SHORT TERM PROBABILISTIC LOAD FORECASTING AT LOCAL LEVEL IN DISTRIBUTION NETWORKS

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In the Department of Electrical and Computer Engineering

University of Saskatchewan

Saskatoon

By

Bingzhi Wang

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Electrical and Computer Engineering

57 Campus Drive

University of Saskatchewan

Saskatoon, Saskatchewan S7N 5A9 Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9 Canada

ABSTRACT

Along with the growing inclusion of smart technologies into the electrical power grids, benefits, which can be originated form advanced metering infrastructure (AMI), have grabbed noticeable attention from distribution utilities. Since the number of meters are severely ample in practical systems, the utilities now is able to create virtual meter data by aggregating loads for distribution substations, feeders, transformers, or regions with the help of geographic information system. Such an important change brought by smart meter rollout is considered as the main factor which motivates this thesis to delve more into the load pattern modeling and forecasting at local level and find approaches which can yield to the enhanced applications in distribution networks. However, low aggregation level leads to high volatile load characteristic. In this regard, this thesis proposes a comprehensive methodology for uncertainty modeling and short-term probabilistic load forecasting (STPLF) in distribution networks.

Existing methods related to uncertainty modeling and forecasting are rarely applied to local level loads and they suffer from over- or under-fitting of data when there is a misfit between the complexity of the model and the amount of data available. These models are limited to specific situations due to the great diversity of loads in distribution networks and need to be tuned every time when the load aggregation level changes. They also need a relatively large data set to support the recovery of the predictive densities. Our proposed method addresses this issue and is based on Bayesian nonparametric model which has unbounded complexity and allow the complexity to automatically grow and be inferred from the observed data. The uncertainty underlying load patterns can be endowed with any type of prior distribution and is given in a nonparametric form, i.e. a mixture model with countably infinite number of mixtures, inferred from the posterior using the

Gibbs Sampling, which is a Markov Chain Monte Carlo (MCMC) technique. All effective samples from the sampling procedure along with the exogenous variables are fed to an ensemble learning machine. The final result of the probabilistic load forecasting (PLF) is averaged on the outputs of all learning models, thus reducing the model variance and enhancing the model consistency. The proposed method is tested on both a public data set and a local data set from the Saskatoon Light &Power AMI Meter Replacement Program which offers electricity consumption at a granularity of 30 minutes of more than 65,000 electricity customers including industrial, commercial and residential sectors in the city of Saskatoon, Canada.

ACKNOWLEDGMENT

I would first like to express my sincere gratitude to my supervisor Prof. C.Y. Chung, for his patience and continuous support during my M.Sc. study. His immense knowledge and guidance greatly facilitate my research. Besides academic study, he also encourages me to actively join industry activities, for instance, taking responsibility for an engage grant project with Saskatoon Light & Power, and joining the Engineer in Training program in SaskPower. These valuable experiences motivate me to have a deeper insight into my research and make my work more practical and beneficial to the power industry.

Besides my supervisor, I would like to thank the rest of my graduation committee, Prof. Chris Zhang, Prof. Xiaodong Liang and Prof. Khan Wahid, for their insightful review comments and suggestions, and also their questions which provide great help in improving my thesis.

I also would like to thank my lab mates, for accompanying me through those very long very long days. We share happiness together, take care of each other and help each other. Without their support I cannot get through all the difficulties from both life and researches.

Last but not the least, I would like to express my sincere thanks to my family: my mother, brother and aunt parents, for their spiritual support throughout my life overbroad. They are the headspring of my self-belief and sense of being loved.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMI	advanced metering infrastructure
AMR	automatic meter reading
ANN	artificial neural network
ARIMA	autoregressive moving average
BNP	Bayesian nonparametric
CDF	cumulative density function
CRP	Chinese Restaurant Process
CRPS	continuous ranked probability score
DP	Dirichlet Process
DPMM	Dirichlet Process mixture model
DT	decision tree
GIS	geographic information system
KDE	kernel density estimation
LCL	Low Carbon London
LTLF	long-term load forecasting
MAE	mean absolute error
MAPE	mean absolute percentage error
MCMC	Markov Chain Monte Carlo
MTLF	mid-term load forecasting
MW	Megawatt

NN	neural network
PF	probabilistic forecast
PDF	probability density function
PI	predictive interval
PLF	probabilistic load forecasting
QR	quantile regression
QRA	quantile regression average
SLP	Saskatoon Light & Power
STLF	short-term load forecasting
STPLF	short-term probabilistic load forecasting
SVM	support vector machine

1. Introduction

1.1. Electric Load Forecasting

One of the main requirements of a secure and reliable power system is that the electricity supplied must always meet the load plus the system loss within the allowable range, due to the lack of storability of electricity. Figure 1.1 illustrates the energy balance in a power system. To make sure that every customer has access to usable energy, over decades electric load forecasting has been playing an essential role in scheduling and dispatching resources in power systems. Load forecasting is basically defined as the prediction of future load for a certain period ahead on a given system. It provides information on when, where and how much energy is demanded and assists utilities to decide operation actions such as adjusting output of generators or interchanging power with neighboring systems, and planning actions such as installing extra generation to meet the increasing demand.

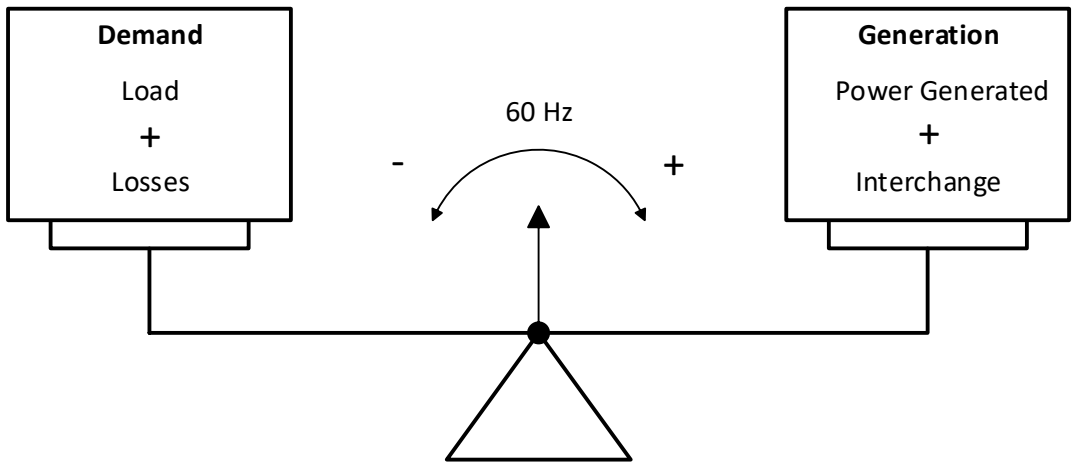


Figure 1.1: Illustration of power system balancing

Accompanying deregulation and privatization into power markets, accurate load forecasting has grabbed increasing noticeable attention on a wide of sections including transmission and distribution

planning, secure and optimal operation, and system investments such as energy cost economy [1]. We can never predict with one hundred percent accuracy due to the randomness of the load and various exogenous factors such as weather condition, holiday events, calendar factors, seasonal factors, etc. Inaccurate load forecasting, both positive and negative errors, will cause an increase in cost. For instance, overestimating the load will require extra generation, which is always unwelcome to the system and increases operating cost by increasing output or committing more units. Underestimating the load will result in failure of providing necessary reserve and stability to the system, which may cause system breakdown [2]. Moreover, it fails to satisfy the demand, the impact of which on end users is even more complicated to evaluate, and buying at the last minute from the power market is super expensive. [3] gives an example that even only 1% increase in the national load forecasting error cost around £10 million a year at 1984 in U.K., due to inefficient plant scheduling. [4] also points out that conservatively speaking, a 1% decrease in load forecasting error for a system with 10,000 MW capacity can save around $1.6 million per year. Therefore, accurate load forecasting is at the core of operating and planning a reliance and secure power system, in an economic way.

1.2. Categories of Electric Load Forecasting

The efforts to categorize load forecasting vary in the literature and they mainly differs in the way of time horizon division. A widely accepted categorization is proposed by [5] in terms of time horizon of varying duration and is known as short-term load forecasting (STLF), medium-term load forecasting (MTLF) and long-term load forecasting (LTLF), which are detailed as follows. Figure 1.2 summaries the categorization.

- **Short-term load forecasting**

STLF concerns forecasts over time intervals from less than one hour up to a day ahead. Accurate STLF is crucial to efficient operations. Utilities use STLF for optimal operating of generation allocation and scheduling, unit commitment, and interchange evaluation. STLF ensures short-term availability and reliability of power supply and its forecast step is usually one hour or half an hour.

- **Medium-term load forecasting**

MTLF ranges from one day to a year ahead. It is mainly used for operational planning and suits a wide range of applications such as maintenance planning, fuel scheduling and switching operation, medium-term hydro thermal coordination, etc. The main concern of MTLF is to assess the adequacy of plant margins during maintenance and breakdown. The forecast step for MTLF is flexible and it often provides prediction of daily peak load, daily total load and monthly total load.

- **Long-term load forecasting**

LTLF covers a long-term time horizon which lies from one year to several years into the future. It is used to check the resource adequacy of a certain region and evaluate both the reliability and economic performance of the growing power system. Based on this, utilities set plans for transmission and distribution networks, as well as generation facilities. LTLF usually predicts annual peak load.

This thesis focuses on the scale of STLF.

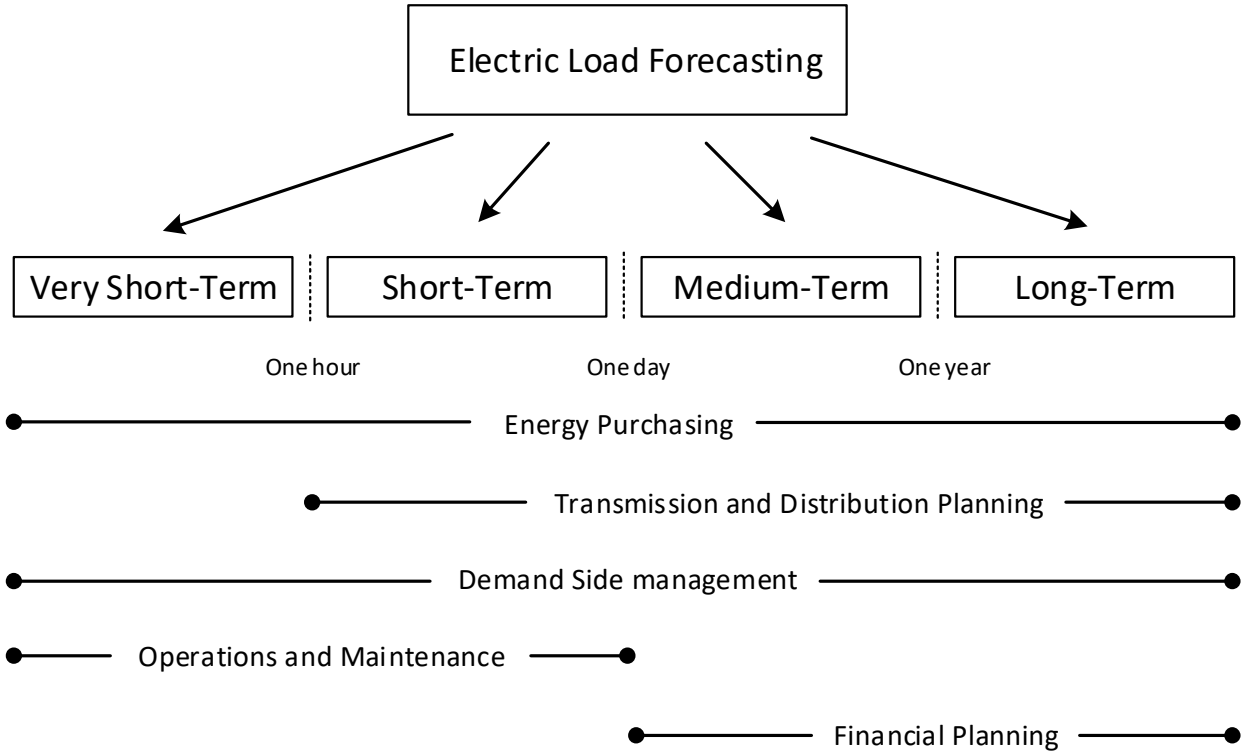


Figure 1.2: Categorization of load forecasting by time horizon

1.3. Smart Meter Rollout and Advanced Metering Infrastructure

In a global trend, climate change, awareness of energy efficiency, more competitive power market, and customer participation are driving the conventional power system to move towards a smarter grid with intelligence integrated for better reliability, efficiency and cost effectiveness. This development raises the need to monitor, control and optimize usage and thus puts forward the requirement to construct a system based on information and communication technologies. As a response, the maturing smart metering technology has addressed this issue and is playing an indispensable role in bridging connection between different elements in modern power systems. It plays as the base infrastructure and evolves coordinately with a developing smart grid.

Smart metering system evolve from the automatic meter reading (AMR) technology which consists of an electronic meter and a communication module [6] and is used only for meter reading and billing as it only allows one-way communication. With more functionality added and two-way communication technology enabled, smart meters along with the corresponding communication system and data management system, known as advanced metering infrastructure (AMI), has grabbed noticeable attention from distributed utilities. Table 1.1 lists the comparison of basic functionalities between AMR and AMI. In early 2013 for the first time, AMI meters outnumbered AMR meters in U.S. which was once the leading market of smart meters before that time, reported by the U.S. Energy Information Administration [7]. Figure 1.3 illustrates the trend. Since 2013, China has taken over the top spot that once belonged to the North American market with 406.9 million smart meter installations by 2018 and the trend will continue due to the policy support for investment in smart meters. During the same period, the U.S. and Japan installed 38.7 million and 36.5 million units respectively, ranking second and third in the world considering the number of installations according to the report from GlobalData. Figure 1.4 shows the projection of smart meter installations by 2020. As predicted by Navigant Research, the global penetration of smart meters is expected to hit about 53% by 2025, with U.S., Japan and South Korea targeting 100% penetration.

Table 1.1: Comparison of basic functionalities between AMR and AMI

AMR	AMI
One-way communication Data collection only (monthly or at most daily) No means for broadcasting command or control messages	Metering measurements hourly or more frequently All information in real time and on demand Remote meter programming Home area network interface Outage/theft detection Power quality

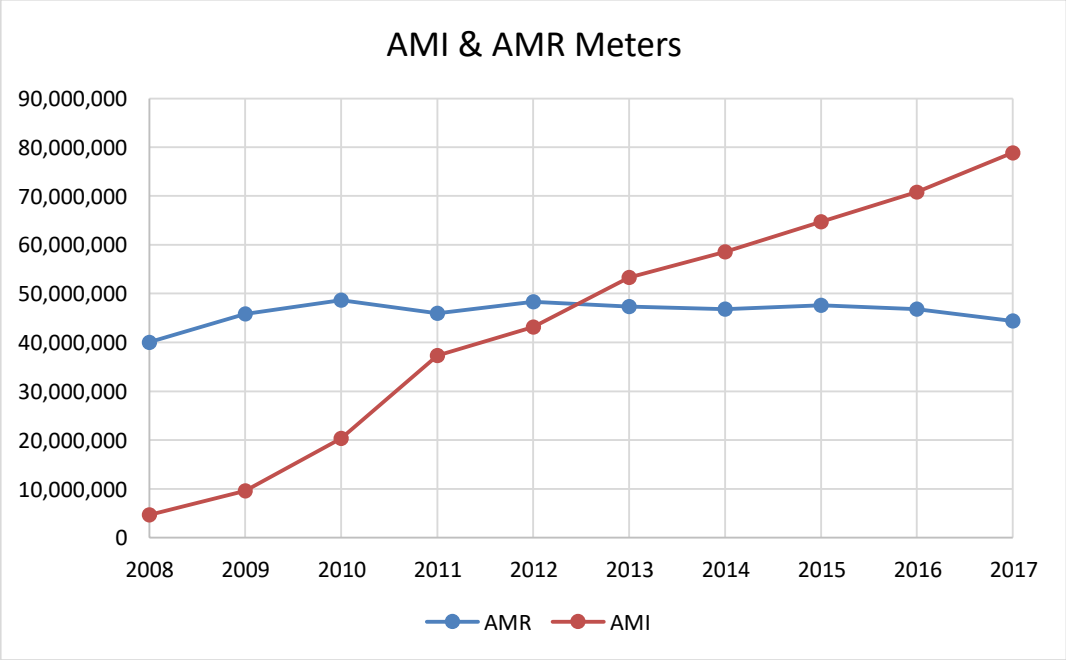


Figure 1.3: Numbers of AMI & AMR meters in U.S.

