ABSTRACT

SINGH, URVIR. Load Estimation for Distribution Feeder Monitoring and Management. (Under the direction of Mesut E Baran).

Load Estimation is an indispensable tool for distribution system studies, since knowledge of load profiles along the feeder has direct influence on system planning and operation activities. The main difficulties in the load modeling result from the random behavior of loads, diverse load shapes at customer sites, limitation and uncertainty in the information on loads.

This thesis explores a new technique of load modeling and estimation on distribution systems. With the AMI technology on the distribution systems, real-time data about customer loads would be available at the control center, hence an estimate of loads on the distribution feeder can be made. With this estimation and the temperature forecast, a load model predicting the real-time load variations, can be made. This thesis elaborates the statistical approach used to build such a harmonics-based model with auto-correlated errors (a time series model).

A time series approach to model and predict the random behavior of distribution feeder loads is explained, by harmonically decomposing the seasonal and daily variation of load consumption. With the historical power data of residential and commercial class

available, statistical tools are used to perform load estimation on distribution feeder using SAS (Statistical Analysis System).

Various load data sets can be grouped or clustered together, using available ‘clustering analysis’ techniques. The data of a meter whose readings are not available at any time instant can be estimated using the proposed time series method and other available meter readings from its respective cluster.

Load Estimation for Distribution Feeder Monitoring & Management

by

Urvir Singh

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Electrical Engineering

Raleigh, North Carolina

2010

APPROVED BY

_______________________________ _______________________________

Dr. Winser E Alexander Dr. Kimberly Weems

____

Dr. Mesut E Baran
Chair of Advisory Committee

BIOGRAPHY

Urvir Singh was born on 22 July 1986 in Allahabad, INDIA. He spent his early childhood in Leuven, Belgium where his father was attending graduate school to study Engineering. He did his schooling in Chandigarh (Punjab); Lucknow (Uttar-Pradesh) and Mumbai .He received his Bachelor of Technology (B.Tech) degree in Electrical Engineering from Veermata Jijabai Technological Institute, VJTI (formerly Victoria Jubilee Technical Institute), Mumbai in 2008. Immediately after finishing undergraduation, he joined North Carolina State University to pursue Master of Science in Electrical Engineering under the guidance of Dr. Mesut Baran.

ACKNOWLEDGEMENTS

I express my deepest gratitude towards the following people who have helped and guided me throughout my course of study.

To Dr Mesut Baran, an ideal advisor who has extensive knowledge in all the areas of Electrical Engineering, that helped me greatly to perform my research. My heartfelt thanks to you for your support and guidance throughout the time I was at NC State.

To Dr Kimberly Weems and NC State’s statistics department, who provided valuable help throughout the course of thesis on Statistical topics.

To my friends: Sumit Dutta, Shane Hutchinson, Hossein Hooshyar and Zhuoning Liu. Thank you guys, I would always cherish your friendship.

My family-my dear parents and brother, whose love has always supported and guided me throughout my life.

In Waheguru’s name, I present my thesis:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

One of the main components of distribution automation is real-time monitoring and control of distribution-level circuits. To achieve the goal of real-time monitoring and control, a distribution circuit state estimator tool which can provide real-time estimates of the states of the system is required. Due to the limitation on the availability of the real-time measurements on the distribution systems, a load modeling technique is required, which can provide real-time estimates of customer load demands.

With the development of distribution automation (DA) and other advanced applications in distribution systems, the real-time monitoring and control of distribution systems becomes possible. Now there are only a limited number of real-time measurements on the distribution systems. The load monitoring and estimation of customers can be an important source of information used by the distribution analysis applications. In recent years, an increasing number of automated meter reading (AMR) systems have been installed. AMR can provide customer consumption information and other data such as confirmations for outages and restoration.

Load Estimation is a challenging field because of the size and complexity of distribution systems and the features then distinguish them. The modeling techniques, which can be found in literature, extend from being simple approaches (based on assumptions such as load (power) is directly proportional to kWhr consumption / transformer kVA ratings) to more statistically intensive oriented methods.

