Energy Conversion and Management 92 (2015) 507–516

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Clustering disaggregated load profiles using a Dirichlet process mixture model

Ramon Granell [a], Colin J. Axon [b,*], David C.H. Wallom [a,1]

[a] Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, United Kingdom
[b] Institute of Energy Futures, Brunel University, Uxbridge, London UB8 3PH, United Kingdom

ARTICLE INFO

Article history:
Received 12 August 2014
Accepted 24 December 2014

Keywords:
Bayesian statistics
Classification algorithms
Data mining
Energy use
Power demand
Smart grids

ABSTRACT

The increasing availability of substantial quantities of power-use data in both the residential and commercial sectors raises the possibility of mining the data to the advantage of both consumers and network operations. We present a Bayesian non-parametric model to cluster load profiles from households and business premises. Evaluators show that our model performs as well as other popular clustering methods, but unlike most other methods it does not require the number of clusters to be predetermined by the user. We used the so-called 'Chinese restaurant process' method to solve the model, making use of the Dirichlet-multinomial distribution. The number of clusters grew logarithmically with the quantity of data, making the technique suitable for scaling to large data sets. We were able to show that the model could distinguish features such as the nationality, household size, and type of dwelling between the cluster memberships.
© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

There are supply and demand-side drivers to better understand power-use patterns to help deliver robust electricity distribution networks. The introduction of advanced metering raises the possibility of exploiting increasing volumes of data with potential benefits (and disadvantages) to customers, retailers, and network operators. Measures to implement demand-side management are becoming technically feasible in both commercial and residential premises, and innovation in deregulated markets is arising from changing customer expectations and usage patterns. Increasing embedded and multi-scale generation, decreasing reliance on base-load generation, the electrification of transport and heating alongside increasing use of cooling loads are factors introducing uncertainty into network management. The ability to recognise types of customer load and to differentiate between them will become an important tool for the design of tariffs to incentivize load-shifting or other changes in consumption patterns, fault recognition and detection, and planning. The clustering of time-series power-use data may provide a useful tool in these respects.

In general, clustering techniques are unsupervised machine learning algorithms to determine the subsets into which a data set can be divided without a priori information. The objective is to detect the data elements that are similar and to ensure that the elements of these clusters are all different to the elements of other clusters. An overview of many of the aspects of the clustering of electricity profiles and a review of the techniques has been made by [1]. Previous work adopted the frequentist approach and used well-known clustering algorithms such as k-means and hierarchical algorithms [2–5], fuzzy k-means [6], a Support Vector Clustering model [7], or iterative self-organised maps [8,5]. In [9], they simply group the data employing environmental characteristic such as months of the year. The disadvantage of these techniques is the need to declare the number of clusters before beginning computation. In [10], the follow-the-leader algorithm was used to cluster load profiles where instead of defining the number of clusters, a distance threshold among the clusters is given by the user. Other approaches segment electricity profiles based on a priori known customer features such as the commercial sector, without applying any clustering method [11].

Our approach uses Bayesian statistics to enable the modelling of unknown parameters that govern the distribution used for explaining the data (which has a distribution itself). Using a distribution of these parameters gives greater flexibility and robustness for managing the uncertainty that data present. In non-parametric algorithms, the number of parameters is not previously established, there being potentially infinite parameters. When clustering with a Bayesian non-parametric method, the number of the resulting

* Corresponding author. Tel.: +44 (0)1895 267932.
E-mail addresses: ramon.granell@oerc.ox.ac.uk (R. Granell), colin.axon@brunel.ac.uk (C.J. Axon), david.wallom@oerc.ox.ac.uk (D.C.H. Wallom).
[1] Tel.: +44 (0)1865 610601.

http://dx.doi.org/10.1016/j.enconman.2014.12.080
0196-8904/© 2015 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Nomenclature

Abbreviations and acronyms

Dir()	Dirichlet distribution
Mult()	multinomial distribution
pdf	probability density function
ANOVA	analysis of variance
CRP	Chinese restaurant process
DP	Dirichlet process
DPMM	Dirichlet process mixture model
MIA	mean index adequacy
PGMA	pair group method average algorithm
PGMC	pair group method centroid algorithm
SI	scatter index
VRC	the variance ratio criterion

Symbols

c_{ij}	jth counter of the ith load profile x_i
d	dimension of the load profiles
i, j, k, l, m	dummy variables
$k^{(I)}$	number of cluster in the Ith iteration of the Gibbs sampling algorithm
m_{ij}	mean of the ith cluster at the jth dimension
n	number of load profiles
n_{kj}	summation of the jth counter of all the data points in kth cluster
t_i	ith component of the prior mean of the Dirichlet-multinomial distribution
z_i	index for ith load profile indicating the cluster or mixture component of the DPMM to which it is assigned
x_i	d-dimensional ith load profile obtained from a smart meter of a house
C_i	summation of all the d counters of object x_i
G	DP-distributed random probability measure
G_0	base probability measure of a DP
I	number of iterations in the Gibbs sampling algorithm for computing the posterior distribution in the DPMM
K	number of clusters
X	set of all the n load profiles to cluster
T	number of iterations in the Gibbs sampling algorithm for estimating α_0
Z	set of all the n cluster indices for the load profiles
α	significance level for T-test and F-test
α_0	precision or scaling parameter of a DP
β	d-dimensional concentration parameter of a Dirichlet distribution
χ	temporal complexity of computing the reallocation probability for one data point on a cluster
π_i	probability of the ith component in Π
ρ	strength of the prior of the Dirichlet-multinomial distribution
θ_{ij}	jth dimensional value of Θ_i
$\Gamma()$	gamma function
Π	$K+1$ probabilities of a multinomial distribution that correspond to mixture components or cluster in the DPMM
Θ_i	d-dimensional parameter of a multinomial distribution governed by a Dirichlet distribution that corresponds to the ith cluster

clusters is determined by the model and the data, it is not fixed by the user. The Bayesian model that we use is the Dirichlet process mixture model (DPMM). These models have been used successfully for solving clustering tasks in diverse areas such as computational biology [12], computational linguistics [13] or marketing [14]. In [15] they used a Bayesian clustering by dynamics method to cluster electricity use time series for load forecasting. Their method models the data dynamics as Markov chains and then applies an agglomerative clustering procedure (our algorithm is not agglomerative) and employs an entropy-based heuristic search strategy to find the most likely partition, which is not needed in our case.

In this paper, we describe the data sets (Section 2), development of the new model and clustering process (Section 3), its application and the extraction of various features from the data set (Section 4), an evaluation with other common clustering techniques. We then conclude and discuss possible further work in this area.

2. Data sets and pre-processing

Our approach to time-series power-use data exploits two unique data sets. The first is a small high-resolution residential data set (with metadata) produced by a European project.[2] Previous studies have used data of 15–30 min resolution [2,3,7]. We suggest that using a higher data rate should improve the usefulness of clustering of power consumption data. The second data set is larger and comprises 30-min resolution data from commercial users.

