada [5]) and greenhouse gas emissions (see, e.g., [19, 24] for United States statistics). Furthermore, in many regions, residential consumers are significant contributors to peak demand; e.g., in Ontario, Canada, residential air conditioning load is a major contributor to peak demand, which occurs in the afternoon of hot summer weekdays [23].

We argue that consumption profile generation is a fundamental smart meter data mining operation that electricity providers, resellers and consultants can perform, with at least the following applications:

- Conducting "virtual energy audits" and making personalized recommendations for saving electricity based on the trends identified in the profiles.
- Clustering households based on the features captured by the profiles. This may be used to understand different classes of consumers and to design targeted energy conservation and peak reduction programs for different classes.
- Generating real-time alerts if new consumption readings do not match the expected consumption predicted by the profiles. A related application is to identify consumers with "suspicious" load profiles that do not fit in any cluster, which could indicate electricity theft, malfunctioning meters or the presence of specialized equipment such as electric vehicle chargers.
- Generating realistic synthetic data based on the available real data. This may be used as input to grid simulation, transformer sizing, forecasting and pricing models, or to create very large realistic data sets for testing the scalability of smart meter data management systems.

[*] Work done while the author was a Postdoctoral Fellow at the University of Waterloo.

1.1 Challenges and Contributions

In order to be useful for the above applications, we argue that consumption profiles must satisfy the following criteria.

- First, they must be *accurate*, i.e., able to describe and predict consumption with reasonable accuracy.

- Second, they must be easily *interpretable*; a complex machine learning model may be accurate, but if it is not interpretable, then actionable energy-saving recommendations cannot be easily inferred from it.

- Third, they must be *practical*, and therefore they should only require data that are easily available to utilities, such as hourly smart meter readings and weather. While household characteristics (e.g., home size and age, number of appliances, number of occupants, etc.) and consumer demographics could be useful, this information is typically not available to utilities due to privacy regulations and cannot be easily obtained without intrusive measurements and questioning.

The challenge in computing accurate and interpretable profiles from smart meter data is that residential electricity consumption depends on many factors, including the time of day, weather and the occupants' daily routines. Broadly speaking, prior work can be divided into two approaches. One is to compute various aggregate statistics from historical consumption data that account for typical daily activity; examples include the average, maximum, minimum and variance of hourly or daily consumption, ratios of night-to-day or morning-to-evening consumption, or identifying the hour of day when peak consumption usually occurs. The other approach has been to correlate consumption with external temperature, e.g., using piecewise linear regression, and use the correlation coefficients as representatives for the cooling and heating efficiency of homes.

In this paper, we propose a simple and practical technique that combines the best features of existing methods, and accounts for both temperature and activity in an accurate and interpretable fashion. The idea is to remove the effects of external temperature and outliers from the raw consumption data[1], and compute typical hourly consumption values from the remaining signal using a time series auto-correlation model. This gives us typical consumption levels of a given home at different times of the day, independent of temperature and robust to "noise" (e.g., time periods when the home was empty or unusually busy).

Specifically, we make the following contributions in this paper:

- We propose a simple and interpretable technique for computing electricity consumption profiles from household smart meter data, which may be used for personalized recommendations, forecasting, classification, and as input to simulation models.

- Using predictive power as a metric for accuracy and hence the representativeness of consumption profiles, we compare the proposed method to several existing approaches. Using a real data set, consisting of a year of hourly smart meter readings from 1000 homes in southern Ontario, Canada, we show that our approach outperforms existing approaches in terms of the root-mean-squared prediction error.

1.2 Roadmap

The remainder of this paper is structured as follows. Section 2 gives the intuition and an overview of our solution; Section 3 presents the details of our consumption profile algorithm; Section 4 describes our experimental results; Section 5 discusses related work; and Section 6 concludes the paper and discusses open problems in smart meter data management.

2. INTUITION AND SOLUTION OVERVIEW

In this section, we give an overview of our solution and we explain the intuition behind it. The input to our problem consists of two time series: 1) periodic (e.g., hourly) timestamped electricity consumption readings from a given home for some period of time (e.g, 6 months or a year), and 2) a corresponding time series with external temperature measurements, with the same granularity and for the same period of time, e.g., from a nearby weather station.

A very simple consumption profile could consist of 24 numbers: the average consumption for each hour of the day, aggregated over some or all of the input data. A simple extension is to compute two such profiles: one for weekdays and one for weekends and holidays. (We could go further and compute separate profiles for every day of the week, but, to keep the model simple and easily interpretable, we will only consider weekday-weekend splits in this paper.)

