Probabilistic Peak Load Estimation in Smart Cities Using Smart Meter Data

Mingyang Sun, *Member, IEEE*, Yi Wang, *Student Member, IEEE*, Goran Strbac, *Member, IEEE*, and Chongqing Kang, *Fellow, IEEE*

***Abstract*—Adequate capacity planning of substations and feeders primarily depends on an accurate estimation of the future peak electricity demand. Traditional coincident peak demand estimation is carried out based on the empirical metric, after diversity maximum demand, indicating individual peak consumption levels and demand diversification across multiple residents. With the privilege of smart meters in smart cities, this paper proposes a data-driven probabilistic peak demand estimation framework using fine-grained smart meter data and sociodemographic data of the consumers, which drive fundamental electricity consumptions across different categories. In particular, four main stages are integrated in the proposed approach: load modeling and sampling via the proposed variable truncated R-vine copulas method, correlation-based customer grouping, probabilistic normalized maximum diversified demand estimation, and probabilistic peak demand estimation for new customers. Numerical experiments have been conducted on real demand measurements across 2639 households in London, collected from Low Carbon London project's smart-metering trial. The mean absolute percentage error and the pinball loss function are used to quantitatively demonstrate the superiority of the proposed approach in terms of the point estimate value and the probabilistic result, respectively.**

***Index Terms*—Coincident peak demand, distribution network planning, probabilistic estimation, R-vine copulas, smart meter.**

I. Introduction

IN FUTURE smart cities, increasingly more smart buildings or parks will be expanded or planned starting from scratch [1]. Accurate peak load estimation is the key driver for determining the capacities of electricity power delivery equipment such as substations and feeders. A balance between reliability and economy should be achieved by optimally matching the demand and supply. Underestimating the peak demand will result in undersized assets and an inability to service load during some periods. Meanwhile, overestimating the peak demand can lead to substantially cost-inefficient decisions, given that the same level of reliability could be provided with less expensive assets of reduced rating [2].

The electricity consumption behaviors of individual consumers exhibit high randomness and volatility [3]. The total coincident peak load is more regular and stable with a larger number of consumers that are aggregated [4]. Therefore, the peak load estimation for a single or multiple buildings supplied by a low-voltage substation and a feeder is much more challenging than for larger power systems. For small groups of consumers, it is imperative to consider the diversity of electricity consumption behaviors of individual consumers. In fact, fundamental electricity consumptions are driven by the different categories of customers (e.g., income levels and the number of occupants). For example, in [5], household size is shown to be clearly correlated with the maximum peak demand and can be a useful proxy for sizing connections of individual households.

In the literature, several metrics have been proposed to measure the demand diversity and to estimate the peak demand for the sizing and siting of substations. In particular, after diversity maximum demand (ADMD) is one of the most widely used metrics, defined as the maximum coincident peak demand per consumer when the number of consumers approaches to infinity [6]. In addition, the diversity factor, defined as the ratio between the sum of individual peak demand in a group of customers and the coincident peak demand of this group, has been used for peak demand estimation in [7]. Other metrics such as the coincidence factor [8] and the conversion factor have also been applied for network planning. It is constructive to note that most of the above-mentioned metrics can only provide a single estimation value obtained based on heuristic formulation and engineering judgment, which is not suitable for the applications considering uncertainties.

The uncertainties of the electricity load are generally depicted by a probabilistic distribution. The skewness of the distribution of typical load current is fitted by a beta distribution in [9]. Both the Weibull distribution and the log-normal distribution are used to model the probability distributions of individual households in [10], where the goodness of fit is quantified by the Kolmogorov–Smirnov test. In addition, sampling and

Manuscript received October 16, 2017; revised January 5, 2018; accepted January 25, 2018. Date of publication February 28, 2018; date of current version September 28, 2018. This work was supported in part by the National Key R&D Program of China under Grant 2016YFB0900100 and in part by the Engineering and Physical Sciences Research Council under GrantEP/N03466X/1 and Grant EP/N030028/1. *(Corresponding author: Mingyang Sun.)*

M. Sun and G. Strbac are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: mingyang.sun11@imperial.ac.uk; g.strbac@imperial.ac.uk).

Y. Wang and C. Kang are with the State Key Laboratory of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: wangyi14@mails.tsinghua.edu.cn; cqkang@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2018.2803732

simulation can be considered as effective approaches to estimate future coincident peak demand.

For example, a Monte Carlo simulation model is established for consumers considering the statistical spread of demand in [11]. The consumption data are sampled from a gamma distribution. In [5], a random sampling method based on a smart meter and demographic data is applied to calculate the ADMA of nine types of consumers. However, the accuracy of the proposed sampling method has not been verified. In addition, the software HOMER is used in [12] to simulate individual household load and then to investigate the diversity of these households. The results show that for a minigrid system, there should be at least 50 households to guarantee the manageability of the demand variance and the economy of the whole system.

With the widespread adoption of smart meters, which provide more fine-grained electricity consumption data of individual consumers [13], substantial efforts have been devoted to the applications of smart meter data including load profiling [14], anomaly detection [15], demand response program implementation [3], phase detection [16], and load forecasting [17]. A comprehensive review on smart meter data analytics is available in [18]. In contrast to traditional load forecasting, which is generally applied to assist the system operation by predicting the future short-term electricity consumptions of the existing customers based on historical data, the coincident peak demand estimation problem has two characteristics: 1) it aims to predict the coincident peak demand of future customers without having existing data at the time of connection to the network; and 2) it is essentially a long-term forecasting problem for sizing infrastructures.

Although various approaches have been proposed in the existing literature for estimating the coincident peak demand, most of the previous works are focused on empirical methods. The influx of smart meter data introduces an intuitive question: *Is it possible to further improve the peak load estimation accuracy by making full use of fine-grained smart meter data and sociodemographic information of consumers*? A few studies have begun to conduct peak load estimation based on smart meter data. For example, a clusterwise weighted constrained regression-based peak load estimation method for a low-voltage substation is proposed in [19], where the contribution factor is developed. The accuracy and stability of this method have been verified using cross validation.

In this paper, we propose a data-driven probabilistic peak load estimation approach based on smart meter data and sociodemographic data. First, the complex nonlinear dependence structures among different consumers are modeled via the proposed variable truncated R-vine copulas (VTRC) method for sampling-based data augmentation. Subsequently, the metrics ADMD and normalized maximum diversified demand (NMDD) are estimated following a consumer grouping procedure, which aims to enhance the efficiency of selecting a determined number of customers at random. Finally, the uncertainties of future peak load are described by a series of quantiles. Note that, in this paper, we do not consider the losses in distribution systems and focus on proposing a novel probabilistic peak load estimation method. In the future work, the accuracy of the estimated coincident peak load can be further improved with the consideration of the losses in the network. As stated above, the key contributions of this paper can be summarized as follows.

1) A composite data-driven probabilistic peak demand estimation framework is proposed based on the information derived from high-resolution smart meter measurements and demographic data that will be useful for designing new buildings in smart cities.
2) A novel VTRC method is proposed to model and sample the high-dimensional demand data with large computation speedups, while capturing their complex nonlinear dependencies, particularly the tail dependence.
3) A correlation-based customer grouping is proposed to be conducted before the NMDD estimation procedure for efficiently selecting the given number of customers.
4) The superior performance of the proposed method has been demonstrated through comparisons with the other tested methods, as indicated by the lower estimation errors. Additionally, the results demonstrate the importance of using the demographic data when designing new buildings with different categories of customers.

The remainder of this paper is organized as follows. Section II formally defines the peak load estimation problem. Section III introduces the framework and technical details of the proposed approaches. Section IV presents the evaluation metrics and the methods to be compared in the case studies. Section V conducts numerical experiments on the Low Carbon London (LCL) dataset. Finally, the conclusion is presented in Section VI.

