COMBINING SISTER LOAD FORECASTS

by

Jiali Liu

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2015

Approved by:

____
Dr. Tao Hong

____
Dr. Badrul Chowdhury

____
Dr. Churlzu Lim

ABSTRACT

JIALI LIU. Combining sister load forecasts. (Under the direction of DR. TAO HONG)

Combining forecasts is a well-known approach to further improving forecast accuracy. In the load forecasting literature, there are only few papers discussing load forecast combination. Most of them are on combining independent forecasts. However, in practice, load forecasters may be able to concentrate on only one or few particular forecasting techniques due to limitations in educational background, time for model development, costs of additional software and so forth. How to conduct forecast combination with these real-world constraints is a challenging problem.

This thesis proposes a novel solution to the aforementioned problem by combining sister load forecasts, which are generated from a family of sister models sharing very similar model structure developed from similar variable selection processes. In this thesis, 13 forecast combination methods are tested on four sister models. Through a comprehensive case study using publicly available data from the Global Energy Forecasting Competition 2014, combining sister forecasts using simple methods is found to outperform each individual forecast. In addition, the regression based combination, which uses a regression model to combine sister forecasts, outperforms the other methods for the aforementioned data set. Comparing with the best individual model, the regression-based combination reduces the forecast mean absolute percentage error (MAPE) by approximately 9%. It also outperforms simple average by 11 %. Note that simple average may not always outperform the best individual forecast, which is shown in this test case.

This thesis starts the first formal investigation on combining sister forecasts, which is shown to be very effective in improving load forecasting accuracy of individual models. The proposed approach is of great practical value in the sense that it leverages existing variable selection processes and does not require additional skill sets from the load forecaster.

ACKNOWLEDGMENTS

First, I want to express my appreciation to Dr. Tao Hong, who helped me explore knowledge and skills in the area of load forecasting. Due to his supervision, I am more confident in facing the challenges of forecasting in the real world. I also want to thank my committee members, Dr. Badrul Chowdhury and Dr. Churlzu Lim, for generously offering their time and guidance through my preparation of this thesis.

The members of Big Data Energy Analytics Laboratory; Bidong Liu, Rain Xie, Lili Zhang, Xiaoqian Lyu, and Jakub Nowotarski have given me many useful suggestions to help me revise this thesis. I especially thank Bidong, who worked with me on projects of load forecasting from the very beginning and always shared his insights with me generously.

I also appreciate the company of my boyfriend, Stephen Howard. The spiritual support and encouragement he gave me helped me through my two-year study.

Last, but most importantly, I want to devote my deepest appreciation and my work to my parents, Rungang Liu and Kunling Deng. Without their unconditional love and financial support, I couldn't have a fruitful academic experience in the U.S.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Many departments of a utility company require load forecasts as an important input to the business decision making processes. For instance, the transmission and distribution planning department needs load forecasts for a city or region to develop plans for building, maintaining and upgrading the infrastructure, such as substations, transmission lines and distribution feeders. The operations and maintenance department needs to know the load forecasts to arrange its schedule of the equipment maintenance in the territory. The finance department also requires load forecasts in order to make revenue projections in the coming years and prepare for rate cases (Gross & Galiana, 1987; Weron, 2006; Hong & Shahidehpour, 2015).

While load forecasting is a critical function since the inception of the electric power industry, the rapid growth and modernization of the electric power industry has been challenging the conventional load forecasting practices. In the late 19th century, light bulb was the only end use of electricity, when access to this convenient form of power was not available to most people. Today the situation has changed with more than 5 billion people around the world having easy access to electricity. Many basic needs are being covered by electricity, such as space heating and cooling, cooking, and cleaning. In recent few decades electricity is also being used in greater proportions for entertainment as people use more power hunger devices such as computers, phones, and televisions. Deployment of smart grid technologies since late 2000s has made electricity consumption easier to monitor and enabled more effective demand side management than before. The diverse and active demand patterns make the load forecasting problem increasingly interesting and challenging (Hong, 2014).

Based on the forecasting horizon, load forecasting can be roughly classified into short-term load forecasting (STLF) (Hong, 2010) and long-term load forecasting (See e.g. Hong, Wilson, & Xie, 2014; Xie, Hong, & Stroud, 2015) with two weeks as the cutoff point. The rationale is based on the accuracy of weather forecasts. Electricity demand is highly dependent upon weather but today's weather forecasting technologies can only provide reliable forecasts within a two week period (Hong & Shahidehpour, 2015). An alternative and more detailed classification is to use one day, two weeks, and three years as cutoff points for very short, short, medium, and long term forecasting (Hong, 2010). This thesis is devoted to STLF, more specifically, 24-hour ahead hourly load forecasting.

Gross & Galiana (1987) gave a summary of the utility applications of STLF:

- Drive the scheduling functions that determine the most economic commitment of generation sources, consistent with reliability requirements, operational constraints and policies, and physical, environmental, and equipment limitations.
- Provide a predictive assessment of the power system security.
- Provide system dispatchers with timely information.

The remaining content in this thesis is organized as follows: Chapter 2 reviews literature of basic forecasting methods and combination methods. Chapter 3 discusses 13 forecast combination methods. Chapter 4 introduces the experiment and analyses the results. Chapter 5 concludes the thesis with recommendations for the future work.

CHAPTER 2: LITERATURE REVIEW

This literature review is conducted from the aspects of STLF and forecast combination. Section 2.1 reviews the techniques and methodologies used in STLF. Section 2.2 reviews the forecast combination methods reported in the forecasting community and the ones specifically used in load forecasting. Section 2.3 highlights the contributions of this thesis.

2.1 Short-Term Load Forecasting

2.1.1 Load Forecasting Reviews

Over the past three decades, there are several notable literature reviews on STLF. They summarized the application and influence of STLF, and introduced different types of STLF techniques.

Gross and Galiana (1987) summarized the application of STLF in details. The authors addressed the four major influence categories of the system load behavior which were economic, time, weather and random effects. They also explained that the basic models for STLF were peak load models and load shape models.

Moghram and Rahman (1989) presented a comparison analysis of five STLF techniques, which were multiple linear regression, stochastic time series, general exponential smoothing, state space and Kalman filter, and knowledge-based. The authors offered some insights about the future development of these.

Hippert, Pedreira, and Souza (2001) reviewed the literature of using neural networks for STLF, with the aim of clarifying the advantage of artificial neural networks and its misuses. They pointed out that many models in the literatures were over-parameterized.

Metaxiotis, Kagiannas, Askounis, and Psarras (2003) had a broad coverage on the development of artificial intelligence technologies (AI) for STLF. The paper summarized the research of expert systems (ESs), artificial neural networks (ANNs), and genetic algorithms (GAs), and analyzed how these techniques were applied in the electric power industry.

Weron (2006) discussed the factors that affected load patterns, such as time factors and weather conditions. It also explained statistical forecasting methods in details with case studies. The methods included similar-day method, exponential smoothing, regression methods, autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and so on.

Taylor and McSharry (2007) tested univariate short-term forecasting methods with 10 time series of intraday electricity demand. They concluded that the best performing method was double seasonal exponential smoothing models, with MAPE approximately 0.1% to 1.5% lower than other forecasting methods based on the differences of forecasting horizon.

Hong (2010) had a comprehensive and critical review of the literature of STLF covering most representative papers prior to 2010. The review included notable literature reviews, papers on statistical and artificial intelligence techniques, and usage of weather and calendar variables.

Hong, Pinson, and Fan (2014) summarized the forecasting methods used by top entries in the Global Energy Forecasting Competition 2012, of which one track was on hierarchical load forecasting. The paper covered the background, problem, data, results, and lessons learned from the competition. The complete dataset of the competition was published in order to encourage reproducible research for energy forecasting.

2.1.2 Load Forecasting Techniques and Methodologies

Most load forecasting techniques fall into one of the two categories, statistical techniques, such as linear regression models, semi-parametric models, autoregressive and moving average (ARMA) models, exponential smoothing models, and artificial intelligence techniques, such as artificial neural networks (ANN), fuzzy regression models, support vector machines (SVMs). Table 2.1 lists the representative techniques and key references for STLF.

Most of the papers in the literatures only discussed load forecasting techniques, such as regression analysis, time series analysis and ANN. Few of them had methodological breakthroughs. Here "methodology" is used to refer to the framework that can be applicable across various techniques. For instance, Hong (2010) proposed a methodology of selecting temperature variables and their interactions with calendar variables for STLF. The methodology was demonstrated to be applicable for three different load forecasting techniques, linear regression model, ANN, and fuzzy regression.

Another emerging area that needs research on fundamental methodology is hierarchical load forecasting, which means forecasting load of different levels in the system. Hong (2008) applied a hierarchical reconciliation method to spatial load forecasting. Fan, Methaprayoon, and Lee (2009) proposed a multiregion load forecasting model which could find optimal region partition with different weather conditions and load conditions. Lai and Hong (2013) proposed a grouping strategy to enhance forecasting accuracy at aggregated levels. The most recent major development in hierarchical load forecasting is from the Global Energy Forecasting Competition 2012 (Hong, Pinson, et al., 2014).

Table 2.1: Representative techniques and references for STLF.

	Techniques	References
Statistical Techniques	Regression analysis	Papalexopoulos and Hesterberg (1990); Hong (2010); Hong, Wang, and Willis (2011); Fan and Hyndman (2012);
	Box-Jenkins approach	Hagan and Behr (1987); Amjady (2001); Weron (2007);
	Exponential Smoothing	Taylor and McSharry (2007);
Artificial Intelligence	ANN	Hippert, Pedreira, and Souza (2001); Metaxiotis et al. (2003); Hippert and Pedreira (2004);
	Fuzzy Systems	Srinivasan, Tan, Chang, and Chan (1998) ; Hong and Wang (2014);
	Support Vector Regression	Chen, Chang, and Lin (2004);

2.2 Forecast Combination

2.2.1 Forecast Combination

Many combination methods have been developed and studied in statistical and economic forecasting for decades. However, load forecasting combination is still an underdeveloped area. Load forecasters now have an opportunity to leverage the forecast combination literature to improve load forecasts.

Granger and Ramanathan (1984) discussed three alternative approaches of linear combination of forecasts. The authors analyzed the testing results of the three approaches and stated that the combination weights of forecasts should be constrained but didn't have to sum up to unity.

Palm and Zellner (1992) introduced Bayesian methods and non-Bayesian methods of combining forecasts. The paper summarized the advantages of using simple average to combine forecasts. One is that no weights of the forecasts need to be estimated. Another one is that the bias from individual forecasts could be averaged out.

Winkler and Clemen (1992) brought the point that combining forecasts could reduce the influence of extremely bad forecast. However, if the combination method was very sensitive and the weights fluctuated widely, the benefit of reducing risks by combining forecasts could be offset by the risk associated with the possibility of extreme weights, such as negative weights or weights larger than one.

Batchelor and Dua (1995) provided two measurements of benefit from combining forecasts. The first one was percentage reduction in expected error variance. The second one was probability of a reduction in the error variance. These two measurements could tell the forecasters the amount of forecasts to combine by keeping certain percentage reduction in expected error variance or certain probability of a reduction in the error variance. The paper also argued that combining forecasts from diverse forecasting methods could be more fruitful than combining forecasts from similar ones due to less positive correlations among errors of diverse forecasting methods.

de Menezes, W. Bunn, and Taylor (2000) summarized seven combination methods from other papers and the criteria of how to combine forecasts, including error variance, distribution asymmetry, and serial correlation.

Armstrong (2001) summarized many issues related to combining forecasts. Armstrong stated that combining forecasts from different forecasting methods could improve accuracy because the methods could capture different components which affected forecasting. The more diverse the methods were, the larger the error reductions were. This book chapter suggested that it was better to combine at least five forecasts, but the rate of improvement was diminishing as more forecasts were included. The book chapter also mentioned that the combination procedure should be formal and fully described, and judgmental weights for different forecasts should be avoided. Forecasts should be given equal weights to combine unless there was strong evidence to support that some forecasts were better than the rest. To reduce the effects of large errors, the paper also suggested using trimmed mean to combine forecasts.

