

# A Nonparametric Bayesian Methodology for Synthesizing Residential Solar Generation and Demand Data

Thomas Power, *Student Member, IEEE*, Gregor Verbič, *Senior Member, IEEE*,  
Archie C. Chapman, *Member, IEEE*,

**Abstract**—The uptake of *behind-the-meter* distributed energy resources in low-voltage distribution networks has reached a level where network issues have started to emerge, which requires new tools for operation and planning. In this paper, we propose a methodology for synthesizing stochastic demand and generation profiles for unobserved customers with rooftop PV, called *prosumers*. The proposed model bridges the gap between the limited available empirical data, and the large amount of high-quality, stochastic demand and generation data required for probabilistic analysis. The approach employs clustering analysis and a Dirichlet-categorical hierarchical model of the features of unobserved prosumers. Based on the data of clusters of prosumers, Markov chain models of demand and generation profiles are constructed from empirical data, and synthetic demand profiles are subsequently sampled from these. The sampled traces are cross-validated and show a good statistical fit to the observed data. Two case studies are considered to confirm the validity of the proposed methodology. The first studies the impact of behavioral differences on the synthetic demand profiles, while the second looks at the impact of varying solar generation penetration on demand profiles.

**Index Terms**—Probabilistic analysis, Markov chains, hierarchical Dirichlet model, Bayesian nonparametrics, rooftop PV, stochastic demand model, smart grid, prosumers.

## I. INTRODUCTION

THE emergence of cost-effective behind-the-meter distributed energy resources, including on-site generation, energy storage, electric vehicles, and flexible loads is changing the way that electricity consumers source and consume electric power. To give some context, installed capacity of rooftop PV has increased globally from approximately 4 GW in 2003 to nearly 227 GW in 2015, driven by government incentives, electricity price increases, and decreasing PV capital cost [1]. In Australia, the Energy Networks Australia and the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) have estimated the projected uptake of solar PV and battery storage in 2050 to be 80 GW and 100 GWh [2], respectively, which will represent between 30%–50% of total demand, a scenario called “Rise of the Prosumer”. They refer to the *prosumer* as a small-scale (residential, commercial and small industrial) electricity consumer with on-site generation.

Thomas Power, Gregor Verbič, and Archie C. Chapman are with the School of Electrical and Information Engineering, The University of Sydney, Sydney, New South Wales, Australia. e-mails: tpow5976, gregor.verbic, archie.chapman@sydney.edu.au.

The rising penetration of PV and battery storage, however, presents a number of challenges with regard to the operation of low-voltage networks. Most existing studies have only focused on the impact of rooftop PV on voltages. Chen et al. [3], for example, investigated the impact of an increasing penetration of distributed generation on the voltage profiles of US distribution networks, while Widen et al. did that for a Swedish network [4]. A recent UK study by Navarro-Espinosa and Ochoa in the UK [5] considered a wider range of *low carbon technologies*, including battery storage. They didn’t however, consider battery scheduling as a result of *home energy management* [6], [7].

A common feature of these studies is that they use probabilistic analysis to properly account for the variation in both PV generation and electricity demand. The latter, in particular, is extremely important in the context of home energy management, as it can have a significant impact on the battery schedule and, consequently, the grid demand. Against this backdrop, this paper proposes a methodology for synthesizing statistically representative demand and PV generation profiles using only a limited amount of observed smart meter data.

### A. Related work on Synthetic Demand and Generation Profiles

Much of the previous study on the synthesis of stochastic demand and generation profiles has centered around a ‘bottom-up’ approach. This particular framework is utilized by Widen, Nilsson and Wackelgard [8] to model building occupancy for the purposes of forecasting lighting demand. Their methodology involves simulating building occupancy as a Markov chain, defined by a state transition probability matrix. Richardson, Thompson and Infield [9] and Page et al. [10] use a similar methodology for sampling building occupancy profiles. McKenna and Thompson extend this methodology [11] in their high-resolution stochastic integrated thermal-electrical domestic demand model (the *CREST* model). Their particular occupancy model is central to the determination of a residence’s demand profile, and allows for a number of distinct day types, whereby it defines separate transition probability matrices for both weekdays and weekends, which captures weekly behavioral patterns. This is effective in that there are significant structural differences in the behavior of household demand between weekends and weekdays. The generation component in the *CREST* model first involves sampling the weather conditions. Similarly to the demand

model, a transition matrix is constructed from prior historical data for the clearness index, which is a measure of the extent to which incident irradiance is blocked from reaching a PV module by cloud cover. In a method similar to that used by Hofmann et al. [12] and Bright et al. [13], initial states are chosen at random, and the clearness index profile is sampled at each timestep through a repeated application of a state transition probability matrix [11]. This is then used to generate stochastic PV generation profiles [11].

### B. Research Overview

This work develops a framework for producing a large collection of residential demand and generation profiles from limited observed *smart meter* data. Previous studies have either involved using a simplified probabilistic model [3], or reconstructing profiles using highly granular observed data, which is computationally costly [11]. The methodology described in this paper requires data on a household level that is somewhat less granular, and more readily accessible. It also aims to retain the variance that is normally observed among observed demand and generation profiles.

The rest of the paper proceeds as follows. Section II details the formulation of the proposed model, while Section III cross-validates the model. Section IV outlines potential applications of the model through two specific case studies. Finally, Section V concludes.

## II. MODEL FORMULATION

We wish to synthesize typical solar generation and demand profiles for a household according to specific features. These features can either be continuous or discrete, and may not be independently distributed within the data. Discrete features generally need no treatment prior to assignment. However, if the number of values taken on by numerical features is impractically large, some clustering analysis may be done, as described next. After this, the method used to identify a Markov chain model of the demand and solar generation profiles is described, followed by the steps used to generate new synthetic demand and generation traces.

