RETAIL ENERGY FORECASTING

by

Jingrui Xie

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. Arnulfo de Castro

ABSTRACT

JINGRUI XIE. Retail energy forecasting.
(Under the direction of DR. TAO HONG)

Deregulation of the electric power industry has created both wholesale markets and retail markets. Most load forecasting studies in the literature are on the wholesale side. Minimal research efforts have been devoted to tackling the challenges on the retail side, such as limited data history and high customer attrition rate. This paper proposes a comprehensive methodology to retail energy forecasting in order to feed the forecasts to a conservative trading strategy. The problem is dissected into two sub-problems: load per customer forecasting and tenured customer forecasting. Regression analysis and survival analysis are applied to each sub-problem, respectively. The proposed methodology has been implemented at a fast growing retailer in the U.S. showing superior performance, in terms of Mean Absolute Percentage Error (MAPE) of both daily and monthly energy, over a commonly used method that assumes constant customer attrition rate.

ACKNOWLEDGMENTS

Dr. Tao Hong was the first person who introduced me to the energy forecasting field, then became the advisor of my master study in the MSEM program and my PhD study in the INES program. Dr. Hong guided me step-by-step in learning the analytical skills in energy forecasting, gathering the capability of using the skills in solving industrial energy forecasting problems, and building up the attitude and continuous learning habits a researcher / a forecaster should have. Dr. Hong always encourages me to continuously challenge myself to get improved and provides me great resource to enable me to grow and to be successful. I would like to express my great appreciate for his mentoring and supports in my research and study.

Dr. Badrul Chowdhury is the first one who systematically introduced me to the field of energy system and energy market. In the energy system course, Dr. Chowdhury always encouraged me to look for my own answers to many energy system and energy market related questions. Thanks to Dr. Chowdhury for helping me build the fundamental background of energy system and energy market analysis.

Dr. Arnulfo de Castro taught me the energy system planning course which is an important part that energy forecasting is used for. I would like to thanks Dr. de Castro for providing great guidance and sufficient freedom for me to learn this field.

Lastly, but the most importantly, I would like to express my deepest appreciation to my family. My husband, Yue Chai, unconditionally loved and encouraged me that supported me to finish my study during my most difficult time while I was also working full-time and having a new-born joining the family. I also would like to thank my parents

Guochun Xie and Junfang Liu, and my parents-in-law, Hongfeng Chai and Sheng Wang, for their support spiritually and financially that enabled me to focus on my study and research.

DEDICATION

To My Family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AEMO	Australia Energy Market Operator
C	dummy variable for censoring
CC_0	customer count at the beginning of the test period
CC_N	customer count at the end of the test period
CC_t	customer count at time t
D_a	the end date of the analysis period
D_{ci}	the service connection date of customer i
Day	day of the week
ERCOT	Electric Reliability Council of Texas
FERC	Federal Energy Regulatory Committee
G&T	generation and transmission
GEFCom2012	Global Energy Forecasting Competition of 2012
h	survival rate from the commonly used method
Hour	hour of the day
IPP	independent power producer
ISO	independent system operator
K-M estimators	Kamplan-Meier estimators
LDC	local distribution and service company
M&A	merge and acquisition
MAPE	mean absolute percentage error
MLR	multiple linear regression
Month	month of the year

NCEMC	North Carolina Electric Membership Corporation
PJM	PJM Interconnection LLC
PURPA	public utility regulatory policies act
REP	retail electric provider
SA1	strata strategy 1
SA2	strata strategy 2
SA3	strata strategy 3
t	time t (in days)
T_i	tenure of customer i
T&D	transmission and distribution
TMP_k	temperature of the k^{th} hour
Z1	zone 1

CHAPTER 1: INTRODUCTION

1.1 The Deregulation of the Electric Market

There are four key components on the utility industry's supply chain: they are *Generation (G), Transmission (T), Distribution (D)* and *Retail Service (R)*. A typical traditional regulated market structure is shown in Figure 1(a): under this structure, each vertically-integrated and regulated utility performs all these four functions as the owner-operator-seller within its retail service territory. In the regulated era, other than this classic structure, some alternatives may also exist: (1) Independent Power Producers (IPPs). The Public Utility Regulatory Policies Act (PURPA) signed in 1978 requires vertically-integrated utilities to purchase power from IPPs that sell electricity at a lower cost than the utility would have the power generated by themselves. Such industry structure is shown in Figure 1 (b). (2) Another alternative is shown in Figure 1 (c) where a bunch of small local distribution and service companies (LDCs) jointly own a G&T company to generate and transfer the power they need. [1][2]

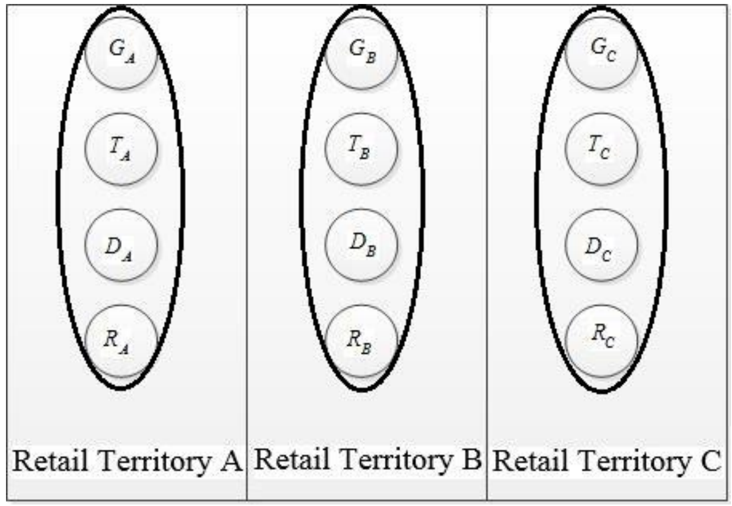


Figure 1 (a): Regulated structure – vertically integrated utilities

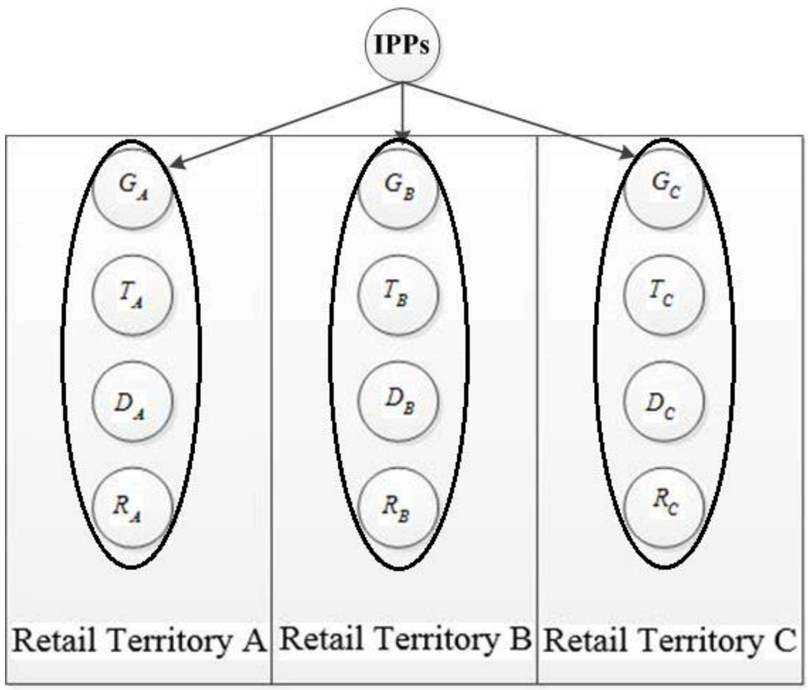


Figure 1 (b): Regulated structure – vertically integrated utilities with IPPs

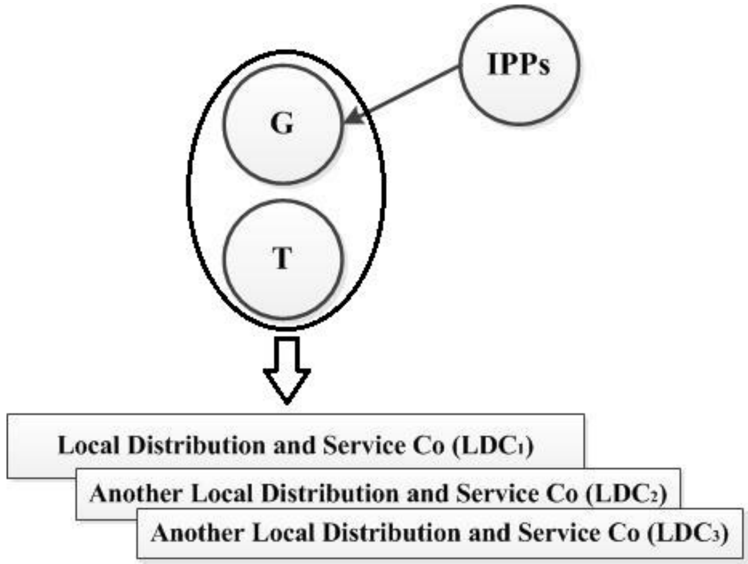


Figure 1 (c): Regulated structure – LDCs own G&T

Such a regulated structure has dominated the U.S. electric market for decades. In order to bring in competition and encourage innovations, a series of laws have been passed since 1978 to de-regulate this market. Figure 2 (a) shows a basic structure of the de-regulated electric market: many generation companies compete in the wholesale electric market; independent system operators (ISOs) run regional transmission grid to provide generators unbiased access to the transmission lines; local distribution companies perform the distribution and retail services function and are still regulated. Even though there may be some variations between this basic structure and the actual structure implemented from state to state within the U.S., this basic structure represents the mainstream that is the generation level is de-regulated but not the retail level. This is mainly because the generation level and the transmission level usually involve the inter-state power supply which is mainly regulated by the Federal Energy Regulatory Committee (FERC). In such case, de-regulation can be ordered by FERC uniformly around the country. But, at the retail level, the state regulatory has the authority to decide if the market should be de-regulated.