Stock and Watson (2004) used economy data from seven countries to test five types of combination forecast methods: simple combination forecasts, discounted mean squared forecast error (MSFE) forecasts, shrinkage forecasts, factor model forecasts, and time-varying-parameter combination forecasts. They used different techniques to apply historical information on combination weights to different combination methods. Simple combination forecasts include mean, trimmed mean, and median. The authors used autoregression forecasts as a benchmark, concluding that combining forecasts increased the accuracy substantially. They also found that the combination method with the least data adaptivity in the weighting scheme performed the best.

Jose and Winkler (2008) took one step further in their investigations of trimmed mean, Winsorized mean, and median. Comparing to mean's sensitivity of extreme values, trimmed mean, Winsorized mean, and median were more robust and more efficient. The paper used family L-estimators as location estimators of the amount of data to trim or winsorize, and uses symmetric mean absolute percentage error (sMAPE) as the measurement of combined methods' performances. From its experimental results, the paper recommended 10% - 30% of data to trim and 15% - 45% of data to winsorize. Trimmed mean and Winsorized mean worked better if the individual forecasts had more variability.

Montgomery, Jennings, and Kulahci (2011) used an equation which had been widely applied to calculate combination weights to combine two individual forecasts. The variances of two forecasts and the correlation of two individual forecast errors were used in the calculation. It had the conclusion that large negative correlation between the two individual forecasts would lead to a variance of the two forecasts which approached zero, leading to a more accurate result. However, Bates and Granger (1969) had a different opinion on this equation: the variance was fixed by using the equation in Montgomery, Jennings, and Kulahci (2011), therefore the combining method couldn't capture the change of errors over time.

Bates and Granger (1969) provided five combining methods. The methods all had the form of weighted average and the weights could change over time. The calculation of weights was based on errors in previous hours. Besides, Montgomery, Jennings, and Kulahci (2011) and Bates and Granger (1969) both explained how to combine two individual methods in detail, but neither of them discussed how to combine more than two forecasts at the same time.

Pesaran and Pick (2011) investigated the combination of the same model's forecasts over different estimation window. The models were linear regression models. The average of forecasts over different estimation windows led to a lower root mean square forecast error, comparing to the one with a single estimation window. Exponential smoothing was also used for comparison, which was too sensitive to the down-weighting parameter than the proposed model.

2.2.2 Load Forecast Combination

Smith (1989) mentioned an equation of calculating weighting vectors by using variance covariance matrix. The writer stated that each forecasting method has its weakness, while combining different methods could make up each one's weakness. This paper didn't provide any specific examples of how to use the combination equation nor prove the efficiency of its combined forecast equation.

To make good use of temperature forecasts from multiple weather stations, Fan, Chen, and Lee (2009) used forecast combination with adaptive coefficients to improve accuracy of forecasting. Weighted average was used in temperature forecast combination and the weights were calculated based on the last observation's performance. The results showed that the ANN model with combined temperature forecasts had the lowest MAPE than the ones with individual temperature forecasts. Later on, Hong, Wang, and White (2015) proposed a method to select weather stations.

To reduce the effect of weather forecasting error in STLF, Fay and Ringwood (2010) proposed a forecast combination algorithm which used weighted average to combine four models. The combination weight was calculated with error covariance matrices. The combined forecast showed improvement of forecasting accuracy. Its MAPE was 2.17% while the lowest MAPE of the four individual models was 2.27%.

Wang, Zhu, Zhang, and Lu (2010) proposed a linear combination method to combine the forecasts from three time series models (ARIMA, exponential smoothing model, and SVM). The combination weight was decided in an optimization process using adaptive particle swarm optimization (APSO). The combination results showed that the combined forecast outperformed the forecasts from individual models.

Taylor (2012) used simple average to combine individual forecasts from six load forecasting models. The results showed that the combined forecast outperformed individual methods by approximately 0.5% to 1% based on the differences of forecasting horizon.

Matijaš, Suykens, and Krajcar (2013) proposed a meta-learning system for multivariate time-series forecasting which included algorithm selection process. The system analyzed 7 load forecasting algorithms. The ranks of algorithms' performances were associated with eulidean distance, classification and regression (CART) decision tree, learning vector quantization (LVQ) network, multilayer perceptron (MLP), automatic multiplayer perceptron (AutoMLP), ε support vector machine (ε-SVM), and Gaussian process (GP). The best-performed algorithm would be selected among them.

2.3 Contribution of This Thesis

Forecast combination has been well studied in academia, but the number of literature related to load forecast combination is limited, of which most combined forecasts from independent individual forecasting models. However, in practice, when forecasters or research groups plan to conduct load forecasting, it is common that educational background, model development time, and costs of additional software may constrain their studies.

This thesis will investigate and present a solution to the aforementioned challenge by combining sister load forecasts. Sister load forecasts are the load forecasts generated from a family of sister models sharing very similar model structure which are developed from similar variable selection processes. Combining sister forecasts can improve the accuracy of individual sister forecasts without requiring load forecaster to have additional skill sets or software investment. This thesis is the first formal study on combining sister load forecast.

CHAPTER 3: FORECAST COMBINATION METHTDOLOGIES

This chapter will provide the theoretical background of the 13 forecast combination methods and explain how each method works. In Chapter 4, the experiment and comparison results of combining sister forecasts using the 13 methods will be presented.

3.1 Overview of the Chapter

According to Armstrong (2001), combining forecasts means the averaging of independent forecasts. These forecasts can be based on different methods or different data, and the rule of averaging can be replicated. The goal of combining forecasts is to reduce the effects of errors from different forecasting methods so the combined forecast is more accurate than using the individual forecasts.

As mentioned in Chapter 2, many combination methods proposed in research papers can only combine two forecasts at one time. In the real world, it is impossible to test all of the combinations of all forecasting models every time to find the best pair to be combined in the future. The process will be very time-consuming and the forecasting results may not be accurate. To reiterate, there is very limited literature that discusses load forecast combination. In practice, most forecasters only use simple averaging when applying combinations. This thesis will provide a reference for forecasters with regards to load forecast combination beyond just a simple average.

Combination methods that will be tested have been separated into 5 groups by their characteristics. These groups are as follows (also, see Figure 3.1): simple average and median, trimmed mean and Winsorized mean, group analysis, rolling combination, and regression combination. MAPE is used to measure the accuracy of forecast combination. Most of the combination methods have equations to combine forecasts. In these equations, S_i $(i = 1,2, \dots, n)$ is used to represent individual forecasts from one model for each hour and each hour has n forecasts in total (in this thesis n equals to 8, see Chapter 4). FC refers to the combined forecast for each hour.

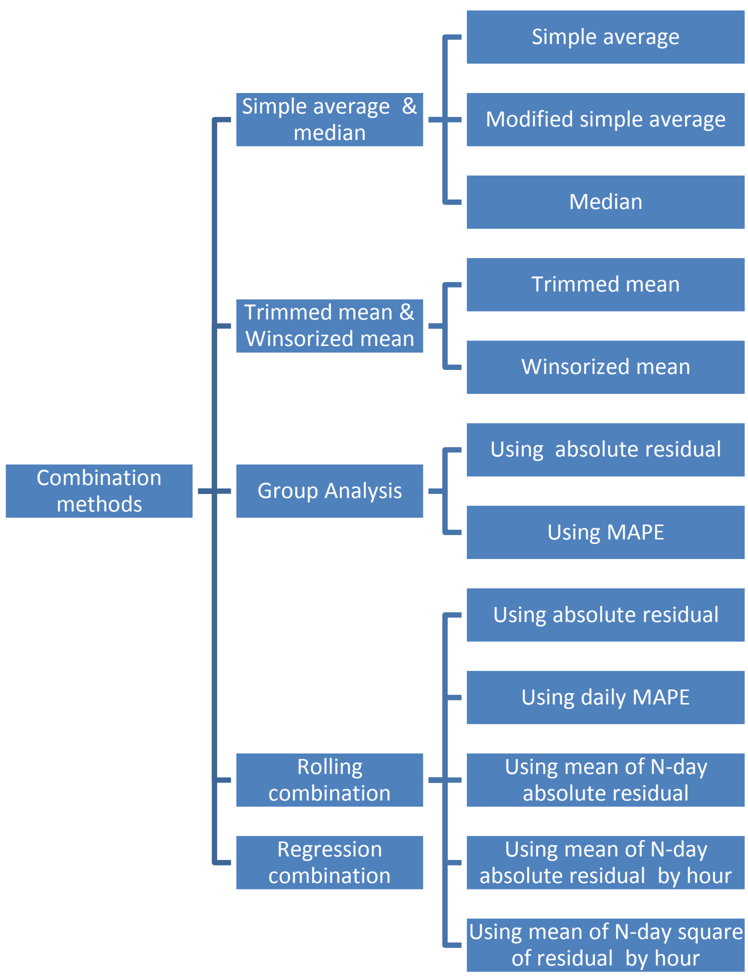

Figure 3.1: Structure of combination methods comparison.

3.2 Simple Average and Median

This section will define the methods used for simple average combination method and median combination method.

3.2.1 Simple Average

When forecasters are uncertain about the weights of different forecasts from different methods, simple average is usually the best choice. Especially when forecasts have close performances to each other, inappropriate weights could lead to less accurate forecasts than using a simple average. The equation for simple average of each hour is given as follows:

$$FC = \frac{1}{n}\sum_{i=1}^{n} S_i \qquad (3.1)$$

3.2.2 Modified Simple Average

In this method, the forecasts from the forecasting model with the largest MAPE are removed and the remaining $n-1$ individual forecasts are averaged for each hour. If one model has larger forecasting MAPE than other models for the validation period, it is reasonable to believe that this model will perform worse than others continuously for the forecasting period and all the forecasts from it should all be removed. The equation for the modified simple average of each hour is given as follows:

$$FC = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} S_i\right) - S_L\right] \qquad (3.3)$$

where S_L is the hourly forecast from Model L which has largest MAPE for validation period.

3.2.3 Median

As mentioned in Jose and Winkler (2008), simple average is sensitive to extreme values, so one alternative can be median. The median of n forecasts for each hour is used as the combined forecast.

3.3 Trimmed Mean and Winsorized Mean

Jose and Winkler (2008) stated that trimmed mean and Winsorized mean can reduce the effect of extreme values on combined forecasts. The accuracy of these combination methods can be slightly better than simple average. Meanwhile, these simple combination methods tend to outperform the complicated ones.

3.3.1 Trimmed Mean

If there are n models and each one of them provide one forecast, the smallest and largest individual forecasts are removed for each hour and the remaining $n-2$ forecasts are averaged afterwards. Using the example in Figure 3.2, for hour 1 the forecasts are sorted. The smallest forecast is S_1 and the largest forecast is $S_{8.}$ In this example these two forecasts are removed and the remaining 6 forecasts are averaged as a trimmed mean. The equation for the trimmed mean model of each hour (Jose & Winkler, 2008) is given as follows:

$$FC = \frac{1}{n-2}\sum_{k=2}^{n-1} S'_k \tag{3.4}$$

where S'_k are sorted forecasts.

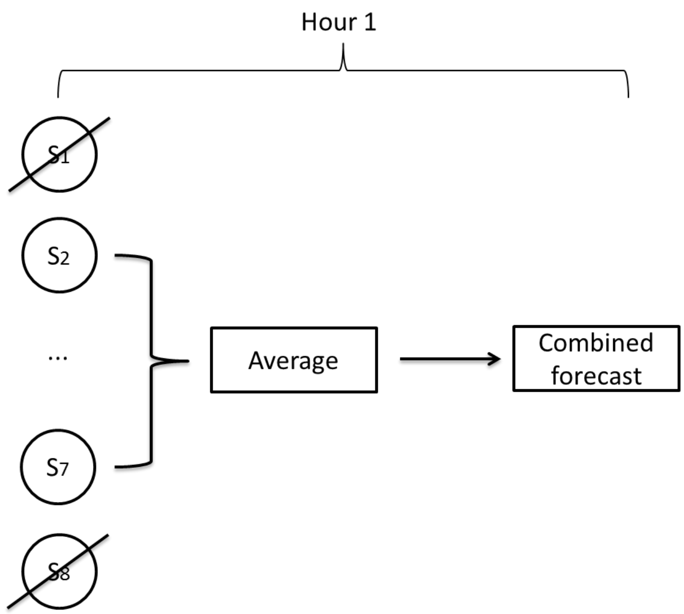

Figure 3.2: Example of trimmed mean. For hour 1 the smallest forecast is S_1 and the largest forecast is S_8. In this example these two forecasts are removed and the remaining 6 forecasts are averaged as a trimmed mean.