### A. Feature-based Cluster Assignment

For this clustering analysis, empirical data was collected during the *Ausgrid Smart-Grid Smart-City* (SGSC) project [14]. Clustering is important because (i) considering each customer as a single category is computationally expensive, and (ii) it provides generalizable statistical information as the demand and PV generation in each set are correlated with their features.

Let  $n \in \mathcal{N}$  and  $m \in \mathcal{M}$  denote the set of observed and unobserved customers, respectively. Clustering analysis is run on empirical data to assign the  $n \in \mathcal{N}$  customers into representative sets, denoted  $k \in \mathcal{K}$  according to their features. The features of demand are the day types (weekday or weekend) and number of residents, while those for PV include the PV capacity, panel orientation and weather information. Clustering is completed using either  $k$ -means clustering [15]

or *maximum a-posteriori Dirichlet process mixtures* (MAP-*DP*) clustering [16], which is useful for instances in which the number of clusters cannot be easily determined.

### B. Estimating the Dirichlet Distribution

After clustering, we could compute the frequencies,  $\{p_s\}_{k \in \mathcal{K}}$ , of each observed customer being a member of a certain cluster. These values can be interpreted as the probability of a new, unobserved customer having certain features. However, they are only an estimate across the observed customers, and directly using them to allocate features fails to properly consider the error in this estimate, which can be significant where the fraction of customers observed is small. Thus, a Bayesian estimation approach is employed.

Specifically, the model uses the count of each  $k \in \mathcal{K}$  in the observed  $\mathcal{N}$  as a hyper-parameter of a Dirichlet distribution, which itself is sampled to yield a *categorical* probability distribution over the features for unobserved customers,  $m \in \mathcal{M}$ . Formally, this is given by:

$$\begin{aligned} \alpha & \quad \text{Vector of cluster counts} \\ \mathbf{q} \mid \alpha & \sim \text{Dir}(\alpha) \\ S_m \mid \mathbf{q} & \sim \text{Cat}(\mathbf{q}) \end{aligned}$$

In more detail, given the clusters, we use the Dirichlet distribution, given by the the probability density function:

$$f(\mathbf{q}_j; \alpha_j) = \frac{\Gamma(\alpha_{j0})}{\prod_{k=1}^m \Gamma(\alpha_{jk})} \prod_{k=1}^m q_{kj}^{\alpha_{jk}-1}. \quad (1)$$

The Dirichlet distribution accounts for variance in the distribution of features across unobservable customers, which overcomes a limitation of previous models.

In this case,  $\mathbf{q}_j$  is the resultant probability measure for feature  $j$  drawn from the Dirichlet distribution, with individual elements  $q_{jk}$ , where  $k \in \{1, 2, \dots, m\}$ , and  $\sum_{k=1}^m q_k = 1$ , while the corresponding random probability measures and elements are given by  $\mathbf{Q}_j$  and  $Q_{jk}$ , respectively. The distribution is parametrized by  $\alpha_j$  which is given by:

$$\alpha_{jk} = \sum_{i=1}^n \delta_{c_{ij}, k}, \quad \forall j \leq u, \quad \forall k \leq m, \quad (2)$$

where  $\alpha_{j0} = \sum_{k=1}^m \alpha_{jk}$ , and  $\delta$  is the Kronecker delta. For each feature,  $\mathbf{q}_j$  parametrises a draw from a categorical distribution, such that for  $n^*$ , unobserved prosumers, we have:

$$\mathbf{A}_j^* = [A_{j1}^*, A_{j2}^*, \dots, A_{jm}^*], \quad (3)$$

which is defined by the probability density function:

$$f(\mathbf{A}_j^*; \mathbf{q}_j) = \frac{\Gamma(\sum_{k=1}^m A_{jk} + 1)}{\prod_{k=1}^m \Gamma(A_{jk} + 1)} \prod_{k=1}^m \mathbf{p}_k^{A_{jk}}. \quad (4)$$

This collection of feature counts defined by  $\mathbf{A}_j$  are then assigned to the unobserved prosumers.

### C. Demand Model

The demand model outlined in Fig. 1 synthesizes stochastic demand profiles for unobserved prosumers using observed data for characteristically similar prosumers. To do this, Markov chain modeling is used, with a state transition matrix defined for each timestep throughout the day. This model is time inhomogeneous, as there are structural differences in the behavior of household energy usage at different points in the day. Once the timestep length and the collection of relevant observable customers have been defined, the state transition matrices are then constructed. For this study, a timestep of 30 minutes is used and as such, 48 different state transition matrices will be required to sample a profile over a 24 hour period. In addition to this, it will need to be specified whether the resulting profile is to be synthesized for a weekend or a weekday, as they have significantly different behavior. Following this, the state transition matrices can be defined:

$$\mathbf{M} = [\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_{48}]. \quad (5)$$

In each matrix  $\mathbf{M}_k$ , the states are defined for

$$m_{i,j,k}, \quad i, j \in \{0, 1, \dots, n\}, \quad (6)$$

and are used to represent the amount of energy consumed for the duration of timestep  $k$ . In the case of the validation data used in this work [14], the data is logged in terms of the number of kWh consumed in each relevant 30 minute period. In order to construct the state transition probability matrices, these readings are discretized with each kWh reading rounded to the nearest 0.01 kWh, and the possible attainable states are defined as being this value multiplied by 100. The state, therefore, of a reading of 1.023 kWh will be 102. These states are defined up to an arbitrary maximum limit  $n$ , which is set well above the maximum observed state. The state transition probability matrices are built up from the observed data by taking:

$$m_{i,j,k} = \sum_{h=1}^N \delta_{x_{hk-1},i} \delta_{x_{hk},j}, \quad (7)$$

over each similar observed prosumer in the available data. Stochastic demand profiles can then be directly sampled from this set of matrices by first calculating the conditional distribution of the initial state, which is given by:

$$\mathbf{p} = [p_1, p_2, \dots, p_n], \quad p_j = \frac{\sum_{i=1}^n m_{i,j,1}}{\sum_{i=1}^n \sum_{j=1}^n m_{i,j,1}}. \quad (8)$$