In the U.S., most of the states are still regulated at the retail level, even though in recent years, some states have also de-regulated the electric market at the retail level where the end users can choose their electric power provider that provides services in the same territory such as shown in Figure 2 (b). [3][4]

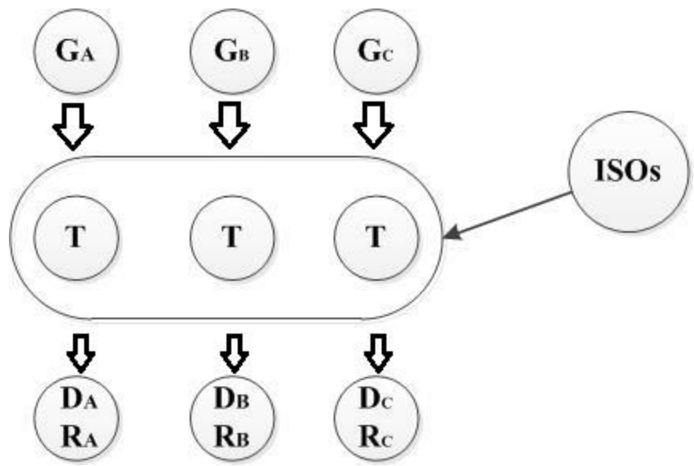


Figure 2 (a): A basic de-regulated structure

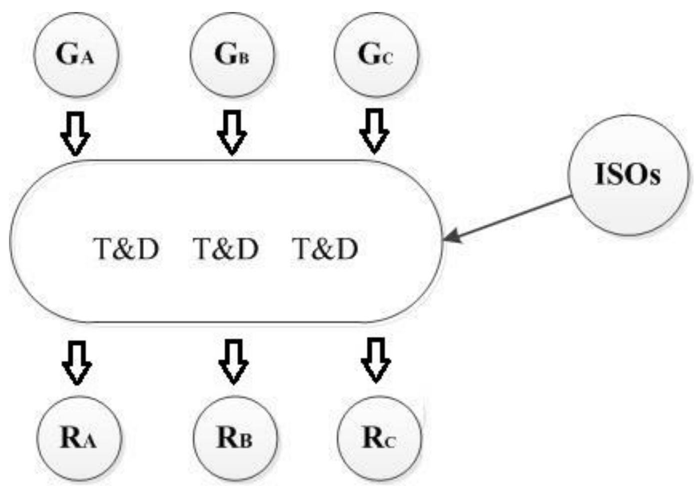


Figure 2 (b): A variation of the basic de-regulated structure

1.2 Electric Load Forecasting

Utilities use load forecasts for many different purposes from planning, operation to other market activities [5]–[7].

(1) Planning. For generation planning, utilities usually need load forecasts at system level in order to plan the generation resources in advance. For transmission and distribution (T&D) planning, the utility need to understand the power flow, analyze the reliability of the system and properly maintain and upgrade the system to meet the demand. The load forecasts are usually provided at substation or small area level for this purpose and are known as spatial load forecasts which answer *when*, *where* and *how much* the load would be.

(2) Operation and Maintenance. One important subject in utilities' daily operation is unit commitment that is the utility needs to plan which generation unit should be shut down and started up at a certain time. To do so, the utility need hour(s)-ahead to a few days ahead load forecasts to schedule the start-up and the shut-down of the generation plants while considering the reliability issue, the environmental issue and the economic cost at the same time. Another subject that is also highly related to the load forecasts used in utilities' operation and maintenance activities is demand side management: in the short run, load forecasts support the utility's daily decision on load control; in the long run, load forecasts with consideration of demand side management helps utilities on understanding the load consumption pattern of the end users and planning the maintenance and upgrade of the system accordingly.

(3) Other market activities. Utilities depend heavily on load forecasts at different time horizon to conduct energy purchasing activities: they can enter bilateral trading to

commit purchase or assets for a certain period, they can purchase a large chunk of power in advance and then adjust a smaller amount of purchase in the day-ahead market. Also, the middle-term or long-term energy forecasts (which are a few months to several decades ahead) are critical information while utilities make decisions about merge and acquisition (M&A) activities, project budgeting and etc.

1.3 Retail Energy Forecasting

As introduced in Section 1.1, deregulation of the electric power industry has created both wholesale markets and retail markets. Most, if not all, load forecasting methodologies and techniques in the literature have been developed for and applied to the transmission and distribution (T&D) companies and independent systems operators on the wholesale side. The research community has not yet devoted many efforts to the forecasting problems in the retail markets.

Load forecasting for retail electric provider (REP) is different from that for T&D companies in the following three aspects among others:

1) Length of history. Since most retailers are younger than the T&D companies, they have shorter data history than that of T&D companies. The length of data history available to a long term retail energy forecasting study typically ranges from a few months to several years. A well-established retailer may have over 10 years of history. Nevertheless, due to the changes in product offerings, the data of the very early years may not be representative to today's customer behavior. Therefore, we do not have the luxury of doing comprehensive sliding simulation or rolling regression in a retail energy forecasting study as what we can do in a typical long term load forecasting study [8].

2) Customer attrition. A T&D company has a low attrition rate of its customer base. There are two common ways that a customer of a T&D company can turn over. The customer can move to a territory operated by a different utility. The customer can also switch some of the functions powered by electricity to the alternatives powered by gas. Either way is a relatively big change to the customer. Therefore, the attrition rate of a T&D company is quite low, i.e., typically less than 3% per year. On the other hand, the customers of a retailer

may choose a different provider at any time, as is common in the insurance industry. As a result, the retailers are often dealing with a ten times higher attrition rate than that of a T&D company. The high attrition rate leads to the volatility in a retailer's customer count. 3) Customer count. A T&D company has a fairly stable customer count. The number of customers may increase or decrease year over year but usually at a modest rate, e.g., less than 2%. Depending upon the level of investment in sales and marketing activities, the number of customers in a given zone of a retailer can increase over 20% in a period of few weeks.

The above characteristics of electricity retailers bring significant challenges to retail energy forecasting. As a result, many REPs deploy a conservative trading strategy. At any given planning cycle, the trading team only secures power for the existing customers throughout the planning horizon. Due to customer attrition, some of the existing customers will leave sometime during the planning horizon. Consequently, the trading team should purchase the contracts to meet the demand with a decreasing trend throughout the planning horizon, i.e., a few months to one or two years. In the next planning cycle, there are usually new customers coming in as a result of sales and marketing efforts. Then the trading team will secure power for these additional customers, again, to meet a decreasing demand trend because of customer attrition.

This trading strategy is illustrated in Figure 3 In the upper plot, at the start of the 1st planning cycle, the trading team only purchases the contracts for a decreasing demand which is the shadow area ABCD. In the lower plot, at the beginning of the 2nd planning cycle, the trading team will re-evaluate the existing customer count. They will then purchase the contracts for a decreasing demand (i.e. area EFGH) minus the amount of power that has been purchased during the 1st planning cycle (i.e. area CDIF), which yields a net purchase shown by the shadow area DJEI plus CGHJ. Between the starting dates of the 1st and the 2nd planning cycles, there is a difference between the actual demand and those purchased at the beginning of the 1st planning cycle (bubble area AIE in the lower plot). This part of energy will be managed through short term transactions, namely the trades placed at the day-ahead markets.

The aggressiveness of the sales and marketing strategy would greatly affect the customer count, which in turn affects the total load. A benefit of deploying this conservative trading strategy is to separate the uncertainties brought by sales and marketing efforts from the energy trading functions. However, this conservative trading strategy brings a research problem of retail energy forecasting: how to forecast the demand of the tenured customers?

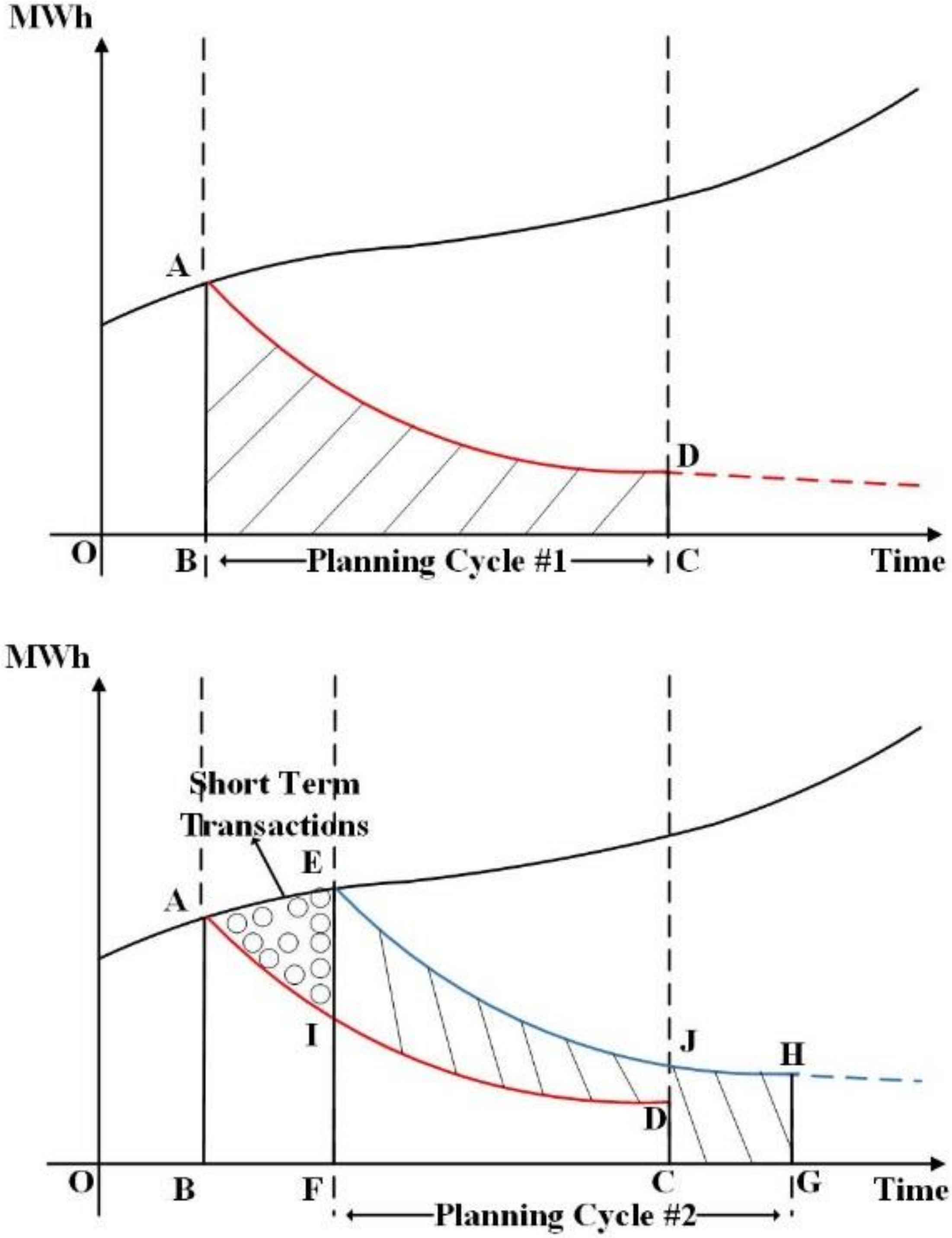


Figure 3: A conservative trading strategy

CHAPTER 2: LITERATURE REVIEW

2.1 Load Forecasting

Load forecasting problems can be categorized into two groups: short term load forecasting and long term load forecasting. Short term load forecasts, of which the forecast horizon is up to two weeks, are primarily used in power systems operations, such as unit commitment and economic dispatch. Long term load forecasts, of which the forecast horizon may range from a few months to several decades, are mostly used in power systems planning and financial planning.