2.1. Residential data set

This data set [16] comprises electrical power consumption data collected between 30-03-2010 and 24-11-2010 from 135 British and 84 Bulgarian dwellings (a total of 219). The meters were recording at between 6 and 8 s resolution. The metadata (Table 1) was: nationality [UK or Bulgaria], number of occupants, number of bedrooms, and the type of dwelling. There are five categories of dwelling: flat or apartment, terraced (a house that is situated in a row of houses sharing side walls with neighbouring properties), semi-detached house (a houses that only shares a single common wall with another house or property), detached house (a dwelling that does not share any walls with any other structure), and other kind of dwelling.

Two pre-process filters removed anomalous readings and meter faults. First, the negative and zero values, and secondly, dwellings whose readings had five or fewer different values for at least half of the total readings. The clean data was transformed to one minute resolution by averaging the number of readings present within each minute. The daily load profile corresponds to the averaged data with minute resolution during a day aggregating all working days of a specific dwelling. Only load profiles that present values for at least 1438 of the 1440 min were used subsequently, with a total of 197 dwellings satisfying all these criteria (125 British: 72 Bulgarian).

2.2. Commercial data set

This data set consists of half-hourly electricity use for 1877 UK business from the entertainment sector during 2009 and 2010.[3] These businesses are categorised as restaurants/cafes, hotels/guest houses, pubs/bars, clubs, and cinemas/other leisure.

The pre-process procedure had four steps. First, negative and zero values were removed. Secondly, for each business, we

[2] www.dehems.eu.

[3] The entertainment sector businesses used in this study are a subset of a commercial data set of 12,000 businesses.

Table 1
Features and categories of the processed data set.

Nationality				
Bulgaria			England	
72 (37%)			125 (63%)	
Number of occupants				
One	Two	Three	Four	Five or more
18 (9%)	49 (25%)	52 (26%)	40 (20%)	38 (19%)
Number of bedrooms				
One	Two	Three	Four	Five or more
29 (15%)	60 (30%)	78 (40%)	20 (10%)	10 (5%)
Type of dwelling				
Flat	Terrace	Semi	Detached	Other
55 (28%)	55 (28%)	54 (27%)	26 (13%)	7 (4%)

removed readings that are over three times the mean plus three times the standard deviation. Thirdly, businesses that did not have a minimum of ten different values in their total readings were removed. Finally, businesses that did not present a minimum of six months of data *i.e.* 8760 readings, were also removed.

After applying these filters created a data set of 1207 businesses with an average of 20,230 readings and a standard deviation of 7706. On average, they present around 60% of the readings during the two years of sampling. Each daily profile has 48 readings.

3. Developing the DPMM

A d-dimensional Dirichlet distribution with concentration parameter $\beta = (\beta_1, \ldots, \beta_d)$ is a continuous distribution that defines a probability measure over the $k-1$-simplex, *i.e.* the domain from the Dirichlet distribution can be seen itself as a d-dimensional discrete distribution.

$$(\theta_1, \ldots, \theta_d) \sim \mathrm{Dir}(\beta) \tag{1}$$

$$\text{where} \quad \theta_i \geqslant 0, \quad \text{for all } i; \text{ and } \prod_{i=1}^{d} \theta_i = 1$$

A Dirichlet process (DP) [17] is a distribution over probability measures that can be seen as an extension of a Dirichlet distribution with infinite dimension. It is composed of two parameters $\mathrm{DP}(G_0, \alpha_0)$, where G_0 is the base probability measure and α_0 is the precision or scaling parameter. Any draw G from the DP ($G \sim \mathrm{DP}(G_0, \alpha_0)$) can be viewed as a discrete distribution with probability of one. Therefore, they can be used as prior probabilities for other discrete distributions or components in a mixture model. There are different representations of the DP such as the stick-breaking construction [17], the Pólya urn scheme [18], and the one used in this work: the Chinese restaurant process (CRP) [19].

The CRP uses the property that if there are $n-1$ independent variables distributed by a probability measure generated by a DP (*i.e.* $\Theta_1, \ldots, \Theta_{n-1} \sim G$ and $G \sim \mathrm{DP}(G_0, \alpha_0)$), the next draw from G *i.e.* Θ_n has a probability that is greater than zero of repeating the value of any of the previous draws [18]. In addition, the draws that appear more times are more likely to appear again than those draws that appear fewer times ('the rich gets richer' effect). These two properties have a clustering effect that can be exploited with a method analogous to allocating spaces in a Chinese restaurant. Imagine a Chinese restaurant with potential infinite number of tables. Consider the data points to cluster as clients of the restaurant, and the clusters as the tables where customers will sit around (assigned to the cluster). The CRP works in the following way, the first client will sit at the first table, and the nth client will sit at:

$$\begin{cases} \text{table } k & \text{with probability } \frac{n_k}{\alpha_0+n-1}, \ 1 \leqslant k \leqslant K \quad (2a) \\ \text{new table } K+1 & \text{with probability } \frac{\alpha_0}{\alpha_0+n-1} \quad (2b) \end{cases} \tag{2}$$

where n_k is the number of customers there are already sitting at the table k, and K is the total number of tables (clusters). Following the analogy, the dishes on the table can be seen as the parameters of the distribution that explains all of the data points in the cluster. Eq. (2b) guarantees that there exists always a small probability to create a new cluster. Once this first allocation of all the customers is carried out, customers can be reallocated between tables. Posterior probabilities of reallocating a customer from one table to another (tables 1 to K) or to a new one ($K+1$) is computed. Computation of these probabilities for our model is described in the following section.

3.1. The model

We used the CRP method to solve the DPMM with a potentially infinite number of components (clusters). We made use of a hierarchical Dirichlet process [20] (the Dirichlet-multinomial distribution). The load profiles to cluster are represented as draws from a Multinomial distribution whose parameters are generated by a Dirichlet distribution. Formally, to cluster n load profiles $X = \{x_1, \ldots, x_n\}$, where each profile x_i is formed by d counters $x_i = (c_{i1}, \ldots, c_{id})$, the DPMM that models our problem can be hierarchically expressed as:

$$_i = (c_{i1}, \ldots, c_{id}) \sim \mathrm{Mult}(\Theta_{z_i}) \ i = 1, \ldots, n \tag{3}$$

$$\Theta_{z_i} = (\theta_{z_i 1}, \ldots, \theta_{z_i d}) \sim \mathrm{Dir}(\beta_1, \ldots, \beta_d) \ 1 \leqslant z_i \leqslant K \tag{4}$$

$$z_i \sim \mathrm{Mult}(\Pi) \ i = 1, \ldots, n \tag{5}$$

$$\Pi = (\pi_1, \ldots, \pi_{K+1}) \sim \mathrm{DP}(G_0, \alpha_0) \tag{6}$$

where the distributions in Eq. (3) model the probability of generating the values of the load profile as a multinomial distribution whose parameters correspond to a cluster with index z_i (*i.e.* $z_i = j$ indicates that ith load profile is assigned to jth cluster). The Dirichlet distribution that generates the parameters of the multinomial distribution of each cluster (prior distribution in Bayesian statistics) is given by Eq. (4), where K is the number of clusters (these K can vary depending on the observed data). The distribution of Eq. (5) models the cluster selection (mixture component) by each of the data points. It corresponds to a draw from one element from a multinomial of parameters Π. These πs are the components probabilities (priors on mixture model) governed by the distribution in Eq. (6). These draws from a DP were calculated using the CRP based on Eqs. (2a) and (2b) where π_{K+1} corresponds to the probability of allocating the data point to a new cluster, *i.e.* creating a new cluster.