Once this initial state  $s_1$  is drawn from the categorical distribution parametrized by  $\mathbf{p}$ , subsequent states for the remainder of the day can be drawn from the following matrices. For each state  $s_k$ , the associated probability measure is derived from  $\mathbf{M}_{s_{k-1},j,k}$  using kernel density estimation. Gaussian kernel density estimation is used to make each state attainable, even if it has not been observed in the data. It is given by:

$$p_{s_{k-1},j,k} = \frac{\sum_{j^*=1}^n m_{s_{k-1},j^*,k} e^{-\frac{(j-j^*)^2}{2h^2}}}{\sum_{j^*=1}^n e^{-\frac{(j-j^*)^2}{2h^2}}}. \quad (9)$$

From this kernel density estimate, the state  $s_k$  can be drawn from a categorical distribution:

$$s_k \sim \text{Cat}(\mathbf{p}_{s_{k-1},k}). \quad (10)$$

Before synthesizing these profiles, it will need to be specified whether the the profiles are to be synthesized for a weekday or a weekend. There are significant structural differences in energy usage behavior between weekdays and weekends [17], and as such, only observed data from weekdays is to be used to synthesize weekday profiles, while only observed data from weekends is to be used to calculate weekend profiles.

### D. Generation Model

The generation model is the third key component in this model. It is intended to synthesize solar generation data in order to compile stochastic daily profiles similar to the demand model mentioned previously. This generation model was developed and described in detail in a previous work [18], and utilizes a Markov chain framework similar to that provided by the CREST model [9]. In this model, solar PV generation is modelled using the following formula:

$$P = \eta \times A \times TIF \times CI \times G, \quad (11)$$

where  $\eta$  is the efficiency of the PV module, and  $A$  is the area of the module. For the purposes of this work,  $TIF$  is the ‘time irradiance factor’, and is a measure of the incident irradiance on a PV module with respect to the latitude of the panel, orientation, time of day and time of year.  $G$  is the extraterrestrial irradiance constant, while  $CI$  is the clearness index, which represents the extent to which cloud cover obstructs irradiance of the panel. This is the stochastic quantity that this model emulates. As detailed in previous work [18], the clearness index for a particular prosumer is sampled from a single state transition probability matrix, which is constructed from observed generation data. This is substituted into (20) to obtain stochastic PV generation values for each timestep of the day.

## III. MODEL VALIDATION

### A. Feature Assignment

The following section details the results of a cross validation of the model using data sourced from Ausgrid’s Smart Grid, Smart City (SGSC) data set [14]. This data contains a three year sample of half hourly household demand and generation measurements for approximately 13,000 households. Following model cross validation, a number of case studies are then conducted in Section IV to demonstrate potential applications.

### B. Cross Validation

The first part of this framework involves the clustering and assignment module. For the SGSC data set, the peak readings for solar installations are the only relevant variable available for clustering, as they act as a proxy for the capacity of the system. The distribution of the 189 observations used for clustering is shown in Fig. 2.

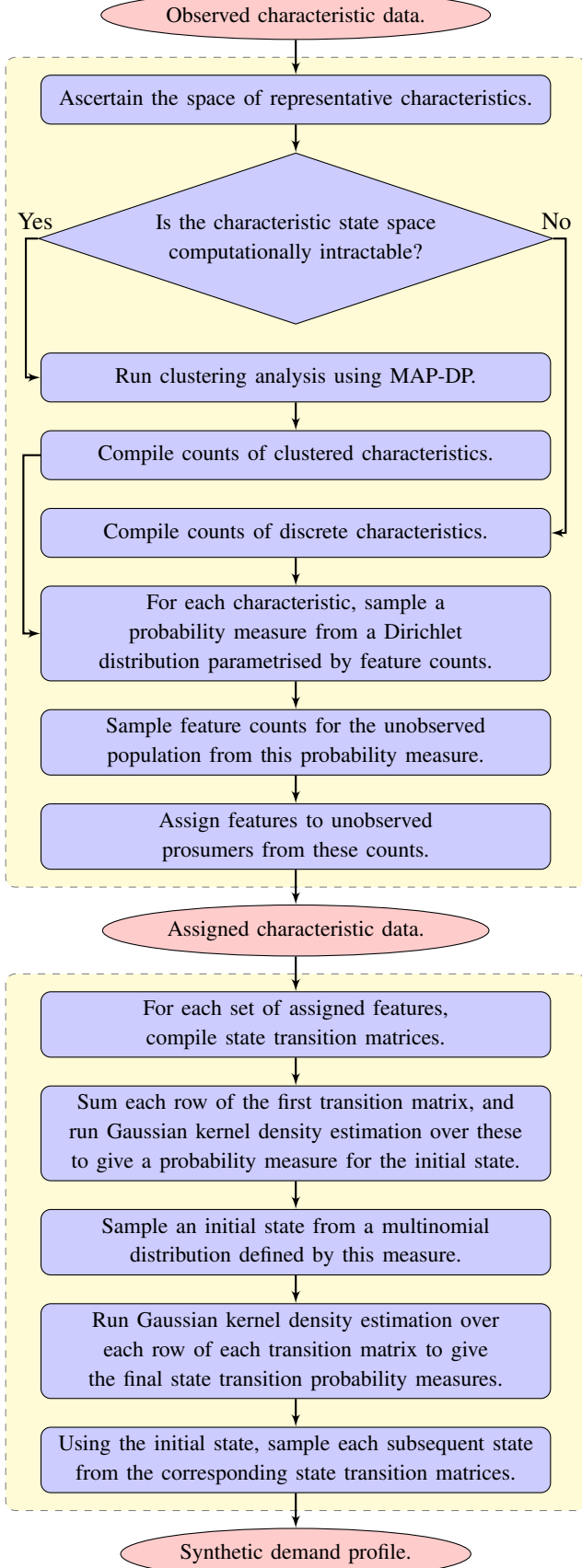


Fig. 1. The characteristic assignment and profile synthesis process.