Hourly averages may reveal some high-level details about the consumption habits of a household, but we can do better. Observe that in climates with summer air conditioning usage and/or winter electric heating, a large part of the electricity consumption is temperature-sensitive; e.g., in the United States, roughly 40 percent of a home's energy consumption servers heating and cooling needs [24]. For example, Figure 1 plots the hourly consumption and external temperature (measured in degrees Celcius) for a sample home in southern Ontario, Canada, between April 2011 and October 2012 (we omit further details about the data source to preserve privacy). Observe that the peak summer consumption of this home is roughly 1.5 kilowatt-hours (kWh) higher than the peak winter consumption, which is likely due to air conditioning usage. If we could quantify the consumption of temperature-sensitive loads and remove it from the original consumption time series, the remaining consumption would give us a better idea of the occupants' routines and activities, and thus a more accurate profile.

The problem is that the relationship between consumption and external temperature is not exact, making it difficult to estimate the consumption of temperature-sensitive appliances from whole-house smart meter data. Figure 2 plots the noon-time energy consumption of the same sample home as a function of temperature; i.e., each point represents the noon-time consumption of this home on some day between April 2011 and October 2012 as well as the temperature at that time. On some days, the noon-time consumption is rel-

[1] Here, by outliers we mean consumption readings that are much lower or higher than the average consumption for the given home, and thus do not correspond to the typical level of activity in this home.

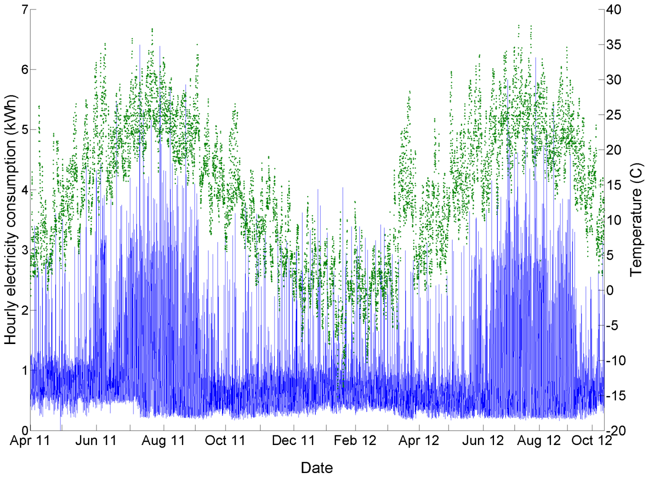

Figure 1: Hourly consumption of a sample home (blue curve with Y-axis on the left) and the temperature (green curve with Y-axis on the right), measured between April 2011 and October 2012.

atively high when the temperature is moderate or relatively low, and vice-versa. Some of these "outliers" may correspond to days when the home was empty or unusually busy. If we could remove these outliers, we could obtain a more accurate estimate of the temperature-sensitive load, and also a more accurate estimate of the remaining (routine and activity) load, which can give a more accurate and robust profile.

Our proposed solution, illustrated in Figure 3, implements the above observations. Given the smart meter and temperature time series, we will compute 48 numbers: the typical daily consumption values for each hour of the day on weekdays and weekends, after accounting for temperature-dependent load and outliers. The details of our solution are presented in the next section.

3. COMPUTING CONSUMPTION PROFILES

We now describe the proposed solution, beginning with an overview of the time series model that we employ (Section 3.1), followed by a discussion of how outliers and temperature effects are taken into account (Section 3.2) and how model parameters are chosen (Section 3.3). We then show how to extract consumption profiles from our time series model (Section 3.4) and we discuss several applications of the proposed profiles (Section 3.5).

3.1 The PARX Model

The main idea behind our solution is to apply a time series autoregression model, specifically Periodic Auto Regression with eXogenous variables (PARX) [17]. Table 1 lists the symbols used in the remainder of the paper; in our case, we have 24 "seasons", each corresponding to a particular hour of the day, as we will be building separate consumption models for each hour.