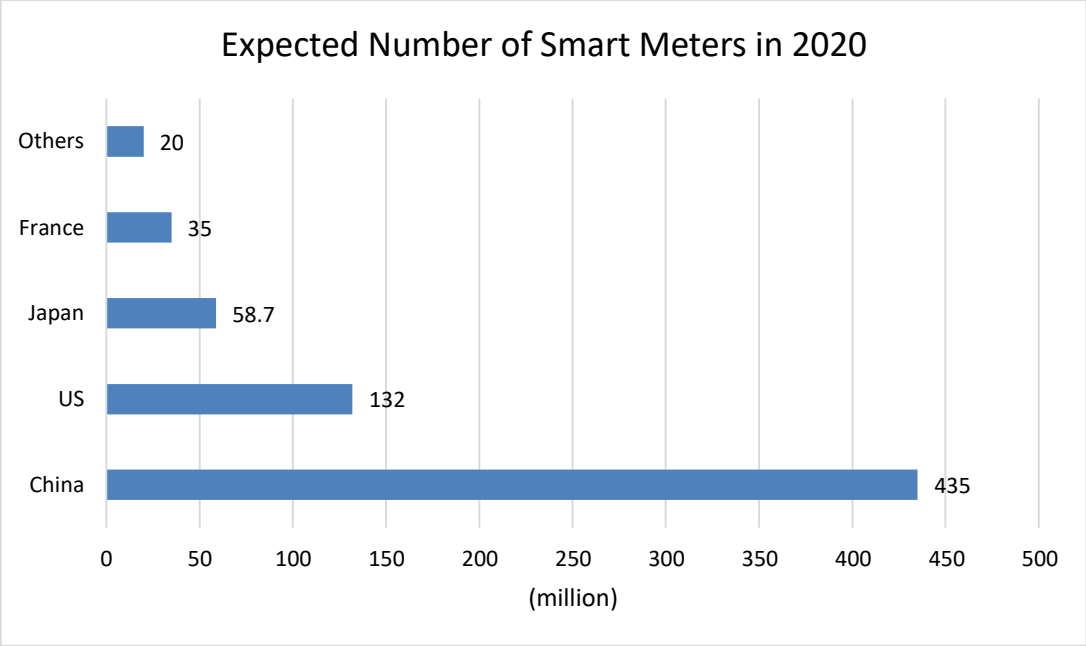


Figure 1.4: Projection of smart meter installations in 2020

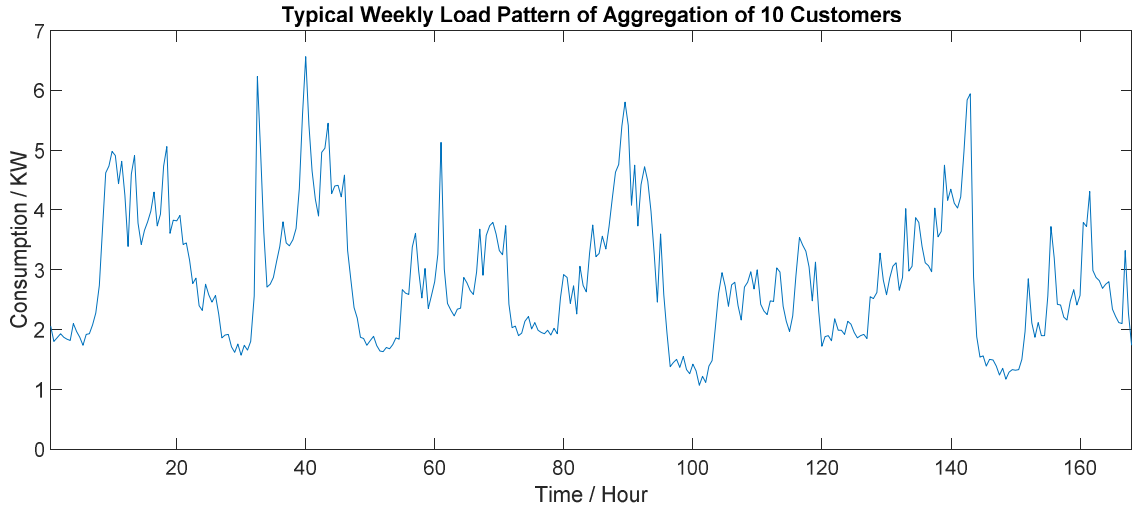
Not only the number of AMI meters explodes, the capability of the meters also develops rapidly to meet the need of various stakeholders and adapt to the fast development of the smart grid. An AMI meter basically has the following capabilities [6][8]:

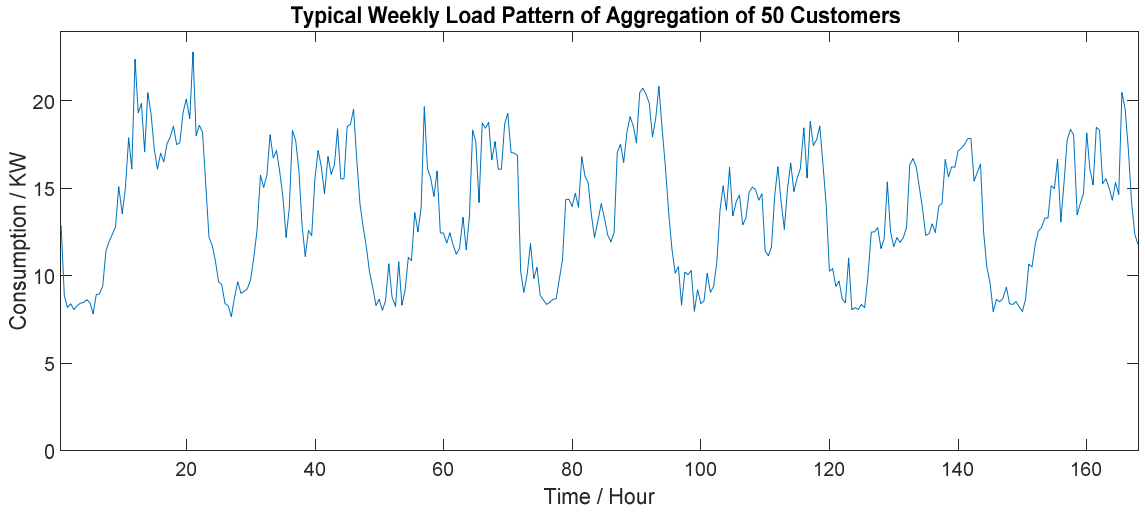
- real-time or near real time recording of usage
- remote on-demand reading, service switching and meter programing
- linking to gas and water supply
- ability to capture events and power quality data

Despite the direct benefits from the functionalities of smart meters, utilities now show unprecedented interest in meter data analytics. Data analytics is defined as the process of analyzing raw data to draw conclusions about that information, such as hidden patterns, unknown correlations, etc. Smart meters generate data in a timely manner and transfer the data to meter data management system through the information infrastructure. This high volume and fine-grained time interval data (15 min to 1h) opens up new opportunities to enhance applications especially in the side of

distribution grid. Also thanks to the increasing penetration of smart meters, utilities now is able to aggregate the data for distribution substations, feeders, transformers, or regions with the help of geographic information system and by doing so, they can conduct analytics at any aggregation level, from the whole system down to a city block. For instance, the meters connected to the same transformer can be aggregated to generate its load pattern and identify overloading conditions. Aggregated load of different components inside large facilities can also help to optimize their building operations and enable energy efficiency strategies. Utilities also show interests in the aggregated load profile of a typical neighborhood, which is important to know when constructing the power grid for a new living area. Understanding the underlying load pattern of a typical neighborhood can help our engineers optimally design the new network, such as the optimal size of transformers, the reasonable capacity of transmission lines, etc.

Such a significant change brought by smart meter rollout is considered as the main factor which inspires researchers working in the area of load forecasting to reexamine existing methods and develop new forecasting tools, in order to cope with the challenge that the available data is more granular and new applications require forecasts at smaller scales in distribution networks, such as the aggregation of a city block, or an individual distributed transformer. Unlike system level load which shows much more regularity, low aggregation level leads to high volatility, making load forecasting at such level more challenging. Figure 1.5 plots three typical weekly load profiles at different aggregation levels in a distribution network.


Typical Weekly Load Pattern of Aggregation of 10 Customers
Consumption / KW
Time / Hour
0
1
2
3
4
5
6
7
20
40
60
80
100
120
140
160


Typical Weekly Load Pattern of Aggregation of 50 Customers
Consumption / KW
Time / Hour
0
5
10
15
20
20
40
60
80
100
120
140
160

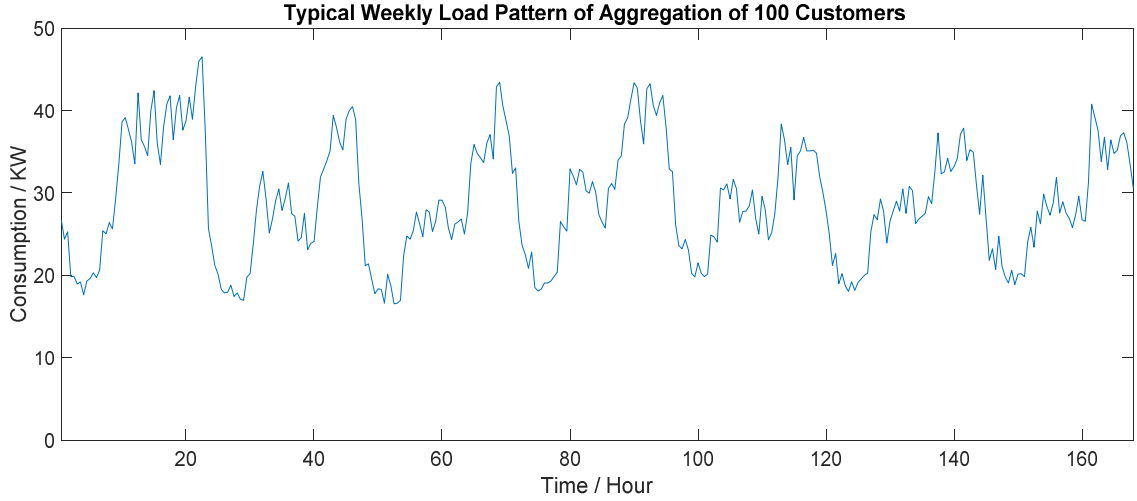

Typical Weekly Load Pattern of Aggregation of 100 Customers
Consumption / KW
Time / Hour
0
10
20
30
40
50
20
40
60
80
100
120
140
160

Figure 1.5: Three typical weekly load profiles at different aggregation levels in a distribution network

1.4. Problem Statement and Research Objectives

The remarkable evolution of smart meters in recent years has significantly increased and improved the monitoring of electricity distribution networks through AMI. The availability of real-time fine-grained energy usage information provides better visibility of the network and is of great interest to distribution network operators for better operation and planning. A wide range of applications in distribution network rely on an accurate short-term load forecast at local level, including optimal operation planning [9], energy storage optimization [10], demand side management [11], distributed energy resources integration [12], energy management systems [13], [14], microgrid applications [15], [16], distribution system state estimation [17], etc.

Traditionally, STLF which predicts the electrical demand over a period from several hours up to one week, has achieved very reliable forecasting performance at the system level (high voltage grid, balancing groups or zones, etc.), load profile of which shows much more regularity and less volatility. There exists a rich literature on STLF at the system domain, developing and improving a wide variety of modeling and forecasting techniques from time series models such as Box-Jenkins models [18] and autoregressive integrated moving average (ARIMA) models [19] to artificial intelligence based techniques such as neural network (NN) [20], support vector machine (SVM) [21], fuzzy systems [22] and hybrid methods [23], etc. A systematic review can be found in [24]. In contrast, STLF techniques at local level were poorly developed in the past due to the lack of available measurements until recently, the worldwide smart meter roll-out has addressed the issue and shifted

the interests from the high voltage domain to low voltage domain to meet the increasing need in load forecasting at local level in distribution networks. Wide-area introduction of smart meters enables utilities to create virtual meters (aggregates of smart meters) for load of distribution substations, feeders, transformers, or regions with the help of geographic information system (GIS). With this advantage, it is possible now that analytics can be conduct at any aggregation level in a distribution network and novel opportunities arise for a wide variety of applications in distribution networks [9]–[17].

However, low aggregation level leads to high volatile load characteristic, making STLF even more challenging. Several works examined the effect of load aggregation on STLF performance [25]–[28]. The authors evaluate state-of-the-art STLF methods for different prediction horizons and data types and demonstrate a strong scaling law relationship between the forecasting errors and the aggregation size. For instance, [27] illustrates that the empirical coefficient of variation varies significantly when the aggregation level is under 10^3. [28] shows that the mean absolute percentage error (MAPE) of STLF based on NNs varies from 7.9% up to 47.6% as the aggregation decreases from 10^3 to 10^1, which covers most of the situations at the local level in a distribution network. These two metrics respectively represent the volatility and variability of load. To handle and consider the corresponding high uncertainty, most recently there has been a significantly increased interest in probabilistic load forecasting (PLF) [29], especially in the scope of short-term probabilistic load forecasting (STPLF). Researchers begin to delve more into the load pattern uncertainty modeling and forecasting at local level and find approaches which can yield to the enhanced applications in distribution networks, such as probabilistic load flow [30], unit commitment [31], reliability planning [32], etc.

The basic objective of this research work is to propose a new method focusing on uncertainty

modeling and STPLF at the local level in distribution network. . Loads at such level show much less regularity than system level aggregation load and are very volatile. Existing methods related to uncertainty modeling and probabilistic forecasting are rarely applied to local level loads and they suffer from over- or under-fitting of data when there is a misfit between the complexity of the model and the amount of data available. They also need a relatively large data set to support the recovery of the predictive densities. Our proposed method addresses this issue and is based on Bayesian nonparametric (BNP) model which has unbounded complexity and allow the complexity to automatically grow and be inferred from the observed data.

1.5. Structure of the Thesis

This thesis consists of five chapters which is briefly described as follows:

Chapter 1 introduces basic concepts of electric load forecasting and its categorization. Opportunities and challenges brought by smart meter roll-out are explained. The problem that we are encountering and trying to solve is described. The research objective is lastly presented in this chapter.

Chapter 2 further discusses PLF. General background and development are presented. Formal definition and preliminary concepts are introduced. A comprehensive review of the state-of-art techniques is also given in this chapter.

Chapter 3 presents the proposed method for STPLF at local level in distribution network. Variables that are selected for forecasting are discussed first. Afterward, the proposed methodology is introduced in two steps, modeling uncertainties underlying load patterns and forecasting probability distribution conditional on exogenous factors. The techniques that are used in the

modeling and forecasting process are introduced. The whole structure of the proposed methodology is summarized at the end of this chapter.

Chapter 4 demonstrates the effectiveness of the proposed method by comparing with the most popular benchmarks using smart meter data. A comprehensive evaluation criteria for PLFs is introduced. Two data sets are tested including a public data set and a local data set. The case studies focus on intro-day PLFs at local level. Various ranges of aggregation levels are verified to enhance the conclusion. The impact of the randomness of stochastic algorithms on the PLFs, which is rarely payed attention to in the literature, is examined in this chapter.

Chapter 5 summarizes the thesis and suggests future research directions.

2. Probabilistic Load Forecasting

2.1. Introduction

Point load forecasting, which gives an expected value at each forecasted time step, has been widely implemented since the early time of power system. With the power industry going through a significant transition process, point load forecasting is becoming unreliable as diversified smart technologies are introduced to the grid, such as distributed energy resources, distributed energy storage systems, plugin electric vehicles, and demand response programs, etc. As a consequence, these technologies add extra uncertainty and complexity to the system, making point load forecasting unreliable. To tackle this problem, researches on PLF have grabbed increasing attention in recent years.

Among the limited literature on PLF in the scope of technical and methodological development, quantile regression (QR) and kernel density estimation (KDE) [33]–[36] are two most widely used methods to directly generate probabilistic forecasts (PFs). PFs can also be indirectly generated from point forecasts, for instance, by modeling and simulating the residuals of the underlying point forecast [33], or by feeding temperature scenarios to point forecasting models [37]. To manifest uncertainty, these methods provide PLFs in the form of confidential intervals, quantiles or the whole probability density function (PDF), which provides more information of the predicted load than point forecasts thus enhancing the decision-making process in operation and planning of the system.

The rest of this chapter goes through the literature and briefly introduces the most used methods of PLFs, including QR, KDE, residual simulation, scenario generation and other techniques.

2.2. Review of State-of-Art Techniques

The literature on PLF is quite limited. In the following subsections, a brief description of the most used techniques and their methodological development are reviewed.

