VTT Technical Research Centre of Finland

Hierarchical Multiplicative Model for Characterizing Residential Electricity Consumption

Kuusela, Pirkko; Norros, Ilkka; Reittu, Hannu; Piira, Kalevi

Published in:
Journal of Energy Engineering

DOI:
10.1061/(ASCE)EY.1943-7897.0000532

Published: 01/03/2018

Document Version
Early version, also known as pre-print

Hierarchical multiplicative model for characterizing residential electricity consumption

Pirkko Kuusela[1] Ph.D. Ilkka Norros[2] Prof. Hannu Reittu[3] Ph.D. and Kalevi Piira[4]

[1]VTT Technical Research Centre of Finland Ltd., P.O. Box 1000,
FI-02044 VTT, Finland. pirkko.kuusela@vtt.fi (corresponding author),

[2]VTT Technical Research Centre of Finland Ltd., P.O. Box 1000,
FI-02044 VTT, Finland. ilkka.norros@vtt.fi,

[3]VTT Technical Research Centre of Finland Ltd., P.O. Box 1000,
FI-02044 VTT, Finland. hannu.reittu@vtt.fi,

[4]VTT Technical Research Centre of Finland Ltd., P.O. Box 1000,
FI-02044 VTT, Finland. kalevi.piira@vtt.fi

Abstract

This work presents a hierarchical multiplicative framework for modeling the energy consumption of households. The constituents of the model are a lognormally distributed annual consumption, an annual consumption profile at week resolution, a mean weekly consumption profile, and a multiplicative lognormally distributed random variation. Further, the annual and weekly profiles of households are shown to fall naturally into a small number of rather homogeneous groups, identified by the Regular Decomposition method. The framework is adapted to monitor and compare populations of electricity consumers. On the other hand, it provides a convenient way to produce synthetic traces of household energy consumption with similar stochastic properties as measured traces. It is also shown how additional household information can be utilized to predict both the annual consumption and the random variation of the consumption of a household.

Keywords: household electricity consumption, mathematical modeling, clustering, profiles, monitoring

INTRODUCTION

Local (district level and building embedded) renewable energy production is growing globally. This causes challenges like how to solve increasing energy grid balance problems; how to design and optimize local energy production, consumption and the use of the energy storage; how to cut or shift consumption; and how to operate with more fluctuating energy prices. In the near future this means new businesses and huge global markets to Information and Communication Technology (ICT) solutions for smart grid management in addition to ICT solutions for smart grid adaptable buildings. These challenges are difficult to solve without reliable and scalable forecasting of energy consumption at household, building, block, and district levels. One important piece in the solution of these challenges is to study models to characterize a customer's power consumption with a few parameters.

By utilizing Automatic Meter Reading (AMR) data of residential energy consumption, this paper focuses on characterizing and comparing populations of consumers. The aim is to develop efficient and illustrative parameters that allow

1. comparison and trending of different consumer populations,
2. communicating all essential elements of consumption, including its volatility, and
3. easy generation of consumption traces with realistic random variation.

The random variation around regular patterns is an essential part of the presented model. In fact, the volatility plays an important role in the control of future low voltage grids, prompting the research beyond the regular consumption patterns. Although the high volatility of the households' energy consumption has been recognized and it is becoming more important in the future grid control, the random variation around consumption patterns has mostly been neglected in modeling.

The motivation of this work is to answer the following research questions: i) How to model and parameterize electricity consumption of households with few parameters in such a way that realistic variability in consumption traces can be generated? ii) How to monitor and compare populations of consumers? iii) How to direct towards automatized handling of Big Data by repeated use of autonomous algorithms after initial tuning and validation?

The idea of the presented approach is to decompose the consumption into the following components per customer: i) total annual consumption, ii) annual profile, iii) mean week profile and iv) multiplicative random variation, which consists of the difference of the actual consumption time series from the repeating annual and weekly profiles arising from the above components. The authors propose lognormal models for elements i) and iv) and, depending on the context, clustering of the profiles ii) and iii).

The main contribution of this work is the identification of multiplicative lognormal noise as a maximally simple way to characterize and monitor the random volatility component by one parameter at minimum. The authors also apply a recently developed grouping (clustering) method that is favorable to handle large amounts of data. Consumption clusters and profiles are illustrative and valuable as such, but this approach integrates them into a consumption modeling and monitoring framework as parameters. The overall approach of this paper is holistic, touching many popular problems of energy consumption modeling.

There is a vast recent literature on electricity consumption, and the authors bring up only those results from AMR data literature that are closely related to the methodology and ideas of this paper.

See McLoughlin et al. (2015) and McLoughlin et al. (2012) for clustering and a review of the same Irish AMR data as utilized in this work. Chicco (2012) presents an overview and performance assessment of the clustering methods for electrical load pattern grouping. A finite mixture model of Gaussian multivariate distributions is introduced in Haben et al. (2016) as an alternative to the popular k-means clustering. This could be an interesting framework to be related to the findings on lognormality, as some of the clustering attributes could benefit from a log-transformation and this work could provide a further attribute describing the random variation. The stability of clustering is studied in Haben et al. (2016) by bootstrapping methods. Clustering of hourly data has been done by utilizing shape dictionaries of consumption patterns and magnitude as a multiplicative factor in Kwac et al. (2014), which is close to the multiplicative decomposition presented here. A multiplicative model in clustering is used in Räsänen et al. (2010) as well, without considering the random variation. Recent developments in clustering include also the clustering of particular time periods and the use of pre-processed load shapes to obtain efficient compression of large data (Kwac et al. 2016; Wang et al. 2017; Haben et al. 2016). Hierarchical methods are used in this context as well, but the focus is directed to grid management, demand response and control, whereas the load prediction for grid control lies outside the scope of this paper. This paper contributes to the methodology of consumption clustering by applying the novel Regular Decomposition clustering method (Reittu et al. 2014; Reittu et al. 2017), which the authors believe to have potential in future needs of clustering, e.g., automated handling of dynamic large data.

The proposed hierarchical multiplicative modeling paradigm is motivated by the reported lognormality of energy consumption at various time scales, see (Kuusela et al. 2015; Mutanen et al. 2012; Kwac et al. 2014; Kolter and Ferrera 2011) and the properties of lognormality in Kuusela et al. (2015). For other approaches, see the review Grandjean et al. (2012) of developing consumption traces either by top-down or bottom-up approaches, where the traces in the popular bottom-up approach are obtained by mimicking appliances and generating user and appliance behaviors in various ways.

This paper also studies the modeling of the random variation of the annual consumption and the volatility component by relating them to other household characteristics, touching the field of energy consumption survey data analysis. For a review of studies and factors affecting electricity consumption, see Jones et al. (2015), Gouveia and Seixas (2016), and Beckel et al. (2014). Clustering and survey are combined in a recent paper by Gouveia and Seixas (2016). This points also to earlier studies on survey methods in electricity consumption. Beckel et al. (2014) extracted 34 features of consumption to reveal household characteristics from the very same data as used in this paper.

The outcomes of this paper can be useful to practitioners in various ways. The analysis section provides relatively simple methods for stochastic simulation of the consumption to be utilized, e.g., in populating network models and designing demand response programs as well as in designing new architectural setups, algorithms and decision support tools to utilize distributed energy resources in meeting the demands. Moreover, the authors take a viewpoint of monitoring household energy consumption and propose an intuitive and efficient collection of variables to be measured and monitored. This provides means for electricity distribution companies to trend, compare and predict the consumption. In this framework, it is possible to study both the profiles and their clustering together with the random variation around cluster profiles. The utilized data provide an opportunity to model a household's total consumption without interference of households' own energy production or demand response. Models for the total consumption are needed in, e.g., smart city research. On the other hand, the energy consumption of households is changing in the near future due to the increase in distributed generation and smart devices. This work provides means to observe this change in different scales via trends in monitoring variables.

The paper is structured as follows. The research data are summarized in Section 2. In Section 3, the four-layer model is presented and its accuracy is studied. Section 4 is devoted to grouping the annual and weekly profiles by the Regular Decomposition method. The problem of monitoring electricity consumption at a population level is addressed in Section 5 by finding suitable consumption monitoring parameters. Possibilities to make inference on electricity consumption based on additional information about households are discussed in Section 6. Finally, the proposed method is validated with unseen data in Section 7. The conclusions are drawn in Section 8.

CHARACTERISTICS OF THE RESIDENTIAL SMART METER DATA AND SURVEY

The developed methods are illustrated with the popular dataset of the Irish Smart Metering Trial Archive (2012). The Irish trial took place during 2009 and 2010. The data include smart meter readings at 30 min intervals and participant background data in a survey format. The data set analyzed in this paper covers 995 households during the 364 first days (i.e., full weeks) of year 2010. The sample was selected by including all customers heating their house with electricity (either by central heating or using plug-in heaters) and randomly picking from the rest homes having uninterrupted records. The authors were interested to study how the different heating methods are reflected in the electricity consumption traces. This amounted to the inclusion of 238 homes with electrical heating and 757 homes heated with other energy sources. Besides the Irish data, the elements of this methodology were developed with Finnish urban consumer data containing both households and small and medium-sized enterprises (SMEs).

HIERARCHICAL ANALYSIS IN A MULTIPLICATIVE FRAMEWORK

The hierarchical analysis of electricity consumption presented in this section will be the modeling framework through the whole paper. The weeks are indexed by $i = 1, \ldots, 52$ and the half-hour time intervals of a week by $t = 1, \ldots, 336$. The electricity consumption C of a household H in half-hour t of week i is then written as

$$C_t^H(i) = W^H \times \frac{y^H(i)}{52} \times \frac{a_t^H}{336} \times \xi_t^H(i), \tag{1}$$

where

$$\begin{aligned}
W^H &= \text{the total annual electricity consumption of the household} \\
y^H(i) &= \text{the weight of week } i \text{ in the household's annual consumption profile} \\
a_t^H &= \text{the weight of half-hour } t \text{ in the household's mean week profile} \\
\xi_t^H(i) &= \text{the relative multiplicative variation of consumption around the mean} \\
& \text{profile in week } i \text{ and time } t.
\end{aligned}$$

The annual and weekly profiles are scaled so that $1/52 \sum_{i=1}^{52} y^H(i) = 1$ and $1/336 \sum_{t=1}^{336} a_t^H = 1$ and, by construction, the irregular variation $\xi_t^H(i)$ has mean 1 as well. Thus, the annual total consumption is the only element with an absolute magnitude, while the annual and weekly profiles present the relative distribution of the consumption in time.

Energy consumption has been reported to follow lognormal distribution at several time scales: annual scale in Kuusela et al. (2015) and Mutanen et al. (2012), daily scale (with lognormal mixtures) in Kwac et al. (2013), and hourly scale in Chen and Cook (2012) and Mutanen et al. (2012). Kolter and Ferrera (2011) present log-log plots of energy consumption vs. living area, together with the lognormality of the former. In general, many natural phenomena are multiplicative and generate lognormal distributions (Limbert et al. 2001). Multiplication preserves lognormality, which in part suggests the chosen multiplicative model and consumption profile approach.

The next step is to analyze these hierarchical elements one by one with the aim to study distributions of variables and suitable models for random variation. In this section, each household H has individual consumption parameters W^H, y^H, a^H, and a model parameter for ξ. For clarity, however, the superscript H will be dropped from the notation in the next section.

