Algorithms for Upgrading the Resolution of Aggregate Energy Meter Data

Harshad Khadilkar[I], Tanuja Ganu[I], Zainul Charbiwala[I], Lim Chee Ming[U], Sathyajith Mathew[U], Deva P Seetharam[I]
[I]IBM India Research Lab [U]University of Brunei Darussalam

ABSTRACT

Metering of the energy supplied to consumers is an important component of operations for utility providers. Several schemes have been employed for this purpose, including traditional postpaid and prepaid metering, and more advanced smart metering technology. Analysis of the data generated by these meters has the potential to provide insights into consumer characteristics and power consumption patterns, including consumer segmentation and anomaly detection. We describe the different types of power purchase and consumption data, as well as the analytics algorithms that can be applied to them. Most applications developed for energy meter data require high resolution information of the type provided by smart meters, thus leaving aggregate prepaid or postpaid meter schemes at a disadvantage. In this paper, we present analytics-based methodologies to upgrade aggregate prepaid and postpaid meter data resolution, which will allow smart meter analytics to be applied without expensive infrastructure upgrades.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Energy Metering, Prepaid, Postpaid, Smart Meters

Keywords

Analytics Algorithms, Data Resolution Upgrade

1. INTRODUCTION

Energy metering technology has been in development since the 1880s [1]. Energy meters can be broadly classified into two types: postpaid and prepaid meters. The traditional business model for electricity retail involves the installation of postpaid meters at customer premises and subsequent billing for the amount of energy consumed during the previous billing period (typically a month or a quarter). Since these meters rarely have access to communication facilities, they must be read and billed manually. Prepaid meters are gaining popularity because they simplify billing operations, especially in areas where utility providers face severe non-payment issues. This metering scheme requires the customers to make advance payment for their energy. Prepaid meters are being used in many countries including Brunei, India, Ireland, South Africa and Sudan [2, 3, 4].

Conventional metering schemes (both prepaid and postpaid) are limited by the aggregate nature of measurement, which does not allow tracking of the rate of energy consumption as a function of time. This information is important for utility providers, since it can be leveraged for applications such as time-of-use pricing, demand response programs and fine-grained spatio-temporal load forecasts [5]. This shortcoming is addressed by the use of smart meters, which can record the amount of energy consumed as a function of time [6]. These meters can directly communicate time-stamped consumption data to meter data management systems. Residual concerns with regard to cost and data privacy have restricted their popularity at present [7]. Consequently, there is a need for low-cost solutions to the problem of obtaining high-resolution data from energy meters. As far as the authors can determine, the algorithm presented in Sec. 4 is the first effort to address the issue with a purely analytical approach. The principal advantage of this methodology is that no hardware changes or retrofitting is required to the aggregate energy meters.

In Sec. 3, we discuss the applicability of various analysis algorithms to different types of energy consumption data. Sec. 4 presents algorithms for extracting fine-grained temporal consumption details from the aggregate consumption data reported by conventional postpaid and prepaid meters. We present a convergence analysis for the proposed algorithms, as well as a derivation of the corresponding error bounds. In Sec. 5, we validate our assumptions using actual empirical data and also evaluate the estimation algorithms using real-world and synthetic datasets.

2. RELATED WORK

Analysis of energy meter data has received wide attention in past literature. The emphasis has been on applications such as segmentation of consumers into groups based on similarity of usage [8, 9], predicting consumer behaviour [10] and setting of power tariffs [11, 12]. Most prior studies focus on smart meter data because of the high resolution that it provides. Such data is important for applications such as fraud detection [13, 14] and real-time consumer feedback [15]. We review some of these applications in Sec. 3.

e-Energy'14, June 11–13, 2014, Cambridge, UK.

ACM 978-1-4503-2819-7/14/067 ...$15.00.
http://dx.doi.org/10.1145/2602044.2602059.

The problem of recovering high-resolution data from low-resolution aggregate data has been addressed in multiple fields of research. Data fusion in wireless networks presents challenges when the sensors are separated spatio-temporally [16]. Upgrade algorithms for sensor data are also developed in the field of compressed sensing [17, 18]. It is necessary to interpolate data being received from low-resolution sensors, in order to combine it effectively with high-resolution sensors. The problem considered in this paper has a similar objective, with aggregate energy meter data being used to estimate dynamic energy consumption.

The concept of combining data from multiple proximal sensors into a single high-resolution data stream is frequently used in climate modelling [19, 20, 21] and image processing [22]. Analogously, in this paper we combine the meter readings of consumers with similar consumption patterns to estimate dynamic energy consumption for each consumer.

3. ALGORITHMS FOR DATA ANALYSIS

As discussed in the previous section, the three major types of metering mechanisms are (i) prepaid meters, (ii) postpaid meters, and (iii) smart meters. The corresponding data formats can be broadly classified as *purchase data* for prepaid meters, *aggregate consumption* for postpaid meters, and *dynamic consumption* for smart meters. In this section, we describe the analysis algorithms that can be applied to each type of data without any interconversion. Algorithms for upgrading the resolution of purchase and aggregate consumption data are described in Sec. 4.

3.1 Analysis of purchase data

Prepaid meter data consists of logs of the number of units purchased by each consumer, with the corresponding time (and possibly location) stamps. In the absence of ground truth about rate of energy consumption, analysis algorithms for this type of data are limited to characterization of the trends in energy purchased. In this paper, we use data from Brunei Darussalam for studying these patterns empirically. The time range for this data covers three years, from January 2010 to December 2012.

It is possible to use the prepaid data to identify segments of nominal purchase patterns, using standard cluster analysis techniques such as k-means [23]. A sample plot with 7 consumer segments is shown in Fig. 1. The X axis in the figure depicts the progression of time from Jan 2010 to Dec 2012, and the Y axis shows the average monthly purchased energy for each consumer segment. Each consumer in the data is mapped to one of the seven segments. This type of consumer segmentation facilitates the automated detection of *outliers*, or consumers whose purchase patterns are not similar to any of the common patterns seen in the data. Such information may be used for further analysis, with fraud detection being an example. Fig. 2 shows a sample plot for two outliers detected in the prepaid data. The two highlighted users can be seen to have stopped purchasing energy midway through the period of analysis. This behaviour could be indicative of illegal activities such as energy purchases from unauthorized dealers. Note that aggregate data only enables us to detect gross changes in consumption patterns. A more effective anomaly detection mechanism - one that is sensitive and that can pinpoint the cause of the anomaly - requires high resolution data. This further motivates the necessity of developing algorithms such as those presented in Sec. 4.

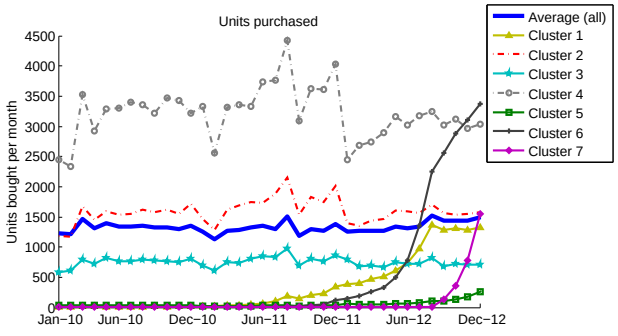

Figure 1: Segments of residential consumers separated into 7 clusters based on unit purchases. Large users can be seen to have reduced energy purchases in 2012.

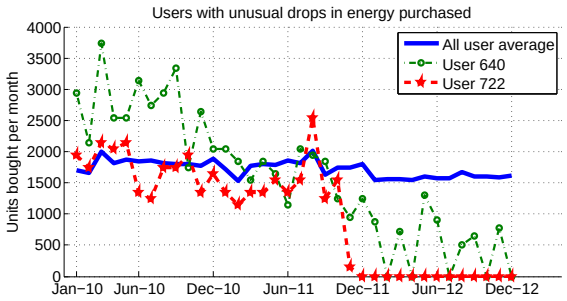

Figure 2: Purchase-based outlier detection.