In general, a PARX model of order p represents a time series in terms of 1) its recent history (the most recent p data points), 2) exogenous variables, and 3) a white noise

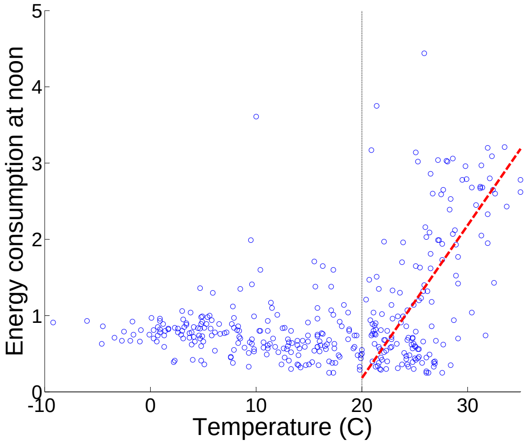

Figure 2: Hourly consumption measured at noon in a sample home versus the external temperature. The dashed line represents best linear fit for temperatures higher than 20° Celsius.

component. In our case, this can be written as

$$Y_t = \sum_{i=1}^{p} \phi_{i_s} Y_{t-i} + \sum_{j=1}^{n} \psi_{j_s} X_t^j + C_s + \epsilon_t, \qquad t \in s \qquad (1)$$

where Y_t is the electricity consumption at a particular hour at time t, n is the number of exogenous variables (*i.e.,* the X^j's), ϵ_t is the value of the white noise component[2] at time t, C_s is an intercept term, and s is the "season" index. The model parameters $C_s, \phi_{i_s}, \psi_{j_s}$, and σ_s^2 depend on the season.

Intuitively, Equation (1) states the following. Pick a "season", i.e., some hour of the day, say, noon. The electricity consumption at noon is a linear function of the consumption at noon on the previous p days[3], and of the n exogenous variables (such as temperature), plus a constant intercept term and an error term. Since each hour of the day is a separate season with its own model, the values of the coefficients ϕ_i and ψ_j may be different for different hours. That is, this method is flexible enough to capture the possibility that at some hours of the day (e.g,. night-time), temperature has a stronger relationship with consumption, whereas at other hours of the day (e.g., dinner-time), the load is more affected by the occupants' activities.

3.2 Exogenous Variables

As mentioned in Section 2, we want to compute the typical hourly consumption of a household, after accounting for temperature and "outliers" corresponding to periods of very low or very high consumption. This is exactly the purpose of exogenous variables. The effects of other unknown factors on the overall consumption (i.e., other appliances, daily routines and patterns) will be captured in the resulting consumption profile via the auto-regressive part of the model.

[2] We assume that the white noise process is a sequence of independent and identically distributed random variables with zero mean and finite variance σ_s^2.

[3] More precisely, it is a function of the previous p weekdays for the weekday profile and the previous p weekends/holidays for the weekend/holiday profile.

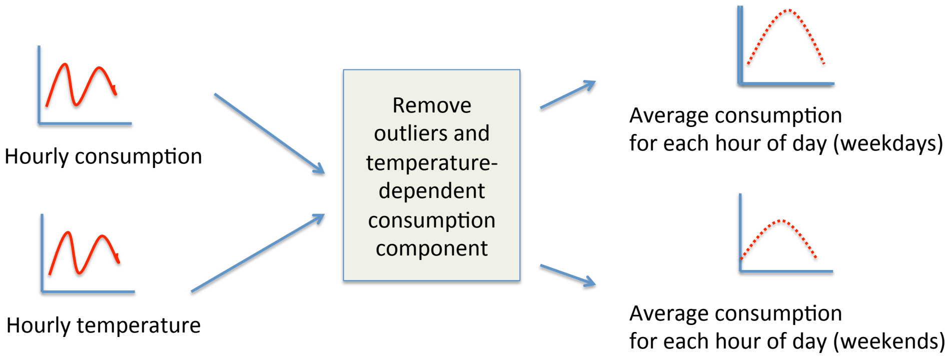
Hourly consumption
Hourly temperature
Remove outliers and temperature-dependent consumption component
Average consumption for each hour of day (weekdays)
Average consumption for each hour of day (weekends)

Figure 3: Overview of the proposed consumption profile approach.

Y_t	The original time series of household electricity consumption at time t
Y_t^*	The time series at time t obtained after removing the effects of exogenous variables, representing temperature and outliers
$XT1$, $XT2$, $XT3$	Temperature related exogenous variables
$XO1$, $XO2$	Occupancy related exogenous variables
σ_s	The standard deviation of season s
ϵ_t	The value of the white noise component at time t
ϕ_{i_s}	The coefficient of Y_{t-i} in season s
ψ_{j_s}	The coefficient of the j^{th} exogenous variable in season s
C_s	The intercept term of season s

Table 1: List of symbols used in this paper

First, we deal with temperature. Recall Figure 2 and notice that the effect of temperature in the summer may be very different to the effect of temperature in the winter. In southern Ontario, the regression line has a positive slope at high temperatures, corresponding to the increasing intensity of air conditioning usage as temperatures climb. On the other hand, the winter effects of temperature are less pronounced, because the majority of homes, including our sample home, mainly use natural gas for heating.