In a DPMM the number of clusters obtained grows logarithmically in relation to the number of input data points [21]. This also depends on the distributions parameters.

3.2. Gibbs sampling algorithm in the DPMM

It is not feasible to compute exactly the posterior $\Pr(\Pi, \Theta, Z|X)$ where $\Theta = \{\Theta_1, \ldots, \Theta_K\}$ and $Z = \{z_1, \ldots, z_n\}$. We used a Gibbs sampling algorithm to approach iteratively the probability of reallocating the object x_i to a new cluster $z\prime_i$ [14] (lines 5 to 11 in Fig. 1). The reallocating posterior distribution $(\pi_1, \ldots, \pi_{K+1})$ from Eq. (6) was computed for each iteration and ith data point in the following way:

$$\pi_k = \Pr(z_i = k|Z_{-i}, X) = \frac{1}{B}\frac{n_k}{\alpha_0 + n - 1}\Pr(x_i|z_i = k), \ 1 \leqslant k \leqslant K \tag{7}$$

$$\pi_{K+1} = \Pr(z_i = K+1|Z_{-i}, X) = \frac{1}{B}\frac{\alpha_0}{\alpha_0 + n - 1}\Pr(x_i|z_i = K+1) \tag{8}$$

where k is the index of the cluster in which point x_i is reallocated and $z_i = K+1$ indicates the creation of a new cluster. B is a normalisation factor that guarantees that probabilities sum to one. Z_{-i} are all indices Z but not including z_i. The marginal probabilities

$\Pr(x_i|z_i = k)$ is the likelihood of the data point x_i, given that the cluster to reallocate this point is the k. Computing this marginal is not straightforward [14]. We computed it analytically, integrating the multinomial parameter θ (Eq. (4)) in the following way:

$$\Pr(x_i|z_i = k) = \int_{\Theta_k} \Pr(x_i|\Theta_k)\Pr(\Theta_k)d\Theta_k \tag{9}$$

where $\Pr(x_i|\Theta_k)$ corresponds to the probability mass function of x_i in the multinomial distribution with parameter Θ_k:

$$\Pr(x_i|\Theta_k) = \frac{C_i!}{c_{i1}!,\ldots,c_{id}!}\prod_{j=1}^{d}\theta_{kj}^{c_{ij}} \tag{10}$$

with $C_i = \sum_{j=1}^{d} c_{ij}$. $\Pr(\Theta_k)$ is the probability density function (pdf) of a Dirichlet distribution with parameter β updated with the previous data points in this cluster:

$$\Pr(\Theta_k) = \mathrm{Dir}(\Theta_k|(\beta_1,\ldots,\beta_d) + (n_{k1},\ldots,n_{kd})) \tag{11}$$

$$= \frac{\Gamma(\sum_{j=1}^{d}(\beta_j + n_{kj}))}{\prod_{j=1}^{d}\Gamma(\beta_j + n_{kj})}\prod_{j=1}^{d}\theta_{kj}^{\beta_j + n_{kj} - 1} \tag{12}$$

where n_{kj} is the summation of the jth counter of all the data points in kth cluster but not x_i (*i.e.* $n_{kj} = \sum_{(x_{i'}|z_{i'}=k)\wedge(x_{i'}\neq x_i)} c_{i'j}$). This update of the parameters is possible because a Dirichlet distribution is the conjugate prior for a multinomial distribution. Substituting Eq. (9) with Eqs. (10) and (12), and after removing the constant term that is equal for all clusters, we obtain:

$$\Pr(x_i|z_i = k) = \frac{\Gamma(\sum_{j=1}^{d}(\beta_j + n_{kj}))}{\prod_{j=1}^{d}\Gamma(\beta_j + n_{kj})}\int_{\Theta_k}\prod_{j=1}^{d}\theta_{kj}^{c_{ij}+\beta_j+n_{kj}-1}d\Theta_k \tag{13}$$

Note that integrating any probability distribution over all possible parameter values should be one, therefore for any Dirichlet distribution:

$$\int_{\Theta_k}\frac{\Gamma(\sum_{j=1}^{d}\beta_j)}{\prod_{j=1}^{d}\Gamma(\beta_j)}\prod_{j=1}^{d}\theta_{kj}^{\beta_j-1}d\Theta_k = 1 \tag{14}$$

$$\frac{\Gamma(\sum_{j=1}^{d}\beta_j)}{\prod_{j=1}^{d}\Gamma(\beta_j)}\int_{\Theta_k}\prod_{j=1}^{d}\theta_{kj}^{\beta_j-1}d\Theta_k = 1 \tag{15}$$

$$\int_{\Theta_k}\prod_{j=1}^{d}\theta_{kj}^{\beta_j-1}d\Theta_k = \frac{\prod_{j=1}^{d}\Gamma(\beta_j)}{\Gamma(\sum_{j=1}^{d}\beta_j)} \tag{16}$$

Applying the results of Eq. (16) in the integral of Eq. (13), we obtain:

$$\Pr(x_i|z_i = k) = \frac{\Gamma(\sum_{j=1}^{d}(\beta_j + n_{kj}))}{\prod_{j=1}^{d}\Gamma(\beta_j + n_{kj})}\frac{\prod_{j=1}^{d}\Gamma(c_{ij} + \beta_j + n_{kj})}{\Gamma(\sum_{j=1}^{d}(c_{ij} + \beta_j + n_{kj}))} \tag{17}$$

By taking the numerator of the first term of Eq. (17) and the denominator of the second term of the same equation and applying the property that $x\Gamma(x) = \Gamma(x+1)$, we obtain:

$$\frac{\Gamma(\sum_{j=1}^{d}(\beta_j + n_{kj}))}{\Gamma(\sum_{j=1}^{d}(c_{ij} + \beta_j + n_{kj}))} = \frac{1}{\prod_{l=0}^{(\sum_{m=1}^{d}c_{im})-1}\left(\sum_{j=1}^{d}(\beta_j + n_{kj}) + l\right)} \tag{18}$$

By taking the denominator of the first term of Eq. (17) and the numerator of the second term of the same equation and applying the same property over the Γ function, we obtain:

$$\frac{\prod_{j=1}^{d}\Gamma(c_{ij} + \beta_j + n_{kj})}{\prod_{j=1}^{d}\Gamma(\beta_j + n_{kj})} = \prod_{j=1}^{d}\prod_{l=0}^{c_{ij}-1}(\beta_j + n_{kj} + l) \tag{19}$$

Joining the results from Eqs. (18) and (19), we obtain an expression that does not contain Θs parameter to compute the probability of Eq. (9) and make use only of simple operations:

$$\Pr(x_i|z_i = k) = \frac{\prod_{j=1}^{d}\prod_{l=0}^{c_{ij}-1}(\beta_j + n_{kj} + l)}{\prod_{l=0}^{(\sum_{m=1}^{d}c_{im})-1}\left(\sum_{j=1}^{d}(\beta_j + n_{kj}) + l\right)} \tag{20}$$

Note that in the case of a new cluster (Eq. (8)) the marginal probability is:

$$\Pr(x_i|z_i = K+1) = \frac{\prod_{j=1}^{d}\prod_{l=0}^{c_{ij}-1}(\beta_j + l)}{\prod_{l=0}^{(\sum_{m=1}^{d}c_{im})-1}\left(\sum_{j=1}^{d}(\beta_j) + l\right)} \tag{21}$$

The stop criteria for the iterative Gibbs sampling algorithm (line 5 of Fig. 1) are stopping when a consecutive number of iterations without changes during the reallocations has occurred (*i.e.* values from Z are constant), or when a maximum number of iterations is reached.