Hong offered an overview of the evolution of load forecasting research and practice over the past century [9]. Most load forecasting papers in the literature are on short term load forecasting, where researchers have tried various techniques such as Artificial Neural Networks [10], multiple linear regression [11], [5], semi-parametric additive models [12], [13], and fuzzy linear regression [14]. A hybrid forecasting framework was proposed to address the bidirectional dependencies between electricity demand and price in price-responsive environment [15]. Recently, the Global Energy Forecasting Competition 2012 brought together several novel ideas to short term hierarchical load forecasting [16], such as parametric and semi-parametric models [17], [18], and gradient boosting machines [19], [20] presented a long term probabilistic load forecasting and normalization methodology implemented at North Carolina Electric Membership Corporation (NCEMC), a large generation and transmission cooperative in the U.S. [8]. The NCEMC case study demonstrated the superiority of using hourly information over the models based on low resolution data at monthly or daily interval. The proposed approach also applied the integrated load forecasting approach mentioned in [5] by leveraging short term load forecasting models in long term load forecasting process.

Comparing with traditional high-voltage level load forecasting for system operators and/or utilities, load forecasting for REPs has two major challenges due to the dynamic nature of the retail market: limited data history and high customer attrition rate. Several papers in the literature tried to address some challenges in a dynamic environment. Charytoniuk et al. proposed an indirect neural network based method utilizing the monthly energy consumption data and the customer demand survey data to forecast the long term electric demand in a deregulated environment [21]. This approach targeted the challenge of limited historical data with the underlying assumption that the customer base would be stable throughout the forecast horizon. Chan et al. introduced a generalization learning strategy on short-term load forecasting [22]. The proposed method utilized the techniques of cross-validation, regularization and pruning to tackle the limited training data issue for an ANN-based model. Kandil et al. presented a long term load forecasting case study for fast developing utilities [23]. Such approach utilized low resolution data at annual interval to fit polynomial regression models. It also took human inputs to try to overcome the uncertainty with the changing environment. Nevertheless, neither of these studies tackled the challenge of customer attrition in retail markets.

2.2 Survival Analysis

Detailed introduction to the theory of survival analysis is covered in [24]. Forecasting using survival analysis is briefly discussed in [25]. Many industries use survival analysis to serve their various analysis purposes: system engineers use survival analysis for the reliability analysis, manufacturing also use survival analysis for equipment failure, and the service industries use survival analysis to model, analyze and manage the customer attrition [26]–[28].

In this paper, a comprehensive solution to retail energy forecasting is proposed: hourly load and weather information are used to overcome the challenge of short data history. The challenge of high customer attrition rate is addressed by applying survival analysis. The significance of this paper includes three aspects. Firstly, an important problem of load forecasting in the retail markets where customers frequently switch among multiple service providers is raised. Secondly, a practical solution to integrate customer attrition modeling into load forecasting process is proposed. Lastly, the proposed solution has been illustrated through a field validation at a fast growing retailer in the United States.

CHAPTER 3: THEORETICAL BACKGROUND

In this chapter, the theoretical background of the techniques used in this thesis will be introduced： Section 3.1 will cover the multiple linear regression method that is related to load forecasting; Section 3.2 will introduce the fundamental of the survival analysis method; A brief summary of the model evaluation technique used in this thesis will be presented in Section 3.3.

3.1 Multiple Linear Regression

Multiple linear regression (MLR) has been widely used in the forecasting field including load forecasting. A comprehensive introduction to MLR can be found in [29]. A general matrix form of the MLR model can be defined as (1) where Y is the $n \times 1$ column vector of observations on the dependent variable, X is the $n \times (p+1)$ matrix representing the one column of ones and p columns of the observations on the independent variables, β is the $(p+1)$ vector of model parameters to be estimated and ε is the $n \times 1$ vector of random errors.

$$Y = X\beta + \varepsilon \tag{1}$$

In this thesis, the dependent variable Y is the hourly electric demand, the independent variables include quantitative variables, qualitative variables and their interactions. An example of the quantitative variable is the hourly temperature: in the summer, as the temperature increases, the electric demand may also increase; in the winter, as the

Temperature increase, the electric demand may decrease. However, this temperature and load relationship may not be linear but nonlinear/curvilinear instead. As a result, the polynomial of the independent variables may be used. Quantitative variables such as month, weekday or hour of day can also be used as independent variables by introducing dummy variables into the model. For example, if month is used as an independent variable in the model, then each of the 12 months will have a dummy variable to represent if the electric demand falls into that month:

$$\begin{bmatrix} X_{1i} \\ X_{2i} \\ \dots \\ X_{12i} \end{bmatrix} = \begin{bmatrix} \begin{cases} 1 \ if \ \ January \\ 0 \ else \end{cases} \\ \begin{cases} 1 \ if \ \ Feburary \\ 0 \ else \end{cases} \\ \dots \\ \begin{cases} 1 \ if \ \ December \\ 0 \ else \end{cases} \end{bmatrix} \tag{2}$$

The interactions between variables should be introduced into the MRL model if the impact of one independent variable depends on the level of another independent variable. Figure 4 shows the different impact of the temperature to the electric demand during weekday and weekend: in this case, the impact of temperature to the electric demand depends on if the demand falls on a weekday or a weekend. The interaction of two independent variables can be represented by multiplying these two variables to generate a new independent variable.

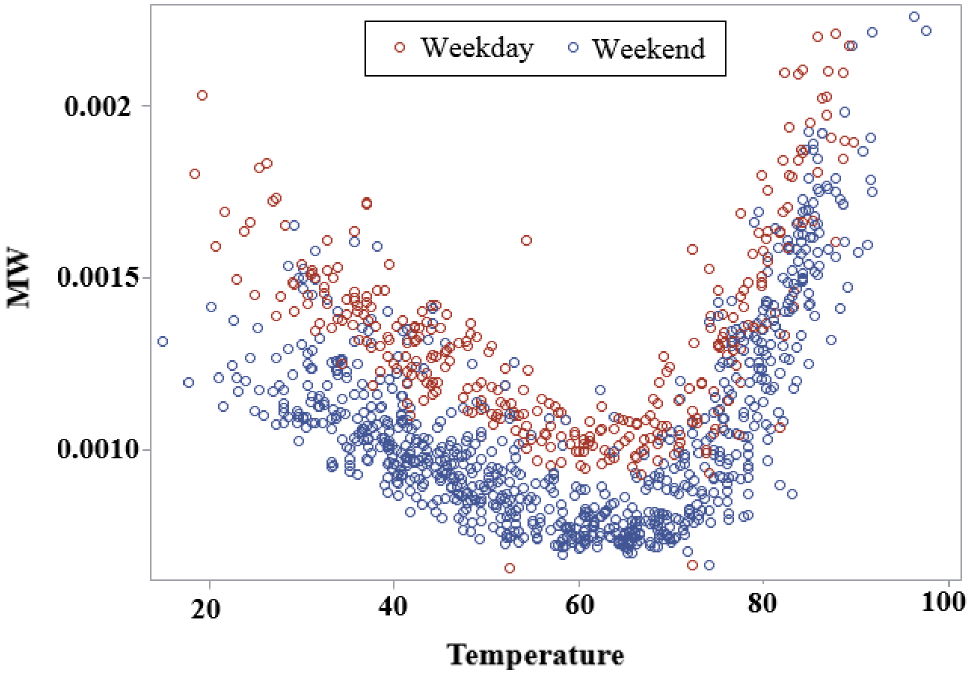


Figure 4: Scatter plot of hourly load and temperature during weekday and weekend

3.2 Survival Analysis

Survival analysis, also known as time-to-event analysis, is a class of statistical methods that deal with analysis of the occurrence and timing of events. It was firstly designed by biostatistician to conduct the study of death and then have been widely used by engineers, economist and sociologist in analyzing survival data which is longitudinal data on the occurrence of events in their fields. For the customer attrition analysis in a service industry, an event can be defined as the customer discontinue with the service. Through analyzing the most important aspect of survival data – tenure (T), which is defined as the time duration from the start of the study to the time the event occurs, survival analysis could tell when the customer would disconnect the service. There is another feature in survival analysis that differentiate it from other statistical methods – censoring: in the case that an event has happened prior to the start of the study, left-censoring is applied; in the case that an event has not happened to some observations when the study ends, these observations will be right-censored. In the customer attrition analysis presented in this study, only right censoring will be applied that is customers who have not discontinued with the service prior to or on the analysis day will be right-censored.

In survival analysis, the survival distribution function can be defined as (3) which is the probability that a subject survives longer than some specified time t and the hazard function can be defined as (4) that is the probability of having the disconnection at tenure t given no prior occurrence of the event is of interest to be estimated:

$$\boldsymbol{S} \equiv \Pr(T > t) \tag{3}$$

$$H_{T} \equiv \Pr(T = t \mid T \geq t) \tag{4}$$

There exist many different methods to estimate the survival distribution function and the hazard function: by specifying a parametric model; by developing an empirical estimate using Kaplan-Meier (K-M) estimators or life table, just to name a few. Both of the Kaplan-Meier method and the life table method do not require pre-assumption on the distribution of the survival time. And, they both use the idea of conditional probability in estimating the survival function, but the K-M method considers the censored at the end of each time interval while the life table method assumes the censored uniformly distributes across each time interval and considers the censored at the middle point of each time interval in estimating the survival function. For a discrete data which is the case in this study where the data is collected and analyzed on a daily basis, the K-M method estimate the survival function as

$$S(t) \equiv \Pr(T > t) = \prod_{m=1}^{t} (1 - \frac{d_m}{s_m}) \tag{5}$$

and the life table method estimate the survival function as:

$$S(t) \equiv \Pr(T > t) = \prod_{m=1}^{t} (1 - \frac{d_m}{s_m - \frac{c_m}{2}}) \tag{6}$$

where d_m is the number of disconnections, s_m is the number of customers that stay into the m^{th} time interval and c_m is the number of censored customers.