The above observation implies that we cannot use a single exogenous variable to account for temperature. Instead, following previous work on modelling the effect of temperature on electricity consumption (e.g., [5, 13]), we use three variables: XT1, XT2, and XT3. They are defined in Equations (2),(3) and (4). The coefficients of these variables represent the cooling (temperature above 20 degrees), heating (temperature below 16 degrees), and overheating (temperature below 5 degrees) slopes, respectively.

$$XT1 = \begin{cases} T - 20 & \text{if } T > 20 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$XT2 = \begin{cases} 16 - T & \text{if } T < 16 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$XT3 = \begin{cases} 5 - T & \text{if } T < 5 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Now, we show how to handle "outliers" corresponding to unusually low or high consumption. The first step is as follows. For each season (i.e., hour of the day), logically we produce a plot similar to that in Figure 2, which illustrates the relationship between temperature and consumption at that particular hour of the day across different days, using all the historical data given as input. For each value of temperature, we then compute the 10th and 90th percentiles of the consumption values. Using these values, we then define two new exogenous variables, $XO1$ and $XO2$, as follows.

- At any time t, $XO1$ is equal to one if the consumption at t is higher than the 90th percentile of the consumption at that hour of the day and that specific temperature, as described above. It is zero otherwise.
- Similarly, $XO2$ is equal to one if the consumption at t is less than the 10th percentile of the consumption at that hour of day and that specific temperature. It is zero otherwise.

We chose the 10th and 90th percentile values heuristically to define $XO1$ and $XO2$, corresponding to very low consumption (when the home may have been empty for a long while) and very high consumption (when the household is unusually busy). We refer to $XO1$ and $XO2$ as occupancy-related variables.

Using all five exogenous variables, our PARX model becomes

$$\begin{aligned} Y_t = &\sum_{i=1}^{p} \phi_{i_s} Y_{t-i} + \psi_{1_s} XT1_t + \psi_{2_s} XT2_t + \psi_{3_s} XT3_t \\ &+ \psi_{4_s} XO1_t + \psi_{5_s} XO2_t + C_s + \epsilon_t, \qquad \text{for } t \in s \end{aligned} \tag{5}$$

3.3 Parameter Estimation

The first parameter that we need to set is p, the number of previous days to include in the auto-regressive part of the model. For each of our 24 seasons (hours of day), we tried different values of p between 1 and 48, and computed the Bayesian Information Criterion (BIC) [25]. Based on our data set of 1000 homes, $p = 3$ gave the best results (i.e., the lowest BIC).

Once we have determined an optimal value of p, we use the standard Ordinary Least Squares (OLS) method to derive the coefficients of the PARX model for each hour of the day (repeating the process for weekdays only, and for weekends/holidays).

3.4 Putting it All Together

Having presented the details of our PARX framework, we are now ready to describe how the consumption profiles are derived. First, we compute our PARX model for each hour of day, separately for weekdays and weekends/holidays. We then generate a new consumption time series by taking the original values and removing the temperature-sensitive consumption as well as the outliers. For each hour of day (and separately for weekdays and weekends/holidays), we do this simply by "reversing" the model and subtracting the effects of exogenous variables. Let Y_t^* be the new consumption time series after removing temperature- and occupancy-sensitive components:

$$\begin{aligned} Y_t^* = Y_t - \psi_{1_s} XT1_t - \psi_{2_s} XT2_t - \psi_{3_s} XT3_t \\ - \psi_{4_s} XO1_t - \psi_{5_s} XO2_t \qquad \text{for } t \in s \end{aligned} \tag{6}$$

That is, what remains in Y_t^* is just the auto-regressive part of the model.

Finally, we take the hourly averages of the corresponding Y_t^*'s, separately for weekdays and weekends/holidays, which completes the discussion of the process shown in Figure 3. Thus, our profiles consist of two vectors of 24 values, where the ith value is the typical consumption level of the given home at the ith hour of the day, after removing the effects of exogenous variables.

Figures 4 and 5 illustrate the weekday and weekend profiles of our sample home. Note that the profiles are easy to interpret and contain a great deal of useful information . For example, it is easy to see that 1) the typical hourly consumption is higher on weekdays than weekends, 2) peak weekday load occurs at 19:00 with a small peak at 9:00, while peak weekend load occurs at 17:00, and 3) the occupants of this home appear to consume more electricity between 8:00 and 11:00 on weekends than weekdays.

