Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges

Yi Wang, *Student Member, IEEE*, Qixin Chen, *Senior Member, IEEE*, Tao Hong, and Chongqing Kang, *Fellow, IEEE*

***Abstract*—The widespread popularity of smart meters enables an immense amount of fine-grained electricity consumption data to be collected. Meanwhile, the deregulation of the power industry, particularly on the delivery side, has continuously been moving forward worldwide. How to employ massive smart meter data to promote and enhance the efficiency and sustainability of the power grid is a pressing issue. To date, substantial works have been conducted on smart meter data analytics. To provide a comprehensive overview of the current research and to identify challenges for future research, this paper conducts an application-oriented review of smart meter data analytics. Following the three stages of analytics, namely, descriptive, predictive, and prescriptive analytics, we identify the key application areas as load analysis, load forecasting, and load management. We also review the techniques and methodologies adopted or developed to address each application. In addition, we also discuss some research trends, such as big data issues, novel machine learning technologies, new business models, the transition of energy systems, and data privacy and security.**

***Index Terms*—Smart meter, big data, data analytics, demand response, consumer segmentation, clustering, load forecasting, anomaly detection, deep learning, machine learning.**

I. Introduction

Smart meters have been deployed around the globe during the past decade. For example, the numbers of smart meters installed in the U.K., the U.S., and China reached 2.9 million [1], 70 million [2], [3], and 96 million, respectively by the end of 2016. Smart meters, together with the communication network and data management system, constitute the advanced metering infrastructure (AMI), which plays a vital role in power delivery systems by recording the load profiles and facilitating bi-directional information flow [4]. The widespread popularity of smart meters enables an immense amount of fine-grained electricity consumption data to be collected. Billing is no longer the only function of smart meters. High-resolution data from smart meters provide rich information on the electricity consumption behaviors and lifestyles of the consumers. Meanwhile, the deregulation of the power industry, particularly on the delivery side, is continuously moving forward in many countries worldwide. These countries are now sparing no effort on electricity retail market reform. Increasingly more participators, including retailers, consumers, and aggregators, are involved in making the retail market more prosperous, active, and competitive [5]. How to employ massive smart meter data to promote and enhance the efficiency and sustainability of the demand side has become an important topic worldwide.

In recent years, the power industry has witnessed considerable developments of data analytics in the processes of generation, transmission, equipment, and consumption. Increasingly more projects on smart meter data analytics have also been established. The National Science Foundation (NSF) of the United States provides a standard grant for cross-disciplinary research on smart grid big data analytics [6]. Several projects for smart meter data analytics are supported by the CITIES Innovation Center in Denmark. These projects investigate machine learning techniques for smart meter data to improve forecasting and money-saving opportunities for customers [7]. The Bits to Energy Lab which is a joint research initiative of ETH Zurich, the University of Bamberg, and the University of St. Gallen, has launched several projects for smart meter data analytics for customer segmentation and scalable efficiency services [8]. The Siebel Energy Institute, a global consortium of innovative and collaborative energy research, funds cooperative and innovative research grants for data analytics in smart girds [9]. Meanwhile, the National Science Foundation of China (NSFC) and the National Key R&D Program of China are approving increasingly more data-analytics-related projects in the smart grid field, such as the National High Technology Research and Development Program of China (863 Program) titled Key Technologies of Big Data Analytics for Intelligent Distribution and Utilization. ESSnet Big Data, a project within the European statistical system (ESS), aims to explore big data applications, including smart meters [10]. The workpackage in the ESSnet Big Data project concentrates on smart meter data access, handling, and deployments of methodologies and techniques for smart meter data analytics. National statistical institutes from Austria, Denmark, Estonia, Sweden, Italy, and Portugal jointly conduct this project.

Apart from academic research, data analytics has already been used in industry. In June 2017, SAS published the

Manuscript received September 2, 2017; revised January 1, 2018; accepted February 15, 2018. Date of publication March 22, 2018; date of current version April 19, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0900100, and in part by the Major Smart Grid Joint Project of National Natural Science Foundation of China and State Grid under Grant U1766212. Paper no. TSG-01277-2017. *(Corresponding author: Chongqing Kang.)*

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

T. Hong is with the Energy Production and Infrastructure Center, University of North Carolina at Charlotte, Charlotte, NC 28223 USA.

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/TSG.2018.2818167

results from its industrial analytics survey [11]. This survey aims to provide the issues and trends shaping how utilities deploy data and analytics to achieve business goals. There are 136 utilities from 24 countries that responded to the survey. The results indicate that data analytics application areas include energy forecasting, smart meter analytics, asset management/analytics, grid operation, customer segmentation, energy trading, credit and collection, call center analytics, and energy efficiency and demand response program engagement and marketing. More and more energy data scientists will be jointly trained by universities and industry bridge the talent gap in energy data analytics [12]. Meanwhile, the privilege of smart meters and deregulation of the demand side are accelerating the birth of many start-ups. These start-ups attempt to collect and analyze smart meter data and provide insights and value-added services for consumers and retailers to make profits. More details regarding industrial applications can be found from the businesses of the data-analytics-based start-ups.

Analytics is known as the scientific process of transforming data into insights for making better decisions. It is commonly dissected into three stages: descriptive analytics (what do the data look like), predictive analytics (what is going to happen with the data), and prescriptive analytics (what decisions can be made from the data). This review of smart meter data analytics is conducted from these three aspects.

A. Bibliometric Analysis

To provide an overview of the existing research in smart meter data analytics, a bibliometric analysis was conducted on 31 December 2017 using the well-established and acknowledged databases, Web of Science (WoS). The query for WoS is as follows: TS=((“smart meter” OR “consumption” OR “demand” OR “load”) AND “data” AND (“household” OR “resident” OR “residential” OR “building” OR “industrial” OR “individual” OR “customer” OR “consumer”) AND (“energy theft” OR “demand response” OR “clustering” OR “forecasting” OR “profiling” OR “classification” OR “abnormal” OR “anomaly”) AND (“smart grid” OR “power system”)).

Fig. 1 shows the number of publications indexed by WoS 2010 to 2017. In total, 200 publications were found in WoS. Before 2011, the number of publications was at a relatively low level, while it increased rapidly beginning in 2012 and reached 60 in WoS, in the year of 2017. This result should not be a surprise. The smart grid initiatives started around the late 2000s. It takes a few years for power companies to collect the data for extensive research and another few years to bring the research findings to journal publications.

Fig. 2 depicts the journals ranked by the number of relevant papers published since 2010 according to WoS. IEEE Transactions on Smart Grid, the youngest journal on this list, has published 28 relevant papers since it was founded in 2012, making it the most popular journal for smart meter data analytics articles. Among the top 5 most popular journals, Energy and Buildings is not a traditional venue for power engineering papers since smart meters are tied to residential homes and office buildings, which makes this journal a natural outlet for smart meter data analytics papers.

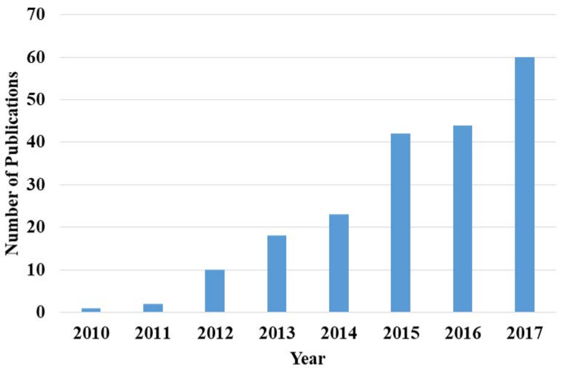


Fig. 1. Number of publications indexed by WoS.

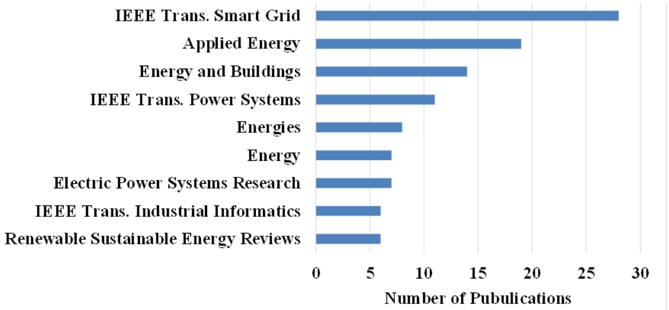


Fig. 2. Number of publications in nine most popular journals.

B. Relevant Review Articles

Here, we conduct a brief review of several review articles related to smart meter data analytics. Note that these existing reviews are either beyond the scope of this paper or not as comprehensive and up-to-date as this paper.

The challenges and opportunities related to the privacy and security aspects of energy data analytics were discussed in [13]. Several reviews of load profiling have been conducted from the aspects of clustering method comparison [14], applications [15] and demand response [16]. A few review articles focused on non-intrusive load monitoring (NILM), such as different disaggregation algorithms [17] and the evaluation criteria [18]. The methods for how to detect non-technical loss (NTL) in power grids based on smart meter data were reviewed in [19], while the NTL challenges were discussed in [20]. The applications of smart meter data analytics on customer behavior were summarized in [21]. Load forecasting is the main application of smart meter data analytics. A natural application of smart meter data is to enhance load forecast accuracy. The recent developments of point and probabilistic load forecasting were critically reviewed in [22]. Several forecasting techniques for building loads were summarized in [23]. A comprehensive review of smart meter analytics was conducted in [24], which covered several large areas, including the smart metering environments, analysis techniques, and potential applications. Our paper differs from [24] by only focusing on data analytics applications and methodologies rather than including data collection and communication. The review papers mentioned above summarized the application of smart meter data analytics from a specific aspect. In addition, under the background of the big data era and further opening of the retail market, smart meter data analytics is still an emerging research field. More data analytics methods have been studied, and more novel problems have been defined. This paper attempts to provide a comprehensive review of smart meter data analytics in wider applications. More emphasis is placed on the research over the past five years. Classical and typical works of literature that were published earlier are also included.

C. Open Load Datasets

Due to many issues, such as privacy and security, many power companies are hesitant to release their smart meter data to the public. This has been a challenge for conducting research in smart meter data analytics and its applications. Nevertheless, several anonymized or semi-anonymized datasets at the household level have been made publicly available over the past few years. Various studies have been conducted based on these smart meter datasets. Several open load datasets are summarized in Table I.

- **Customer Behavior Trials**: The Commission for Energy Regulation (CER), the regulator for the electricity and natural gas sectors in Ireland, launched a smart metering project to conduct customer behavior trials (CBTs) to determine how smart metering can help shape energy usage behaviors across a variety of demographics, lifestyles, and home sizes [25].
- **Low Carbon London**: Similar to CBTs, a trial as a part of the Low Carbon London project involved over five thousand households in the London area [38]. Smart meter data, time-of-use tariff data, and survey data were collected to investigate the impacts of a wide range of low carbon technologies on London's electricity distribution network.
- **PecanStreet**: The dataset is supported by the Pecan Street experiment in Austin, TX. The dataset contains minute-level electricity consumption data from 500 homes (both the whole-home level and individually monitored appliance circuits) [41].
- **Building Data Genome**: The dataset is from The Building Data Genome Project. There are 507 whole-building electrical meters in this collection, and the majority are from buildings on university campuses [42].
- **UMass Smart**: This dataset was collected by the Mass Smart Microgrid project, which contains consumption data of 400 homes at the one-minute granularity of a whole day [43].
- **Ausgrid Resident**: The smart meter data combined with rooftop PV generation data of 300 residential consumers in an Australian distribution network over 3 years have been recorded by the Ausgrid distribution network [46].
- **Ausgrid Substation**: Ausgrid also makes the load profiles of approximately 180 zone substations publicly available from 2005 [47]. The dataset is continuously updated.
- **GEFCom2012**: About four and a half years (Janurary 2004 to July 2008) of hourly load and temperature data were provided. The temperature readings were from 11 weather stations. The top ranked methods in the competition were summarized in [48].
- **GEFCom2014**: The dataset contains the hourly load data and 25 weather station data during 2005 and 2010. The detailed description of the dataset and many practical and high-rank forecasting methods in this competition are summarized in [57].
- **ISO New England**: ISO New England publishes the system-level load data and corresponding temperature data of 9 zones every month. More information can be found in [65].

As shown, the datasets including Low Carbon London, Ausgrid Residents, and Ausgrid Substation have seldom been used in the existing literature. In addition, Building Data Genome is a newly released dataset. More works can be performed based on these datasets.

D. Taxonomy

Fig. 3 depicts the five major players on the demand side of the power system: consumers, retailers, aggregators, distribution system operators (DSO), and data service providers.

For retailers, at least four businesses related to smart meter data analytics need to be conducted to increase the competitiveness in the retail market. 1) Load forecasting, which is the basis of decision making for the optimization of electricity purchasing in different markets to maximize profits. 2) Price design to attract more consumers. 3) Providing good service to consumers, which can be implemented by consumer segmentation and characterization. 4) Abnormal detection to have a cleaner dataset for further analysis and decrease potential loss from electricity theft. For consumers, individual load forecasting, which is the input of future home energy management systems (HEMS) [70], can be conducted to reduce their electricity bill. In the future peer-to-peer (P2P) market, individual load forecasting can also contribute to the implementation of transactive energy between consumers [71], [72]. For aggregators, they deputize a group of consumers for demand response or energy efficiency in the ancillary market. Aggregation level load forecasting and demand response potential evaluation techniques should be developed. For DSO, smart meter data can be applied to distribution network topology identification, optimal distribution system energy management, outage management, and so forth. For data service providers, they need to collect smart meter data and then analyze these massive data and provide valuable information for retailers and consumers to maximize profits or minimize cost. Providing data services including data management, data analytics is an important business model when increasingly more smart meter data are collected and to be processed.