If the whole time period could be separated into precise intervals, the sample size is not very large and grouping is not necessary, the K-M method is a good method in providing the survival estimation. But, if the time interval is crude in the data (e.g. the data was collected every quarter/year, 5 years, etc.), the life table estimate is more suitable because it could consider censor in the mid-point. In addition, in the case that the sample size is large such that grouping is needed, the life table estimate would be useful. In the case studied in this research, the target forecasting interval is at daily level and the data was collected at daily level, so both of these two methods could be applicable. However, while implementing the study using SAS PROC LIFETEST procedure [30], the life table estimate also provides estimates for hazard function and it would be useful in estimating the conditional probability of survival, the life table method will be implemented. A simplified example using discrete data and right censoring to estimate the survival distribution and hazard function is presented below:

Table 1: Example of estimating survival function and hazard function using life table method

Interval		**Beg. Total (s_m)**	**Disconnection (d_m)**	**Censored (c_m)**	**Survival**	**Hazard**
0	1	10000	200	0	1.0000	0.0202
1	2	9800	150	10	0.9173	0.0154
2	3	9640	100	20	0.9032	0.0104

3.3 Model Evaluation

Mean absolute percentage error (MAPE) as defined in (7) has been widely used in evaluating the forecasting performance of a model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\,Actual_i - Forecast_i\,|}{|\,Actual_i\,|}\times 100\% \tag{7}$$

A series of MAPE values are of interest to a load forecaster:

(1) Hourly load of a specific period which is usually a year depending on the availability of the data and the purpose of the forecasting.

(2) Daily, monthly or annual energy which can be calculated by aggregating the hourly load to daily, monthly or annual level.

(3) Daily, monthly or annual peak load which is the maximum load of a day, a month or a year.

In this thesis, the hourly MAPE is used to evaluate the load forecasting model while other matrix such as daily energy MAPE and monthly energy MAPE will also be presented, and the daily MAPE is used to evaluate the customer count forecasting. All evaluations are made on an out-of-sample basis for a fixed forecast origin and forecast horizon. [31]

CHAPTER 4: METHODOLOGY

We can naturally dissect the retail energy problem into the following two components as shown in Figure 5: A) modeling and forecasting load per customer; B) modeling customer attrition and forecasting tenured customers. Finally, by multiplying the load per customer forecast with the tenured customer forecast, we will obtain the forecast of tenured demand.

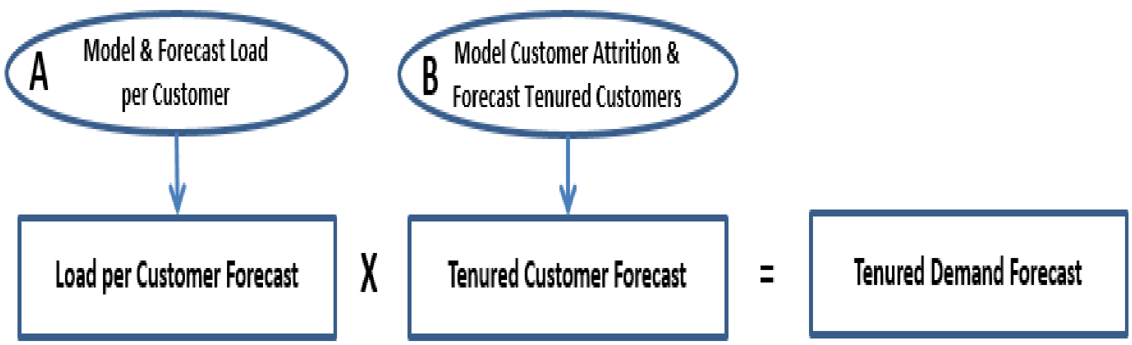


Figure 5: Solution Architecture

4.1 Load per Customer Forecasting

REPs may serve both residential and business customers. Because residential electricity demand has good response to the weather conditions and working schedules, the primary demand of interest for a REP to forecast is from residential customers. Since the electricity consumption pattern of an average residential customer does not change much over few years, we can safely forecast load per customer with the assumption of stable load profile within a year or two.

To develop the model for load per customer forecasting, we follow the variable selection procedure proposed by Hong [5] as summarized below:

1) Start with a benchmark model. A benchmark model should be easy to implement and interpret [11]. A well-established benchmark model is Tao's Vanilla Benchmark Model used in GEFCom2012 [16]. It includes seven main effects (Trend, TMP_k, TMP_k^2, TMP_k^3, Month, Day, Hour) and seven cross effects (Day*Hour, TMP_k*Hour, TMP_k^2*Hour, TMP_k^3*Hour, TMPk*Month, TMP_k^2*Month, TMP_k^3*Month).

2) Model recency effect. Recency effect refers to the effect of past few hours of temperatures on the present demand. Recency effect is modeled by including temperatures of the preceding hours, weighted temperatures of the preceding hours, their polynomials and interactions with calendar variables such as Hour and Month.

3) Model weekend effect. Weekend effect refers to the different demand patterns during different days of a week. Weekend effect is modeled by grouping adjacent days of a week together. The purpose of modeling weekend effect is to reduce the total number of parameters while maintaining similar or better forecasting accuracy.

4) Model holiday effect. Holiday effect refers to the special demand patterns during holidays and their surrounding days. Holiday effect is modeled by treating some of the holidays as weekend days, and the surrounding days of some holidays as weekend or alternative weekdays.

4.2 Customer Attrition Modeling And Tenured Customer Forecasting

The features of the retail customer behavior data (i.e. the survival data of analysis) are illustrated in Figure 6: Each line in the figure represents the lifetime of one customer account. Not all lines start from the origin, because customers are gradually acquired by the REP at different time during the analysis period. Each customer account has a connect date denoted as D_{ci}. If the account is disconnected on or prior to the end of the analysis period (D_a), it would have a disconnect date denoted as D_{di}. Otherwise, the disconnect date will fall into an unknown future date that is beyond the end of the analysis period (the shadow area in Figure 6). The vertical bar shows the tenure of the customer, namely the number of days this customer i has been serviced by the REP:

$$T_i = \begin{cases} D_{di} - D_{ci} \ , \ if \ \ customer \ \mathrm{i} \ already \ disconnected \ on \ or \ prior \ to \ D_a \\ D_a - D_{ci} \ , \ if \ \ customer \ \mathrm{i} \ has \ not \ disconnected \ on \ D_a \end{cases} \tag{8}$$

The customers that stay connected beyond the end of the analysis period will be used for right-censoring. In this analysis, Censor (C) is a dummy variable defined as:

$$C = \begin{cases} 1 \quad if \ \ the \ event \ (disconnection) \ has \ not \ happened \ on \ D_a \\ 0 \quad if \ \ the \ event \ (disconnection) \ has \ already \ happened \ on \ or \ prior \ to \ D_a \end{cases} \tag{9}$$

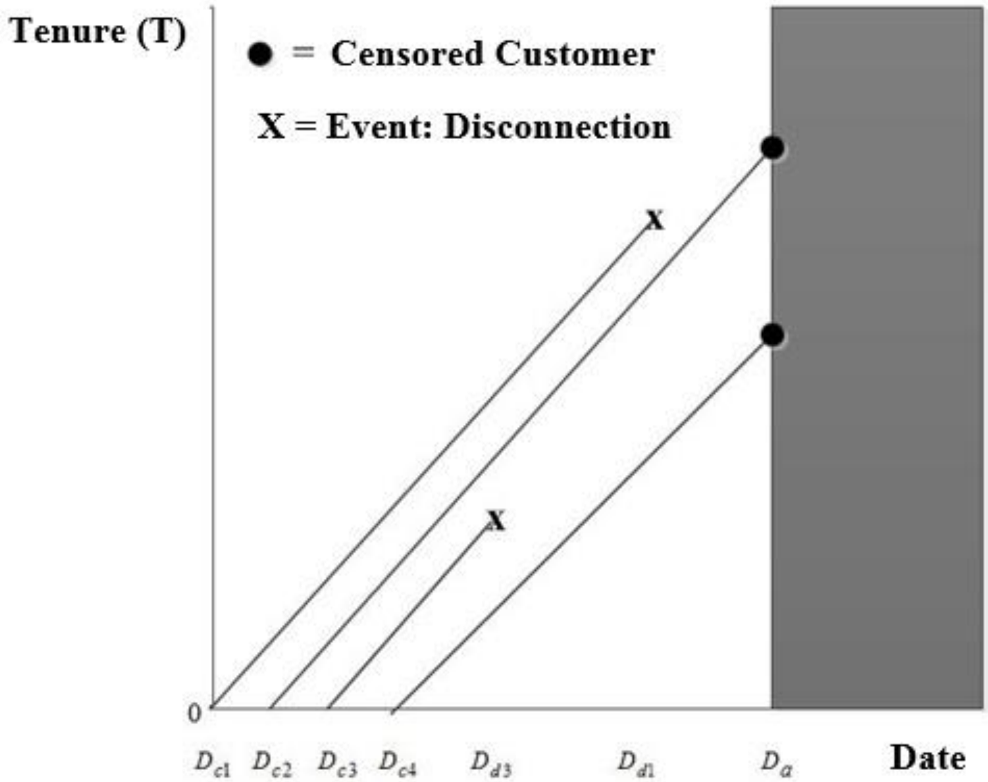


Figure 6: Tenure, event and censor in survival analysis

To forecast the number of tenured customers in the future, we should first estimate the hazard function. We can then derive the conditional probability of a customer i with the tenure of days to stay connected for additional k days (i.e. the forecast horizon):

$$S_{T_i+k|T_i} = \prod_{j=0}^{k-1}(1 - H_{T_i+j}) \tag{10}$$

Finally, the tenured customer forecast, namely the expected number of existing event-free customers that will stay connected for additional k days, can be obtained as:

$$Tenured\ Customer\ Forecast = \sum_{i=1}^{n}\prod_{j=0}^{k-1}(1 - H_{T_i+j}) \tag{11}$$

where n is the number of event-free customers at the forecast origin. It does not account for new customers acquired after the forecast origin.