2.2.1. Quantile Regression

As a classical statistical method, linear regression [38] estimates the conditional mean function that indicates the relation between predictor variables and a response variable by minimizing a loss function, e.g., the summation of the squared residuals. As an alternative, median regression estimates the conditional median function via least absolute distance estimation. These two methods both give an evaluation of the relation between the predictors and the response, while median regression better interprets the relation when the distribution is highly skewed.

Quantiles are defined as points dividing a sample into equal-sized groups, e.g. the median is the 0.5^{th} quantile showing the central location of the entire sample. Formally, the qth quantile denotes the value below which the proportion of the data points to the entire population is q. A quantile is a continuous value between the range [0,1], so any position of a distribution can be calculated given the sample and a predefined quantile. Figure 2.1 illustrates a standard normal distribution with the 4-quantiles (also known as quartiles) shown. The 4-quantiles consists of three quantiles, which divide the distribution into four sections with the same area, i.e., the area below the PDF curve is the same in intervals $(-\infty, q_1)$, (q_1, q_2), (q_2, q_3), and $(q_3, +\infty)$.

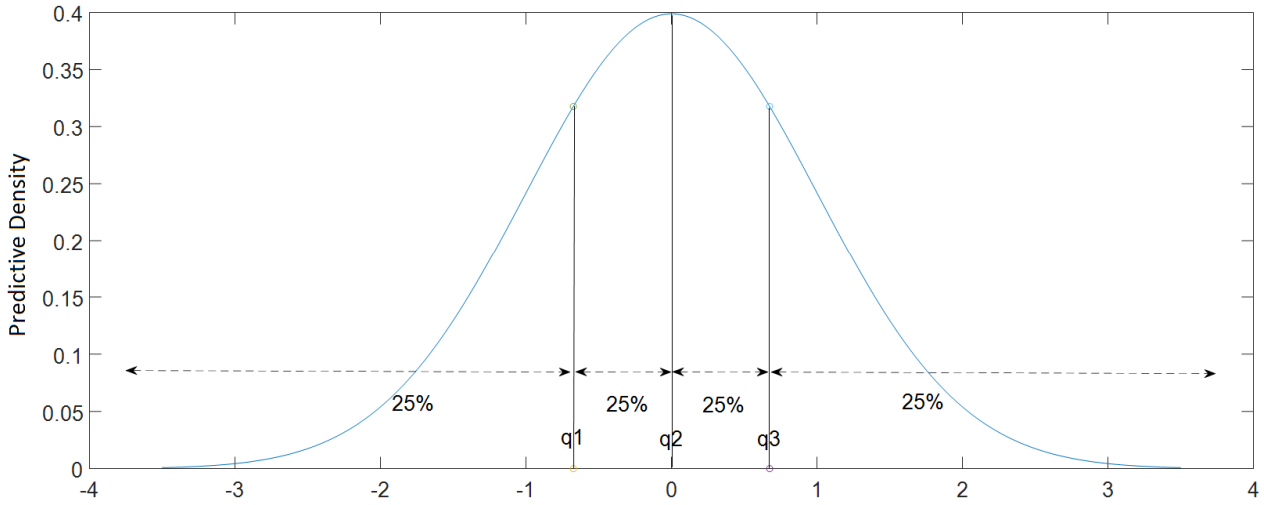


Figure 2.1: Quartiles of standard normal distribution

First introduced in 1978, QR [39] models conditional quantiles as functions of predictors, providing a more comprehensive view of the impact of the predictors on the response variable. Suppose that the load and predictors have a multiple linear relation. A typical form proposed by [40] is shown below:

$$Load = \boldsymbol{\beta} \boldsymbol{X}_t$$

$$\boldsymbol{X}_t = (1, Trend, Hour, Weekday, Month, T, T^2, T^3, Hour * Weekday, \quad (2.1)$$
$$T * Hour, T^2 * Hour, T^3 * Hour, T * Month, T^2 * Month, T^3 * Month)$$

where $Trend$ is a natural number indicating a linear trend; $Hour$ is a categorical variable denoting the 24 hours of a day; $Weekday$ is a categorical variable denoting the 7 days of a week; $Month$ is also a categorical variable denoting the 12 months of a year; and T represents the temperature. In the case of PLF, the linear quantile regression problem can be expressed as

$$f(q|\boldsymbol{X}_t) = \boldsymbol{\beta}_q \boldsymbol{X}_t \quad (2.2)$$

where $f(q|\boldsymbol{X}_t)$ is the conditional qth quantile of the load distribution y_t at time t, $\boldsymbol{X}_t$ are the predictors (regressors) at time t, and $\boldsymbol{\beta}_q$ is the parameter vector. For a given quantile q, the parameter

vector $\boldsymbol{\beta_q}$ is estimated by

$$\boldsymbol{\beta_q} = \arg\min_{\boldsymbol{\beta_q}} \sum_t l_q(y_t - \boldsymbol{X}_t\boldsymbol{\beta_q}) \tag{2.3}$$

where $l_q(\bullet)$ is the loss function given by

$$l_q(s) = |s(q - \mathbb{1}(s))| \tag{2.4}$$

where $\mathbb{1}$ is the Heaviside step function whose value is 1 if the real argument is non-negative otherwise the value is 0.

Classical linear QR considers a large set of covariates, increasing both computation burden and model complexity. Several methods have been proposed to address the problem. [41] treats the quantiles of the load as a linear combination of several principle components, using functional principle component analysis [42] to reduce the dimensions. QR can also be combined with machine learning methods. A quantile gradient boosting method is proposed by [33], which has more flexibility than linear models and incorporates an embedded variable selection process. QR can also be combined with tree-based models, such as QR forest [43], since tree-based models have a natural ability to deal with both numerical and categorical inputs and its recursive portioning algorithm is quite efficient and has well-accepted performance when dealing with high-dimensional inputs. QR is among the most popular models in recent literature related to probabilistic forecasting, not only PLF but also wind forecasting, solar power forecasting, energy price forecasting, etc. QR gives a non-parametric way to estimate conditional quantiles and recover the entire conditional distribution by computing a large set of quantiles.

2.2.2. Kernel Density Estimation

KDE belongs to the branch of non-parametric statistics and gains popularity due to its ability of estimating the PDF of a continuous variable without relying on any particular parametric family of probability distributions, such as Gaussian distribution. As a non-parametric estimator, KDE does not have fixed structure or functional form, and automatically learn the density shape depending on all the sample points.

Denote $X_1, \dots, X_n \in \mathbb{R}$ as random samples independently and identically drawn from an unknown distribution with density f. The estimated PDF of f via KDE can be expressed as

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.5}$$

where K is a smooth function called the kernel function and h is the smoothing bandwidth which is a positive real number. More intuitively, consider that the density at the point x is estimated by taking the local density of the sample points within the distance h, namely, the size of the neighborhood around x. If equal weight is assigned to each point within the neighborhood, the estimated PDF will be bumpy as a histogram. To fix this, a kernel function K is introduced to control the weight decreasing towards zero continuously as the distance between X_i and x increases. In this way, a large value of the PDF is estimated for neighborhoods with many observations, while a small value is given for neighborhoods with only a few observations. Figure 2.2 plots the effect of different kernel function K on the shape of the estimated density curve for the same data. It can be seen that these density estimates varies slightly but overall comparable. Only the curve using box kernel function lacks smoothness. The smoothness of the density curve can be controlled by the choice of bandwidth value h. Figure 2.3 illustrates an example by using a Gaussian kernel smoothing function with three different bandwidth values. The default bandwidth is the optimal value for estimating normal distribution. It can be clearly seen that the choice of the bandwidth value has a big impact

on the shape of the estimated density curve.

Many literatures use KDE as the final step to estimate the PDF given a set of conditional samples, where KDE is only employed as an auxiliary unconditional estimator. However, the literature is quite limited regarding the applications of conditional KDE. To the most of the author's knowledge, the newest and most relative paper using conditional KDE for PLF is [36], which comprehensively compares the forecasting accuracy of KDE methods conditional on different impacting factors (period of week, period of day, lagged consumption, etc.).

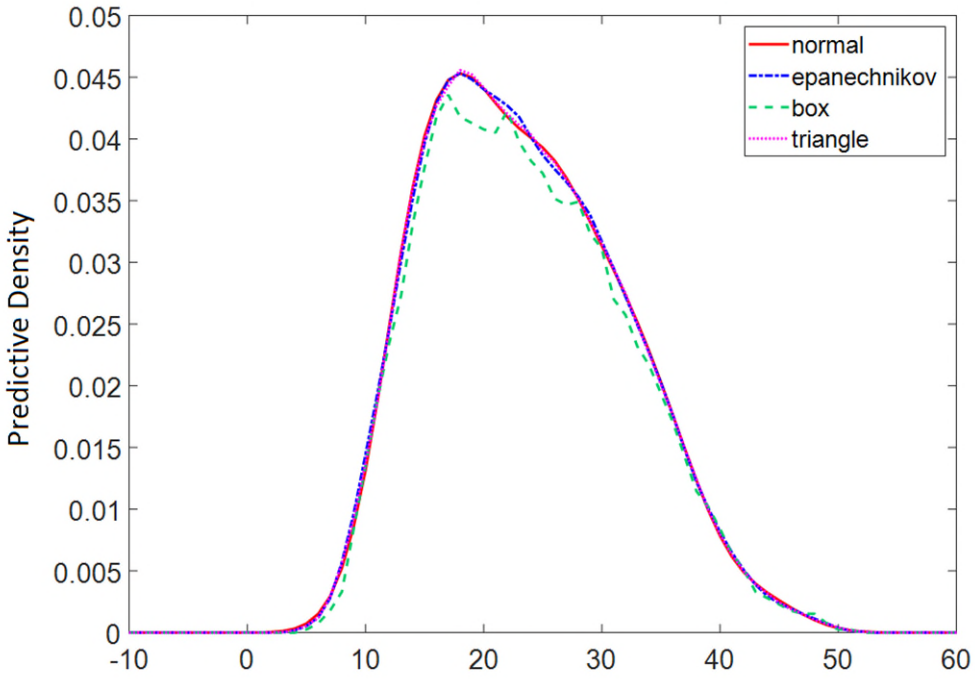


Figure 2.2: Density plot with different kernel functions

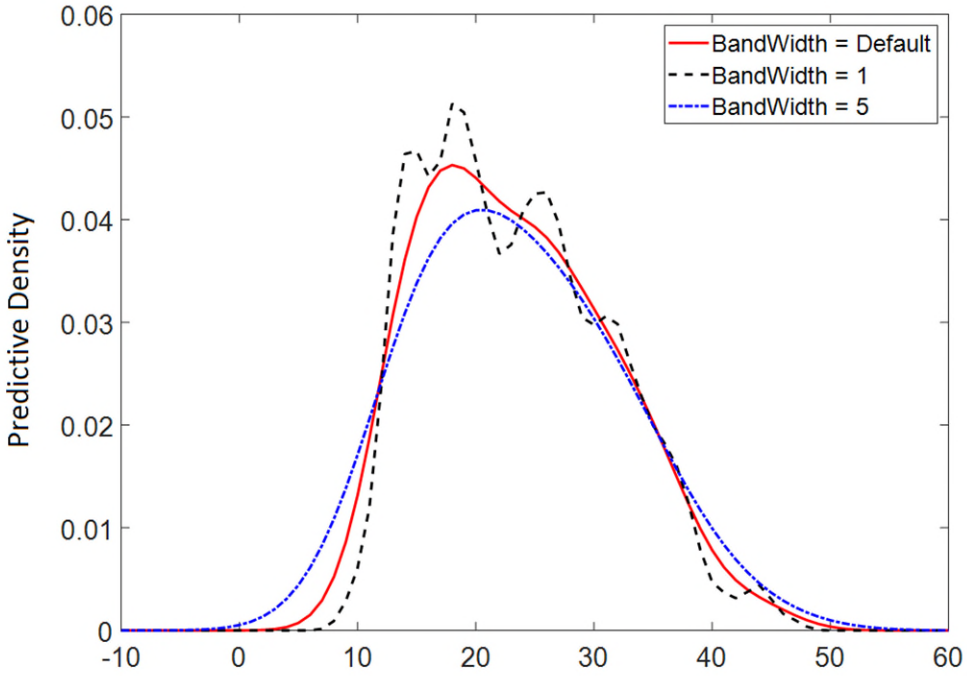


Figure 2.3: Density plot with different values of bandwidth

2.2.3. Residual Simulation

In the area of time series forecasting, the differences between the forecasts and the actuals are called the residuals. Ideally the residuals should be a random noise which shows a bell shape and its peak centered at zero in the residual distribution plot, indicating that the forecasting method is a good fit and unbiased with no discernible pattern unmodeled. However practically, the residual errors can have trends, bias or even seasonality that the model fails to capture. Therefore, the analysis of residuals is widely used in regression problem to help validate the model. The temporal structure of residuals can also be directly modeled and predicted to help lift the performance of the model.

In the literature of load forecasting, most papers make normality assumptions for the residual distribution to generate PLFs, as discussed in [29]. A comprehensive examination on the normality assumption for load forecasting residuals has been conducted by [45]. The result shows that none of

the residual series passed the K-S test based on certain significance level and critical value, indicating that the normality assumption is not sufficient. Another important conclusion of this paper is that adding residuals simulated from normal distribution can improve the performance of deficient models but the improvement diminishes for models with more predictive power.

In fact, residuals of load forecasting do not necessarily follow any well-defined parametric distribution due to the complex relation between consumption and exogenous factors. To avoid relying on unverified distribution of forecasting residuals, researchers seek help from non-parametric methods such as QR to model the residuals, which are further integrated with the point forecast to generate PLFs. The newest relevant work can be found in [46], which applies a non-parametric QR model to model the probabilistic residual forecast and the point forecasts are used as an input along with other related exogenous factors.

2.2.4. Scenario Generation

Compared with other approaches to generate PLFs, scenario generation is more commonly accepted and widely used by industry because of its simplicity and interpretability. The basic idea of this method is that multiple scenarios are first generated by simulating the input variables, and afterwards these scenarios are fed to a point forecasting model to obtain a series of forecasts which are used to estimate the final PLFs. Figure 2.4 gives a schematic view of this idea. Since calendar variables are fixed and weather is the major factor that drives the load demand, most of the literature on this topic focus on simulating temperature to generate input scenarios. Different methods are reported and can be basically classified into four typical categories, namely fixed-date, shifted-date, bootstrap and surrogate methods, as briefly described as below. A more detailed introduction about

the four methods can be found in [37].

1) Fixed-date method: this method creates a temperature scenario of a future period by assigning the temperature profile of a past period date by date.
2) Shifted-date method: this method generates a new scenario by shifting the temperature profile of a past period forward or backward by one or more days.
3) Bootstrap method: this method divides the temperature profile of each past year into equal length samples, which are then repeatedly drawn with replacement to obtain a new temperature profile.
4) Surrogate method: this method generates new temperature profiles by taking the Fourier transform of the past temperature series. The advantage of this method is that it can keep the distribution and autocorrelation of the original temperature series.

This paper also formally compares and quantitatively evaluates these methods based on some criteria and proposes a practical guideline for model selection when using these temperature scenario generation methods.

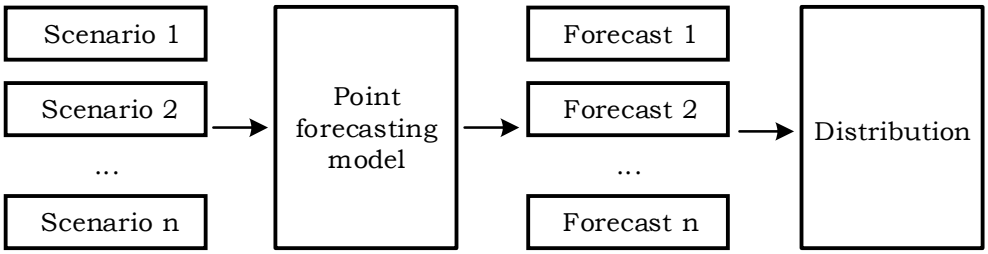


Figure 2.4: Schematic view of scenario generation method

2.2.5. Other Techniques

Generally, the methods of generating PLFs can be classified into two main types. One strand provides PLFs directly, such as quantile regression and KDE. Another main strand attempts to

extend point forecasts to obtain PLFs, such as residual simulation and scenario generation. Based on these methods, researchers also have developed a few other techniques by making various hybrid models, combining multiple forecasts or applying new methodologies. [47] includes residual simulation as a post-processing component to improve the original PLFs generated by scenario generation. Instead of applying QR on point forecasts from one single model, [34] applies the quantile regression averaging (QRA) technique to a set of sister point forecasts, which are predictions generated from the same family of models. These models are obtained by varying input features through a variable selection process. By averaging on different forecasts from a set of models, QRA greatly reduces the risk of making poor decisions over model selection, thus achieving significantly enhanced quantile forecasts. Some researchers seek help from stochastic processes since such models are well-established to address problems associated with uncertainties. [48] incorporates the Gaussian process into QR to construct a non-linear quantile regression model and has reached satisfactory forecasting performance, however with the sacrifice of computation efficiency. To address this issue, [49] proposes a heteroscedastic Gaussian process model using $l_{1/2}$ regularization which gives more sparse solutions, thus significantly reducing the computational complexity. Researchers have been consistently taking efforts to enrich the very limited PLF literature. They keep improving the current methods and trying to open new directions. It is believed that load forecasting is a hot topic for the time being and will still be in the future.