Model elements for a single customer

As already mentioned, lognormal modeling of the annual consumption W was studied comprehensively in Kuusela et al. (2015), and the authors adopt it in this paper as well. Besides a mostly good fit with the body of the empirical distribution, lognormal modeling allows heavy tails as well as straightforward transfer of the simple characterization of dependencies in multivariate Gaussian models. Figure 1 presents the body and tail fits (inset) of the present data with both lognormal and Weibull distributions. The inset shows that the Weibull distribution underestimates large consumptions considerably. Due to other reported results on lognormality of electricity consumption in various time scales in Kuusela et al. (2015), Mutanen et al. (2012), Chen and Cook (2012), Kwac et al. (2013), and Kolter and Ferrera (2011) as well as the ease of modeling with lognormal distributions, the lognormal model is preferred.

Let us then consider the annual and weekly profiles of a household. Recall that in this paper the term 'profile' means a vector with mean one. Define a customer's year profile $y(\cdot)$

as the vector of the consumption in each of the 52 weeks divided by the total consumption W and multiplied by 52 so that the mean of the vector's components is 1. Similarly, define for each week i the household's profile $\lambda_{\cdot}(i)$ as the vector of the half-hour consumptions divided by the total consumption in week i, multiplied by 336. The mean week profile of the household is then defined as

$$a_t = \frac{1}{52}\sum_{i=1}^{52} \lambda_t(i). \tag{2}$$

The ratio

$$\xi_t(i) = \frac{\lambda_t(i)}{a_t}, \quad i = 1, \ldots, 52, \ t = 1, ..., 336 \tag{3}$$

can now be defined as the multiplicative random variation of the household's energy consumption around its mean profile during week i. Thus, the profiles have been decomposed multiplicatively as $\lambda_t(i) = a_t \xi_t(i)$. Note also that $\frac{1}{52}\sum_{i=1}^{52} \xi_t(i) = 1$ for every t.

Although $\xi_{\cdot}(i)$ depends on i, it was noticed that most households have rather stable week profiles in the sense that the process $\xi_t(i)$ retains its character over varying i. Remarkably, the overall marginal distribution of $\xi_{\cdot}(\cdot)$ was found to be close to lognormal for most households. The Kolmogorov-Smirnov distance (maximum deviation between distribution functions) between each household's random variation distribution and a fitted lognormal distribution is illustrated in Figure 2. The distance varies in [0,0.1] with mean 0.049, but with few large deviations. Although the deviation is big in some individual cases, the marginal distribution of the multiplicative random variation is mostly well approximated by a lognormal distribution.

Since the random variation ξ has mean 1 by construction, its approximating lognormal distribution is characterized by a single parameter, for example by $\mathrm{Var}\left(\log(\xi)\right)$. Moreover, Figure 3 shows that the parameters $\mathrm{Var}\left(\log(\xi)\right)$ of households are themselves lognormally distributed.

In the following, $\xi_t(i)$ will be modeled by a stationary process with a lognormal marginal distribution. Before doing this, it is worth of considering the nature of this simplification in detail. Most households behave qualitatively similarly as the following example (household

29). The left plot of Figure 4 shows the whole process $\log \xi_t(i)$ over a year: a steady random "cloud". The visual homogeneity is, however, deceptive, because the variance of $(\log \xi_t(\cdot))$ turns out to vary with t with a strong daily pattern. The right plot of Figure 4 presents, for each $t \in \{1, \ldots, 336\}$, the empirical variance of $(\log \xi_t(i))_{i=1}^{52}$ (blue curve). The mean week profile of the household is shown for comparison (red curve). Note that the variance is not a monotone function of the mean profile value. Moreover, their shapes differ widely from household to household. However, rough lognormality holds also in this detailed level — the parameter of each lognormal variable then just depends on t, and this picture is very similar for most households. Such a model would, however, be unattractive for practical purposes. We leave now the challenge of more accurate modeling of the $\xi_\cdot(\cdot)$ processes for the future and look for maximally simple models.

Most mean profiles a_t vary strongly in t (see examples in Section 4), and their marginal distributions can be rather considered as approximately lognormal, with mean one, than approximately Gaussian. An important observation made in this work is that the variation $\xi_t(i)$ depends very weakly on the weekly mean variation a_t. Both time series are close to lognormal, so their logarithms are close to Gaussian, and their dependence is well captured by the respective correlation. The uncorrelatedness of $\log a$ and $\log \xi$ is equivalent to the equality $\mathrm{Var}\,(\log \lambda) = \mathrm{Var}\,(\log a) + \mathrm{Var}\,(\log \xi)$. (The values of $\mathrm{Var}\,(\log \lambda)$ and $\mathrm{Var}\,(\log \xi)$ are estimated using all the weeks, and one can use $\log 0 = 0$ when needed.) Figure 5 shows to what extent this holds. The numbers $\mathrm{Var}\,(\log \lambda)$, $\mathrm{Var}\,(\log a)$, and $\mathrm{Var}\,(\log a) + \mathrm{Var}\,(\log \xi)$ are plotted for all households in the order of increasing $\mathrm{Var}\,(\log \lambda)$.The first is almost always equal to or a bit smaller than the third, i.e., logarithms of the mean profile a and the multiplicative random variation ξ are slightly negatively correlated. Figure 6 shows these correlations in the same order as the previous figure, and they range between $[-0.2, 0]$, with mean -0.089.

A study of the temporal behavior of the variation process ξ indicates that, in the mean, the random variation (relative to the household's mean week profile) that happens at a time point is almost uncorrelated to what happens after 12 hours, but clearly (0.2) positively correlated to what happens after 24 hours and even after 48 hours again. This is illustrated in Figure 7.

In order to take into account the, albeit small, dependence between ξ and a, as well as a part of the time correlation, the authors propose modeling the ξ-processes as being conditioned on the mean week profile process a, and fitting the lag 1 cross-correlations. Note that although a is non-random, its variance and lag 1 correlation may be computed as for any time series. By forming for each household H time series ξ_n and a_n over all measurements, $n = 1, \ldots, 336 \times 52$, the empirical covariance matrix $\mathrm{Cov}\,(\log \xi_n, \log \xi_{n+1}, \log a_n, \log a_{n+1}))$ is calculated, where notation n and $n+1$ refers to studying consecutive measurements. By averaging all such covariance matrices over all households, the mean covariance matrix

$$\begin{aligned} &\mathrm{Mean}(\mathrm{Cov}\,(\log \xi_n, \log \xi_{n+1}, \log a_n, \log a_{n+1})) \\ &= \begin{bmatrix} 0.635 & 0.338 & -0.042 & -0.039 \\ 0.338 & 0.635 & -0.029 & -0.042 \\ -0.042 & -0.029 & 0.391 & 0.357 \\ -0.039 & -0.042 & 0.357 & 0.391 \end{bmatrix} \end{aligned} \tag{4}$$

is obtained. As $\mathrm{Cov}\,(\xi_n, a_{n+1})$ and $\mathrm{Cov}\,(\xi_{n+1}, a_n)$ differ, the process is not invertible in time.

Synthesis of simulated consumption

The mean covariance matrix (4) suggests that the random variation ('noise') processes $\log \xi_t^H(i)$ be almost independent from the mean profiles $\log a_t^H$, whereas the processes $\log \xi_t^H(i)$ have strongly positive lag 1 autocorrelation and differ therefore clearly from white noise. Note that (4) presents the mean of all household-specific covariance matrices. In order to assess the significance of the temporal dependence structure of the variation processes, two sets of simulated traces were generated for each household: one where the true process $\log \xi_t^H(i)$ was replaced by mean 1 lognormal i.i.d. random variables with the household-specific variance, and one using instead the household-specific empirical covariance matrix $\mathrm{Cov}\left(\xi^H, a^H\right) := \mathrm{Cov}\left(\log \xi^H_n, \log \xi^H_{n+1}, \log a^H_n, \log a^H_{n+1}\right)$. Figure 8 compares the variability in measured and model-generated consumptions at 30 min intervals.

The correlated random variation traces produces the same amount of variance for non electric heating consumers (index range 1 - 757). For heaters, the model overestimates the variability. The simple i.i.d. model produces larger variability than the one in measured traces, but the difference is not dramatic, and also this model could be satisfactory for some purposes. Figure 9 provides details on how well the minimum, median, and maximum of the weekly consumption maximum are reproduced. The correlated model is slightly better than the i.i.d. one in predicting the median maximum consumption, while both clearly tend to produce too large overall maxima. Figure 8 suggests that electric heaters might form a special group in the modeling.

GROUPING OF ANNUAL AND WEEKLY PROFILES

In the previous section, the consumption traces of households were modeled by the hierarchical multiplicative model with parameters

$$(W^H, y^H, a^H, \mathrm{Cov}\left(\xi^H, a^H\right)) \quad \text{or} \quad (W^H, y^H, a^H, \mathrm{Var}\left(\log \xi^H\right)), \tag{5}$$

where the household-specific profiles y^H and a^H are vectors with lengths 52 and 336, respectively. There is no a priori theoretical model that would generate the observed variety of annual and weekly mean profiles of a household population. However, the number of model parameters can be reduced drastically by replacing the individual annual and average week profiles by mean profiles of relatively homogeneous subgroups of the population. By doing so, two similarly grouped populations can be compared with each other by comparing the relative sizes of corresponding groups in the two populations.

In order to test the stability of the proposed methodology, the original data were split into two populations in such a way that both populations had about 24% of heaters. This gives rise to a monitoring development population of 746 households and a test population of 249 households. The latter will be used in Section 7 to validate the outcome of the present section.

Grouping of consumption profiles by Regular Decomposition

The Regular Decomposition method

Clustering algorithms typically divide a data set into groups of elements that are near each other according to some metric. In contrast, the recently developed Regular Decomposition method (Reittu et al. 2014; Reittu et al. 2017; Pelillo et al. 2016) aims at a grouping that is optimal in terms of an information-theoretic criterion, the Minimum Description Length Principle, Grünwald (2007). Consider a set of customers $\mathcal{C}$, each having a non-negative time series $(x_t^{(c)})_{t\in T}$, $c \in \mathcal{C}$. Let $\mathcal{P}$ be a finite partition of $\mathcal{C}$. As explained in (Reittu et al. 2014; Reittu et al. 2017), the quantity

$$Comp(x_\cdot^{(\cdot)}|\mathcal{P}) = \sum_{B\in\mathcal{P}} \sum_{c\in B} \sum_{t\in T} D\left(x_t^{(c)} \middle\| \frac{1}{|B|} \sum_{c'\in B} x_t^{(c')} \right), \tag{6}$$

where $D(\beta\|\alpha) = \alpha - \beta + \beta \log(\beta/\alpha)$ is the Kullback-Leibler divergence between the distributions Poisson(α) and Poisson(β), estimates the dominant term of the bit length of a code that describes the data assuming that the partition $\mathcal{P}$ captures all structure (non-randomness) present in it (The full code contains also other terms with lesser order of magnitude). For each positive integer k, the partition

$$\mathcal{P}_k^* = \underset{|\mathcal{P}|=k}{\arg\min}\, Comp(x_\cdot^{(\cdot)}|\mathcal{P}) \tag{7}$$

presents the best grouping into k blocks. Finally, a practically optimal k can be identified as the smallest k for which the improvement $Comp(x_\cdot^{(\cdot)}|\mathcal{P}_k^*) - Comp(x_\cdot^{(\cdot)}|\mathcal{P}_{k+1}^*)$ remains below some small threshold value. Note that popular clustering methods like k-means lack an inherent principle for the selection of k.