Any practical distribution state estimator needs a load modeling technique that can provide load estimates, where the measurements are not available. Because of the limited real-time measurements in the distribution systems, the state estimator cannot acquire enough real-time measurements, so pseudo-measurements are necessary for a distribution system state estimator. Since most of the load modeling techniques rely on just the historical kWhr data or static data (kVA ratings) to predict the load on the system. Any load-modeling technique which models the load, considering the real time system data and weather conditions also, is bound to be more accurate the former.

1.2 Related Work

Various load-modeling techniques found in literature were reviewed .They range from simple estimates for simple planning purpose using transformer kVA and billing kWh to more sophisticated approaches for operation studies which take advantage of statistical analysis techniques, power flow tools and available SCADA information.[10][11][12][13][14]

A major drawback of traditional load modeling procedures has been their inability to provide a measure of uncertainty regarding its estimates. Lubkeman [2] proposed a probabilistic load modeling technique, based on daily load curves, illustrating need for the time of day dependency. Time of day variation is incorporated by building daily load curves. Charytoniuk and Chen [6] discussed the application of non-parametric probability density estimation to the problem of customer demand forecasting, using the data available at utilities. They use the demand survey information (energy data of a sample number of customers) and temperature conditions to build a probabilistic model, which denotes both the random nature of demand and its temperature dependence. The main input is the energy usage and outside temperature. The accuracy of forecast depends upon the quality of customer classification, size of sampling populations and composition and size of the estimated group. Chen, Hwang

[8] and others proposed a load survey system to determine the load characteristics of various classes, served by a utility company, followed by a statistical analysis on the acquired data to build a power consumption model of each class.

The number of AMR systems on the distribution side has been increasing, and a method is proposed by Schulz and Wang[1] , based on AMR data and customer class curves. They suggest a procedure to estimate the real and reactive loads at various nodal points (where distribution transformers are connected), in the distribution system, with data (kWh) available from AMR technology. Energy meters would transmit meter readings at various intervals. Average real-time power at time ‘t’ can be estimated, based on two consecutive meter readings and the time interval between them. The shorter the interval, the better the estimation of real time power.

1.3 Thesis Objective

Usually on the distribution feeder, measurements for all the loads are not available all the times (meters are not installed at all the customer sites or due to some meter failures). Hence in that case a load estimation technique is required which can estimate the missing data about the customers. A load estimation technique aims to model and predict those missing values based on the available historical data from those customers and other real-time data pertaining to those factors, which influence power consumption.

In this thesis, a time series approach to model and predict the random behavior of distribution feeder loads is explained, for load-monitoring purposes. With the historical power data of residential and commercial class available, statistical tools are used to perform load estimation of meters on distribution feeder.

Observing the trend of changes occurring in a day and seasonal changes, we propose (hypothesize) a model using harmonic components based on the concept of Fourier series. The power consumption variations would be broken down into its harmonic parts.

Furthermore, application of a clustering algorithm is presented, which is used to cluster or group customers based on similar consumption pattern.

Consider a line section on a distribution feeder with 'x' customers. On a specific day and time, usually all the customers data is not known. In that case, a load estimation technique helps to estimate the missing customer data. Based on the historical data of all the customers, available with the utility, a 'clustering' algorithm can be used to find groups or 'clusters' of customers. There would be some real time data (AMI) available within each cluster and some of the data may be unavailable. So within each cluster the Load Estimation technique can be used to predict the missing customer's data, using the available data within that cluster.