3.3.2 Winsorized Mean

Winsorized mean in this thesis is facilitated by replacing the smallest and largest individual forecast with second smallest and second largest forecast, and averaging the data afterwards. In Figure 3.3, the forecasts are sorted. S_1 and S_8 are the smallest and largest one. To get Winsorized mean, they are replaced with S_2 and S_7 which are the second smallest and second largest forecasts, then all the forecasts are averaged for hour 1. The equation for the Winsorized mean model of each hour (Jose & Winkler, 2008) is given as follows:

$$FC = \frac{1}{n}\left[S_2 + \sum_{k=2}^{n-1} S'_k + S_{n-1}\right] \quad (3.5)$$

where S'_k are sorted forecasts.

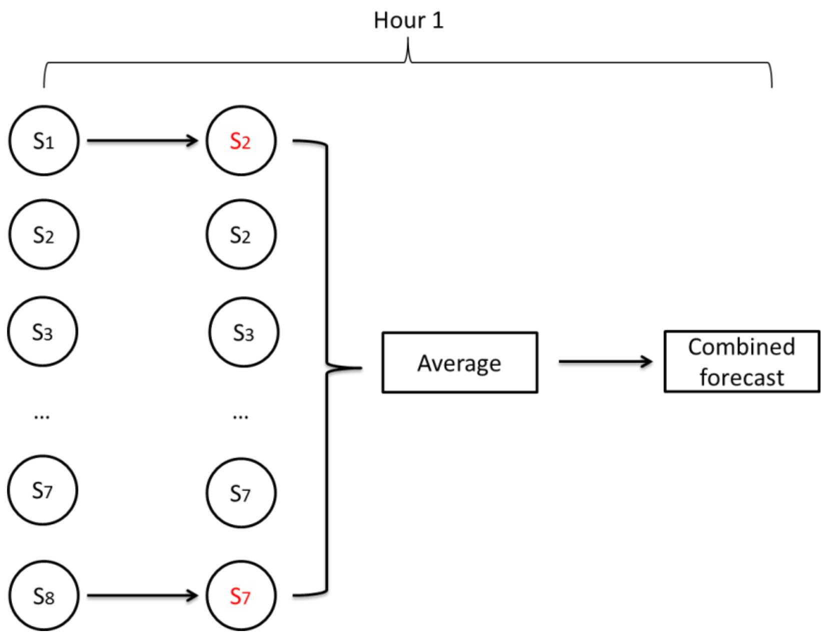

Figure 3.3: Example of Winsorized mean. For hour 1 S_1 and S_8 are the smallest and largest one which are replaced with S_2 and S_7, then all the forecasts are averaged as Winsorized mean.

3.4 Group Analysis

In group analysis, the individual forecasts for each hour are in one group and the forecast with the smallest absolute residual is selected as the combined forecast for that hour. It is more like a forecast selection process than a forecast averaging process because only the best forecast for each hour is selected and the remaining forecasts are not averaged or in use. The load forecasting method in this thesis is 24-hour ahead forecasting so it is impossible to know the residuals or MAPEs of forecasts for the forecasting period. In this method the absolute residuals from previous days are used to estimate combination weights. Using N as the number of days ahead: if N-day ago, one forecast has the smallest residual or daily MAPE for hour 1, it could be used as a reference for hour 1 N-day later by assuming this hour or this day repeats a similar load pattern. Therefore the forecast selected for each hour or each day could be different from its adjacent hours or days in the forecasting period. For the days with no previous residuals or MAPEs to use (like January 1st), simple average is used to combine the forecasts. The structure of combination equation 3.6 is used through section 3.4 and section 3.5 while the ways of calculating weights are different.

3.4.1 Residual Based Binary-Weight Combination

This method will use previous one day's absolute residual (absolute error) to generate the weight for each forecast, therefore N days are tested to check the combination accuracy. In the model below, if N days ago one forecast has the smallest absolute residual for a certain hour, its coefficient α_i for that hour of the forecasting period will be 1, and otherwise it will be 0. This method is trying to pick the most accurate individual forecast as the combined forecast for each hour. The equation for the residual based binary-weight combination model of each hour is given as follows:

$$FC = \sum_{i=1}^{n} \alpha_{i,N} S_i \tag{3.6}$$

where $residual_{min,N}$ denotes the hourly smallest absolute residual and

$$\alpha_{i,N} = \begin{cases} 1, & if\ residual_{i,N} = \min\{residual_{i,N}\} \\ 0, & otherwise \end{cases}.$$

For example, let N=1 so absolute residuals of January 1st are used to judge the weights for the forecasts of January 2nd. In Figure 3.4, R is denoted as absolute residual. The absolute residuals and corresponding forecasts are listed from the smallest to the largest. The smallest forecasting absolute residual is R_2 at 10 a.m. January 1st, so the weight for S_2 at 10 a.m. January 2nd is 1. The weights for the other forecasts are zero.

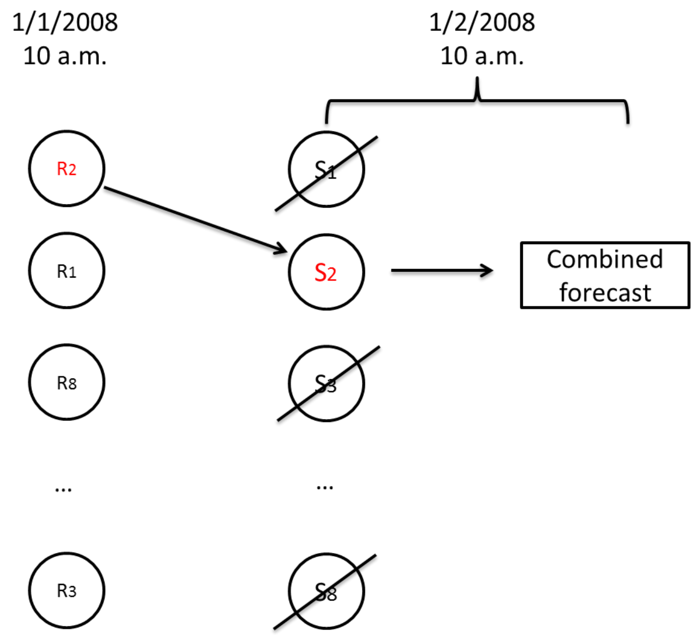

Figure 3.4: Example of residual based binary-weight combination (N=1). The smallest forecasting absolute residual is R_2 at 10 a.m. January 1st so the weight for S_2 at 10 a.m. January 2nd is 1 and the weights for other forecasts are zero.

3.4.2 MAPE Based Binary-Weight Combination

Instead of using absolute residuals, this method uses daily MAPE for each hour to estimate the weight of forecasts. The structure of the equation for the MAPE based binary-weight combination of each hour is the same as equation 3.6, but $MAPE_{min,N}$ denotes the smallest daily MAPE. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \begin{cases} 1, & MAPE_{i,N} = \min\{MAPE_{i,N}\} \\ 0, & otherwise \end{cases}$.

Figure 3.5 shows an example of this method. M is denoted as MAPE. The MAPEs and corresponding forecasts are listed from the smallest to the largest. Since S_2 from January 1st has the lowest MAPE, the forecast from this model will be chosen as combined forecast for January 2nd.

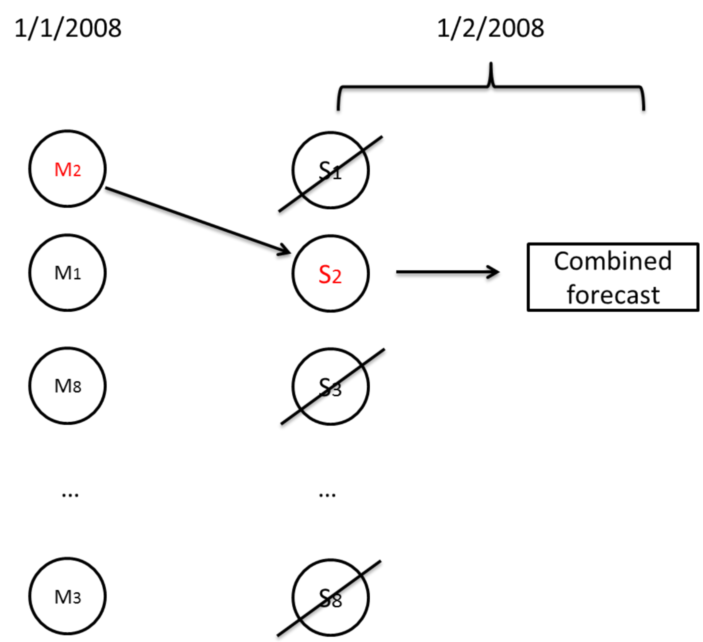

Figure 3.5: Example of MAPE based binary-weight combination (N =1). S_2 from January 1st has the lowest MAPE, the forecast from this model will be chosen as the combined forecast for January 2nd.

3.5 Rolling Combination (RC)

The structure of the combination equations in this section is the same as equation 3.6, but the weights are not binary. Rolling combination method's weight is calculated with residuals or MAPE which are from previous N-day forecasts or the mean of previous N -day's absolute residuals or MAPEs. This method is called rolling combination because the days used to calculate weights change as the date changes. For example, if two-day ago (N=2) residuals are used, weights for the models of January 5th are calculated from the residuals of January 3rd. For January 6th, the models' weights are calculated from the residuals of January 4th. For the days with no previous residuals or MAPEs to use (like January 1st), simple average is used to combine the forecasts.

With a rolling procedure, the weights can be more suitable for the present situation because load patterns are changing over time. If the combination origin is fixed, it is possible that a heavy weight could be given to a model which has accurate forecasts for last month but has bad forecasts for the current month. The rolling combination origin can reduce the errors caused by this type of situation.

All the rolling combination models share the same equation structure from equation 3.6. To different the five rolling combination models in section 3.5.1 to section 3.5.5, each method is given a name.

3.5.1 Absolute Residual (RC-A)

In this method, to avoid negative weights, absolute residuals ($|e_i|$) instead of residuals are used. Negative weights could lead to less accurate forecasts. N-day ago daily mean absolute residuals are used to calculate weights for the individual forecasting models. Let $E_{i,N}$ denotes mean absolute residuals of one day, which equation is $E_{i,N} = \frac{1}{24}\sum_{h=1}^{24}|e_{i,N,h}|$ and $i, j, l,$ and k denotes the order of forecasts. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \frac{\sum_{l}^{n} E_{l,N}}{E_{i,N}\sum_{j=1}^{n}\left(\frac{\sum_{k=1}^{n} E_{k,N}}{E_{j,N}}\right)}$.

Using N=3 as an example. There are 8 models to provide 8 forecasts for each hour. If the forecasting time is January 4th, mean absolute residuals of January 1st are used to calculate weights (see Figure 3.6). For January 5th, the residuals of January 2nd are used to calculate the weights.