II. Problem Statement

In smart cities, it is imperative to perform an accurate future coincident peak demand estimation for the electricity network design, which has the following main challenges.

1) Different types of properties with various future customers exhibit different consumption behaviors. Although these customers can be categorized based on their demographic data, for each category, it is still challenging to estimate the coincident peak demand without having existing data at the time of connection to the network.
2) The demand diversity among customer loads significantly increases the difficulties in estimating the group peak demand, particularly when the number of new connected customers n is considerably lower than the number of existing ones in the network. Note that demand diversification refers to the effect that the coincident peak demand exhibits reduced sensitivity to the attributes of individual consumers with an increasing number of n, and vice versa. This is because not all of the customers consume their peak demand simultaneously.

Mathematically, the problem can be defined as follows. Given that there are G customer categories defined by the demographic data (e.g., household occupancy and wealth level) and N new customers need to be connected to the network, for each category g, the number of new connections is denoted by n_g, where $\sum_{g=1}^{G} n_g = N$. In addition, let m_g denote the number of existing customers in category g with smart meter data and

demographic data for $g = 1, ..., G$. The target of this problem is to estimate the total coincident peak demand

$$\hat{C}_N = \sum_{g=1}^{G} \hat{C}_{n_g}. \tag{1}$$

For each category, we estimate $\hat{C}_{n_g}$ via the metric ADMD, which is traditionally defined as the coincident peak electrical demand per customer when n approaches infinity. In this work, we extend the concept of ADMD to be a function of n, denoted by ADMD^n. Consequently, based on the smart data of existing customers $D_g \in \mathbb{R}^{T\times m_g}$, ADMD^n for $n \in \{1, m_g\}$ can be calculated for each category as follows:

$$\text{ADMD}^n = \max_{t=1,...,T} \left(\sum_{i=1}^{n} D_{t,i} \right) /n. \tag{2}$$

Note that it is computationally impractical to calculate ADMD^n for all the $n!/n!(m_g - n)!$ possible household combinations, particularly when $n << m_g$. To this end, it is more efficient to randomly select n customers out of m_g as many as required. For each category, we obtain the ADMD^n curve from $n = 1$ to m_g and then normalize it based on the highest value of each curve. In this way, the NMDD value NMDD^n in [0, 1] can be further used to estimate $\hat{C}_{n_g}$ as follows:

$$\hat{C}_{n_g} = n_g \times \text{NMDD}^{n_g} \times D_g^{0.95\max} \tag{3}$$

where $D_g^{0.95\max}$ represents the 95th percentile peak demand of D_g. Finally, the total coincident peak demand of the future N customers, $\hat{C}_N$, can be obtained by calculating the sum of $\hat{C}_{n_g}$ for $g = 1, ..., G$. Note that the output $\hat{C}_N$ is a probabilistic distribution rather than a single-point value due to the large number of combinations that we select while calculating ADMD^n.

III. Methodologies

A. Proposed Framework

As illustrated in Section II, the challenges related to the coincident peak demand estimation are the unknown data of future customers and the influence of demand diversity among customer loads. To address these challenges, a novel probabilistic peak demand estimation framework shown in Fig. 1 is introduced in this paper. This framework consists of four main stages.

1) *Load modeling and sampling stage:* Given the input smart meter data $D_g \in \mathbb{R}^{T\times M_g}$, the existing customers of category g are randomly partitioned into training customers $D_g^{\text{train}} \in \mathbb{R}^{T\times m_g}$ and test customers $D_g^{\text{test}} \in \mathbb{R}^{T\times n_g}$, where $m_g + n_g = M_g$ and $m_g \approx 0.8 \times M_g$. Subsequently, an accurate statistical model is constructed based on D_g^{train} at the VTRC modeling stage. Given the input number of samples T_s, the output of the VTRC sampling stage is $\hat{D}_g^{\text{train}} \in \mathbb{R}^{T_s\times m_g}$.
2) *Customer grouping stage:* This stage can be regarded as a pregrouping stage for the probabilistic ADMD^n estimation stage. In particular, given the number of clusters K, all the m_g training customers are clustered based on their nonlinear correlations, quantified by the calculated Spearman's rank correlation coefficient matrix R_g. The output of this stage is the constructed K groups of customers, where $\hat{D}_{\text{train}}^k \in \mathbb{R}^{T_s\times d_k}$, for $k = 1, ..., K$.
3) *Probabilistic NMDD estimation stage:* In this stage, we calculate the distribution of ADMD^n for each $n \in \{1, .., m_g\}$ based on the constructed groups and obtain the probabilistic ADMD^n curve. Then, the probabilistic NMDD^n curve is obtained by normalizing the ADMD^n curve.
4) *Probabilistic peak demand estimation stage for new customers:* Using the probabilistic NMDD^n curve, given the number of future customers n_g, a probabilistic coincident peak demand estimation result $\hat{C}_{n_g}$ can be computed using (3). Compared with the actual coincident peak demand C_{n_g}, the performance of the proposed method can be evaluated via quantitative metrics. In particular, the mean absolute percentage error (MAPE) and the pinball loss function are used to assess the mean value and the distribution of $\hat{C}_{n_g}$, respectively.

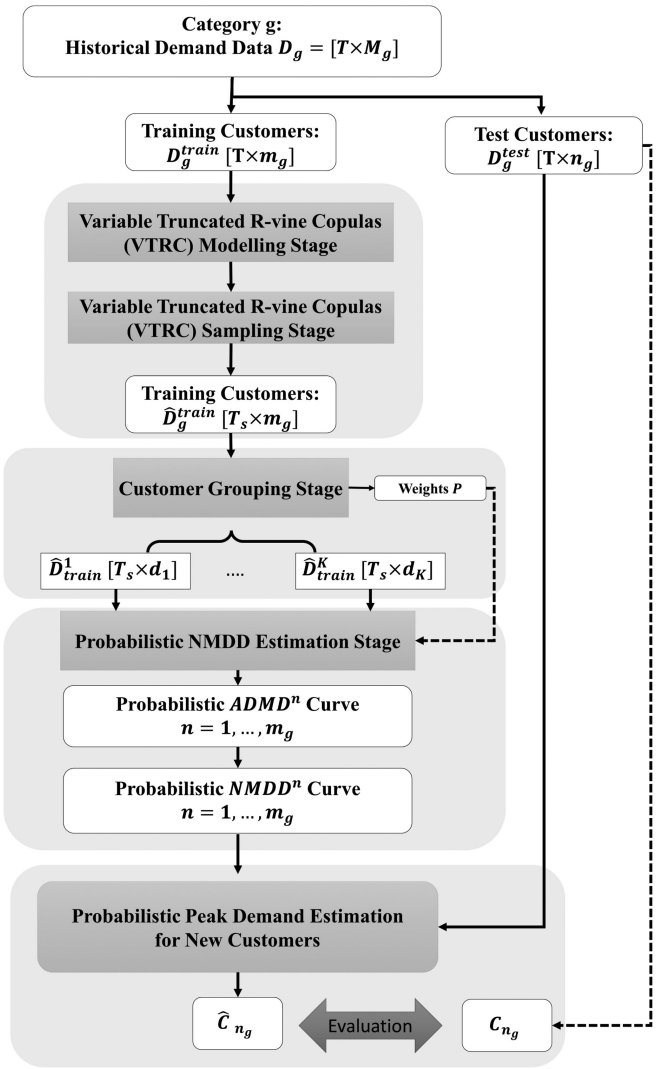


Fig. 1. Proposed probabilistic coincident peak demand estimation framework for a single category g.