To support the businesses of retailers, consumers, aggregators, DSO, and data service providers, following the three stages of analytics, namely, descriptive, predictive and prescriptive analytics, the main applications of smart meter data analytics are classified into load analysis, load forecasting, load managements, and so forth. The detailed taxonomy is illustrated in Fig. 4. The main machine learning techniques used for smart meter data analytics include time series, dimensionality reduction, clustering, classification, outlier detection, deep learning, low-rank matrix, compressed sensing, online

TABLE I
BASIC INFORMATION OF SEVERAL OPEN LOAD DATASETS

Name	Brief Description	Number	Frequency	Duration	References
Customer Behavior Trials [25]	Smart meter read data; Pre- and post-trial survey data;	6445	Every 30 min	2009/9-2011/1	[26][27][28][29] [30][31][32][33] [34][35][36][37]
Low Carbon London [38]	Smart meter read data; Electricity price data; Appliance and attitude survey data;	5567	Every 30 min	2013/1-2013/12	[39][40]
PecanStreet [41]	Residential electricity consumption data; Electric vehicle charging data;	500	Every 1 min	2005/5-2017/5	
Building Data Genome [42]	Non-residential building smart meter data; Area, weather, and primary use type data;	507	Every 1 hour	2014/12-2015/11	
UMass Smart [43]	Residential electricity consumption data;	400	Every 1 min	One day	[44][45]

Ausgrid Residents [46]	General consumption data;
	Controlled load consumption data;
	PV output data;
		300	Every 30 min	2010/7-2013/6	

Ausgrid Substation [47]	Substation metering data;	177	Every 15 min	2005/5-	

GEFCom 2012 [48]	Zonal load data;
	Temperature data;
		20	Hourly	2003/1-2008/6	[49][50][51][52]
					[53][54][55][56]

GEFCom 2014 [57]	Zonal load data;
	25 weather station data;
		1	Hourly	2005/1-2010/9	[58][59][60][61]
					[62][63][64]

ISO New England [65]	System load data;
	Temperature data;
	Locational marginal pricing data;
		9	Hourly	2003/1-	[66][67][68][69]

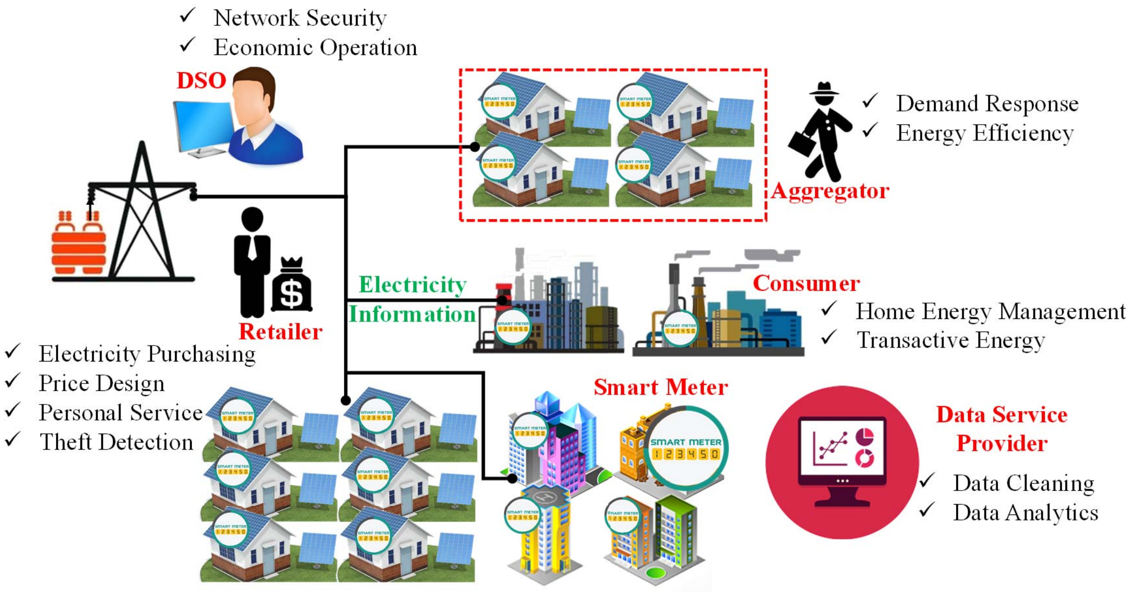


Fig. 3. Participators and their businesses on the demand side.

learning, and so on. Studies on how smart meter data analytics works for each application and what methodologies have been applied will be summarized in the following sections.

E. Contributions and Organization

This paper attempts to provide a comprehensive review of the current research in recent years and identify future challenges for smart meter data analytics. Note that every second or higher frequency data used for NILM are very limited at present due to the high cost of communicating and storing the data. The majority of smart meters collect electricity consumption data at a frequency of every 15 minutes to each hour. In addition, several comprehensive reviews have been conducted on NILM. Thus, in this review paper, works about NILM are not included.

The contributions of this paper are as follows:

1) Conducting a comprehensive literature review of smart meter data analytics on the demand side with the newest developments, particularly over the past five years.
2) Providing a well-designed taxonomy for smart meter data analytics applications from the perspective of load analysis, load forecasting, load management, and so forth.
3) Discussing open research questions for future research directions, including big data issues, new machine learning technologies, new business models, the transition of energy systems, and data privacy and security.

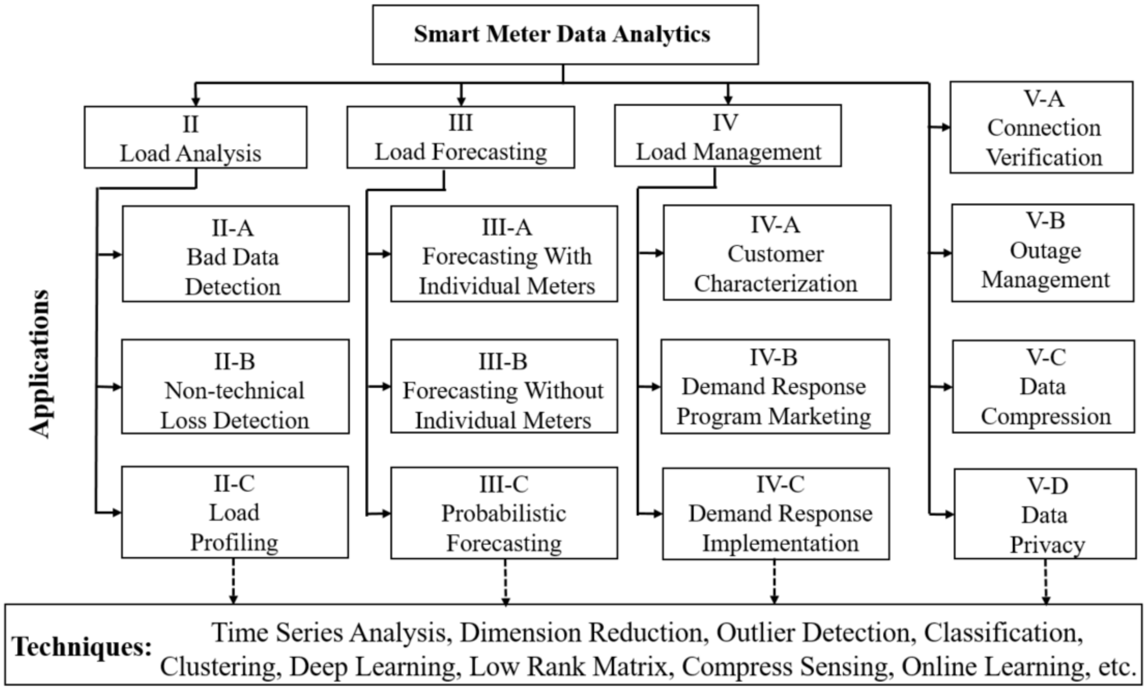

Smart Meter Data Analytics
Applications
II Load Analysis
III Load Forecasting
IV Load Management
V-A Connection Verification
II-A Bad Data Detection
III-A Forecasting With Individual Meters
IV-A Customer Characterization
V-B Outage Management
II-B Non-technical Loss Detection
III-B Forecasting Without Individual Meters
IV-B Demand Response Program Marketing
V-C Data Compression
II-C Load Profiling
III-C Probabilistic Forecasting
IV-C Demand Response Implementation
V-D Data Privacy
Techniques: Time Series Analysis, Dimension Reduction, Outlier Detection, Classification, Clustering, Deep Learning, Low Rank Matrix, Compress Sensing, Online Learning, etc.

Fig. 4. Taxonomy of smart meter data analytics.

The remainder of the paper is organized as follows. Sections II–IV conduct the survey on smart meter data analytics for load analysis, load forecasting, and load management, respectively. Section V provides a miscellany of smart meter data analytics in addition to the three aspects above. Section VI discusses several future research issues. Section VII draws the conclusions.

II. Load Analysis

Fig. 5 shows eight typical normalized daily residential load profiles obtained using the simple k-means algorithm in the Irish resident load dataset. The load profiles of different consumers on different days are diverse. Having a better understanding of the volatility and uncertainty of the massive load profiles is very important for further load analysis. In this section, the works on load analysis are reviewed from the perspectives of anomaly detection and load profiling. Anomaly detection is very important because training a model such as a forecasting model or clustering model on a smart meter dataset with anomalous data may result in bias or failure for parameter estimation and model establishment. Moreover, reliable smart meter data are important for accurate billing. The works on anomaly detection in smart meter data are summarized from the perspective of bad data detection and NTL detection (or energy theft detection). Load profiling is used to find the basic electricity consumption patterns of each consumer or a group of consumers. The load profiling results can be further used for load forecasting and demand response programs.

A. Bad Data Detection

Bad data as discussed here can be missing data or unusual patterns caused by unplanned events or the failure of data collection, communication, or entry. Bad data detection can be divided into probabilistic, statistical, and machine learning methods [73]. The methods for bad data detection in other areas can be directly applied to smart meter data. Only the works closely related to smart meter bad data detection are surveyed in this subsection. According to the modeling methods, these works are summarized as time-series-based methods, low-rank matrix technique-based methods, and time-window-based methods.

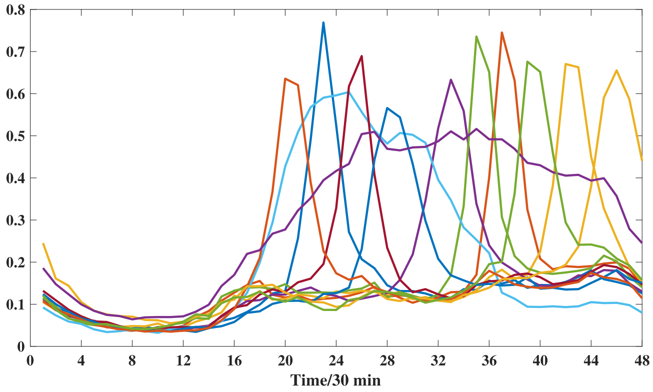


Fig. 5. Typical normalized daily residential load profiles.

Smart meter data are essentially time series. An optimally weighted average (OWA) method was proposed for data cleaning and imputation in [74], which can be applied to offline or online situations. It was assumed that the load data can be explained by a linear combination of the nearest neighbor data, which is quite similar to the autoregressive moving average (ARIMA) model for time series. The optimal weight was obtained by training an optimization model. While in [75], the nonlinear relationship between the data at different time periods and exogenous inputs was modeled by combining autoregressive with exogenous inputs (ARX) and artificial neural network (ANN) models where the bad data detection was modeled as a hypothesis testing on the extreme of the residuals. A case study on gas flow data was performed and showed an improvement in load forecasting accuracy after ARX-based bad data detection. Similarly, based on the autoregression (AR) model, the generalized extreme Studentized deviate (GESD) and the Q-test were proposed to detect the outliers when the number of samples is more and less than ten, respectively, in [76]. Then, canonical variate analysis (CVA) was conducted to cluster the recovered load profiles, and a linear discriminate analysis (LDA) classifier was further used to search for abnormal electricity consumption. Instead of detecting bad data, which forecasting method is robust to the cyber attack or bad data without bad data detection was investigated in [77].

The electricity consumptions are spatially and temporally correlated. Exploring the spatiotemporal correlation can help identify the outliers and recover them. A low-rank matrix fitting-based method was proposed in [216] to conduct data cleaning and imputation. An alternating direction method of multipliers (ADMM)-based distributed low-rank matrix technique was also proposed to enable communication and data exchange between different consumers and to protect the privacy of consumers. Similarly, to produce a reliable state estimation, the measurements were first processed by low-rank matrix techniques in [78]. Both off-line and on-line algorithms have been proposed. However, the improvement in state estimation after low-rank denoising has not been investigated. Low-rank matrix factorization works well when the bad data are randomly distributed. However, when the data are unchanged for a certain period, the low-rank matrix cannot handle it well. More data preparation works were conducted to detect these types of bad data before singular value thresholding (SVT)-based low-rank matrix-based bad data identification and recovery in [79].

Rather than detecting all the bad data directly, strategies that continuously detect and recover a part within a certain time window have also been studied. A clustering approach was proposed on the load profiles with missing data in [80] and [81]. The clustering was conducted on segmented profiles rather than the entire load profiles in a rolling manner. In this way, the missing data can be recovered or estimated by other data in the same cluster. A collective contextual anomaly detection using a sliding window framework was proposed in [82] by combining various anomaly classifiers. The anomalous data were detected using overlapping sliding windows. Since smart meter data are collected in a real-time or near real-time fashion, an online anomaly detection method using the Lambda architecture was proposed in [83]. The proposed online detection method can be parallel processed, having high efficiency when working with large datasets.

B. Energy Theft Detection

Strictly speaking, smart meter data with energy theft also belong to bad data. The bad data discussed above are unintentional and appear temporarily, whereas energy theft may change the smart meter data under certain strategies and last for a relatively long time. Energy theft detection can be implemented using smart meter data and power system state data, such as node voltages. The energy theft detection methods with only smart meter data are summarized in this part from two aspects: supervised learning and unsupervised learning.

Supervised classification methods are effective approaches for energy theft detection, which generally consists of two stages: feature extraction and classification. To train a theft detection classifier, the non-technical loss was first estimated in [26]. K-means clustering was used to group the load profiles, where the number of clusters was determined by the silhouette value [84]. To address the challenge of imbalanced data, various possible malicious samples were generated to train the classifier. An energy theft alarm was raised after a certain number of abnormal detections. Different numbers of abnormal detections resulted in different false positive rates (FPR) and Bayesian detection rates (BDR). The proposed method can also identify the energy theft types. Apart from clustering-based feature extraction, an encoding technique was first performed on the load data in [85], which served as the inputs of classifiers including SVM and a rule-engine-based algorithm to detect the energy theft. The proposed method can run in parallel for real-time detection. By introducing external variables, a top-down scheme based on decision tree and SVM methods was proposed in [86]. The decision tree estimated the expected electricity consumption based on the number of appliances, persons, and outdoor temperature. Then, the output of the decision tree was fed to the SVM to determine whether the consumer is normal or malicious. The proposed framework can also be applied for real-time detection.

Obtaining the labeled dataset for energy theft detection is difficult and expensive. Compared with supervised learning, unsupervised energy theft detection does not need the labels of all or partial consumers. An optimum-path forest (OPF) clustering algorithm was proposed in [87], where each cluster is modeled as a Gaussian distribution. The load profile can be identified as an anomaly if the distance is greater than a threshold. Comparisons with frequently used methods, including k-means, Birch, affinity propagation (AP), and Gaussian mixture model (GMM), verified the superiority of the proposed method. Rather than clustering all load profiles, clustering was only conducted within an individual consumer to obtain the typical and atypical load profiles in [88]. A classifier was then trained based on the typical and atypical load profiles for energy theft detection. A case study in this paper showed that extreme learning machine (ELM) and online sequential-ELM (OS-ELM)-based classifiers have better accuracy compared with SVM. Transforming the time series smart meter data into the frequency domain is another approach for feature extraction. Based on the discrete Fourier transform (DFT) results, the features extracted in the reference interval and examined interval were compared based on the so-called *Structure & Detect* method in [89]. Then, the load profile can be determined to be a normal or malicious one. The proposed method can be implemented in a parallel and distributed manner, which can be used for the on-line analysis of large datasets. Another unsupervised energy theft detection method is to formulate the problem as a load forecasting problem. If the metered consumption is considerably lower than the forecasted consumption, then the consumer can be marked as a malicious consumer. An anomaly score was given to each consumption data and shown with different colors to realize visualization in [90].