CHAPTER 5: IMPLEMENTATION

5.1 Data Description

The case study of this paper is from a fastest growing REP in the U.S., which operates in several electricity markets such as ISO New England, PJM and ERCOT. Load forecasts directly drive important decisions around energy trading and financial planning at this REP. Its territory can be divided into 14 zones. In the real-world implementation, the methodology proposed in Chapter 4 was deployed to forecast load of each zone individually. In this paper, we use one zone, denoted as Z1, to illustrate the proposed methodology to avoid verbose presentation.

This REP offers two major plans: fixed plan and variable plan. The fixed plan, which carries a constant rate throughout a contract period, is designed for long term customers. It includes two products based on the length of contract period: 6 months and 12 months. Additional fees will be applied to early termination. The variable plan is designed for short term customers. It carries a rate fluctuating based on market conditions. The customer can terminate the contract any time without penalty. The trading decisions are made separately for these two plans. Therefore, the business needs require two separate load forecasts, one for each plan.

The original dataset includes the contract start and end dates for each customer, hourly load history of total customers, and hourly temperature history of the associated weather stations. Figure 7 shows the daily customer counts from 01/01/2011 (the connect date of the first customer in Z1) to 12/31/2013 (the end of the analysis period for this paper), where Term 6, Term 12 and Term 0 represent the 6-month product, the 12-month product and the variable product respectively. Term 6 is almost invisible, because the number of customers enrolled in this plan is very small.

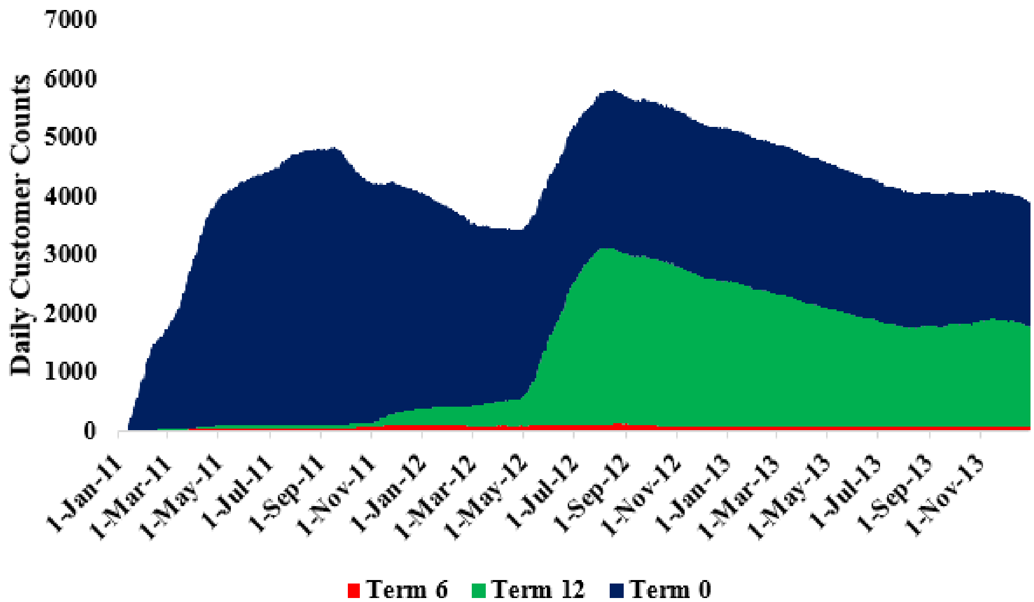


Figure 7: Daily customer counts over the analysis period

When testing the proposed methodology, the data are sliced into three pieces, training, validation and test:

(1) Training data are used to estimate the parameters of the model. All available data on or prior to 9/30/2012 is used to train the model.

(2) Validation data are used as holdout sample to select the best model from the candidates. In this paper, 6 months of data from 10/1/2012 to 3/31/2013 are used for model selection.

(3) Testing data are the data completely blind from training and validation. In other words, the information in the test data is not used for parameter estimation or to model selection.

The purpose of using this test data is to provide an evaluation of the ex post forecasting accuracy without peeking future information. In this paper, we use the last 9 months from 4/1/2013 to 12/31/2013 as testing data.

In sum, among 1096 days of history in Z1, we are using 640 days for training, 181 days for validation, and 275 days for testing. The summary statistics of Z1's load in these three periods are shown in Table 2.

Table 2: Summary statistics on hourly load of Z1

Period	**Days**	**Min (MW)**	**Mean (MW)**	**Maximum (MW)**
Training	640	0.002	5.015	19.165
Validation	181	0.774	5.795	10.586
Test	275	2.265	4.815	13.965

5.2 Forecasting Load per Customer

Dividing the total load of Z1 by the daily customer counts, we can obtain a series of hourly load per customer over the whole analysis period, as shown in Figure 8. The scatter plot in Figure 9 illustrates the relationship between hourly load per customer and hourly temperature of 2011. The load per customer series preserves similar features as those of a typical load series at high voltage level, such as seasonal patterns and the high correlation to the temperature.

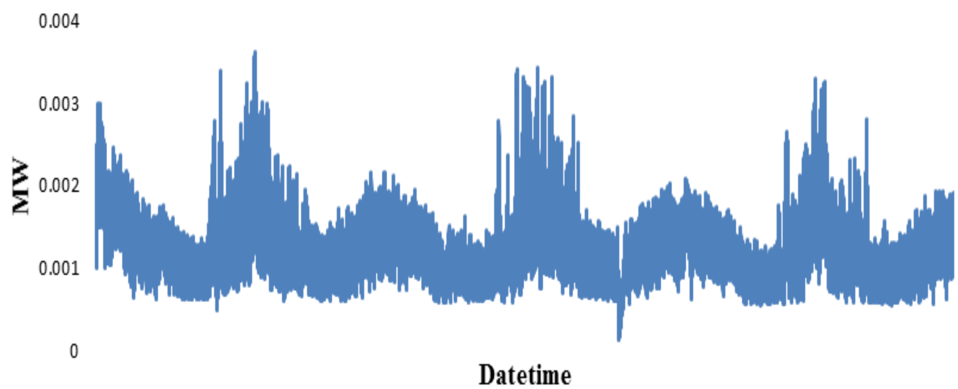


Figure 8: Hourly load per customer (Jan 2011 – Dec 2013)

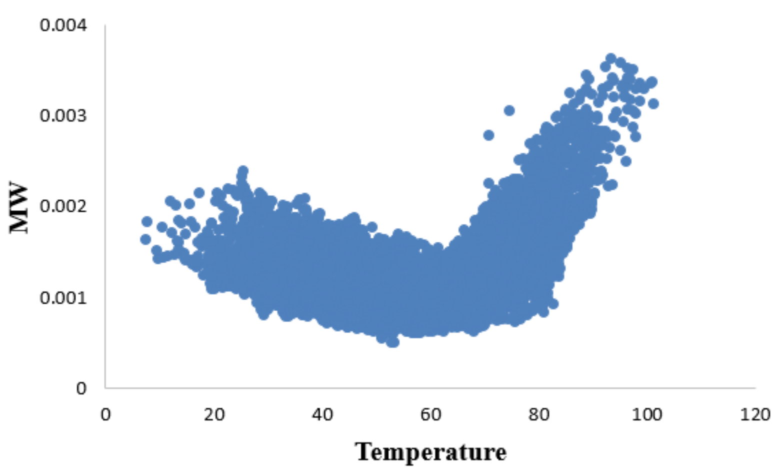


Figure 9: Scatter plot of load per customer and temperature (Year = 2011)

In this paper, Hong's load forecasting methodology is applied, which was proposed in [4] and summarized in Section 4.1, to forecast load per customer. The final model selected based on the validation period is presented in Table 3. The modifications to the day of the week code and holiday code are presented in Table 4 as the result of modeling weekend and holiday effects. The General Linear Model procedure (PROC GLM) of SAS [32] is used to estimate the parameters.

Table 3: Main and cross effects of the general linear model (GLM)

Main Effects	**Cross Effects**
$Trend$, TMP_h, TMP_h^2, TMP_h^3, TMP_{h-1}, TMP_{h-1}^2, TMP_{h-1}^3, TMP_{h-2}, TMP_{h-2}^2, TMP_{h-2}^3, TMP_{avg}, TMP_{avg}^2, TMP_{avg}^3, $Month$, Day, $Hour$	$TMP_h\ Month$, $TMP_h^2\ Month$, $TMP_h^3\ Month$, $TMP_h\ Hour$, $TMP_h^2\ Hour$, $TMP_h^3\ Hour$, $TMP_{h-1}\ Month$, $TMP_{h-1}^2\ Month$, $TMP_{h-1}^3\ Month$, $TMP_{h-1}\ Hour$, $TMP_{h-1}^2\ Hour$, $TMP_{h-1}^3\ Hour$, $TMP_{h-2}\ Month$, $TMP_{h-2}^2\ Month$, $TMP_{h-2}^3\ Month$, $TMP_{h-2}\ Hour$, $TMP_{h-2}^2\ Hour$, $TMP_{h-2}^3\ Hour$, $TMP_{avg}\ Month$, $TMP_{avg}^2\ Month$, $TMP_{avg}^3\ Month$, $TMP_{avg}\ Hour$, $TMP_{avg}^2\ Hour$, $TMP_{avg}^3\ Hour$, $Day\ Hour$

Table 4: Modifications to the days of year for model

Days of a Year (Original)	**Day Code (Modified)**
Tuesday (regular)	Monday
Wednesday (regular)	Monday
Thursday (regular)	Monday
New Year's Day	Friday → Saturday; else → Sunday
Memorial Day	Sunday
Day After Memorial Day	Friday
Day Before Independence Day	Friday
Independence Day	Friday → Saturday; else → Sunday
Day After Independence Day	Friday
Labor Day	Sunday
Day After Labor Day	Thursday
Thanksgiving Day	Special Weekday 8
Day After Thanksgiving Day	Saturday
Christmas Day	Friday → Saturday; else → Sunday

5.3 Model Customer Attrition

Using historical customer account information, we can derive the hazard function to describe customer attrition. The LIFETEST Procedure (PROC LIFETEST) of SAS is used to derive the hazard functions, while the life table method is used for parameter estimation [30], [33].