It is remarkable that such a grouping can be found in a computationally efficient way. The algorithm presented in Reittu et al. (2014) starts with a random grouping into k blocks and proceeds as a greedy optimization algorithm. As discussed in Reittu et al. (2017), Regular Decomposition has its roots in the mathematics of large structures like graphs and tensors, suggesting a generic applicability of this approach in the separation of structure and randomness in large data. The authors prefer to use the word *group* in the context of Regular Decomposition, as the word *cluster* suggests that the cluster members be close to each other in some metric, which need not always hold.

Grouping of annual profiles

A regular decomposition of the annual profiles suggested six groups denoted by `A`... `F`, see Figure 10. The model development data contains about 180 households heating with electricity, but only the group `C` with 80 members has a large difference between summer and winter consumption. (Recall that the consumption values are scaled, so the profiles show how a household's total consumption spreads throughout the year.) The second largest group `B` has almost steady consumption throughout the year. The small groups `D, E,` and `F` are similar, but `D` and `E` show an increase in consumption levels for the third quarter Q3 suggesting cooling or other summer time usage. The last three groups are quite small, but the authors wanted to keep these. The analysis in Section 3 showed that the heaters differ to some extent from non-heaters, and the authors hoped to catch the group of heaters by detecting a usage pattern that differs from the majority (the outcome will be examined later in this paper).

The obtained profile shapes are remarkably similar to the ones in Gouveia and Seixas (2016), where a grouping was done by Ward's method involving both the pattern and the magnitude of the consumption, contrary to the present method that separates those two. The degree of independence of the total consumption level and profile grouping will be examined in Section 5.

Grouping of average week profiles

Figure 11 illustrates the regular decomposition of the average week profiles and the rich variety in the mean profiles in each group, denoted by `a` ... `i`. Now the optimal number of groups is clearly higher than what was needed for the annual profiles, and there are no very small groups. Most profiles show Mon-Fri vs. Sat-Sun patterns, and these reveal different weekly rhythms of the households' activities.

In contrast to McLoughlin et al. (2015) and Kwac et al. (2014), the grouping was done for full weeks instead of individual days. In McLoughlin et al. (2015), the median was used to select a daily profile that a household used most of the time, putting more weight on weekday patterns. Kwac et al. (2014) had a day shape dictionary of 1000 shapes, and they addressed the variability of day shapes by performing an entropy analysis. The groupings in Haben et al. (2016) and Kwac et al. (2016) use particular time periods of day to group the consumption with one European and one US dataset. The key time periods vary somewhat depending on the consumer population.

The authors found that households have rather constant weekly rhythms, and the consumption evolution through the days of a week is by itself interesting. The authors have also performed an unpublished analysis of urban consumer data containing households and SMEs over 20 districts that illustrated different characteristic weekly rhythms in residential, commercial, and SME industrial districts.

Comparison of measured and groupwise synthesized traces

This section examines at household level the impact of replacing a household's individual annual and weekly profiles by the ones obtained as the average of the profiles within its annual and weekly profile group, respectively. An example is shown in Figure 12 with 30 min consumption traces of a two week period. The measured traces at the top are compared with two alternative synthetic counterparts. The middle row presents the simplest multiplicative model of Section 3 that models the random variation ξ^H by the i.i.d. random variable. The bottom row replaces the individual profiles by the means of their groups. Both synthesized traces are similar to each other as the magnitude of the random variation exceeds the impact of the difference in the profile component. As expected, both models produce larger peak consumption values than the measured ones as illustrated in Figure 9. In addition, the measured traces show more clearly the underlying regular profile shape than the synthesized traces. A shortcoming of the random variation model is seen at the maximum consumption level, and there is also too large variability when the consumption is small or moderate. The authors leave the further tuning of the random variation component model for future work and continue here with the monitoring approach.

MONITORING THE PARAMETERS OF A POPULATION

The authors propose that a population of energy consumers could be monitored by calculating the following variables from the AMR data:

1. total annual consumptions: the parameters of a lognormal distribution
2. annual profile: profiles and the frequencies of profile groups
3. average weekly profile: profiles and the frequencies of profile groups
4. random variation around the profiles: the parameters of lognormal distributions.

This would result in four variables per consumer, i.e., total consumption W, annual profile group, week profile group and a model parameter of random variation, ξ, such as $\mathrm{Var}(\log(\xi))$. In addition to these, there would be $N \times 52$ and $M \times 336$ matrices containing annual and weekly group profile vectors, respectively, with grouping the population into N annual and M weekly groups. By following and comparing these variables, the essential characteristics of residential electricity consumption can be captured. These form a feasible set of monitoring parameters in the following sense: i) they have the power to represent relevant aspects of consumption realistically, ii) the set is minimal and the variables are almost independent from each other (see below), iii) the estimation of parameters is robust, and iv) the comparison of consumer populations is easy. The comparison of populations can be done by comparing lognormal distributions and the frequencies of annual/weekly profiles. It is also easy to generate artificial populations for network models and demand response studies by picking consumer parameters independently from each other.

Independence of the total consumption level and the annual profile group: Chi square testing of the total consumption (taken with a granularity of 5 MW) and the annual consumption groups shows a dependence between variables due to the three very small groups (that, moreover, have low total consumption levels). When those groups, comprising only 34 members, are removed, the chi square test value becomes 0.84. Thus, for the rest of the data, the annual profile group and the total annual consumption are independent of each other.

Independence of the annual and weekly profile groups: The annual and weekly profile groups show weak dependence in the model development data (and independence in the test data). By studying the mutual information values between partitions and the expected information between corresponding random partitions, the authors conclude that the annual and weekly profile groups are not informative on each other and can be considered as independent from each other.

RELATING HOUSEHOLD CHARACTERISTICS TO CONSUMPTION PARAMETERS

This section takes advantage of the associated survey data in order to model the total annual consumption and the random variation. Relating the household characteristics to the consumption is not necessary for monitoring purposes, but such models would allow deeper understanding and offer more possibilities in the generation of new realistic consumption populations. The authors attempted to find a profile classifier based on the household characteristics. However, no valuable linkage was found, even for central heaters. This outcome is in line with Gouveia and Seixas (2016), McLoughlin et al. (2012), and McLoughlin et al. (2015). A low correlation between energy usage behavior and geodemographics is also reported in Haben et al. (2013).

Stochastic models for the total annual consumption and the random variation parameter

The number of persons, the number of rooms, and the home floor area increase a household's energy consumption (Gouveia and Seixas 2016; Jones et al. 2015). These will be related to the total annual consumption and the amount of random variation. The authors have applied these characteristics successfully in earlier research with Finnish and Irish data to model the total annual consumption (Kuusela et al. 2015). Moreover, using such data is practical as it is typically available, and it is close to housing district planning data as well. Data mining methods applied to the Irish data in Beckel et al. (2014) were successful in inferring the occupancy, the number of persons and the number of appliances and, with some difficulties, the floor area and the number of bedrooms from 34 energy consumption features derived from the dataset.

This paper utilizes the multivariate lognormal model and the notation from Kuusela et al. (2015) to derive multivariate lognormal distributions for the vectors (P, F, B, W) and (P, F, B, V), where P=number of persons + 0.5, F=home floor area, and B=number of bedrooms + 0.5, and W= total consumption in MWh, $V = \mathrm{Var}(\log(\xi))$ (the addition of 0.5 to P and B is only for plotting purposes). The multivariate lognormal distribution is parameterized by $\boldsymbol{\mu}$, the vector of mean values of the log-transformed variables, and Γ, the covariance matrix of the log-transformed variables. The estimated model parameter $\boldsymbol{\mu}$ equals $(1.185, 5.014, 1.4254, 2.172)$ and the parameter Γ equals

$$\begin{bmatrix} 0.190, & 0.053, & 0.032, & 0.110 \\ 0.053, & 0.160, & 0.053, & 0.083 \\ 0.032, & 0.053, & 0.047, & 0.046 \\ 0.110, & 0.083, & 0.046, & 0.280 \end{bmatrix}, \tag{8}$$

for the vector (P, F, B, W). The respective parameters for the (P, F, B, V)-vector are

$(1.185, 5.014, 1.425, -0.435)$ and

$$\begin{bmatrix} 0.190, & 0.053, & 0.032, & -0.015 \\ 0.053, & 0.160, & 0.053, & -0.045 \\ 0.032, & 0.053, & 0.047, & -0.021 \\ -0.015, & -0.045, & -0.021, & 0.211 \end{bmatrix}. \tag{9}$$

The estimation results are listed for the two estimations as the interest will be to estimate W or V, given P, F, and B.

Figure 13 presents the marginal densities of multivariate lognormal fits to the target variables at the top row. The lognormal distribution fits very well to W and V. Lognormal fits are rather good for P and F as well, but the variable B is skewed to the opposite direction in comparison to the other variables. The granularity and the concept of a bedroom might be a bit problematic, see the discussion in Kuusela et al. (2015).

VALIDATION WITH THE TEST POPULATION

This section studies i) the stability of the grouping of the annual and weekly consumption profiles and ii) the ability to predict the total annual consumption and the random variation parameter by household characteristics. Also, the predicted consumption traces are compared with the measured ones.

Stability of annual and weekly profiles

In the Regular Decomposition method the number of groups as well as the annual and weekly profile vectors were fixed to those obtained from the model development data. Then the same classification algorithm was run to group the annual and average week profiles from the validation data.

This grouping with a fixed scheme works well also for the new data; the previously fixed profiles and the averages of profiles among group members are very close to each other. In the four largest annual groups, the fixed schema provides a very good match. Naturally, one should not include groups of insufficient size in population monitoring.

Figure 14 illustrates the largest difference in weekly profiles. The differences in profiles are associated with the group size and hence with the averaging over member profiles. Since even the profile pair with the largest difference captures well the essential consumption pattern, the authors conclude that grouping the unseen validation data with fixed weekly profile function works well and allows to compare customer populations by recording the frequencies of profiles in the population.

In this validation data, the annual and weekly groups are independent of each other (chi square independence test value 0.13).

Grouping with fixed vs. free profiles

What results if only the number of annual and weekly clusters is fixed, and the cluster profiles are let to be optimal for the test data? This kind of analysis provides information on the goodness of grouping with fixed profiles. Firstly, one needs to verify that the group profiles resulting from optimization are close to the fixed profiles. It is also interesting how the consumers form the groups. This question is examined with the weekly grouping, where all the groups have substantial sizes.