3.3. Parameter estimation

There are two unique input parameters of the model: (1) the Dirichlet prior or concentration parameters ($\beta = (\beta_1,\ldots,\beta_d)$ in Eq. (4)) that control the distributions that govern the data points in clusters, and (2) the precision parameter prior in the DPMM (α_0 in Eq. (6)) that controls the number of clusters. This second

1: Initial assignment of data points $X = (x_1,\ldots,x_n)$ to clusters $Z^{(0} = (z_1^{(0},\ldots,z_n^{(0})$ using Eq. (2a) and Eq. (2b)
2: Initial value to α_0
3: **while** Not stop criteria is achieved for α_0 **do**
4: $I = 1$
5: **while** Not stop criteria is achieved for Z **do**
6: **for** $i = 1$ to n **do**
7: Reallocating points x_i into cluster $z_i^{(I}$ based on $z_i^{(I-1}$ using distribution given by Eq. (7) and Eq. (8)
8: **end for**
9: $k^{(I}$ is equal to the number of clusters in $Z^{(I}$
10: $I = I + 1$
11: **end while**
12: α_0 is recomputed using $k^{(0},\ldots,k^{(I}$ (see Eq. (22))
13: **end while**

Fig. 1. The Gibbs sampling algorithms used for clustering load profiles and for estimating α_0.

parameter can be computed with its Maximum Likelihood Estimator and practically estimated using a Gibbs sampling algorithm [22]. This algorithm will also include the sampling algorithm used for the CRP with the DPMM (see Fig. 1). The new value of α_0 in each iteration of this new sampling algorithm was computed by solving:

$$\frac{1}{I}\sum_{i=1}^{I} k^{(i} = \sum_{j=1}^{n} \frac{\alpha_0}{\alpha_0 + j + 1} \tag{22}$$

where I corresponds to the number of iterations of the sampling algorithm used for the CRP (line 10 of Fig. 1) and $k^{(i}$ is the number of clusters in the ith iteration (line 9 of Fig. 1). The stop criteria in this new sampling algorithm (line 3 of Fig. 1) is that either α_0 remains almost constant, or reaches a maximum number of iterations. To solve Eq. (22), the Newton–Raphson method was used.

In addition to the cyclical updates of the precision parameter α_0, we also update the concentration parameter of each cluster so they utilise all members of the cluster. To perform this every 20 iterations, parameter Θ_z for each cluster (see Eq. (4)) is redrawn, but from a new Dirichlet distribution whose parameters are updated taking into account the counters of the data points of the cluster (see Eq. (11)).

The time complexity of the whole process depends on the number of iterations for both the Gibbs sampling algorithms and the cost of reallocating all points in each iteration: $\mathcal{O}(T \cdot I^* \cdot n \cdot k^* \cdot \chi)$ where T is the number of iterations of the sampling algorithm for estimating α_0 (loop starting in line 5 of Fig. 1). I^* is the maximum number of iterations in the sampling algorithm for computing the posterior distribution (loop starting in line 5 of Fig. 1). k^* is the maximum number of clusters in all iterations, since the reallocation probability is computed for each data point and cluster. This number cannot exceed the number of data points n and usually it is low (see Section 4). χ is the time complexity of computing the reallocation probability for one data point on a cluster that is given by Eqs. (20) and (21): $\mathcal{O}(c^* * d^2)$ where c^* is the maximum counter for all data points and dimension. The final time complexity is $\mathcal{O}(T \cdot I^* \cdot n^2 \cdot c^* \cdot d^2)$.

3.3.1. Estimating the concentration parameter

Estimating the concentration (prior parameter) of the Dirichlet-multinomial distribution is an open research problem for which there are two basic approaches. First, *informative prior* in which the parameters are estimated taking into account the data composition *i.e.* which counters are more likely than others, and secondly *non-informative prior* in which no a priori information of the data counts is used.

A non-informative prior approach can simply be the direct assignment of the same constant value to each $\beta_i, 1 \leqslant i \leqslant d$ if it is not know which dimensions should received more weight. Other solutions [23] propose to divide β_i into $\beta_i = \rho \cdot t_i$, where $t_i, 0 < t_i < 1, \sum_{i=0}^{d} t_i = 1$ is the prior mean and ρ is a constant called the strength of the prior information. Then, they give a value to t_i *e.g.* $t_i = 1/d$ and set ρ to a constant such as $\rho = 1, \rho = d/2$ or $\rho = 1/d$.

In the case of informative prior estimation, [24] proposed a maximum-likelihood approach, employing fixed-point and Newton methods. However, a common approach [12,25] is to divide again $\beta_i = \rho \cdot t_i$ computing the prior mean $t_i, 1 \leqslant i \leqslant d$ as the sample mean:

$$t_i = \frac{\sum_{j=1}^{n} c_{ji}}{\sum_{l=1}^{d}\sum_{j=1}^{n} c_{jl}} \tag{23}$$

and estimating ρ in a similar way as non-informative prior or from a draw from a new finite mixture distribution. In [12], they combined an exponential and uniform density, estimating ρ with a Metropolis sampling algorithm. The estimation can be performed before starting the Gibbs sampling algorithm shown in Fig. 1.

Whether using the informative or non-informative prior, [23] recommended that any prior selection should be reduced by the data dimension. We test different informative and non-informative approaches in the experimental section.

4. Experiments

The DPMM model was implemented in C++ with all experiments performed over the processed data set. The clusters obtained with the DPMM algorithm were compared with other well-known algorithms using various validity evaluators. The experiments were performed using an Intel Core2 Quad CPU Q9650 at 3.00 GHz with 4 Gb of memory.

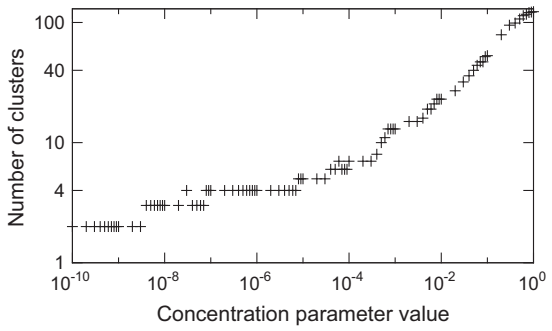
Each input data point is represented as an array of d counters $(x_i = (c_{i1}, \ldots, c_{id}))$ that correspond to a draw from a multinomial distribution. To obtain these counters, each of the averaged values of the load profiles is normalised and transformed into its closest integer value.