2.3. Summary

In this chapter a brief review of the state-of-art PLF techniques is provided. More efforts are given to the most widely used methods, i.e., QR, KDE, residual simulation and scenario generation. Other techniques are also introduced with a few highlighted.

3. Proposed Method for Short-Term Probabilistic Load Forecasting

3.1. Introduction

The majority of the literature on PLF still focus on the system level forecasting and comparatively, researches at local level are sparse and limited to household/single smart meter level or a relatively high aggregation level. To the best of the authors' knowledge, the only paper that falls in the scope of PLF at local level to the most recent, is [24] who considers using QR to generate PLFs and analyses the effect of aggregation on predictive intervals (PIs) at a lower level, aggregating from 1 up to 240 customers in steps of 30.

Though the methods mentioned in the previous chapter can possibly be applied to local level loads, they all suffer from their own limitations and may not be the best fit. QR gives PLFs in the form of PIs by computing specific quantiles. The predictive distribution can only be recovered provided a large set of quantiles are computed [33]. There are also a few issues associated with the kernel density approach, i.e. the kernel shape, the kernel bandwidth and possibly a larger sample size than what you might need for a parametric method [50]. Residual simulation methods heavily rely on unverified distributions of point forecast residuals such as normal distribution, as reported in [45]. Among the weather scenario generation methods, though they are practical and widely used in the industry, their methodological foundation is not yet solid, resulting in ad-hoc, judgmental and indefensible choices during the scenario generation procedure [37].

In this regard, this thesis adds a new methodology exclusive from the above methods to the PLF literature, focusing on day-ahead PLF at local level in distribution network. Loads at such level show much less regularity than system level aggregation load and are very volatile, but are still more

predictable than household/single smart meter level load. The proposed methodology solves the STPLF problem in two steps, modeling uncertainties underlying load patterns and forecasting probability distribution conditional on exogenous factors. A more detailed description of the proposed methodology is provided in the following sections.

3.2. Uncertainty Modeling

The first step of the proposed method is presented in this section. The basic idea of modeling uncertainty is to apply Bayesian inference on mixture models. The Dirichlet process mixture model (DPMM) is proposed in this research. The DPMM is represented by the Chinese restaurant process (CRP) and then inferred by Gibbs sampling, which is an MCMC algorithm. The detailed methodology is introduced in the following subsections.

3.2.1. Time Series Uncertainty Representation - Mixture Models

It is widely implemented that simplifying assumptions are made when doing data analysis. It is usually assumed that the observations are drawn from one specific distribution. However, though such assumption may work in many cases, it is too restrictive and not sufficient when the data to be modeled are more complex than normal cases, such as multimodal data. To address this issue, mixture models, which are a discrete or continuous weighted combination of distributions, are introduced.

In mixture models, it is assumed that data are generated from different sources, each of which is called a component with a unique pattern of output. For example, each component of a Gaussian mixture model is Gaussian distributed. Mixture models have been well documented in literature and

widely used to model various random phenomena, including time series modeling [51]. A load pattern is basically a time series. Mathematically, given a time series of length N, a typical mixture model is formulated in terms of latent variables and observables, denoted by $\boldsymbol{z} = \{z_1, \dots, z_N\}$ and $\boldsymbol{y} = \{y_1, \dots, y_N\}$ respectively. Each observable is distributed according to a mixture of components specified by K latent variables. Considering $i = 1, \dots, N$, $z_i \in \{1, \dots, K\}$, it is assumed that the series is generated by a simple process in which the latent variable z_i is first sampled and then the observable y_i is sampled from a distribution conditional on z_i, i.e.

$$p(z_i, y_i) = p(z_i)p(y_i|z_i) \tag{3.1}$$

where $p(z_i)$ is always multinomial, and all $p(y_i|z_i)$ usually belong to the same parametric family of distributions but with different parameters. Figure 3.1 illustrates a mixture model with K components.

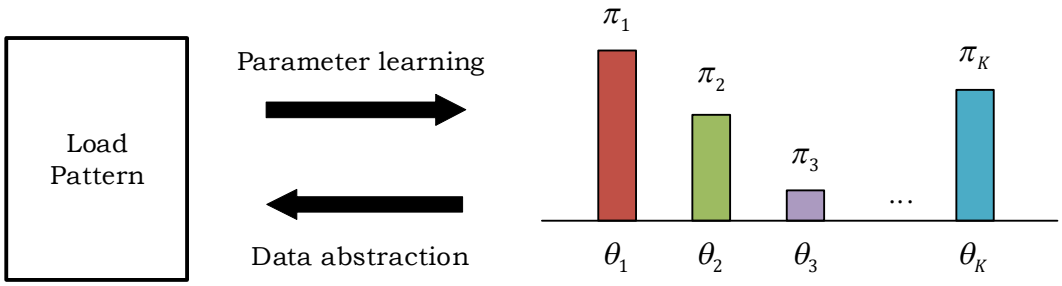


Figure 3.1: Illustration of mixture model

Selecting an appropriate number of the mixture components, K is a commonly reported issue while employing the mixture models. In normal practices K is usually predetermined or given a finite limit based on some criteria, which however has a significant effect on the modeling accuracy [52]. To tackle with this issue, a Bayesian nonparametric approach, DPMM, by which the finite limit for K is taken to infinity, is proposed in this research.

3.2.2. Bayesian Posterior Inference

For better understanding BNP and the proposed DPMM, we first briefly introduce the Bayesian inference, which is a powerful statistical method to model random variables. In Bayesian statistics, the model parameters are considered uncertain and to be drawn from some probability distributions. The essence of Bayesian inference is encapsulated by Bayes' Theorem of conditional probabilities:

$$P(\boldsymbol{\mu}|\boldsymbol{D}) = P(\boldsymbol{\mu})P(\boldsymbol{D}|\boldsymbol{\mu})/P(\boldsymbol{D}) \tag{3.2}$$

- D is the data.
- $\boldsymbol{\mu}$ is the vector of model parameters.
- $P(\boldsymbol{\mu})$ is the probability of the parameters before seeing the data.
- $P(\boldsymbol{\mu}|\boldsymbol{D})$ is the probability of the parameters given the data.
- $P(\boldsymbol{D}|\boldsymbol{\mu})$ is the probability of the data under the parameters.
- $P(\boldsymbol{D})$ is the probability of data under any parameters.

Formally, $P(\boldsymbol{\mu})$ is called the prior distribution, $P(\boldsymbol{\mu}|\boldsymbol{D})$ is called the posterior distribution and $P(\boldsymbol{D}|\boldsymbol{\mu})$ is called the likelihood function. $P(\boldsymbol{D})$ is considered as a normalizing constant. The prior distribution is chosen based on our domain-knowledge of the parameter to be estimated, before any data sample is considered. The likelihood is the probability of observing the data given the prior hypothesis. The normalizing constant, $P(\boldsymbol{D})$, functions to ensure that the integral of the resulting posterior distribution equals to one. However, the computation of this constant shows extremely high complexity. In most cases equation (3.2) is written as

$$P(\boldsymbol{\mu}|\boldsymbol{D}) \propto P(\boldsymbol{\mu})P(\boldsymbol{D}|\boldsymbol{\mu}) \tag{3.3}$$

Equation (3.3) shows that the posterior is proportional to the likelihood times the prior.

Assuming that we want to infer an underlying and unknown distribution from a given observed

data set, the Bayesian inference addresses this problem by placing a prior over the unknown distribution and computing the posterior. The procedure generally consists of three steps:

1) Choose a prior distribution that represents our beliefs about probability of the parameters;
2) Choose a statistical model that represents the conditional relation about the data given the parameters;
3) Calculate the posterior distribution by updating the prior with the likelihood given the observations.

3.2.3. Bayesian Non-parametric Approach and DPMM

Choosing a model at an appropriate level of complexity, e.g., the number of mixture components for mixture models or the number of factors for factor analysis, based on the data or system is paramount. As stated in the specialized literature, BNP approach can be one of the best solutions to address this problem [53]. A BNP model is basically defined on an infinite-dimensional parameter space chosen as the set of all possible solutions for a given learning problem; however, on a finite sample, it is evaluated by using a finite subset of available parameters to explain the sample [54]. In other words, BNP model can adapt model complexity to the data; when more data is available, the model complexity grows by including more parameters, e.g. mixing components in the mixture models.

The Dirichlet Process (DP) is a stochastic process and is one of the most popular process used in BNP models [55]. Similar to the Gaussian process which has Gaussian distributed finite dimensional marginal distributions, the DP has Dirichlet distributed finite dimensional marginal distributions. The Dirichlet distribution belongs to the family of continuous multivariate probability distributions,

and is often used as a prior distribution in Bayesian inference as the conjugate prior of the multinomial distribution. The formal definition of the Dirichlet distribution is defined as follows. Let $L^k = (L_1, \dots, L_k)$ be a vector with k non-negative components and $\sum_{i=1}^{k} L_i = 1$. Let $\alpha^k = (\alpha_1, \dots, \alpha_k)$ be a vector with k positive components. The PDF for Dirichlet distribution is expressed as

$$f(l^k) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} l_i^{\alpha_i - 1} \tag{3.4}$$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, and $l_i > 0$, $\sum_{i=1}^{k} l_i = 1$. In the following contents, the Dirichlet distribution is denoted by $Dirichlet(\alpha_1, \dots, \alpha_k)$.

The original definition of DP is given by [56]. A random distribution is said to follow a DP prior with base distribution G_0 being a distribution over the parameter space Θ and concentration parameter α, denoted $G|\alpha, G_0 \sim DP(\alpha, G_0)$, if

$$\big(G(A_1), \dots, G(A_K)\big) \sim Dirichlet(\alpha G_0(A_1), \dots, \alpha G_0(A_K)) \tag{3.5}$$

where $\{A_1, \dots, A_K\}$ is an arbitrary partition of the parameter space Θ.

The above mathematical expression is obscure. To better understand the DP, consider the following distribution:

$$\pi \sim \lim_{K \to \infty} Dirichlet(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}) \tag{3.6}$$

For each draw from the above Dirichlet distribution, a draw from the base distribution is associated:

$$\theta_k \sim G_0 \ \ for \ \ k = 1, \dots, \infty \tag{3.7}$$

Let $G \coloneqq \sum_{k=1}^{\infty} \pi_k \, \delta_{\theta_k}$,where δ_{θ_k} represents a point mass at position θ_k, then G can be seen as an infinite discrete distribution over the parameter space Θ. In this way the above procedure is considered as a DP, written as:

$$G \sim DP(\alpha, G_0) \tag{3.8}$$

It can be clearly seen that the samples drawn from the DP are discrete distributions, with the parameters associated with each point mass drawn from the base distribution G_0, and the corresponding weight drawn from an infinite dimensional Dirichlet distribution. Simply put, a DP is a distribution over distributions.

In the application of inferring mixture models, the non-parametric nature of DP allows the designer to have a mixture model with countably infinite mixture components, known as the DPMM with specific prior on the weights and the component-specific parameters. Simply put, the actual number of components is not fixed or limited by a maximum number, and can be automatically inferred from the data with the limit taken into infinity. Formally, letting θ_i be the distribution parameters of the component associated with the i^{th} data point, the time series $\boldsymbol{y}$ is generated by the following DPMM:

$$\begin{gathered} y_i|\theta_i \sim f(y_i; \theta_i) \\ \theta_i|G \sim G \\ G|\alpha, G_0 \sim DP(\alpha, G_0) \end{gathered} \tag{3.9}$$

where $DP(\alpha, G_0)$ denotes a DP which is parameterized by a concentration parameter α, and a base distribution G_0 which is the prior distribution over component parameters. It might be helpful to mention that α is strictly positive and determines the extent of the discretization of the output distribution.

To effectively represent a DP and introduce the later sampling procedure, the DPMM is employed via the CRP [57]. Consider a Chinese restaurant with potentially infinite number of tables, each of which can seat unlimited number of customers. In this representation, each observable is equivalent to a customer and the corresponding mixture component is thus equivalent to the table where the customer sits. Imagine that new customers stream in one by one and each sits at a table.

The first customer always sits at the first table. Generally, with $n-1$ customers already in the restaurant, the n^{th} customer will either choose an occupied table k with probability $\frac{n_k}{n-1+\alpha}$ or an unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$, where n_k is the number of customers sitting at table k, and α determines how likely a customer will sit at a new table. The procedure is illustrated in Figure 3.2.

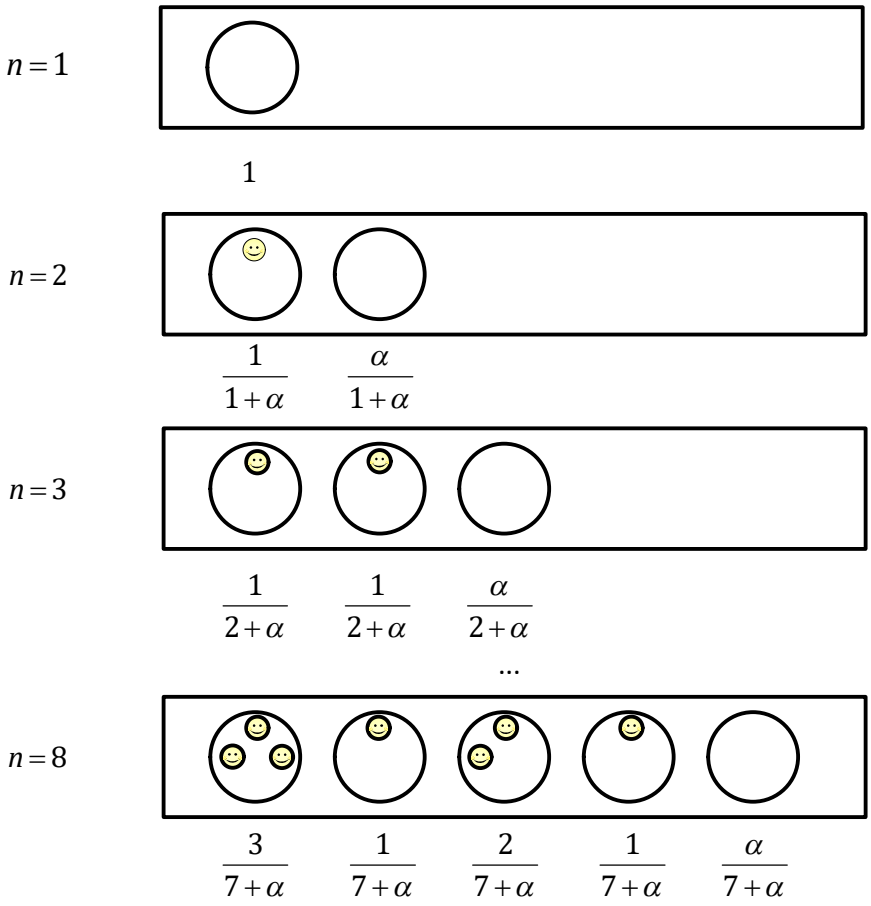


Figure 3.2: Graphic example of the CRP

Let latent variable z_i denotes the table index of the i^{th} customer, the above process can be summarized as:

$$P(z_n|z_1,\dots,z_{n-1}) = \begin{cases} \dfrac{n_k}{n-1+\alpha}, & if\ occupied\ table \\ \dfrac{\alpha}{n-1+\alpha}, & if\ unoccupied\ table \end{cases} \tag{3.10}$$

Consider that each table k is associated with a mixture component with parameter θ_k^* drawn from G_0, and each observable y_i is a customer seated at table z_i with its value drawn from $f(z_i|\theta_{z_i}^*)$. Then, (3.9) can be equivalently expressed as the following process

$$\begin{aligned} z_i &\sim CRP(\alpha) \\ \theta_k^*|G_0 &\sim G_0 \\ y_i|z_i,\{\theta_k^*\} &\sim f(\theta_{z_i}^*) \end{aligned} \tag{3.11}$$

where θ_k^* are parameters corresponding to component k and $\theta_i = \theta_{z_i}^*$. Figure 3.3 visually shows a graphical model representation of the DPMM, where each node in the graph is associated with a random variable and each rectangle denotes replication of the process within the rectangle.