C. Load Profiling

Load profiling refers to the classification of load curves or consumers according to electricity consumption behaviors. In this subsection, load profiling is divided into direct-clustering-based and indirect-clustering-based approaches. Various clustering techniques, such as K-means, hierarchical clustering, and self-organizing map (SOM), have been directly implemented on smart meter data [14]–[16]. Two basic issues about direct clustering are first discussed. Then, the works on indirect clustering are classified into dimensionality reduction, load characteristics, and variability and uncertainty-based methods according to the features that are extracted before clustering.

There are some basic issues associated with direct clustering. The first issue is the resolution of smart meter data. In [91], three frequently used clustering techniques, namely, k-means, hierarchical algorithms, and the Dirichlet process mixture model (DPMM) algorithm, were performed on the smart meter data with different frequencies varying from every 1 minute to 2 hours to investigate how the resolution of smart meter data influences the clustering results. The results showed that the smart meter data with a frequency of at least every 30 minutes is sufficiently reliable for most purposes. The second issue is that the smart meter data are essentially time-series data. In contrast to traditional clustering methods for static data, k-means modified for dynamic clustering was proposed in [92] to address time-dependent data. The dynamic clustering allows capturing the trend of clusters of consumers. A two-stage clustering strategy was proposed in [45] to reduce the computational complexity. In the first stage, K-means was performed to generate the local representative load profiles; in the second stage, clustering was further performed on the clustering centers obtained in the first stage at the central processor. In this way, the clustering method can be performed in a distributed fashion and largely reduce the overall complexity.

Apart from direct clustering, increasingly more literatures are focusing on indirect clustering, i.e., feature extraction is conducted before clustering. Dimensionality reduction is an effective way to address the high dimensionality of smart meter data. Principal component analysis (PCA) was performed on yearly load profiles to reduce the dimensionality of original data and then k-means was used to classify consumers in [93]. The components learned by PCA can reveal the consumption behaviors of different connection point types. Similarly, PCA was also used to find the temporal patterns of each consumer and spatial patterns of several consumers in [94]. Then, a modified K-medoids algorithm based on the Hausdorff distance and Voronoi decomposition method was proposed to obtain typical load profiles and detect outliers. The method was tested on a large real dataset to prove the effectiveness and efficiency. Deep-learning-based stacked sparse auto-encoders were applied for load profile compression and feature extraction in [95]. Based on the reduced and encoded load profile, a locality sensitive hashing (LSH) method was further proposed to classify the load profiles and obtain the representative load profiles.

Insights into the local and global characteristics of smart meter data are important for finding meaningful typical load profiles. Three new types of features generated by applying conditional filters to meter-resolution-based features integrated with shape signatures, calibration and normalization, and profile errors were proposed in [96] to cluster daily load curves. The proposed feature extraction method was of low computational complexity, and the features were informative and understandable for describing the electricity usage patterns. To capture local and global shape variations, 10 subspace clustering and projected clustering methods were applied to identify the contact type of consumers in [97]. By focusing on the subspace of load profiles, the clustering process was proven to be more robust to noise. To capture the peak load and major variability in residential consumption behavior, four key time periods (overnight, breakfast, daytime, and evening) were identified in [98]. On this basis, seven attributes were calculated for clustering. The robustness of the proposed clustering was verified using the bootstrap technique.

The variability and uncertainty of smart meter data have also been considered for load profiling. Four key time periods, which described different peak demand behaviors, coinciding with common intervals of the day were identified in [98], and then a finite mixture-model-based clustering was used to discover ten distinct behavior groups describing customers based on their demand and variability. The load variation was modeled by a lognormal distribution, and a Gaussian mixture model (GMM)-based load profiling method was proposed in [99] to capture dynamic behavior of consumers. A mixture model was also used in [39] by integrating the C-vine copula method (C-vine copula-based mixture model) for the clustering of residential load profiles. The high-dimensional nonlinear correlations among consumptions of different time periods were modeled using the C-vine copula. This method has an effective performance in large data sets. While in [29], a Markov model was established based on the separated time periods to describe the electricity consumption behavior dynamics. A clustering technique consisting of fast search and find of density peaks (CFSFDP) integrated into a divide-and-conquer distributed approach was proposed to find typical consumption behaviors. The proposed distributed clustering algorithm had higher computational efficiency. The expectation maximization (EM)-based mixture model clustering method was applied in [100] to obtain typical load profiles, and then the variabilities in residential load profiles were modeled by a transition matrix based on a second-order Markov chain and Markov decision processes. The proposed method can be used to generate pseudo smart meter data for retailers and protect the privacy of consumers.

D. Remarks

Table II provides the correspondence between the key techniques and the surveyed references in smart meter data analytics for load analysis.

For bad data detection, most of the bad data detection methods are suitable for business/industrial consumers or higher aggregation level load data, which are more regular and have certain patterns. The research on bad data detection on the individual consumer is still limited and not a trivial task because the load profiles of an individual consumer show more variation. In addition, since bad data detection and repairing are the basis of other data analytics application, how much improvement can be made for load forecasting or other applications after bad data detection is also an issue that deserves further investigation. In addition, smart meter data are essentially streaming data. Real-time bad data detection for some real-time applications, such as very-short-term load forecasting, is another concern. Finally, as stated above, bad data may be brought from data collection failure. Short period anomaly usage patterns may also be identified as bad data even through it is "real" data. More related factors such as sudden events need to be considered in this situation. Redundant data are also good sources for "real" but anomaly data identification.

For energy theft detection, with a longer time period of smart meter data, the detection accuracy is probably higher because more data can be used. However, using longer historical smart meter data may also lead to a detection delay, which means that we need to achieve a balance between the detection accuracy and detection delay. Moreover, different private data and simulated data have been tested on different energy theft detection methods in the existed literature. Without the same dataset, the superiority of a certain method cannot be guaranteed. The research on this area will be promoted if some open datasets are provided. Besides, in most cases, one paper proposes one energy theft detection method. Just like ensemble learning for load forecasting, can we propose an ensemble detection framework to combine different individual methods?

For load profiling, the majority of the clustering methods are used for stored smart meter data. However, the fact is that smart meter data are streaming data. Sometime, we need deal with the massive streaming data in a real-time fashion for specific applications. Thus, distributed clustering and incremental clustering methods can be further studied in the field of load profiling. Indirect load profile methods extract features first and then conduct clustering on the extracted features. Some clustering methods such as deep embedding clustering [101] that can implement feature extraction and clustering at the same time, have been proposed outside the area of electrical engineering. It is worth trying to apply these state-of-the-art methods to load profiling. Most load profiling methods are evaluated by clustering-based indices, such as similarity matrix indicator (SMI), Davies-Bouldin indicator (DBI) and Silhouette Index (SIL) [102]. More application-oriented matrices such as forecasting accuracy are encouraged to be used to guide the selection of suitable clustering methods. Finally, how to effectively extract meaningful features before clustering to improve the performance and efficiency of load profiling is another issue that needs to be further addressed.

III. Load Forecasting

Load forecasts have been widely used by the electric power industry. Power distribution companies rely on short- and long-term forecasts at the feeder level to support operations and planning processes, while retail electricity providers make pricing, procurement and hedging decisions largely based on the forecasted load of their customers. Fig. 6 presents the normalized hourly profiles of a week for four different types of loads, including a house, a factory, a feeder, and a city. The loads of a house, a factory, and a feeder are more volatile than the city-level load. In reality, the higher level the load is measured at, the smoother the load profile typically is. Developing a highly accurate forecast is nontrivial at lower levels.

Although the majority of the load forecasting literature has been devoted to forecasting at the top (high voltage) level, the information from medium/low voltage levels, such as distribution feeders and even down to the smart meters, offer some opportunities to improve the forecasts. A recent review of load forecasting was conducted in [22], focusing on the transition from point load forecasting to probabilistic load forecasting. In this section, we will review the recent literature for both point and probabilistic load forecasting with the emphasis on the medium/low voltage levels. Within the point load forecasting literature, we divide the review based on whether the smart meter data is used or not.

A. Forecasting Without Smart Meter Data

Compared with the load profiles at the high voltage levels, the load profiles aggregated to a customer group or medium/low voltage level are often more volatile and sensitive to the behaviors of the customers being served. Some of them, such as the load of a residential community, can be very responsive to the weather conditions. Some others, such as the load of a large factor, can be driven by specific work schedules. Although these load profiles differ by the customer composition, these load forecasting problems share some common challenges, such as accounting the influence from the competitive markets, modeling the effects of weather variables, and leveraging the hierarchy.

In competitive retail markets, the electricity consumption is largely driven by the number of customers. The volatile customer count contributes to the uncertainties in the future load profile. A two-stage long-term retail load forecasting method was proposed in [103] to take customer attrition into consideration. The first stage was to forecast each customer's load using multiple linear regression with a variable selection method. The second stage was to forecast customer attrition using survival analysis. Thus, the product of the two forecasts provided the final retail load forecast. Another issue in the retail market is the consumers' reactions to the various demand response programs. While some consumers may respond to the price signals, others may not. A nonparametric test was applied to detect the demand responsive consumers so that they can be forecasted separately [104]. Because the authors did not find

TABLE II
BRIEF SUMMARY OF THE LITERATURE ON LOAD ANALYSIS

Load Analysis	Key words	References
Bad Data Detection	Time Series Analysis Low Rank Matrix Time Window	[74] [75] [76] [216] [79][78] [80] [81] [82] [83]
Energy Theft Detection	Supervised Learning Unsupervised Learning	[26] [86] [85] [87] [88] [89] [90]
Load Profiling	Direct Clustering Dimension Reduction Local Characteristics Variability and Uncertainty	[14] [15] [16] [91] [92] [45] [14] [93] [94] [95] [96] [97] [98] [98] [29] [99] [39] [100]

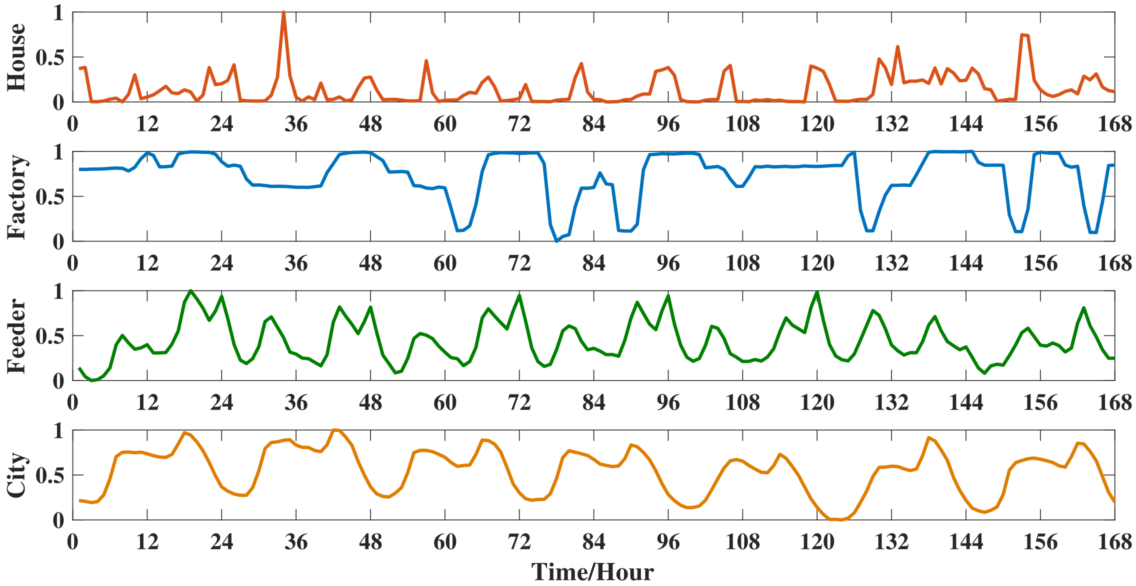


Fig. 6. Normalized hourly profiles of a week for four types of loads.

publicly available demand data for individual consumers, the experiment was conducted using aggregate load in the Ontario power gird.

Since the large scale adoption of electrical air conditioning systems in the 1940s, capturing the effects of weather on load has been a major issue in load forecasting. Most load forecasting models in the literature include temperature variables and their variants, such as lags and averages. How many lagged hourly temperatures and moving average temperatures can be included in a regression model? An investigation was conducted in [66]. The case study was based on the data from the load forecasting track of GEFCom2012. An important finding is that a regression-based load forecasting model estimated using two to three years of hourly data may include more than a thousand parameters to maximize the forecast accuracy. In addition, each zone may need a different set of lags and moving averages.

Not many load forecasting papers are devoted to other weather variables. How to include humidity information in load forecasting models was discussed in [67], where the authors discovered that the temperature-humidity index (THI) may not be optimal for load forecasting models. Instead, separating relative humidity, temperature and their higher order terms and interactions in the model, with the corresponding parameters being estimated by the training data, were producing more accurate load forecasts than the THI-based models. A similar investigation was performed for wind speed variables in [68]. Comparing with the models that include wind chill index (WCI), the ones with wind speed, temperature, and their variants separated were more accurate.

The territory of a power company may cover several micro-climate zones. Capturing the local weather information may help improve the load forecast accuracy for each zone. Therefore, proper selection of weather stations would contribute to the final load forecast accuracy. Weather station selection was one of the challenges designed into the load forecasting track of GEFCom2012 [48]. All four winning team adopted the same strategy: first deciding how many stations should be selected, and then figuring out which stations to be selected [51], [53]–[55]. A different and more accurate method was proposed in [56], which follows a different strategy, determining how many and which stations to be selected at the same time instead of sequentially. The method includes three steps: rating and ranking the individual weather stations, combining weather stations based on a greedy algorithm, and rating and ranking the combined stations. The method is currently being used by many power companies, such as North Carolina Electric Membership Corporation, which was used as one of the case studies in [56].

The pursue of operational excellence and large-scale renewable integration is pushing load forecasting toward the grid edge. Distribution substation load forecasting becomes another emerging topic. One approach is to adopt the forecasting techniques and models with a good performance at higher levels. For instance, a three-stage methodology, which consists of preprocessing, forecasting, and postprocessing, was taken to forecast loads of three datasets ranging from distribution level to transmission level [49]. A semi-parametric additive model was proposed in [105] to forecast the load of Australian National Electricity Market. The same technique was also applied to forecast more than 2200 substation loads of the French distribution network in [106]. Another load forecasting study on seven substations from the French network was reported in [107], where a conventional time series forecasting methodology was used. The same research group then proposed a neural network model to forecast the load of two French distribution substations, which outperformed a time series model [108].

Another approach to distribution load forecasting is to leverage the connection hierarchy of the power grid. In [109], The load of a root node of any subtree was forecasted first. The child nodes were then treated separately based on their similarities. The forecast of a "regular" node was proportional to the parent node forecast, while the "irregular" nodes were forecasted individually using neural networks. Another attempt to make use of the hierarchical information for load forecasting was made in [110]. Two case studies were conducted, one based on New York City and its substations, and the other one based on PJM and its substations. The authors demonstrated the effectiveness of aggregation in improving the higher level load forecast accuracy.