It turns out that 64% of the test data is grouped so that there is a very close profile from the fixed development data group profile set. Overall, the new group average profiles are quite similar to the fixed profiles (although less smooth due to the smaller number of samples in the averaging). However, it is not easy to identify a mapping to the whole data set that takes a grouping with fixed profiles to the grouping with free profiles. An interesting observation is that the consumer groups do not remain unchanged when the profiles are let to be free. The variation in households' individual average week profiles is still large and hence the memberships of the groups are not always obvious. However, the resulting group average profiles are rather stable. Thus, one should not follow the group membership labels of individual consumers in time, but what kind of groups the consumers form. 78% of the validation data is covered by the five largest groups, and it is rather easy to find a mapping between the fixed group mean profiles (from the model development data) and the new group mean profiles (from the validation data). The closest profile can be chosen unequivocally in four cases, and the remaining one has a few rather close profile candidates. The best profile matches are illustrated in Figure 15.

Prediction of the annual consumption and the random variation

In this section, the total annual consumption W and the random variation parameter V are estimated by conditioning each on the household size P, the home floor area F, and the number of bedrooms B. The estimators are the conditional expectation of W given (P, F, B) and that of V given (P, F, B), derived in Kuusela et al. (2015). The conditional distribution of W given (P, F, B), denoted as $W|(P, F, B)$, is lognormally distributed, and similarly for V. Formulas for the expectations and the variances of $W|(P, F, B)$ and $V|(P, F, B)$ can be written by equations (2) and (3) of Kuusela et al. (2015). It turns out that the conditional expected value cannot predict the target variables accurately at the household level. This is due to the large variability of households: the estimator is the expected value of the conditional consumption. Instead, the models can reproduce a similar random variation in the target values as that existing in the test population (see the discussion in Kuusela et al. (2015)). However, the selection of consumers for this paper results to a worse model than the one analyzed more deeply in Kuusela et al. (2015). For each validation observation (P, F, B, W), the conditional distribution $W|(P, F, B)$ and its 95% confidence interval was formed. In this validation sample, 18% of the W values were outside of the 95% confidence intervals compared to less than 5% in a population of the same Irish data utilized in Kuusela et al. (2015). Note that the conditional distribution is a function of (P, F, B) so that the confidence intervals are also functions of these variables.

The random variation component is studied with the group profiles obtained by fixing the annual and weekly profiles to those obtained from the model development data. In less than 4% of the observed test data, the random parameter values are outside the 95% confidence interval of the random value parameter estimator. The pair of curves in Figure 16 illustrates the distributions of the observed random variation parameter values and the corresponding model-generated values obtained by picking 50 samples from the conditional distribution given the household characteristics of each validation data consumer, i.e., $V|(P, F, R)$. The observed random variation parameter values tend to be larger than the ones generated by the developed model, although the difference is not huge. However, it will be visible in the model-generated consumption traces shown in Figure 17, where the model tends to predict a smaller random variation than the observed variation. One possible reason could be that by the grouping with predefined annual and weekly profiles, the random component includes an impact of the non-optimal group profiles in addition to the pure random variation. Thus, the most realistic random variation scheme should use the households' individual profiles.

CONCLUSIONS

This work contributed to the field of electricity consumption modeling and monitoring by analyzing a multiplicative modeling framework consisting of i) total annual consumption, ii) annual consumption profile, iii) average weekly consumption profile, and iv) random variation around the repeated mean consumption profiles. The variation of consumption is a natural element in the model and very easy to monitor in this framework. This modeling intuition stemmed from the lognormality of the electricity consumption. Section 3 showed that the model was able to sufficiently capture the amount of random variation around the repeated consumption patterns, and the generated consumption traces accurately reproduce the minimum and the median of a consumer's weekly consumption maxima. However, the random variation model would benefit from further tuning at low and, in particular, at peak consumption levels.

Then the interest was turned towards monitoring a population of electricity consumers and the properties of the proposed monitoring parameters. For that purpose, the recently developed Regular Decomposition method was utilized to group the annual and weekly profiles. It turned out that the monitoring parameters were essentially independent from each other. The validation showed good stability of the groups. The authors propose to direct research interest towards the random variation around regular patterns as the amount of randomness exceeds small differences in profiles. The grouping of profiles would benefit from efficient methods to handle dynamic large data.

The data provide an opportunity to model the households' total electricity consumption as household energy systems were rare in Ireland during the trial period. When the households' energy production and smart energy systems will become common, it will be very difficult to assess the actual energy consumption of a household, as the energy companies only see the amount of energy required to meet the total consumption. The rapid evolution in household energy equipment and the offered energy products as well as the tariffs also have an impact on the data collected by the energy companies, and the modeling of households' total consumption will become increasingly difficult.

The developed household consumption model offers a relatively simple method to simulate the stochastic variation of electricity consumption to populate network models or to design new architectural setups, algorithms, and decision support tools to utilize distributed energy resources in meeting the demands.

References

Archive, I. S. S. D. (2012). "CER Smart Metering Project. `http://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

Beckel, C., Sadamori, L., Staake, T., and Santini, S. (2014). "Revealing household characteristics from smart meter data." *Energy*, 78, 397 – 410.

Chen, C. and Cook, D. J. (2012). "Behavior-based home energy prediction." *Intelligent Environments (IE), 2012 8th International Conference on*, IEEE, 57–63.

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

Gouveia, J. P. and Seixas, J. (2016). "Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys." *Energy and Buildings*, 116, 666–676.

Grandjean, A., Adnot, J., and Binet, G. (2012). "A review and an analysis of the residential electric load curve models." *Renewable and Sustainable Energy Reviews*, 16(9), 6539–6565.

Grünwald, P. (2007). *Minimum Description Length Principle*. The MIT Press.

Haben, S., Rove, M., Greethm, D., Grindord, P., Holderbaum, W., Potter, B., and Singleton, C. (2013). "Mathematical solutions for electricity networks in a low carbon future." *22nd International Conference on Electricity Distribution (CIRED), Stockholm, 10-13 June 2013*.

Haben, S., Singleton, C., and Grindrod, P. (2016). "Analysis and clustering of residential customers energy behavioral demand using smart meter data." *IEEE Transactions on Smart Grid*, 7(1), 136–144.

Jones, R. V., Fuertes, A., and Lomas, K. J. (2015). "The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings." *Renewable and Sustainable Energy Reviews*, 43, 901 – 917.

Kolter, J. Z. and Ferrera, J. J. (2011). "A large-scale study on predicting and contextualizing building energy usage." *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 7-11 August 2011, San Francisco, California, USA. AAAI Press, 2011.*

Kuusela, P., Norros, I., Weiss, R., and Sorasalmi, T. (2015). "Practical lognormal framework for household energy consumption modeling." *Energy and Buildings*, 108, 223–235.

Kwac, J., Flora, J., and Rajagopal, R. (2014). "Household energy consumption segmentation using hourly data." *Smart Grid, IEEE Transactions on*, 5(1), 420–430.

Kwac, J., Flora, J., and Rajagopal, R. (2016). "Lifestyle segmentation based on energy consumption data." *IEEE Transactions on Smart Grids*, PP(99).

Kwac, J., Tan, C.-W., Sintov, N., Flora, J., and Rajagopal, R. (2013). "Utility customer segmentation based on smart meter data: Empirical study." *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, 720–725 (Oct).

Limbert, E., Stahel, W., and Abbt, M. (2001). "Log-normal distributions across the sciences: keys and clues." *BioScience*, 51(5), 341 – 352.

McLoughlin, F., Duffy, A., and Conlon, M. (2012). "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study." *Energy and Buildings*, 48, 240–248.

McLoughlin, F., Duffy, A., and Conlon, M. (2015). "A clustering approach to domestic electricity load profile characterisation using smart metering data." *Applied Energy*, 141, 190 – 199.

Mutanen, T., Sorasalmi, T., and Weiss, R. (2012). "Electricity usage type selection and model validation based on electricity usage measurements." *Industrial Conference on Data Mining 2012. IbaI Publishing (2012)* (July).

Pelillo, M., Elezi, I., and Fiorucci, M. (2016). "Revealing structure of large graphs: Szemerédi's regularity lemma and its use in pattern recognition." *Pattern Recognition Letters*.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., and Kolehmainen, M. (2010). "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.

Reittu, H., Bazsó, F., and Norros, I. (2017). "Regular Decomposition: an information and graph theoretic approach to stochastic block models arXiv:1704.07114.

Reittu, H., Weiss, R., and Bazso, F. (2014). "Regular decomposition of multivariate time series and other matrices." *Structural, Syntactic, and Statistical Pattern Recognition*, P. Fränti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, eds., number 8621 in LNCS, Berlin-Heidelberg, Springer-Verlag, 424–433.

Wang, Y., Chen, Q., Kang, C., Xia, Q., and Luo, M. (2017). "Sparse and redundant representation-based smart meter data compression and pattern extraction." *IEEE Transactions in Power Systems*, 32(3).

List of Figures

1 Histogram of the total consumption distribution (blue) and its fits with a lognormal distribution (red) and a Weibull distribution (green), inset shows the tail probability fit in $\log_{10}$ scale. 27

2 Kolmogorov-Smirnov distances (maximum deviation between distribution functions) between each household's random variation distribution and a fitted lognormal distribution. 28

3 Histogram of the log-variances of the random variation processes of all customers, and its fit with a lognormal distribution (red). 29

4 (a): The process $\log \xi_t(i)$ of household 29 over a whole year. (b): The empirical variance of $(\log \xi_t(i))_{i=1}^{52}$ for $t \in \{1, \ldots, 336\}$ (blue curve), and the mean week profile of household 29 (red curve). 30

5 The numbers $\mathrm{Var}\,(\log \lambda)$ (black, thick), $\mathrm{Var}\,(\log a))$ (blue), and $\mathrm{Var}\,(\log a) + \mathrm{Var}\,(\log \xi))$ (red) for each household, plotted in the order of increasing $\mathrm{Var}\,(\log \lambda)$. 31

6 The correlations between $\log a$ and $\log \xi$ for each household, plotted in the order of increasing $\mathrm{Var}\,(\log \lambda)$. 32

7 The average, over all households, of the logarithmic autocorrelations of the random variation processes for half-hour lags $1, \ldots, 96$. 33

8 Variances of the normalized consumption processes of all customers, presented as a cumulative plot. Blue: the true values. Green: synthetic traces with i.i.d. random variation. Red: synthetic traces with correlated random variation. Customers with numbers 758-995 use electrical heating. 34

9 The smallest, the largest, and the median values of weekly maxima of the normalized consumption processes of all customers, presented as a cumulative plot. Colors and indexing as in Fig. 8. 35

10 Annual consumption profiles `A`. . . `F` of the model development data with group size in the parenthesis. 36

11 Group mean profiles and sizes in the grouping of weekly average profiles in the model development data. 37
12 Comparison of observed,(a), and two modeled, (b) and (d), traces as well as the non-random components of the two models, (c) and (e). The non-random components consist of annual and weekly profiles. The figures (b) and (c) utilize individual profiles of the selected customer whereas figures (d) and (e) illustrate results by utilizing the group mean annual and weekly profiles. . . 38
13 Marginals of multivariate lognormal modeling. Figure (a) illustrates the annual consumption, W, in MWh and (b) the variance around consumptions patterns, V. These are the target variables. The remaining figures illustrate the household characteristics persons in household, P in figure (c), home area in m^2, F in (d), and the number of bedrooms, B in (e). These household characteristics are related to each of the target variables in turn. 39
14 Illustration of the largest difference between the fixed week profile (blue) and the mean of group profile (orange) in the validation data set. This group had 16 members. 40
15 The best profile matches between weekly profiles of the model development data (blue) and new group mean profiles (red) from free classification of households' average weeks into nine groups in the validation data. 41
16 The distributions of the observed (blue) and model generated (red) random variation parameter values in the validation data. 42
17 An example of comparison of consumption traces in the validation data. Profile figure illustrates the non-random component from consumer's annual and weekly classification profiles. 43