3.5 Applications

We conclude this section with a brief description of how the proposed consumption profiles can be used for two of the motivating applications listed in Section 1.

Personalized recommendations for saving electricity

Normally, we expect the hourly consumption of a typical household to decrease at night, when there is little to no activity in the home. If the consumption profile suggests that the nightly consumption of a given household remains high, then we can recommend a new refrigerator or another appliance that is always on. Note that this recommendation makes sense because we have removed temperature-sensitive load before computing the profiles. Otherwise, we would not know if the high nightly load is caused by heating and cooling or by another appliance. Similarly, if the profile shows high consumption during expensive on-peak hours regardless of temperature, then we can recommend shifting some activities (such as laundry) to off-peak hours.

Furthermore, we can generate *comparative* feedback by clustering similar consumers based on the hourly loads contained in their profiles and/or the coefficients of their exogenous variables. For instance, if a household belongs to a cluster in which other households have similar hourly loads but lower coefficients of the temperature-related exogenous variables, then we can hypothesize that this household has an inefficient air conditioning system.

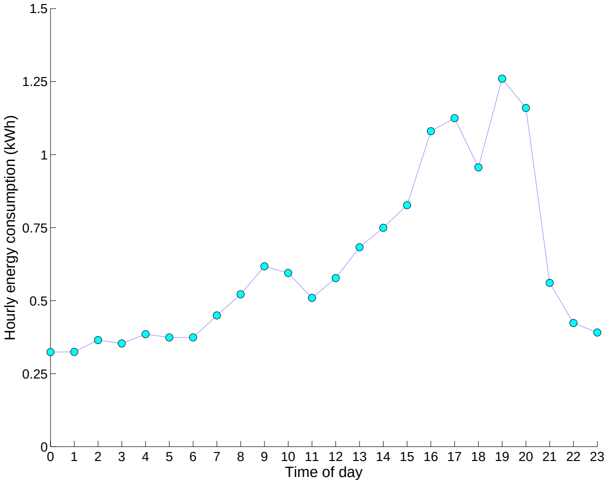

Figure 4: Weekday consumption profile of a sample home.

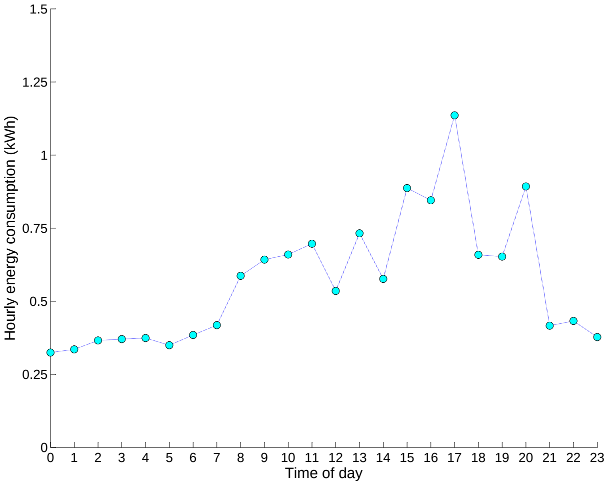

Figure 5: Weekend/holiday consumption profile of a sample home.

Generating realistic synthetic data as input to forecasting models or to test the scalability of smart meter data management systems

Here, the objective is to create realistic synthetic "households" based on the profiles computed from a real data set. One way to do this is as follows. First, we use a clustering algorithm such as k-means to group together similar profiles. To generate a new consumption time series, we randomly choose a cluster and use its centroid as the consumption profile of the new household. We can then choose the exogenous variable coefficients from a random member of this cluster, generate a weather forecast time series, and use this information to create the new consumption time series.

4. EXPERIMENTS

In this section, we evaluate the predictive power of our consumption profiles and compare it to the predictive power of representative profiling techniques from prior work—one that focuses on hourly consumption aggregates, one that uses temperature alone, and one that uses both temperature and hourly averages. We implemented the algorithms in Matlab.

Our dataset is comprised of aggregate hourly electricity consumption levels of 1000 homes from a city in southern Ontario, Canada. Measurements were taken between March 2011 and October 2012. We also obtained the ambient air temperature data of that region from the Environment Canada Website.

4.1 Methodology

We use two thirds of the consumption dataset of each home as the training data set for building the model, and the rest, including 170 days from April 2012 to October 2012, as the testing data set for evaluating its predictive power. For instance, to evaluate the predictive power of a model on April 1, 2012, we use consumption measurements from March 2011 to March 2012 as the training set. We extend the training set by adding days from the test set which are prior to the day for which we evaluate the predicative power. For example, April 1, 2012 is added to the training set when we evaluate predictive power for April 2, 2012.