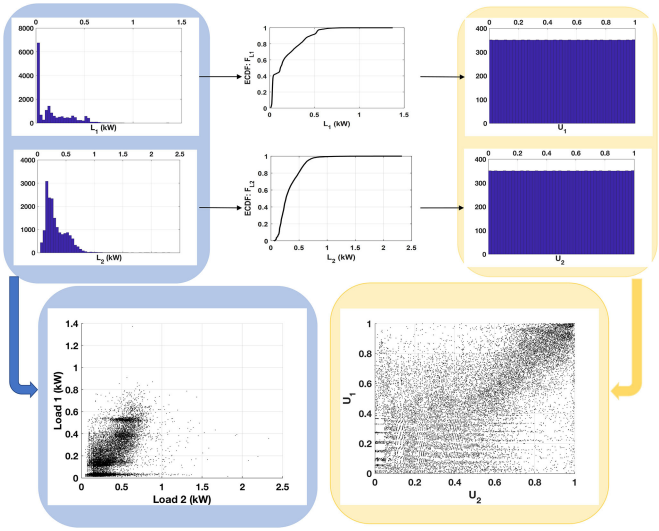


Fig. 2. Bivariate example of histograms of marginals and scatter plots of active energy consumption data (kW) for two randomly selected households from the LCL smart meter trail, measured between January 2013 and December 2013. Additionally, this figure illustrates the procedure for transforming the data from original domain kW to uniform domain [0, 1] through ECDFs.

B. Load Modeling and Sampling

Given the training customers' data $D_g^{\text{train}} \in \mathbb{R}^{T\times m_g}$, the first step of the proposed framework is to construct a statistical model for the residential electricity demand. The aim of this stage is to understand current consumers' behavior and to model their dependence structure by interpolating and extrapolating the historical smart meter measurements. Then, the constructed model can be used to generate as many samples as required to represent the potential consumption behavior for future customers with unknown data. To illustrate the challenges of modeling high-dimensional smart meter data, an example of the demand data of two customers randomly selected from the Adverse and 1 occupant households is presented in Fig. 2. It can be observed that the demand data exhibit highly non-Gaussian marginal distributions and complex nonlinear dependence structure between different customers even though they are in the same category. To address these problems, a novel VTRC model is proposed to perform the load modeling and sampling procedure for each category's households while retaining the empirical marginal distributions and capturing the multivariate nonlinear dependencies.

1) Proposed VTRC Modeling Approach: Let f, F, c, and C denote the probability density function (PDF) and the cumulative distribution function and their copula versions, respectively. Consider m_g random variables $D_g^{\text{train}} = (D_1, ..., D_{m_g}) \in \mathbb{R}^{T\times m_g}$ with marginal density functions $f_i(d_i)$ and distribution functions $F_i(s_i)$ for $i = 1, ..., m_g$. According to Sklar's theorem [20], the joint PDF can be expressed as follows:

$$f(d_1, d_2, ..., d_{m_g}) = \left(\prod_{i=1}^{m_g} f_i(d_i)\right) \times c_{1\cdots m_g}(F_1(d_1), ..., F_{m_g}(d_{m_g})) \quad (4)$$

where the copula density function $c_{1...m_g} : [0, 1]^{m_g}$ represents the dependence structure between uniform random variables $\{U_1, ..., U_{m_g}\} = \{F_1(D_1), ..., F_{m_g}(D_{m_g})\}$. Note that the empirical cumulative distribution function (ECDF) and its inverse version ECDF^{-1} are used to model the marginal distributions and to transform it between the actual domain (i.e., kW) and the [0, 1] domain, as shown in Fig. 2. Regarding the correlation among the stochastic variables, various classes of copula functions are capable of describing complex dependence structures, but most of them are limited to the bivariate case. To this end, the pair-copula construction method was proposed in [21] to decompose a multivariate copula into the product of a cascade of bivariate copulas, introducing flexibility in capturing complex dependence structures, particularly different characters of tail dependence.

In this work, we consider one of the most flexible graphical models, regular vine (R-vine), which consists of a nested set of $m_g - 1$ trees $\Upsilon = (T_1, ..., T_{m_g-1})$ such that the edges E_j of tree T_j become the nodes N_{q+1} of tree T_{q+1}, for $j = 1, ..., m_g - 1$. Based on the definition provided in [22], for m_g random variables, Υ of an R-vine on m_g random variables needs to satisfy the following conditions.

1) The first tree T_1 consists of nodes $N_1 = 1, ..., m_g$ and a set of edges denoted by E_1.
2) T_i consists of edges E_i and nodes $N_i = E_{i-1}$, for $i = 2, ..., m_g$.
3) For $i = 2, ..., m_g - 1$, $\{j, k\} \in E_i$ must hold that $\#(j \cap k) = 1$, where $j = \{j_1, j_2\}$ and $k = \{k_1, k_2\}$.

Note that condition 3 represents that two nodes in T_{q+1} are only connected by an edge if they share a common node in T_q. For a regular vine Υ, let $S_e = \{\nu \in N_1 | \exists e_i \in E_i, i = 1, ..., m_g - 1, \text{ with } \nu \in e_1 \in ... \in e_{m_g-1} \in e\}$ denoting the complete union of an edge $e = \{j, k\} \in E_q$ in tree T_q. The conditioning set and the conditioned sets associated with edge $e = \{j, k\}$ are defined as $\Psi_e := S_j \cap S_k$ and $\{\Omega_{e,j} = S_j \backslash \Psi_e, \Omega_{e,k} = S_k \backslash \Psi_e\}$, respectively, where $(-)\backslash(*) := (-) \cap (*)^C$ and $(*)^C$ is the complement of $(*)$. Following the above-mentioned definitions, the density function $f(d_1, ..., d_{m_g})$ can be decomposed as follows:

$$\prod_{i=1}^{m_g} f_i(d_i) \times \prod_{i=1}^{m_g-1} \prod_{e\in E_i} c_{\Omega_{e,j},\Omega_{e,k}|\Psi_e}(F_{\Omega_{e,j}|\Psi_e}(.), F_{\Omega_{e,k}|\Psi_e}(.)) \quad (5)$$

where $e = \{j, k\}$ and $F_{\Omega_{e,j}|\Psi_e}(.) = F_{\Omega_{e,j}|\Psi_e}(d_{\Omega_{e,j}}|d_{\Psi_e})$, which is denoted as an h-function. The detailed formulation of h-function for R-vine is presented in [22].

Note that there are a series of nodes–edges–trees combinations that meet the R-vine conditions. Therefore, the key challenges related to constructing an appropriate R-vine include the following.

1) Select the optimal sets of $\{\Omega_{e,j}, \Omega_{e,k}|\Psi_e\}$ for all edges.
2) For each edge of the selected trees, determine the optimal bivariate copula family.
3) Estimate corresponding parameters for each bivariate copula.

Regarding the selection of the bivariate copula, the Akaike information criterion (AIC) is used in this work to choose the

best-fit copula, which is indicated by the smallest AIC values. In addition, a Kendall's τ-based sequential method, an automated strategy, was proposed in [23] to select the R-vine tree that maximizes the sum of absolute empirical Kendall's τ. The output of the sequential method is an R-vine specification matrix M. The detailed definition and explanation of the specification matrix and sequential method are presented in [23]. Note that although the sequential method cannot ensure a global optimum, the selected model can be regarded as a reasonable one because it models the original variables and the following conditional variables from higher to relatively lower dependencies. When the variables in higher tree exhibit independence, the truncation method and the simplification method proposed in [22] can be used to accelerate the modeling procedure. However, the number of possible R-vines on m_g variables (i.e., $m_g!/2 \times 2^{\binom{m_g-2}{2}}$) and the number of estimated pair copulas (i.e., $n_f \times m_g(m_g-1)/2$, where n_f is the number of considered bivariate copula families) rapidly increase with m_g, resulting in an impractical issue for very high-dimensional cases (e.g., 10 000 households) due to the high computational complexity.