3.2 Analysis of aggregate consumption data

Utilities collect aggregate consumption data from meters primarily for billing. The meter reading interval typically coincides with the billing period, which may range from a few weeks to a few months. Although aggregate meter readings are temporally sparse, they can nevertheless be used for some analytical applications. One example is that of inferring connectivity models of electricity grids. The connectivity model of a distribution network provides the underlying interconnection between various assets (such as transformers) and customers in the grid. Prior literature has shown that meter readings from a subset of the distribution points are sufficient to estimate grid connectivity [24]. The accuracy of this information deteriorates over time due to repairs, maintenance, and balancing efforts. Partial or incorrect connectivity information leads to delays and higher costs in identifying the true location of a malfunction.

Analysis of aggregate consumption data can reveal insights about consumer segments, patterns in their behavior, and potential theft [25]. It can also be used at a macro scale to determine economic development metrics. For example, the World Bank uses national energy and electricity production in their World Development Reports [26]. Additionally, Lorenz curves (commonly used by economists to estimate income inequality) are sometimes used to combine energy access and consumption into a single equity metric [27, 28]. Using data from Norway, the US, El Salvador, Thailand, and Kenya, prior studies show that the distribution of energy across consumers in industrialized nations is far more

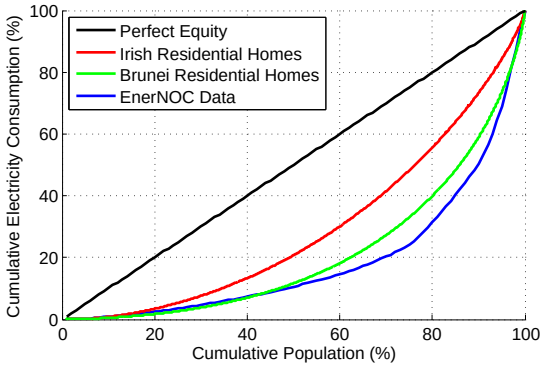

Figure 3: Lorenz curve comparison between commercial (EnerNOC) and residential (Brunei Darussalam and Ireland) aggregate consumption data.

uniform than in developing nations. Fig. 3 shows a comparison of the Lorenz curves for 100 commercial consumers in the US and 57156 and 782 residential consumers from Brunei Darussalam and Ireland respectively. The curves illustrate the higher disparity in commercial enterprises than in residential consumers, owing to the wider spectrum of companies served by the utility.

3.3 Analysis of dynamic consumption data

The proliferation of real-time sensing and feedback in electricity grids has enabled their evolution into *smart grids*. Smart meters perform sensing functions at the individual household level in smart grids [29]. These meters record energy consumption with fine granularity (5 minute to 30 minute intervals). The use of two-way communication between smart meters and utilities has allowed the implementation of applications such as outage detection, identification of demand response (DR) potential and the detection of consumption anomalies and energy theft [30, 31].

Demand response programs aim to provide higher system reliability by altering consumer demand in response to available supply and economic conditions [32, 33, 34]. These programs identify the target set of consumers based on criteria such as the day of week, time of day, peak loads and demand variability. For example, Fig. 4 shows representative smart meter data for three residential consumers in Ireland, measured over a period of six months. The variability of demand on collated on an hourly basis is shown in the form of box plots. The higher the variability of demand, the greater is the potential flexibility of the consumer. From Fig. 4, it is seen that consumer A is the most suitable for demand response in the morning, while both consumers A and B can be targeted in the evening. Note that such insights are only available through high resolution consumption data. They cannot be directly derived from traditional power purchase data or aggregate postpaid consumption data.

Anomalous consumption patterns could indicate energy theft, and are a major concern for utility providers. A common vector for energy theft is to bypass the energy meters at certain times of the day (such as late nights) or on certain days of the week (such as weekends) [35]. It is observed that only a fraction of non-technical losses due to fraud are ever detected using historical aggregated consumption data. However, dynamic consumption data can be used for

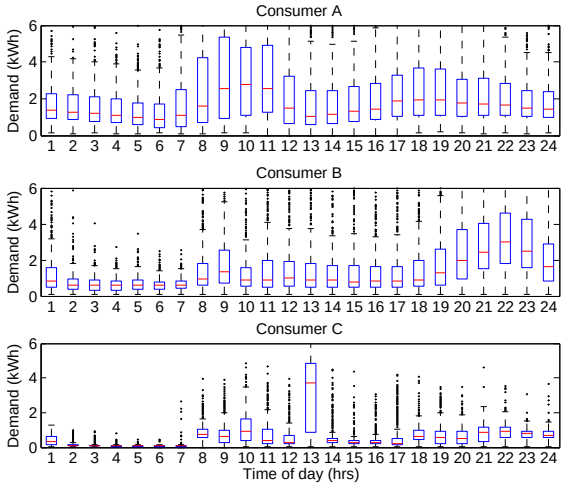

Figure 4: Selecting consumers for specific DR events based on their hourly dynamic consumption data

theft detection through advanced data analysis techniques [36, 37]. Similarly, smart meter data combined with context data (weather, public events) and demographic attributes can also used for consumer segmentation [38]. For the end-users, real-time feedback about energy usage [13, 14] can help reduce energy costs by taking advantage of time of use pricing [15]. In summary, insights from dynamic smart meter data enable utility providers to maintain efficient and reliable grid operations, while also allowing consumers to use energy more effectively.

4. UPGRADING DATA RESOLUTION

The analytics algorithms that are applicable to different types of meter data were described in Sec. 3. It was shown that the data produced by smart meters is the most useful for deep analytics and for real-time applications. The underlying property of smart meter data that makes it more useful than postpaid or prepaid meter data, is its high resolution. However, this richness of information is accompanied by higher cost, because of the accompanying communication and sensing infrastructure that needs to be installed.

In this section, we present algorithms that can upgrade the resolution of prepaid and postpaid meter data without any additional hardware installation. These algorithms are based on the similarity of consumption patterns across consumers within a single segment. As such, the temporal resolution that they can achieve is a function of the minimum time interval between data samples across different consumers. The traditional postpaid regime receives meter readings for different consumers tagged with the day of reading. Thus, the best temporal distinction between data samples is one day. As a result, this section focuses on converting aggregate meter readings to an estimate of daily consumption. We emphasize that the algorithms themselves are applicable to any two time scales, as long as the stated assumptions regarding the underlying consumption patterns are satisfied. For example, the same algorithms presented in Sec. 4.1 and Sec. 4.2 can be used to upgrade daily consumption data to an estimate of 15-minute consumption.

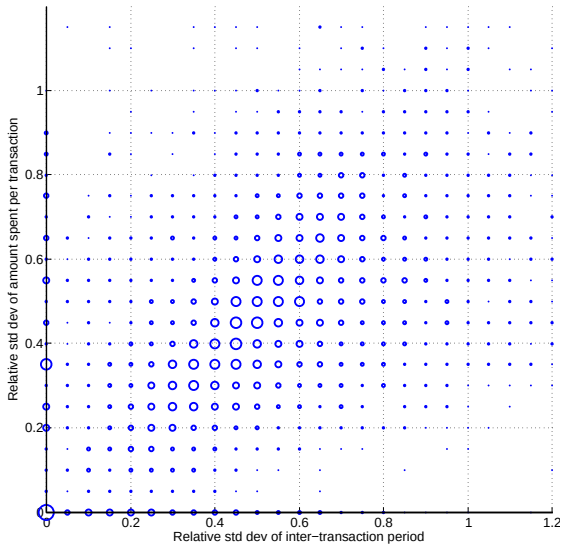

Figure 5: Transaction amount and inter-transaction period characteristics in prepaid data. Both axes are normalized by the corresponding averages for each customer.

4.1 Estimation of aggregate consumption from purchase data

Energy utilities using the prepaid meter model typically only have access to transaction data. Any conversion from monetary transactions to high-resolution consumption estimates necessarily involves an intermediate step where aggregate consumption is estimated. Therefore, in the following development, we only describe an algorithm to form a noisy estimate of aggregate consumption from prepaid transaction information. Sec. 4.2 develops the subsequent method for converting noisy aggregate consumption data to estimates of daily consumption.

Fig. 5 shows the transaction behaviour of consumers in Brunei Darussalam for the years 2010 and 2011. The Y axis shows the standard deviation in the amount per transaction for residential consumers, normalized by the average amount per transaction. The X axis shows the analogous statistic for the inter-transaction period. The size of each circle is proportional to the frequency with which the corresponding (x, y) coordinate is seen in the data. It can be seen that the behaviour across all consumers varies considerably, with the most common cases falling in the region where both sets of standard deviation are 50% of the mean. It is, however, possible to determine some estimate of the aggregate consumption based on the following assumption.