B. Forecasting With Smart Meter Data

The value that smart meters bring to load forecasting is two-fold. First, smart meters make it possible for the local distribution companies and electricity retailers to better understand and forecast the load of an individual house or building. Second, the high granularity load data provided by smart meters offer great potential for improving the forecast accuracy at aggregate levels.

Because the electricity consumption behaviors at the household and building levels can be much more random and volatile than those at aggregate levels, the traditional techniques and methods developed for load forecasting at an aggregate level may or may not be well suited. To tackle the problem of smart meter load forecasting, the research community has taken several different approaches, such as evaluating and modifying the existing load forecasting techniques and methodologies, adopting and inventing new ones, and a mixture of them.

A highly cited study compared seven existing techniques, including linear regression, ANN, SVM and their variants [111]. The case study was performed based on two datasets: one containing two commercial buildings and the other containing three residential homes. The study demonstrated that these techniques could produce fine forecasts for the two commercial buildings but not the three residential homes. A self-recurrent wavelet neural network (SRWNN) was proposed to forecast an education building in a microgrid setting [112]. The proposed SRWNN was shown to be more accurate than its ancestor wavelet neural network (WNN) for both building-level load forecasting (e.g., a 694 kW peak education building in British Columbia, Canada) and state- or province-level load forecasting (e.g., British Columbia and California).

Some researchers tried deep learning techniques for the household- and building-level load forecasting. Conditional Restricted Boltzmann Machine (CRBM) and Factored Conditional Restricted Boltzmann Machine (FCRBM) were assessed in [113] to estimate energy consumption for a household and three submetering measurements. FCRBM achieves the highest load forecast accuracy compared with ANN, RNN, SVM, and CRBM. Different resolutions ranging from one minute to one week have been tested. A pooling-based deep recurrent neural network (RNN) was proposed in [28] to learn spatial information shared between interconnected customers and to address the over-fitting challenges. It outperformed ARIMA, SVR, and classical deep RNN on the Irish CER residential dataset.

Spartsity is a key character in household level load forecasting. A spatio-temporal forecasting approach was proposed in [114], which incorporated a large dataset of many driving factors of the load for all surrounding houses of a target house. The proposed method combined ideas from Compressive Sensing and data decomposition to exploit the low-dimensional structures governing the interactions among the nearby houses. The Pecan Street data was used to evaluate the proposed method. Sparse coding was used to model the usage patterns in [115]. The case study was based on a dataset collected from 5000 households in Chattanooga, TN, where Including the sparse coding features led to 10% improvements in forecast accuracy. A least absolute shrinkage and selection (LASSO)-based sparse linear method was proposed to forecast individual consumption in [116]. The consumer's usage patterns can be extracted from the non-zero coefficients, and it was proven that data from other consumers contribute to the fitted residual. Experiments on real data from Pacific Gas and Electric Company showed that the LASSO-based method has low computational complexity and comparable accuracy.

A commonly used method to reduce noise in smart meter data is to aggregate the individual meters. To keep the salient features from being buried during aggregation, clustering techniques are often used to group similar meters. In [30], next-day load forecasting was formulated as a functional time series problem. Clustering was first performed to classify the historical load curves into different groups. The last observed load curve was then assigned to the most similar cluster. Finally, based on the load curves in this cluster, a functional wavelet-kernel (FWK) approach was used to forecast the next-day load curve. The results showed that FWK with clustering outperforms simple FWK. Clustering was also conducted in [117] to obtain the load patterns. Classification from contextual information, including time, temperature, date, and economic indicator to clusters, was then performed. Based on the trained classifier, the daily load can be forecasted with known contextual information. A shape-based clustering method was performed in [118] to capture the time drift characteristic of the individual load, where the cluster number was smaller than those obtained by traditional Euclidean-distance-based clustering methods. The clustering method is quite similar to k-means, while the distance is quantified by dynamic time warping (DTW). Markov models were then constructed to forecast the shape of the next-day load curve. Similar to the clustering method proposed in [118], a k-shape clustering was proposed in [119] to forecast building time series data, where the time series shape similarity was used to update the cluster memberships to address the time-drift issue.

The fine-grained smart meter data also introduced new perspectives to the aggregation level load forecasting. A clustering algorithm can be used to group the customers. Each customer group can then be forecasted with different forecasting models. Finally, the aggregated load forecast can be obtained by summing the load forecast of each group. Two datasets including the Irish CER residential dataset and another dataset from New York were used to build the case study in [31]. Both showed that forecast errors can be reduced by effectively grouping different customers based on their energy consumption behaviors. A similar finding was presented in [120] where the Irish CER residential dataset was used in the case study. The results showed that cluster-based forecasting can improve the forecasting accuracy and that the performance depends on the number of clusters and the size of the consumer.

The relationship between group size and forecast accuracy based on Seasonal-Nave and Holt-Winters algorithms was investigated in [121]. The results showed that forecasting accuracy increases as group size increases, even for small groups. A simple empirical scaling law is proposed in [122] to describe how the accuracy changes as different aggregation levels. The derivation of the scaling law is based on Mean Absolute Percentage Error (MAPE). Case studies on the data from Pacific Gas and Electric Company show that MAPE decreases quickly with the increase of the number of consumers when the number of consumers is less than 100,000. When the number of consumers is more than 100,000, the MAPE has a little decrease.

Forecast combination is a well-known approach to accuracy improvement. A residential load forecasting case study showed that the ensembles outperformed all the individual forecasts from traditional load forecasting models [123]. By varying the number of clusters, different forecasts can be obtained. A novel ensemble forecasting framework was proposed in [124] to optimally combine these forecasts to further improve the forecasting accuracy.

Traditional error measures such as MAPE cannot reasonably quantify the performance of individual load forecasting due to the violation and time-shifting characteristics. For example, MAPE can easily be influenced by outliers. A resistant MAPE (r-MAPE) based on the calculation of the Huber M-estimator was proposed in [125] to overcome this situation. The mean arctangent absolute percentage error (MAAPE) was proposed in [126] to consider the intermittent nature of individual load profiles. MAAPE, a variation of MAPE, is a slope as an angle, the tangent of which is equal to the ratio between the absolute error and real value, i.e., absolute percentage error (APE). An error measure designed for household level load forecasts was proposed in [127] to address the time-shifting characteristic of household level loads. In addition to these error measures, some modifications of MAPE and mean absolute error (MAE) have been used in other case studies [115], [116].

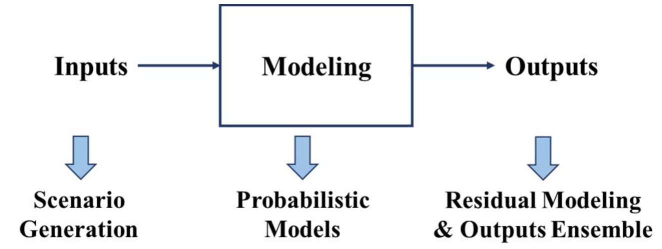


Fig. 7. From point forecasting to probabilistic forecasting.

C. Probabilistic Forecasting

A probabilistic forecast provides more information about future uncertainties that what a point forecast does. As shown in Fig. 7, a typical point forecasting process contains three parts: data inputs, modeling, and data outputs (forecasts). As summarized in [22], there are three ways to modify the workflow to generate probabilistic forecasts: 1) generating multiple input scenarios to feed to a point forecasting model; 2) applying probabilistic forecasting models, such as quantile regression; and 3) augmenting point outputs to probabilistic outputs by imposing simulated or modeled residuals or making ensembles of point forecasts.

On the input side, scenario generation is an effective way to capture the uncertainties from the driving factors of electricity demand. Various temperature scenario generation methods have been proposed in the literature, such as direct usage of the previous years of hourly temperatures with the dates fixed [128], shifting the historical temperatures by a few days to create additional scenarios [129], and bootstrapping the historical temperatures [130]. A comparison of these three methods based on pinball loss function was presented in [69]. The results showed that the shifted-date method dominated the other two when the number of dates being shifted is within a range. An empirical formula was also proposed to select parameters for the temperature scenario generation methods. The idea of generating temperature scenarios was also applied in [131]. An embedding based quantile regression neural network was used as the regression mode instead of MLR model, where the embedding layer can model the effect of calendar variables. In this way, the uncertainties of both future temperature and the relationship between temperature and load can be comprehensively considered. The scenario generation method was also used to develop a probabilistic view of power distribution system reliability indices [132].

On the output side, one can convert point forecasts to probabilistic ones via residual simulation or forecast combination. Several residual simulation methods were evaluated in [133]. The results showed that the residuals do not always

TABLE III
BRIEF SUMMARY OF THE LITERATURE ON LOAD FORECASTING

Load Forecasting	Key words	References
Without Individual Meters	Consumer Attrition / Demand Response	[103][104]
	Weather Modeling & Selection	[66] [67] [68] [48] [51] [53] [54] [55] [56]
	Traditional High Accurate Model	[49] [105] [106] [107] [108]
	Hierarchical Forecasting	[109] [110]
With Individual Meters	Traditional Methods	[111] [112]
	Sparse Coding / Deep Learning	[114] [115] [116] [113] [28]
	Clustering	[30] [117] [118] [119]
	Aggregation Load	[31] [120][137] [123] [121] [138][124]
	Evaluation Criteria	[125] [127] [115] [116] [32] [126]
Probabilistic Forecasting	Scenario Generation	[128] [129] [130] [69] [131][132]
	Residual Modeling & Output Ensemble	[133] [50] [66] [134]
	Probabilistic Forecasting Models	[22] [64] [32] [136]

follow a normal distribution, though group analysis increases the passing rate of normality tests. Adding simulated residuals under the normality assumption improves probabilistic forecasts from deficient models, while the improvement is diminishing as the underlying model improves. The idea of combining point load forecasts to generate probabilistic load forecasts was first proposed in [50]. The quantile regression averaging (QRA) method was applied to eight sister load forecasts, a set of point forecasts generated from homogeneous models developed in [66]. A constrained QRA (CQRA) was proposed in [134] to combine a series of quantiles obtained from individual quantile regression models.

Both approaches mentioned above rely on point forecasting models. It is still an unsolved question whether a more accurate point forecasting model can lead to a more skilled probabilistic forecast within this framework. An attempt was made in [135] to answer this question. The finding is that when the two underlying models are significantly different w.r.t. the point forecast accuracy, a more accurate point forecasting model would lead to a more skilled probabilistic forecast.

Various probabilistic forecasting models have been proposed by statisticians and computer scientists, such as quantile regression, Gaussian process regression, and density estimation. These off-the-shelf models can be directly applied to generate probabilistic load forecasts [22]. In GEFCom2014, a winning team developed a quantile generalized additive model (quantGAM), which is a hybrid of quantile regression and generalized additive models [64]. Probabilistic load forecasting has also been conducted on individual load profiles. Combining the gradient boosting method and quantile regression, a boosting additive quantile regression method was proposed in [32] to quantify the uncertainty and generate probabilistic forecasts. Apart from the quantile regression model, kernel density estimation methods were tested in [136]. The density of electricity data was modeled using different implementations of conditional kernel density (CKD) estimators to accommodate the seasonality in consumption. A decay parameter was used in the density estimation model for recent effects. The selection of kernel bandwidths and the presence of boundary effects are two main challenges with the implementation of CKD that were also investigated.

D. Remarks

Table III provides the correspondence between the key techniques and the surveyed references in smart meter data analytics for load forecasting.

Forecasting the loads at aggregate levels is a relatively mature area. Nevertheless, there are some nuances in the smart grid era due to the increasing need of highly accurate load forecasts. One is on the evaluation methods. Many forecasts are being evaluated using widely used error measures such as MAPE, which does not consider the consequences of over- or under-forecasts. In reality, the cost to the sign and magnitude of errors may differ significantly. Therefore, the following research question rises: how can the costs of forecast errors be integrated into the forecasting processes? Some research in this area would be helpful to bridge the gap between forecasting and decision making. The second one is load transfer detection, which is a rarely touched area in the literature. Distribution operators may transfer the load from one circuit to another permanently, seasonally, or on an ad hoc basis, in response to maintenance needs or reliability reasons. These load transfers are often poorly documented. Without smart meter information, it is difficult to physically trace the load blocks being transferred. Therefore, a data-driven approach is necessary in these situations. The third one is hierarchical forecasting, specifically, how to fully utilize zonal, regional, or meter load and local weather data to improve the load forecast accuracy. In addition, it is worth studying how to reconcile the forecasts from different levels for the applications of aggregators, system operators, and planners. The fourth one is on the emerging factors that affect electricity demand. The consumer behaviors are being changed by many modern technologies, such as rooftop solar panels, large batteries, and smart home devices. It is important to leverage the emerging data sources, such as technology adoption, social media, and various marketing surveys.

To comprehensively capture the uncertainties in the future, researchers and practitioners recently started to investigate in probabilistic load forecasting. Several areas within probabilistic load forecasting would need some further attention. First, distributed energy resources and energy storage options often disrupt the traditional load profiles. Some research is needed to generate probabilistic net load forecasts for the system with high penetration of renewable energy and large scale storage.

Secondly, forecast combination is widely regarded in the point forecasting literature as an effective way to enhance the forecast accuracy. There is a primary attempt in [134] to combine quantile forecasts. Further investigations can be conducted on combining other forms probabilistic forecasts, such as density forecasts and interval forecasts. Finally, the literature of probabilistic load forecasting for smart meters is still quite limited. Since the meter-level loads are more volatile than the aggregate loads, probabilistic forecasting has a natural application in this area.

IV. Load Management

How smart meter data contribute to the implementation of load management is summarized from three aspects in this section: the first one is to have a better understanding of sociodemographic information of consumers to provide better and personalized service. The second one is to target the potential consumers for demand response program marketing. The third one is the issue related to demand response program implementation including price design for price-based demand response and baseline estimation for incentive-based demand response.

A. Consumer Characterization

The electricity consumption behaviors of the consumers are closely related to their sociodemographic status. Bridging the load profiles to sociodemographic status is an important approach to classify the consumers and realize personalized services. A naive problem is to detect the consumer types according to the load profiles. The other two issues are identifying sociodemographic information from load profiles and predicting the load shapes using the sociodemographic information.

Identifying the type of consumers can be realized by simple classification. The temporal load profiles were first transformed into the frequency domain in [139] using fast Fourier transformation (FFT). Then the coefficients of different frequencies were used as the inputs of classification and regression tree (CART) to place consumers in different categories. FFT decomposes smart meter data based on a certain sine function and cosine function. Another transformation technique, sparse coding, has no assumption on the base signal but learns them automatically. Non-negative sparse coding was applied to extract the partial usage patterns from original load profiles in [27]. Based on the partial usage patterns, linear SVM was implemented to classify the consumers into residents and small and medium-sized enterprises (SME). The classification accuracy is considerably higher than discrete wavelet transform (DWT) and PCA.