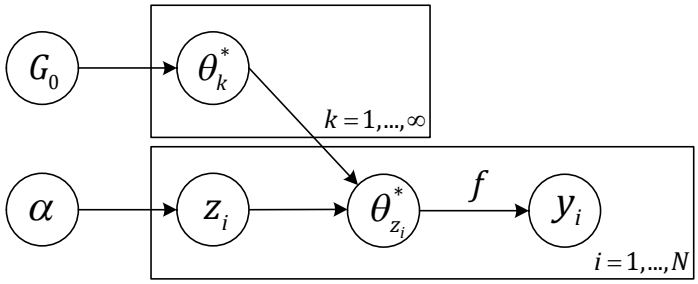


Figure 3.3: Schematic representation of DPMM

3.2.4. Model Inference by MCMC – Gibbs Sampling

In this research the Bayesian inference is employed to estimate the proposed DPMM. By applying Bayes' Theorem, the posterior distribution of latent variables can be calculated by:

$$P(z_i = z|z_{-i}, y_i, \theta_z^*) \propto P(y_i|\theta_z^*, z_i, z_{-i})P(z_i = z|z_{-i}) \tag{3.12}$$

$$= \begin{cases} \frac{N_{-i,z}}{N-1+\alpha} f(y_i;\theta_z^*) & for\ existing\ z \\ \frac{\alpha}{N-1+\alpha} \int f(y_i;\theta^*) dG_0(\theta^*) & for\ new\ z \end{cases}$$

Then we obtain the parameters θ^* associated with component z

$$\begin{aligned} P(\theta^*|y_{i:z_i=z}) &\propto \prod_{i:z_i=z} P(y_i|\theta^*)P(\theta^*) \\ &= \prod_{i:z_i=z} f(y_i;\theta^*)G_0(\theta^*) \end{aligned} \tag{3.13}$$

The computation of the above posterior distributions is intractable considering the limiting case $K \to \infty$. Rather, we adopt Gibbs sampling [58] which is an MCMC technique to sample from the posterior distributions. The basic idea of MCMC is to do independent and identically distributed sampling from a target distribution Ω via Markov chain mechanism. After N samples $\{\boldsymbol{s}^{(i)}\}_{i=1}^N$ obtained from the sampling procedure, the target distribution can be approximated by the following empirical point-mass function:

$$\Omega_N(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{s}^{(i)}}(\boldsymbol{s}) \tag{3.14}$$

and any description of the target distribution (some expected value of a function f) can be computed by:

$$E[f(\boldsymbol{s})]_\Omega \approx \frac{1}{N}\sum_{i=1}^{N} f\left(\boldsymbol{s}^{(i)}\right) \tag{3.15}$$

In this spirit, the Monte Carlo estimation of the model is given by averaging on a set of effective samples each of which represents one possible realization of the model.

As an MCMC sampling algorithm, the Gibbs sampler updates each variable in turn by sampling from its posterior conditional on other variables. Formally, given a D-dimensional variable vector

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_D)$ and a prior distribution q_0, a generic Gibbs sampler can be described by Algorithm 1:

Algorithm 1: Gibbs sampler

set $t = 0$ and initialize $\boldsymbol{\mu}^{(0)} \sim q_0$

for $t = 1, \dots, T$ repeat

for each dimension $i = 1, \dots, D$

draw $\boldsymbol{\mu}_i^{(t)} \sim P(\boldsymbol{\mu}_i | \boldsymbol{\mu}_1^{(t)}, \dots, \boldsymbol{\mu}_{i-1}^{(t)}, \boldsymbol{\mu}_{i+1}^{(t)}, \dots, \boldsymbol{\mu}_D^{(t)})$

end

end

To apply Gibbs sampling to the inference for DPMM, conjugate prior [59] is used over the component parameters θ^* so that it only needs to integrate out the parameters and sample latent variables z_i. This method is known as Collapsed Gibbs Sampling first proposed by [60]. As the power consumption series is only one-dimensional, we assume that the distribution associated with each component is Gaussian and is endowed with a Normal-Inv-Gamma (NIG) prior [61], i.e.

$$\begin{aligned} P_0(\mu, \sigma^2) &= NIG(m_0, V_0, a_0, b_0) \\ &= N(\mu | m_0, \sigma^2 V_0) IG(\sigma^2 | a_0, b_0) \end{aligned} \tag{3.16}$$

with the following substitutions

$$m_0 = \mu_0 \quad V_0 = \frac{1}{\kappa_0} \quad a_0 = \frac{\nu_0}{2} \quad b_0 = \frac{\nu_0 \sigma_0^2}{2} \tag{3.17}$$

Due to conjugacy, the posterior is also NIG, i.e.

$$P(\mu, \sigma^2 | \boldsymbol{D}) = NIG(m_n, V_n, a_n, b_n) \tag{3.18}$$

with

$$\begin{aligned} V_n^{-1} &= V_0^{-1} + n \\ \frac{m_n}{V_n} &= V_0^{-1} m_0 + n\bar{x} \end{aligned} \tag{3.19}$$

$$a_n = a_0 + \frac{n}{2}$$

$$b_n = b_0 + \frac{1}{2}(m_0^2 V_0^{-1} + \sum_{i=1}^{n} x_i^2 - m_n^2 V_n^{-1})$$

where $\boldsymbol{D} = \{x_1, \dots, x_n\}$. With these settings, the collapse Gibbs sampler is implemented for the proposed DPMM by following the CRP, as shown in Algorithm 2.

Algorithm 2: Collapse Gibbs sampler

set $t = 0$ and initialize $\boldsymbol{z}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\sigma}^{(0)}$

for $t = 1, \dots, T$ in each iteration repeat

  $\boldsymbol{z} = \boldsymbol{z}^{(t-1)}$

for $i = 1, \dots, N$ repeat

  Remove observable y_i from component z_i.

  If y_i is the only point in its current component, remove

  z_i and decrease K by 1.

  for $k = 1, \dots, K$ repeat

    Draw a new sample for z_i from

    $P(z_i = k, k \leq K) \propto \frac{N_{-i,z}}{N-1+\alpha} N(y_i|\mu_k, \sigma_k^2)$

    $P(z_i = K+1) \propto \frac{\alpha}{N-1+\alpha} \int N(y_i|\mu, \sigma^2) P_0(\mu, \sigma^2) d\mu d\sigma$

    If $z_i = K + 1$, sample parameters for this new

    Component from $P_0(\mu, \sigma^2)$ and increase K by 1.

  end

  set $\boldsymbol{z}^{(t)} = \boldsymbol{z}$

end

end

It should be noted that the sampler is initialized with random values; under this circumstance, the first few samples should be discarded because they may not represent the actual posterior distribution. Such discarded iterations are known as burn-in period [62]. All effective samples will

then be used as input of the next training procedure. It is also worth mentioning that the proposed algorithm can efficiently realize suitable extensibility by selecting not only Gaussian distribution but any conjugate priors. Besides, uncertainty modeling can be incorporated in higher dimensional TS for example by placing a multivariate conjugate prior, for example, a Normal-Inverse-Wishart Distribution prior [59]. This paper only verifies the proposed algorithm with one-dimensional TS, though other forms of TS can be employed without generality.

3.3. Probabilistic Forecasting

Load forecasting is generally a regression problem based on previous electric load patterns and relevant variables. Instead, our proposed DPMM considers each load point to be drawn from the distribution associated with a specific mixture component represented by discrete latent variables. This representation inspires us to artfully transform a regression problem to a classification problem. In this spirit, an ensemble tree-based classification learning method is proposed to capture the mapping relation between the component assignment, $\boldsymbol{z}$, and the relevant exogenous variables, also called as predictors, denoted by $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ where each $\boldsymbol{x}_i$ being a vector of exogenous variables. Moreover, the Bagging aggregation technique is employed in this paper as it gives the exact probability scores for classification. Taking advantage of the probabilistic classification scores, the forecasted results can be expressed as a Gaussian mixture distribution by aggregating each component proportional to its forecasted probability. The final probabilistic forecast is averaged on all MCMC samples. The proposed forecasting method is detailed in the next sub chapter.

3.3.1. Ensemble Learning and Bagging Algorithm

Ensemble learning [63] is a machine learning process that com-bines diverse set of models together to achieve better learning performance than just using a single model. Figure 3.4 graphically presents the basic idea of ensemble learning. In ensemble learning, each model is called a base learner and is generated by a base learning algorithm usually selected from splines, decision tree, ANN, etc. The mechanism of training and aggregating base learners leads to different ensemble methods, where Bagging is implemented in this paper.

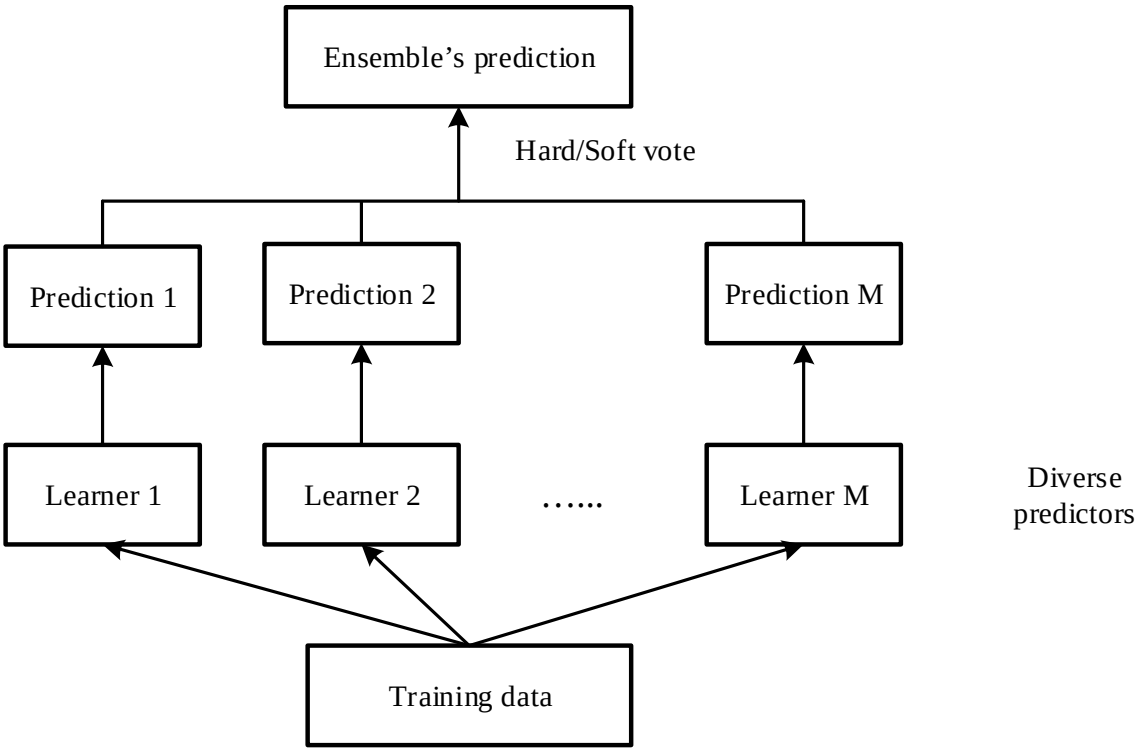


Figure 3.4: Graphical representation of general ensemble learning

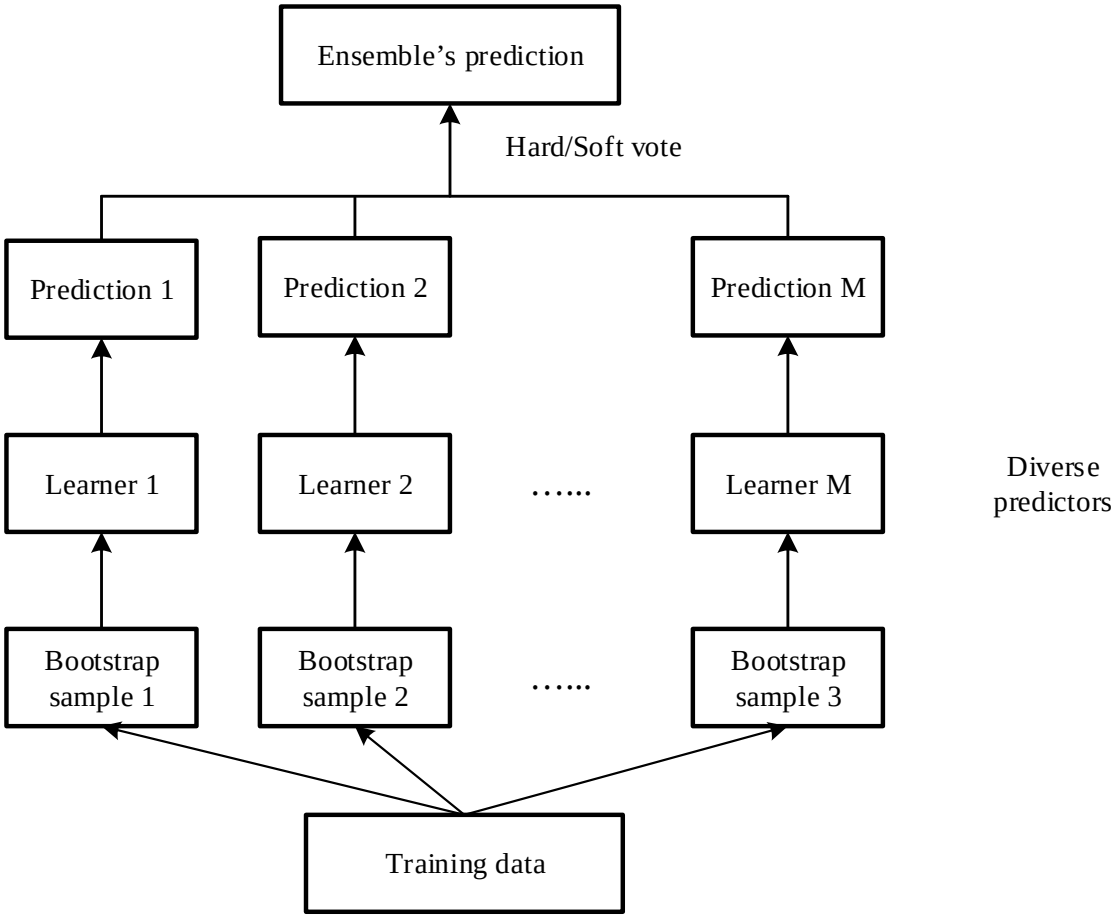


Figure 3.5: Graphical representation of Bagging learning method

Bagging, which is also known as bootstrap aggregating, is one of the earliest, most intuitive and simplest ensemble algorithms. Bagging trains each base learner with a randomly drawn subset of the training set with replacement by a base learning algorithm. The learners are then combined by having each in the ensemble vote with equal weight and the most voted class is predicted. Thus, the associated voting weight of each class reflects its prediction probability [64]. The implementation of bagging is very simple as provided by Algorithm 3, and illustrated in Figure 3.5.

Algorithm 3: Bagging

Input Data set $\boldsymbol{\mathcal{D}}$, Base learning algorithm $\boldsymbol{\mathcal{C}}$, Ensemble size $\boldsymbol{T}$, Percent $\boldsymbol{\mathcal{R}}$ to create bootstrapped training data

for $\boldsymbol{t = 1, \ldots, T}$ repeat

Generate a bootstrapped sample $\boldsymbol{\mathcal{D}_t = bootstrap(\mathcal{D}, \mathcal{R})}$

Train a base learner $\boldsymbol{c_t = \mathcal{C}(\mathcal{D}_t)}$

end

Obtain total vote received by each class $\boldsymbol{k}$

$$\boldsymbol{V_k = \sum_{t=1}^{T} \mathbb{I}(z = c_t(X))}$$

where $\mathbb{I}(*)$ gives 1 if * is true and 0 otherwise

It is shown in the specialized literature that the Bagging technique achieves better performance when a base learner with high variance and low bias is used [65]. Bias are caused by the simplifying assumptions that are made by the model and variance is caused by the use of different training data which will change the estimated target function. It is obvious that decision trees (DTs) are a typical example of model with high variance and low bias, since DTs make almost no assumptions about the target function and are very sensitive to variance in training data. Thus, the DT is utilized as the base learner in this work, though any other method can be used without loss of generality.