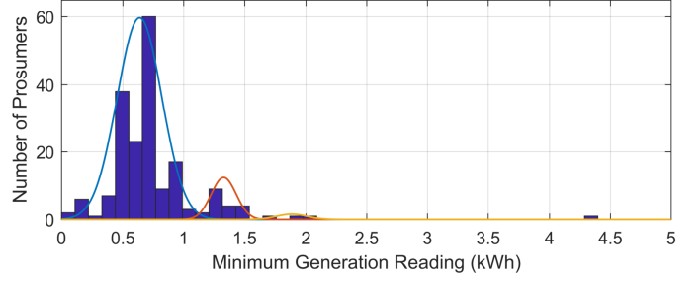


Fig. 2. Histogram of peak solar readings for all observed generators.

TABLE I  
RESULTS OF MAP-DP CLUSTERING WITH A DIRICHLET PARAMETER OF 9

| Cluster ( $i$ )          | 1      | 2      | 3      | 4      | Error |
|--------------------------|--------|--------|--------|--------|-------|
| Location ( $\bar{x}_i$ ) | 0.6368 | 1.3293 | 1.8843 | 4.3960 | -     |
| Variance ( $s_i$ )       | 0.0338 | 0.0104 | 0.016  | 0      | -     |
| Population ( $n_i$ )     | 165    | 20     | 3      | 1      | 5.83  |

One of the limitations of  $k$ -means clustering is that the number of clusters and prior estimates for cluster locations are required before running the algorithm. Therefore we use the MAP-DP clustering algorithm proposed by Raykov et al. [16]. For this algorithm, an initial cluster mean and variance is specified, along with a prior estimate for the variance of new clusters and the Dirichlet parameter  $\alpha$ , which represents the propensity of the algorithm to pick new clusters for assignments. In the case of the peak solar reading data, this clustering algorithm was run with the naive prior cluster mean of 0.7428, and variance of 0.1700.

Using these values, the MAP-DP algorithm was run using a Dirichlet parameter of 9, which yielded the 4 clusters shown in Table I. The error is comparable to  $k$ -means clustering, which used an informed prior. This therefore demonstrates the efficacy of MAP-DP under a lack of prior information.

A superposition of these clusters comprises a Gaussian mixture model, which is then normalized for assignments to unobserved prosumers, as shown in Fig. 2. The Gaussian mixture model is given by the following probability mass function:

$$f(x) = \frac{\sum_{i=1}^N \frac{1}{\sqrt{2\pi s_i^2}} e^{-\frac{(x-\bar{x}_i)^2}{2s_i^2}}}{\sum_{i=1}^N n_i}, \quad (12)$$

where  $\bar{x}_i$  is the mean,  $s_i$  is the variance and  $n_i$  is the population of cluster  $i$ . The MAP-DP clustering analysis has therefore revealed a set of three Gaussian distributed clusters from which this peak solar data is distributed. In contrast to the  $k$ -means algorithm, this analysis did not require prior knowledge of cluster centroids or variances, but only an estimate for the Dirichlet parameter. This makes the MAP-DP algorithm much more versatile where little is known about the subject data. Once these features are clustered, the resulting set of clusters can be used to parametrize a Dirichlet distribution. This defines a distribution for the probability measure for the assignments, thereby giving it a variance as well as an expected value, as illustrated in Table II.

TABLE II  
A SEQUENCE OF DRAWS FROM A DIRICHLET DISTRIBUTION  
PARAMETRISED BY MAP-DP CLUSTERED PEAK SOLAR OUTPUT DATA

| Draw     | $q_1$   | $q_2$   | $q_3$   | $q_4$   |
|----------|---------|---------|---------|---------|
| 1        | 0.8802  | 0.1104  | 0.0049  | 0.0046  |
| 2        | 0.8370  | 0.1503  | 0.0106  | 0.0021  |
| 3        | 0.8567  | 0.1237  | 0.0053  | 0.0143  |
| 4        | 0.8990  | 0.0925  | 0.0079  | 0.0005  |
| 5        | 0.8426  | 0.1459  | 0.0088  | 0.0027  |
| 6        | 0.8691  | 0.1006  | 0.0197  | 0.0105  |
| 7        | 0.8819  | 0.0805  | 0.0321  | 0.0054  |
| 8        | 0.8758  | 0.0967  | 0.0239  | 0.0036  |
| 9        | 0.8042  | 0.1659  | 0.0211  | 0.0087  |
| 10       | 0.8539  | 0.1270  | 0.0137  | 0.0054  |
| Mean     | 0.86004 | 0.11935 | 0.01480 | 0.00578 |
| Variance | 0.00075 | 0.00079 | 0.00008 | 0.00001 |

Following the sampling from a Dirichlet distribution, the feature counts drawn from a categorical distribution parametrized by these samples reflect the mean and variance of the probability measure itself, as shown in Table II, which emulates the variance in the estimation of the distribution of features over the whole population.