There are still consumers without smart meter installations. External data, such as the sociodemographic status of consumers, are applied to estimate their load profiles. Clustering was first implemented to classify consumers into different energy behavior groups, and then energy behavior correlation rate (EBCR) and indicator dominance index (IGD) were defined and calculated to identify the indicators higher than a threshold [33]. Finally, the relationship between different energy behavior groups and their sociodemographic status was mapped. Spectral clustering was applied to generate typical load profiles, which were then used as the inputs of predictors such as random forests (RF) and stochastic boosting (SB) in [140]. The results showed that with commercial and cartographic data, the load profiles of consumers can be accurately predicted. Stepwise selection was applied to investigate the factors that have a great influence on residential electricity consumption in [141]. The location, floor area, the age of consumers, and the number of appliances are main factors, while the income level and home ownership have little relationship with consumption. A multiple linear regression model was used to bridge the total electricity consumption, maximum demand, load factor, and ToU to dwelling and occupant socioeconomic variables in [142]. The factors that have a great impact on total consumption, maximum load, load factor, and ToU were identified. The influence of socioeconomic status of consumers' electricity consumption patterns was evaluated in [143]. Random forest (RF) regression was proposed to combine socioeconomic status and environmental factors to predict the consumption patterns.

More works focus on how to mine the sociodemographic information of consumers from the massive smart meter data. One approach is based on a clustering algorithm. DPMM was applied in [144] for household and business premise load profiling where the number of clusters was not required to predetermined. The clustering results obtained by the DPMM algorithm have a clear corresponding relation with the metadata of dwellings, such as the nationality, household size, and type of dwelling. Based on the clustering results, multinomial logistic regression was applied to the clusters and dwelling and appliance characteristics in [34]. Each cluster was analyzed according to the coefficients of the regression model. Feature extraction and selection have also been applied as the attributes of the classifier. A feature set including the average consumption over a certain period, the ratios of two consumptions in different periods, and the temporal properties was established in [35]. Then, classification or regression was implemented to predict the sociodemographic status according to these features. Results showed that the proposed feature extraction method outperform biased random guess. More than 88 features from consumption, ratios, statistics, and temporal characteristics were extracted, and then correlation, KS-test, and η^2-based feature selection methods were conducted in [145]. The so-called extend CLASS classification framework was used to forecast the deduced properties of private dwellings. A supervised classification algorithm called dependent-independent data classification (DID-Class) was proposed to address the challenges of dependencies among multiple classification-relevant variables in [146]. The characteristics of dwellings were recognized based on this method, and comparisons with SVM and traditional CLASS proposed in [35] were conducted. The accuracy of DID-Class with SVM and CLASS is slightly higher than those of SVM and CLASS. To capture the intra-day and inter-day electricity consumption behavior of the consumers, a two-dimensional convolutional neural network (CNN) was used in [37] to make a bridge between the smart meter data and sociodemographic information of the consumers. The deep learning method can extract the features automatically and outperforms traditional methods.

B. Demand Response Program Marketing

Demand response program marketing is to target the consumers who have a large potential to be involved in demand response programs. On one hand, 15 minute or half-hour smart meter data cannot provide detail information on the operation status of the appliance; on the other hand, the altitude of consumers towards demand response is hard to model. Thus, the demand response potential cannot be evaluated directly. In this subsection, the potential of demand response can be indirectly evaluated by analyzing the variability, sensitivity to temperature, and so forth.

Variability is a key index for evaluating the potential of demand response. A hidden Markov model (HMM)-based spectral clustering was proposed in [147] to describe the magnitude, duration, and variability of the electricity consumption and further estimate the occupancy states of consumers. The information on the variability, occupancy states, and inter-temporal consumption dynamics can help retailers or aggregators target suitable consumers at different time scales. Both adaptive k-means and hierarchical clustering were used to obtain the typical load shapes of all the consumers within a certain error threshold in [148]. The entropy of each consumer was then calculated according to the distribution of daily load profiles over a year, and the typical shapes of load profiles were analyzed. The consumers with lower entropy have relatively similar consumption patterns on different days and can be viewed as a greater potential for demand response because their load profiles are more predictable. Similarly, the entropy was calculated in [29] based on the state transition matrix. It was stated that the consumers with high entropy are suitable for price-based demand response for their flexibility to adjust their load profile according to the change in price, whereas the consumers with low entropy are suitable for incentive-based demand response for their predictability to follow the control commands.

Estimation of electricity reduction is another approach for demand response potential. A mixture model clustering was conducted on a survey dataset and smart meter data in [100] to evaluate the potential for active demand reduction with wet appliances. The results showed that both the electricity demand of wet appliances and the attitudes toward demand response have a great influence on the potential for load shifting. Based on the GMM model of the electricity consumption of consumers and the estimated baseline, two indices, i.e., the possibility of electricity reduction greater than or equal to a certain value and the least amount of electricity reduction with a certain possibility, were calculated in [149]. These two indices can help demand response implementers have a probabilistic understanding of how much electricity can be reduced. A two-stage demand response management strategy was proposed in [150], where SVM was first used to detect the devices and users with excess load consumption and then a load balancing algorithm was performed to balance the overall load.

Since appliances such as heating, ventilation and air conditioning (HVAC) have great potential for demand response, the sensitivity of electricity consumption to outdoor air temperature is an effective evaluation criterion. Linear regression was applied to smart meter data and temperature data to calculate this sensitivity, and the maximum likelihood approach was used to estimate the changing point in [151]. Based on that, the demand response potentials at different hours were estimated. Apart from the simple regression, an HMM-based thermal regime was proposed to separate the original load profile into the thermal profile (temperature-sensitive) and base profile (non-temperature-sensitive) in [152]. The demand response potential can be calculated for different situations, and the proposed method can achieve much more savings than random selection. A thermal demand response ranking method was proposed in [153] for demand response targeting, where the demand response potential was evaluated from two aspects: temperature sensitivity and occupancy. Both linear regression and breakpoint detection were used to model the thermal regimes; the true linear response rate was used to detect the occupancy.

C. Demand Response Implementation

Demand response can be roughly divided into price-based demand response and incentive-based demand response. Price design is an important business model to attract consumers and maximize profit in price-based demand response programs; baseline estimation is the basis of quantifying the performance of consumers in incentive-based demand response programs. The applications of smart meter data analytics in price design and baseline estimation are summarized in this subsection.

For tariff design, an improved weighted fuzzy average (WFA) K-means was first proposed to obtain typical load profiles in [154]. An optimization model was then formulated with a designed profit function, where the acceptance of consumers over price was modeled by a piecewise function. The similar price determination strategy was also presented in [155]. Conditional value at risk (CVaR) for the risk model was further considered in [156] such that the original optimization model becomes a stochastic one. Different types of clustering algorithms were applied to extract load profiles with a performance index granularity guided in [157]. The results showed that different clusterings with different numbers of clusters and algorithms lead to different costs. GMM clustering was implemented on both energy prices and load profiles in [158]. Then, ToU tariff was developed using different combinations of the classifications of time periods. The impact of the designed price on demand response was finally quantified.

For baseline estimation, five naive baseline methods, HighXofY, MidXofY, LowXofY, exponential moving average, and regression baselines, were introduced in [159]. Different demand response scenarios were modeled and considered. The results showed that bias rather than accuracy is the main factor for deciding which baseline provides the largest profits. To describe the uncertainty within the consumption behaviors

TABLE IV
BRIEF SUMMARY OF THE LITERATURE ON LOAD MANAGEMENT

Load Management	Key words	References
Consumer Characterization	Consumer Type	[27] [139]
	Load Profile Prediction	[33] [140] [141] [142] [143]
	Socio-demographic Status Prediction	[144] [34] [35][37] [145] [146]
Demand Response Program Marketing	Variability	[147] [148] [29]
	Electricity Reduction	[100] [149][150]
	Temperature Sensitivity	[151] [152][153]
Demand Response Implementation	Tariff Design	[154][155] [156] [157] [158]
	Baseline Estimation	[159] [160] [161] [162]

of consumers, Gaussian-process-based probabilistic baseline estimation was proposed in [160]. In addition, how the aggregation level influences the relative estimation error was also investigated. K-means clustering of the load profiles in non-event days was first applied in [161], and a decision tree was used to predict the electricity consumption level according to demographics data, including household characteristics and electrical appliances. Thus, a new consumer can be directly classified into a certain group before joining the demand response program and then simple averaging and piecewise linear regression were used to estimate to baseline load in different weather conditions. Selecting a control group for baseline estimation was formulated as an optimization problem in [162]. The objective was to minimize the difference between the load profiles of the control group and demand response group when there is no demand response event. The problem was transformed into a constrained regression problem.

D. Remarks

Table IV provides the correspondence between the key techniques and the surveyed references in smart meter data analytics for load management.

For consumer characterization, it is essentially a high dimensional and nonlinear classification problem. There are at least two ways to improve the performance of consumer characterization: 1) conducting feature extraction or selection; 2) developing classification models. In most existing literature, the features for consumer characterization are manually extracted. A data-driven feature extraction method might be an effective way to further improve the performance. The classification is mainly implemented by the shallow learning models such as ANN and SVM. We can try different deep learning networks to tackle the high nonlinearity. We also find that the current works are mainly based on the Irish dataset [25]. Low Carbon London dataset may be another good choice. More open datasets are needed to enrich the research in this area.

For demand response program marketing, evaluating the potential for load shifting or reduction is an effective way to target suitable consumers for different demand response programs. Smart meter data with a frequency of 30 minutes or lower cannot reveal the operation states of the individual appliance; thus, several indirect indices, including entropy, sensitivity to temperature and price, are used. More indices can be further proposed to provide a comprehensive understanding of the electricity consumption behavior of consumers. Since most papers target potential consumers for demand response according to the indirect indices, a critical question is why and how these indices can reflect the demand response potential without experimental evidence? More real-world experimental results are welcomed for the research.

For demand response implementation, all the price designs surveyed above are implemented with a known acceptance function against price. However, the acceptance function or utility function is hard to estimate. How to obtain the function has not been introduced in existing literature. If the used acceptance function or utility function is different from the real one, the obtained results will deviate from the optimal results. Sensitivity analysis of the acceptance function or utility function assumption can be further conducted. Except for traditional tariff design, some innovative prices can be studied, such as different tariff packages based on fine-grained smart meter data. For baseline estimation, in addition to deterministic estimation, probabilistic estimation methods can present more future uncertainties. Another issue is how to effectively incorporate the deterministic or probabilistic baseline estimation results into demand response scheduling problem.

V. MISCELLANIES

In addition to the three main applications summarized above, the works on smart meter data analytics also cover some other applications, including power network connection verification, outage management, data compression, data privacy, and so forth. Since only several trials have been conducted in these areas and the works in the literature are not so rich, the works are summarized in this miscellanies section.

A. Connection Verification

The distribution connection information can help utilities and DSO make the optimal decision regarding the operation of the distribution system. Unfortunately, the entire topology of the system may not be available especially at low voltage levels. Several works have been conducted to identify the connections of different demand nodes using smart meter data.

Correlation analysis of the hourly voltage and power consumption data from smart meters were used to correct connectivity errors in [163]. The analysis assumed that the voltage magnitude decreases downstream along the feeder. However, the assumption might be incorrect when there is a large amount of distributed renewable energy integration. In addition to consumption data, both the voltage and current data were used in [164] to estimate the topology of the distribution system secondary circuit and the impedance of each branch. This estimation was conducted in a greedy fashion rather than an exhaustive search to enhance computational efficiency. The topology identification problem was formulated as an optimization problem minimizing the mutual-information-based Kullback-Leibler (KL) divergence between each two voltage time series in [165]. The effectiveness of mutual information was discussed from the perspective of conditional probability. Similarly, based on the assumption that the correlation between interconnected neighboring buses is higher than that between non-neighbor buses, the topology identification problem was formulated as a probabilistic graph model and a Lasso-based sparse estimation problem in [166]. How to choose the regularization parameter for Lasso regression was also discussed.

The electricity consumption data at different levels were analyzed by PCA in [167] for both phase and topology identification where the errors caused by technical loss, smart metering, and clock synchronization were formulated as Gaussian distributions. Rather than using all smart meter data, a phase identification problem with incomplete data was proposed in [168] to address the challenge of bad data or null data. The high-frequency load was first obtained by a Fourier transform, and then the variations in high-frequency load between two adjacent time intervals were extracted as the inputs of saliency analysis for phase identification. A sensitivity analysis on smart meter penetration ratios was performed and showed that over 95% accuracy can be achieved with only 10% smart meters.

B. Outage Management

A power outage is defined as an electricity supply failure, which may be caused by short circuits, station failure, and distribution line damage [169]. Outage management is viewed as one of the highest priorities of smart meter data analytics behind billing. It includes outage notification (or last gasp), outage location and restoration verification.

How the outage management applications work, the data requirements, and the system integration considerations were introduced in [170]. The outage area was identified using a two-stage strategy in [171]. In the first stage, the physical distribution network was simplified using topology analysis; in the second stage, the outage area was identified using smart meter information, where the impacts of communication were also considered. A smart meter data-based outage location prediction method was proposed in [172] to rapidly detect and recover the power outages. The challenges of smart meter data utilization and required functions were analyzed. Additionally, as a way to identify the faulted section on a feeder or lateral, a new multiple-hypothesis method was proposed in [173], where the outage reports from smart meters were used as the input of the proposed multiple-hypothesis method. The problem was formulated as an optimization model to maximize the number of smart meter notifications. A novel hierarchical framework was established in [174] for outage detection using smart meter event data rather than consumption data. It can address the challenges of missing data, multivariate count data, and variable selection. How to use data analytics method to model the outages and reliability indices from weather data was discussed in [132]. Apart from the data analytics method for outage management, more works on smart meter data-based outage managements have been adopted to the corresponding communication architectures [175], [176].

C. Data Compression

Massive smart meter data present more challenges with respect to data communication and storage. Compressing smart meter data to a very small size and without (large) loss can ease the communication and storage burden. Data compression can be divided into lossy compression and lossless compression. Different compression methods for electric signal waveforms in smart grids are summarized in [177].