3.3.2. Probabilistic Forecasting

Assume M effective samples are generated by the DPMM fitting process, each of which is denoted by $\boldsymbol{\mathcal{D}_j}$, $j = 1, \ldots, M$. For every $\boldsymbol{\mathcal{D}_j}$, train an ensemble DT based on the bagging algorithm introduced in Section III.A, and let $V_{j,l,k}$ denote the vote score of class k, which is associated with the l^{th} step ahead forecast. Then the predictive density of l^{th} step ahead prediction value, y_{N+l}, for the j^{th} DPMM sample can be expressed as a Gaussian mixture, i.e.,

$$P^{(j)}(y_{N+l}|y_{1:N}, \boldsymbol{x}_{1:N+l}) = \sum_{k=1}^{K} \frac{V_{j,l,k}}{\sum_{k=1}^{K} V_{j,l,k}} \cdot \mathcal{N}(\mu_k, \sigma_k^2) \tag{3.20}$$

In the spirit of MCMC, the l^{th} step ahead probabilistic load forecast can be approximated using equation (3.14), i.e.

$$P(y_{N+l}|y_{1:N}, \boldsymbol{x}_{1:N+l}) = \frac{1}{M} \cdot \sum_{j=1}^{M} P^{(j)}(y_{N+l}|y_{1:N}, \boldsymbol{x}_{1:N+l}) \tag{3.21}$$

The full structure of our proposed methodology is illustrated graphically in Figure 3.3. As it can be seen in this flowchart, the final PLF is formulated by averaging a set of PLFs, each of which corresponds to a mixture sampled from the MCMC procedure and is generated through the same classification ensemble model. Such output combination via averaging reduces the risk of overfitting and selecting a poor mixture model.

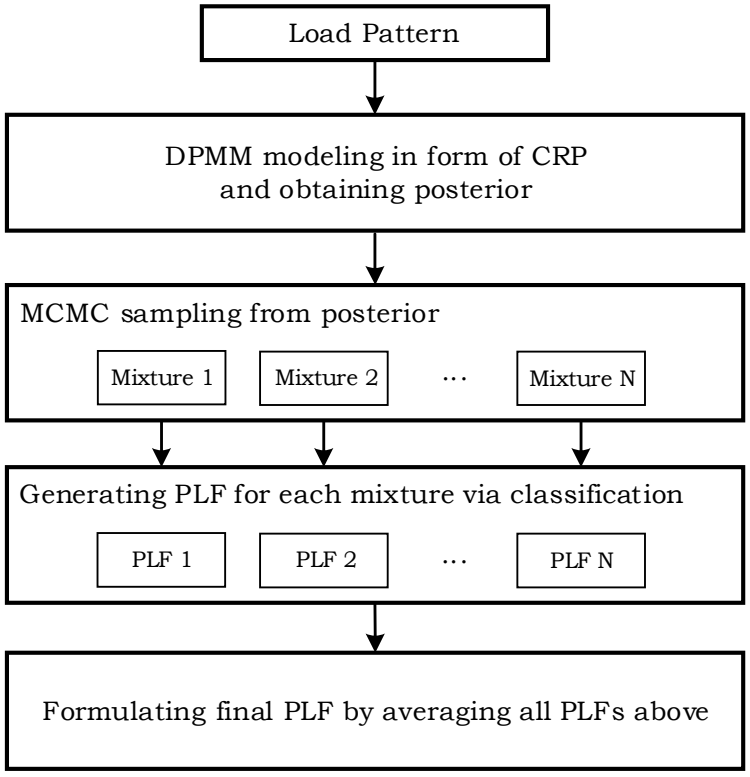


Figure 3.6: Schematic structure of the proposed methodology

3.4. Summary

In this chapter, the proposed method for short-term probabilistic load forecasting at local level is presented. The DPMM, which is a Bayesian nonparametric model, is introduced to model the uncertainty underlying the load pattern. The Bayesian inference is proposed to infer the DPMM and the Gibbs sampler, which is an MCMC sampling technique, is proposed to sample from the posterior. The obtained mixture model representation is then utilized to train an ensemble learning model to generate PLFs. The basic concepts and mathematical expressions of the above-mentioned methods are briefly discussed. It is believed that the DPMM can adjust the model complexity to the data so that overfitting is avoided, and moreover, the averaging over all effective MCMC samples can greatly reduce the variance of the result, thus enhancing the model consistency, which is examined in the next chapter.

4. Intro-Day Probabilistic Load Forecasting at Local Level Using Smart Meter Data

4.1. Introduction

In this chapter, case studies are carried out to examine the performance of the proposed approach. Two smart meter data sets are used, and different aggregation levels of the smart meter data are tested. The detail of the two data sets is described in the following charters. A 3-month period from 1st May to 1st August, which is covered by the whole summer season, is picked for simulation for each data set, so that seasonality is neglected in the following case studies. Calendar variables, weather conditions and lagged demands are considered as exogenous factors. To evaluate the performance on STPLF, the case studies in this chapter focuses on day-ahead forecasts. All forecasting methods in the following experiments are tested with Matlab scripts, using a personal computer featured Intel 3.4–GHz CPU with 16 GB of RAM.

4.2. Smart Meter Data Description and Test Settings

The first data set is a public data set initiated by the UK Power Networks led Low Carbon London (LCL) project [66]. We pick a 3-month period from 1st May to 1st August of 2013. Historical hourly local weather data is downloaded from [67]. It is assumed that weather conditions stay fixed within each hour. For calendar effects, the period-of-day, day-of-week and holidays, which are categorical, are included.

The second data set is a local data set from Saskatoon Light &Power (SLP) AMI Meter Replacement Program which offers electricity consumption at a granularity of 30 minutes of

more than 65,000 electricity customers including industrial, commercial and residential sectors in the city of Saskatoon, Canada. More than 99% of the customers of SLP have smart meters installed. As a match, we also pick the same 3-month summer period of 2018 and include the same type of exogenous variables. The local weather data is collected from [68].

The first 3-month data (from 1st May to 31st July) is used as training set. The training set contains 4416 data points (half-hourly data) and is firstly used to fit the DPMM. 800 samples are generated from the posterior and the last 100 samples are selected for training the ensemble learning model. The remaining data (1st August) is used as a test set to evaluate foresting performance. The public data set is first examined and the local data set is then tested to confirm our conclusion.

4.3. Benchmarks and Evaluation Criteria

4.3.1. Benchmarks for PLF

We verify our proposed model performance for PLF. Due to the high volatility, both naïve methods and the most popular state-of-the-art methods are used as benchmarks.

1) *Unconditional Empirical Distribution*: The empirical distribution is computed based on all historical observations without conditioning on any exogenous variables, and is used as the most naïve prediction. This method is denoted UNED.

2) *Empirical Distribution Conditional on Time-of-Day*: The whole data set is divided into several subsets based on the segmentation of the time-of-day. The empirical distribution is computed for each subset and is used as the prediction for each time period of the forecasted day. This method is denoted CED.

3) *Conditional KDE on Time-of-Day and Day-of-Week*: We also include one of the most used nonparametric method, conditional KDE as a benchmark to formulate and forecast probability densities conditional on time-of-day and day-of-week. This method is denote CKDE.

4) *Simulation of ARIMA Model Using Inferred Residuals*: In this benchmark the ARIMA model is used to model the time series and sample paths are simulated by using the observed series and inferred residuals. This method is denoted ARIMARES.

5) *QR Tree Ensemble*: One good implementation of conducting nonlinear QR is using quantile random forest which is a regression tree ensemble. This method estimates conditional quantiles of the load given predictor data. We recover the predictive distribution from the quantiles by computing a large set of quantiles. This method is denoted QRTE.

4.3.2. Evaluation Criteria

Evaluating the forecasting accuracy of PLF requires specific numerical measures such as the Brier score, the Winkler score, the pinball loss function, etc. In this paper, we use the continuous ranked probability score (CRPS) which is defined as the integral of the difference between the cumulative distribution function (CDF) $F(y)$ of the generated density and the observable y^*. Formally,

$$CRPS(F, y^*) = \int_{-\infty}^{+\infty} \big(F(y) - \mathbb{1}(y - y^*)\big)^2 dy \tag{4.1}$$

where $\mathbb{1}$ is the Heaviside step function whose value is 1 if the real argument is non-negative otherwise the value is 0. Figure 4.1 presents the graphical representation of the CRPS. The area of the shaded part is exactly the value of CRPS for the illustrated An alternative representation is given by

$$CRPS(F, y^*) = E_F|Y - y^*| - \frac{1}{2}E_F|Y - Y'| \tag{4.2}$$

where Y and Y' are independent random variables distributed according to the same cumulative distribution F. Eq. (4.2) shows that the CRPS generalizes the mean absolute error (MAE) to the case of probabilistic forecasts. In contrast with other probabilistic forecast measures, the CRPS considers the distribution of forecasts as a whole instead of focusing on specific points of the probabilistic forecasts. It quantifies both the calibration and sharpness [69] of the predictive distribution thus providing a comprehensive evaluation. By representing F through an L-ensemble $y_{i=1,\dots,L}$, Eq. (4.2) leads to the following estimator

$$\widehat{CRPS}_{NRG}(F, y^*) = \frac{1}{L}\sum_{i=1}^{L}|y_i - y^*| - \frac{1}{2L^2}\sum_{i,j=1}^{L}|y_i - y_j| \tag{4.3}$$

To make comparisons with benchmarks which give probabilistic forecasts in the form of quantiles, we recover the predictive distribution from these quantiles using linear interpolation and draw samples via inverse transform sampling, i.e. $Y = F_Y^{-1}(U)$ and U is a continuous uniform variable in [0,1].

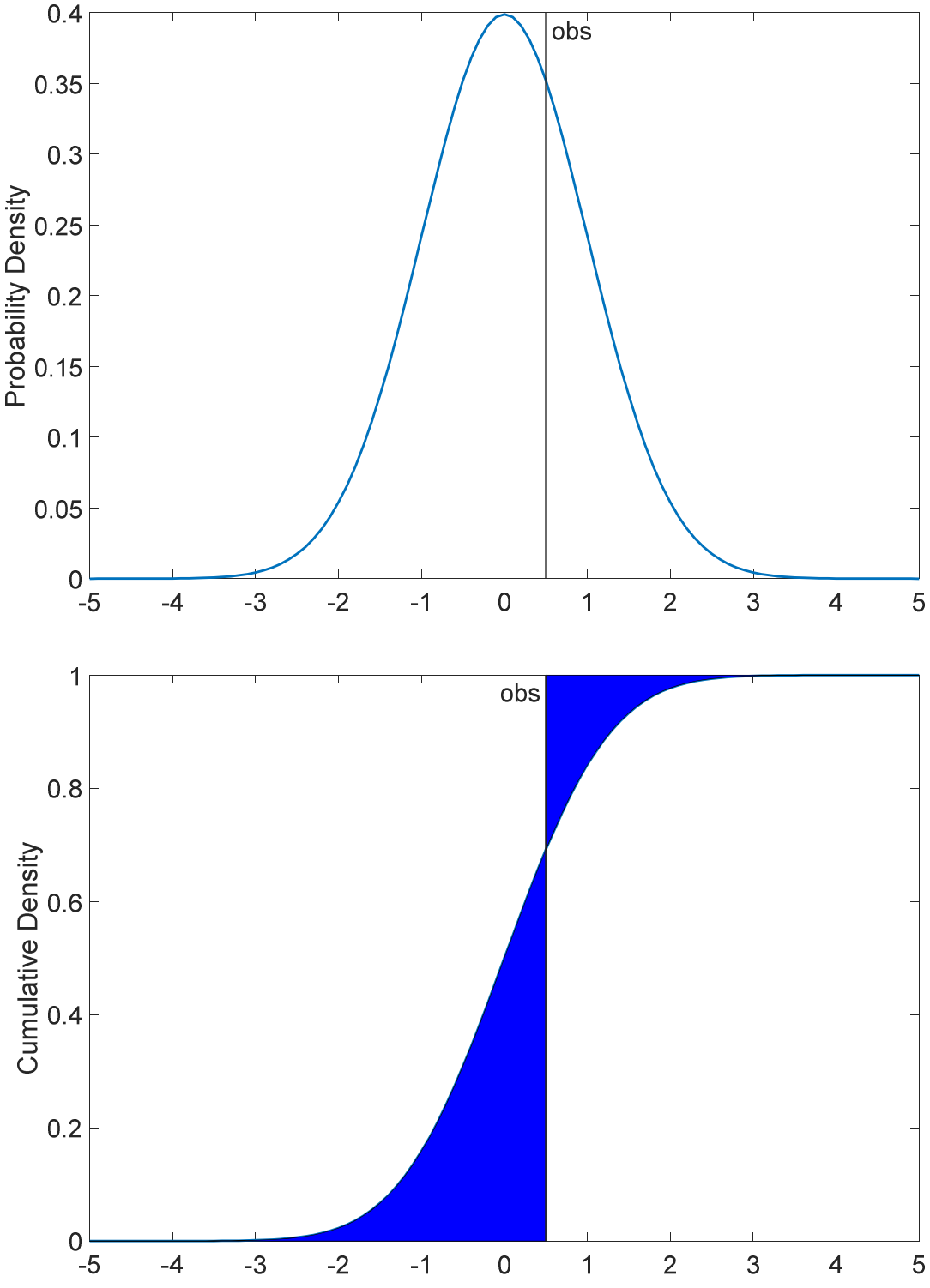


Figure 4.1: Graphic representation of the calculation of CRPS

4.4. Case Studies and Results of the LPL Data set

We randomly choose five groups of smart meters and form aggregations since no location information is provided for this data set. The five aggregation levels are set to include 20, 50, 100, 200 and 400 individual customers, respectively. Table 4.1 shows the comparison of the CRPS of the five different level forecasts between the proposed method and the benchmarks, averaged over $H = 48$ steps. As it can be seen in this table, the proposed method performs slightly better than QRTE and significantly outperforms the rest of the benchmarks in terms of the overall accuracy measured by average CRPS, at all tested aggregation levels. By conditioning on time-of-day, CED greatly improves the accuracy compared to UNED, which indicates that time-of-day is a good variable for STPLF. CKDE and ARIMARES show similar or even worse result compared to CED, which is probably due to that CKDE is limited by the very small amount of samples and ARIMARES only considers lags of dependent variable and does not incorporate exogenous variables that play a very important role in electric load forecasting. ARIMARES also shows larger errors at late time steps, which is not plot here. To compare between different aggregation levels, we normalize the CRPS by true observation and averaged over H steps, similar to the transformation from MAE to mean absolute percentage error (MAPE), i.e.

$$CRPS_{ER} = \frac{100\%}{H} \cdot \sum_{h=1}^{H} \left| \frac{CRPS_h}{y_h^*} \right| \tag{4.4}$$

In terms of the CRPS percentage error, it is shown that the forecasting error drops when the aggregation level increases. Besides, one can notice that the gap between the performance of the proposed method and QRTE increases when the aggregation level decreases, indicating that the proposed method performs even better at low aggregation levels where load patterns show more

variability and volatility. Figure 4.2 plots the estimated probability distribution of the forecast at the 19th step ahead at aggregation level 20 for all tested methods. It is obvious that ARIMARES, CKDE, CED and UNED lack sharpness. In other words, the predictive probability distributions of these four methods are more spread than necessary. The proposed method and QRTE have similar sharpness but QRTE fails to satisfy the calibration as it has lower probability density at the real value point, compared to the proposed method.