### C. Demand Model Validation

The function of the demand module in this model is to take a collection of demand data from observed prosumers as an input, and to synthesize either daily or multiple day profiles based on this data. Following the assignment of features to unobserved prosumers, the set  $\mathbf{M}$ , of state transition matrices can then be populated using observed data from prosumers sharing similar features. In this case, the set of matrices is limited to  $700 \times 700 \times 48$ , which corresponds to the maximum demand being 7 kWh over a half-hour period. In order to conduct a cross-validation of this model, the synthetic profiles are compared to those observed from a separate subset of the data to check for similarity. In order for this cross validation to be valid, subsets of the data need to be taken such that they do not contain any nested statistical factors. As such, the data random selection of 1000 prosumers are used for the model construction, and a separate random subset of another 1000 prosumers is used for cross validation. The matrices calculated using this process are large and sparse, and as such, and as such, the state transition matrix at 6:00pm on weekdays is visualized for states 0 to 100 as a heatmap in Fig. 3.

As these matrices are sparse, the following logarithmic transform of the counts is used for  $m_{ijk} \in M_i$ :

$$m_{ijk}^* = \log_{10}|m_{ijk}| \quad i, j \in \{1, 2, \dots, 100\}. \quad (13)$$

The heatmap shown in Fig. 3, uses red to denote a high count, while blue indicates a low count. Each pixel represents a distinct state transition count, with the previous state, or state before 6:00pm denoted by the row, and the state after 6:00pm denoted by the column.

It can therefore be seen from Fig. 3 that a majority of transitions take place from states near zero, to similar states near zero. There are two major features to this data, the first of which is that transitions to higher states tend to be less likely than transitions to lower states. The second feature of

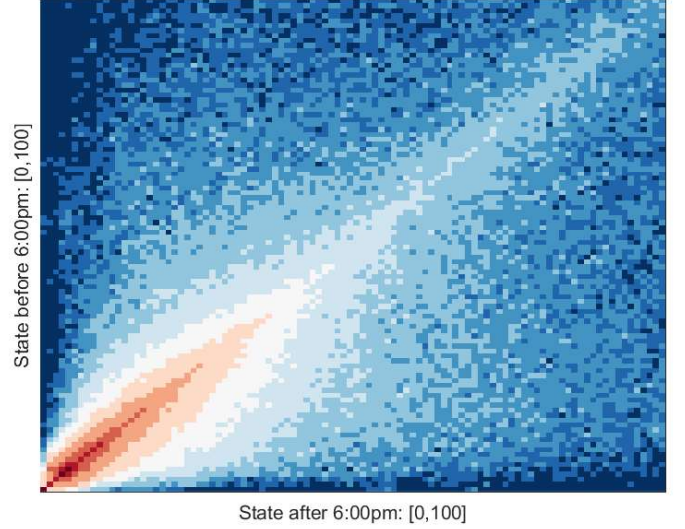


Fig. 3. Logarithmic heatmap of the Markov state transition matrix calculated at 6:00pm for 1000 prosumers.

this visualization is that transitions to similar states are more likely than transitions to dissimilar states. From this set of matrices, profiles can be sampled directly, as shown in Fig. 4 (top). When aggregated, however, it is apparent that the mean of this synthetic data matches that of the observed data. This is illustrated in Fig. 4 (bottom). It is visually clear that the data synthesized by the model shows similar behavior to the observed data.

In order to better quantify this, the mean absolute error of the estimate given by the synthetic profiles is calculated. This is given by:

$$\text{Mean Absolute Error} = \frac{1}{48} \sum_{i=1}^{48} \frac{|x_i - x_i^*|}{x_i}, \quad (14)$$

where 48 is the number of samples in the daily profile,  $x_i$  is the observed value at timestep  $i$ , and  $x_i^*$  is the synthetic value at timestep  $i$ . The mean absolute error is calculated for the weekday profile to be 9.80%. Given that the samples range between approximately 200 kWh and 500 kWh over the course of a day, this error rate is small compared to the range of the observed data itself. It is therefore evident that the model can accurately represent the behavior of observed data. It is also evident that a large amount of information is lost when modeling is done using the aggregate, as the information on the variability of individual profiles is not retained.

### D. Multiple Day Profiles

One important aspect of the Dirichlet process is its reinforcing property as detailed by MacQueen [15]. This allows for the generation of samples that have a tendency to follow recurring trajectories and sequences of states. In this way, states and sequences of states that have previously been sampled more often have a higher chance of being sampled again, and recurring behaviors will start to arise. To emulate the recurring behavior of an actual consumer, we propose a variation of the process described in Section II-C. This



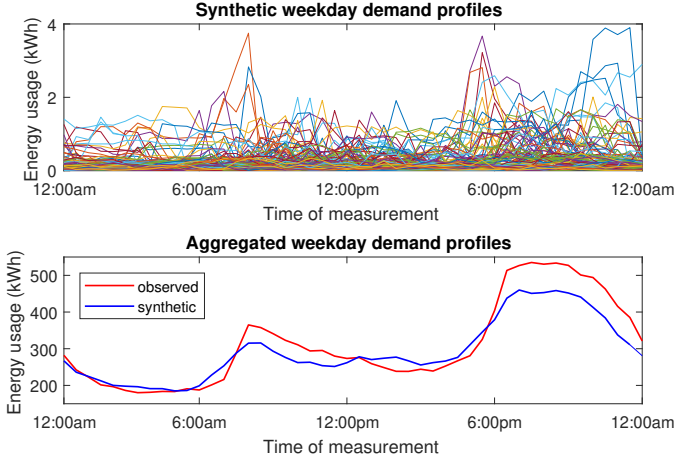


Fig. 4. Sample synthetic demand profiles (top) and their average value compared to the observed average profile (bottom).