Since a customer can choose one of the following three contract options, 6-month fixed plan, 12-month fixed plan and variable plan, there are multiple options to group the customers for the purpose of deriving hazard functions. In this paper, the following three different strata strategies are compared: 1) create one group of all customers (SA1); 2) create two groups, one for fixed plan, and the other for variable plan (SA2); 3) separate the three groups (SA3). Figure 10 shows the resulting survival probability plots for each option.

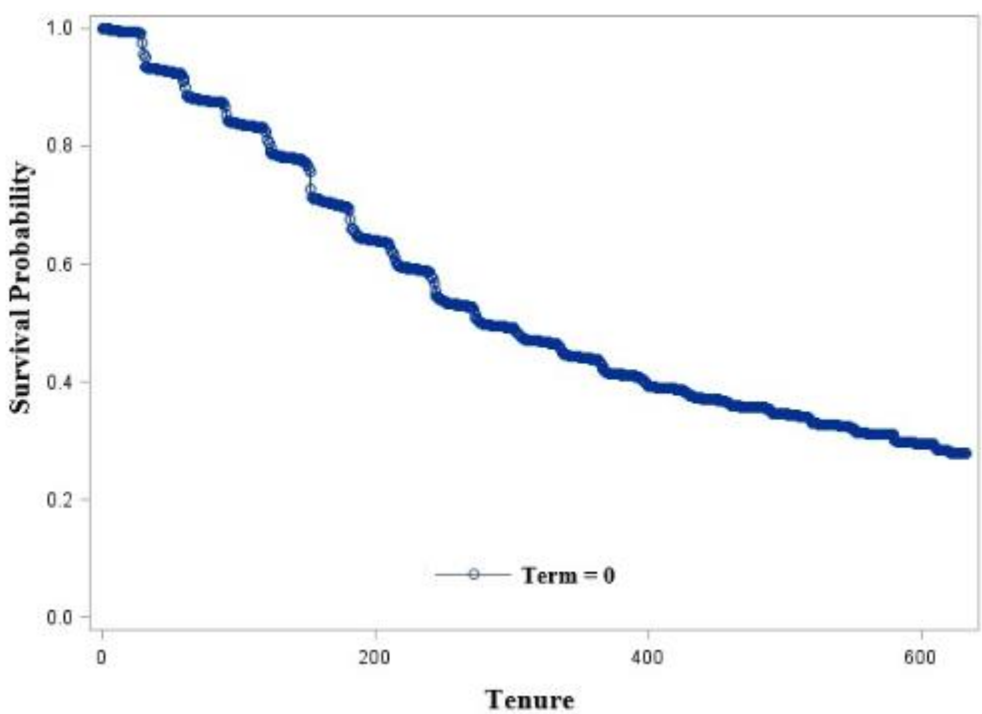
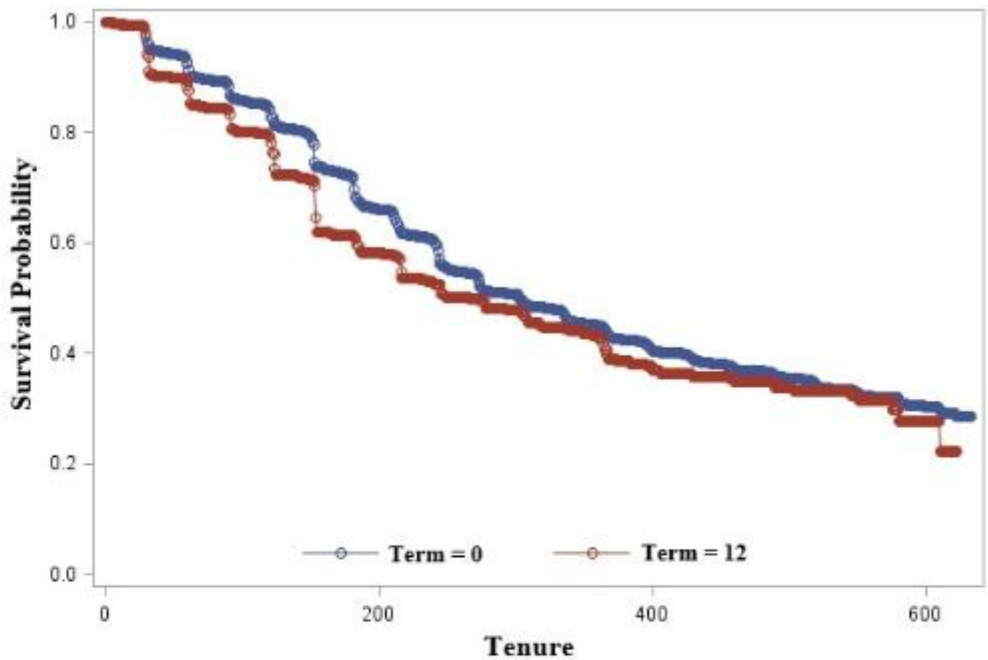


Figure 10: Survival probability plots for the three strata strategies



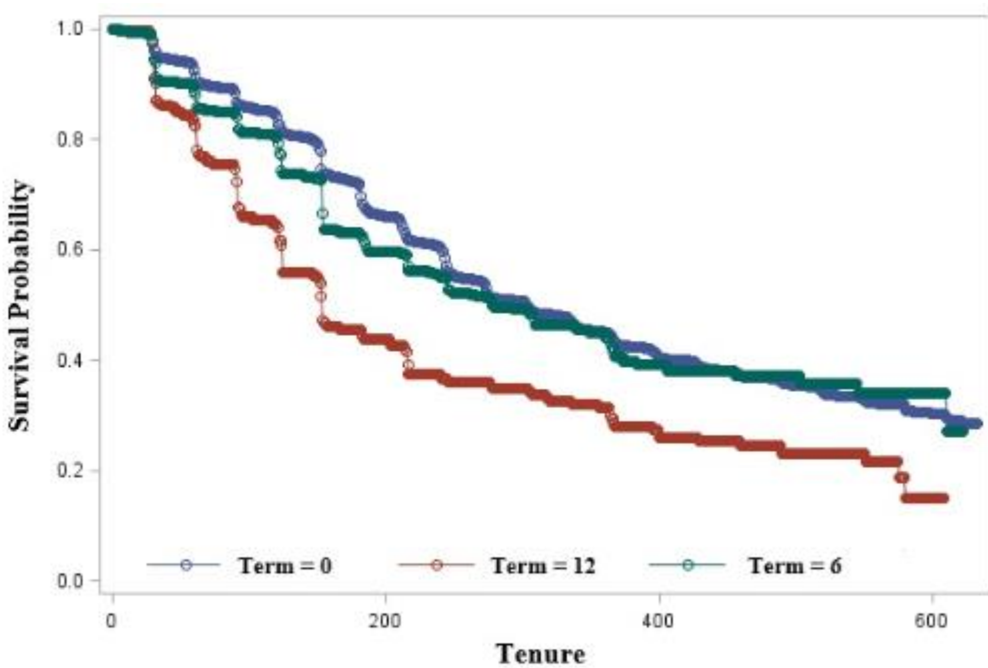


Figure 10 (Continued)

5.4 Forecast Tenured Customer

With the hazard functions obtained above, the tenured customer forecast for each strategy using the methodology described in Section 4.2 is developed. Table 5 lists the MAPE of daily tenured customer count forecast of these three models for the validation period, while Figure 11 shows the line plots of the three forecasts versus actual customer counts for each plan. For both of the fixed and the variable plans, the best strategy, which yields the lowest out-of-sample MAPE, is to combine the two fixed-term products together. Therefore, we will apply this strategy to forecast tenured customer count in the testing period for both plans.

Table 5: Tenured customer count forecasting

	Contract Term Grouping	**MAPE (%) Fixed Plan**	**MAPE (%) Variable Plan**
SA_1	Variable => 0 6-month => 0 12-month => 0	2.54	5.69
SA_2	Variable => 0 6-month => 12 12-month => 12	**0.64**	**5.31**
SA_3	Variable => 0 6-month => 6 12-month => 12	0.76	5.31