To fundamentally accommodate the problem of high dimensionality, a variable truncated R-vine copula method is proposed in this work to make the R-vine modeling and sampling procedure executed in a lower dimensional space constructed with important features. In [24], locality preserving projection (LPP) has been demonstrated as an efficient dimensionality reduction technique for C-vine and D-vine copulas, which are two representative subclasses of R-vine copulas. In particular, LPP is a sparse-spectral linear feature extraction technique that aims to identify a lower dimensional dataset that preserves the local neighborhood structure of the original data manifold [25]. Note that the number of reduced dimensions is determined by the user-defined information retainment threshold in [0, 1], which is defined based on eigenvalues for quantifying the proportion of variance that can be retained using a reduced number of transformed variables. Also, the issues of missing data and outliers could be addressed by fitting the data to a sophisticated statistical model via the proposed copula-based approach [14]. To summarize, the proposed VTRC modeling and sampling algorithm is outlined in Algorithm 1.

Algorithm 1: VTRC Modeling and Sampling.

Input: Historical demand data: $D_g^{\text{train}} = [D_1, ..., D_{m_g}] \in \mathbb{R}^{T \times m_g}$, Information retainment threshold: IR, Number of samples: T_s.

Output: Sampled demand data: $\hat{D}_g^{\text{train}} = [\hat{D}_1, ..., \hat{D}_{m_g}] \in \mathbb{R}^{T_s \times m_g}$.

Step 1: *Transform D_g^{train} from the original domain to $V = [V_1, ..., V_{m_g}]$ in the rank-uniform domain $[0,1]^{m_g}$ via the ECDFs of D_g^{train}:*

$V_i = F_i(D_i)$, for $i = 1, ..., m_g$

Step 2: *Given the input parameter IR, perform LPP to extract important features $X_g \in \mathbb{R}^{T_s \times l_g}$, where $l_g < m_g$. Also, the solution matrix A, where $X = A^T V$, needs to be stored for the sampling procedure*:

$[X, A] = \text{LPP}(V, IR)$

Step 3: *Perform the uniform transformation again through the ECDFs of X_g and obtain $U = [U_1, ..., U_{l_g}]$ in the unit domain $[0,1]^{l_g}$.*

$U_i = F_i(X_i)$, for $i = 1, ..., l_g$

Step 4: *Perform the sequential method to determine the optimal R-vine specification Υ, indicated by matrix $M \in \mathbb{R}^{l_g \times l_g}$, the best-fit bivariate copula families, stored in matrix $B \in \mathbb{R}^{(l_g-1)\times(l_g-1)}$ as well as their estimated parameters $\Theta \in \mathbb{R}^{(l_g-1)\times(l_g-1)}$.*

$[M, B, \Theta] = TheSequenceMethod(U)$

Step 5: *Given the required number of samples T_s, simulate the R-vine specification to generate samples $U_s \in \mathbb{R}^{T_s \times l_g}$.*

$\hat{U} = SimulationRvine(M, B, \Theta, T_s)$

Step 6: *Transform the generated samples $\hat{U}$ from the uniform domain back to the domain of extracted features X, thus obtaining $\hat{X} \in \mathbb{R}^{T_s \times l_g}$ via the inverse empirical distribution functions (ECDF $^{-1}$) of X.*

$\hat{X}_i = F_i^{-1}(\hat{U}_i)$, for $i = 1, ..., l_g$

Step 7: *Back-project $\hat{X}$ from the lower dimensional space to the original dimensional space and obtain $\hat{V} \in \mathbb{R}^{T_s \times m_g}$. Finally, obtain the samples $\hat{D}_g^{\text{train}} \in \mathbb{R}^{T_s \times m_g}$ via the ECDF $^{-1}$ s of D_g^{train}.*

$\hat{V} = AX$

$\hat{D}_i = F_i^{-1}(\hat{V}_i)$, for $i = 1, ..., m_g$

C. Customer Grouping

When calculating the ADMD^n for n customers at the next stage, the existence of demand diversity necessitates selecting a large amount of replicates that consist of different combinations of n customers. Instead of randomly selecting n customers from all the m_g trained customers, it may be more efficient to select the customers from different groups of customers, each of which has similar consumption patterns. To this end, the customer grouping stage is proposed in the framework to construct K groups based on the correlations among different customers. It is constructive to mention that the reason for employing correlation-based clustering, rather than traditional distance-based clustering, is because the occurrence of the coincident peak demand is highly related to the correlations of their electricity consumption behaviors. For example, if a group of customers are highly correlated, then they will have very similar consumption patterns across different points in time, and the influence of demand diversity will be decreased when calculating their coincident peak demand. Given the sampled demand data $\hat{D}_g^{\text{train}} \in \mathbb{R}^{T_s \times m_g}$, the first step of this stage is to calculate the correlation matrix $R = (r_{i,j})_{i,j=1,...,m_g}$. According to the definition of Spearman's correlation coefficient, which is one of the most widely used nonparametric measures of rank correlation, for each pair of $\hat{D}_i$ and $\hat{D}_j$,

Algorithm 2: Probabilistic NMDD Estimation.

Input: Total number of customers in category g: m_g; Number of replicates: N_r; Constructed groups: $(\hat{D}^k_{\text{train}})_{k=1,\dots,K}$; Probability for each group: $(P_k)_{k=1,\dots K}$.

Output: Probabilistic NMDD^n curve, $n = 1, ..., m_g$: NMDD

for $n = 1{:}m_g$ **do**
  for $i = 1{:}N_r$ **do**
    for $k = 1{:}K$ **do**
      $D^k_n \in \mathbb{R}^{T_s \times n_k}$: randomly select $n_k = n \times P_k$ customers from d_k customers in the kth group $\hat{D}^k_{\text{train}}$.
    end for
    $D_{n,i} = \{D^1_n, ..., D^K_n\} \in \mathbb{R}^{T_s \times n}$
    $\text{CP}_{n,i} = \max\limits_{t \in \{1,..,T_s\}} \{\sum_{j=1}^{n} D^j_{t,i}\}$: calculate the coincident peak demand for replicate i.
  end for
  $\text{CP} = \{\text{CP}_{n,1},...,\text{CP}_{n,N_r}\}^T \in \mathbb{R}^{N_r \times 1}$
  $\text{ADMD}^n = \text{CP/n}$
end for
$\text{ADMD} = \{\text{ADMD}^1,...,\text{ADMD}^{m_g}\} \in \mathbb{R}^{N_r \times m_g}$
for $i = 1{:}n$ **do**
  $\text{NMDD}^n_i = \text{ADMD}^n_i / \max(\text{ADMD}^1_i)$
end for
$\text{NMDD} = \{\text{NMDD}^1,...,\text{NMDD}^{m_g}\} \in \mathbb{R}^{N_r \times m_g}$

we have

$$r_{i,j} = \text{cov}(\text{rank}(\hat{D}_i), \text{rank}(\hat{D}_j))/(\sigma_{\text{rank}(\hat{D}_i)}\sigma_{\text{rank}(\hat{D}_j)}) \quad (6)$$

where rank(.) indicates the rank variable, cov(rank($\hat{D}_i$), rank($\hat{D}_j$)) is the covariance of the rank variables, and $\sigma_{\text{rank}(.)}$ represents the standard deviation of the rank variable. Subsequently, an agglomerative hierarchical clustering method with single linkage is performed to cluster the customers based on the new "distance" matrix $1 - R$. The output of this stage is $\hat{D}^k_{\text{train}} \in \mathbb{R}^{T_s \times d_k}$ with corresponding probability $P_k = d_k/m_g$ for $k = 1, ..., K$. The reason of employing the agglomerative hierarchical clustering method with single linkage to establish different groups of customers can be concluded as follows.