A-1 Consumers recharge their prepaid meters when the residual units in the meter drop below a certain threshold. The threshold itself may vary from consumer to consumer.

It is reasonable to expect that most consumers will have a threshold below which they judge themselves to be in danger of running out of electricity, and will therefore have an urge to recharge their meters. This assumption is clearly not valid in all cases, as seen from Fig. 5. However, in absence of accompanying consumption data with the prepaid transaction data set, we will proceed to estimate aggregate energy consumption based on **A-1**. In steady state, the 'threshold' number of units will always be present in the meter. Each fresh transaction will simply add to this number and the meter will subsequently count down back to the threshold value. The value will act like a 'displaced zero' with the following consequence:

P-1 All units purchased during one transaction are consumed by the time when the next transaction takes place. If u_1 units are purchased at time t_1 and u_2 units are subsequently purchased at time t_2, then it follows that u_1 units are consumed over the period from t_1 to t_2.

The estimate of aggregate consumption obtained using **P-1** can be fed into the algorithm described in Sec. 4.2 to estimate dynamic energy consumption. In effect, this two-step procedure converts prepaid transaction data into estimates of dynamic energy consumption.

4.2 Estimation of dynamic consumption from aggregate consumption data

In the following development, we focus on estimating daily consumption from an aggregate consumption measurement period of M days. As noted earlier, the same algorithm is applicable to any other pair of time scales. We use data from EnerNOC, a US-based utility provider, to test the validity of the assumptions made in the following treatment. Commercial energy consumption data from this provider is freely available online for a set of 100 consumers and a duration of one year [39]. The raw data is available from the utility in the form of 5-minute energy consumption information for each consumer. This is artificially aggregated in order to emulate postpaid meter readings, thus allowing us to compare the results of the estimation algorithm with ground truth.

4.2.1 Background assumptions

We will make the following assumptions before we begin the discussion of using aggregate consumption data for forming an estimate of dynamic consumption.

A-2 Monthly meter readings are taken for different consumers on different days, each aggregated over M days.

We will denote the aggregate consumption by $F_c(i)$, where $c \in \{1, 2, \ldots, N\}$ is the index of the consumer and i is the day when the reading is taken.

A-3 The consumption for each user c can be mapped to a single consumption pattern through some one to one mapping function. In this paper, we will use a simple scaling factor a_c.

This factor may be calculated based on context information such as area of premises and occupancy, or by base-lining average consumption over a previous time period. Fig. 6 depicts empirical daily energy consumption for consumers subscribing to EnerNOC. The upper plot in the figure shows the actual spread of consumption for each of the 100 users over a 60 day period. The value of the scaling constant a_c is assumed to be equal to the median daily consumption for each consumer, and the lower plot shows the resulting normalized values. The variability of the daily consumption values can be seen to be substantially smaller in the lower plot, compared to the upper plot. Fig. 7 quantifies this improvement by comparing the spread in the original data

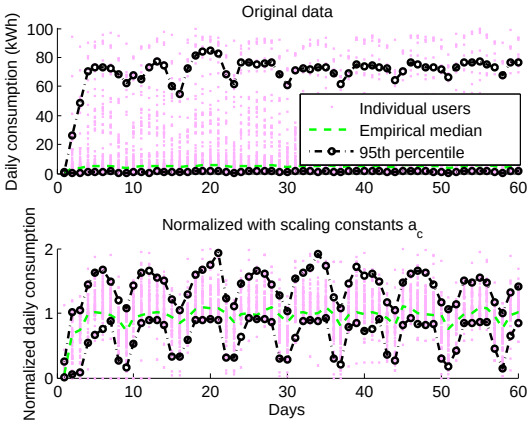

Figure 6: Comparison of the spread in daily energy consumption in the original data (top plot) and the spread in data normalized using the scaling constants a_c.

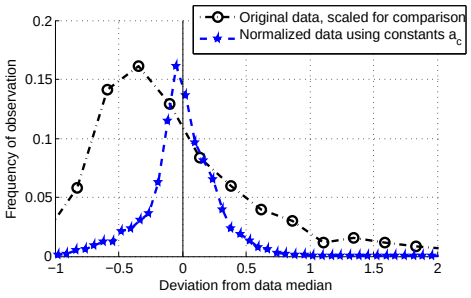

Figure 7: The distribution of normalized data around the median of the normalized data. The original data has been scaled by the original median for comparison.

with that in the normalized data. The 'original data' has been scaled by a single value (median daily consumption) across all consumers, while the 'normalized data' has been scaled using the estimated constants a_c. It can be seen that the scaling factors substantially reduce the differences across consumers, allowing the estimation of a common consumption pattern for the entire data set.

The normalized daily consumption will be denoted by $f(i)$ with i being any integer (negative values for past days), and the aggregate reading for each consumer can be calculated by the following relation.

$$F_c(i) = \sum_{j=i-M+1}^{i} a_c \; f(j).$$

The normalized *aggregate* consumption is the same across all users (according to **A-3**) and is denoted by $y(i) = \frac{F_c(i)}{a_c}$. We will now proceed to make an assumption that will allow us to build an algorithm for estimating daily consumption.

A-4 The underlying consumption function $f(i)$ is periodic with a (possibly unknown) period T. We will denote the daily consumption values by $\{z_1, z_2, \ldots, z_T\}$, and note that $z_1 = f(1) = f(T+1)$ and so on. As a consequence, the normalized aggregate consumption also becomes periodic with $y(1) = y(T+1)$ and so on.

The periodicity assumption allows us to define linear relations between the daily consumption and aggregate consumption values. Periodicity is a property seen in real-world empirical energy consumption data, as shown in Sec. 5. These relations can be solved either by directly inverting the coefficient matrix (denoted by $\mathcal{B}$ in Sec. 4.2.4) or through the iterative procedure described in Sec. 4.2.2, subject to the following three properties.

P-2 If the period T is exactly equal to the measurement period M, it is not possible to construct an algorithm for estimating daily consumption from aggregate data.

This follows from the observation that $T = M$ gives us linear equations with the same set of variables (the T unknowns $\{z_1, z_2, \ldots, z_T\}$) and the same coefficients for each equation. The equations are linearly dependent, the coefficient matrix is singular, and no estimation of the individual variables is possible [40]. A minor extension of this linear dependency argument gives us two additional properties.

P-3 Estimation of daily consumption is also not possible if T and M have one or more common prime factors.

P-4 Exact estimation of daily consumption is possible if the period T is known and the measurement period M has no common prime factors with T. The number of aggregate readings required for this estimation is T.

If there are no common prime factors between T and M, we can write T linearly independent equations with T readings $\{y(1), \ldots, y(T)\}$, and simply invert the coefficient matrix to calculate the exact values of each of the unknowns $\{z_1, z_2, \ldots, z_T\}$ [40]. We will make one final assumption before proceeding to develop an iterative estimation algorithm for daily consumption.

A-5 The period T can be independently calculated from prior empirical data.

This is a reasonable assumption to make, with one possible approach for calculating T involving matching the aggregate supply and demand characteristics for the utility provider.

4.2.2 *Iterative estimation algorithm*

If the period T is known exactly, and there is no noise in the system and in the measurements, then inverting the linear relation between daily and aggregate consumption is the simplest way of estimating daily consumption. However, these conditions are unlikely to be satisfied in a realistic setting. When there is noise in the data and/or trends in consumption, it is desirable to implement the estimation procedure over a rolling window of size T. For large T, it is expected that an iterative algorithm initialized with the estimates from the previous window will be computationally faster than solving the full set of linear equations each time. The functional period T can be very large when real-time estimation of consumption is required. For example, 15-minute resolution estimates with a consumption periodicity of one day (1440 minutes) will imply $T = \frac{1440}{15} = 96$. The computational time for the algorithm is even more important if real-time demand response signals are to be generated based on the consumption estimates.