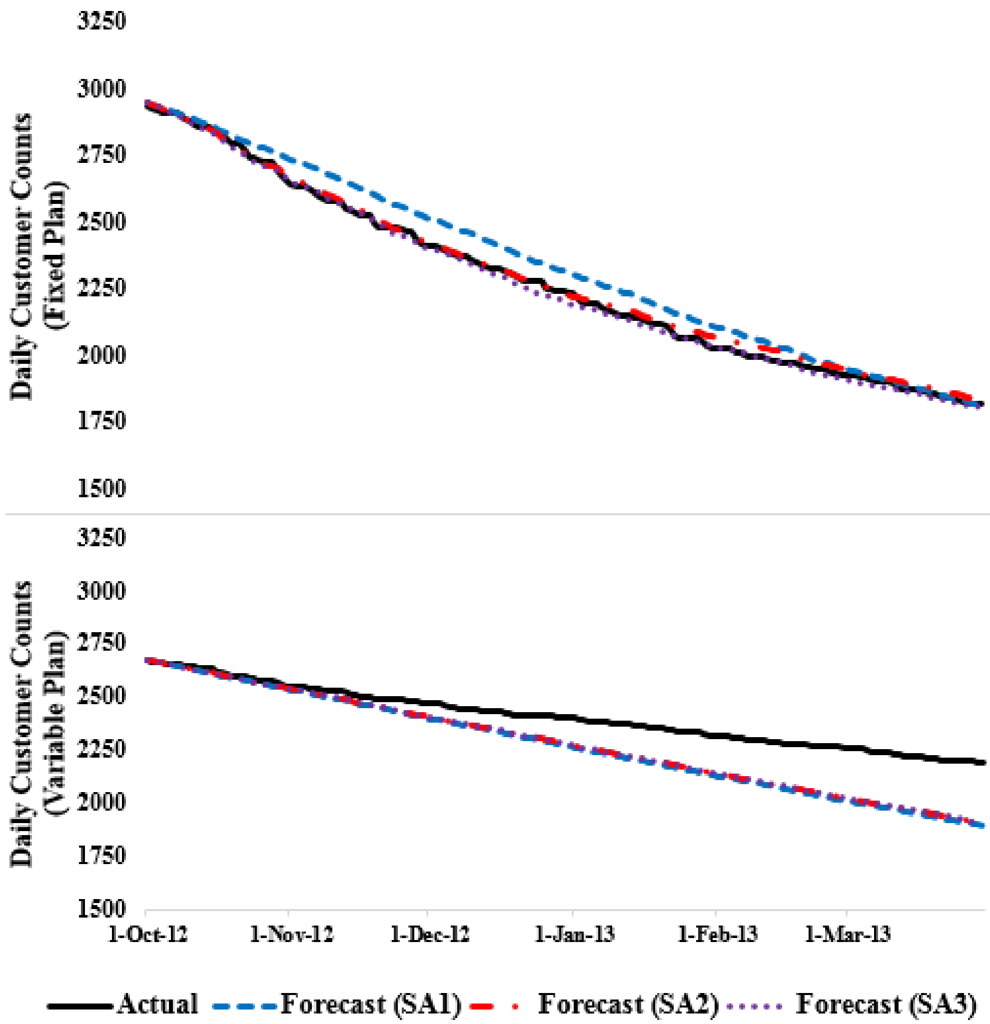


Figure 11: Tenured customer forecast

CHAPTER 6: FORECASTING RESULTS

In this chapter, the load per customer forecast model selected from Section 5.2 is applied to all the available historical data prior to the testing period to derive the load per customer forecast. Using the same historical period, the strata strategy selected from Section 5.4 is applied to derive the tenured customer forecast in the testing period. To show the effectiveness of the proposed method, the forecasting results are compared with those of a common method to forecast the count of tenured customers, which does not involve survival analysis.

6.1 A Commonly Used Method

The common method is based on the assumption that the probability that a customer with tenure t disconnects at tenure t+1 stays constant for any t in a given N-day analysis period for all customers at the beginning of this period. We denote this probability as h, which can be solved from the following equation:

$$CC_N = CC_0 * (1 - h)^N \tag{12}$$

The customer count data prior to the testing period are used to derive the survival rate. Then this daily customer survival rate is used to forecast the daily customer count for the testing period:

$$CC_t = CC_0' * (1 - h)^t \tag{13}$$

where CC_0' is the customer count at the beginning of the test period, and t = 1, 2, …, 275. Implementation of this method does not require advanced statistical software. MS Excel is capable enough to calculate h and forecast the daily customer count.

6.2 Comparison

Both the proposed method and the common method include the same load per customer forecasts developed using the procedure as described in Section 5.2. The resulting MAPE values of hourly load, daily energy and monthly energy in the testing period are 11.56%, 10.03% and 7.75% respectively. Figure 12 shows the daily energy per customer forecast versus the actual daily energy per customer for the testing period.

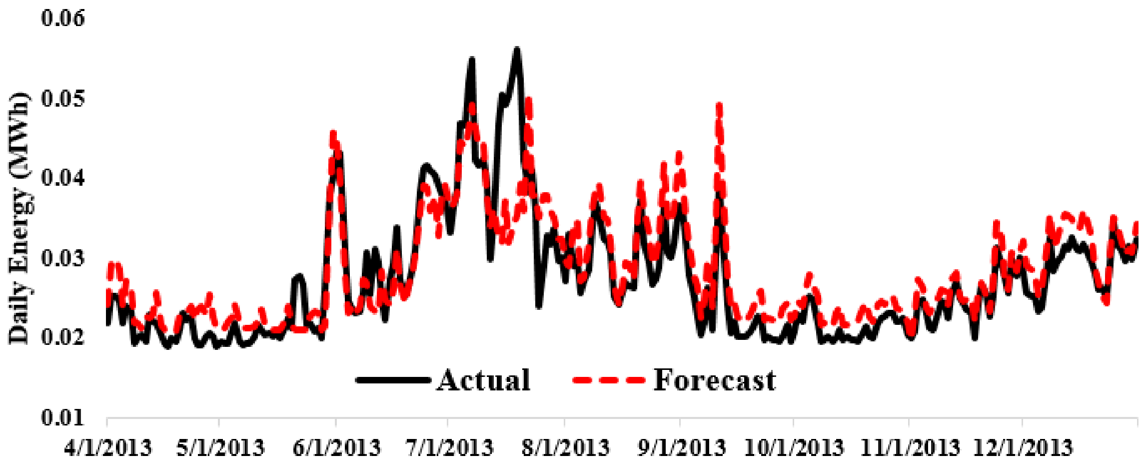


Figure 12: Daily energy per customer

Figure 13 shows the comparison between the tenured customer forecast based on survival analysis and the one based on the common method. For the survival analysis method, the MAPE values of daily customer count are 8.99% and 10.31% for the fixed plan and the variable plan respectively.

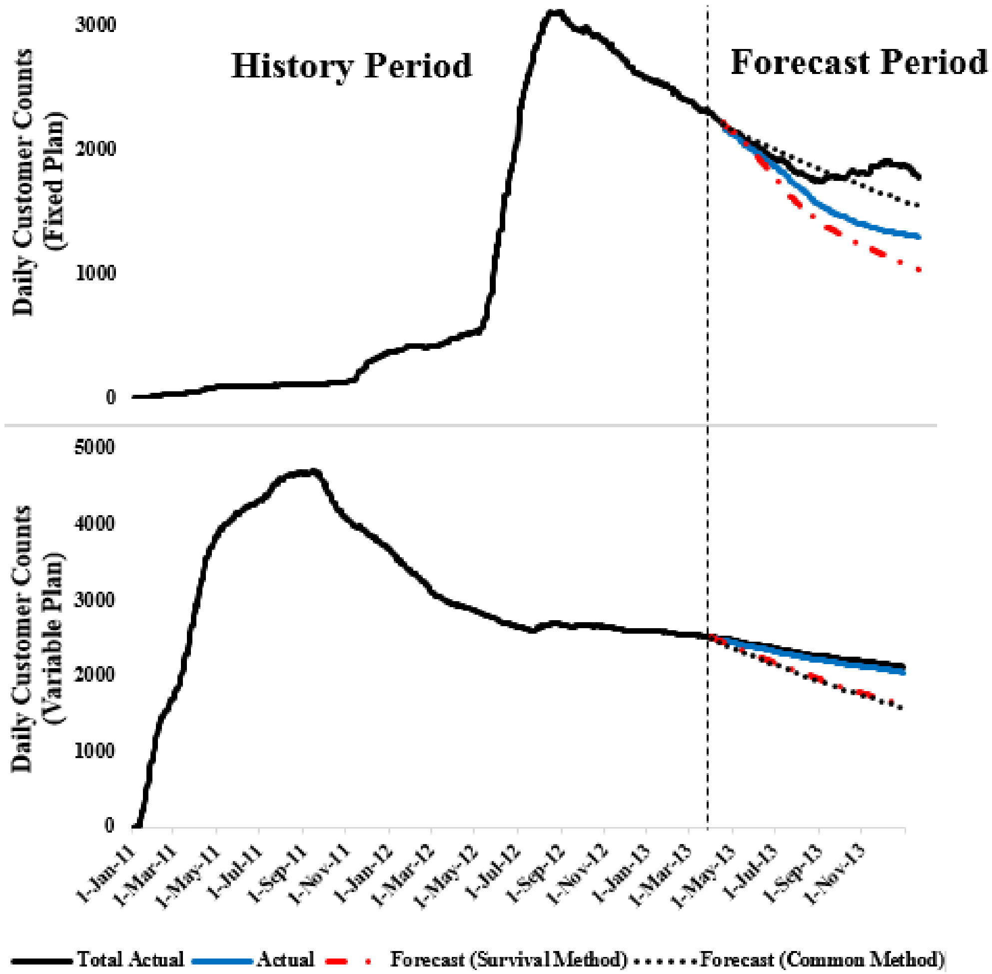


Figure 13: Tenured customer forecast

By multiplying the load per customer forecast with the tenured customer forecast, we can obtain the load forecast of those tenured customers. Figure 14 shows the actual monthly energy and the forecasted ones from the survival analysis method and the common method.

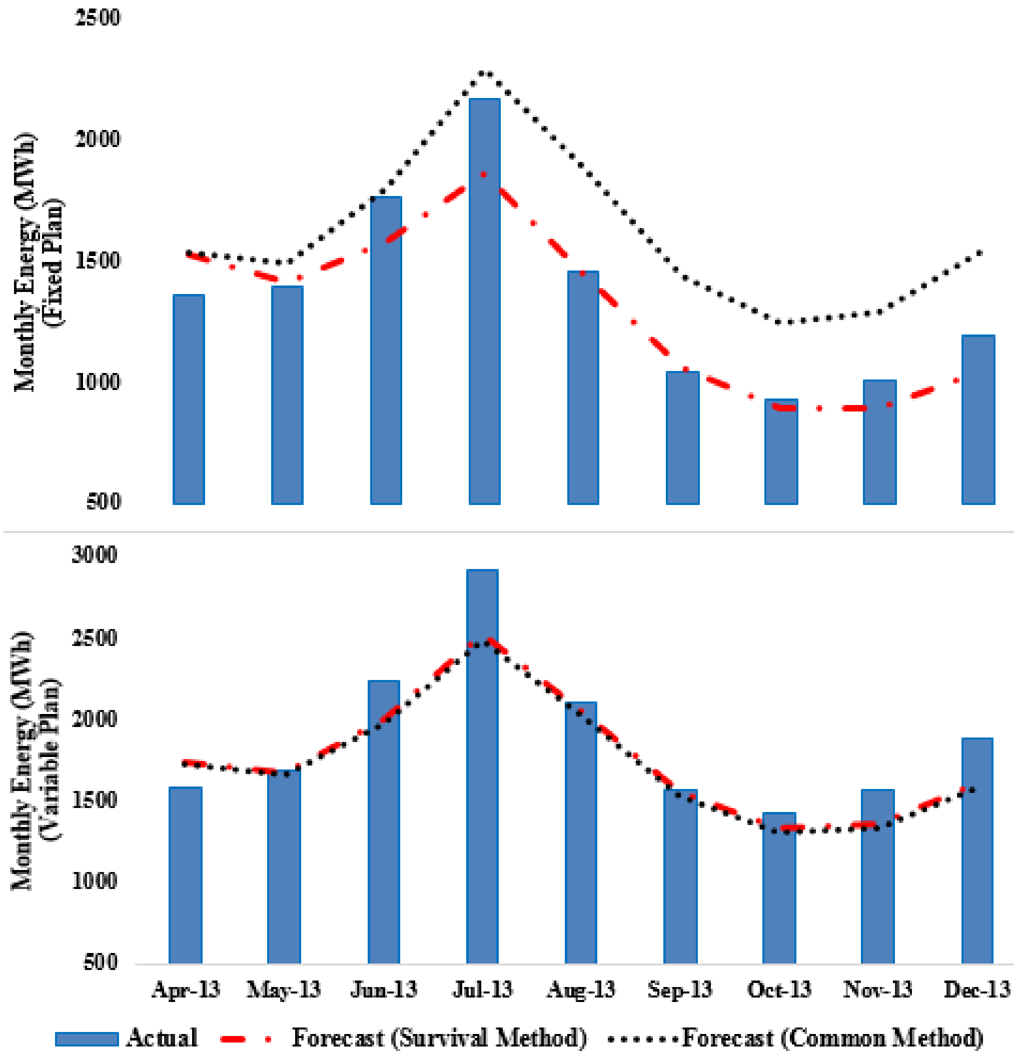


Figure 14: Comparison of actual and forecasted monthly energy

6.3 Discussion

The forecasting results for Z1 are summarized in Table 6, which shows that the proposed method is dominantly better than the common one on the accuracy of tenured customer forecasting and load forecasting for both plans. There are two other points that worth highlighting.