We predict the consumption of each home using the following four profiling approaches, assuming that the hourly temperature forecast is available one day in advance.

The first is our approach, labeled *PARX*. We predict the hourly consumption on the test day, for a given hour h, as follows. First, we look up the average hourly consumption for that hour from the profile, call it P_h. We then add in the contribution of exogenous variables for that hour using the coefficients of that hour's model. This gives us an estimate of the consumption for that hour, call it $\hat{Y}_h$:

$$\begin{aligned}\hat{Y}_h = Y^*_h + {\psi_1}_h XT1_h + {\psi_2}_h XT2_h + {\psi_3}_h XT3_h \\ + {\psi_4}_h \widehat{XO1_h} + {\psi_5}_h \widehat{XO2_h}\end{aligned} \tag{7}$$

Note that in order to use our consumption profiles for predicting future consumption, we must use *estimated* values of the occupancy-related exogenous variables, denoted $\widehat{XO1}$ and $\widehat{XO2}$, since we obviously do not know their true values when making the prediction. Here, we simply use the observed value of these variables in the previous hour ($\widehat{XO1_h} = XO1_{h-1}$ and $\widehat{XO2_h} = XO2_{h-1}$), although more sophisticated methods could be used to estimate these variables and further improve the predictive power of our approach.

The second approach represents methods that compute hourly aggregates from the consumption time series. In particular, we compute hourly averages over the training set and use these for prediction. We call this approach *Hourly Mean*.

The third approach represents methods that focus on the correlation between consumption and temperature [5]. This algorithm fits a three-piece linear regression model after removing very-low and very-high consumption values and uses the temperature of the test day to predict consumption. We refer to this technique as *3-Line*.

Finally, the fourth approach, proposed in [13], uses a time series model similar to PARX and also takes temperature into account (but does not account for outliers, which we do). We call this algorithm *Convergent Vector* since it computes typical hourly consumption by finding the convergent vector of the input time series.

The real value of the hourly electricity consumption on the test day is then compared with the predicted values to compute the root-mean-square error (RMSE) of each approach for each day.

4.2 Results

We compared the predictive power of the above four approaches using all 1000 homes in our dataset. Our findings are as follows.

- PARX outperformed Hourly Mean for 982 homes, 3-Line for 960 homes, and Convergent Vector model for 901 homes.
- The average RMSE was 0.70 for PARX, 0.81 for Hourly Mean, 0.94 for 3-Line, and 0.77 for Convergent Vector. This means that our model's RMSE was 14 percent lower than that of Hourly Mean, 26 percent lower than that of 3-Line, and 9 percent lower than that of Convergent Vector.

Thus, although 3-Line and Convergent Vector obtained a slightly lower prediction error for a few homes, on average their prediction error is considerably higher than that of PARX. This confirms that, in most cases, incorporating historical hourly consumption, temperature dependence, and occupancy dependence results in a more representative model. Nevertheless, in some cases, temperature is highly correlated with electricity consumption, and therefore incorporating occupancy does not improve prediction accuracy. This is most likely because the bulk of the electricity consumption of these homes serves heating and cooling needs, and thus temperature alone is a very good predictor.

Figure 6 shows the average RMSE on the testing days for 20 randomly selected homes along with the RMSE values averaged over all 1000 homes on the testing days. The RMSE of PARX is lower than the RMSE of the other three approaches for all but two of these homes.

5. RELATED WORK

A considerable body of previous work has developed various consumption modelling techniques from household