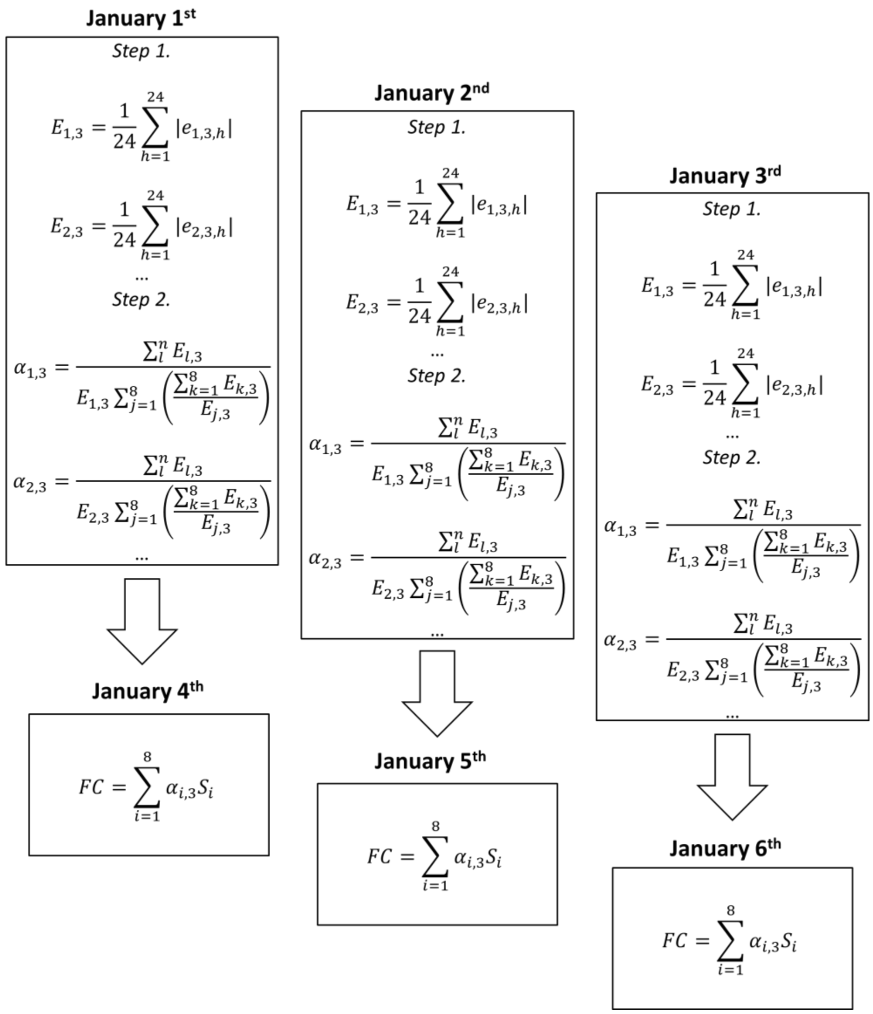

Figure 3.6: Example of rolling combination-A (N=3). To calculate combination weights for January 4th, mean absolute residuals of January 1st are used. Each day's weights are calculated with residuals from different previous days.

3.5.2 MAPE (RC-B)

The only difference between RC-A and RC-B is that the daily MAPE is used to replace the absolute residual. MAPE is a method to measure error, which could capture different characteristics in forecasts than absolute residuals. The MAPEs are from the forecasts N-day ago.

In this method $M_{i,N}$ denotes daily MAPE of one day and $i, j, l,$ and k denotes the order of forecasts. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \frac{\sum_l^n M_{l,N}}{M_{i,N} \sum_{j=1}^n \left(\frac{\sum_{k=1}^n M_{k,N}}{M_{j,N}} \right)}$.

3.5.3 Mean of N-day Absolute Residuals (RC-C)

Different from previous two rolling combination methods, this method uses the mean of preceding N days' residuals to calculate weights for individual forecasts of each hour. The advantage of this method is that it improves stability of weights by using more than one day's forecasting residuals. Let $A_{i,N}$ denotes mean absolute residuals of N days which equation is $A_{i,N} = \frac{1}{24N} \sum_{t=1}^{24N} |e_t|$and $i, j, l,$ and k denotes the order of forecasts. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \frac{\sum_l^n A_{l,N}}{A_{i,N} \sum_{j=1}^n \left(\frac{\sum_{k=1}^n A_{k,N}}{A_{j,N}} \right)}$.

Using the calculation of 8 models' weights on January 3rd as an example: if N=2, the mean absolute residuals from January 2nd and January 1st are used to calculate the weight (see Figure 3.7).

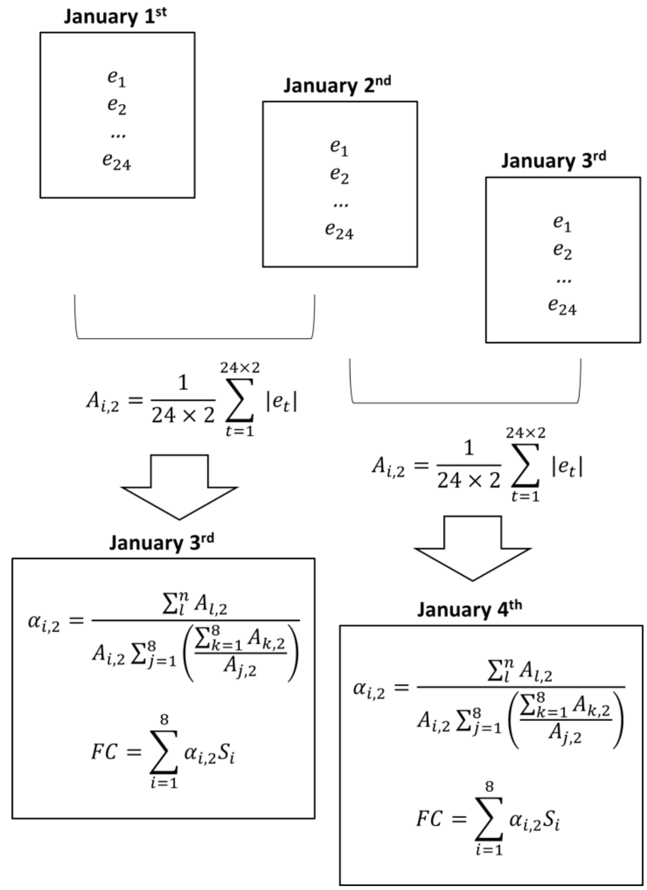

Figure 3.7: Example of using rolling combination model-C (N=2). For January 3rd, the mean of absolute residuals from previous two days is used to calculate weights.

3.5.4 Mean of N-day Absolute Residuals by Hour (RC-D)

This method is an extended version of RC-A. Instead of using one day's mean of absolute residuals to calculate weights, the new method first slices the data into 24 pieces by each hour, and then uses the mean of preceding N hours' absolute residuals to calculate weights for the present hour's forecasts. Let $H_{i,N}$ denotes mean of previous N days' hourly absolute residual which equation is $H_{i,N} = \frac{1}{N}\sum_{t=1}^{N} |e_t|$ and $i, j, l,$ and k denotes the order of forecasts. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \frac{\sum_l^n H_{l,N}}{H_{i,N} \sum_{j=1}^{n} \left(\frac{\sum_{k=1}^{n} H_{k,N}}{H_{j,N}} \right)}$.

For example, to calculate the weights of the 8 individual forecasts at 10 a.m. January 3rd: if N=2, the mean of absolute residuals at 10 a.m. of January 1st and January 2nd are used to calculate the weights (see Figure 3.8).

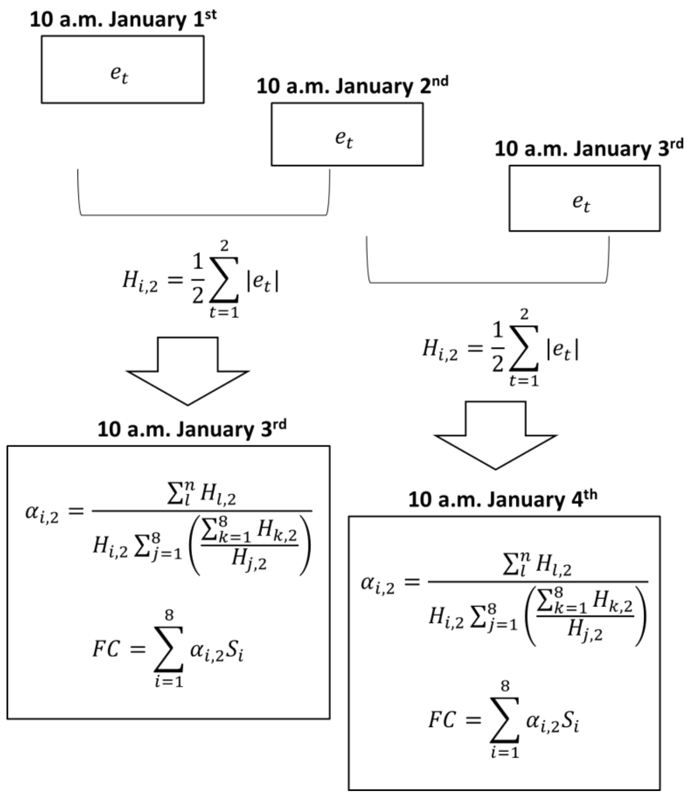

Figure 3.8: Example of rolling combination model-D (N=2). The mean of absolute residuals at 10 a.m. of January 1st and January 2nd are used to calculate the combination weights for the models at 10 p.m. January 3rd.

3.5.5 Mean of N-day Square of Residuals by Hour (RC-E)

RC-E is a modified version of RC-D, which uses the mean of preceding N days' square residuals to calculate weights for individual forecasts of each hour. Let $Q_{i.N}$ denotes mean hourly square residuals of previous N days which equation is $Q_{i.N} =$

$\frac{1}{N}\sum_{t=1}^{N} e_t^2$ and $i, j, l,$ and k denotes the order of forecasts. The weight in equation 3.6 is calculated as $\alpha_{i,N} = \frac{\sum_{l}^{n} Q_{l,N}}{Q_{i,N}\sum_{j=1}^{n}\left(\frac{\sum_{k=1}^{n} Q_{k,N}}{Q_{j,N}}\right)}$.

3.6 Regression Combination

Regression models can be used to calculate weights for combination, instead of using absolute residuals and MAPEs. The advantage of this method is that weights may be more objective and may capture the characteristics which are not involved in residuals and MAPEs. The disadvantage could be its longer processing time than other methods. It uses ordinary least squares to minimize the sum of squares of forecasting errors. To generate accurate results, past N-day actual load and forecasts are used as historical data to estimate parameters. The forecasting has a moving window like the rolling combination method mentioned previously. For the first N days which don't have historical data and combined results, simple average is used to combine the forecasts. For example, if N=10, then the first 10-day actual load and forecasts are used as historical data to estimate parameters for the linear model. Therefore these 10 days don't have combined forecasts and simple average is used to combine them. The equation for the regression model of each hour is given as follows:

$$Load = \beta_0 + \sum_{i=1}^{n} \beta_i S_i \tag{3.7}$$

Where β_0 refers to constant, β_i refers to parameters, and S refers to forecasts.

CHAPTER 4: EXPERIMENT AND RESULTS

This chapter will show the process of generating sister forecasts, combining sister forecasts and analyzing the combination results. The strategy of selecting sister models will be described. A summary and discussion of combination results will be addressed in the last section of this chapter.

4.1 Background

4.1.1 Data Description

In this thesis, the load data is from Global Energy Forecasting Competition 2014. Temperature data for each hour are provided from 25 weather stations (the location of these weather stations is not provided). The arithmetic average of the 25 temperature data is used as the representative temperature for each hour. Hong, Wang et al. (2015) proposed an advanced method to select temperature data to combine, which is weather station selection.

Figure 4.1 shows the relationship of temperature and load from year 2008. The scatter shape takes the form of a check mark or the letter ‘v’. The reason for this pattern is that people tend to consume more electricity in summer and winter for the purpose of cooling and heating. Seasons with mild temperatures (around 50°F) have characteristically lower energy usage.