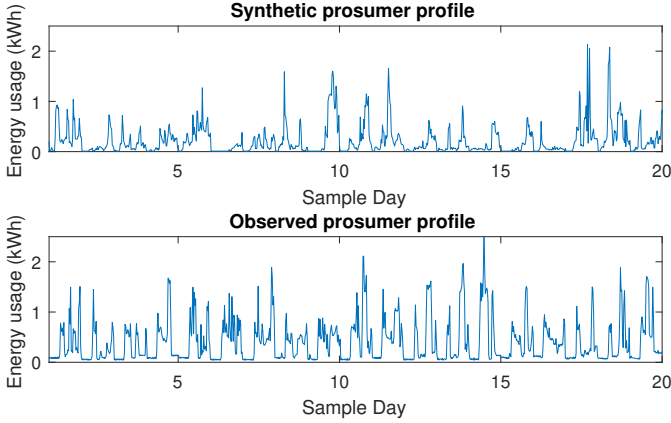


Fig. 5. Synthetic (top) and observed (bottom) multi-day prosumer profile.

process transforms the original matrices of counts uniquely for each prosumer to be modeled, by drawing from a Dirichlet distribution parametrized by each row.

Fig. 5 shows two multi-day prosumer profiles: a synthetic one (top) and a profile observed in the SGSC dataset (bottom). Some degree of habitual daily behavior is evident in each of these profiles. To quantify this effect, autocorrelation is calculated for each sample, which measures the correlation between each day and the day preceding it:

$$R(\mathbf{x}, n) = \frac{1}{(N-n)s^2} \sum_{t=1}^{N-n} (x_t - \bar{x})(x_{t+n} - \bar{x}), \quad (15)$$

where  $n$  is the lag in timesteps,  $s^2$  is the sample variance, and  $N$  is the total number of samples. The autocorrelation in this case calculates the correlation between any sample  $x_{t+n}$  and the corresponding sample  $x_t$  at  $n$  steps beforehand. Using  $n = 48$  allows for the correlation between daily behaviors to be quantified. In this particular case, the observed profile had a calculated daily autocorrelation of 0.3066, while the synthetic profile was 0.1993. This shows that in this case, the synthetic profile does emulate habitual daily behavior to a large extent, but does not entirely reproduce the habitual

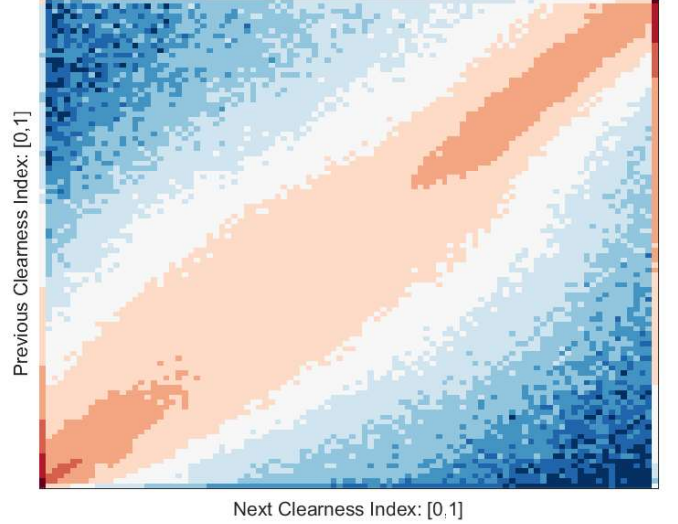


Fig. 6. Clearness index state transition probability heatmap for 90 observed prosumers.

behavior in the observed sample. This is also evident in an average autocorrelation taken over ten synthetic samples and ten observed samples, which yielded 0.1429 and 0.3584 respectively.

#### E. Generation Model Validation

Similarly to the demand model, the generation model is cross-validated using the SGSC data set. Using the methodology outlined in previous work [18], a clearness index state transition matrix is generated using data from 90 prosumers over the three-year sample period. This is illustrated in Fig. 6, in which a number of key features are evident. Firstly, it is apparent that the distribution of clearness index state transitions is bi-modal, with transitions being more likely from low indices to low indices, and high indices to high indices, shown in red. There is also a discontinuity at clearness indices 0 and 1 exactly, at the extreme left and right columns of the figure respectively. Transitions to these states are relatively likely, and transitions from these states are relatively unlikely. This shows that the clearness index has a tendency to transition towards 0 and 1.

Following the population of the clearness index state transition matrix, profiles can be synthesized. In Fig. 7, the aggregate of 100 synthetic profiles is shown in red, generated for the approximate location of Sydney city in March. These profiles have been generated similarly to those detailed in [18]. These are shown along with 100 aggregated observed profiles, which are shown in blue. From Fig. 7, it is clear that the synthetic solar profiles follow a similar curve to the observed profiles over the course of the day. Given that only 100 profiles were used in this case, the variance is higher than in the demand model cross validation mentioned previously. The maximum absolute error in the estimate occurs at 8:00am, and is 11.6 kWh. Maximum absolute error is used in this case as the mean absolute error metric defined for the generation module cross validation is not effective in this case. This is due

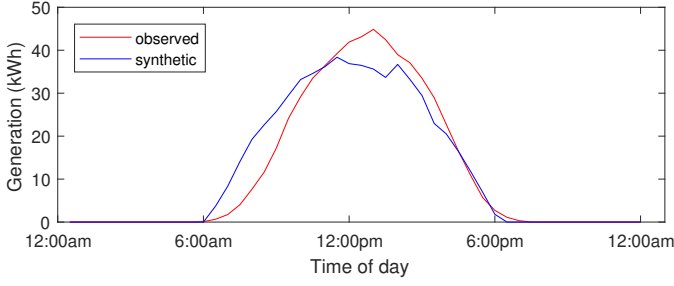


Fig. 7. 100 aggregated synthetic solar profiles (blue), and 100 aggregated observed solar profiles (red).

to the fact that the values approach zero before approximately 6:00am and after approximately 6:00pm, and the percentage error in these cases does not effectively represent the accuracy of the estimation.