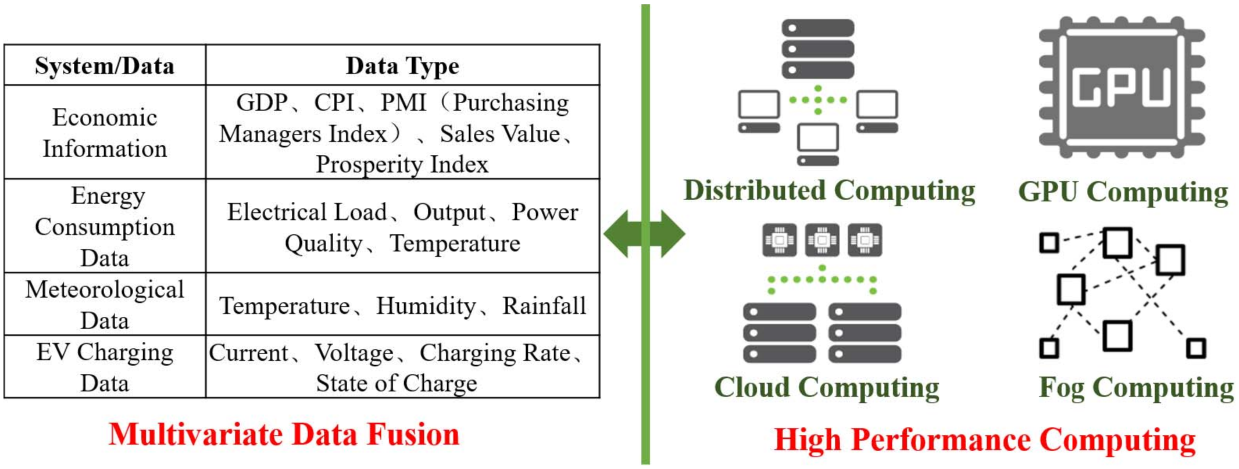
Some papers exist that specifically discuss the smart meter data compression problem. Note that the changes in electricity consumption in adjunct time periods are much smaller than the actual consumption, particularly for very high-frequency data. Thus, combining normalization, variable-length coding, and entropy coding, and the differential coding method was proposed in [178] for the lossless compression of smart meter data. While different lossless compression methods, including IEC 62056-21, A-XDR, differential exponential Golomb and arithmetic (DEGA) coding, and Lempel Ziv Markov chain algorithm (LZMA) coding, were compared on REDD and SAG datasets in [179]. The performances on the data with different granularities were investigated. The results showed that these lossless compression methods have better performance on higher granularity data.

For low granularity (such as 15 minutes) smart meter data, symbolic aggregate approximation (SAX), a classic time series data compression method, was used in [29] and [180] to reduce the dimensionality of load profiles before clustering. The distribution of load profiles was first fitted by generalized extreme value in [36]. A feature-based load data compression method (FLDC) was proposed through defining the base state and stimulus state of the load profile and detecting the change in load status. Comparisons with the piecewise aggregate approximation (PAA), SAX, and discrete wavelet transform (DWT) were conducted. Non-negative sparse coding was applied to transform original load profiles into a higher dimensional space in [27] to identify the partial usage patterns and compress the load in a sparse way.

D. Data Privacy

One of the main oppositions and concerns for the installation of smart meters is the privacy issue. The sociodemographic information can be inferred from the fine-grained smart meter data, as introduced in Section IV. There are several works in the literature that discuss how to preserve the privacy of consumers.

A study on the distributed aggregation architecture for additive smart meter data was conducted in [181]. A secure communication protocol was designed for the gateways placed at the consumers premises to prevent revealing individual data information. The proposed communication protocol can be implemented in both centralized and distributed manners.



System/Data	Data Type
Economic Information	GDP、CPI、PMI（Purchasing Managers Index）、Sales Value、Prosperity Index
Energy Consumption Data	Electrical Load、Output、Power Quality、Temperature
Meteorological Data	Temperature、Humidity、Rainfall
EV Charging Data	Current、Voltage、Charging Rate、State of Charge

Fig. 8. Big data issues with smart meter data analytics.

A framework for the trade-off between privacy and utility requirement of consumers was presented in [182] based on a hidden Markov model. The utility requirement was evaluated by the distortion between the original and the perturbed data, while the privacy was evaluated by the mutual information between the two data sequences. Then, a utility-privacy trade-off region was defined from the perspective of information theory. This trade-off was also investigated in [183], where the attack success probability was defined as an objective function to be minimized and ε-privacy was formulated. The aggregation of individual smart meter data and the introduction of colored noise were used to reduce the success probability.

Edge detection is one main approach for NILM to identify the status of appliances. How the data granularity of smart meter data influences the edge detection performance was studied in [184]. The results showed that when the data collection frequency is lower than half the on-time of the appliance, the detection rate dramatically decreases. The privacy was evaluated by the F-score of NILM. The privacy preservation problem was formulated as an optimization problem in [185], where the objective was to minimize the sum of the expected cost, disutility of consumers caused by the late use of appliances, and information leakage. Eight privacy-enhanced scheduling strategies considering on-site battery, renewable energy resources, and appliance load moderation were comprehensively compared.

VI. Open Research Issues

Although smart meter data analytics has received extensive attention and rich literature studies related to this area have been published, developments in computer science and the energy system itself will certainly lead to new problems or opportunities. In this section, several works on smart meter data analytics in the future smart grid are highlighted.

A. Big Data Issues

Substantial works in the literature have conducted smart meter data analytics. Two special sections about big data analytics for smart grid modernization were hosted in IEEE Transactions on Smart Grid in 2016 [186] and IEEE Power and Energy Magazine in 2018, respectively [187]. However, the size of the dataset analyzed can hardly be called big data. How to efficiently integrate more multivariate data with a larger size to discover more knowledge is an emerging issue. As shown in Fig. 8, big data issues with smart meter data analytics include at least two aspects: the first is multivariate data fusion, such as economic information, meteorological data, and EV charging data apart from energy consumption data; the second is high-performance computing, such as distributed computing, GPU computing, cloud computing, and fog computing.

1) Multivariate Data Fusion: The fusion of various data is one of the basic characteristics of big data [188]. Current studies mainly only focus on the smart meter data itself or even electricity consumption data. Very few papers consider weather data, survey data from consumers, or some other data. Integrating more external data, such as crowd-sourcing data from the Internet, weather data, voltage and current data, and even voice data from service systems may reveal more information. The multivariate data fusion needs to deal with structured data with different granularities and unstructured data. We would like to emphasis that big data is a change of concept. More data-driven methods will be proposed to solve practical problems that may traditionally be solved by model-based methods. For example, with redundant smart meter data, the power flow of the distribution system can be approximated through hyperplane fitting methods such as ANN and SVM. In addition, how to visualize high dimensional and multivariate data to highlight the crucial components and discover the hidden patterns or correlations among these data is a very seldom touched area [189].

2) High-Performance Computation: In addition, a majority of smart meter data analytics methods that are applicable to small data sets may not be appropriate for large data sets. Highly efficient algorithms and tools such as distributed and parallel computing and the Hadoop platform should be further investigated. Cloud computing, an efficient computation architecture that shares computing resources on the Internet, can provide different types of big data analytics services, including Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructure as a Service (IaaS) [190]. How to make full use of cloud computing resources for smart meter data analytics is an important issue. However, the security problem introduced by cloud computing should be addressed [191]. Another high-performance computation approach is GPU computation. It can realize highly efficient parallel computation [192]. Specific algorithms should be designed for the implementation of different GPU computation tasks.

B. New Machine Learning Technologies

Smart meter data analytics is an interdisciplinary field that involves electrical engineering and computer science, particularly machine learning. The development of machine learning has had great impacts on smart meter data analytics. The application of new machine learning technologies is an important aspect of smart meter analytics. The recently proposed clustering method in [193] has been used in [29]; the progress in deep learning in [194] has been used in [195]. When applying one machine learning technology to smart meter data analytics, the limitations of the method and the physical meaning revealed by the method should be carefully considered. For example, the size of data or samples should be considered in deep learning to avoid overfitting.

1) Deep Learning and Transfer Learning: Deep learning has been applied in different industries, including smart grids. As summarized above, different deep learning techniques have been used for smart meter data analytics, which is just a start. Designing different deep learning structures for different applications is still an active research area. The lack of label data is one of the main challenges for smart meter data analytics. How to apply the knowledge learned for other objects to the research objects using transfer learning can help us fully utilize various data [196]. Many transfer learning tasks are implemented by deep learning [197]. The combination of these two emerging machine learning techniques may have widespread applications.

2) Online Learning and Incremental Learning: Note that smart meter data are essentially real-time stream data. Online learning and incremental learning are varied suitably for handling these real-time stream data [198]. Many online learning techniques, such as online dictionary learning [199] and incremental learning machine learning techniques such as incremental clustering [200], have been proposed in other areas. However, existing works on smart meter data analytics rarely use online learning or incremental learning, expect for several online anomaly detection methods.

C. New Business Models in Retail Market

Further deregulation of retail markets, integration of distributed renewable energy, and progress in information technologies will hasten various business models on the demand side.

1) Transactive Energy: In a transactive energy system [201], [202], the consumer-to-consumer (C2C) business model or micro electricity market can be realized, i.e., the consumer with rooftop PV becomes a prosumer and can trade electricity with other prosumers. The existing applications of smart meter data analytics are mainly studied from the perspectives of retailers, aggregators, and individual consumers. How to analyze the smart data and how much data should be analyzed in the micro electricity market to promote friendly electricity consumption and renewable energy accommodation is a new perspective in future distribution systems.

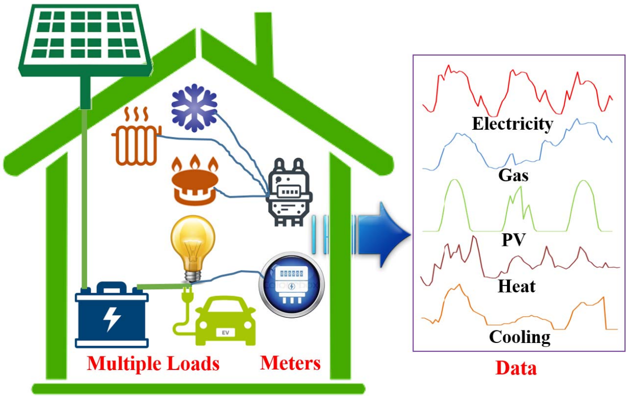


Fig. 9. Transition of energy systems on the demand side.

2) Sharing Economy: For the distribution system with distributed renewable energy and energy storage integration, a new business model sharing economy can be introduced. The consumers can share their rooftop PV [203] and storage [204] with their neighborhoods. In this situation, the roles of consumers, retailers, and DSO will change when playing the game in the energy market [205]. Other potential applications of smart meter data analytics may exist, such as changes in electricity purchasing and consumption behavior and optimal grouping strategies for sharing energy.

D. Transition of Energy Systems

As shown in Fig. 9, the integration of the distributed renewable energy and multiple energy systems is an inevitable trend in the development of smart grids. A typical smart home has multiple loads, including cooling, heat, gas, and electricity. These newcomers such as rooftop PV, energy storage, and EV also change the structure of future distribution systems.

1) High Penetration of Renewable Energy: High penetration of renewable energy such as behind-the-meter PV [206], [207] will greatly change the electricity consumption behavior and will significantly influence the net load profiles. Traditional load profiling methods should be improved to consider the high penetration of renewable energy. In addition, by combining weather data, electricity price data, and net load data, the estimation of renewable energy capacity and output can be estimated. In this way, the original load profile can be recovered. Energy storage is widely used to stabilize renewable energy fluctuations. However, the charging or discharging behavior of storage, particularly the behind-the-meter storage [208], is difficult to model and meter. Advanced data analytical methods need to be adopted for anomaly detection, forecasting, outage management, decision making, and so forth in high renewable energy penetration environments.

2) Multiple Energy Systems: Multiple energy systems integrate gas, heat, and electricity systems together to boost the efficiency of the entire energy system [209]. The consumptions for electricity, heat, cooling, and gas are coupled in the future retailer market. One smart meter can record the consumptions of these types of energy simultaneously. Smart meter data analytics is no longer limited to electricity consumption data. For example, joint load forecasting for electricity, heating, and cooling can be conducted for multiple energy systems.

E. Data Privacy and Security

As stated above, the concern regarding smart meter privacy and security is one of the main barriers to the privilege of smart meters. Many existing works on the data privacy and security issue mainly focus on the data communication architecture and physical circuits [210]. How to study the data privacy and security from the perspective of data analytics is still limited.

1) Data Privacy: Analytics methods for data privacy is a new perspective except for communication architecture, such as the design of privacy-preserving clustering algorithm [211] and PCA algorithm [212]. A strategic battery storage charging and discharging schedule was proposed in [213] to mask the actual electricity consumption behavior and alleviate the privacy concerns. However, several basic issues about smart meter data should be but have not been addressed: Who owns the smart meter data? How much private information can be mined from these data? Is it possible to disguise data to protect privacy and to not influence the decision making of retailers?

2) Data Security: For data security, the works on cyber physical security (CPS) in the smart grid such as phasor measurement units (PMU) and supervisory control and data acquisition (SCADA) data attack have been widely studied [214]. However, different types of cyber attacks for electricity consumption data such as NTL should be further studied [215].

VII. Conclusion

In this paper, we have provided a comprehensive review of smart meter data analytics in retail markets, including the applications in load forecasting, abnormal detection, consumer segmentation, and demand response. The latest developments in this area have been summarized and discussed. In addition, we have proposed future research directions from the prospective big data issue, developments of machine learning, novel business model, energy system transition, and data privacy and security. Smart meter data analytics is still an emerging and promising research area. We hope that this review can provide readers a complete picture and deep insights into this area.

References

[1] "Smart meters, quarterly report to end December 2016, great Britain," Dept. Bus. Energy Ind. Strategy, London, U.K., Rep., 2017.

[2] *How Many Smart Meters are Installed in the United States, and Who Has Them?* Accessed: Jul. 31, 2017. [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=108&t=3

[3] *Number of Electric Smart Meters Deployed in the U.S. From 2008 to 2016*. Accessed: Jul. 31, 2017. [Online]. Available: https://www.statista.com/statistics/499704/number-of-smart-meters-in-the-united-states/

[4] R. R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure," *Int. J. Elect. Power Energy Syst.*, vol. 63, pp. 473–484, Dec. 2014.

[5] J. Yang, J. Zhao, F. Luo, F. Wen, and Z. Y. Dong, "Decision-making for electricity retailers: A brief survey," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2651499.

[6] National Science Foundation. (2016). *Smart Grids Big Data*. [Online]. Available: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1636772&HistoricalAwards=false

[7] X. Liu, A. Heller, and P. S. Nielsen, "CITIESData: A smart city data management framework," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 699–722, 2017.

[8] *Bits to Energy Lab Projects*. Accessed: Jul. 31, 2017. [Online]. Available: http://www.bitstoenergy.ch/home/projects/

[9] Siebel Energy Institute. (2016). *Advancing the Science of Smart Energy*. [Online]. Available: http://www.siebelenergyinstitute.org/

[10] *WP3 Overview*. Accessed: Jul. 31, 2017. [Online]. Available: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_overview

[11] "Utility analytics in 2017: Aligning data and analytics with business strategy," SAS, Cary, NC, USA, Rep., 2017.

[12] T. Hong, D. W. Gao, T. Laing, D. Kruchten, and J. Calzada, "Training energy data scientists: Universities and industry need to work together to bridge the talent gap," *IEEE Power Energy Mag.*, vol. 16, no. 3, pp. 66–73, May/Jun. 2018.

[13] J. Hu and A. V. Vasilakos, "Energy big data analytics and security: Challenges and opportunities," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2423–2436, Sep. 2016.

[14] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, 2012.

[15] K.-L. Zhou, S.-L. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 24, pp. 103–110, Aug. 2013.

[16] Y. Wang *et al.*, "Load profiling and its application to demand response: A review," *Tsinghua Sci. Technol.*, vol. 20, no. 2, pp. 117–129, Apr. 2015.

[17] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Trans. Consum. Electron.*, vol. 57, no. 1, pp. 76–84, Feb. 2011.