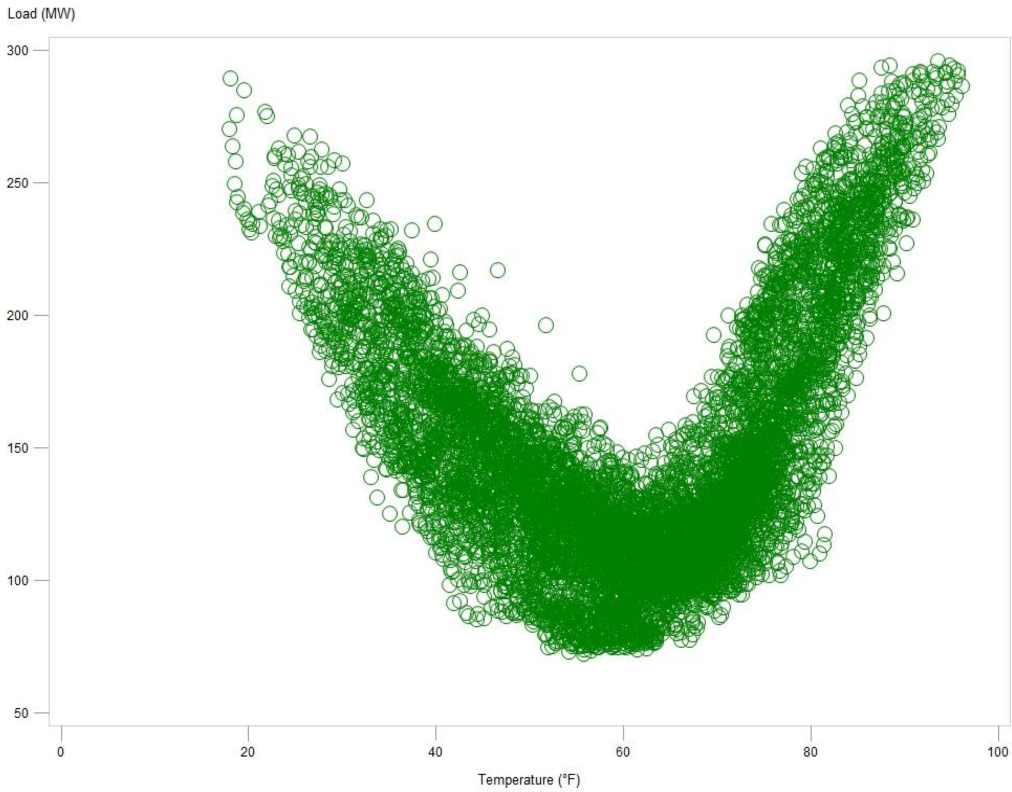

Figure 4.1: Load-temperature relationship of 2008. Summer and winter have higher electricity consumption than spring and autumn.

Figure 4.2 shows consumers of electricity use energy at different rates depending on the time of the day. This verifies the importance of time of day as a variable for load forecasting. Note that while the load vs. temperature curves exhibit the same basic shape, they are shifted depending on the time of day.

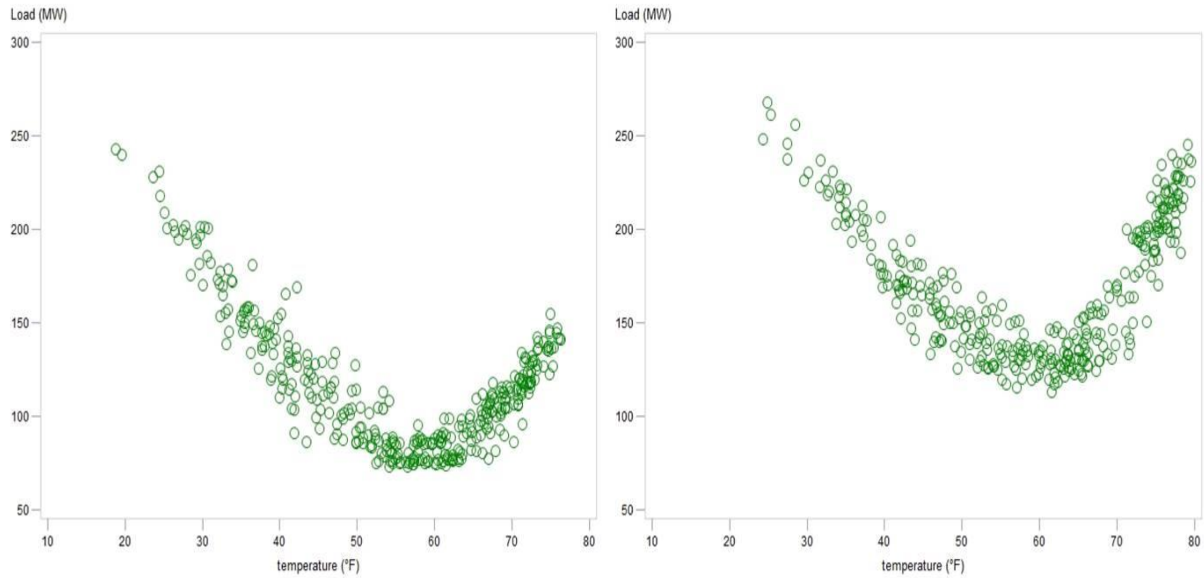

Figure 4.2: Comparison of load patterns (versus temperature) at 4 a.m. (left graph) and 9 p.m. (right graph). Electricity consumption changes due to the different time.

4.1.2 Experiment Process

There are three steps to accomplish the experiment.

1) The first one is to use different selection methods to select sister models. The sister forecasts generated from this step are used for the two sub-steps in step 2.
2) First sub-step is to compare the sister forecasts and out-of-sample sister forecasts to verify the usefulness of sister models. Second sub-step is using the sister forecasts to compare combination methods so that the best one can be selected among them.
3) In step 3, the selected best combination method will be tested with the out-of-sample sister forecasts (see Figure 4.3). This process can ensure the reliability of both sister forecasts and the best combination methods are checked with out-of-sample test.

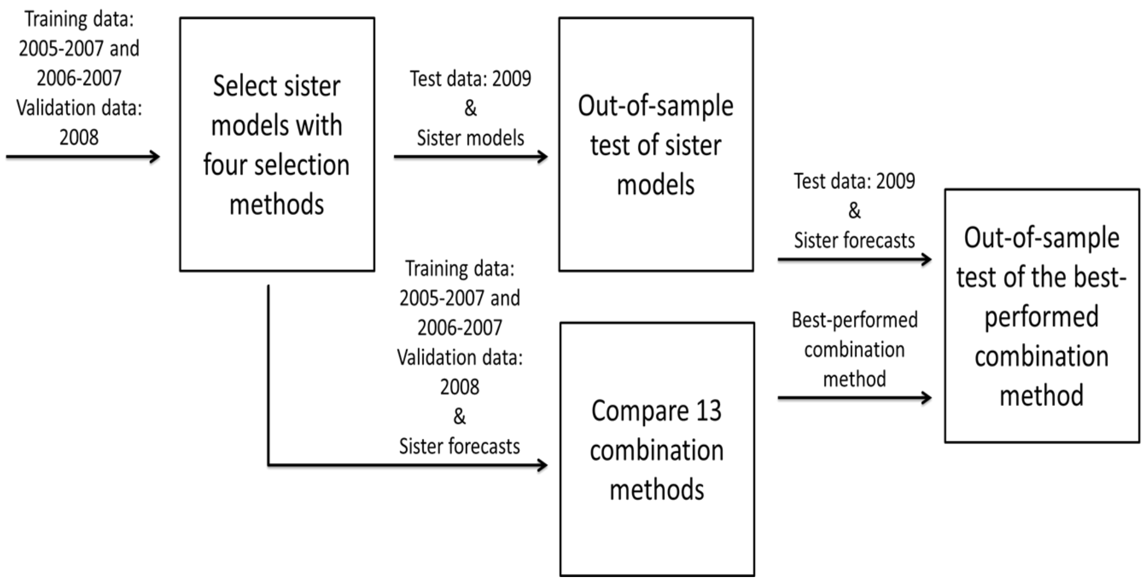

Figure 4.3 Experiment process. The process includes 3 stages to ensure the reliability of sister forecasts and combination methods

4.2 Sister Load Forecast

There are two types of forecasting: ex-ante forecasting and ex-post forecasting. Ex-ante forecasting uses forecast data in the model to forecast load. Ex-post forecasting uses real data to forecast load and it is also called “backcasting”. For example, using previous forecasts of temperatures to forecast tomorrow’s load is ex-ante forecasting. Using yesterday’s temperature data to forecast yesterday’s load is ex-post forecasting. For real-time load forecasting, ex-ante forecasting is used because the actual temperature is unknown and forecasts of temperature must be used. Ex-post forecasting is commonly used to understand modeling error with actual information of the independent variables. In this thesis, ex-post forecasting is used.

4.2.1 Basic Model

The basic model used in this thesis to develop sister models is Tao’s vanilla benchmark model (Hong, Wang, et al., 2015; Hong, Pinson, et al., 2014). It is a multiple linear regression model and was first proposed in Hong (2010). It is also the benchmark model in GEFCom2012 (Hong, Pinson, et al., 2014). Class variables are month, weekday, and hour in the model, which is shown as follows:

$$Load = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t) \tag{4.1}$$

where T denotes as temperature, β denotes the parameter of the linear regression model, and the function $f(T)$ is defined as follows:

$$f(T_t) = \beta_5 T_t + \beta_6 {T_t}^2 + \beta_7 {T_t}^3 + \beta_8 T_t M_t + \beta_9 {T_t}^2 M_t + \beta_{10} {T_t}^3 M_t + \beta_{11} T_t H_t + \beta_{12} {T_t}^2 H_t + \beta_{13} {T_t}^3 H_t \tag{4.2}$$

4.2.2 Forecasting Process

To improve the accuracy of benchmark model, recency effects are added in the model. Recency effect is used to describe the effects of temperatures in preceding hours to the present hours' load forecasts. Lagged temperature and moving average of temperature are the two components used in recency effects. The process of adding recency effects into the benchmark model has been explained in Hong, Liu, and Wang (2015). In this thesis, the difference between the sister models is that they use different rules to select the number of lagged temperature and moving average of temperature to add in the benchmark models.

Year-ahead forecasting is used to test the effect of adding lagged temperature variables and moving average temperature variables in the benchmark model. This type of forecasting uses historical data to forecast one year's load at one time. The lag-average pair (lagged variable - moving average variable pair) with the lowest MAPE under different test scenarios will be chosen. Lagged temperature (lag) is an hourly lagged variable. The interval for the moving average (avg) of temperature is 24 hours.

Day-ahead rolling forecasting is used to validate sister models. The selected lag-average pair will be used on day-ahead load forecasting which uses historical data to forecast the next day's load. The forecasting origin is not fixed so that the forecasting has moving window (Tashman, 2000). For example, Monday to Friday's actual temperature and actual load data are used to forecast Saturday's load, then Tuesday to Saturday's actual temperature and actual load data are used to forecast Sunday's load. This forecasting method is similar to methods used in electric power industry for STLF. The model of adding lagged temperature and moving average of temperature is shown as follows:

$$\text{Load} = \beta_0 + \beta_1 \mathrm{M}_t + \beta_2 \mathrm{W}_t + \beta_3 \mathrm{H}_t + \beta_4 \mathrm{W}_t \mathrm{H}_t + f(T_t) + f\left(\sum_d T_{Avg,d}\right) + f\left(\sum_i T_{Lag,i}\right) \tag{4.3}$$

Where the function $T_{Avg,d}$ is as follows:

$$T_{Avg,d} = \frac{1}{24}\sum_{i=24d-23}^{24d} T_{Lag,i} \tag{4.4}$$

4.2.3 Methodology of Sister Forecasting

Different periods of training data are used to estimate parameters of the models. One period uses data from 2005, 2006, and 2007 as training data, which is denoted as L_1. The other period uses data from 2006 and 2007 as training data, which is denoted as L_2. The data from 2008 is used as validation data. The test data is from 2009. This test data includes hourly load and hourly temperature. The forecasting in this chapter is day-ahead ex-post forecasting. The process of sister load forecasting can be seen in Figure 4.4.