Firstly, we used non-informative priors, performing a scanning process to check sensitivity to the concentration parameter $\beta = (\beta_1, \ldots, \beta_{1440})$ (Eq. (4)). Using the Gibbs sampling algorithm, the dependency of the number of clusters obtained for values of β for the residential data set are shown in Fig. 2a. The number of clusters increases logarithmically with β, reaching a constant number of output clusters for a wide range of values of this parameter. Furthermore, the cluster size is reasonably stable with respect to the load profiles that formed them. The number of clusters obtained ranged between three and six for a large part of the β parameters space (from 10^{-9} to 10^{-4}). The most repeatable result is four clusters (from $\beta = 8 \cdot 10^{-8}$ to $7 \cdot 10^{-6}$). Two different sets of four clusters were obtained, differing only in the allocation of two profiles. A similar behaviour was observed with the commercial data set (Fig. 2b), but with outputs that contain five and six clusters. To obtain a reduced number of clusters the values of the parameter should be smaller than for the residential data. This is due to the higher number of profiles in the set. This robustness to the variance of the concentration parameter is advantageous with respect to other clustering algorithms such as k-means that have strong dependency on the initialisation conditions where the user has to fix the number of clusters.

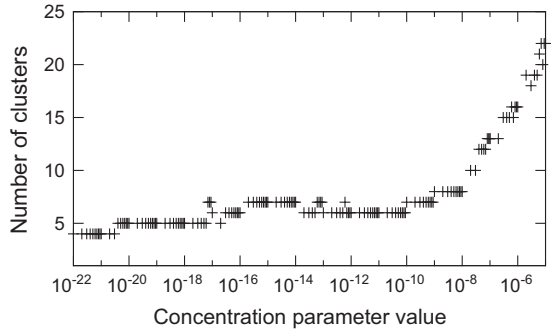
Using an informative prior based on the sampling mean reduced by the dimension (*i.e.* $\rho = 1/d$) to estimate the concentration parameter, the resulting output of the DPMM algorithm is also four clusters (Fig. 3). Applying the same technique to estimate the concentration parameter of the commercial data set, we obtained 36 clusters. However, applying a technique based on [12] and reducing by the dimension as [23] suggested, the number of clusters falls to 16. The parameter estimation techniques proposed in [24] produce an non-practical number of clusters for both data sets (gt 70 for the residential data set).

4.1. Analysis of clusters

Each of the four clusters obtained for the residential data set exhibits distinct behaviour. Of the 197 complete profiles 94 were in cluster one (Fig. 3a), 57 in cluster two (Fig. 3b), 38 in cluster three (Fig. 3c) and 8 in cluster four (Fig. 3d). From each of these the load profile representing the centroid of each cluster is plotted as shown in Fig. 4. Analysing the shape of the profile in each gives us some basic information about the general characteristics of energy use. The majority of the load profiles of cluster one present two peaks: the first one around 6am which is also a ramping up from the lower overnight load level, and a more sustained second



(a) Residential data set



(b) Commercial data set

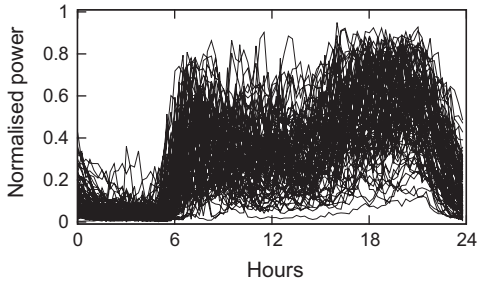
Fig. 2. Number of clusters depending on the concentration parameter β.

one at around 6 pm. These two peaks can also been seen in the profiles of cluster two. However the morning peak starts earlier, and the evening peak is usually shorter, than those of cluster one. The peak energy consumption of both peaks in cluster two appear to be similar. The profiles of cluster three present a single peak in the evening. In cluster four, there are only eight load profiles and they show a unique long peak from 6am to 6 pm. There is a sharp ramp from the overnight low consumption period for clusters 1 and 2, the centroid with the latest peak corresponds to cluster three with its single peak. The centroid of cluster four with its unique long peak has a different shape and is worthy of further examination.

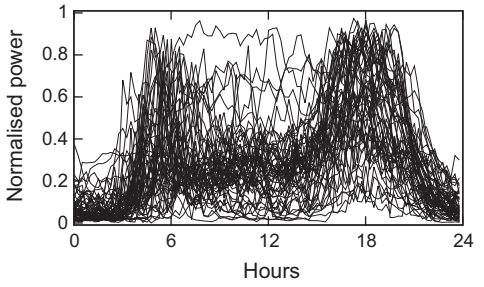
Analysing the clusters using the metadata in Table 1 we observe:

1. There is a clear division of the load profiles according to nationality (Fig. 5a). In cluster one there are a majority of profiles of English houses. Approximately 70% of the profiles in cluster two are Bulgarian, while cluster four is 100%. Cluster three is the only one where there is no clear dominant group.
2. For the number of bedrooms of the property (Fig. 5b), cluster 1 has a majority of which are three bedroom properties and cluster 4 are all single bedroom properties. The other two clusters give no clear differentiation with this feature.
3. For the type of dwelling (Fig. 5c) cluster one has the majority made up of terrace or semi properties, whereas in cluster two a clear majority are flats. Cluster four is also all flats (or Other).

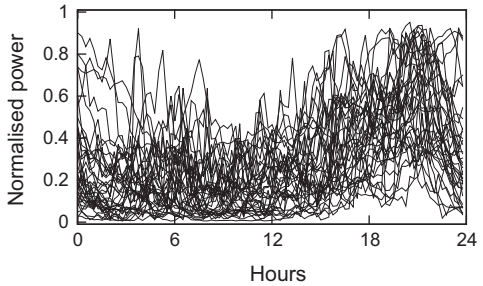
For the commercial data set, the centroids (for six clusters) show heterogeneous behaviour (Fig. 6); the composition histogram is shown in Fig. 7. The centroid of cluster one (the largest cluster) shows a double peak—the first at midday and the second around 7 pm. This cluster is mostly pubs and restaurants and shows the lunch and dinner time business peaks. In contrast, although cluster four contains a similar number of profiles and composition the lunchtime peak is clearly smaller. The evening peak is also shifted a little later, which is most likely due to the slightly larger number of clubs in this cluster. This demonstrates that quite subtle differences can be detected. Cluster two, whose centroid has a single large peak at midday, is mainly composed of restaurants and cafes which we interpret that are only open during normal office hours. The centroid of cluster three shows double peak around 7 am and 7 pm. As hotels and guest houses are the main categories of this cluster, it indicates the business activities at breakfast and dinner times. The small number of profiles of cluster five are mainly pubs



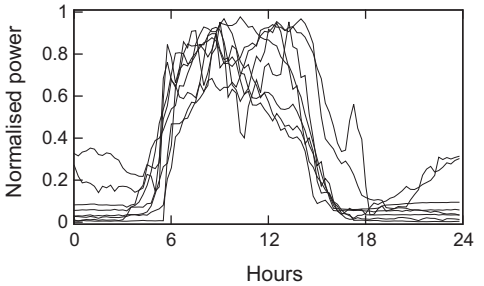
(a) Cluster 1, formed by 94 load profiles.