1) Hierarchical clustering method can handle nonspherical data.
2) The constructed hierarchical clusters are independent to their initial points, thus leading to its deterministic nature.
3) No prior knowledge of K is required.

D. Probabilistic NMDD Estimation

Using (2), the distribution of ADMD^n for each $n \in \{1, .., m_g\}$ can be calculated based on the constructed groups, thus obtaining the probabilistic ADMD^n curve. Then, the probabilistic NMDD^n curve is obtained by normalizing the ADMD^n curve. The calculation procedure is outlined in Algorithm 2.

E. Probabilistic Peak Demand Estimation

Given n_g new customers from category g, the distribution of the coincident peak demand $\hat{C}_{n_g} \in \mathbb{R}^{N_r \times n_g}$ can be estimated using (3) based on the obtained probabilistic NMDD curve. Compared with the actual coincident peak demand C_{n_g}, the performance of the proposed framework can be assessed via the evaluation metrics, introduced in the next section.

IV. Evaluation Metrics and Comparisons

A. Evaluation Metrics

1) Mean Absolute Percentage Error: In this work, we use the MAPE to evaluate the performance of the estimation model in terms of the expectation value of the estimated probabilistic distribution:

$$\text{MAPE}(C_n, \bar{\hat{C}}_n) = 100\% \times |C_n - \bar{\hat{C}}_n|/C_n \quad (7)$$

where C_n and $\bar{\hat{C}}_n$ are the actual coincident peak demand and the expected estimated coincident peak demand for the tested n customers, respectively.

2) Pinball Loss Function: In general, three main factors, namely, calibration, sharpness, and reliability, should be evaluated for the probabilistic estimation model [26]. The pinball loss function is considered in this work with the benefits of providing a comprehensive metric value for all three factors. Let C_n and $\hat{C}_{n,q}$ denote the actual coincident peak demand and the estimated coincident peak demand at the qth percentile for the tested n customers, respectively. Then, the pinball loss function can be expressed as

$$\text{Pinball}(C_n, \hat{C}_{n,q}, q) = \begin{cases} (\hat{C}_{n,q} - C_n)(1-q), & \hat{C}_{n,q} > C_n \\ (C_n - \hat{C}_{n,q})q, & \hat{C}_{n,q} < C_n. \end{cases} \quad (8)$$

In this work, the average of all the Pinball$(C_n, \hat{C}_{n,q}, q)$ s for $q = 0.01, 0.02, ..., 0.99$ is used to evaluate the overall performance of the probabilistically estimated coincident peak demand of n customers. Note that a lower pinball loss indicates a better quantile model.

B. Comparisons

To illustrate the benefits of each stage of the proposed framework, a series of tested methods are compared in this paper. Note that M0 is a point estimation method, whereas M1, M2, M3, and M4 are all probabilistic methods.

1) M0: Empirical ADMD: Traditionally, the coincident peak demand of future customers without existing data is generally estimated using empirical ADMD, which been widely used in distribution network operator (DNO)'s planning guidelines [5]. Given that n new customers will be connected to the network, $\hat{C}_n$ can be calculated via the following equation:

$$\hat{C}_n = 0.7 \times n \times \text{ADMD} \times (1 + 12/(\text{ADMD} \times n)). \quad (9)$$

According to the network design manual [27], we set ADMD = 2 kW for properties with up to four bedrooms and gas heating. Note that the scaling factor 0.7 is multiplied to take the effect of demand diversification into account.

TABLE I
TESTED METHODS

	Empirical	Historical	VRTC	Customer Grouping
M0	√	–	–	–
M1	–	√	–	–
M2	–	√	–	√
M3	–	–	√	–
M4	–	–	√	√

2) M1: Historical NMDD^n: M1 estimates the coincident peak demand based on NMDD^n curves, obtained by using the historical smart meter data with random customer selection approach.

3) M2: Historical NMDD^n and Customer Grouping: Based on M1, M2 is appended with the proposed correlation-based customer grouping stage.

4) M3: VTRC NMDD^n: A large amount of samples, generated via the proposed VTRC method, are used to estimate $\hat{C}_n$ based on the framework of M1.

5) M4: VTRC NMDD^n and Customer Grouping: M4 is the proposed probabilistic peak demand estimation method that integrates both VTRC sampling and correlation-based customer grouping.

To summarize, all the tested methods are concluded in Table I.

V. CASE STUDY

A. Data Description

The LCL smart meter trials provide a valuable opportunity to understand residential electricity consumption behaviors and to assess the benefits from exploiting smart metering for distribution network design. In the LCL project, Landis and Gyr (L+G) E470 electricity meters were installed in 2639 residential households across the Mayor of London's Low Carbon Zones and the London Power Networks distribution network license area operated by U.K. Power Networks [5]. Specifically, the Engineering Instrumentation Zones of the LCL trial include the areas of Brixton, Merton, and Queen's Park. Regarding the collected data, the LCL demand dataset consists of 17 520 half-hourly measurements of demand across 2639 customers in kW for a full calendar year from January 1, 2013, to December 31, 2013. In addition, various socioeconomic data of the participating households were also recorded in the dataset. In this paper, we focus on the data pertaining to household occupancy (i.e., the number of people living in the property) and wealth level (i.e., determined based on mapping all participating households to ACORN groups). In particular, we consider three types of occupancy, namely, 1 occupant, 2 occupants, and 3 occupants, as well as three wealth classes, namely, Adverse, Comfortable, and Affluent in an increasing order. Consequently, nine categories (i.e., $G = 9$) are defined as the combinations of occupancy and wealth level, and the number of customers for each category is given in Table II. Note that more detailed descriptions of the tested LCL dataset can be found in the literature (e.g., [28]).

TABLE II
NUMBER OF HOUSEHOLDS ACROSS NINE CUSTOMER CATEGORIES

	1 occupant	2 occupants	3 occupants
Adverse	315	278	234
Comfortable	240	304	214
Affluent	431	400	223

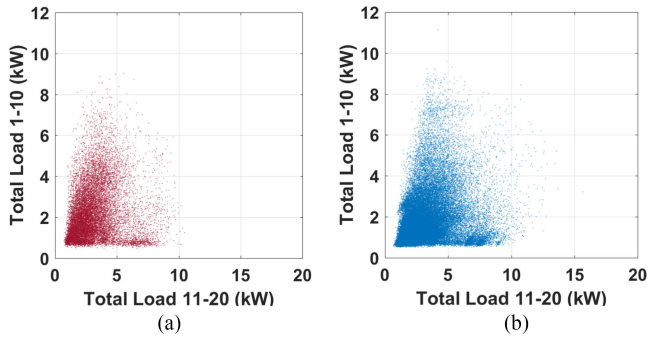


Fig. 3. Scatter plots of the total load of the randomly selected customers 1–10 and customers 11–20. (a) Historical data. (b) Sampled data.