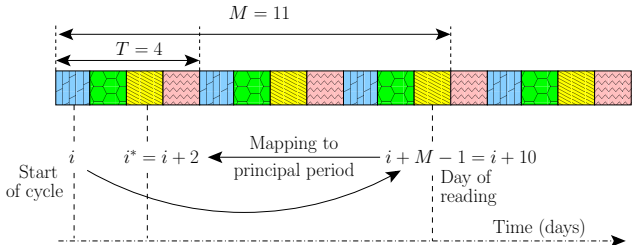

Figure 8: Relation between the start of the measurement cycle i, and the reading date mapped to the principal period i^*. For illustration, it is assumed that $T = 4$ and $M = 11$.

The uniqueness of the solution to a non-singular linear system of equations [40] ensures that the problem formulation satisfies one of the necessary conditions for convergence of iterative update algorithms [41]:

P-5 If conditions **A-4** and **A-5** are met and T and M have no common prime factors, then there is only one unique solution $\{z_1, z_2, \ldots, z_T\}$ to an observed aggregate consumption pattern $\{y(1), \ldots, y(T)\}$.

We now proceed to define the structure of the estimation algorithm. Let us assume that the period T is known, and denote the estimated daily consumption variables by $\hat{z}_i$ and the predicted normalized aggregate consumption by $\hat{y}(j)$. We further assume that the measurement period is M and that $n = \lfloor \frac{M}{T} \rfloor$ is the number of full functional cycles in each measurement period. Consider a measurement cycle that begins on day i. The normalized reading for this cycle will be $y(i + M - 1)$, taken on day $(i + M - 1)$. By the periodicity assumption, this reading will be equal to $y\left((i + M - 1) - T\lfloor \frac{i+M-1}{T} \rfloor\right)$, which falls within the principal period $\{1, \ldots, T\}$. For simplicity, we will use the symbol i^* to denote the mapping of $(i + M - 1)$ to the principal period. Fig. 8 clarifies the relation between i and i^*. We now define the iterative update algorithm for each variable $\hat{z}_i$ to be,

$$\hat{z}_i^+ = \hat{z}_i^- + k\,[y(i^*) - \hat{y}(i^*)], \tag{1}$$

where $\hat{z}_i^-$ is the estimate of z_i before the update step and $\hat{z}_i^+$ is the estimate after the update step. Only one $\hat{z}_i$ is updated at a time. The predicted measurement $\hat{y}(i^*)$ is calculated using the relation,

$$\hat{y}(i^*) = n \sum_{j=1}^{T} \hat{z}_j + \sum_{m=i}^{i^*} \hat{z}_m. \tag{2}$$

Note that the second summation may involve a roll-over if $i^* < i$. In that case, the summation will be evaluated over $\hat{z}_m$ with $m \in \{i, i+1, \ldots, T, 1, 2, \ldots, i^*\}$. The estimation mechanism in Eq. (1) is to compensate for the prediction error by increasing the estimate of the first day in the measurement cycle, with a gain equal to k. **P-5** guarantees that termination can occur only with the unique solution.

4.2.3 *Selection of the update gain k*

There are two types of error associated with the iterative algorithm given by Eq. (1). The first kind is introduced by noise in the aggregate measurements, and fundamentally limits the accuracy of the estimated daily consumption (independently of the estimation algorithm). A treatment of this kind of error is given in Sec. 4.2.4. The second type of error relates to successive iterations of Eq. (1) for the same set of T aggregate readings, and is a property of the applied estimation algorithm. We can show that this can be driven to zero through a judicious selection of the update gain k. The post-update estimation error can be computed by subtracting Eq. (1) from the true value z_i:

$$\begin{aligned}
z_i - \hat{z}_i^+ &= z_i - \hat{z}_i^- - k\,[y(i^*) - \hat{y}(i^*)] \\
&= z_i - \hat{z}_i^- - k\left[\left(n\sum_{j=1}^{T} z_j + \sum_{m=i}^{i^*} z_m\right) - \right. \\
&\quad \left.\left(n\sum_{j=1}^{T} \hat{z}_j + \sum_{m=i}^{i^*} \hat{z}_m\right)\right] \\
&= z_i - \hat{z}_i^- - k\left[n\sum_{j=1}^{T}(z_j - \hat{z}_j) + \sum_{m=i}^{i^*}(z_m - \hat{z}_m)\right] \\
z_i - \hat{z}_i^+ &= [1 - (n+1)\,k]\,(z_i - \hat{z}_i^-) \\
&\quad - (n+1)\,k \sum_{m=i+1}^{i^*} (z_m - \hat{z}_m) \\
&\quad - n\,k\left[\sum_{j=1}^{i-1}(z_j - \hat{z}_j) + \sum_{p=i^*+1}^{T} (z_p - \hat{z}_p)\right].
\end{aligned} \tag{3}$$

The last step follows from the fact that there are $(n + 1)$ copies of individual error terms $(z_j - \hat{z}_j)\ \forall j \in \{i, \ldots, i^*\}$, and n copies for all other j. Consider evaluating Eq. (3) for $i = 1$, and let 1^* be the equivalent of i^*. If we denote the error terms $(z_j - \hat{z}_j)$ by e_{rj} where $j \in \{1, \ldots, T\}$ and r is the iteration number, then the error for $j = 1$ after the first iteration is given by,

$$e_{11} = \begin{bmatrix}(1-(n+1)k) & -(n+1)k & \ldots & -nk\end{bmatrix} \begin{bmatrix} e_{01} \\ e_{02} \\ \vdots \\ e_{0T}\end{bmatrix}. \tag{4}$$

For ease of representation, consider the simple case where $T = 3$ and $M = 2$, which means that the consumption pattern repeats every third day and measurements are taken on every second day. The number of complete cycles involved in each reading is $n = 0$ and $1^* = 2$ (last day of the measurement period which started on day 1). Instantiating Eq. (4) for this example, the error after the first update is,

$$\begin{aligned}
e_{11} &= \begin{bmatrix}(1-k) & -k & 0\end{bmatrix} \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03}\end{bmatrix} \\
\Rightarrow \begin{bmatrix} e_{11} \\ e_{02} \\ e_{03}\end{bmatrix} &= \begin{bmatrix}(1-k) & -k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1\end{bmatrix} \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03}\end{bmatrix} = \mathcal{A}_1 \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03}\end{bmatrix}.
\end{aligned} \tag{5}$$

The next update corresponds to the aggregate reading $z_3 + z_1$, and is equivalent to substituting $i = 3$ in Eq. (1). This update will use the updated version of $\hat{z}_1$ with the corresponding error e_{11}. Therefore, the post-update error vector in Eq. (5) is now modified to,

$$\begin{bmatrix} e_{11} \\ e_{02} \\ e_{13}\end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -k & 0 & (1-k)\end{bmatrix} \begin{bmatrix} e_{11} \\ e_{02} \\ e_{03}\end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} e_{11} \\ e_{02} \\ e_{13} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -k & 0 & (1-k) \end{bmatrix} \mathcal{A}_1 \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03} \end{bmatrix} = \mathcal{A}_3\, \mathcal{A}_1 \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03} \end{bmatrix}.$$

The final update in the first stage of iteration will be for $\hat{z}_2$, and the error terms after this step will be given by,

$$\begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & (1-k) & -k \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{02} \\ e_{13} \end{bmatrix} = \mathcal{A}_2 \begin{bmatrix} e_{11} \\ e_{02} \\ e_{13} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \end{bmatrix} = \mathcal{A}_2\, \mathcal{A}_3\, \mathcal{A}_1 \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03} \end{bmatrix} = \mathcal{A} \begin{bmatrix} e_{01} \\ e_{02} \\ e_{03} \end{bmatrix}. \quad (6)$$

In the last step, we have denoted the product of the three error update matrices by the matrix $\mathcal{A}$. The entries of this matrix are only a function of k and are well known. All further iterations involve the same three update steps, and are equivalent to pre-multiplying Eq. (6) by $\mathcal{A}$ each time.

P-6 The convergence of the estimation algorithm is determined by the eigenvalues of $\mathcal{A}$. As the system is discrete, errors e_{rj} are guaranteed to decay to 0 iff the dominant eigenvalue is smaller than 1 in magnitude [42].