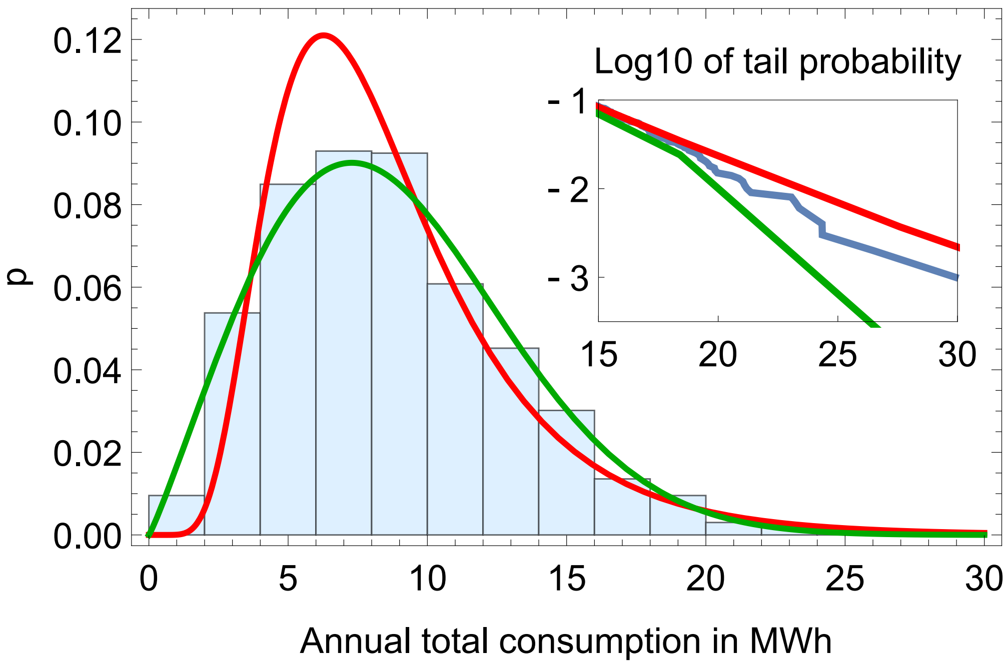

Fig. 1. Histogram of the total consumption distribution (blue) and its fits with a lognormal distribution (red) and a Weibull distribution (green), inset shows the tail probability fit in $\log_{10}$ scale.

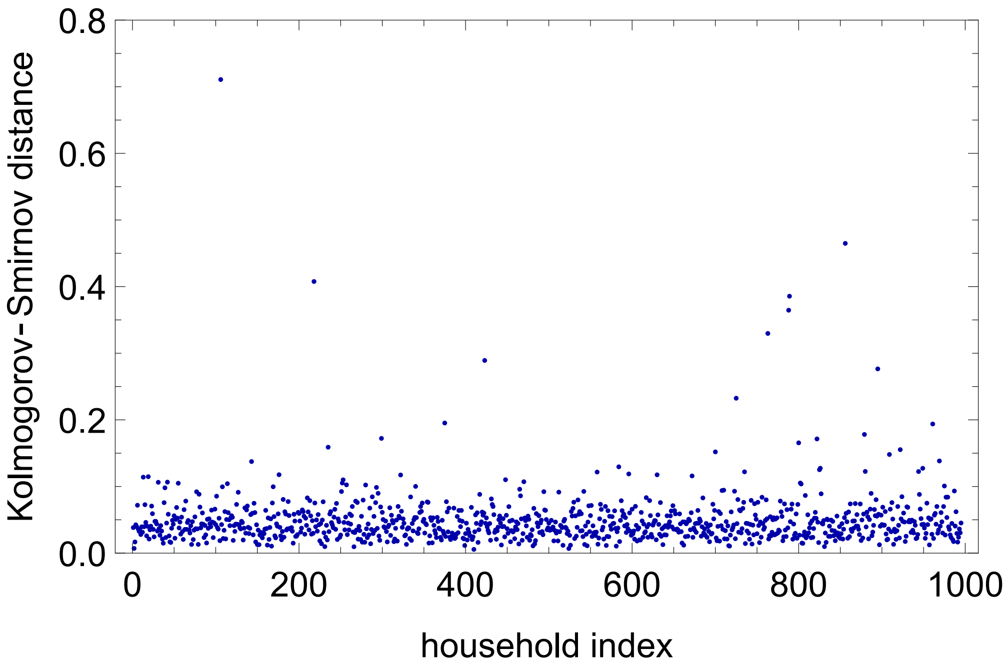

Fig. 2. Kolmogorov-Smirnov distances (maximum deviation between distribution functions) between each household's random variation distribution and a fitted log-normal distribution.

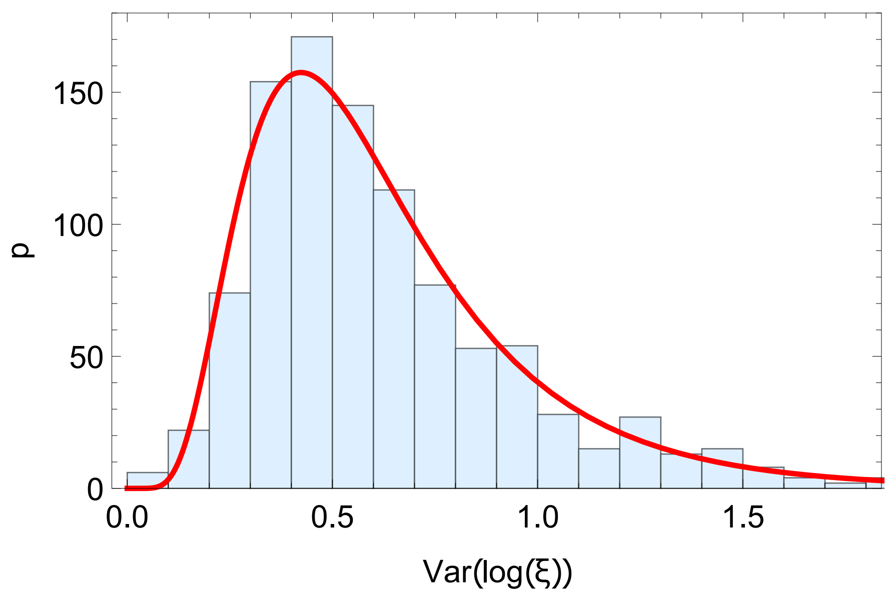

Fig. 3. Histogram of the log-variances of the random variation processes of all customers, and its fit with a lognormal distribution (red).

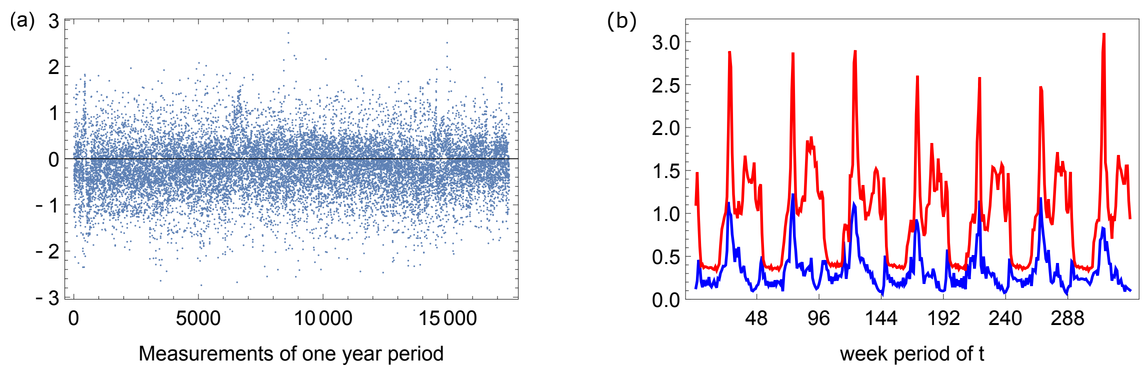

Fig. 4. (a): The process $\log \xi_t(i)$ of household 29 over a whole year. (b): The empirical variance of $(\log \xi_t(i))_{i=1}^{52}$ for $t \in \{1, \ldots, 336\}$ (blue curve), and the mean week profile of household 29 (red curve).

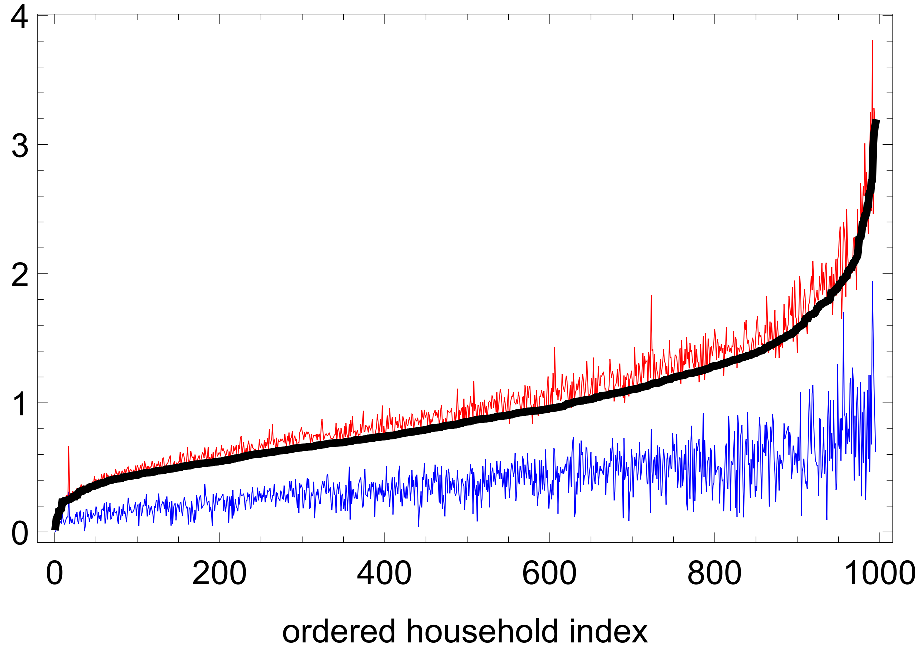

Fig. 5. The numbers $\mathrm{Var}\,(\log\lambda)$ **(black, thick),** $\mathrm{Var}\,(\log a))$ **(blue), and** $\mathrm{Var}\,(\log a) + \mathrm{Var}\,(\log\xi))$ **(red) for each household, plotted in the order of increasing** $\mathrm{Var}\,(\log\lambda)$**.**

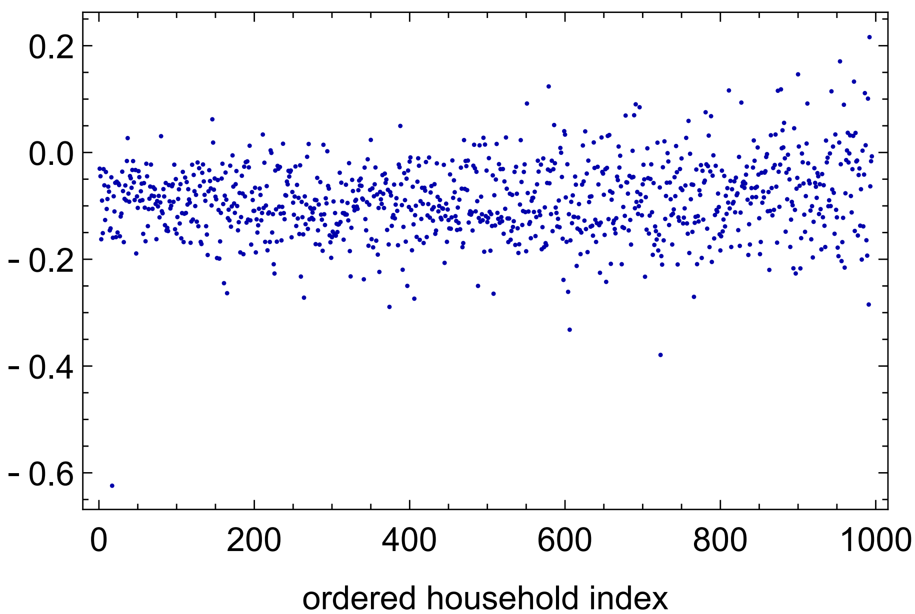

Fig. 6. The correlations between $\log a$ **and** $\log \xi$ **for each household, plotted in the order of increasing** $\mathrm{Var}\,(\log \lambda)$.