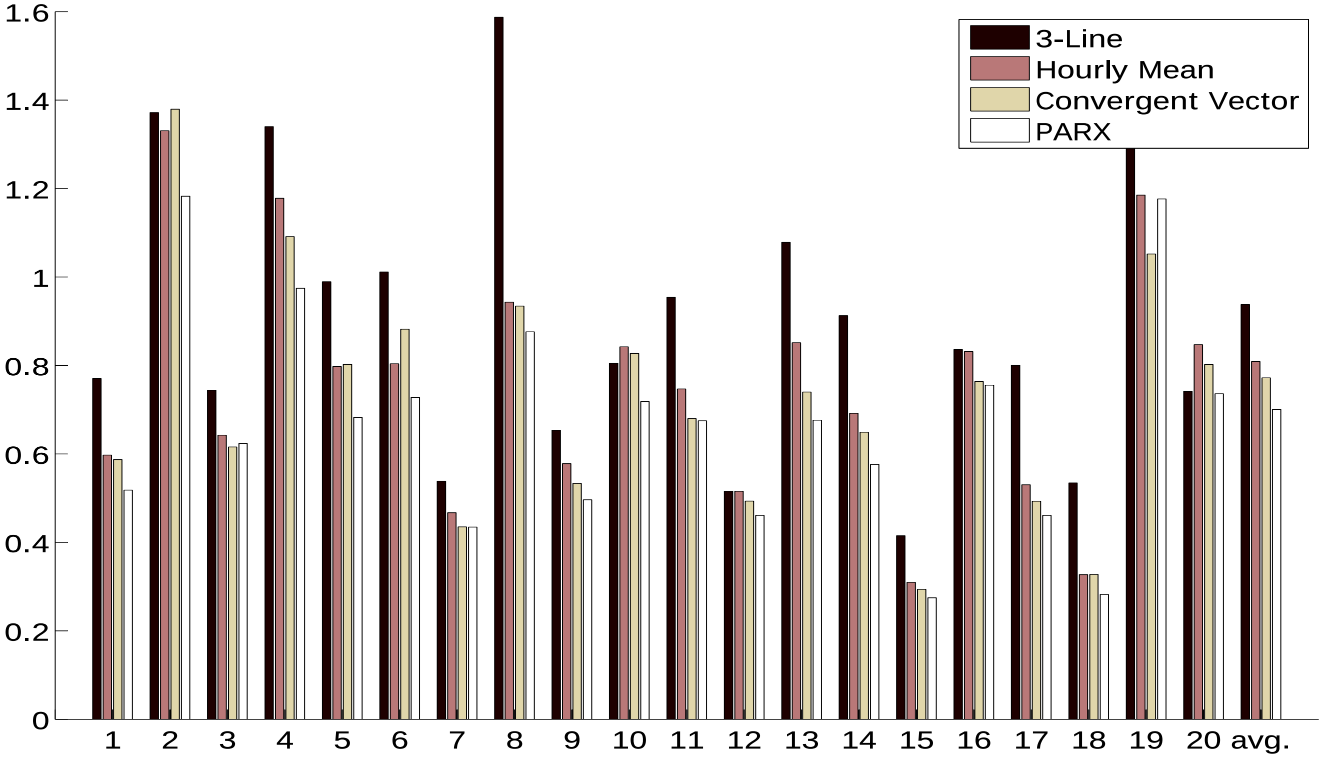

Figure 6: Average RMSE of the four approaches on 20 randomly selected homes measured on 170 testing days.

smart meter data, with applications ranging from clustering similar consumers, planning and forecasting, tariff design, electricity loss and theft detection, to providing personalized feedback on how to save electricity. Our technique may be used in many of these applications in regions where some fraction of the electricity consumption is correlated with temperature.

From a technical standpoint, previous work has approached the consumption profiling problem from two directions. One was to examine historical consumption values and compute various representative aggregates; see, e.g., [3, 8–10, 14, 20, 22]. The other direction has been to focus on the relationship between electricity consumption and temperature; see, e.g., [1, 5]. There are also techniques that combine aggregated values with temperature correlations, e.g., [13]. In this paper, our goal was to design a simple and interpretable but also accurate profiling algorithm by combining the best features of existing methods.

In particular, the two approaches most closely related to ours are by Espinoza et al. [13] and Birt et al. [5]. Our approach combines the time series modelling approach of [13] with the temperature model of [5], and additionally takes outliers into account. As we experimentally showed in Section 4, by combining and enhancing the best features of prior models, our techniques resulted in a lower prediction error.

In general, energy data management is an emerging field of study, with recent work on smart grid data management and analytics [6, 15], using Hadoop to manage smart meter data [11], imputing missing data in smart meter time series [18], and symboling representation of smart meter time series [26].

6. CONCLUSIONS AND OPEN PROBLEMS

In this paper, we described a simple and interpretable technique for computing electricity consumption profiles from residential smart meter data combined with temperature data. Our solution relies on auto-regressive time series modelling with exogenous variables to take into account various factors influencing electricity consumption, such as temperature and the occupants' daily habits. Experimental evaluation using a real data set of smart meter readings from 1000 homes revealed the advantages of our method over previous work in terms of prediction accuracy.

One limitation of the proposed approach is that it is effective only for regions where some fraction of household electricity consumption is correlated with temperature, such as those with heavy air conditioning use during the summer. If this is not the case, simpler profiling techniques may be used, such as computing the average electricity consumption in each hour of the day.