(b) Cluster 2, formed by 57 load profiles.



(c) Cluster 3, formed by 38 load profiles.



(d) Cluster 4, formed by 8 load profiles.

Fig. 3. Clusters obtained using the DPMM clustering algorithm for the residential data set.

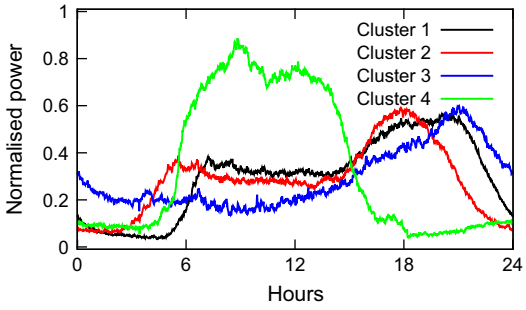
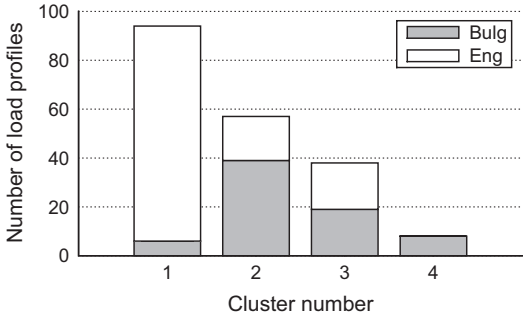
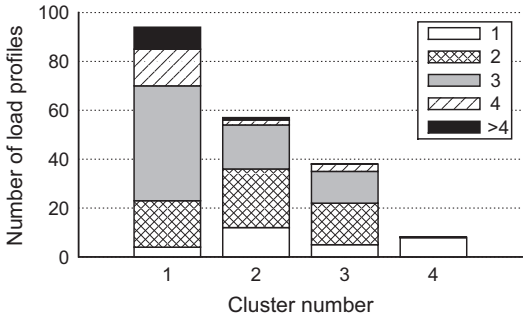


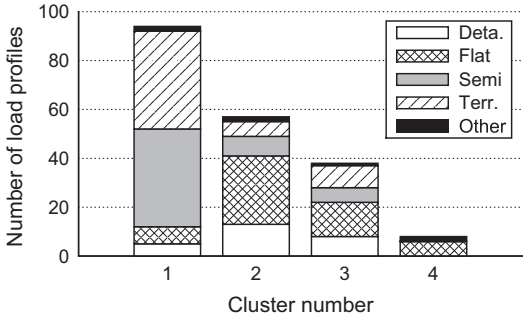
Fig. 4. Centroids of the clusters of the residential data set.



(a) Nationality feature.



(b) Number of bedrooms feature.



(c) Type of dwelling feature.

Fig. 5. Clusters analysed using features of the metadata (residential data set).

and bars that show a peak at around 10 pm and significant activity into the early hours of the following day. This suggests that they have so-called 'late' licences. Cluster six is very small and shows modest peaks at 3am and midday.

For the residential data, we examined the statistical differences between the clustered profiles using ANOVA tests (Section 4.2). This analysis suggests that we can make some further observations, although the data set is not large enough to be conclusive. The cluster four profiles are all one-bedroom dwellings with one occupant (Fig. 5b). Most of the profiles that correspond to houses with four or more occupants are in cluster one. Additionally, cluster two and three are formed by a majority of profiles from dwellings with one or two occupants. Taking into account the type of dwelling (Fig. 5c), the clearest division is that cluster one is mainly formed by load profiles whose house is a terrace or a semi. Profiles from cluster one (majority English) have the morning peak later than profiles from cluster two where the majority are Bulgarian. Whether this is statistically representative of Bulgaria is not clear, but in this data set the feature is systematic. Cluster four profiles deviate significantly from normal behaviour and their peak load is also high.

4.2. Clustering evaluation

The resulting clusters need to be compared with those obtained using other well-known techniques. To assure the validity of the comparison the output data should be in the same format (granularity and normalisation) and only results with the same number of clusters can be compared.

In the DPMM algorithm, the number of clusters is not one of the input parameters as in most other clustering algorithms. Therefore we cannot compare the results for all the possible numbers of clusters, only the ones obtained after scanning the concentration parameter (Fig. 2). The algorithms selected for comparison were: k-means, single link, complete link, pair group method average (PGMA), pair group method centroid (PGMC), and the Ward or minimum variance algorithm [26]. Most evaluators are based on computing the similarity of the data elements within each cluster, and the difference among elements of the other clusters. We used three evaluators [3,4,7]:

- the mean index adequacy (MIA) measures the distance of all the load profiles of the cluster with its cluster centre,
- the variance ratio criterion (VRC) (so-called Calinski-Harabasz Index) that is based on a ratio between intra-cluster and inter-cluster factors, and
- the scatter index (SI) makes use of the distance of data points and centre with the mean of all data points.

For the MIA and SI evaluators, lower values suggest better clustering results; it is the opposite for the VRC. The scores are shown in Fig. 8. Only results below 20 clusters are shown as above this number the population of each cluster becomes too small to make meaningful comparisons. For the MIA evaluator (Fig. 8a and b) the DPMM performed slightly worse than other techniques as the number of clusters increased. For the VRC evaluator (Fig. 8c and d) the DPMM algorithm performed moderately well, especially over the commercial data set. Using the SI evaluator (Fig. 8e and

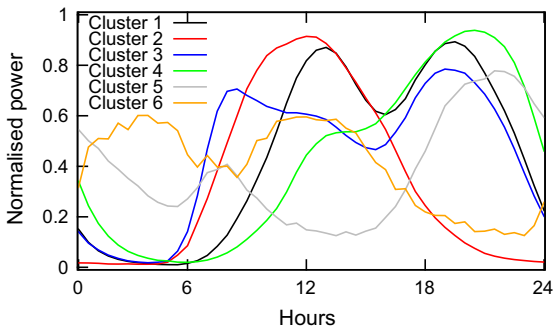


Fig. 6. Centroids of the clusters of the commercial data set.