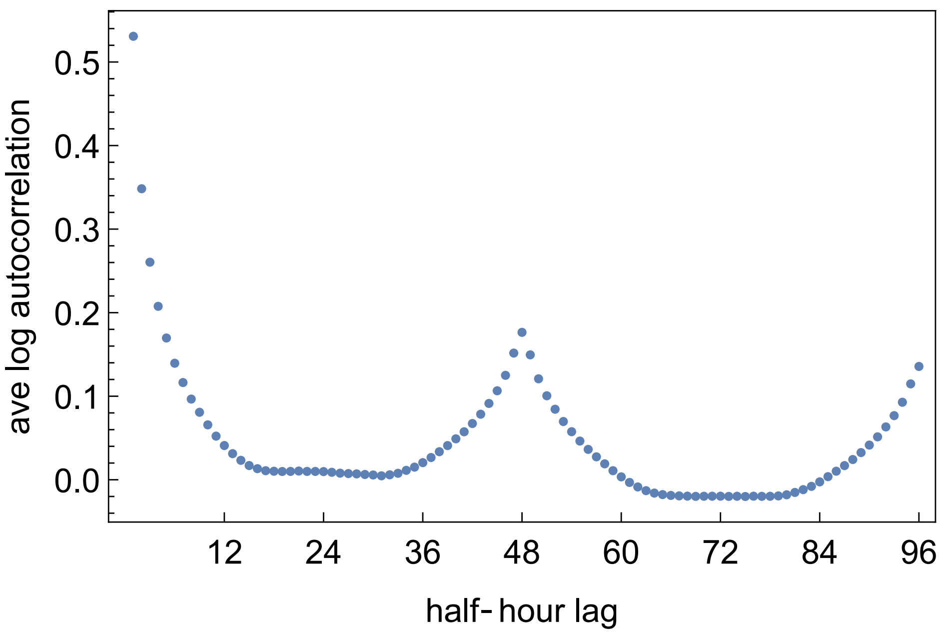

Fig. 7. The average, over all households, of the logarithmic autocorrelations of the random variation processes for half-hour lags $1, \ldots, 96$.

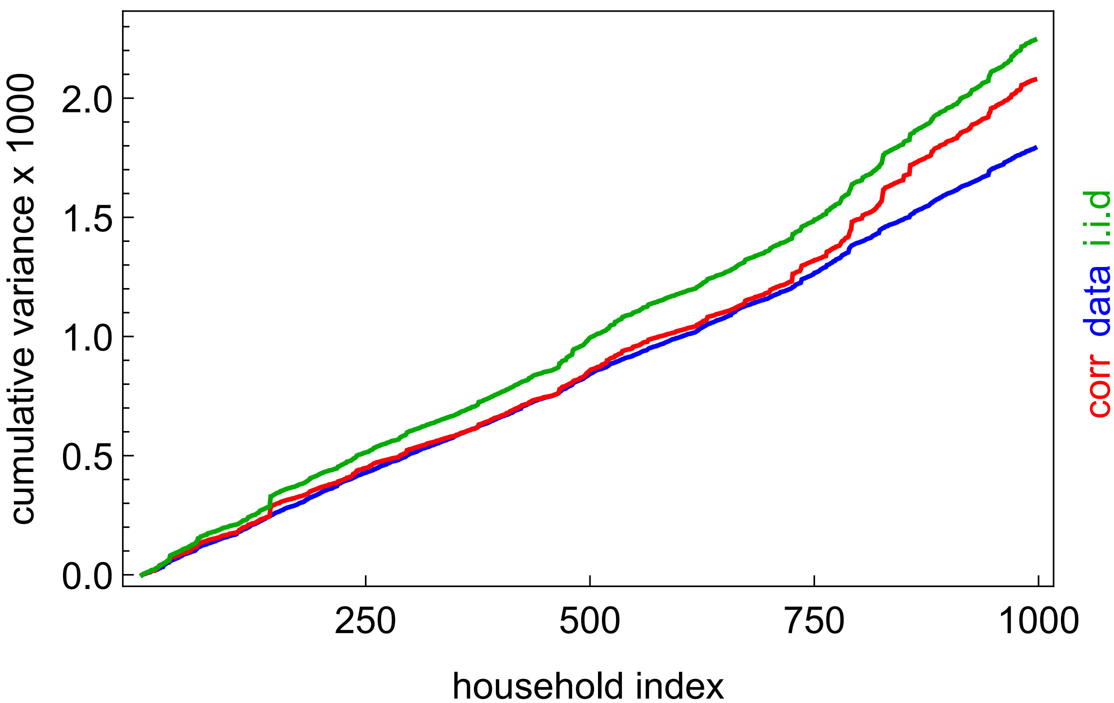

Fig. 8. Variances of the normalized consumption processes of all customers, presented as a cumulative plot. Blue: the true values. Green: synthetic traces with i.i.d. random variation. Red: synthetic traces with correlated random variation. Customers with numbers 758-995 use electrical heating.

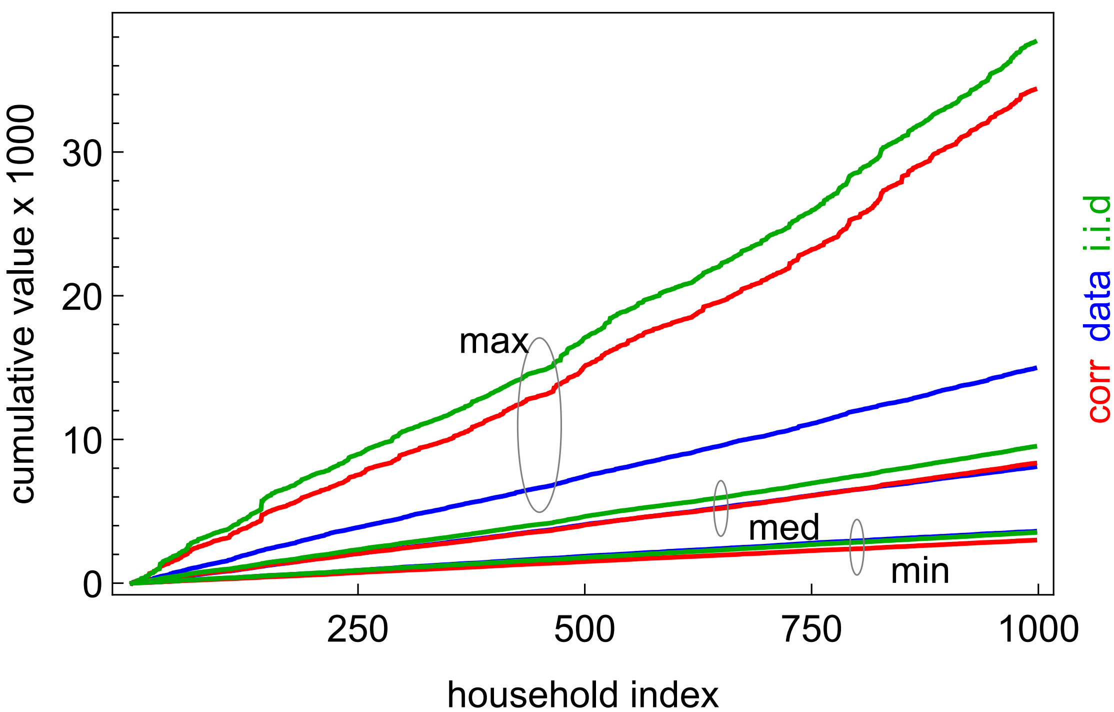

Fig. 9. The smallest, the largest, and the median values of weekly maxima of the normalized consumption processes of all customers, presented as a cumulative plot. Colors and indexing as in Fig. 8.

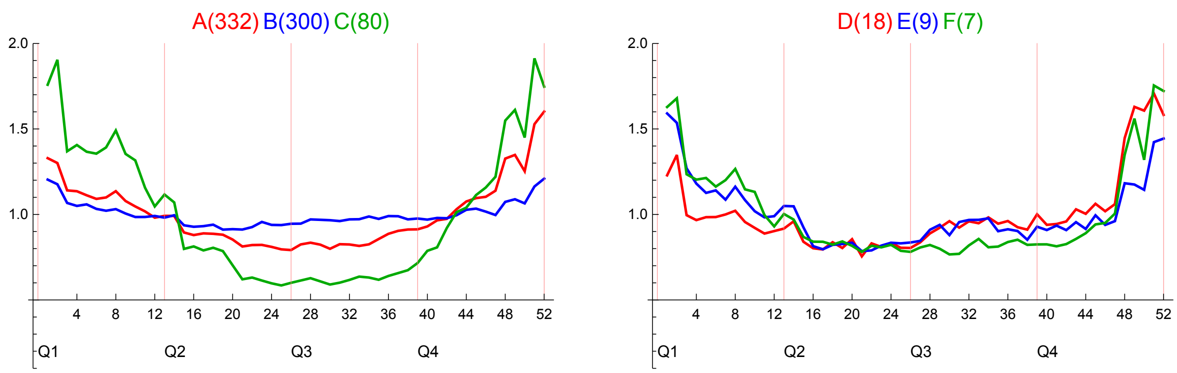

Fig. 10. Annual consumption profiles `A... F` **of the model development data with group size in the parenthesis.**