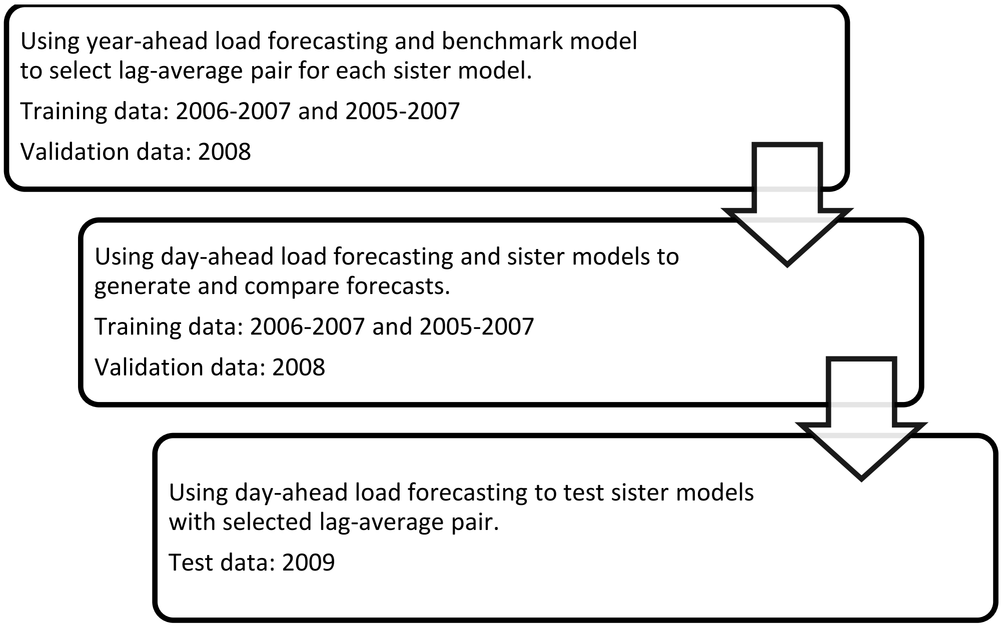

Figure 4.4: Process of sister load forecasting. The stages include selecting lag-average pair for sister models, using sister models to generate and compare forecasts, and testing sister models with new dataset.

Load usage has seasonal and hourly changes. The pattern of electricity consumption of summer is different from the one of winter. People also tend to use more electricity in the daytime and night (when they are most active) than in early morning. Therefore much of the literature in this area slices data into 24 pieces so that each hour has its own parameter estimation of the forecasting model. On the other hand, most of time all of the historical data are not sliced and directly used to estimate the parameters of one model for all the hours. The test scenario of selecting sister models is based on these two data separation strategies and their accuracy of forecasting will be analyzed.

To find the best recency effect, combinations of lag variables and moving average variables are tested. The number of lag variables varies from 0 to 48 and the number of moving averages varies from 0 to 7, so there are 392 combinations in total. Each sister model has its own way to select a lag-average pair (see Table 4.1), which is explained in

Liu, Liu, and Hong (2015). The forecasting procedure is finished by using SAS software and its macro language.

The scenarios of selecting lag-average pair for sister models:

1) Model A uses completed training data to estimate parameters of one model. It tries all the lag-average pairs and uses the one with lowest MAPE as the final model.
2) Model B uses sliced training data and sliced validation data. These data are sliced by hours so that each hour has its own parameter estimation of the model with recency effect. MAPE is calculated on the base of 24 hours and the lag-average pair with lowest MAPE is selected as the final pair for a sister model.
3) Model C uses completed training data and sliced validation data. Completed training data is used to estimate parameters for hourly validation data and this process repeats 24 times. The lag-average pair with lowest MAPE for each hour is selected as the final pair for that hour. There are 24 pairs in total.
4) Model D uses sliced training data and sliced validation data. Hourly training data and hourly validation data are used to estimate parameters. This process repeats 24 times. Instead of calculating MAPE on the base of 24 hours, hourly MAPE is calculated and the lag-average pair with lowest MAPE is selected as the final pair for that hour. There are 24 pairs in total.

Table 4.1: Recency effect modeling of sister models.

model	**Training**	**Validation**	**Lag-average Pair**
A	1	1	1
B	24	24	1
C	1	24	24
D	24	24	24

For Model A and Model B, because they only choose one lag-average pair so every hour has the same pair. Model A-L_1 selects 0 lag variable and 17 moving average variables and Model A-L_2 selects 1 lag variable and 4 moving average variables. Model B-L_1 selects 1 lag variable and 1 moving average variable and Model B-L_2 selects 1 lag variable and 0 moving average variable. Model C and Model D have 24 lag-average pairs for 24 hours so each hour has different pairs, which are shown in Table 4.2.

Table 4.2: Model C and Model D with selected lag-average pair.

Hour	Model C				Model D			
	L_1		L_2		L_1		L_2	
	Lag	**Avg**	**Lag**	**Avg**	**Lag**	**Avg**	**Lag**	**Avg**
1	0	16	1	5	1	0	1	0
2	0	17	1	1	1	1	1	0
3	0	18	1	1	1	1	1	0
4	0	19	1	2	1	0	1	1
5	0	20	1	3	2	1	2	1
6	0	21	1	12	2	2	1	0
7	1	13	1	13	2	1	2	1
8	0	13	1	12	2	1	2	1
9	0	28	1	9	2	1	1	1
10	0	31	1	0	1	1	1	1
11	2	0	1	0	1	0	1	0
12	0	19	1	1	1	0	1	0
13	1	2	1	2	0	2	1	1
14	1	12	1	3	2	0	0	2
15	1	13	1	4	0	0	0	0
16	1	5	1	6	0	0	0	0
17	1	9	1	8	0	1	1	0
18	1	23	1	7	0	1	0	1
19	1	13	1	17	0	2	0	2
20	0	10	1	9	1	3	0	3
21	0	5	1	5	1	0	1	1
22	1	6	1	6	1	0	1	1
23	1	6	1	3	1	0	1	0
24	2	8	1	4	1	0	1	0

4.2.4 Analysis of Forecasting Results

Table 4.3 shows the MAPE of sister forecasts for 2008. Model A with three-year training data (A-L_1) has the lowest MAPE. This result is also tested on the data of 2009 and the conclusion is the same (see Table 4.4). However, the model with lowest MAPE (Lowest MAPE is highlighted in bold) for each hour is different for 2008 and 2009. For instance, for 2008, the model with lowest MAPE for hour 3 to hour 8 is A-L_1 while for 2009 the models with lowest MAPE for the same period of time are B-L_1, D-L_1, D-L_1, B-L_1, D-L_1, and D-L_1, respectively. Forecasting contains uncertainty and the past isn't repeatable, so for different periods of time, the best forecasting model could be different. On the other hand, this experiment also shows a common flaw: separating data by hour always leads to more accurate forecasts. This is of course not always true. In this experiment, Model B and Model D don't outperform Model A.

The sister models perform better than benchmark model. Model A-L_1's MAPE is 4.575%. Comparing to the benchmark model which has a MAPE of 5.641%, Model A-L_1 has reduced MAPE by 23%. The results from using sister models on test data shows that MAPE of Model A-L_1 is 5.013%, which is the lowest one among MAPEs of the eight models.

Table 4.3: MAPE (%) of sister forecasts for 2008.

Hour	**A**		**B**		**C**		**D**	
	L_1	**L_2**	**L_1**	**L_2**	**L_1**	**L_2**	**L_1**	**L_2**
1	**4.229**	4.402	4.268	4.512	4.592	4.492	4.324	4.515
2	**4.419**	4.571	4.457	4.666	4.853	4.675	4.530	4.695
3	**4.533**	4.750	4.647	4.754	4.984	4.728	4.762	4.756
4	**4.560**	4.747	4.735	4.810	5.049	4.701	4.892	4.798
5	**4.482**	4.653	4.740	4.596	4.972	4.581	4.768	4.552
6	**4.411**	4.545	4.770	4.548	4.759	4.496	4.761	4.504
7	**4.755**	4.849	5.195	5.002	4.809	4.778	5.181	4.981
8	**4.065**	4.123	4.675	4.283	4.188	4.119	4.379	4.287
9	3.930	**3.852**	4.432	4.054	3.973	3.901	4.061	4.131
10	4.170	**4.059**	4.347	4.234	4.287	4.067	4.264	4.279
11	4.509	4.461	4.521	4.496	4.468	4.491	**4.347**	4.357
12	4.682	4.653	4.855	4.730	4.736	**4.637**	4.638	4.653
13	**4.808**	4.821	5.107	4.817	4.849	4.895	4.827	5.003
14	4.945	**4.927**	5.253	5.071	4.993	5.043	5.079	4.982
15	5.129	**5.095**	5.327	5.309	5.122	5.128	5.219	5.169
16	5.258	5.232	5.456	5.404	**5.217**	5.331	5.261	5.240
17	5.366	5.346	5.510	5.245	5.203	5.305	5.324	**5.158**
18	5.742	5.566	5.453	5.210	5.566	5.587	**5.022**	5.119
19	5.223	5.000	5.054	5.111	5.017	5.118	**4.665**	4.814
20	4.507	4.509	4.644	4.666	4.604	4.612	**4.442**	4.536
21	**3.879**	4.009	4.109	4.023	4.001	4.052	3.918	4.140
22	3.921	4.031	4.053	4.018	3.928	3.929	**3.886**	4.115
23	**4.018**	4.196	4.060	4.179	4.059	4.049	4.072	4.178
24	4.263	4.442	**4.202**	4.473	4.366	4.408	4.276	4.441
Average MAPE	**4.575**	4.618	4.745	4.675	4.691	4.630	4.621	4.642

Table 4.4: MAPE (%) of sister forecasts for 2009.

Hour	A		B		C		D	
	L_1	**L_2**	**L_1**	**L_2**	**L_1**	**L_2**	**L_1**	**L_2**
1	**4.973**	5.801	5.119	5.133	5.553	5.704	5.126	5.289
2	**5.126**	5.922	5.247	5.280	5.932	5.872	5.255	5.289
3	5.398	6.201	**5.316**	5.511	6.205	6.063	5.325	5.522
4	5.395	6.125	5.273	5.417	6.164	6.023	**5.170**	5.560
5	5.466	6.268	5.194	5.493	6.228	6.144	**5.163**	5.450
6	5.136	5.796	**4.933**	5.231	5.489	5.686	4.936	5.327
7	5.390	5.811	5.201	5.335	5.653	5.929	**5.184**	5.282
8	4.567	4.846	4.715	4.676	5.013	5.046	**4.517**	4.568
9	**4.252**	4.570	4.546	4.449	4.570	4.631	4.286	4.347
10	**4.361**	4.710	4.671	4.677	4.854	4.765	4.652	4.685
11	**4.758**	5.171	5.063	5.071	5.087	5.029	5.062	5.056
12	**4.986**	5.393	5.495	5.336	5.286	5.312	5.494	5.433
13	**5.233**	5.665	5.912	5.784	5.653	5.637	5.703	6.015
14	**5.306**	5.598	5.937	5.818	5.685	5.613	6.138	5.835
15	**5.455**	5.687	5.930	5.975	5.786	5.638	6.633	6.633
16	**5.531**	5.684	5.965	5.808	5.680	5.677	6.714	6.736
17	5.588	5.701	5.941	5.590	5.599	**5.562**	5.923	6.070
18	5.829	5.891	5.717	**5.095**	5.937	5.527	5.611	5.634
19	5.215	5.531	5.283	**4.985**	5.430	5.399	5.421	5.369
20	**4.347**	4.837	4.992	4.538	4.698	4.937	4.836	4.528
21	**4.406**	4.949	4.586	4.553	4.687	5.182	4.558	4.713
22	4.464	5.007	4.493	**4.401**	4.726	4.940	4.496	4.671
23	**4.490**	5.088	4.552	4.558	4.849	5.114	4.565	4.738
24	**4.643**	5.299	4.785	4.747	5.031	5.272	4.793	4.921
Average MAPE	**5.013**	5.481	5.203	5.144	5.408	5.446	5.232	5.320

4.3 Analysis of Forecast Combination Results

This section will analyze the results of combining sister forecasts. In this investigation the MAPE of Model A-L_1 is used as benchmark 1 and the MAPE of a simple average is used as benchmark 2. All of the combination methods in Chapter 3 will be tested on the forecasts of 2008. The one that has the lowest MAPE and also beats the benchmarks is selected as the representative combination method, and will be used on the forecasts of 2009. SAS software and its macro language are used to generate combined forecasts.