The $\mathcal{A}$ matrix can be calculated for any period T once the order of updates is known ($\hat{z}_1 \rightarrow \hat{z}_3 \rightarrow \hat{z}_2$ in Eq. (6)). The update gain can be selected by calculating the eigenvalues of the corresponding $\mathcal{A}$ matrix as a function of k, and choosing any value that guarantees stability, as shown in Sec. 5.

4.2.4 Confidence bounds for estimation errors

In the following development, we consider the effect of error in the aggregate measurements on the estimates of daily consumption. We assume a structure for the introduction of noise into the system, and then proceed to analyse the problem for various properties of the noise function.

A-6 There is an underlying periodic function z_i with period T ($i \in \{1, 2, \ldots, T\}$), and the actual daily consumption is the sum of z_i and some form of additive noise.

The assumption of periodicity is not restrictive, because (i) periodicity is seen in most real-world data sets for energy consumption, and (ii) the noise terms can account for small variations in the estimation of T.

The simplest case to analyse is that for a single consumer, where the estimation algorithm uses T successive aggregate readings to compute $\hat{z}_i$. Since each reading period is M days long, the total time period before these estimates can be computed is equal to MT. We make the following assumption for analysing the statistical properties of the estimated daily consumption for a single consumer:

A-7 Additive noise in consumption on day i is equal to w_i, zero-mean, independent and identically distributed (i.i.d). The consumption on day i is thus $(z_i + w_i)$.

We have not specified the distribution of w_i beyond the i.i.d assumption, because:

P-7 As long as the noise is zero-mean i.i.d, the central limit theorem ensures that the sum of the noise variables approaches a zero-mean normal distribution with variance equal to the variance of $\sum w_i$ [43].

The j^{th} normalized aggregate measurement for a single consumer will be given by the relation,

$$y(jM) = \sum_{i=(j-1)M+1}^{jM} (z_i + w_i),$$

and will be composed of M days. The periodicity assumption implies that $z_{j+T} = z_j$, but this relation need not hold true for the noise terms w_j. The set of values $\{\hat{z}_1, \ldots, \hat{z}_T\}$ will be estimated from the measurements $\{y(M), \ldots, y(MT)\}$. There are MT i.i.d noise terms involved with no overlap between readings, thus implying the following property:

$$\begin{aligned} \text{var}(y(jM)) &= \text{var}\left(\sum_{i=(j-1)M+1}^{jM} (z_i + w_i) \right) \\ &= \underbrace{\text{var}\left(\sum_{i=(j-1)M+1}^{jM} z_i \right)}_{\text{constants}} + \underbrace{\text{var}\left(\sum_{i=(j-1)M+1}^{jM} w_i \right)}_{M \text{ i.i.d variables}} \\ &= 0 + M\, \text{var}\,(w_i). \end{aligned}$$

P-8 The variance of each aggregate reading $y(jM)$ is M times the variance of each day's consumption.

Let us denote the nominal daily consumption $[z_1, \ldots, z_T]'$ by Z, where $'$ is the matrix transpose operator. Similarly, let the estimated daily consumption be $\hat{Z} = [\hat{z}_1, \ldots, \hat{z}_T]'$, and the vector of aggregate readings be $Y = [y(M), \ldots, y(MT)]'$. Regardless of the method of estimation (matrix inversion or the iterative algorithm), the estimated daily consumption is the unique solution of the relation $\mathcal{B}\hat{Z} = Y$, where $\mathcal{B}$ is composed of the coefficients of $\hat{z}_i$ in Eq. (2). The entries of this matrix are equal to the number of copies of each z_i in the aggregate readings, and are equal to n or $(n+1)$. **P-4** tells us that $\mathcal{B}$ is non-singular and hence invertible. The estimated daily consumption is thus given by,

$$\hat{Z} = \mathcal{B}^{-1}\, Y.$$

The error bounds for the estimation algorithm are related to the statistical properties of $\hat{Z}$:

P-9 The estimation procedure is unbiased since the expectation of the estimate of daily consumption is given by,

$$\mathbb{E}\,[\hat{Z}] = \mathbb{E}\,[\mathcal{B}^{-1}\, Y] = \mathcal{B}^{-1}\, \mathbb{E}\,[Y] = Z.$$

The last equality follows from **A-7** (the noise in each aggregate reading is zero-mean). To calculate the variance of $\hat{Z}$, we first calculate the expectation of the outer product $\hat{Z}\hat{Z}'$:

$$\begin{aligned} \mathbb{E}\,[\hat{Z}\hat{Z}'] &= \mathbb{E}\,\left[(\mathcal{B}^{-1}\, Y)(\mathcal{B}^{-1}\, Y)' \right] \\ &= \mathcal{B}^{-1}\, \mathbb{E}\,[YY'](\mathcal{B}^{-1})'. \end{aligned} \quad (7)$$

The right hand side of Eq. (7) is easy to compute numerically, because the matrix $\mathcal{B}$ is well known and $\mathbb{E}[YY']$ can be approximated using observed aggregate readings. Equivalently, the left hand side can be directly calculated from the empirical mean of $\hat{Z}\hat{Z}'$. In either case, we note the following property for single-customer consumption estimation.

P-10 While the estimate $\hat{Z}$ can be computed from T aggregate readings, several more sets of aggregate readings are required to compute the error bounds on $\hat{Z}$.

The variance of each z_i can be calculated from the diagonal elements of $\mathbb{E}[\hat{Z}\hat{Z}']$ using the following relation.

$$\text{var}(\hat{z}_i) = \mathbb{E}[\hat{z}_i^2] - (\mathbb{E}[\hat{z}_i])^2.$$

A combination of **P-9** and the central limit theorem allows us to equate the variance of the estimates $\hat{z}_i$ to the variance of the errors $(z_i - \hat{z}_i)$.

P-11 For a sufficiently large number of aggregate readings, the error $(z_i - \hat{z}_i)$ is Gaussian with zero mean and standard deviation $\sigma_i = \sqrt{\text{var}(\hat{z}_i)}$.

When there are multiple consumers in the data, the analysis is different from that of the single-consumer case. We show that the variability across consumers tends to increase the estimation error, but if the consumer segmentation has been correctly carried out, increasing the number of consumers in the data set helps drive the error down. We will assume that **A-3** nominally holds true, but that there may be a temporary increase/decrease in each customer's consumption relative to the normalized mean (from **A-3**, the normalization constant for each consumer c is a_c). Therefore, the noise in the aggregate readings across consumers and over the measurement time period M will have two sources.

A-8 Individual trends in consumption (not accounted for by the normalization by a_c) for each consumer c are represented by the i.i.d random variable v_c. In addition, the random variation around the nominal pattern for consumer c on day i is represented by w_{ci}, also i.i.d with $i \in \{1, 2, \ldots, M\}$. The total energy consumption for consumer c on day i is given by $z_{ci} = z_i + v_c + w_{ci}$, where z_i is the mean nominal consumption for that particular consumer segment.

As a consequence of this assumption, the normalized aggregate reading for consumer c for a measurement period starting on day i is given by,

$$\begin{aligned} y_c(i+M-1) &= \sum_{j=i}^{i+M-1} z_{cj} \\ &= \underbrace{\sum_{j=i}^{i+M-1} z_j}_{\text{Mean trend}} + \underbrace{M\, v_c}_{\text{Cust. specific}} + \underbrace{\sum_{j=i}^{i+M-1} w_{cj}}_{\text{Random variation}}. \end{aligned}$$

In order to be able to estimate the mean trend of consumption, we compute the average across consumers of all readings taken on day $(i+M-1)$. Let $\mathcal{C}_i$ denote the set of consumers whose meter readings are taken on this day, and let $|\mathcal{C}_i|$ be the number of such consumers. Then the average normalized reading $y(i+M-1)$ is given by,

$$\begin{aligned} y(i+M-1) &= \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} y_c(i+M-1) \\ &= \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \sum_{j=i}^{i+M-1} z_j + \frac{M}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} v_c \\ &\quad + \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \sum_{j=i}^{i+M-1} w_{cj}. \end{aligned}$$

Since the mean trend z_j is the same across consumers, the first summation opens out and we get,

$$\begin{aligned} y(i+M-1) = &\sum_{j=i}^{i+M-1} z_j + \frac{M}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} v_c \\ &+ \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \sum_{j=i}^{i+M-1} w_{cj}. \end{aligned} \quad (8)$$