Consider a 33 node sample feeder (Figure 1);

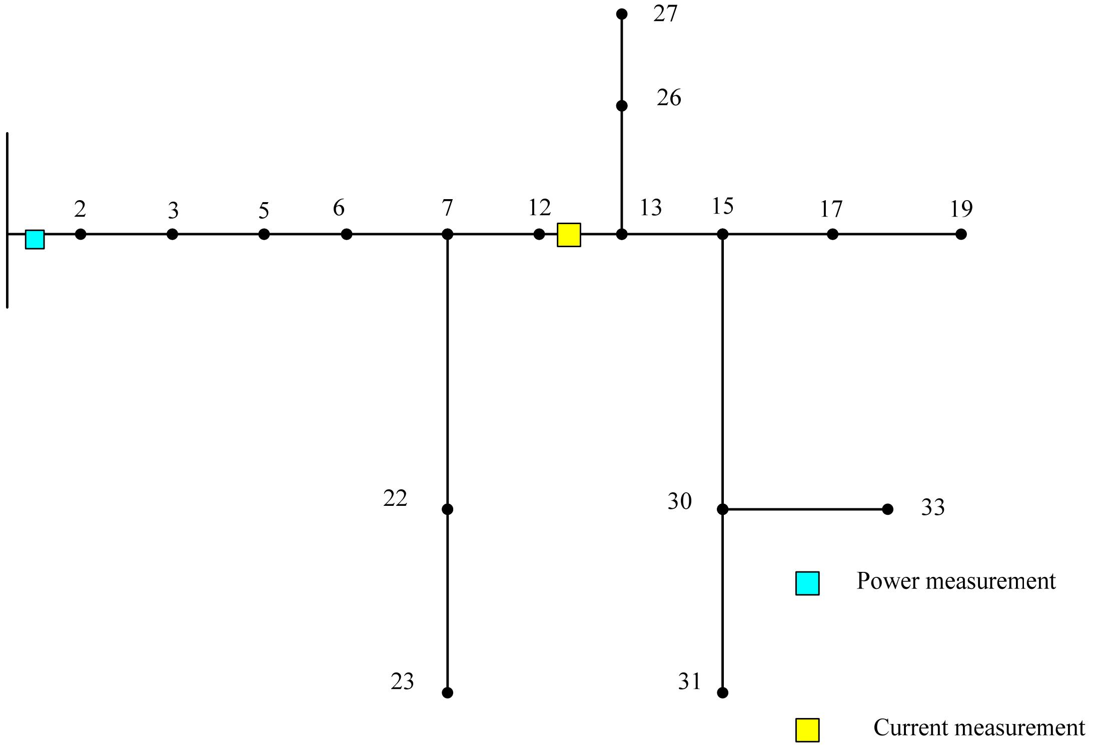


Figure 1. Sample feeder

For example: the data of some of the customers on the line section between node 2and 3 (real time power consumption) would be not available due to a failed meter or any other reason. A 'clustering technique' (Chapter 4) would be implemented to cluster these customers into various homogenous groups. In a particular cluster, there may be some customers with missing data. The load modeling technique proposed in Chapter 3 would used to estimate the missing data of the affected customers from their historical data and other customers in their cluster.

1.4 Organization

The thesis is organized as follows. In Chapter 2, a method developed by Noel Schulz[1] is implemented and extended using statistical tests. The proposed algorithm makes use of the information that AMR provides as its input. It also incorporates the use of the basic customer class load curve to improve the accuracy of individual customer real-time load estimates. In Chapter 3 a novel method of load modeling based on harmonic decomposition of power consumption and use of an autocorrelation model to predict the load consumption is described and implemented. Chapter 4 focuses on ‘load clustering’ i.e. grouping consumers based on similarities. The load modeling technique proposed in Chapter 3 is modified to estimate the missing data of the affected customers from their historical data and real-time data of other customers in their cluster.

1.5 Glossary

AMI : Automated Meter Infrastructure is an intelligent technology that includes metering systems capable of recording and reporting energy consumption and other measurements at more frequent intervals that the customer’s billing cycle.

AMR : **Automatic Meter Reading** is the technology of automatically collecting consumption, diagnostic, and status data from metering devices (water, gas, electric) and transferring that data to a central database for billing, troubleshooting, and analyzing.