4.3.1 Comparison of Simple Average Methods and Median Method

Comparing to sister model A-L_1's MAPE of 4.575%, simple average's MAPE (benchmark 2) has decreased 4%. Modified simple average model has an even lower MAPE than simple average (see Table 4.5). When forecasters are not sure about weighted combination, simple average and its modified models could be helpful to generate basic combined results. The method of using median of the sister forecasts perform better than benchmark 1 (model A-L_1), but doesn't outperform simple average. Its MAPE is 4.440%.

Table 4.5: MAPE (%) comparison of simple average methods and median method.

Method	**MAPE**
Simple average (Benchmark 2)	4.385
Modified simple average	**4.382**
Median	4.440

Simple average also can be used as a benchmark to measure the performance of combination methods used in most literatures. For example, Montgomery et al. (2011)

provides an equation to calculate weight by using correlation and variances of different forecasting methods (see Equation 4.1) which can only combine two forecasts at one time. Table 4.6 shows MAPE's of combining two models at one time using Equation 4.1, for every sister load forecast. Notice that most of the MAPE of combined forecasts are lower than the lowest MAPE of sister models, but it is much higher than using a simple average, which is the most commonly used combination method for load forecasting.

$$Combination\ weight = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{4.5}$$

Table 4.6: MAPE (%) comparison of using Equation 4.1. The darker color means the MAPE is smaller.

Model	A-L_1	B-L_1	C-L_1	D-L_1	A-L_2	B-L_2	C-L_2	D-L_2
A-L_1		4.580	4.604	4.452	4.479	4.484	4.487	4.447
B-L_1			4.536	4.585	4.493	4.615	4.493	4.596
C-L_1				4.484	4.563	4.535	4.563	4.492
D-L_1					**4.415**	4.553	4.419	4.559
A-L_2						4.477	4.548	4.424
B-L_2							4.493	4.575
C-L_2								4.462
D-L_2								

4.3.2 Comparison of Trimmed Mean and Winsorized Mean Methods

The MAPE of Winsorized mean is 4.389% while the MAPE of trimmed mean is 4.440%. Winsorized mean has better performance than trimmed mean, but its MAPE is still higher than benchmark 2.

4.3.3 Comparison of Group Analysis Methods

To get accurate combined forecasts, the effect of using one day ago residuals or MAPE to 28-day ago residuals is tested. The lowest MAPE for the method using absolute residual is 4.583% and the lowest MAPE for the method using MAPE is 4.545% (see Table 4.7). The method using 9 days ago forecasts' MAPE to calculate the weight for present hour is the one outperforms others.

Table 4.7: MAPE (%) comparison between residual based binary-weight method and MAPE based binary-weight method.

***N* days**	**MAPE (residual-base method)**	**MAPE (MAPE-based method)**
1	4.591	4.562
2	4.595	4.562
3	4.597	4.568
4	4.618	4.554
5	4.635	4.566
6	4.600	4.557
7	4.588	4.551
8	4.603	4.564
9	4.608	**4.545**
10	4.622	4.559
11	4.611	4.562
12	4.592	4.561
13	4.602	4.562
14	4.595	4.569
15	4.619	4.560
16	4.611	4.555
17	4.600	4.555
18	4.609	4.560
19	4.611	4.557
20	4.624	4.563
21	4.617	4.565
22	**4.583**	4.572
23	4.592	4.565
24	4.629	4.557
25	4.591	4.556
26	4.611	4.563
27	4.608	4.565
28	4.614	4.560

4.3.4 Comparison of Rolling Combination Methods

Table 4.8 shows the combination results of the five rolling combination methods. Both model A and model B have the same MAPE and they barely beat benchmark 2. The effect of using one day ago absolute residuals or MAPEs to 28-day ago absolute residuals or MAPEs is tested.

Table 4.8: MAPE (%) comparison of five rolling combination methods.

***N* days**	**Model-A**	**Model-B**	**Model-C**	**Model-D**	**Model-E**
1	4.401	4.402	4.408	4.413	4.473
2	4.396	4.397	4.405	4.408	4.460
3	4.392	4.393	4.401	4.400	4.452
4	4.397	4.398	4.403	4.402	4.455
5	4.411	4.413	4.402	4.402	4.455
6	4.391	4.395	4.401	4.399	4.450
7	4.396	4.400	4.401	4.399	4.453
8	4.404	4.406	4.401	4.400	4.456
9	4.393	4.396	4.400	4.399	4.454
10	4.397	4.401	4.400	4.399	4.454
11	4.401	4.403	4.399	4.399	4.452
12	4.390	4.392	4.398	4.398	4.450
13	4.395	4.397	4.398	4.398	4.449
14	4.401	4.403	4.399	4.398	4.449
15	4.402	4.402	4.398	4.398	4.450
16	4.385	4.385	4.397	4.397	4.448
17	4.398	4.398	4.398	4.396	**4.446**
18	4.398	4.398	4.398	4.397	4.448
19	4.397	4.397	4.398	4.397	4.448
20	4.398	4.398	4.398	4.397	4.449
21	**4.385**	**4.385**	4.397	4.396	4.447
22	4.389	4.389	**4.396**	4.396	4.447
23	4.400	4.400	4.397	4.396	4.447
24	4.403	4.403	4.397	4.397	4.448
25	4.404	4.404	4.397	4.396	4.448
26	4.401	4.401	4.398	4.397	4.448
27	4.396	4.396	4.397	4.396	4.448
28	4.389	4.389	4.397	**4.395**	4.447

4.3.5 MAPE of Regression Combination method

Different from the methods above, this method uses a regression model instead of forecasting residuals to estimate the weight for each sister forecast. The regression model needs historical data so the number of days it needs to estimate parameters is tested. The results show using 50 days to estimate parameters has the lowest MAPE, which is 4.299%. This MAPE is lower than benchmark 2. The comparison of MAPEs of using N-day actual load as historical data is in Table 4.9.

Table 4.9: MAPE (%) comparison of using N-day actual load as historical data for regression combination.

N	**MAPE**	***N***	**MAPE**	***N***	**MAPE**	***N***	**MAPE**
1	5.341	22	4.457	43	4.349	64	4.328
2	5.088	23	4.450	44	4.344	65	4.323
3	4.832	24	4.438	45	4.335	66	4.325
4	4.839	25	4.423	46	4.332	67	4.329
5	4.845	26	4.417	47	4.325	68	4.327
6	4.753	27	4.412	48	4.320	69	4.321
7	4.687	28	4.410	49	4.313	70	4.319
8	4.653	29	4.420	**50**	**4.297**	71	4.318
9	4.622	30	4.428	51	4.299	72	4.317
10	4.606	31	4.435	52	4.302	73	4.313
11	4.578	32	4.412	53	4.303	74	4.316
12	4.541	33	4.401	54	4.310	75	4.317
13	4.539	34	4.396	55	4.315	76	4.320
14	4.511	35	4.389	56	4.310	77	4.322
15	4.492	36	4.380	57	4.305	78	4.326
16	4.499	37	4.364	58	4.306	79	4.331
17	4.510	38	4.361	59	4.308	80	4.332
18	4.529	39	4.356	60	4.312	81	4.331
19	4.516	40	4.348	61	4.319	82	4.328
20	4.500	41	4.344	62	4.327	83	4.328
21	4.477	42	4.346	63	4.329	84	4.327

4.4 Summary and Discussion

Figure 4.5 and Table 4.10 show the MAPEs of all the techniques. In the figure GA denotes group analysis method and RC denotes rolling combination method. Nearly every method manages to beat benchmark 1. Only the modified simple average and regression combination beat the benchmark 2. The regression combination method outperforms others in this test case. In rolling combination methods, model C, model D, and model E all outperform model A and model B. Many people probably think it could generate more accurate forecasts by using more than one day's absolute residuals to calculate weights. However, in rolling combination of this test case, model A and model B outperform model C, model D, and model E which shows using one day's absolute residuals could lead to better forecasts.

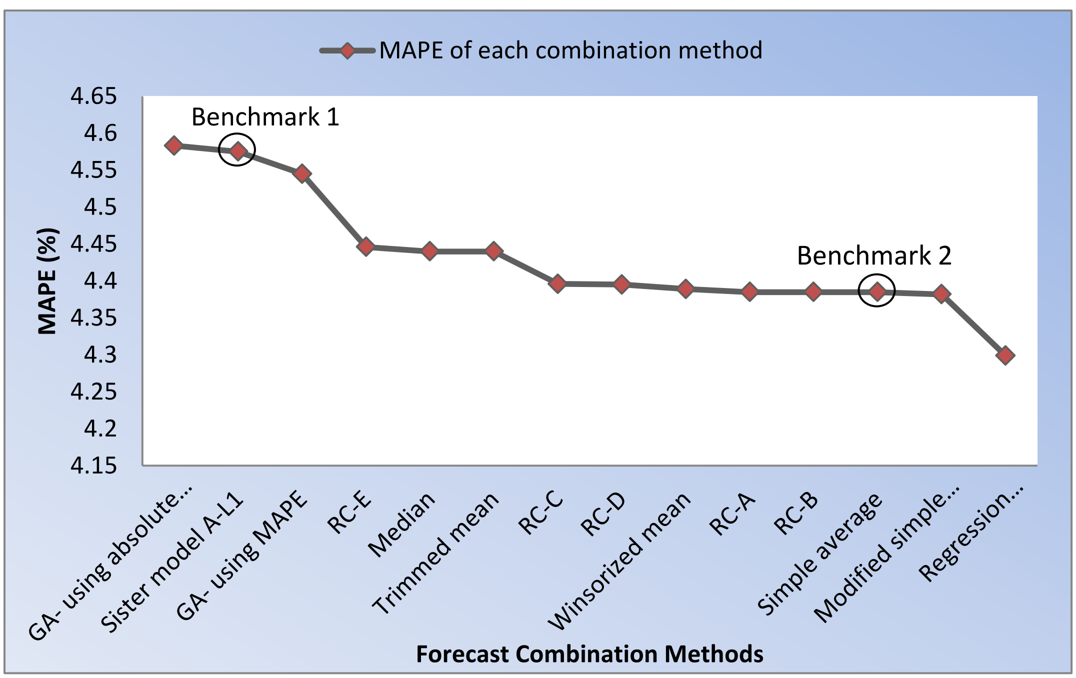

Figure 4.5: Comparison of 13 combination methods and benchmarks. The individual sister model A-L_1 has the largest MAPE while regression combination has the smallest MAPE.

Table 4.10: MAPE (%) comparison of 13 methods with benchmarks.

Methods	**MAPE**
Sister model A-L_1 (benchmark 1)	4.575
Simple average (benchmark 2)	4.385
Modified simple average	4.382
Median	4.440
Trimmed mean	4.440
Winsorized mean	4.389
Group analysis-using absolute residual	4.583
Group analysis-using MAPE	4.545
Rolling combination model-A	4.385
Rolling combination model-B	4.385
Rolling combination model-C	4.396
Rolling combination model-D	4.395
Rolling combination model-E	4.446
Regression combination	**4.299**

To verify the usefulness of the regression combination, its best model (N=50) is tested on the data of 2009. For 2009, the sister forecasting model with lowest MAPE, which is also benchmark 1, is model A-L_1. Its MAPE is 5.013%. The MAPE of benchmark 2 is 5.081% (see table 4.11). By using regression combination with 50-day historical data, the MAPE of the combined forecasts is 4.545%, which is a decrease of 9% compared to benchmark 1 and 11% compared to benchmark 2. The test result of 2009 also shows a fact that sometimes simple average can’t provide a better forecast than using individual forecast. It is necessary to have other combination methods when simple average is no longer suitable for the situation.