Each batch of T variables for the mean trend $\{\hat{z}_i, \ldots, \hat{z}_{i+T-1}\}$ is computed using the set of T consumer-averaged readings $\{y(i+M-1), \ldots, y(i+M+T-2)\}$. Computing the variance of (8) provides useful insights into the estimation accuracy:

$$\begin{aligned} \text{var}(y(i+M-1)) &= \text{var}\left(\sum_{j=i}^{i+M-1} z_j\right) + \text{var}\left(\frac{M}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} v_c\right) \\ &\quad + \text{var}\left(\frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \sum_{j=i}^{i+M-1} w_{cj}\right) \\ &= 0 + \frac{M^2}{|\mathcal{C}_i|^2} \text{var}\left(\sum_{c \in \mathcal{C}_i} v_c\right) \\ &\quad + \frac{1}{|\mathcal{C}_i|^2} \text{var}\left(\sum_{c \in \mathcal{C}_i} \sum_{j=i}^{i+M-1} w_{cj}\right) \\ &= \frac{M^2\,|\mathcal{C}_i|}{|\mathcal{C}_i|^2} \text{var}(v_c) + \frac{M\,|\mathcal{C}_i|}{|\mathcal{C}_i|^2} \text{var}(w_{cj}). \end{aligned}$$

P-12 The variance of each consumer-averaged aggregate reading is thus given by,

$$\text{var}(y(i+M-1)) = \frac{M^2}{|\mathcal{C}_i|} \text{var}(v_c) + \frac{M}{|\mathcal{C}_i|} \text{var}(w_{cj}). \quad (9)$$

Note the difference between **P-8** and **P-12**. While the variance in the former case scaled linearly with M, the variance in the latter case scales quadratically with M. However, it is possible to drive this variance down by increasing the number of readings $|\mathcal{C}_i|$ taken on any given day. The measurement period M at which the first term in Eq. (9) starts dominating the second term depends on the relative ratio between $\text{var}(v_c)$ and $\text{var}(w_{cj})$. This fact emphasizes the necessity of accurate consumer segmentation: $|\mathcal{C}_i|$ should not be increased at the cost of an increase in $\text{var}(v_c)$.

The same procedure as demonstrated in Eq. (7) can be used for estimating the confidence bounds on $\hat{z}_i$. The only modification required is that the empirical values of Y used for calculating the expectation will be the consumer-averaged aggregate readings, instead of the time-averaged readings for a single consumer.

5. EXPERIMENTAL EVALUATION

We now validate the data resolution upgrade algorithm described in Sec. 4.2. Simulated data is used for evaluating the iterative estimation algorithm and the confidence bounds for the single consumer case. Empirical data from EnerNOC, a utility based in the United States, is used for evaluating the confidence bounds for estimation of daily consumption for multiple consumers. We reiterate that the raw data has a resolution of 5 minutes, and is artificially aggregated in order to simulate postpaid meter readings. Fig. 9 shows that data from consumers subscribing to this utility has strong periodicity with a principal period of one week

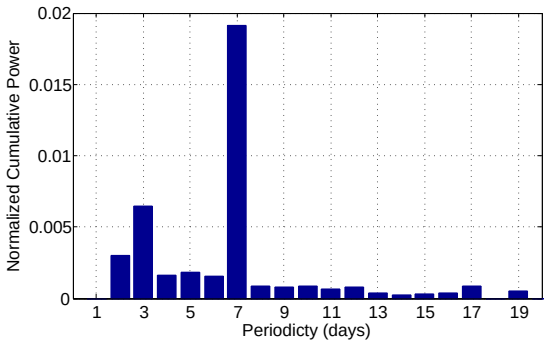

Figure 9: Validation of the assumption of periodicity in energy consumption, for the EnerNOC data set.

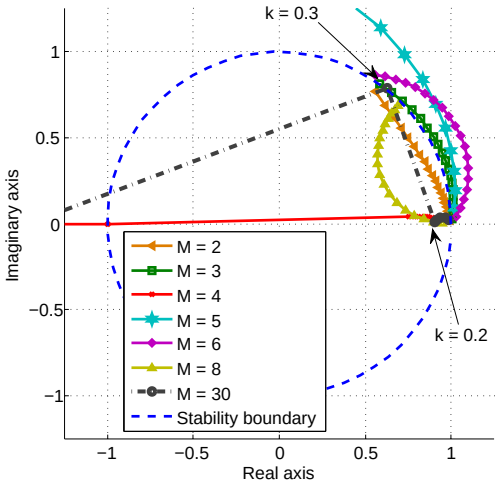

Figure 10: The evolution of the dominant eigenvalue of $\mathcal{A}$ for a weekly consumption cycle, and various values of M. Locations of $k = 0.2$ and $k = 0.3$ for $M = 30$ are marked.

($T = 7$). It depicts the cumulative normalized power spectral distribution for daily energy consumption for all the users in the data set, and indicates that the consumption patterns across users are not only periodic, but have the same periodicity.

Fig. 10 illustrates the implications of **P-6** for the case of a weekly consumption cycle and for various potential values of the aggregate measurement period M. Each curve traces the locus of the dominant eigenvalue of $\mathcal{A}$, parametrized by the gain k. A monthly measurement cycle can be approximated by the curve for $M = 30$. The unit circle in Fig. 10 marks the boundary of the stable region for the iterative algorithm. All the curves originate at $(1, 0)$ for $k = 0$. This value represents the open-loop case, when the coefficient matrix $\mathcal{A}$ is the identity matrix. Choosing a value of k such that the dominant eigenvalue falls within the unit circle ensures convergence of the algorithm (1), and vice versa. Conformance to this property is demonstrated for simulated data in Fig. 11, where the iterative algorithm is seen to converge for $k = 0.2$ (within the unit circle in Fig. 10) but not for $k = 0.3$ (outside the unit circle).

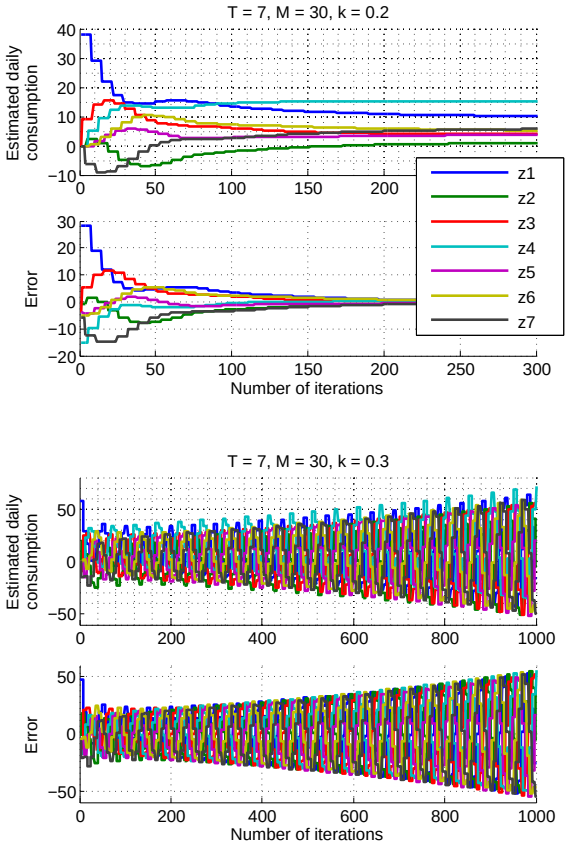

Figure 11: Simulated estimation runs for $T = 7$ and $M = 30$. The case with $k = 0.2$ converges to the correct estimates, while that with $k = 0.3$ does not. The magnitudes of the corresponding dominant eigenvalues of $\mathcal{A}$ are 0.914 and 1.008.

An illustration of the confidence bounds derived in **P-11** for the single-consumer case is shown in Fig. 12. The nominal daily consumption was assumed to be $\{z_1, \ldots, z_7\} = \{50, 20, 80, 40, 110, 90, 140\}$, with a measurement cycle of 30 days. Uncorrelated random noise $w_j \sim \mathcal{N}(0, 0.5)$ was added to the nominal values, and the algorithm (1) was used to estimate the daily consumption. $\mathbb{E}[YY']$ was calculated empirically by taking aggregate consumption readings over several cycles. The 99.7% error bounds in Fig. 12 are equivalent to ± 3 standard deviations of the estimated noise in $\hat{Z}$. Since the confidence intervals for individual $\hat{z}_i$ are nearly equal to each other, only one set is shown.