Table 4.1: Probabilistic load forecasting CRPS summary of the LCL data set

Methods	Proposed	QRTE	ARIMARES	CKED	CED	UNED
20	0.417	0.443	0.521	0.505	0.506	0.682
	14.25%	15.39%	19.28%	20.18%	20.42%	31.41%
50	0.791	0.860	1.156	1.117	0.976	1.362
	10.76%	11.67%	14.88%	14.79%	12.81%	19.91%
100	1.240	1.296	1.822	1.813	1.795	2.674
	8.24%	8.70%	11.71%	11.61%	11.57%	20.01%
200	1.910	2.047	2.894	2.994	2.837	5.097
	6.36%	6.80%	9.15%	9.44%	8.96%	18.81%
400	2.549	2.673	2.887	3.032	2.854	5.028
	4.37%	4.63%	9.06%	9.56%	8.99%	18.51%

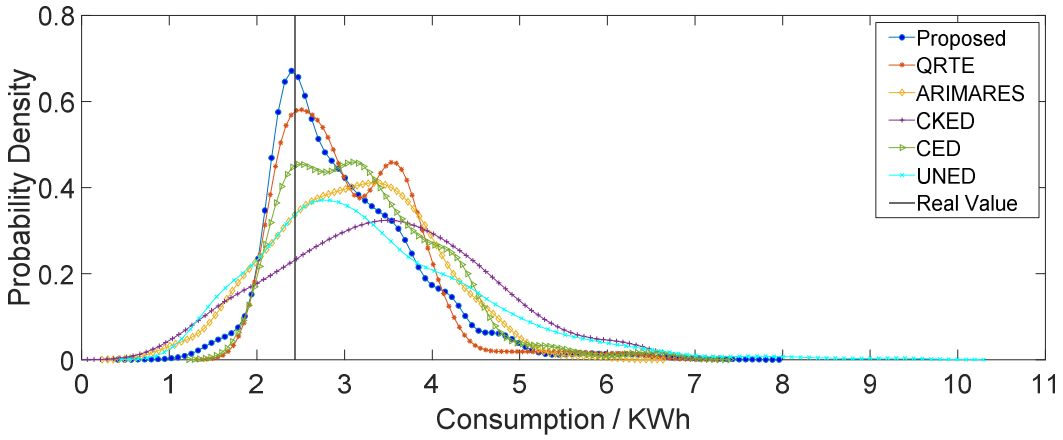
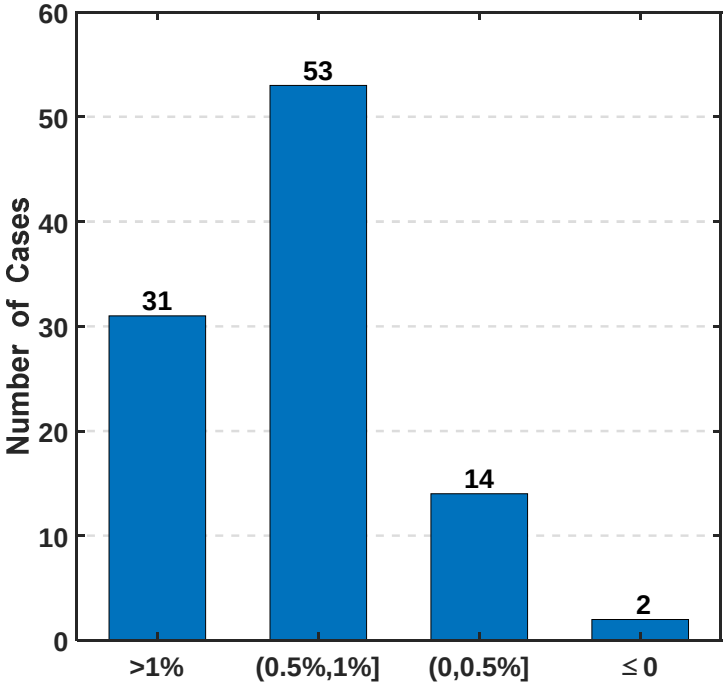
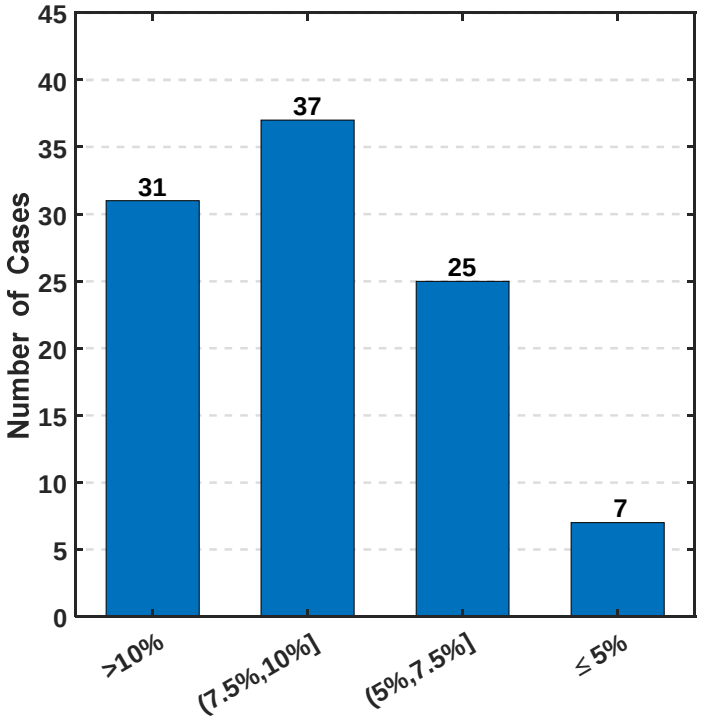


Figure 4.2: Estimated probability distribution of the forecast at the 19th step ahead at aggregation level 20 for all tested methods

4.5. Case Studies and Results of the SLP Data set

A deeper look is taken on the SLP data set, which contains extra geological information of each individual customer. Two simulations are conducted in this subsection. The first simulation compares the performance between the best two methods from the previous subsection, the proposed method and QRTE. This simulation focuses on aggregation of 20 customers (both residential and commercial customers included) at 100 random locations scattered around the city, with the help of the GIS system. 20 customers can form the size of a neighborhood feeder effectively. Next, 4 cases, each of which is randomly picked from each of the four groups shown in Figure 4.3, are selected to examine the performance of the two methods over the forecast horizon. The second simulation is designed to compare the performance of the two methods considering the random nature of stochastic algorithms, which are applied in both methods. Stochastic algorithms generate different results with different runs on the same inputs. To compare different stochastic algorithms, it is important to compare their populations of measures and report the mean and standard deviation of performance, rather than just comparing two single results. In this simulation, each of the two methods is tested 100 times on one specific aggregation of 20 customers.


Number of Cases
60
50
40
30
20
10
0
31
53
14
2
>1%
(0.5%,1%]
(0,0.5%]
≤ 0


Number of Cases
45
40
35
30
25
20
15
10
5
0
31
37
25
7
>10%
(7.5%,10%]
(5%,7.5%]
≤ 5%

(a) statistics of $CRPS_{ER1} - CRPS_{ER2}$ (left chart)

(b) statistics of $-(CRPS_{ER1} - CRPS_{ER2})/CRPS_{ER2}$ (right chart)

Figure 4.3: CRPS percentage error comparison between the proposed method and QRTE on 100 test cases.

Table 4.2: Mean Values Averaged on the 100 Test Cases

$\overline{CRPS_{ER1}}$	9.63%
$\overline{CRPS_{ER2}}$	10.54%
$\overline{-(CRPS_{ER1} - CRPS_{ER2})/CRPS_{ER2}}$	9.51%

The statistics of the first simulation are depicted in Figure 4.3. All analyses in this paragraph are in terms of the CRPS percentage error. It is shown in Figure 4.3 (a) that among the 100 test cases only in 2 of them the proposed method is outperformed by QRTE. The error of the proposed method is more than 0.5% less, compared to the error of QRTE in 84 of the test cases, with 30 cases showing more than 1% less error. Figure 4.3 (b) shows the relative relation of the accuracy of the two methods based on the relative percentage $\rho = \frac{-(CRPS_{ER1} - CRPS_{ER2})}{CRPS_{ER2}}$, where the subscript 1 represents the proposed method and the subscript 2 represents QRTE. As it can be seen in this chart, regarding the abovementioned relative percentage, the maximum number of cases fall in the range of [7.5%, 10%], followed by the index of nearly one third of the test cases exceeding 10%. Figure 4.3 (b) also indicates that in 93 of the 100 test cases the accuracy of the proposed method is more than 5% better, relative to the performance of QRTE. Several important mean values averaged on the 100 test cases are given in Table4.2. It can be easily calculated that the proposed method reduces the error by 0.92% on average, compared to QRTE, which can be considered significant regarding the improvement of forecasting accuracy. The performance of the proposed method is overall 9.51% better than QRTE. The abovementioned statistics

comprehensively confirm the effectiveness of our proposed method regarding its performance on PLF of low aggregation level loads.

The performance of the two methods over the forecast horizon is reported in Figure 4.3. It can be seen that in all the 4 cases, these two methods have very close forecasting performance at points where the CRPS values are relatively small, while the performance of QRTE worsens severely at points with relatively large CRPS values. It can also be noted that large CRPS value mostly happens during peak hours, while the CRPS value is small when the load curve is low and flat. The major error occurs mostly during the peak load with high variation. This result indicates that though these two methods have close performance during off-peak hours which cover most hours of the day, the proposed method performs better than QRTE when the load is associated with high uncertainty. Another noticeable point that can be observed from Figure 4.3 is that the higher the value of ρ is, the more variation the load curve possesses. In other words, the performance difference between the two methods becomes larger when more uncertainty is associated, enhancing the conclusion that our proposed method performs better than QRTE in modeling high uncertainties in low aggregation level load, which contains more variations.

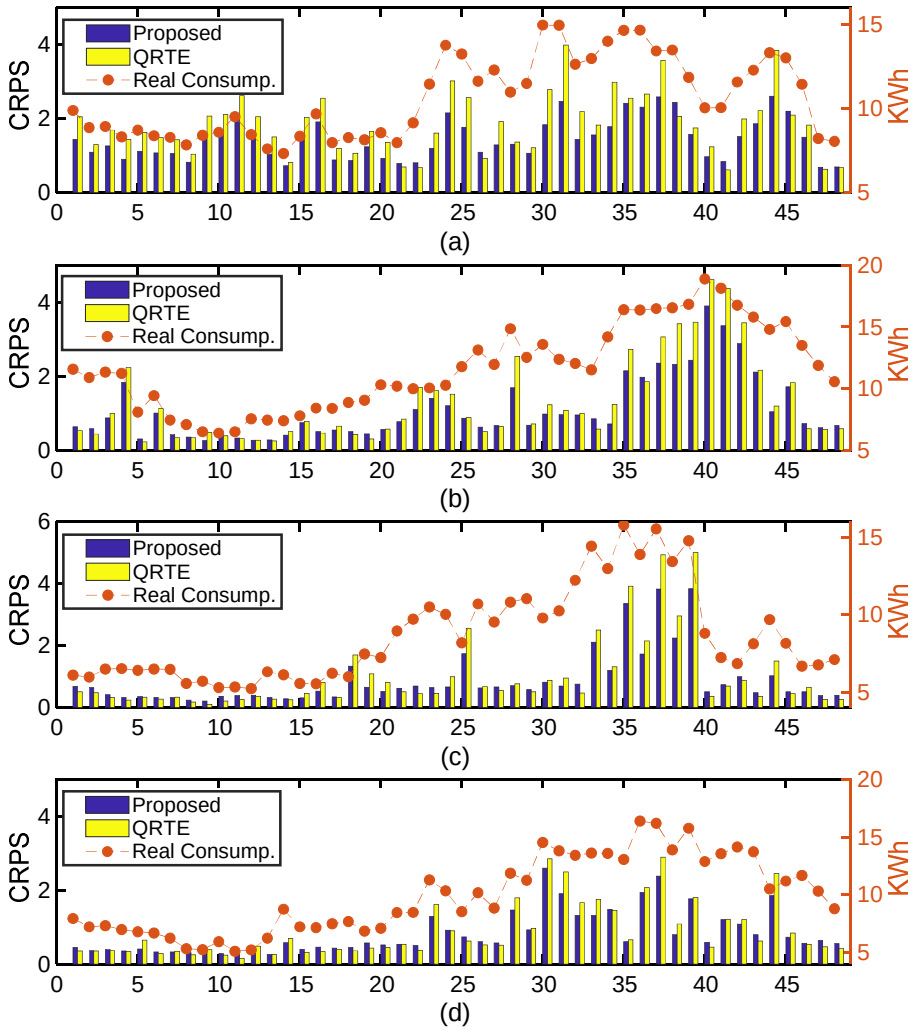


Figure 4.4: CRPS of the proposed method and QRTE over the forecast horizon:

(a) $\rho = 27.60\%$ (b) $\rho = 8.01\%$ (c) $\rho = 6.97\%$ (d) $\rho = 2.56\%$

The results of the last simulation are depicted in Figure 4.5, Figure 4.6 and Table 4.3. In this simulation, the effect of the random nature of the stochastic algorithms is examined. Figure 4.5 is a normalized histogram plot which is an estimate of the probability distribution of the CRPS percentage error of the 100 repeated experiments on the two tested methods. It can be easily observed that the measurements of the proposed method are much more narrowly centered than the measurements of QRTE which are spread out over a wider range with a few highly probable

outliers noticed. This histogram plot indicates that the proposed method can provide more consistent result than QRTE, while the wide spread range of the output of QRTE makes its result unreliable. Further, Figure 4.6 illustrates the results via a notched boxplot representation. It can be seen that the boxplot of the proposed method could sharply bound all the data points, indicating high reliability of the results. However, from the boxplot of QRTE, it can be observed that even the interquartile range box is much taller than the whole range of the plot of the proposed method, and the long whiskers betray a relatively heavy tailed population. Plus, one data point is identified as an outlier, representing an extreme case that has more than three times the height of the box. This bad data point will greatly reduce the forecasting accuracy and affect the decision-making process such as daily operations if it is outputted and accepted. As one may notice that the notches in the boxplot do not overlap, it can be concluded that the true medians of the two group of samples differ with 95% confidence. Since the median of the proposed method is much smaller than that of QRTE, it proves that the proposed method generates better result than QRTE in a certain significance. Table 4.3 reports some basic statistics calculated from the sample points. It can be seen that the standard deviation of the proposed method is quite small while that of QRTE is about 500 times larger, confirming that the proposed method holds better consistency. The mean and the median of the proposed method is almost the same. For QRTE, the median is lower than the mean, indicating that there are a few data points with extreme values that are elevating the mean. In other words, in several cases the QRTE generates results with unusually high error. All analyses in this simulation proves that the impact of the random nature of the stochastic algorithm on the proposed method is quite limited, while the QRTE may produce unreliable prediction due to the randomness of the algorithm itself.

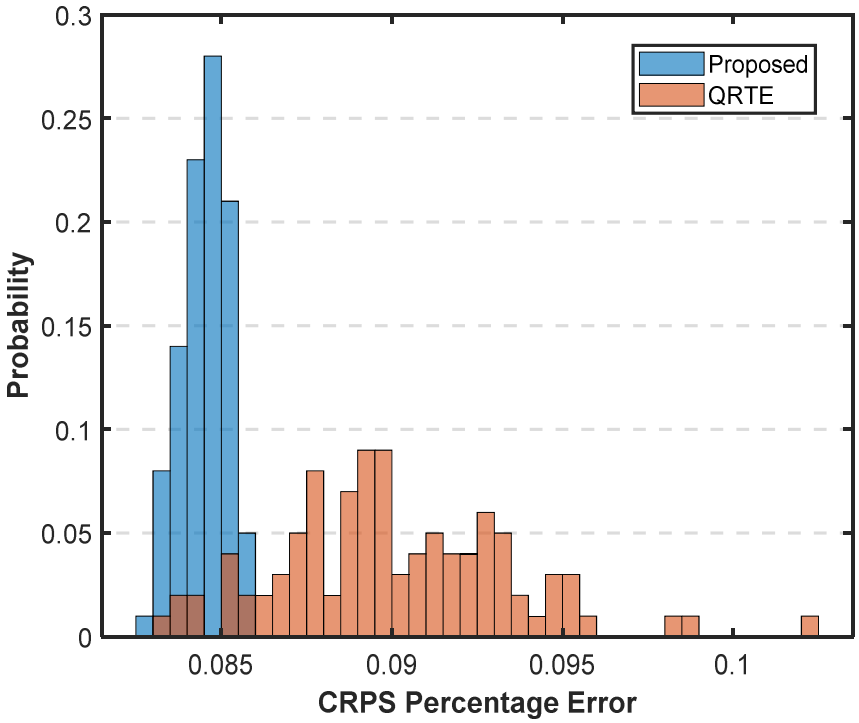


Figure 4.5: Normalized histogram plot of the CRPS percentage error of the 100 repeated experiments on the two tested methods

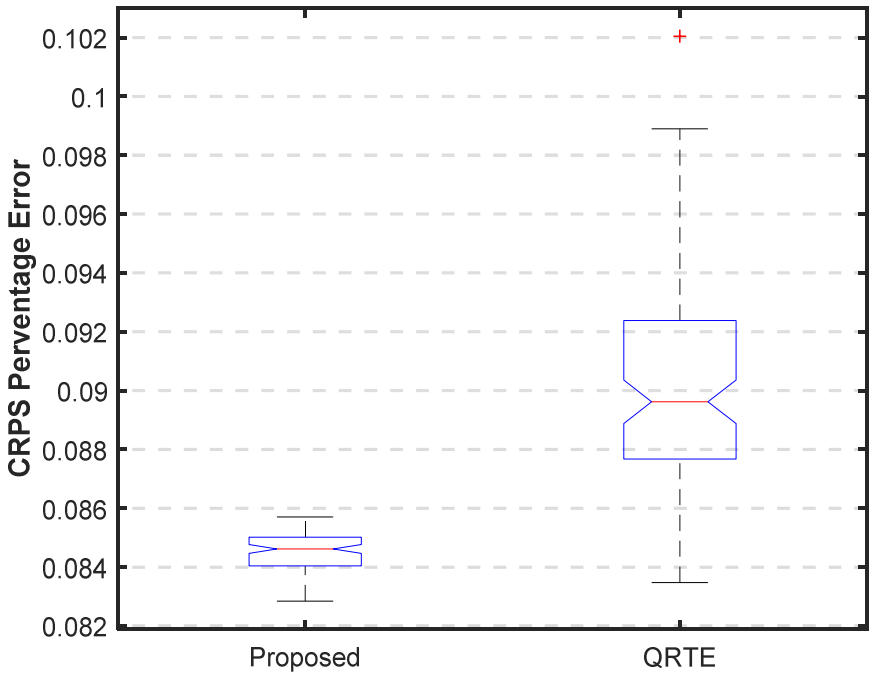


Figure 4.6: Boxplot of the CRPS percentage error of the 100 repeated experiments on the two tested methods

Table 4.3: Statistics of The 100 Repeated Runs of the Two Tested Methods

100 repeated runs	Proposed	QRTE
Median	0.0846	0.0896
Mean	0.0845	0.0901
Std	6.75e-4	0.0034

4.6. Summary

Comprehensive case studies aimed at examining the performance of the proposed method have been conducted in this chapter. Two data sets including one public data set and one local data set are tested. The first data set is used to compare the performance of the proposed method with several benchmarks including two naïve benchmarks and three most popular state-of-the-art methods. The second data set is then used to further examine the performance of the proposed method and QRTE, which are the two best methods among all the benchmarks in the first case study, specifically at the aggregation level of 20 customers. The performance of the two methods over the forecast horizon and their model consistency are also reported.