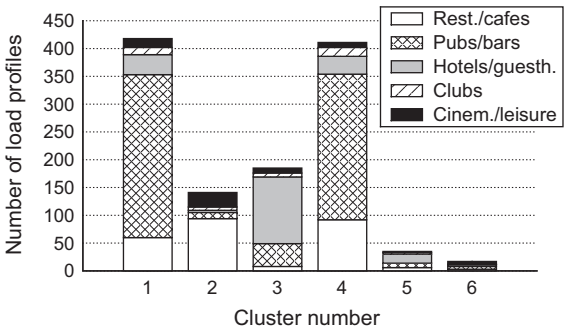
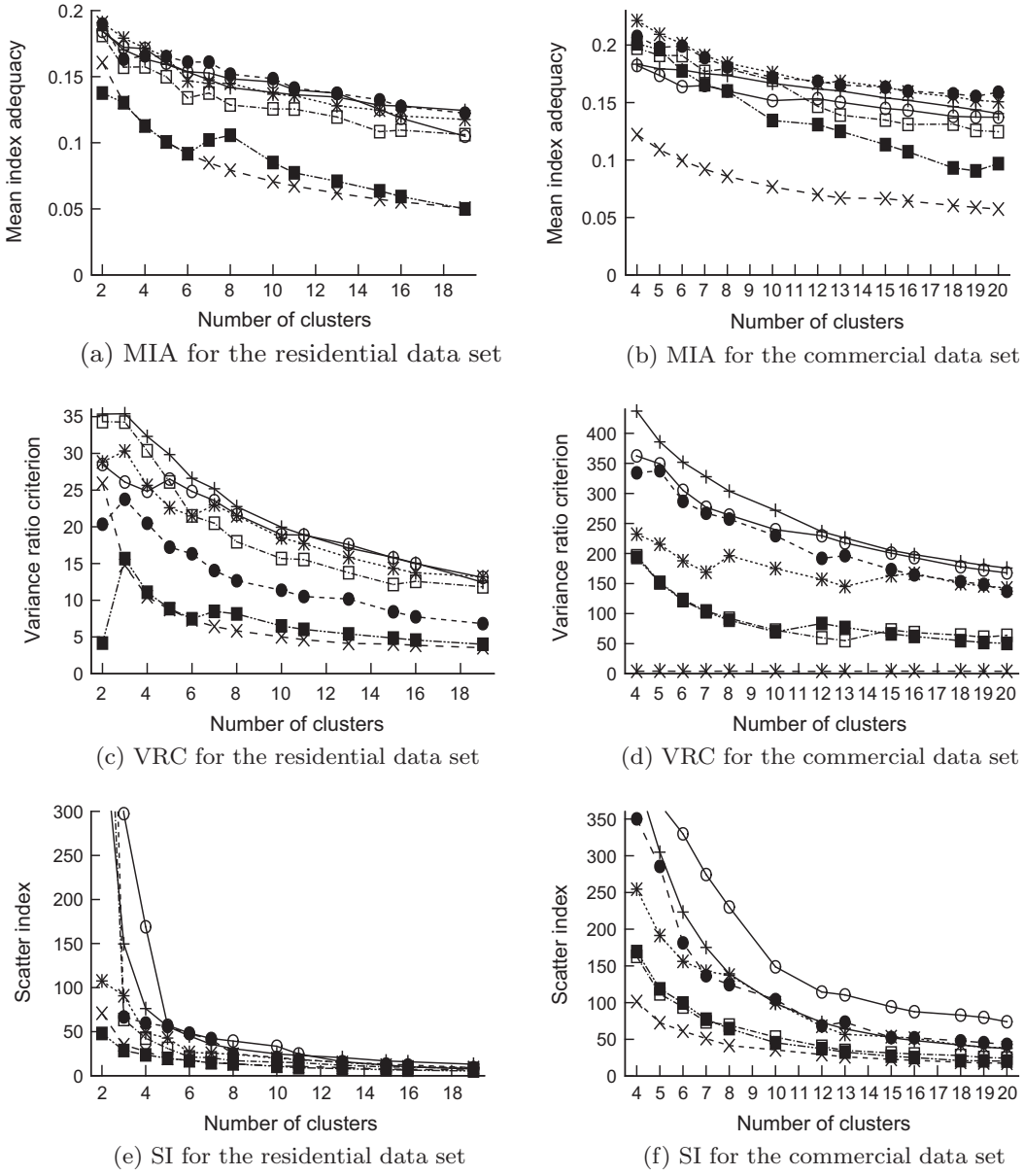


Fig. 7. Clusters analysed using their category composition (commercial data set).

f) the DPMM algorithm performed well for three or more clusters for the residential data set, and performances converged for all algorithms at higher cluster numbers for both data sets.

As an additional test of the statistical difference between the load profiles of the clusters, we conducted ANOVA tests (T-test and F-test with a significance level $\alpha = 0.05$). For the experiments using the residential data set, for each minute we tested the hypothesis that all the clusters have the same mean (*i.e.* $m_{1i} = m_{2i} = \cdots = m_{Ki}$ where m_{ji} is the mean of the jth cluster at the ith minute, and K is the number of clusters). Fig. 4 shows the means of the four clusters. The failure of the test would imply that there is at least one cluster mean that is statistically different for this particular minute. For the residential data set the results indicate that there is at least one different cluster for almost all the minutes (row 1 of Table 2). It implies that the profiles of the



(a) MIA for the residential data set

(b) MIA for the commercial data set

(c) VRC for the residential data set

(d) VRC for the commercial data set

(e) SI for the residential data set

(f) SI for the commercial data set

Fig. 8. Evaluators scores comparing different clustering techniques.

Table 2
Percentage of minute and 30-min intervals that are considered statistically different with a significance level $\alpha = 0.05$ for the residential and commercial data set, respectively.

Number of clusters	2	3	4	5	6
Percentage of minutes with at least one cluster with a different mean (residential data set)	72.9	89.7	97.1	100.0	99.9
Percentage of minutes with all the cluster means different from each other (residential data set)	72.9	57.8	45.8	52.8	67.5
Percentage of 30-min intervals with at least one cluster with a different mean (commercial data set)	–	–	100.0	100.0	100.0
Percentage of 30-min intervals with all the cluster means different from each other (commercial data set)	–	–	93.7	93.7	83.3

clusters present some degree of separation. However, it does not mean that all the cluster means are different each other, excepting for $K = 2$ where 72.9% of means during the 1440 min are statistical different between the two clusters. For this reason, a second more strict test was performed with the following condition $\bigwedge_{j,l} m_{ji} \neq m_{li}$ for $1 \leqslant j, l \leqslant K \wedge j \neq l$ for each minute $1 \leqslant i \leqslant 1440$. It implies that all cluster means are different each other. Results show (second row of Table 2) that the number of minutes that fulfil the condition changes with the number of clusters, *e.g.* six clusters seem to divide the load profiles in more different groups than four or five. This is due to the creation of subgroups with more specific behaviour. Nevertheless, the most important fact is that clusters obtained with the DPMM algorithm present a significant number of minutes that are statistically different among all the clusters, indicating that the division is justified.

Similarly, using the commercial data the less stringent test (third column of Table 2), all numbers of clusters had 100% of the minutes with at least one cluster with a different mean. When comparing the more strict condition, where each cluster mean is different from each other (fourth column of Table 2), the scores are higher than the ones obtained for the residential data set. This may due to: (1) the commercial profiles have greater variety than those of the residential set, (2) the lower temporal resolution of the commercial profiles (consumption is aggregated over a longer time).