We are currently building a prototype smart meter data management system, in which the proposed consumption profiling method will play a central role. We highlight several open problems in this area that we intend to study:

- Smart meter data quality: missing values are common, and unusually low or high values may indicate failing meters that need to be replaced.

- Efficient and scalable smart meter analytics: there is very little work that focuses on exploiting modern data analytics platforms, such as Hadoop, data stream engines or time series databases, for smart meter data.

- In addition to smart electricity meters, smart water meters are being introduced in many jurisdictions, including Toronto, Canada, as described at torontowatermeterprogram.ca. This will enable large-scale water data analytics and require smart meter data management system to handle water data in addition to electricity data.

7. REFERENCES

[1] A. Albert and R. Rajagopal. Building dynamic thermal profiles of energy consumption for individuals and neighborhoods. In *IEEE Big Data Conf.*, 2013

[2] AutoGrid. www.auto-grid.com

[3] C. Beckel, L. Sadamori, and S. Santini. Towards automatic classification of private households using electricity consumption data. In *4th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, pages 169-176, 2012.

[4] Big Data Energy Services. www.bigdataenergyservices.com

[5] B. J. Birt, G. R. Newsham, I. Beausoleil-Morrison, M. M. Armstrong, N. Saldanha, and I. H. Rowlands. Disaggregating categories of electrical energy end-use from whole-house hourly data. *Energy and Buildings*, 50(0):93–102, 2012.

[6] M. Boehm, L. Dannecker, A. Doms, E. Dovgan, B. Filipic, U. Fischer, W. Lehner, T. B. Pedersen, Y. Pitarch, L. Siksnys and T. Tusar. Data Management in the MIRABEL Smart Grid System. In *1st Workshop on Energy Data Management (EnDM)*, 2012.

[7] C3 Energy. www.c3energy.com

[8] G. Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012.

[9] G. Chicco and I.-S. Ilie. Support vector clustering of electrical load pattern data. *IEEE Trans. on Power Systems*, 24(3):1619–1628, 2009.

[10] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *IEEE Trans. on Power Systems*, 18(1):381-387, 2003.

[11] L. dos Santos, A. da Silva, B. Jacquin, M.-L. Picard, D. Worms, C. Bernard. Massive Smart Meter Data Storage and Processing on top of Hadoop, In *VLDB BigData Workshop*, 2012.

[12] eSmart Systems. www.esmartsystems.com

[13] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor. Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. *IEEE Trans. on Power Systems*, 20(3):1622–1630, 2005.

[14] V. Figueiredo, F. Rodrigues, Z. Vale, and J. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Trans. on Power Systems*, 20(2):596–602, 2005.

[15] U. Fischer, D. Kaulakiene, M. E. Khalefa, W. Lehner, T. B. Pedersen, L. Siksnys and C. Thomsen. Real-time Business Intelligence in the MIRABEL Smart Grid System. In *VLDB Workshop on Business Intelligence for the Real-Time Enterprise (BIRTE)*, 2012.

[16] Green Button. www.greenbuttondata.org

[17] H. Hurd and A. Miamee. *Periodically correlated random sequences: spectral theory and practice*, volume 355. Wiley-Interscience, 2007.

[18] R.-S. Jeng, C.-Y. Kuo, Y.-H. Ho, M.-F. Lee, L.-W. Tseng, C.-L. Fu, P.-F. Liang, L.-J. Chen. Missing Data Handling for Meter Data Management System. In *e-Energy Conf.*, pages 275-276, 2013.

[19] J. Miller. North american power plant air emissions. Technical Report 2-923358-11-2, Commission for Environmental Cooperation of North America, 2004.

[20] A. Nizar and Z. Dong. Identification and detection of electricity customer behaviour irregularities. In *IEEE/PES Power Systems Conference and Exposition*, pages 1–10, 2009.

[21] OPower. www.opower.com

[22] T. Rasanen, D. Voukantsis, H. Niska, K. Karatzas and M. Kolehmainen. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*, 87(11):3538-3545, 2010.

[23] I. Rowlands. Demand response in Ontario: exploring the issues. A report for the Independent Electricity System Operator (IESO) of Ontario, 2008.

[24] U.S. Annual Energy Outlook 2012. www.eia.gov/forecasts/aeo/pdf/0383(2012).pdf.

[25] W. W.-S. Wei. *Time series analysis: univariate and multivariate methods*, pages 156–157. Addison-Wesley, 2nd edition, 2006.

[26] T. K. Wijaya, J. Eberle and K. Aberer. Symbolic representation of smart meter data. In *2nd Workshop on Energy Data Management (EnDM)*, pages 242-248, 2013.