Analogous bounds for consumption estimates based on data from multiple consumers are shown in Fig. 13. The estimates are based on a real-world data set consisting of 100 consumers subscribing to the utility EnerNOC. The data was available for a period of one year. The normalization constant a_c for each consumer was calculated by computing the average daily consumption for that consumer over the first four months in the data. The remaining eight months were used for testing the estimation algorithm.

Aggregate meter readings were simulated based on the measured daily consumption for each consumer. The date of reading i_c for each consumer c was generated from a uniform distribution on the range $[1, M]$, where M was the length of the measurement period. The normalized aggregate read-

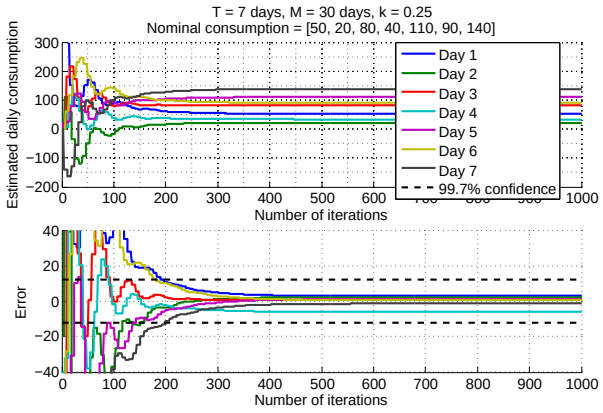

Figure 12: Dynamics of the estimation algorithm and error bounds for a single consumer.

ings for that consumer were calculated by summing the daily consumption over the time periods $[i_c - M + 1, i_c]$, $[i_c + 1, i_c + M]$ and so on, and then dividing each sum by the constant a_c. Since the data set was limited both in terms of number of consumers and the available time period, the number of readings $|\mathcal{C}_i|$ generated for any particular day i was inversely proportional to the measurement time period M. As a consequence, Eq. (9) suggests that the estimation quality for the mean as well as the tightness of the confidence bounds will deteriorate rapidly with increasing M.

Fig. 13 illustrates this property by using three sample values for M. The ground truth for daily consumption for all 100 consumers is depicted by the light pink dots, while the actual mean consumption across all consumers is shown by a dashed red line. The estimated consumption, based on readings aggregated over M days, is shown with a solid black line and the corresponding 95% confidence intervals (± 2 standard deviations) by hollow black circles. Eq. (9) indicates that the confidence intervals are related to the variation v_c *across consumers* rather than the random noise w_{ci} around the *mean trend*. The estimates as well as the bounds are based on the empirical value of $\mathbb{E}[YY']$, and can therefore be improved by increasing the number of readings (either by increasing the number of consumers or by taking more readings over the course of time). The case with $M = 8$ is seen to produce the tightest confidence intervals, with 95% of all consumers falling within the corresponding confidence bounds on each day.

The dominance of the noise term v_c over w_{cj} in Eq. (9) can be seen when recovering the consumption estimates for each consumer, scaled back to the original values. Since the trend v_c is expected to dominate the random variation w_{cj}, the estimate for each consumer is computed by shifting the estimated segment means $\hat{z}_i$ by an amount proportional to the difference in the individual readings $y_c(i)$ and the mean reading $y(i)$. Empirical results show that this adjustment reduces the RMS error in daily consumption for 94% of consumers in the data set. If the trend v_c were not dominating w_{cj}, this procedure would increase/decrease the RMS error in a purely random fashion: a 50% performance. The significantly better results seen in empirical data show that the $M^2 \operatorname{var}(v_c) \gg M \operatorname{var}(w_{cj})$ assumption in **P-12** is valid.

Finally, we demonstrate the application of the estimation algorithm for identifying and analysing anomalous consump-

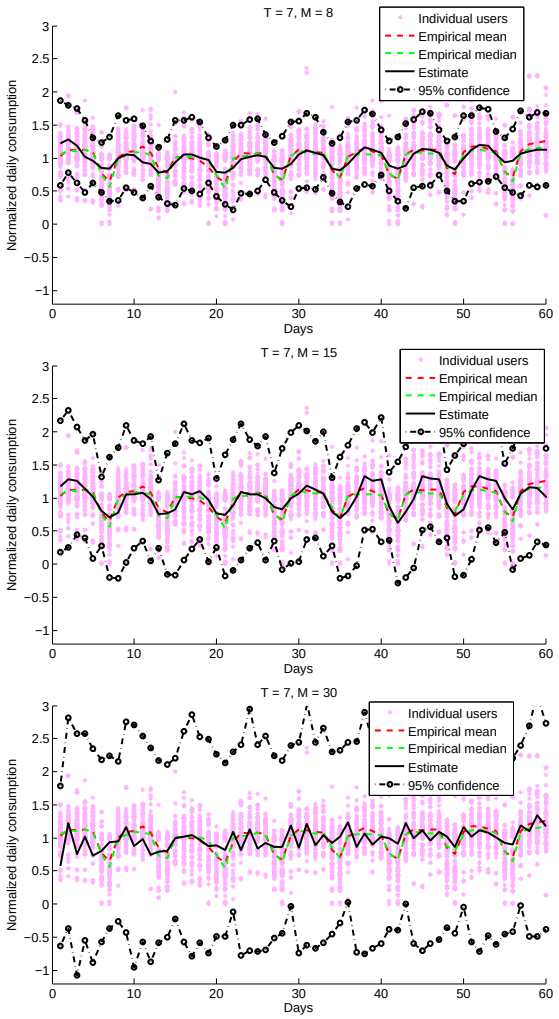

Figure 13: Estimation of the mean trend and error bounds for measurement periods $M = 8$, 15 and 30 days. The same number of aggregate readings are taken per day, in all three cases. It can be seen that the confidence intervals grow larger with increasing M.

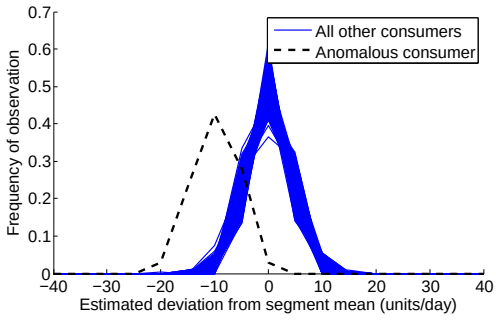

Figure 14: Distribution of estimated consumer-specific deviation v_c.

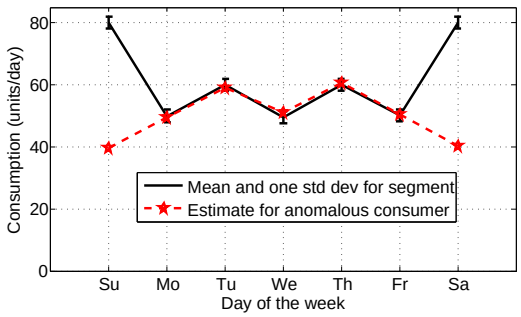

Figure 15: Estimated daily consumption across all consumers, as well as for the anomalous consumer.

tion patterns. We simulate daily consumption for a set of 1400 consumers with a periodicity of $T = 7$, and a nominal daily consumption pattern of $[80, 50, 60, 50, 60, 50, 80]$ units. Meter readings are taken in a staggered fashion, with a measurement time period of $M = 30$ days for each consumer. The consumer-specific deviation v_c in each measurement cycle is drawn from the normal distribution $\mathcal{N}(0, 1)$ while the daily random variation w_{cj} is drawn from the distribution $\mathcal{N}(0, 10)$. Only one of the 1400 consumers is assumed to have a consumption of 40 units on the first and last day of each weekly consumption cycle, as opposed to 80 units. The estimation algorithm is run on simulated data for several measurement cycles, and the estimated consumer-specific variation v_c is noted. It can be seen from Fig. 14 that the anomalous consumer shows a clear difference from the rest of the consumers, even though the standard deviation of w_{cj} is an order of magnitude larger than that of v_c. Estimation of daily consumption based only on aggregate readings from the anomalous consumer is shown in Fig. 15. It clearly emphasizes the differentiation capabilities of the algorithm presented in Sec. 4.2.