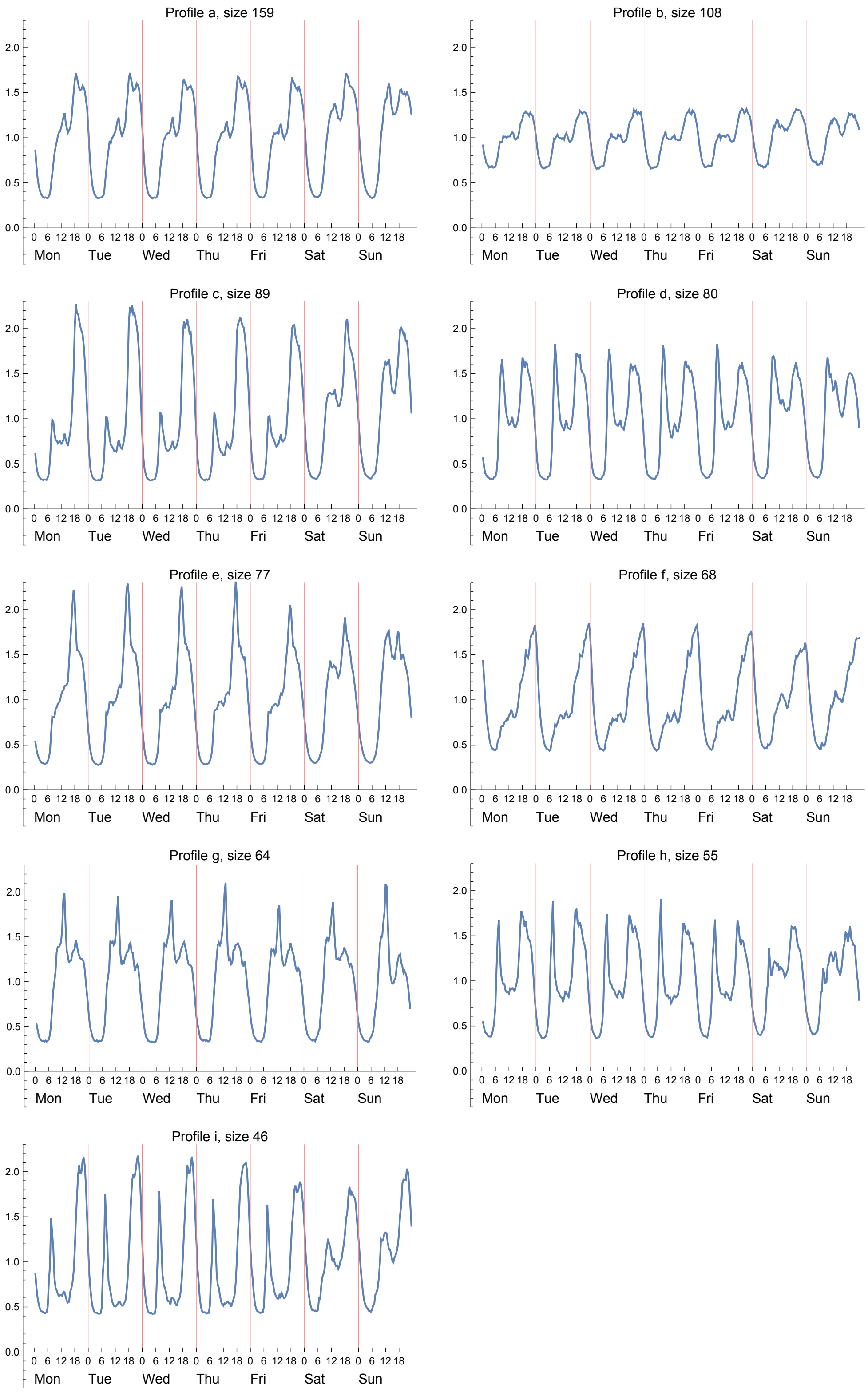

Fig. 11. Group mean profiles and sizes in the grouping of weekly average profiles in the model development data.

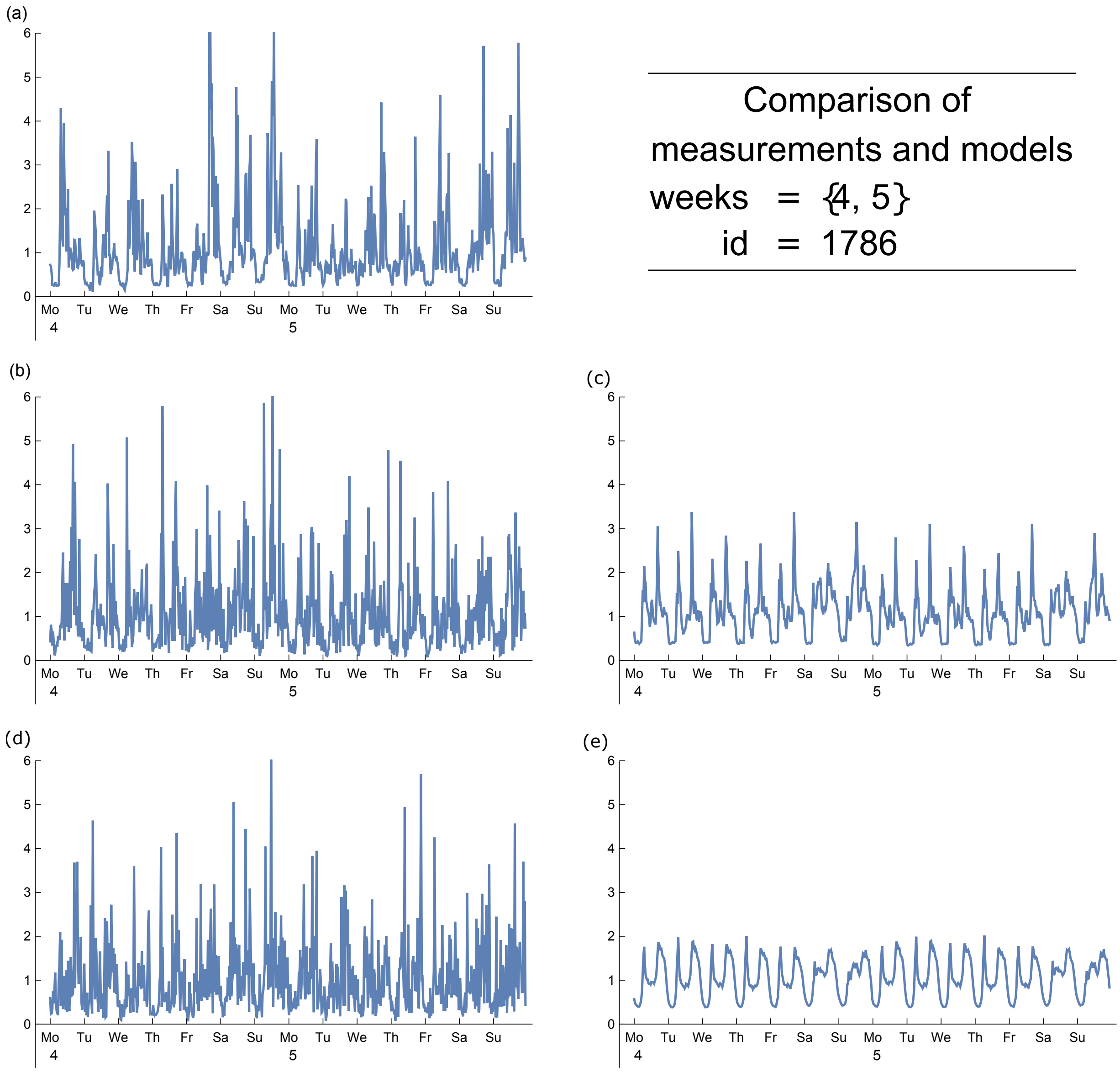

Fig. 12. Comparison of observed,(a), and two modeled, (b) and (d), traces as well as the non-random components of the two models, (c) and (e). The non-random components consist of annual and weekly profiles. The figures (b) and (c) utilize individual profiles of the selected customer whereas figures (d) and (e) illustrate results by utilizing the group mean annual and weekly profiles.

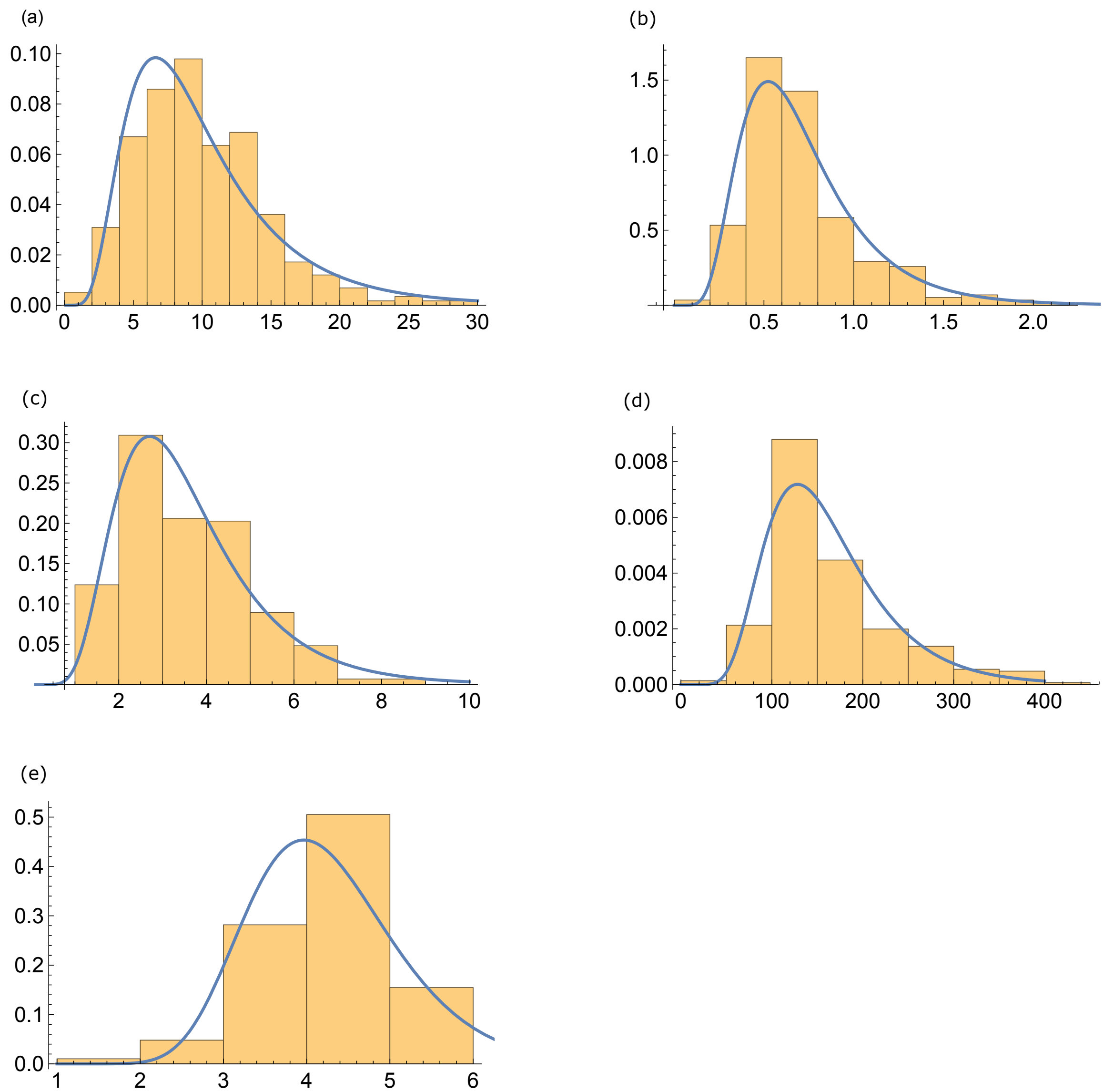

Fig. 13. Marginals of multivariate lognormal modeling. Figure (a) illustrates the annual consumption, W, in MWh and (b) the variance around consumptions patterns, V. These are the target variables. The remaining figures illustrate the household characteristics persons in household, P in figure (c), home area in m^2, F in (d), and the number of bedrooms, B in (e). These household characteristics are related to each of the target variables in turn.