Firstly, the error of the load forecast is lower than the multiplication of the error of the load per customer forecast and that of the tenured customers forecast. In other words, multiplying the forecasted values does not imply multiplying the errors.

Secondly, the errors in the testing period are significantly higher than those of the validation period. This is largely due to the significant discrepancy between the actual loads and the forecasted loads in mid-July of 2013 as shown in Figure 14. The “actual” loads during this period are not the actual observations from the meters. Instead, they are imputed by the utility during the load settlement process. Figure 15 shows the hourly settled load for a week overlaid with observed temperatures. Although the temperature profiles are varying over these days, the load profiles are identical across this period. Due to the defective load settlement process, the settled loads cannot reflect the actual energy consumption. This also leads to the large MAPE value in July 2013.

Table 6: Forecasting results for Z1

Forecast	Interval	MAPE (%) Fixed Plan		MAPE (%) Variable Plan	
		Proposed	Common	Proposed	Common
Load per Customer Forecast	Hourly	11.56	11.56	11.56	11.56
	Daily	10.03	10.03	10.03	10.03
	Monthly	7.75	7.75	7.75	7.75
Tenured Customer Forecast	Daily	**8.99**	14.45	**10.31**	11.60
	Monthly	**8.95**	14.42	**10.28**	11.57
Load	Hourly	**10.24**	24.63	**10.56**	11.18
	Daily	**8.99**	23.49	**9.51**	10.31
	Monthly	**7.45**	20.59	**7.96**	9.17

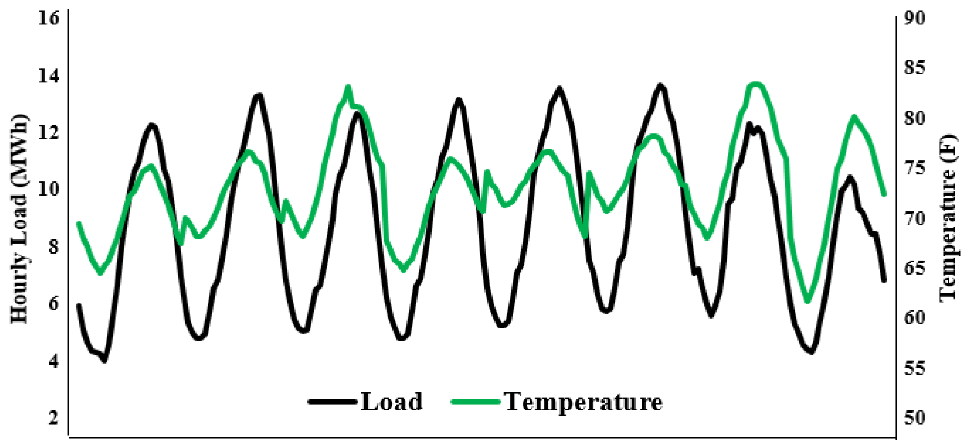


Figure 15: Hourly load and temperature (14 July 2013 – 21 July 2013)

CHAPTER 7: CONCLUSION

Load forecasting is crucial to many REPs in the course of managing their risk. There are two major challenges with retail energy forecasting: short data history and customer attrition. This paper proposed a practical methodology to forecast the load to meet a conservative trading strategy at REPs. High resolution data at hourly interval is used to model load per customer, while survival analysis is used to model customer attrition and forecast future customer count. The effectiveness of the proposed approach is demonstrated through a field implementation at a fast growing REP. Comparing with a commonly used method that assumes constant attrition rate, the proposed approach results in less MAPE of both customer count forecast and final load forecast.

As the first formal study in the academic literature that tackles retail energy forecasting and applies survival analysis to load forecasting, this paper is not trying to address all possible questions in the field. Instead, it lays the ground for several possible future research directions, such as forecasting with consideration of competition among retailers, probabilistic retail energy forecasting, and applications of other techniques in forecasting retail customers.

REFERENCES

[1] L. Philipson and H. L. Willis, "Understanding electric utilities and de-regulation," 2nd ed., CRC Press, 2006, pp. 1–28.

[2] U.S. Department of Energy, "A primer on electric utilities, deregulation, and restructuring of U.S. electric market," 2002, pp. 45–54.

[3] L. Philipson and H. L. Willis, "Understanding electric utilities and de-regulation," 2006, pp. 29–69.

[4] U.S. Department of Energy, "A primer on electric utilities, deregulation, and restructuring of U.S. electric market," 2002, pp. 55–74.

[5] T. Hong, "Short term electric load forecasting," Ph.D. Dissertaion, Graduate Program of Operation Research and Dept. of Electrical and Computer Engineering, North Carolina State University, 2010.

[6] T. Hong, "Electric load forecasting: fundamentals and best practice." [Online]. Available: https://support.sas.com/edu/schedules.html?id=1326.

[7] D. Kirschen and G. Strbac, "Fundamentals of power system economics," John Wiley & Sons, Ltd, 2004.

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

[9] T. Hong, "Energy forecasting : past , present , and future," *Foresight:The International Journal of Applied Forecasting*, no. 32, pp. 43–48, 2014.

[10] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, 2001.

[11] T. Hong, P. Wang, and H. Willis, "A naïve multiple linear regression benchmark for short term load forecasting," *2011 IEEE Power and Energy Society General Meeting, Detroit, MI.*

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

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

[14] T. Hong and P. Wang, "Fuzzy interaction regression for short term load forecasting," *Fuzzy Optimization and Decision Making*, vol. 13, no. 1, pp. 91–103, Sep. 2013.

[15] A. Motamedi, H. Zareipour, and W. D. Rosehart, "Electricity price and demand forecasting in smart grids," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 664–674, 2012.

[16] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357–363, Aug. 2013.

[17] N. Charlton and C. Singleton, "A refined parametric model for short term load forecasting," *International Journal of Forecasting*, vol. 30, no. 2, pp. 364–368, Apr. 2014.

[18] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models," *International Journal of Forecasting*, vol. 30, no. 2, pp. 375–381, Apr. 2014.

[19] J. R. Lloyd, "GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes," *International Journal of Forecasting*, vol. 30, no. 2, pp. 369–374, Apr. 2014.

[20] S. Ben Taieb and R. J. Hyndman, "A gradient boosting approach to the Kaggle load forecasting competition," *International Journal of Forecasting*, vol. 30, no. 2, pp. 382–394, Apr. 2014.

[21] W. Charytoniuk, E. D. Box, W.-J. Lee, M.-S. Chen, P. Kotas, and P. Van Olinda, "Neural-network-based demand forecasting in a deregulated environment," *IEEE Transactions on Industry Applications*, vol. 36, no. 3, pp. 893–898, 2000.

[22] Z. S. H. Chan, H. W. Ngan, a. B. Rad, a. K. David, and N. Kasabov, "Short-term ANN load forecasting from limited data using generalization learning strategies," *Neurocomputing*, vol. 70, no. 1–3, pp. 409–419, Dec. 2006.

[23] M. S. Kandil, S. M. El-Debeiky, and N. E. Hasanien, "Long-term load forecasting for fast developing utility using a knowledge-based expert system," *IEEE Transactions on Power Systems*, vol. 17, no. 2, pp. 491–496, 2002.

[24] "Survival Analysis : Introduction," *American Statistics Association, http://amstat.org/chapters/northeasternillinois/pastevents/presentations/summer05_Ibrahim_J.pdf*, 2005.

[25] G. S. Linoff and M. J. A. Berry, “Data Mining Techniques: for marketing, sales and customer relationship management,” Third Edit., Indianapolis, IN: Wiley Pub., 2011, pp. 357–394.

[26] B. Larivière and D. Van den Poel, “Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services,” *Expert Systems with Applications*, vol. 27, no. 2, pp. 277–285, Aug. 2004.

[27] D. Van den Poel and B. Lariviere, “Customer attrition analysis for financial services using proportional hazard models,” *European Journal of Operational Research*, vol. 157, no. 1, pp. 196–217, Aug. 2004.

[28] M. Guillén, J. Nielsen, T. H. Scheike, and A. M. Pérez-Marín, “Time-varying effects in the analysis of customer loyalty: A case study in insurance,” *Expert Systems with Applications*, vol. 39, no. 3, pp. 3551–3558, Feb. 2012.

[29] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied Regression Analysis A Regression Tool*. Springer, 2001.

[30] “PROC LIFETEST Online Reference,” *SAS Institute Inc, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_lifetest_sect004.htm%20*.

[31] L. J. Tashman, “Out-of-sample tests of forecasting accuracy: an analysis and review,” *International Journal of Forecasting*, pp. 437–450, 2000.

[32] “PROC GLM Online Reference,” *SAS Institute Inc, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#glm_toc.htm*.

[33] P. D. Allison, *Survival Analysis Using SAS: A Practical Guide*, Second Edi. Cary, NC: SAS Institute Inc., 2010.