6. CONCLUSION

Our objective in writing this paper was to present a comparative study of the technical limits for analysing each type of meter data (prepaid, postpaid and smart meters). We characterized the usefulness of several algorithms for analysing aggregate as well as dynamic consumption data. These functions include consumer segmentation as well as automated detection of anomalous consumption patterns. The depth of insight available from the data was shown to improve with data resolution. Consequently, we also described analytical approaches to extract fine-grained temporal consumption details from coarse-grained aggregate consumption data. The proposed algorithms were analysed from the point of view of stability and the statistical properties of estimation error. Empirical data was used to show the efficacy of the estimation algorithms for several applications, including the detection and characterization of anomalous consumption patterns. We believe that the methodology presented in this paper can be applied effectively by utility providers for acquiring high quality consumption data from existing infrastructure, at relatively low cost.

Acknowledgements

The authors would like to thank Abdul Salam Bin Hj Abdul Wahab and Norihramniza Hj Ramlee from the Energy Department at the Prime Minister's Office, Brunei Darussalam and Siti Kadzijjah Binti Abd Latiff and Pg Jamra Weira Bin Pg Hj Petra from the Department of Electrical Services, and Abdul Aziz bin Haji Mohamad Ali from the University of Brunei Darussalam for providing prepaid purchase data.

7. REFERENCES

[1] D. Dahle, "A brief history of meter companies and meter evolution," http://www.watthourmeters.com/history.html.

[2] Electric Ireland, "Pay as you go meters," https://www.electricireland.ie/ei/residential/manage-your-account/pay-as-you-go.jsp.

[3] A. Raj, "Prepaid meters to plug power theft and loss of revenue," http://www.telegraphindia.com/1130319/jsp/bihar/story_16687604.jsp#.UtuqL3nT62w.

[4] Standard Transfer Specification (STS) Committee, "STS membership growth," http://www.sts.org.za.

[5] F. E. R. Commission, "Assessment of demand response and advanced metering," http://www.ferc.gov/legal/staff-reports/2013/oct-demand-response.pdf.

[6] "Smart Grid Deployment Tracker 3Q13," http://www.navigantresearch.com/research/smart-grid-deployment-tracker-3q13.

[7] D. Cornish, "The case against smart meters," http://www.wired.co.uk/news/archive/2012-12/21/smart-meters.

[8] S. J. Moss, M. Cubed, and K. Fleisher, "Market segmentation and energy efficiency program design," *California Institute for Energy and Environment Berkeley*, 2008.

[9] S. Ramos, Z. Vale, J. Santana, and J. Duarte, "Data mining contributions to characterize MV consumers and to improve the suppliers-consumers settlements," in *IEEE Power Engineering Society General Meeting*, 2007, pp. 1–8.

[10] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–12, 2013.

[11] C. Flath, D. Nicolay, T. Conte, C. Dinther, and L. Filipova-Neumann, "Cluster analysis of smart metering data," *Business & Information Systems Engineering*, vol. 4, no. 1, pp. 31–39, 2012.

[12] T. Räsänen and M. Kolehmainen, "Feature-based clustering for electricity use time series data," in *Proceedings of the 9th International Conference on Adaptive and Natural Computing Algorithms*, ser. ICANNGA'09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 401–412.

[13] OPOWER, "Opower: Energy reporting," http://opower.com/products/energy-reporting.

[14] Bidgely, "Itemize your energy bill," http://www.bidgely.com/.

[15] T. Bapat, N. Sengupta, S. K. Ghai, V. Arya, Y. B. Shrinivasan, and D. Seetharam, "User-sensitive scheduling of home appliances," in *Proceedings of the*

2nd ACM SIGCOMM workshop on Green networking. ACM, 2011, pp. 43–48.

[16] D. Ganesan, S. Ratnasamy, H. Wang, and D. Estrin, "Coping with irregular spatio-temporal sampling in sensor networks," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 1, pp. 125–130, January 2004.

[17] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[18] Y. Tsaig and D. Donoho, "Extensions of compressed sensing," *Signal Processing*, vol. 86, no. 3, pp. 549–571, 2006.

[19] A. Moberg, D. Sonechkin, K. Holmgren, N. Datsenko, and W. Karlen, "Highly variable northern hemisphere temperatures reconstructed from low- and high-resolution proxy data," *Letters to Nature*, vol. 433, pp. 613–617, December 2004.

[20] T. Mitchell and P. Jones, "An improved method of constructing a database of monthly climate observations and associated high-resolution grids," *International Journal of Climatology*, vol. 25, no. 6, pp. 693–712, May 2005.

[21] P. Jones, T. Osborn, K. Briffa, C. Folland, E. Horton, L. Alexander, D. Parker, and N. Rayner, "Adjusting for sampling density in grid box land and ocean surface temperature time series," *Journal of Geophysical Research: Atmospheres*, vol. 106, no. D4, pp. 3371–3380, February 2001.

[22] D. Long, P. Hardin, and P. Whiting, "Resolution enhancement of spaceborne scatterometer data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 3, pp. 700–715, 1993.

[23] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Symposium on mathematical statistics and probability*, Berkeley, CA, 1967.

[24] V. Arya, T. Jayram, S. Pal, and S. Kalyanaraman, "Inferring connectivity model from meter measurements in distribution networks," in *Proceedings of the fourth international conference on Future energy systems*. ACM, 2013, pp. 173–182.

[25] C. Bandim, J. Alves Jr, A. Pinto Jr, F. Souza, M. Loureiro, C. Magalhaes, and F. Galvez-Durand, "Identification of energy theft and tampered meters using a central observer meter: a mathematical approach," in *Transmission and Distribution Conference and Exposition, 2003 IEEE PES*, vol. 1. IEEE, 2003, pp. 163–168.

[26] World Bank, "World development reports," http://www.worldbank.org.

[27] A. Jacobson, A. D. Milman, and D. M. Kammen, "Letting the (energy) gini out of the bottle: Lorenz curves of cumulative electricity consumption and gini coefficients as metrics of energy distribution and equity," *Energy Policy*, vol. 33, no. 14, pp. 1825–1832, 2005.

[28] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The Review of Economics and Statistics*, vol. 54, no. 3, pp. 306–316, 1972.

[29] S. Harrison, "Smart Metering Projects Map," http://bit.ly/GSs1o5, 2013.

[30] Z. Pollock, "The True ROI of Smart Meter Deployments," http://bit.ly/1dK33jW, July 2013.

[31] S. Depuru, L. Wang, and V. Devabhaktuni, "Smart meters for power grid: challenges, issues, advantages and status," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 6, pp. 2736–2742, 2011.

[32] N. Armaroli and V. Balzani, "The future of energy supply: challenges and opportunities," *Angewandte Chemie International Edition*, vol. 46, no. 1-2, pp. 52–66, 2007.

[33] U.S. Department of Energy, "Benefits of demand response in electricity markets and recommendations for achieving them, 2006," http://eetd.lbl.gov/ea/emp/reports/congress-1252d.pdf.

[34] K. Spees and L. Lave, "Impacts of responsive load in PJM: load shifting and real time pricing," *The Energy Journal*, vol. 29, no. 2, pp. 101–122, 2008.

[35] K Rowland, "Detecting theft and enhancing billing processes only two of many potential applications," http://www.intelligentutility.com/article/12/03/meter-data-applied, March 2012.

[36] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," *Critical Information Infrastructures Security*, vol. 6027, pp. 176–187, 2010.

[37] S. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007–1015, 2011.

[38] T. Wijaya, T. Ganu, D. Chakraborty, K. Aberery, and D. P. Seetharam, "Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data," in *SIAM International Conference on Data Mining*, 2014.

[39] EnerNOC, "2012 Commercial Energy Consumption Data," http://open.enernoc.com/data/.

[40] E. Kreyszig, *Advanced Engineering Mathematics*. John Wiley & Sons, 2010.

[41] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.

[42] R. Williams and D. Lawrence, *Linear State-Space Control Systems*. John Wiley & Sons, 2007.

[43] D. Bertsekas and J. Tsitsiklis, *Introduction to Probability*. Athena Scientific, 2008.