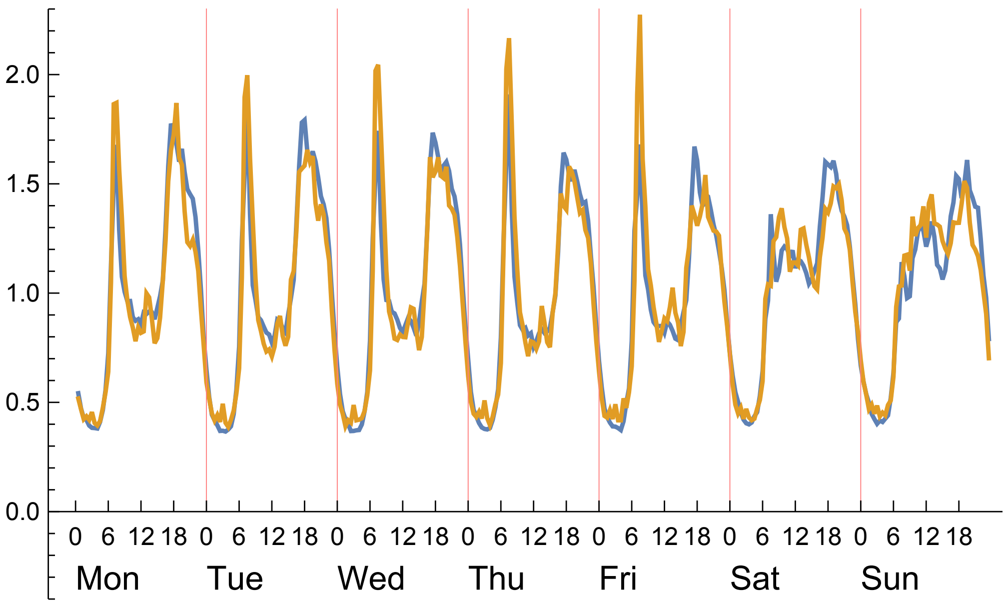

Fig. 14. Illustration of the largest difference between the fixed week profile (blue) and the mean of group profile (orange) in the validation data set. This group had 16 members.

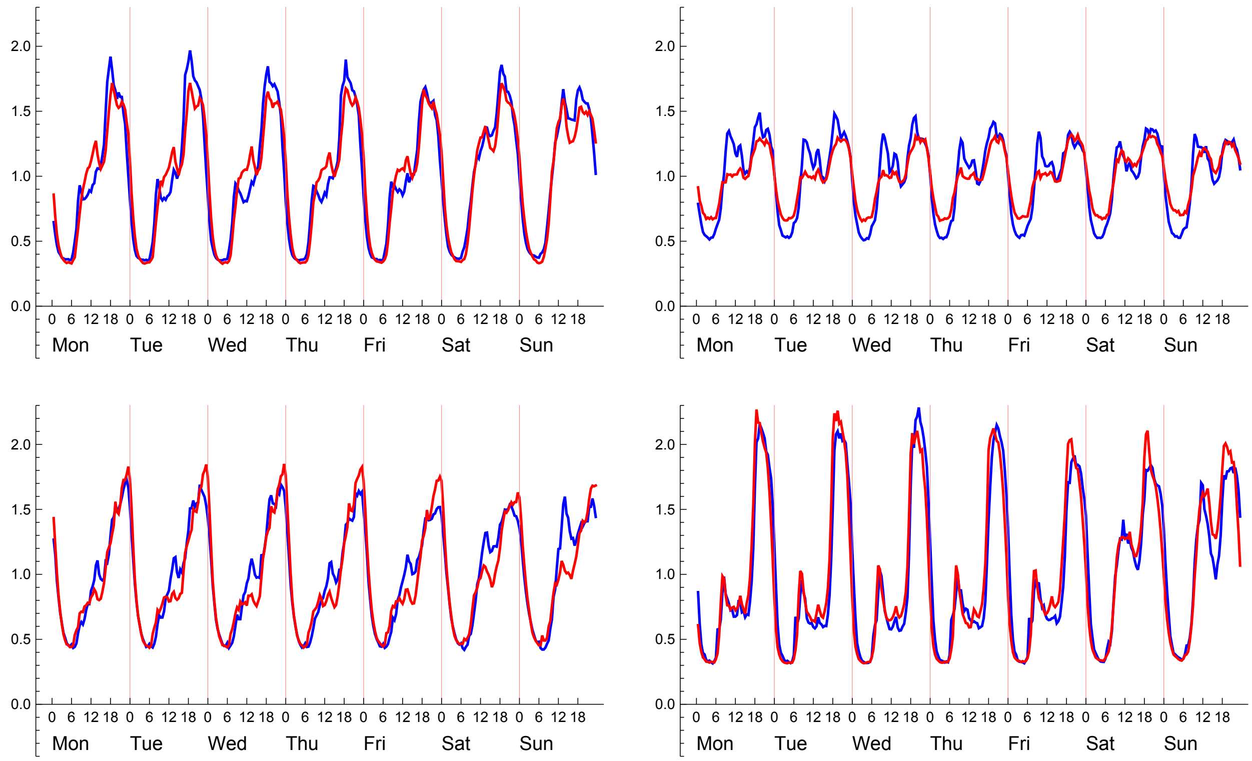

Fig. 15. The best profile matches between weekly profiles of the model development data (blue) and new group mean profiles (red) from free classification of households' average weeks into nine groups in the validation data.

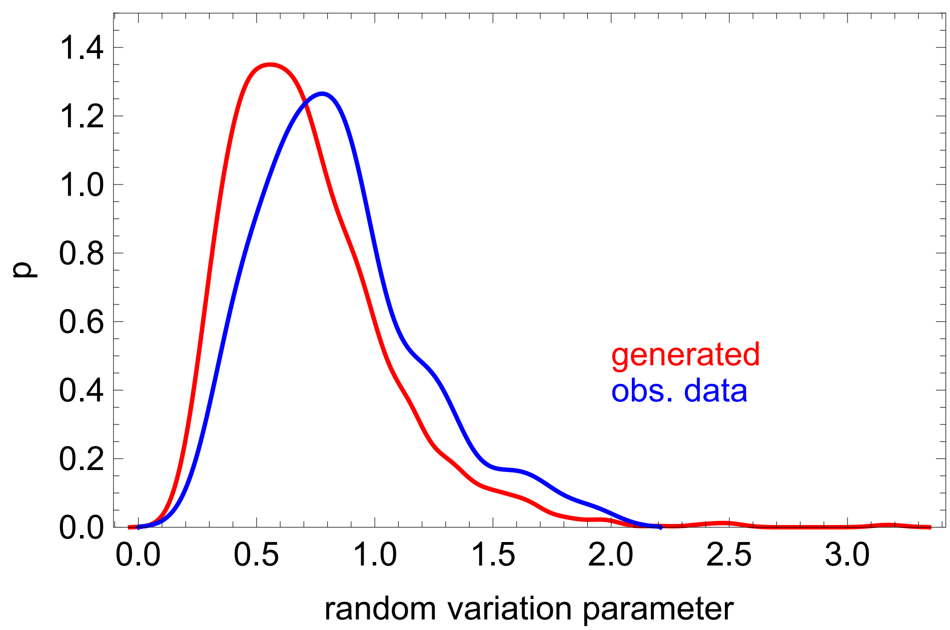

Fig. 16. The distributions of the observed (blue) and model generated (red) random variation parameter values in the validation data.

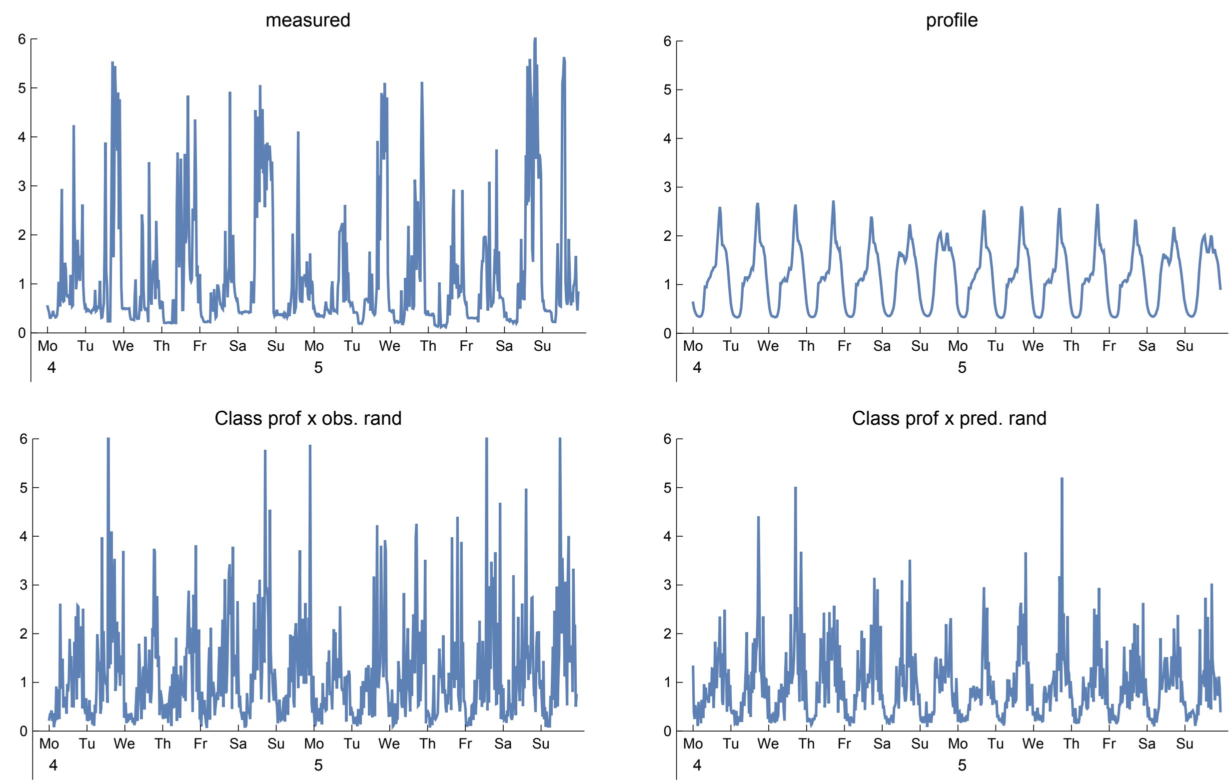

Fig. 17. An example of comparison of consumption traces in the validation data. Profile figure illustrates the non-random component from consumer's annual and weekly classification profiles.