B. Visual Comparison

An example of 20 randomly selected customers from Adverse1 is given in this part to visually inspect the superior performance of the proposed R-vine copulas sampling method. Given the historical dataset of size $17\,520 \times 20$, the output of the proposed VTRC modeling and sampling stages is a larger set of sampled data of size $100\,000 \times 20$ (i.e., $T_s = 100\,000$) with the information retainment threshold IR $= 0.5$. In order to present and compare these two sets of 20-D data, the total load of the first ten customers (vertical axis) and the rest ten customers (horizontal axis) is shown in Fig. 3. As can be seen, the sum of the sampled load data [see Fig. 3(b)] can accurately represent the complex nonlinear dependence structure of the historical data [see Fig. 3(a)] with only 50% information retained. Also, it is important to note that more extreme values can be obtained via using the proposed sampling method, which may improve the accuracy of the final estimated coincident peak demand. Regarding the central processing unit times, the proposed VTRC method was implemented in MATLAB 2017b and run on an Intel Xeon E5-2690 PC with eight cores. For the case of IR $= 1$, which means no dimensionality reduction is performed, the total time including modeling and generating 100 000 samples is about 4429.32 s, whereas it only takes 810.66 s for IR $= 0.5$ (i.e., R-vine is performed on a 9-D dataset). Both the great capability of capturing the dependence structure and the significantly reduced computational cost demonstrate the efficiency and effectiveness of the proposed VTRC approach.

C. Results Across Different Numbers of Customers

In this section, the performance of the proposed probabilistic peak demand estimation method is evaluated based on a single category of customers, Adverse and 1 occupant, with different

TABLE III
MAPE FOR DIFFERENT NUMBERS OF TESTED CUSTOMERS (%)

	$n_g = 10$	$n_g = 20$	$n_g = 40$	$n_g = 50$	$n_g = 60$
M1	20.70	22.97	24.12	24.33	23.76
M2	17.92	21.23	23.05	23.50	24.39
M3	10.31	7.57	6.95	6.94	6.69
M4	**8.97**	**6.37**	**5.68**	**5.51**	**5.35**

TABLE IV
PINBALL LOSS FOR DIFFERENT NUMBERS OF TESTED CUSTOMERS (kW)

	$n_g = 10$	$n_g = 20$	$n_g = 40$	$n_g = 50$	$n_g = 60$
M1	0.7538	1.4111	2.7220	3.3824	4.0219
M2	0.6554	1.2903	2.5472	3.2021	3.8437
M3	0.4687	0.5682	0.8392	0.9661	1.0705
M4	**0.4429**	**0.5360**	**0.7708**	**0.8753**	**0.9640**

numbers of future customers. As illustrated in Fig. 1, all 315 customers in this category are first randomly partitioned into training customers of size $m_g = 249$ and test customers of size $n_g = [10, 20, 40, 50, 60]$. For each number of n_g, in order to obtain the distribution of the actual coincident peak demand, we randomly select 100 000 sets of customers from the whole set of customers except for the training ones. For the tested methods M3 and M4, as introduced in Section IV, the first step is to model the historical load data of the $m_g = 249$ training customers via the proposed VTRC method. Then, a large number of samples are generated via simulating the constructed model. Note that the input parameters of this stage are given as follows: number of samples $T_s = 100\,000$ and information retainment threshold for LPP IR $= 0.5$. Afterwards, for M2 and M4 that include the correlation-based consumer grouping stage, we set the number of clusters for this category to $K_g = 4$. For the stage of probabilistic NMDD estimation, we set the number of replicates $N_r = 100\,000$ for all the methods. In terms of the evaluation metrics, we evaluate the point estimation results and the probabilistic estimation results of M1, M2, M3, and M4 via the metrics MAPE and the pinball loss, shown in Tables III and IV, respectively. Note that, for each n_g, the benchmark value of the estimated peak demand is the mean value of the actual peak demand distribution when calculating the pinball loss and the MAPE. Also, for the MAPE, the point estimated peak demand is the mean value of the probabilistic estimated values.

In particular, the results presented in Tables III and IV both indicate that the proposed framework M4, consisting of the VTRC method and the correlation-based grouping strategy, exhibits superior performance to the other methods. This is evidenced by the gradually reduced MAPE and pinball loss values from historical data-based estimation to sample-based estimation, from random customer selection to clustering-based selection. In terms of the average performance across all the n_g s, when including the customer group stage, M2 has approximately 6.06% and 6.12% improvements when compared with M1 in terms of the point estimation result and the probabilistic estimation result, respectively. Additionally, the introduction of the load modeling and sampling stage in M3 presents 66.99% and 68.17% reductions in the MAPE and pinball loss values of M1. Finally, the combination of both sampling and grouping stages makes further enhancements based on M3, as indicated by a 17.09% reduced MAPE value and a 8.27% reduced pinball loss value.

In addition, Fig. 4 presents the boxplots and the PDF plots of the estimated probabilistic peak demand for each method when $n_g = [10, 20, 40, 50, 60]$. It can be observed that the sample-based methods (i.e., M3 and M4) exhibit better estimation than the historical-data-based methods (i.e., M1 and M2) in terms of the distribution of the estimated peak demand. Additionally, the range of the estimated values obtained via M3 and M4 can almost cover the actual values, whereas parts of extreme actual values are out of the range of the estimated values obtained using M1 and M2. This also demonstrates the importance of the proposed load modeling and sampling method, VTRC, which can capture the tail dependencies among various consumers and obtain more probable coincident peak demand via using the vast number of samples rather than the limited number of historical measurements.

Another critical aspect that impacts the accuracy of the estimated coincident peak demand is the time resolution of smart meter data. Theoretically, low time resolution of the smart meter data will result in an inherent underestimation of the aggregated peak load. In order to demonstrate the importance of high-resolution data, an additional case study based on lower resolution data (i.e., 1-h resolution data obtained based on the 30-min data) is conducted by use M1 to estimate the peak demand across different numbers of tested customers. Fig. 5 presents the bar plots of the MAPE values for the cases of 30-min resolution data and 1-h resolution data. Note that in this case, we do not take the absolute value of $C_n - \hat{C}_n$ when calculating the MAPE using (7). Therefore, positive and negative MAPE values represent the underestimation and the overestimation, respectively. As can be seen, in terms of the mean value of the estimated peak demand, the lower time resolution indeed leads to higher positive estimation errors than those of the high-resolution case, which indicates a more severe underestimation problem when using the dataset with 1-h resolution. However, as the proposed method is applicable to arbitrary time resolution, the accuracy of the estimated peak demand can be further improved if more fine-grained data can be employed.

D. Results for Each Category

Although the previous results have demonstrated the superior performance of the proposed approach in a single category, it is also imperative to extend the analysis to the other categories. For each category, five replicates of training and test sets are constructed by randomly partitioning the customers of each category into 80% and 20%, respectively. The parameters of the VTRC model in Section V-B are retained in this case, whereas the number of clusters at the customer grouping stage is set to different values for different categories, as follows: $K = [K_1, K_2, ..., K_9] = [5, 10, 30, 6, 20, 4, 4, 20, 20]$. Note that all the K_g s are determined by a series of performance

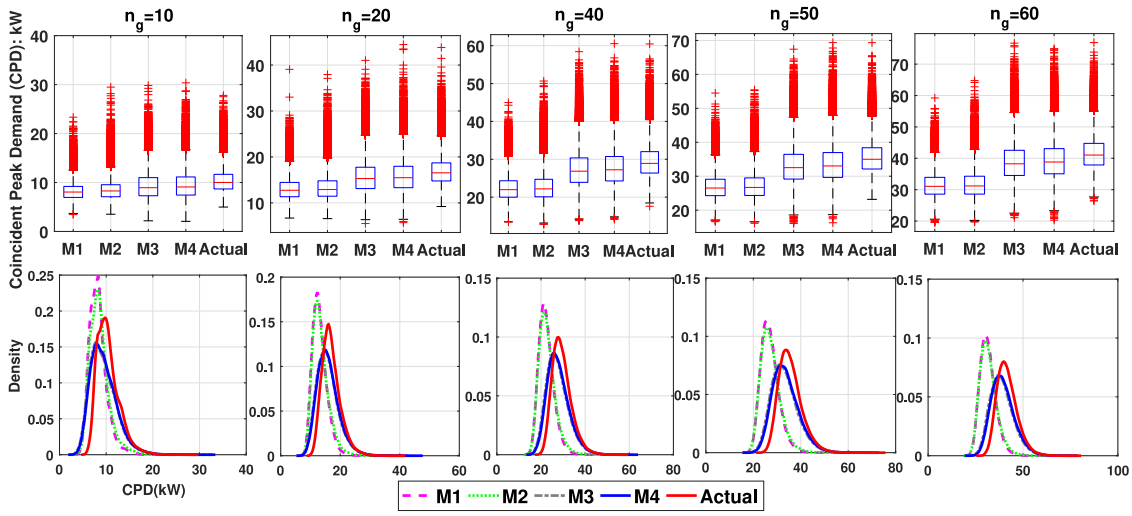


Fig. 4. Boxplots and PDF plots of the estimated probabilistic peak demand and the actual peak demand for $n_g = [10, 20, 40, 50, 60]$.