#### IV. CASE STUDIES

In order to demonstrate some of the potential applications of this model, two specific case studies have been conducted. The first one makes use of the variation in geography evident in the SGSC data set. The model is used to investigate if there are structural behavioral differences between demand and generation profiles in dense urban, suburban, and rural settings and compares and contrasts these. The second case study investigates the effect on aggregated generation and net demand profiles of an increasing penetration of distributed generation systems.

##### A. Case Study: Geographical Variation

The first case study illustrates differences in demand behavior between rural, urban and suburban postcodes on both weekdays and weekends. For this study, three postcodes were selected. These are the urban Sydney suburb of Surry Hills, with 276 observed prosumers, Suburban Newcastle, with 832 prosumers, and the rural town of Pokolbin, with 266 prosumers. Using data from these prosumers, demand state transition matrices were compiled for each postcode and 1000 synthetic profiles were generated for each. The aggregated results for weekday and weekend demand are detailed in Fig. 8. It is apparent from this figure that the 1000 simulated residences in Pokolbin (green) have a generally higher demand than those in Newcastle (red) and Surry Hills (blue). It is evident that the morning peak in each profile occurs at approximately the same time, being 7:00am - 8:00am. There is a difference, however, in the afternoon peak, which occurs just after 6:00pm for Pokolbin, with Newcastle peaking just after, and Surry Hills peaking later, at about 8:00pm. In addition to this, the Pokolbin profile appears to have a much bigger morning peak relative to the level of demand over the middle of the day, while the morning peak in Newcastle and Surry Hills is lower.

During the weekend, the demand during the middle of the day is higher. Both Newcastle and Pokolbin have a morning peak at a similar time to weekdays, however Surry Hills has

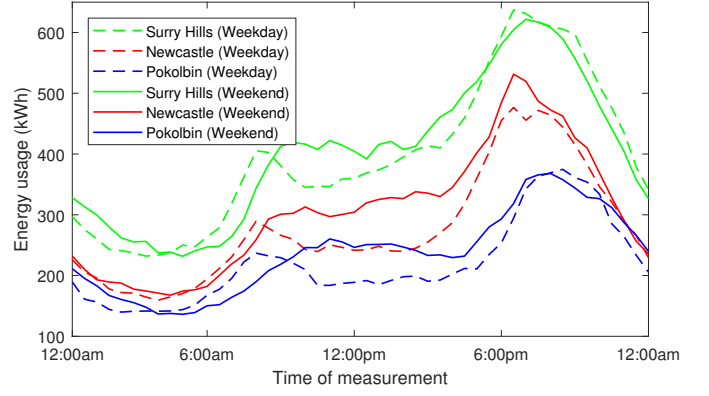


Fig. 8. 1000 aggregated weekday and weekend synthetic demand profiles for Surry Hills (Blue), Newcastle (Red) and Pokolbin (Green).

a lowered morning peak, which occurs at around 10:00am-11:00am. For each of the three profiles, demand does not drop significantly after the morning peak, but instead stabilises, or rises throughout the day. In contrast to this, the evening peak and overnight profile does not differ significantly from the weekday profiles.

This case study demonstrates the application of this model to the generation of load profiles in specific areas. The model was therefore able to synthesize 1000 daily profiles for each postcode, using less than 300 prosumers to populate the Pokolbin and Surry Hills models, and approximately 830 prosumers to populate the Newcastle model. This method has applications to the analysis of aggregated loads at the ends of distribution feeders, or even certain zone substations, where only a limited amount of observed data are available.

##### B. Case Study: Increasing Solar Penetration

The purpose of this case study is to demonstrate the function of the model to simulate the distribution of solar generation amongst a population, given a small set of observed data, and to investigate the effect on net demand profiles with increasing penetrations of installed solar generation. For this case study, the postcode 2280 is used, this corresponds to the southern suburbs in Newcastle. This postcode is an excellent candidate for this study as it has a large amount of observed demand data available in the SGSC dataset, and in addition, also has 61 prosumers with observed solar data available. This means that of the 225 observed customers with data available, 27.11% of these have installed solar generation. For this case study, a set of 1000 unobservable prosumers is used.

In the sampling of the aggregate solar load profile, the same clearness index transition profile was used for every prosumer, as they are tightly grouped geographically. This replicates the effect of the suburb experiencing the similar transitions in cloud cover. Following the formulation of an aggregate generation profile, this is subtracted from the demand profile to yield a net demand profile.

Fig. 9 shows a set of ten synthetic demand profiles for a distributed generation penetration of 27.1% (top), 80% (middle) and the average for six different PV penetration levels (bottom). Observe that for the 80% penetration level,

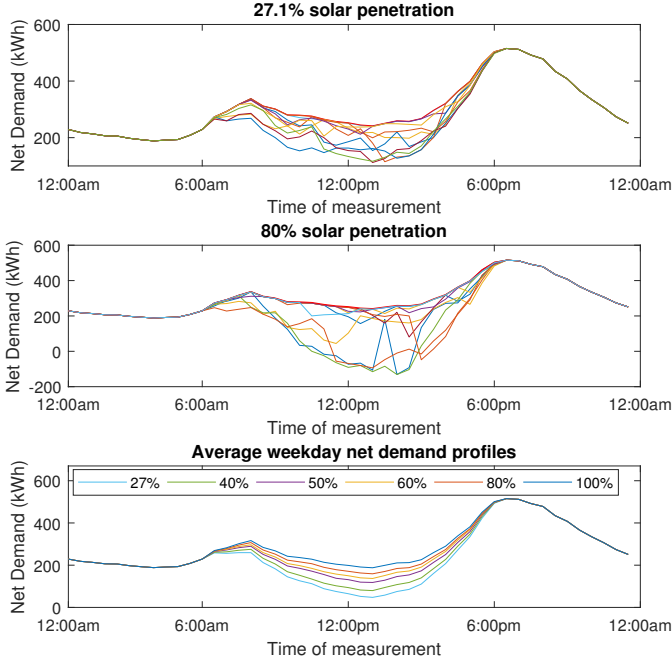


Fig. 9. The impact of PV penetration on the demand profiles: individual demand profiles for two different PV penetration levels, 27.1% (top) and 80% (middle), and the average of 1000 demand profiles for six different PV penetration levels (bottom).

the midday trough is more pronounced, with net demand becoming negative on some days. One feature which remains common to these charts is that the evening peak is largely unchanged from its gross demand value. It is therefore evident that the level of PV penetration has little impact on the daily peak residential demand, which is the evening peak. In this case study, the solar profiles are generated for early March, and as such it is expected that the evening peak will be similarly unaffected from March to September.