[18] N. F. Esa, M. P. Abdullah, and M. Y. Hassan, "A review disaggregation method in non-intrusive appliance load monitoring," *Renew. Sustain. Energy Rev.*, vol. 66, pp. 163–173, Dec. 2016.

[19] T. Ahmad, "Non-technical loss analysis and prevention using smart meters," *Renew. Sustain. Energy Rev.*, vol. 72, pp. 573–589, May 2017.

[20] P. Glauner, J. A. Meira, P. Valtchev, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, pp. 760–775, 2017.

[21] G. Chicco, "Customer behaviour and data analytics," in *Proc. Int. Conf. Expo. Elect. Power Eng. (EPE)*, 2016, pp. 771–779.

[22] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecast.*, vol. 32, no. 3, pp. 914–938, 2016.

[23] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, Jul. 2017.

[24] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 425–436, Feb. 2016.

[25] Irish Social Science Data Archive. (2012). *Commission for Energy Regulation (CER) Smart Metering Project*. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[26] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[27] Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, "Sparse and redundant representation-based smart meter data compression and pattern extraction," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2142–2151, May 2017.

[28] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2686012.

[29] 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.

[30] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, Jan. 2014.

[31] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015.

[32] 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.

[33] X. Tong, R. Li, F. Li, and C. Kang, "Cross-domain feature selection and coding for household energy behavior," *Energy*, vol. 107, pp. 9–16, Jul. 2016.

[34] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. Energy*, vol. 141, pp. 190–199, Mar. 2015.

[35] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, Dec. 2014.

[36] X. Tong, C. Kang, and Q. Xia, "Smart metering load data compression based on load feature identification," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2414–2422, Sep. 2016.

[37] Y. Wang *et al.*, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2018.2805723.

[38] J. R. Schofield *et al.* (2016). *Low Carbon London Project: Data From the Dynamic Time-of-Use Electricity Pricing Trial*. Accessed: Aug. 8, 2017. [Online]. Available: https://discover.ukdataservice.ac.uk/catalogue?sn=7857, doi: 10.5255/UKDA-SN-7857-2.

[39] 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.

[40] M. Sun, Y. Wang, G. Strbac, and C. Kang, "Probabilistic peak load estimation in smart cities using smart meter data," *IEEE Trans. Ind. Electron.*, to be published, doi: 10.1109/TIE.2018.2803732.

[41] Pecan Street. (2012). *Real Energy. Real Customers. in Real Time*. [Online]. Available: http://www.pecanstreet.org/energy/

[42] C. Miller and F. Meggers, "The building data genome project: An open, public data set from non-residential building electrical meters," *Energy Procedia*, vol. 122, pp. 439–444, Sep. 2017.

[43] (2017). *Umass Smart* Dataset*. [Online]. Available: http://traces.cs.umass.edu/index.php/Smart/Smart

[44] A. K. Marnerides, P. Smith, A. Schaeffer-Filho, and A. Mauthe, "Power consumption profiling using energy time-frequency distributions in smart grids," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 46–49, Jan. 2015.

[45] O. Y. Al-Jarrah, Y. Al-Hammadi, P. D. Yoo, and S. Muhaidat, "Multi-layered clustering for power consumption profiling in smart grids," *IEEE Access*, vol. 5, pp. 18459–18468, 2017.

[46] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray, "Residential load and rooftop PV generation: An Australian distribution network dataset," *Int. J. Sustain. Energy*, vol. 36, no. 8, pp. 787–806, 2017.

[47] Ausgird. *Distribution Zone Substation Information Data to Share*. Accessed: Jul. 31, 2017. [Online]. Available: http://www.ausgrid.com.au/Common/About-us/Corporate-information/Data-to-share/DistZone-subs.aspx#.WYD6KenauUl

[48] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecast.*, vol. 30, no. 2, pp. 357–363, 2014.

[49] B. A. Høverstad, A. Tidemann, H. Langseth, and P. Öztürk, "Short-term load forecasting with seasonal decomposition using evolution for parameter tuning," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1904–1913, Jul. 2015.

[50] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic load forecasting via quantile regression averaging on sister forecasts," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 730–737, Mar. 2017.

[51] N. Charlton and C. Singleton, "A refined parametric model for short term load forecasting," *Int. J. Forecast.*, vol. 30, no. 2, pp. 364–368, 2014.

[52] V. Thouvenot, A. Pichavant, Y. Goude, A. Antoniadis, and J.-M. Poggi, "Electricity forecasting using multi-stage estimators of nonlinear additive models," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3665–3673, Sep. 2016.

[53] J. R. Lloyd, "GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes," *Int. J. Forecast.*, vol. 30, no. 2, pp. 369–374, 2014.

[54] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models," *Int. J. Forecast.*, vol. 30, no. 2, pp. 375–381, 2014.

[55] S. B. Taieb and R. J. Hyndman, "A gradient boosting approach to the Kaggle load forecasting competition," *Int. J. Forecast.*, vol. 30, no. 2, pp. 382–394, 2014.

[56] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *Int. J. Forecast.*, vol. 31, no. 2, pp. 286–295, 2015.

[57] T. Hong *et al.*, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecast.*, vol. 32, no. 3, pp. 896–913, 2016.

[58] A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi, "A prediction interval for a function-valued forecast model: Application to load forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 939–947, 2016.

[59] V. Dordonnat, A. Pichavant, and A. Pierrot, "GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1005–1011, 2016.

[60] J. Xie and T. Hong, "GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1012–1016, 2016.

[61] S. Haben and G. Giasemidis, "A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1017–1022, 2016.

[62] E. Mangalova and O. Shesterneva, "Sequence of nonparametric models for GEFCom2014 probabilistic electric load forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1023–1028, 2016.

[63] F. Ziel and B. Liu, "Lasso estimation for GEFCom2014 probabilistic electric load forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1029–1037, 2016.

[64] P. Gaillard, Y. Goude, and R. Nedellec, "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1038–1050, 2016.

[65] ISO New England. (2017). *ISO New England Zonal Information*. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info

[66] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: A big data approach," *Int. J. Forecast.*, vol. 32, no. 3, pp. 585–597, 2016.

[67] J. Xie, Y. Chen, T. Hong, and T. D. Laing, "Relative humidity for load forecasting models," *IEEE Trans. Smart Grid*, vol. 9, no. 1, pp. 191–198, Jan. 2018.

[68] J. Xie and T. Hong, "Wind speed for load forecasting models," *Sustainability*, vol. 9, no. 5, p. 795, 2017.

[69] J. Xie and T. Hong, "Temperature scenario generation for probabilistic load forecasting," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1680–1687, May 2018, doi: 10.1109/TSG.2016.2597178.

[70] C. Keerthisinghe, G. Verbic, and A. C. Chapman, "A fast technique for smart home management: ADP with temporal difference learning," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2016.2629470.

[71] A. Pratt *et al.*, "Transactive home energy management systems: The impact of their proliferation on the electric grid," *IEEE Electrific. Mag.*, vol. 4, no. 4, pp. 8–14, Dec. 2016.

[72] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nat. Energy*, vol. 3, no. 2, pp. 94–101, 2018.

[73] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[74] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Minneapolis, MN, USA, 2016, pp. 1–5.

[75] H. N. Akouemo and R. J. Povinelli, "Data improving in time series using ARX and ANN models," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3352–3359, Sep. 2017.

[76] 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.

[77] J. Luo, T. Hong, and M. Yue, "Real-time anomaly detection for very short-term load forecasting," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 2, pp. 235–243, 2018.

[78] H. Huang *et al.*, "False data separation for data security in smart grids," *Knowl. Informat. Syst.*, vol. 52, no. 3, pp. 815–834, 2017.

[79] Y. Wang, D. Li, and C. Kang, "Application of low-rank matrix factorization to bad data identification and recovering for bus load," *Power Syst. Technol.*, vol. 41, no. 6, pp. 1972–1979, Oct. 2016.

[80] A. Al-Wakeel, J. Wu, and N. Jenkins, "K-means based load estimation of domestic smart meter measurements," *Appl. Energy*, vol. 194, pp. 333–342, May 2017.

[81] A. Al-Wakeel, J. Wu, and N. Jenkins, "State estimation of medium voltage distribution networks using smart meter measurements," *Appl. Energy*, vol. 184, pp. 207–218, Dec. 2016.

[82] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. M. Capretz, and G. Bitsuamlak, "An ensemble learning framework for anomaly detection in building energy consumption," *Energy Build.*, vol. 144, pp. 191–206, Jun. 2017.

[83] X. Liu, N. Iftikhar, P. S. Nielsen, and A. Heller, "Online anomaly energy consumption detection using Lambda architecture," in *Proc. Int. Conf. Big Data Anal. Knowl. Disc.*, Porto, Portugal, 2016, pp. 193–209.

[84] K. Wang, B. Wang, and L. Peng, "CVAP: Validation for cluster analyses," *Data Sci. J.*, vol. 8, pp. 88–93, Apr. 2009.

[85] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and R. C. Green, "High performance computing for detection of electricity theft," *Int. J. Elect. Power Energy Syst.*, vol. 47, pp. 21–30, May 2013.

[86] A. Jindal *et al.*, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.

[87] L. A. P. Júnior *et al.*, "Unsupervised non-technical losses identification through optimum-path forest," *Elect. Power Syst. Res.*, vol. 140, pp. 413–423, Nov. 2016.

[88] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[89] V. Botev *et al.*, "Detecting non-technical energy losses through structural periodic patterns in AMI data," in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, 2016, pp. 3121–3130.

[90] H. Janetzko, F. Stoffel, S. Mittelstädt, and D. A. Keim, "Anomaly detection for visual analytics of power consumption data," *Comput. Graph.*, vol. 38, pp. 27–37, Feb. 2014.

[91] R. Granell, C. J. Axon, and D. C. H. Wallom, "Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3217–3224, Nov. 2015.

[92] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *Int. J. Elect. Power Energy Syst.*, vol. 55, pp. 437–448, Feb. 2014.

[93] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen, "Clustering of connection points and load modeling in distribution systems," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1255–1265, May 2013.

[94] C. Chelmis, J. Kolte, and V. K. Prasanna, "Big data analytics for demand response: Clustering over space and time," in *Proc. IEEE Int. Conf. Big Data*, Santa Clara, CA, USA, 2015, pp. 2223–2232.

[95] E. D. Varga, S. F. Beretka, C. Noce, and G. Sapienza, "Robust real-time load profile encoding and classification framework for efficient power systems operation," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1897–1904, Jul. 2015.

[96] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, "Feature construction and calibration for clustering daily load curves from smart-meter data," *IEEE Trans. Ind. Informat.*, vol. 12, no. 2, pp. 645–654, Apr. 2016.

[97] M. Piao, H. S. Shon, J. Y. Lee, and K. H. Ryu, "Subspace projection method based clustering analysis in load profiling," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2628–2635, Nov. 2014.

[98] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016.

[99] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 88–96, Feb. 2014.

[100] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1561–1569, Aug. 2013.

[101] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[102] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 153–160, Feb. 2012.

[103] J. Xie, T. Hong, and J. Stroud, "Long-term retail energy forecasting with consideration of residential customer attrition," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2245–2252, Sep. 2015.

[104] W. Hoiles and V. Krishnamurthy, "Nonparametric demand forecasting and detection of energy aware consumers," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 695–704, Mar. 2015.

[105] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 134–141, Feb. 2012.

[106] Y. Goude, R. Nedellec, and N. Kong, "Local short and middle term electricity load forecasting with semi-parametric additive models," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 440–446, Jan. 2014.

[107] N. Ding, Y. Bésanger, and F. Wurtz, "Next-day MV/LV substation load forecaster using time series method," *Elect. Power Syst. Res.*, vol. 119, pp. 345–354, Feb. 2015.

[108] N. Ding, C. Benoit, G. Foggia, Y. Bésanger, and F. Wurtz, "Neural network-based model design for short-term load forecast in distribution systems," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 72–81, Jan. 2016.

[109] X. Sun *et al.*, "An efficient approach to short-term load forecasting at the distribution level," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2526–2537, Jul. 2016.

[110] C. E. Borges, Y. K. Penya, and I. Fernández, "Evaluating combined load forecasting in large power systems and smart grids," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1570–1577, Aug. 2013.

[111] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy Build.*, vol. 49, pp. 591–603, Jun. 2012.

[112] H. Chitsaz, H. Shaker, H. Zareipour, D. Wood, and N. Amjady, "Short-term electricity load forecasting of buildings in microgrids," *Energy Build.*, vol. 99, pp. 50–60, Jul. 2015.

[113] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling, "Deep learning for estimating building energy consumption," *Sustain. Energy Grids Netw.*, vol. 6, pp. 91–99, Jun. 2016.

[114] A. Tascikaraoglu and B. M. Sanandaji, "Short-term residential electric load forecasting: A compressive spatio-temporal approach," *Energy Build.*, vol. 111, pp. 380–392, Jan. 2016.

[115] C.-N. Yu, P. Mirowski, and T. K. Ho, "A sparse coding approach to household electricity demand forecasting in smart grids," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 738–748, Mar. 2017.

[116] P. Li, B. Zhang, Y. Weng, and R. Rajagopal, "A sparse linear model and significance test for individual consumption prediction," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4489–4500, Nov. 2017.

[117] Y.-H. Hsiao, "Household electricity demand forecast based on context information and user daily schedule analysis from meter data," *IEEE Trans. Ind. Informat.*, vol. 11, no. 1, pp. 33–43, Feb. 2015.

[118] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2683461.

[119] J. Yang *et al.*, "K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy Build.*, vol. 146, pp. 27–37, Jul. 2017.

[120] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *Proc. IEEE Int. Conf. Big Data*, Santa Clara, CA, USA, 2015, pp. 879–887.

[121] 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.

[122] R. Sevlian and R. Rajagopal, "Short term electricity load forecasting on varying levels of aggregation," *arXiv preprint arXiv:1404.0058*, 2014.

[123] B. Stephen, X. Tang, P. R. Harvey, S. Galloway, and K. I. Jennett, "Incorporating practice theory in sub-profile models for short term aggregated residential load forecasting," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1591–1598, Jul. 2017.

[124] Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia, "An ensemble forecasting method for the aggregated load with sub profiles," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2018.2807985.

[125] J. J. M. Moreno, A. P. Pol, A. S. Abad, and B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013.

[126] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecast.*, vol. 32, no. 3, pp. 669–679, 2016.

[127] S. Haben, J. Ward, D. V. Greetham, C. Singleton, and P. Grindrod, "A new error measure for forecasts of household-level, high resolution electrical energy consumption," *Int. J. Forecast.*, vol. 30, no. 2, pp. 246–256, 2014.

[128] T. Hong, J. Wilson, and J. Xie, "Long term probabilistic load forecasting and normalization with hourly information," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 456–462, Jan. 2014.

[129] "PJM load forecast report January 2015 prepared by PJM resource adequacy planning department," PJM, Norristown, PA, USA, Rep., 2015. [Online]. Available: https://www.pjm.com/-/media/library/reports-notices/load-forecast/2015-load-forecast-report.ashx?la=en

[130] R. J. Hyndman and S. Fan, "Density forecasting for long-term peak electricity demand," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 1142–1153, May 2010.