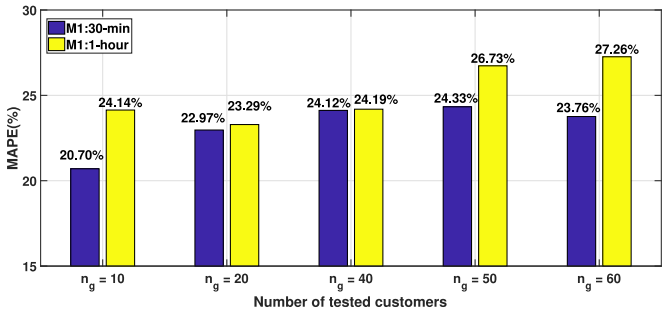


Fig. 5. MAPE for different numbers of tested customers based on the datasets of 30-min resolution (blue bar) and 1-h resolution (yellow bar).

TABLE V
PINBALL LOSS FOR EACH CATEGORY OF CUSTOMERS (kW)

	M1	M2	M3	M4
Adverse1	5.2849	5.2401	2.1653	**2.1332**
Adverse2	2.0718	2.1549	1.2651	**1.0700**
Adverse3+	7.1984	4.1119	4.0572	**3.3464**
Comfortable1	9.7770	9.4943	5.5652	**3.6345**
Comfortable2	8.2052	7.4132	6.6503	**6.2836**
Comfortable3+	4.6918	4.7407	3.3053	**3.2805**
Affluent1	7.6162	7.3931	4.7740	**4.4668**
Affluent2	3.2321	2.9902	1.9698	**1.5258**
Affluent3+	11.2875	11.2735	9.0511	**8.9089**

tests, and they may not be the optimal number for each group. For all the tested methods, the probabilistic estimation performance for each category is shown in Table V, as evaluated by the average pinball loss value of all five replicates.

As shown above, the order of the probabilistic estimation's performance across different methods is highly consistent with the results shown in Table IV. The proposed method M4 always has the lowest pinball loss values for all nine categories, whereas the categories with 3+ occupants exhibit higher pinball loss values than the other types of occupancy. Meanwhile, for most of the tested methods, the average pinball loss displays increasing values with an ascending order of wealth levels from Adverse to Affluent. Both of these results may be because for larger and wealthier households, consumers' habits present higher diversity with more types of electrical appliances (e.g., electric vehicle). Additionally, for some specific categories (e.g., Adverse2 and Comfortable3+), M2 presents slightly higher pinball loss values than M1. This issue could be caused by the inappropriate number of clusters determined in the customer grouping stage. Therefore, future work could be focused on investigating how to determine the optimal or suitable number of groups based on the performance of estimation.

To visually inspect the ADMD curve and then applied to the case of multiple categories, an example of M4 is presented in Fig. 6 that shows the maximum, minimum, and mean ADMD^n values as a function of households, for all the nine categories; the specific numbers provided on the plots are $n = [1, 5, 10, ..., 100, 150, m_g]$ households. As can be seen in the figure, for all the categories, the estimated ADMD^n values across an increased number of households exhibit the reduced sensitivity to the attributes of individual customers due to the effect of demand diversification. Additionally, in most of the cases (e.g., categories of Comfortable and categories of 2 occupants), households of increasing wealth and occupancy level exhibit higher estimated ADMD^n values for each n. It can be concluded that wealth and occupancy information are both useful proxies in inferring the diversified demand for the network design in smart cities. In addition, with the increasing number of considered households n, there is a significant reduction in the diversified peak demand for each category (e.g., ADMD^∞ is approximately ten times lower than ADMD^1 for Affluent3+), thus highlighting the importance of considering demand diversity while estimating the peak demand for practical applications such as operation and planning. Finally, compared with the empirical value (i.e., $\text{ADMD}^\infty = 2\,\text{kW}$), which has been widely used in DNO's network design [27], lower values for different categories are presented in Fig. 6, indicating the overestimation problem of using the heuristic value and demonstrating the benefits of employing the proposed data-driven approach.

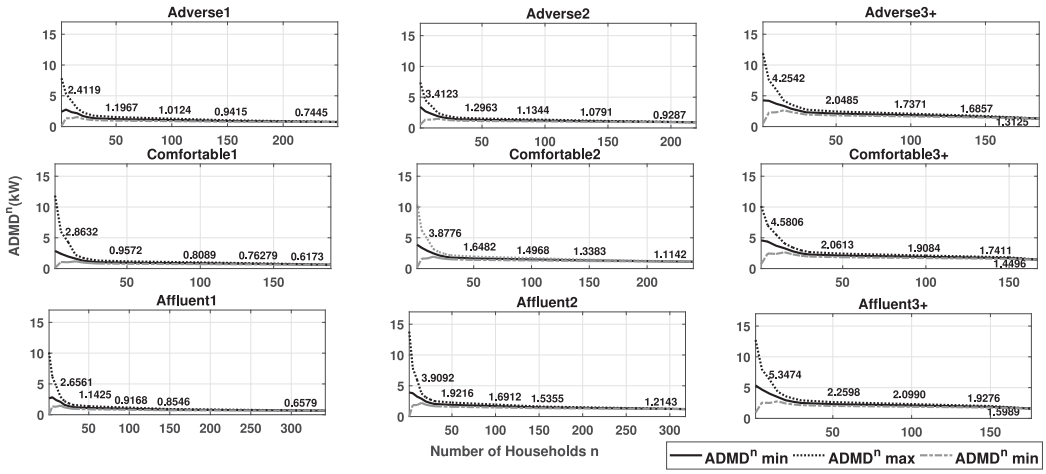


Fig. 6. ADMD^n curve for all the nine categories estimated via M4.

TABLE VI
PEAK DEMAND ESTIMATION PERFORMANCE COMPARISON FOR MULTIPLE CATEGORIES

	M0	M1	M2	M3	M4
MAPE (%)	102.11	14.14	12.87	10.02	**8.12**
Pinball (kW)	–	3.2509	2.9312	2.2236	**1.8787**

E. Results for Multiple Categories

The analyses in the previous sections have illustrated that the tested methods for each of the nine categories show significant differences in the performance of peak demand estimation. However, it is a more realistic scenario for a DNO to perform the network design across multiple categories with different numbers of future customers. Consequently, in this section, an example of a multicategory case is described as follows: a new development of 50 new one-bedroom flats, 30 new two-bedroom flats, and ten new three-bedroom houses in the areas that fall within the ACORN categories designated as Adverse, Comfortable, and Affluent, respectively. The total coincident peak demand for all the tested methods can be estimated following the steps introduced in Section II by using the probabilistic NMDD values, obtained by normalizing ADMD ^{n} curves shown in Fig. 6. Comparing the estimation performance of the different tested methods, it is apparent that the gradually decreased MAPE and pinball loss values shown in Table VI indicate the advantages of considering the load modeling and sampling stage, the customer grouping stage, and their combinations during the peak demand estimation procedure. Specifically, in terms of the MAPE, M4 has an approximately 92% improvement when compared with the empirical method M0. Additionally, from M4 to M1, there is an approximately 42.21% reduction in the calculated pinball loss value. Note that the pinball loss value cannot be calculated for the empirical method M0 because it is a deterministic estimation method.