The bottom plot shows the average net demand profile for, respectively, 27%, 40%, 50%, 60%, 80% and 100% PV penetration. Observe that unlike for the individual demand profiles, the average demand profile never goes negative, even for the 100% PV penetration, which highlights the importance of individual demand profiles.

## V. CONCLUSION

In this paper, we have proposed a Bayesian framework for probabilistic generation of synthetic residential demand and generation data using a limited set of observed customer data. The use of the Dirichlet multinomial distribution overcomes a limitation of previous models, which is they do not account for variance in the distribution of characteristics across unobservable customers. We have shown that the methodology performs well even when the number of unobservable customers greatly exceeds the number of observable customers. One of the issues with sparse observed data is that some states could be calculated to have a zero probability of being sampled, because they have not been observed in the input data, which we overcome by using Gaussian kernel density estimation. Finally, we have proposed a Dirichlet process which allows

for the generation of samples that have a tendency to follow recurring trajectories and sequences of states. This sampling process involves reinforcement, where the state sampled is added into the initial matrix of counts. This results in recurring behavior, which is particularly important in synthesizing residential consumption data. The proposed methodology has been successfully applied to study the impact of network tariffs in low voltage distribution networks [19] and the impact of PV-battery system impacts on low-voltage distribution networks [20], respectively.

## REFERENCES

- [1] International Energy Agency (IEA), "Energy prices and taxes quarterly statistics," Tech. Rep., 2018.
- [2] CSIRO and Energy Networks Australia, "Electricity Network Transformation Roadmap," Final Report, 2017.
- [3] P. Chen, R. Salcedo, Q. Zhu, F. de Leon, D. Czarkowski, Z. Jiang, V. Spitsa, Z. Zabar, and R. Uosef, "Analysis of voltage profile problems due to the penetration of distributed generation in low voltage secondary distribution networks," *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 2020–2028, 2012.
- [4] J. Widen, E. Wackelgard, J. Paatero, and P. Lind, "Impacts of distributed photovoltaics on network voltages: Stochastic simulations of three swedish low-voltage distribution grids," *Electric Power Systems Research*, vol. 80, no. 12, pp. 1562–1571, 2010.
- [5] A. Navarro-Espinosa and L. Ochoa, "Probabilistic impact assessment of low carbon technologies in lv distribution systems," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 2192–2203, 2016.
- [6] H. Tischer and G. Verbič, "Towards a smart home energy management system - a dynamic programming approach," *2011 IEEE PES Innovative Smart Grid Technologies*, no. 4, pp. 1–7, 2011.
- [7] C. Keerthisinghe, G. Verbič, and A. Chapman, "A Fast Technique for Smart Home Management: ADP with Temporal Difference Learning," *IEEE Transactions on Smart Grid*, 2016.
- [8] J. Widen, A. Nilsson, and E. Wackelgard, "A combined Markov-chain and bottom up approach to modelling of domestic lighting demand," *Energy and Buildings*, vol. 41, pp. 1001–1012, 2009.
- [9] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high resolution energy demand model," *Energy and Buildings*, vol. 42, no. 10, pp. 1878–1887, 2010.
- [10] J. Page, D. Robinson, N. Morel, and J. Scartezzini, "A generalised stochastic model for the simulation of occupant presence," *Energy and Buildings*, vol. 40, pp. 83–98, 2008.
- [11] E. McKenna and M. Thompson, "High-resolution stochastic integrated thermal-electric domestic demand model," *Applied Energy*, vol. 165, pp. 445–461, 2016.
- [12] M. Hofmann, S. Riechelmann, C. Crisosto, R. Mubarak, and G. Seckmeyer, "Improved Synthesis of Global Irradiance with One-Minute Resolution for PV System Simulations," *International Journal of Photoenergy*, vol. 165, pp. 445–461, 2017.
- [13] J. Bright, C. Smith, P. Taylor, and R. Crook, "Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data," *Solar Energy*, vol. 115, pp. 229–242, 2015.
- [14] Ausgrid, "Smart Meter Infrastructure: SGSC Technical Compendium," Tech. Rep., 2014.
- [15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.
- [16] Y. Raykov, A. Boukouvalas, F. Baig, and M. Little, "What to do when k-means fails: a simple yet principled alternative algorithm," *PLoS ONE*, 2016.
- [17] D. Jenkins, S. Patidar, and S. A. Simpson, "Synthesising electrical demand profiles for UK dwellings," *Energy and Buildings*, vol. 76, pp. 605–614, 2014.
- [18] T. Power and G. Verbič, "A nonparametric bayesian model for forecasting residential solar generation," in *Australasian Universities Power Engineering Conference (AUPEC)*, 2017.
- [19] D. Azuatalam, A. C. Chapman, and G. Verbič, "Probabilistic Impact Assessment of Network Tariffs in Low Voltage Distribution Networks," *arXiv*, 2018, arXiv:1810.02013.
- [20] Y. Ma, G. Verbič, and A. C. Chapman, "Probabilistic Assessment of PV-Battery System Impacts on LV Distribution Networks," 2018, arXiv:1809.07488.