Table 4.11: MAPE (%) comparison of regression combination method with benchmarks on data of 2008 and 2009.

Techniques	**MAPE (validation period)**	**MAPE (test period)**
Sister model A-L_1 (benchmark 1)	4.575	5.013
Simple average (benchmark 2)	4.385	5.081
Regression combination	4.299	4.545

In addition, the methods are compared using their one-year overall MAPEs. The combination method which has the smallest MAPE doesn’t mean it has the smallest MAPE for every hour. For instance, comparing MAPEs of the top five combination methods with low MAPEs (see Table 4.12), the one has the smallest MAPE for each hour is not always regression combination method. It is possible that regression combination can’t outperform other methods if different error measurement and different dataset are used. However, forecasters can borrow the strategy of applying and comparing the 13 combination methods. Forecasters and researchers could check if it is better to judge a combination method by its average performance than by its hourly performance.

Some scholars have stated that the more diverse the forecasting methods are, the greater the expected improvement in accuracy (Armstrong, 2001). The results of combining sister load forecasts prove that combining the forecasts from similar models can also improve the accuracy of forecasting significantly. As mentioned previously, due to many constrains forecasters or research groups can only concentrate on studying particular forecasting method. Combining sister forecasts could be more suitable to real-world cases than combining diverse forecasts.

Table 4.12: Hourly MAPEs (%) of top five combination methods. The smallest MAPE is highlighted in bold for each hour.

Hour	**Simple average**	**Modified simple average**	**Winsorized mean**	**Rolling Combination model-A**	**Regression Combination**
1	4.195	4.181	4.251	4.124	**4.017**
2	4.341	4.345	4.430	4.306	**4.155**
3	4.491	4.495	4.600	4.437	**4.238**
4	4.554	4.554	4.666	4.486	**4.302**
5	4.399	4.383	4.488	4.347	**4.224**
6	4.319	4.320	4.427	4.289	**4.197**
7	4.628	4.649	4.725	4.640	**4.522**
8	3.999	4.007	4.108	3.967	**3.834**
9	3.820	3.838	3.862	3.793	**3.701**
10	4.042	4.040	4.114	4.031	**3.872**
11	4.256	4.252	4.313	4.313	**4.087**
12	4.489	4.483	4.535	4.495	**4.394**
13	4.650	4.644	4.694	4.632	**4.617**
14	4.702	**4.684**	4.757	4.713	4.701
15	4.858	**4.842**	4.889	4.869	4.851
16	4.959	4.950	4.982	4.982	**4.887**
17	5.000	4.995	5.043	5.041	**4.932**
18	4.974	**4.933**	5.073	5.132	5.048
19	4.652	4.643	4.755	4.788	**4.558**
20	4.263	4.242	4.340	4.300	**4.212**
21	3.819	3.819	3.878	**3.800**	3.832
22	3.791	3.814	3.852	**3.774**	3.877
23	3.897	3.912	3.951	**3.866**	3.993
24	4.139	4.148	4.189	**4.107**	4.112

CHAPTER 5: CONCLUSION AND FUTURE WORK

There is a large amount of literature devoted to the improvement of load forecasting accuracy, but with few on load forecast combination. Most papers on load forecast combination only focus on combining forecasts from independent forecasting techniques. The combination method commonly used is simple average. This thesis investigates combining sister load forecasts, which are developed from a group of sister models with similar model structure. The proposed method can help forecasters improve forecasting accuracy with reasonable efforts and resource requirements. 13 combination methods are tested using data from the Global Energy Forecasting Competition 2014. The best-performing combination method in this test case is regression combination. The out-of-sample test shows that it decreases MAPE approximately 10% comparing with the two benchmarks. The aforementioned forecasting method by using recency effect has become the state-of-the-art forecasting method in the load forecasting area. By testing the combination methods on the sister forecasts which generated by using recency effect, the forecasting accuracy has been improved one-step further. This test also proves that because load is very stochastic, sometimes using simple average to combine forecasts may not improve the accuracy so it is better to have backup load forecast combination methods. For different scenarios the best method may vary, but the overall load forecast combination methodology and strategy in this thesis can be borrowed and applied. This thesis is also the first formal investigation on combining sister forecasts.

Although the representative combination methods have been selected and tested thoroughly, there is still room for improvement. The parameters of the regression models in regression combination are not constrained, which means that it is possible that 1) they can be negative and/or 2) the sum of them could be larger or smaller than 1. From the perspective of improving forecasting accuracy, the value of these parameters is not analyzed in detail. It is possible that positive and sum-to-unity weights may further enhance the forecast combination. Additional research could be conducted to analyze constrained parameters for regression models.

Besides the research direction mentioned above, there are also three topics deserve some further analysis. Firstly, the forecasts used in this thesis are all point forecasts. Forecasters could combine probabilistic forecasts and investigate the best combination method for them (see e.g. Wallis, 2011; Hall & Mitchell, 2007). Secondly, the tested combination methods are quite basic and there are more weighted averaging schemes that may be beneficial for load forecast combination. Forecasters may test additional combination methods on sister forecasts and check if accuracy of forecasting could be improved. Examples of other combination methods include ANN or Least Absolute Deviation method (Nowotarski, Raviv, Trück, & Weron, 2014). Lastly, combining sister electricity price forecasts can be another direction to pursue.

REFERENCES

Amjady, N. (2001). Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Transactions on Power Systems*, *16*(3), 798–805.

Armstrong, S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners* (p. 417- p. 439). Springer Science & Business Media.

Batchelor, R., & Dua, P. (1995). Forecaster Diversity and the Benefits of Combining Forecasts. *Management Science*, *41*(1), 68–75.

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Operational Research Quarterly*, *20*(4), 451–468.

Chen, B.-J., Chang, M.-W., & Lin, C.-J. (2004). Load Forecasting Using Support Vector Machines: A Study on EUNITE Competition 2001. *IEEE Transactions on Power Systems*, *19*(4), 1821–1830.

De Menezes, L. M., W. Bunn, D., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204.

Fan, S., Chen, L., & Lee, W.-J. (2009). Short-term load forecasting using comprehensive combination based on multimeteorological information. *IEEE Transactions on Industry Applications*, *45*(4), 1460–1466.

Fan, S., & Hyndman, R. J. (2012). Short-Term Load Forecasting Based on a Semi-Parametric Additive Model. *IEEE Transactions on Power Systems*, *27*(1), 134–141.

Fan, S., Methaprayoon, K., & Lee, W. J. (2009). Multiregion load forecasting for system with large geographical area. *IEEE Transactions on Industry Applications*, *45*(4), 1452–1459.

Fay, D., & Ringwood, J. V. (2010). On the Influence of Weather Forecast Errors in Short-Term Load Forecasting Models. *IEEE Transactions on Power Systems*, *25*(3), 1751–1758.

Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*(2), 197–204.

Gross, G., & Galiana, F. D. (1987). Short-term load forecasting. *Proceedings of the IEEE*, *75*(12), 1558–1573.

Hagan, M. T., & Behr, S. M. (1987). The Time Series Approach to Short Term Load Forecasting. *IEEE Transactions on Power Systems*, *2*(3), 785–791.

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*(1), 1–13. http://doi.org/10.1016/j.ijforecast.2006.08.001

Hippert, H. S., & Pedreira, C. E. (2004). Estimating temperature profiles for short-term load forecasting: neural networks compared to linear models. *IEE Proceedings - Generation, Transmission and Distribution*. IET Digital Library.

Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55.

Hong, T. (2008). *Long-Term Spatial Load Forecasting Using Human-Machine Co-construct Intelligence Framework - NCSU Digital Repository*.

Hong, T. (2010). *Short Term Electric Load Forecasting*.

Hong, T. (2014). Energy forecasting: Past, present, and future. *Foresight: The International Journal of Applied*, (32), 43–48.

Hong, T., Liu, B., & Wang, P. (2015). *Electric Load Forecasting with Recency Effect: a Big Data Approach*. www.onedrive.live.com/redir.aspx?cid=bca4c2dc57a4cb2a&resid=BCA4C2DC57A4CB2A!1939

Hong, T., Pinson, P., & Fan, S. (2014). Global Energy Forecasting Competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., & Shahidehpour, M. (2015). *Load Forecasting Case Study*. www.naruc.org/Grants/Documents/Load%20Forecasting%20Case%20Study%20_%20UNCC.pdf

Hong, T., & Wang, P. (2014). Fuzzy interaction regression for short term load forecasting. *Fuzzy Optimization and Decision Making*, *13*(1), 91–103.

Hong, T., Wang, P., & White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting*, *31*(2), 286–295.

Hong, T., Wang, P., & Willis, H. L. (2011). A Naïve multiple linear regression benchmark for short term load forecasting. In *2011 IEEE Power and Energy Society General Meeting* (pp. 1–6). IEEE.

Hong, T., Wilson, J., & Xie, J. (2014). Long Term Probabilistic Load Forecasting and Normalization With Hourly Information. *IEEE Transactions on Smart Grid*, *5*(1), 456–462.

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*(1), 163–169.

Lai, S.-H., & Hong, T. (2013). When One Size No Longer Fits All: Electric Load Forecasting with a Geographic Hierarchy. *SAS White Paper*.

Liu, B., Liu, J., & Hong, T. (2015). *Sister models for load forecast combination*. *HSC Research Reports*. Hugo Steinhaus Center, Wroclaw University of Technology.

Matijaš, M., Suykens, J. A. K., & Krajcar, S. (2013). Load forecasting using a multivariate meta-learning system. *Expert Systems with Applications*, *40*(11), 4427–4437.

Metaxiotis, K., Kagiannas, A., Askounis, D., & Psarras, J. (2003). Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management*, *44*(9), 1525–1534.

Moghram, I., & Rahman, S. (1989). Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on Power Systems*, *4*(4), 1484–1491.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2011). *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons.

Nowotarski, J., Raviv, E., Trück, S., & Weron, R. (2014). An empirical comparison of alternative schemes for combining electricity spot price forecasts. *Energy Economics*, *46*, 395–412.

Palm, F. C., & Zellner, A. (1992). To combine or not to combine? issues of combining forecasts. *Journal of Forecasting*, *11*(8), 687–701.

Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*(4), 1535–1547.

Pesaran, M. H., & Pick, A. (2011). Forecast Combination Across Estimation Windows. *Journal of Business & Economic Statistics*, *29*(2), 307–318.

Smith, D. G. C. (1989). Combination of forecasts in electricity demand prediction. *Journal of Forecasting*, *8*(3), 349–356.

Srinivasan, D., Tan, S. S., Chang, C. S., & Chan, E. K. (1998). Practical implementation of a hybrid fuzzy neural network for one-day-ahead load forecasting. *IEE Proceedings - Generation, Transmission and Distribution*, *145*(6), 687.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450.

Taylor, J. W. (2012). Short-Term Load Forecasting With Exponentially Weighted Methods. *IEEE Transactions on Power Systems, 27*(1), 458–464.

Taylor, J. W., & McSharry, P. E. (2007). Short-Term Load Forecasting Methods: An Evaluation Based on European Data. *IEEE Transactions on Power Systems*, *22*(4), 2213–2219.

Wallis, K. F. (2011). Combining forecasts – forty years later. *Applied Financial Economics*, *21*(1-2), 33–41.

Wang, J., Zhu, S., Zhang, W., & Lu, H. (2010). Combined modeling for electric load forecasting with adaptive particle swarm optimization. *Energy*, *35*(4), 1671–1678.

Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach* (p. 192). John Wiley & Sons.

Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of Weights in Combining Forecasts. *Operations Research*, *40*(3), 609–614.

Xie, J., Hong, T., & Stroud, J. (2015). Long-Term Retail Energy Forecasting With Consideration of Residential Customer Attrition. *IEEE Transactions on Smart Grid*, *PP*(99), 1–1.