SCADA : Supervisory Control And Data Acquisition systems are used to monitor and control power system in a wide range of applications like power station control, transmission, distribution automation.

SE : State Estimation as a mathematical analysis tool acts as a noise filter to eliminate errors in data. In practices, other conveniently measured quantities such as P, Q line flows are available, but they cannot be used in conventional power-flow calculations. these limitations can be removed by state estimation.

SAS: 'Statistical Analysis System', statistical software used to perform regression analysis and implement the clustering algorithm.

Chapter 2

Using 2 days AMR data for load estimation

In this chapter the method developed in [1], which estimates the missing load of a meter based on the latest available AMR data of that meter and its historical power consumption pattern is implemented and extended through statistical tests. These estimates of load can be used as pseudo-measurements to account for the meters whose readings are not available. The proposed algorithm makes use of the information that AMR provides as its input. It also incorporates the use of the basic customer class load curve to improve the accuracy of individual customer real-time load estimates. This method demonstrates how AMR data can be used for other functions besides billing.

In recent years an increasing number of Automated Meter Reading (AMR) systems have been installed, which can provide consumer consumption information and other useful data such as outages and restorations. The collection of data is done remotely over telecommunication, power, radio lines etc. This information can be useful while developing any load modeling algorithm.

With yearly power (residential & commercial) data available from Pacific Gas & Electric Company, for a specific location, following procedure was followed (on the lines of the above proposed method):

2.1 Method :

- Generate the load curve – If the meters (installed at various customer location) transmit energy with an interval of time of Δt,

$$P_t = \frac{kWh_t - kWh_{t-\Delta t}}{\Delta t}$$

Where,

Δt is the time period

kWh_t is the meter reading at time t

So, as the time period between two-meter readings Δt is reduced, Pt would closely approximate the real time power. This procedure is applied on many customers, individual load profiles are obtained, and then averaging all of customer load profiles, a general class based load curve can be generated.

- Generate the real load for two days –The load values on the curve are the mean values and assuming normal distribution for the loads, so along with the standard deviation σ, two day's load is generated.

- Generate the meter readings for two days-Accumulate the load for some time interval Δt and get the meter consumption reading kWh

- Estimate the load for two days –If the data from the meters can be transmitted to the utility 'n' times per day, then one day can divided into 'n' intervals. If the time of

$$kWh_{i,today} = \frac{kWh_{i-1,today}}{kWh_{i-1,day\_before}}.kWh_{i,day\_before}$$

Here day before is the day before today in weekday or weekend.

Then the estimated load is

$$P_t = \frac{kWh_{i,today}}{S_i}.P_{i,norm}$$

Here,
S_i is the kWh of interval i in normalized load curve
P_t is the power of time t in normalized load curve

If there is an outage, then $P_t = 0$

- The error in the estimation procedure is reflected by the RRMSE :

$$\text{RRMSE=}\sqrt{\frac{\sum_{t=1}^{1440}(\hat{P}_t - P_t)}{\sum_{t=1}^{1440}(\hat{P}_t)}}$$

Here,
$\hat{P}_t$ is the simulated actual load at time 't'
P_t is the estimated load at time 't'

the number of terms in the summation are 1440 if loads are generated/estimated at one-minute interval.

2.2 Implementation

With yearly power (residential & commercial) data available from Pacific Gas & Electric Company, for a specific location, following procedure was followed (on the lines of the above proposed method):

- Actual residential and commercial load data was collected .(source-Pacific Gas & Electric company)
- Load profiles corresponding to two residential and one commercial class was generated based on the above collected data, which shows the expected value along with the standard deviation, for 30 minute intervals in a day.
- For testing purposes, the load data was generated for customers of these three classes, based on the mean and standard deviation values available from the load data.
- Energy data (which would be available through AMR technology, in actual implementation) is generated for these two days, at 30 minutes and 10 minutes intervals.