VI. CONCLUSION

In future smart cities, the coincident peak demand estimation of new customers is one of the key challenges for designing power equipment such as substations and power delivery lines. This paper proposed a data-driven probabilistic peak demand estimation method that includes four main stages. In the load modeling and sampling stage, a novel VTRC method was proposed to capture the complex dependencies among consumers. In addition, a correlation-based hierarchical clustering method was implemented in the customer grouping stage to improve the performance of peak demand estimation. In this work, a vast number of fine-grained smart meter data and the corresponding sociodemographic data across 2639 customers were provided by the LCL smart meter trial. Nine customer categories were constructed according to the information of household occupancy (i.e., 1, 2, and 3+) and wealth level (i.e., Adverse, Comfortable, and Affluent). Subsequently, M1, M2, M3, and M4 were performed and compared in the context of different numbers of n_g. Moreover, this analysis was conducted for each category. Finally, the case of multiple categories peak demand estimation was investigated. Comparing the calculated MAPE and pinball loss values of the tested methods, major conclusions stemming from the analysis are that the performance of the estimated peak demand exhibits gradual improvements from the empirical method to the data-driven method, from using historical data to employing generated samples, and from random customer selection to cluster-based selection.

Further research could focus on developing the proposed peak load estimation method for a more complex situation with distributed photovoltaic, storages, and demand response.

REFERENCES

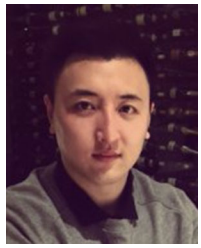
[1] T. Strasser, P. Siano, and V. Vyatkin, "New trends in intelligent energy systems—An industrial electronics point of view," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2420–2423, Apr. 2015.

[2] C. Cecati, J. Kolbusz, P. Różycki, P. Siano, and B. M. Wilamowski, "A novel RBF training algorithm for short-term electric load forecasting and comparative studies," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6519–6529, Oct. 2015.

[3] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016.

[4] P. G. Da Silva, D. Ilic, and S. Karnouskos, "The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 402–410, Jan. 2014.

[5] M. Sun, I. Konstantelos, and G. Strbac, "Analysis of diversified residential demand in London using smart meter and demographic data," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2016, pp. 1–5.
[6] C. Barteczko-Hibbert, "After diversity maximum demand (ADMD) report," Report for the "Customer-Led Network Revolution" project, Durham Univ., Durham, U.K., Rep. CLNR-L217, 2015.
[7] U. C. Chukwu, O. A. Nworgu, and D. O. Dike, "Impact of V2G penetration on distribution system components using diversity factor," in *Proc. IEEE SOUTHEASTCON*, 2014, pp. 1–8.
[8] C. J. Ziser, Z. Dong, and T. K. Saha, "Probabilistic modelling of demand diversity and its relationship with electricity market outcomes," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, 2007, pp. 1–6.
[9] R. Herman and C. T. Gaunt, "A practical probabilistic design procedure for LV residential distribution systems," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 2247–2254, Oct. 2008.
[10] J. Munkhammar, J. Rydén, and J. Widén, "Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data," *Appl. Energy*, vol. 135, pp. 382–390, 2014.
[11] D. H. McQueen, P. R. Hyland, and S. J. Watson, "Monte Carlo simulation of residential electricity demand for forecasting maximum demand on distribution networks," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1685–1689, Aug. 2004.
[12] P. Boait, V. Advani, and R. Gammon, "Estimation of demand diversity and daily demand profile for off-grid electrification in developing countries," *Energy Sustain. Develop.*, vol. 29, pp. 135–141, 2015.
[13] Z. Wan, G. Wang, Y. Yang, and S. Shi, "SKM: Scalable key management for advanced metering infrastructure in smart grids," *IEEE Trans. Ind. Electron.*, vol. 61, no. 12, pp. 7055–7066, Dec. 2014.
[14] M. Sun, I. Konstantelos, and G. Strbac, "C-vine copula mixture model for clustering of residential electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2382–2393, May 2017.
[15] X. Li, C. P. Bowers, and T. Schnier, "Classification of energy consumption in buildings with outlier detection," *IEEE Trans. Ind. Electron.*, vol. 57, no. 11, pp. 3639–3644, Nov. 2010.
[16] H. Yang and S.-Y. R. Hui, "Nonintrusive power measurement method with phase detection for low-cost smart meters," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3962–3969, May 2017.
[17] S. B. Taieb, R. Huser, R. J. Hyndman, and M. G. Genton, "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2448–2455, Sep. 2016.
[18] G. Chicco, "Customer behaviour and data analytics," in *Proc. Int. Conf. Expo. Elect. Power Eng.*, 2016, pp. 771–779.
[19] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templates—Part II: Peak load estimation by clusterwise regression," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3045–3052, Nov. 2015.
[20] R. B. Nelsen, "Introduction," in *An Introduction to Copulas*. New York, NY, USA: Springer, 1999, pp. 1–4.
[21] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Math. Econ.*, vol. 44, no. 2, pp. 182–198, 2009.
[22] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data," *Can. J. Statist.*, vol. 40, no. 1, pp. 68–85, 2012.
[23] J. Dissmann, E. C. Brechmann, C. Czado, and D. Kurowicka, "Selecting and estimating regular vine copulae and application to financial returns," *Comput. Statist. Data Anal.*, vol. 59, pp. 52–69, 2013.
[24] M. Sun, I. Konstantelos, S. Tindemans, and G. Strbac, "Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems," in *Proc. IEEE Power Syst. Comput. Conf.*, 2016, pp. 1–8.
[25] X. He and P. Niyogi, "Locality preserving projections," in *Proc. 16th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 153–160.
[26] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecast.*, vol. 32, pp. 896–913, 2016.
[27] T. Haggis, *Network Design Manual*, E.on, Essen, Germany, 2006.
[28] J. R. Schofield, R. Carmichael, S. H. Tindemans, M. Bilton, M. Woolf, and G. Strbac, "Low Carbon London project: Data from the dynamic time-of-use electricity pricing trial, 2013," 2015.



Mingyang Sun (M'16) received the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2017.

He is currently a Research Associate at Imperial College London. His research interests include big data analytics in power systems.

Yi Wang (S'14) received the B.S. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014. He is currently working toward the Ph.D. degree in electrical engineering with Tsinghua University, Beijing, China.

He is also a Visiting Student Researcher with the University of Washington, Seattle, WA, USA. His research interests include data analytics in smart grids and multiple energy systems.

Goran Strbac (M'95) received the Ph.D. degree in electrical engineering from the University of Belgrade, Yugoslavia, in 1993.

He is a Professor of Electrical Energy Systems at Imperial College London, London, U.K. His current research interests include operation, planning and market design of flexible, low carbon energy systems.

Chongqing Kang (M'01–SM'08–F'17) received the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 1997.

He is currently a Professor with Tsinghua University. His research interests include power system planning, power system operation, renewable energy, low-carbon electricity technology, and load forecasting.