In Fig. 9, we show the execution times of the different algorithms for experiments using the residential data set as reference. As expected from the complexity of the both Gibbs sampling algorithms used for the DPMM (Section 3.3), its running times are longer than the other clustering methods. The number of iterations (I^*) to converge the algorithm is the most important element for the DPMM and is the reason for the longer execution times. We need to be aware that these other methods start with the important advantage of knowing the number of clusters, meanwhile the DPMM algorithm has to converge to the solution that best fits the data given the model with a unique input parameter. If we compare the running time of just one iteration of the DPMM algorithm we appreciate that the time is not far from that of the PGMC and WARD algorithms for small number of clusters. The execution time and the number of iterations to converge the algorithm increase with the number of clusters. This is governed by the stop criterion explained in Section 3.2 where the profiles should remain in the same cluster for some consecutive iterations.

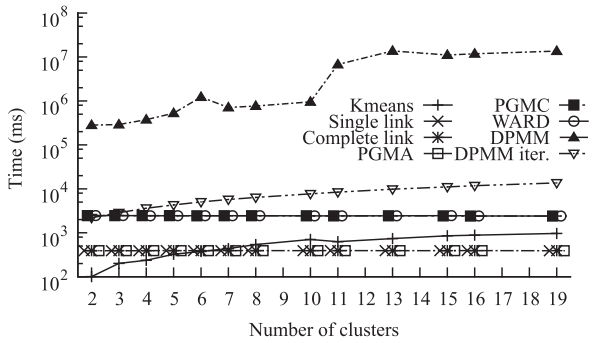


Fig. 9. Times of execution of the algorithms for the residential data set.

5. Conclusions

We have shown that a clustering algorithm based on a Bayesian non-parametric model, the DPMM, can distinguish between electrical power use profiles. The flexibility and robustness for managing uncertainty in real data of Bayesian statistics enabled us to model the unknown parameters that governed the distribution used for explaining the differences between load profile types. This method has the advantage that the number of clusters does not need to be determined before computation is initiated as there are techniques to estimate all of the model parameters. These estimation techniques are important for the resulting clusters, therefore their evolution with varying amounts of data should be taken into account to obtain a robust method. Although the computational performance of the DPMM was found to be slower than other techniques, the difference was not significant for this application. Furthermore, it may be possible to reduce computational complexity by parallelising some of the Gibbs sampling algorithm steps or allowing more relaxed convergence conditions.

Our model was tested using two different real data sets. One comprised residential energy consumption data with one minute resolution and the second of 30-min commercial profiles. The DPMM generated four and six clusters for the residential and commercial data sets respectively. In both cases this was a small enough number to be credible, yet sufficient to present meaningful and distinct load profile types. In particular, using the metadata of the dwellings our analysis showed that we could assign statistically significant features such as the nationality, household size, and type of dwelling to the cluster memberships.

From the residential data set it is apparent that the measurement devices used were not of sufficient quality as they produced unexplained high levels of device failure and data corruption, causing us to discard 10% of the available raw input data. As larger and better quality data become available through widespread deployment of smart meters, our technique can be tested more extensively with potential application to network operations.

Acknowledgements

We wish to thank Opus Energy (www.opusenergy.com) for allowing us access to the data set. This work was supported by an UK Engineering and Physical Science Research Council Grant (EP/I000194/1).

References

[1] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012;42(1):68–80.
[2] Rodrigues F, Duarte J, Figueiredo V, Vale Z, Cordeiro M. A comparative analysis of clustering algorithms applied to load profiling. In: Proceedings of the 3rd MLDM. Springer-Verlag; 2003. p. 73–85.
[3] Chicco G, Napoli R, Piglione F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21(2):933–40.
[4] Tsekouras G, Hatziargyriou N, Dialynas E. Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans Power Syst 2007;22(3):1120–8. http://dx.doi.org/10.1109/TPWRS.2007.901287.
[5] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87(11):3538–45.

[6] Gerbec D, Gasperic S, Smon I, Gubina F. Allocation of the load profiles to consumers using probabilistic neural networks. IEEE Trans Power Syst 2005;20(2):548–55. http://dx.doi.org/10.1109/TPWRS.2005.846236.
[7] Chicco G, Ilie I-S. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009;24(3):1619–28.
[8] Amin-Naseri M, Soroush A. Combined use of unsupervised and supervised learning for daily peak load forecasting. Energy Convers Manag 2008;49(6):1302–8. http://dx.doi.org/10.1016/j.enconman.2008.01.016.
[9] Keyno HS, Ghaderi F, Azade A, Razmi J. Forecasting electricity consumption by clustering data in order to decline the periodic variables affects and simplification the pattern. Energy Convers Manag 2009;50(3):829–36. http://dx.doi.org/10.1016/j.enconman.2008.09.036.
[10] Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Customer characterization options for improving the tariff offer. IEEE Trans Power Syst 2003;18(1):381–7.
[11] Andersen F, Larsen H, Boomsma T. Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers. Energy Convers Manag 2013;68(0):244–52. http://dx.doi.org/10.1016/j.enconman.2013.01.018.
[12] Reich BJ, Bondell HD. A spatial Dirichlet process mixture model for clustering population genetics data. Biometrics 2010:381–90. http://dx.doi.org/10.1111/j.1541-0420.2010.01484.x.
[13] Vlachos A, Ghahramani Z, Briscoe T. Active learning for constrained Dirichlet process mixture models. In: Proceedings of GEMS. ACL; 2010. p. 57–61.
[14] Murugiah S, Sweeting T. Selecting the precision parameter prior in Dirichlet process mixture models. J Stat Planning Inference 2012;142(7):1947–59. http://dx.doi.org/10.1016/j.jspi.2012.02.013.
[15] Fan S, Chen L, Lee W-J. Machine learning based switching model for electricity load forecasting. Energy Convers Manag 2008;49(6):1331–44. http://dx.doi.org/10.1016/j.enconman.2008.01.008.
[16] Sowden R, Tommis M. D7.7 project cycle analysis report for cycle 3, project Dehems; 2011.
[17] Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Stat 1973;1(2):209–30.
[18] Blackwell D, MacQueen JB. Ferguson distributions via Polya urn schemes. Ann Stat 1973;1(2):353–5.
[19] Aldous D. Exchangeability and related topics. In: Ecole d'Ete de probabilities de saint-flour XIII 1983. Springer; 1985. p. 1–198.
[20] Teh YW, Jordan MI, Beal MJ, Blei DM. Sharing clusters among related groups: hierarchical Dirichlet processes. In: Advances in neural information processing systems. MIT Press; 2005. p. 1385–92.
[21] Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 1974;2(6):1152–74. http://dx.doi.org/10.2307/2958336.
[22] McAuliffe JD, Blei DM, Jordan MI. Nonparametric empirical Bayes for the Dirichlet process mixture model. Stat Comput 2006;16(1):5–14. http://dx.doi.org/10.1007/s11222-006-5196-2.
[23] De Campos CP, Benavoli A. Inference with multinomial data: why to weaken the prior strength. In: Proceedings of the 22nd IJCAI. AAAI Press; 2011. p. 2107–12.
[24] Minka TP. Estimating a Dirichlet distribution. Tech. rep., MIT; 2012.
[25] Congdon P. Bayesian statistical modelling. 2nd ed. Chichester: Wiley; 2007.
[26] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Morgan Kaufman; 2006.