5. Summary and Conclusions

It is believed that in future networks much more variability and volatility in load patterns will be seen, both temporarily and spatially. To meet the increasing needs of stakeholders, load forecasting needs to be improved and be probabilistic, providing ranges or even the entire conditional distribution of the future load. This problem is even more challenging at local level in distribution networks.

Chapter 1 gives a brief introduction to electric load forecasting and the smart meter rollout in recent years. The challenges and opportunities that the smart meter rollout may bring to the industry especially in the area of load forecasting are discussed. Chapter 2 introduces the PLF and briefly reviews the state-of-art techniques on this problem. Both advantages and limitations of these techniques, including QR, KDE, residual simulation, scenario generation and some other methods, are discussed.

Chapter 3 presents the proposed method, which is designed into two main steps. In the first step, a load pattern, which is basically a time series, is modeled by the proposed DPMM to capture the underlying uncertainty. The DPMM is able to automatically adjust its model complexity to data, thus avoiding extra model selection process and the problem of under-fitting or overfitting. As a Bayesian model, the DPMM is inferred by the proposed Collapse Gibbs sampling, yielding a set of effective samples from the posterior. Each sample is a mixture model and is fed into the proposed ensemble learning method, along with the exogenous variables including weather and calendar variables, and lagged consumption values. Taking advantage of the Bagging algorithm, the output of each learning machine is a Gaussian mixture distribution. The final PLF is obtained by averaging on all effective samples. Simply put, the final output of the proposed model is a

mixture of Gaussian mixtures, which is believed to greatly reduce the model variance and enhance the model consistency.

Comprehensive case studies are carried out in Chapter 4 to validate the performance of the proposed method. The simulations have been conducted on both a public data set and a local distribution company data set with GIS information. Based on the results of the simulations on the first public data set, the comparison between our method and the five benchmarks show that only QRTE has close overall performance to the proposed method. It is also worth mentioned that the gaps between the performance of QRTE and the proposed method slightly increases when the aggregated level decreases, which indicates that the proposed method is more capable of handling highly volatile load patterns. The rest of the benchmarks which work well with system level load struggle in disaggregated load forecasting, and also suffer from very limited length of training phase. More detailed examinations are carried out in simulations on the local data set. The performance of the two methods over the forecast horizon shows that QRTE is outperformed by the proposed method at points with relatively large errors though they have almost the same accuracy at low error points, proving that our proposed method performs better than QRTE in modeling high uncertainties in low aggregation level load, which contains more variations. The last simulation is carried out to examine the impact of the random nature of the stochastic algorithms on the final output. The simulation results indicate that the proposed method shows better model consistency with the outputs much more narrowly centered and the sample mean smaller than those of QRTE.

There is still plenty of room for extending our method under the background of Bayesian theory. As for future works, we seek jointly modeling multiple correlated load patterns by placing multivariate priors over our proposed model, and we will examine more conjugate priors other

than Gaussian or Gaussian-Wishart distribution to look for a better fit for each load pattern among diverse types of patterns.

REFERENCES

[1] W.C. Hong, “Intelligent Energy Demand Forecasting,” London, UK: Springer, 2013

[2] A. S. Khwaja, M. Naeem, A. Anpalagan, A. Venetsanopoulos, and B. Venkatesh, “Improved short-term load forecasting using bagged neural networks,” *Electr. Power Syst. Res.*, vol. 125, pp. 109–115, Aug. 2015.

[3] D. W. Bunn and E. D. Farmer, *Comparative Models for Electrical Load Forecasting.* New York, NY, USA: Wiley, 1985, Chichester.

[4] B. F. Hobbs, S. Jitprapaikulsarn, S. Konda, V. Chankong, K. A. Loparo and D. J. Maratukulam, "Analysis of the value for unit commitment of improved load forecasts," in *IEEE Transactions on Power Systems*, vol. 14, no. 4, pp. 1342-1348, Nov. 1999.

[5] D. Srinivasan and M. A. Lee, "Survey of hybrid fuzzy neural approaches to electric load forecasting," *1995 IEEE International Conference on Systems,* Man and Cybernetics. Intelligent Systems for the 21st Century, Vancouver, BC, Canada, 1995, pp. 4004-4008 vol.5.

[6] D. R. Strong, “Impacts of diffusion policy: determinants of early smart meter diffusion in the US electric power industry,” *Industrial and Corporate Change*, vol. 28, no. 5, pp. 1343-1363, Mar. 2019.

[7] U.S. Energy Information Administration [Online]. Available: https://www.eia.gov/.

[8] D. Alahakoon and X. Yu, "Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey," in *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 425-436, Feb. 2016.

[9] J. C. López, M. J. Rider and Q. Wu, "Parsimonious Short-Term Load Forecasting for Optimal Operation Planning of Electrical Distribution Systems," in *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1427-1437, March 2019.

[10] M. Rowe, T. Yunusov, S. Haben, C. Singleton, W. Holderbaum and B. Potter, "A Peak Reduction Scheduling Algorithm for Storage Devices on the Low Voltage Network," in *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 2115-2124, July 2014.

[11] N. Ahmed, M. Levorato and G. P. Li, "Residential Consumer-Centric Demand Side Management," in *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4513-4524, Sept. 2018.

[12] S. O. Muhanji, A. Muzhikyan and A. M. Farid, "Distributed Control for Distributed Energy Resources: Long-Term Challenges and Lessons Learned," in *IEEE Access*, vol. 6, pp. 32737-32753, 2018.

[13] B. P. Hayes and M. Prodanovic, "State Forecasting and Operational Planning for Distribution Network Energy Management Systems," in *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1002-1011, March 2016.

[14] C. J. Bennett, R. A. Stewart, and J. W. Lu, “Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system,” *Energy*, vol. 67, pp. 200–212, 2014.

[15] N. Amjady, F. Keynia and H. Zareipour, "Short-Term Load Forecast of Microgrids by a New Bilevel Prediction Strategy," in *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 286-294, Dec. 2010.

[16] R. Atia and N. Yamada, "Sizing and Analysis of Renewable Energy and Battery Systems in Residential Microgrids," in *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1204-1213, May 2016.

[17] B. P. Hayes, J. K. Gruber and M. Prodanovic, "A Closed-Loop State Estimation Tool for MV Network Monitoring and Operation," in *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 2116-2125, July 2015.

[18] Pankratz, A. "Introduction to Box-Jenkins Analysis of a Single Data Series." In *Forecasting with*

Univariate Box-Jenkins Models, pp. 24-44. John Wiley & Sons, Inc., 2008.

[19] M. Y. Cho, J. C. Hwang and C. S. Chen, "Customer short term load forecasting by using ARIMA transfer function model," *Proceedings 1995 International Conference on Energy Management and Power Delivery EMPD '95*, Singapore, 1995, pp. 317-322 vol.1.

[20] K. Nose-Filho, A. D. P. Lotufo and C. R. Minussi, "Short-Term Multinodal Load Forecasting Using a Modified General Regression Neural Network," in *IEEE Transactions on Power Delivery*, vol. 26, no. 4, pp. 2862-2869, Oct. 2011.

[21] N. I. Sapankevych and R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey," in *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24-38, May 2009.

[22] A. Khosravi, S. Nahavandi, D. Creighton and D. Srinivasan, "Interval Type-2 Fuzzy Logic Systems for Load Forecasting: A Comparative Study," in *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1274-1282, Aug. 2012.

[23] M. Hanmandlu and B. K. Chauhan, "Load Forecasting Using Hybrid Models," in *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 20-29, Feb. 2011.

[24]] C. Kuster, Y. Rezgui, and M. Mourshed, "Electrical load forecasting models: A critical systematic review," *Sustainable cities and society*, vol. 35, pp. 257–270, 2017.

[25] C. Yu, P. Mirowski and T. K. Ho, "A Sparse Coding Approach to Household Electricity Demand Forecasting in Smart Grids," in *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 738-748, March 2017.

[26] P. Goncalves Da Silva, D. Ilić and S. Karnouskos, "The Impact of Smart Grid Prosumer Grouping on Forecasting Accuracy and Its Benefits for Local Electricity Market Trading," in *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 402-410, Jan. 2014.

[27] R. Sevlian and R. Rajagopal, "A scaling law for short term load forecasting on varying levels of

aggregation," *Int. J. Electr. Power Energy Syst.*, vol. 98, pp. 350–361, Jun. 2018.

[28] Zufferey, T., Ulbig, A., Koch, S. and Hug, G., "Forecasting of Smart Meter Time Series Based on Neural Networks," in *International Workshop on Data Analytics for Renewable Energy Integration*, Riva del Garda, Italy, 2016, pp. 10–21.

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

[30] P. Chen, Z. Chen and B. Bak-Jensen, "Probabilistic load flow: A review," *2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*, Nanjing, 2008, pp. 1586-1591.

[31] B. Wang, Y. Li and J. Watada, "Supply Reliability and Generation Cost Analysis Due to Load Forecast Uncertainty in Unit Commitment Problems," in *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2242-2252, Aug. 2013.

[32] K. Chandrasekaran and S. P. Simon, "Reserve management in bilateral power market for composite system with load forecast uncertainty," *2011 International Conference on Recent Advancements in Electrical, Electronics and Control Engineering*, Sivakasi 2011, pp. 5–12.

[33] S. Ben Taieb, R. Huser, R. J. Hyndman and M. G. Genton, "Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression," in *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2448-2455, Sept. 2016.

[34] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 730–737, Mar. 2017.

[35] S. Haben and G. Giasemidis, "A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting," *International Journal of Forecasting*, vol. 32,

no. 3, pp. 1017–1022, Jul. 2016.

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

[37] J. Xie and T. Hong, "Temperature Scenario Generation for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1680–1687, May 2018.

[38] D. Montgomery, "Introduction to linear regression analysis," *Journal of Royal Statistical Society*, vol. 170, no. 3, pp. 856–857, Apr. 2007.

[39] R. Koenker and K. Hallock, "Quantile regression: An introduction," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 43–56, 2001.

[40] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *International Journal of Forecasting*, vol. 31, no. 2, pp. 286–295, 2015.

[41] B. L. Cabrera and F. Schulz, "Forecasting Generalized Quantiles of Electricity Demand: A Functional Data Approach," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 127–136, Jan. 2017.

[42] H. L. Shang, "A survey of functional principal component analysis," *AStA Advances in Statistical Analysis*, vol. 98, no. 2. Springer Verlag, pp. 121–142, 2014.

[43] M. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[44] J. Martin, D. D. R. de Adana, and A. G. Asuero, "Fitting Models to Data: Residual Analysis, a Primer," in *Uncertainty Quantification and Model Calibration*, Rijeca, Croatia: InTech, 2017, pp. 133-173.

[45] J. Xie, T. Hong, T. Laing, and C. Kang, "On Normality Assumption in Residual Simulation for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1046–1053,

May 2017.

[46] Y. Wang, Q. Chen, N. Zhang and Y. Wang, "Conditional Residual Modeling for Probabilistic Load Forecasting," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7327–7330, Nov. 2018.

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

[48] Y. Yang, S. Li, W. Li, and M. Qu, "Power load probability density forecasting using Gaussian process quantile regression," *Applied Energy*, vol. 213, pp. 499–509, Mar. 2018.

[49] P. Kou and F. Gao, "A sparse heteroscedastic model for the probabilistic load forecasting in energy-intensive enterprises," *International Journal of Electrical Power Energy System*, vol. 55, pp. 144–154, 2014.

[50] N. Altman and C. Léger, "Bandwidth selection for kernel distribution function estimation," *Journal of Statistical Planning and Inference*, vol. 46, no. 2, pp. 195–214, Aug. 1995.

[51] E. Eirola and A. Lendasse, "Gaussian mixture models for time series modelling, forecasting, and interpolation," *International Symposium on Intelligent Data Analysis*, London, UK., 2013, pp. 162–173.

[52] T. Huang, H. Peng, and K. Zhang, "Model Selection for Gaussian Mixture Models," *Statistica Sinica*, vol. 27, no. 1, pp. 147-169, Jan. 2017.

[53] S. J. Gershman and D. M. Blei, "A Tutorial on Bayesian Nonparametric Models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1-12, Feb. 2012.

[54] P. Müller and R. Mitra, "Bayesian nonparametric models," in *Nonparametric Bayesian Inference in Biostatistics*, Springer International Publishing, 2015, pp. 3-24.

[55] B. A. Frigyik, A. Kapila, M. R. Gupta, and U. W. Uw, “Introduction to the Dirichlet Distribution and Related Processes,” Dept. Electrical Engineering, Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2010-0006, Dec. 2010.

[56] T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, vol. 1, no. 196, pp. 209–230, 1973.

[57] D. J. Aldous, “Exchangability and Related Topics.” in *École d'Été de Probabilités de Saint-Flour XIII-1983*, Berlin, Germany: Springer, 1985, pp. 1–198.

[58] I. Yildirim, “Bayesian Inference: Gibbs Sampling,” Dept. Brain Cogn. Sci., Univ. Rochester, Rochester, NY, USA, Tech. Rep., 2012.

[59] D. Görür and C. E. Rasmussen, “Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution,” *Journal of Computer Science Technology*, vol. 25, no. 4, pp. 615–626, Jul. 2010.

[60] R. Problem Author, J. S. Liu, and J. S. L, “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem,” J. Amer. Statist. Assoc., vol. 89, no. 427, pp. 958–966, 1994.

[61] K. P. Murphy, “Conjugate Bayesian Analysis of the Gaussian Distribution,” *Def*, vol. 1, no. 7, pp. 1–29, Oct. 2007.

[62] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, “An introduction to MCMC for machine learning,” Machine Learning, vol. 50, no. 1, pp. 5–43, Jan. 2003.

[63] C. Zhang and Y. Ma, “Ensemble Machine Learning: Methods and Applications,” New York, NY, USA: Springer, 2012.

[64] W. Fan, E. Greengrass, J. McCloskey, P. S. Yu, and K. Drummey, “Effective estimation of posterior probabilities: Explaining the accuracy of randomized decision tree approaches,” in

Proceedings – IEEE International Conference on Data Mining, ICDM, 2005, pp. 154–161.

[65] J. Abellán and A. R. Masegosa, "Bagging decision trees on data sets with classification noise," in *Proc. 6th Int. Conf. Found. Inf. Knowl. Syst.*, Sofia, Bulgaria, Feb. 2010, pp. 248–265.

[66] SmartMeter Energy Consumption Data in London Households [Online]. Available: https://data.london.gov.uk/data set/smartmeter-energy-use-data-in-london-households.

[67] OpenWeather [Online]. Available: home.openweathermap.org.

[68] Government of Canada [Online]. Available: https://www.canada.ca/en/services/environment/weather.html.

[69] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," J. R. Statist. Soc. B, vol. 69, pp. 243–268, Apr. 2007.