- If the data from AMR is available ‘n’ times in a day, then ‘n’ intervals are present in day and if the time of estimation is in the i’th interval then

$$kWh_{i,today} = \frac{kWh_{i-1,today}}{kWh_{i-1,day\_before}}.kWh_{i,day\_before}$$

,where day_before is the day before today in either weekday/weekend

Estimated load (on the assumption that if the interval energy is same for today and the day before, then the load ‘P ’follows the same pattern as the load curve of the class, to which the customer belongs, and is given by:

$$P_t = \frac{kWh_{i,today}}{S_i}.P_{i,norm}$$

where Si is the kWh is the i’th interval

Pt is the estimated load at time ‘t’

For our case n = t, i.e the time of estimation is same as the time of AMR data recording

Following RRMSE values were obtained for 30 minutes and 10 minutes interval for varying number of customers:

Table 1. 30 Minute Interval (one residential class only)

Number of customers	Average RRMSE (100 runs)
1	0.1869858
2	0.185000095
5	0.184372514
10	0.184398356
15	0.184777484
20	0.185538033

Table 2. 10 Minute Interval (one residential class only)

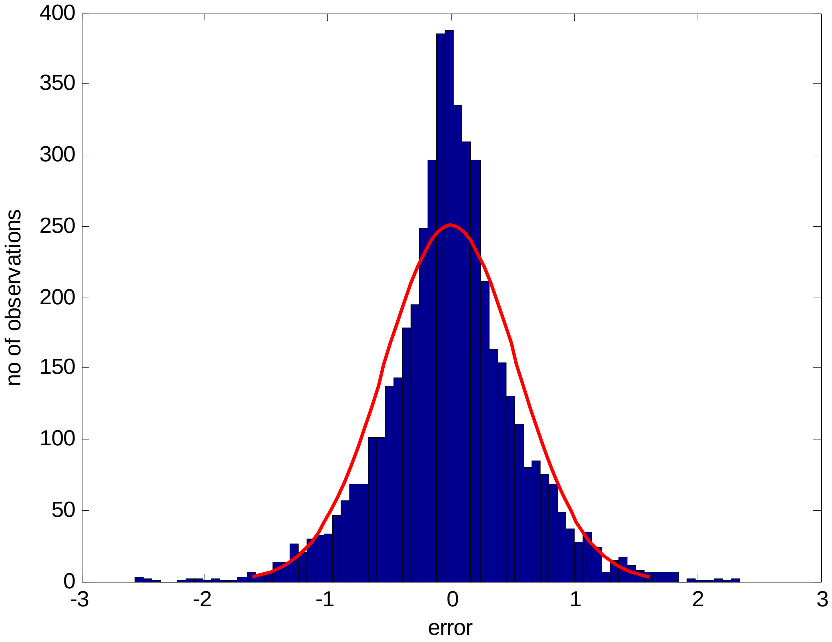
Number of customers	Average RRMSE (100 runs)
1	0.1869858
2	0.185000095
5	0.184372514
10	0.184398356
15	0.184777484
20	0.185538033

The number of customers was increased, but there was insignificant effect on the RRMSE, it remained around 18.5% throughout. Also, decreasing the time-interval had no effect on the error, which was expected as all the 10 minute values were obtained from interpolation of the 30 minute interval

2.3 Statistical tests (Extension)

To check if the estimates are biased or not, a histogram of the errors between the estimated values and the actual values of load is plotted and as seen from the figure 2. the errors are biased, they don't fit in a normal distribution. The mean of the errors was found to be -.0053, which is towards the negative side and also the distribution doesn't follow the normal distribution completely.

Figure 2. Histogram of errors



The estimated vector is an unbiased estimate of the actual values if the expected value of the estimated vector is equal to the actual vector.

$$E(\hat{x}) = x_t ,$$

Where $\hat{x}$ is the vector of the estimated load values

x_t is the vector of the actual load values

‘t’ test :

One of the uses of a t-test is to determine whether the means of two groups (populations) are statistically different from each other. The test statistic follows a students ‘t’ distribution , if the null hypothesis is true, where students ‘t’ distribution is a probability distribution ,which arises while estimating the mean of a normally distributed population ,**when the sample size is small**, (n<30).