[131] D. Gan, Y. Wang, S. Yang, and C. Kang, "Embedding based quantile regression neural network for probabilistic load forecasting," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 2, pp. 244–254, 2018.

[132] J. Black, A. Hoffman, T. Hong, J. Roberts, and P. Wang, "Weather data for energy analytics: From modeling outages and reliability indices to simulating distributed photovoltaic fleets," *IEEE Power Energy Mag.*, vol. 16, no. 3, pp. 43–53, May/Jun. 2018.

[133] J. Xie, T. Hong, T. Laing, and C. Kang, "On normality assumption in residual simulation for probabilistic load forecasting," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1046–1053, May 2017.

[134] Y. Wang *et al.*, "Combining probabilistic load forecasts," *IEEE Trans. Smart Grid*, to be published.

[135] J. Xie and T. Hong, "Variable selection methods for probabilistic load forecasting: Empirical evidence from seven states of the united states," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2702751.

[136] S. Arora and J. W. Taylor, "Forecasting electricity smart meter data using conditional kernel density estimation," *Omega*, vol. 59, pp. 47–59, Mar. 2016.

[137] P. Zhang, X. Wu, X. Wang, and S. Bi, "Short-term load forecasting based on big data technologies," *CSEE J. Power Energy Syst.*, vol. 1, no. 3, pp. 59–67, Sep. 2015.

[138] S. Humeau, T. K. Wijaya, M. Vasirani, and K. Aberer, "Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households," in *Proc. Sustain. Internet ICT Sustain. (SustainIT)*, 2013, pp. 1–6.

[139] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2434–2441, Sep. 2015.

[140] D. Vercamer, B. Steurtewagen, D. V. den Poel, and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3693–3701, Sep. 2016.

[141] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, pp. 184–194, Jun. 2013.

[142] F. McLoughlin, A. Duffy, and M. Conlon, "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study," *Energy Build.*, vol. 48, pp. 240–248, May 2012.

[143] Y. Han, X. Sha, E. Grover-Silva, and P. Michiardi, "On the impact of socio-economic factors on power load forecasting," in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, 2014, pp. 742–747.

[144] R. Granell, C. J. Axon, and D. C. H. Wallom, "Clustering disaggregated load profiles using a Dirichlet process mixture model," *Energy Convers. Manag.*, vol. 92, pp. 507–516, Mar. 2015.

[145] K. Hopf, M. Sodenkamp, I. Kozlovkiy, and T. Staake, "Feature extraction and filtering for household classification based on smart electricity meter data," *Comput. Sci. Res. Develop.*, vol. 31, no. 3, pp. 141–148, 2016.

[146] M. Sodenkamp, I. Kozlovskiy, and T. Staake, "Supervised classification with interdependent variables to support targeted energy efficiency measures in the residential sector," *Decis. Anal.*, vol. 3, no. 1, p. 1, Dec. 2016.

[147] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4019–4030, Nov. 2013.

[148] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 420–430, Jan. 2014.

[149] Y. Bai, H. Zhong, and Q. Xia, "Real-time demand response potential evaluation: A smart meter driven method," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Boston, MA, USA, 2016, pp. 1–5.

[150] A. Jindal, N. Kumar, and M. Singh, "A data analytical approach using support vector machine for demand response management in smart grid," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Boston, MA, USA, 2016, pp. 1–5.

[151] M. E. H. Dyson, S. D. Borgeson, M. D. Tabone, and D. S. Callaway, "Using smart meter data to estimate demand response potential, with application to solar energy integration," *Energy Policy*, vol. 73, pp. 607–619, Oct. 2014.

[152] A. Albert and R. Rajagopal, "Thermal profiling of residential energy use," *IEEE Trans. Power Syst.*, vol. 30, no. 2, pp. 602–611, Mar. 2015.

[153] A. Albert and R. Rajagopal, "Finding the right consumers for thermal demand-response: An experimental evaluation," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 564–572, Mar. 2018.

[154] N. Mahmoudi-Kohan, M. P. Moghaddam, M. K. Sheikh-El-Eslami, and E. Shayesteh, "A three-stage strategy for optimal price offering by a retailer based on clustering techniques," *Int. J. Elect. Power Energy Syst.*, vol. 32, no. 10, pp. 1135–1142, 2010.

[155] S. Joseph and J. E. Abdu, "Real-time retail price determination in smart grid from real-time load profiles," *Int. Trans. Elect. Energy Syst.*, vol. 28, no. 3, pp. 1–11, Mar. 2018.

[156] N. Mahmoudi-Kohan, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Elect. Power Syst. Res.*, vol. 80, no. 9, pp. 1042–1048, 2010.

[157] Maigha and M. L. Crow, "Clustering-based methodology for optimal residential time of use design structure," in *Proc. North Amer. Power Symp. (NAPS)*, Pullman, WA, USA, 2014, pp. 1–6.

[158] R. Li, Z. Wang, C. Gu, F. Li, and H. Wu, "A novel time-of-use tariff design based on Gaussian mixture model," *Appl. Energy*, vol. 162, pp. 1530–1536, Jan. 2016.

[159] T. K. Wijaya, M. Vasirani, and K. Aberer, "When bias matters: An economic assessment of demand response baselines for residential customers," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1755–1763, Jul. 2014.

[160] Y. Weng and R. Rajagopal, "Probabilistic baseline estimation via Gaussian process," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Denver, CO, USA, 2015, pp. 1–5.

[161] Y. Zhang, W. Chen, R. Xu, and J. Black, "A cluster-based method for calculating baselines for residential loads," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2368–2377, Sep. 2016.

[162] L. Hatton, P. Charpentier, and E. Matzner-Løber, "Statistical estimation of the residential baseline," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1752–1759, May 2016.

[163] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.

[164] J. Peppanen, S. Grijalva, M. J. Reno, and R. J. Broderick, "Distribution system low-voltage circuit topology estimation using smart metering data," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo.*, Dallas, TX, USA, 2016, pp. 1–5.

[165] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2682–2694, Jul. 2017.

[166] Y. Liao, Y. Weng, and R. Rajagopal, "Urban distribution grid topology reconstruction via Lasso," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Boston, MA, USA, 2016, pp. 1–5.

[167] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying topology of low voltage (LV) distribution networks based on smart meter data," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2680542.

[168] M. Xu, R. Li, and F. Li, "Phase identification with incomplete data," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2016.2619264.

[169] V. C. Gungor *et al.*, "A survey on smart grid potential applications and communication requirements," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 28–42, Feb. 2013.

[170] H. Tram, "Technical and operation considerations in using smart metering for outage management," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo.*, Chicago, IL, USA, 2008, pp. 1–3.

[171] Y. He, N. Jenkins, and J. Wu, "Smart metering for outage management of electric power distribution networks," *Energy Procedia*, vol. 103, pp. 159–164, Dec. 2016.

[172] K. Kuroda, R. Yokoyama, D. Kobayashi, and T. Ichimura, "An approach to outage location prediction utilizing smart metering data," in *Proc. 8th Asia Model. Symp. (AMS)*, Taipei, Taiwan, 2014, pp. 61–66.

[173] Y. Jiang, C.-C. Liu, M. Diedesch, E. Lee, and A. K. Srivastava, "Outage management of distribution systems incorporating information from smart meters," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 4144–4154, Sep. 2016.

[174] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2697440.

[175] J. Zheng, D. W. Gao, and L. Lin, "Smart meters in smart grid: An overview," in *Proc. IEEE Green Technol. Conf.*, Denver, CO, USA, 2013, pp. 57–64.

[176] T. Andrysiak, Ł. Saganowski, and P. Kiedrowski, "Anomaly detection in smart metering infrastructure with the use of time series analysis," *J. Sensors*, vol. 2017, p. 15, 2017, doi: 10.1155/2017/8782131.

[177] M. P. Tcheou *et al.*, "The compression of electric signal waveforms for smart grids: State of the art and future trends," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 291–302, Jan. 2014.

[178] A. Unterweger and D. Engel, "Resumable load data compression in smart grids," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 919–929, Mar. 2015.

[179] A. Unterweger, D. Engel, and M. Ringwelski, "The effect of data granularity on load data compression," in *Proc. DA-CH Conf. Energy Informat.*, Karlsruhe, Germany, 2015, pp. 69–80.

[180] A. Notaristefano, G. Chicco, and F. Piglione, "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *IET Gene. Transm. Distrib.*, vol. 7, no. 2, pp. 108–117, Feb. 2013.

[181] C. Rottondi, G. Verticale, and C. Krauss, "Distributed privacy-preserving aggregation of metering data in smart grids," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1342–1354, Jul. 2013.

[182] L. Sankar, S. R. Rajagopalan, and S. Mohajer, "Smart meter privacy: A theoretical framework," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 837–846, Jun. 2013.

[183] M. Savi, C. Rottondi, and G. Verticale, "Evaluation of the precision-privacy tradeoff of data perturbation for smart metering," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2409–2416, Sep. 2015.

[184] G. Eibl and D. Engel, "Influence of data granularity on smart meter privacy," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 930–939, Mar. 2015.

[185] C. E. Kement, H. Gultekin, B. Tavli, T. Girici, and S. Uludag, "Comparative analysis of load shaping based privacy preservation strategies in smart grid," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3226–3235, Dec. 2017.

[186] T. Hong *et al.*, "Guest editorial big data analytics for grid modernization," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2395–2396, Sep. 2016.

[187] T. Hong, "Big data analytics: Making the smart grid smarter," *IEEE Power Energy Mag.*, vol. 16, no. 3, pp. 12–16, May/Jun. 2018.

[188] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017.

[189] R. J. Hyndman, X. A. Liu, and P. Pinson, "Visualizing big energy data: Solutions for this crucial component of data analysis," *IEEE Power Energy Mag.*, vol. 16, no. 3, pp. 18–25, May/Jun. 2018.

[190] J. Baek, Q. H. Vu, J. K. Liu, X. Huang, and Y. Xiang, "A secure cloud computing based framework for big data information management of smart grid," *IEEE Trans. Cloud Comput.*, vol. 3, no. 2, pp. 233–244, Apr./Jun. 2015.

[191] S. Bera, S. Misra, and J. J. P. C. Rodrigues, "Cloud computing applications for smart grid: A survey," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 5, pp. 1477–1494, May 2015.

[192] S. Mittal, "A survey of techniques for architecting and managing GPU register file," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 1, pp. 16–28, Jan. 2017.

[193] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[194] F. A. Gers, J. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[195] D. L. Marino, K. Amarasinghe, and M. Manic, "Building energy load forecasting using deep neural networks," in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Florence, Italy, 2016, pp. 7046–7051.

[196] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[197] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 17–36.

[198] T. Diethe and M. Girolami, "Online learning with (multiple) kernels: A review," *Neural Comput.*, vol. 25, no. 3, pp. 567–625, 2013.

[199] Y. Xie *et al.*, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 539–553, Apr. 2014.

[200] Q. Zhang *et al.*, "An incremental CFS algorithm for clustering large data in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1193–1201, Jun. 2017.

[201] F. A. Rahimi and A. Ipakchi, "Transactive energy techniques: Closing the gap between wholesale and retail markets," *Electricity J.*, vol. 25, no. 8, pp. 29–35, 2012.

[202] K. Kok and S. Widergren, "A society of devices: Integrating intelligent distributed resources with transactive energy," *IEEE Power Energy Mag.*, vol. 14, no. 3, pp. 34–45, May/Jun. 2016.

[203] B. Celik, R. Roche, D. Bouquain, and A. Miraoui, "Decentralized neighborhood energy management with coordinated smart home energy sharing," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2710358.

[204] N. Liu, X. Yu, C. Wang, and J. Wang, "Energy sharing management for microgrids with PV prosumers: A Stackelberg game approach," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1088–1098, Jun. 2017.

[205] G. Ye, G. Li, D. Wu, X. Chen, and Y. Zhou, "Towards cost minimization with renewable energy sharing in cooperative residential communities," *IEEE Access*, vol. 5, pp. 11688–11699, 2017.

[206] H. Shaker, H. Zareipour, and D. Wood, "Estimating power generation of invisible solar sites using publicly available data," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2456–2465, Sep. 2016.

[207] Y. Wang *et al.*, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3255–3264, May 2018, doi: 10.1109/TPWRS.2017.2762599.

[208] H. Chitsaz, P. Zamani-Dehkordi, H. Zareipour, and P. Parikh, "Electricity price forecasting for operational scheduling of behind-the-meter storage systems," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2717282.

[209] T. Krause, G. Andersson, K. Frohlich, and A. Vaccaro, "Multiple-energy carriers: Modeling of production, delivery, and consumption," *Proc. IEEE*, vol. 99, no. 1, pp. 15–27, Jan. 2011.

[210] Y. Wu *et al.*, "False load attack to smart meters by synchronously switching power circuits," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2018.2806896.

[211] K. Xing, C. Hu, J. Yu, X. Cheng, and F. Zhang, "Mutual privacy preserving *k*-means clustering in social participatory sensing," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2066–2076, Aug. 2017.

[212] L. Wei, A. D. Sarwate, J. Corander, A. Hero, and V. Tarokh, "Analysis of a privacy-preserving PCA algorithm using random matrix theory," in *Proc. IEEE Glob. Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, 2016, pp. 1335–1339.

[213] S. Salehkalaibar, F. Aminifar, and M. Shahidehpour, "Hypothesis testing for privacy of smart meters with side information," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2787838.

[214] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on cyber security for smart grid communications," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 998–1010, 4th Quart., 2012.

[215] Z. Zhang *et al.*, "Achieving privacy-friendly storage and secure statistics for smart meter data on outsourced clouds," *IEEE Trans. Cloud Comput.*, to be published, doi: 10.1109/TCC.2017.2685583.

[216] G. Mateos and G. B. Giannakis, "Load curve data cleansing and imputation via sparsity and low rank," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2347–2355, Dec. 2013.

Yi Wang (S'14) received the B.S. degree from the Department of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2014.

He is currently pursuing the Ph.D. degree with Tsinghua University. He is also a visiting student researcher with the University of Washington, Seattle, WA, USA. His research interests include data analytics in smart grid and multiple energy systems.

Qixin Chen (M'10–SM'15) received the Ph.D. degree from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2010.

He is currently an Associate Professor with Tsinghua University. His research interests include electricity markets, power system economics and optimization, low-carbon electricity, and power generation expansion planning.

Tao Hong received the B.Eng. degree in automation from Tsinghua University, Beijing, China, in 2005 and the Ph.D. degree in operation research and electrical engineering from North Carolina State University, Raleigh, NC, USA, in 2010.

He is the Director of the Big Data Energy Analytics Laboratory, and an Associate Professor of System Engineering and Engineering Management with the University of North Carolina at Charlotte, Charlotte, NC, USA. He is the Founding Chair of the IEEE Working Group on Energy Forecasting and the General Chair of the Global Energy Forecasting Competition.

Chongqing Kang (M'01–SM'08–F'17) received the Ph.D. degree from the Department of Electrical Engineering, 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.