As the sample size increases, the ‘t’ distribution approaches a normal distribution, and it doesn’t matter whether to use a ‘Z’ test or a ‘t’ test.

Confidence interval for μ1-μ2 (population variances different and unknown): An approximate (1-α) 100% confidence interval for μ1-μ2 is ,

$$(x1 - x2) - t_{\alpha/2}.\sqrt{(\frac{s_1^{\ 2}}{n1} + \frac{s_2^{\ 2}}{n2})} < \mu_1 - \mu_2 < (x1 - x2) + t_{\alpha/2}\sqrt{(\frac{s_1^{\ 2}}{n1} + \frac{s_2^{\ 2}}{n2})}$$

Where,
x_1 and $s_1^{\ 2}$ are the mean and sample variance from population 1

x_2 and $s_2^{\ 2}$ are the mean and sample variance from population 2

$t_{\alpha/2}$ is the value of the 't' distribution with degrees of freedom 'v' given by the expression

$$\frac{(\frac{s_1^{\ 2}}{n1} + \frac{s_2^{\ 2}}{n2})^2}{\{[\frac{(s_1^{\ 2}/n_1)}{n_1 - 1}] + [\frac{(s_2^{\ 2}/n_2)}{n_2 - 1}]\}}$$

The expression is always > 29 hence a value of tα/2 (v=infinity)= 1.645

Samples from two populations are taken, and based on the statistics of those samples, conclusions about the parameters of the parent populations can be made (with some degree of confidence). There are two approaches to test a research hypothesis:

1) Calculating the statistic and compare it with a threshold value of the test statistic (based on a particular value of level of confidence), to either accept or reject the hypothesis.
2) Compute an interval for the difference of population means (based on some level of confidence) and check if the desired difference (stated in the research hypothesis) is

In our case, we follow the second approach as follows:

- The two populations are the –

  Population1 -measurement data (AMR data) of whole class

  Population 2 -estimated data of whole class

  At any time‘t’

- Mean of 20 customer’s measured data is one sample of population 1
- Mean of 20 customer’s estimated data is one sample of population 2
- If a 24 day is divided into 30 minutes interval, then we would have 49 such samples, and ‘t’ test can be applied at all these intervals, as follows :
- A (1-α)*100% confidence interval for μ1-μ2 is given by,

$$(x1-x2)-t_{\alpha/2}\cdot\sqrt{(\frac{s_1^{\ 2}}{n1}+\frac{s_2^{\ 2}}{n2})}<\mu_1-\mu_2<(x1-x2)+t_{\alpha/2}\sqrt{(\frac{s_1^{\ 2}}{n1}+\frac{s_2^{\ 2}}{n2})}$$

Where ,

μ1-μ2 =mean difference between measured and estimated load estimates of whole class (population)

x1= mean of measured load of 20 customers value at time ‘t’ x2= mean of estimated load of 20 customers at time ‘t’

1-α =confidence coefficient

σ1=mean of standard deviation of measured data of 20 customers at time ‘t’

σ2=mean of estimated standard deviation of 20 customers at time ‘t’

n1=n2 = 20

- If the acceptable difference between the measured and estimated load values lies in the obtained confidence interval, then the estimates are acceptable

- So this method, would give an estimate of how good the load estimation process is, from a class point of view, based on samples of some customers, tested throughout a day.

2.4 Conclusion

To check the effect of number of customers on the average RRMSE, the number of customers was increased, but there was insignificant effect on the RRMSE, it remained around 18.5% throughout.

The results of testing the estimates through the‘t’ tests, are unsatisfactory, as most of the intervals do not contain ‘zero’, indicating considerable difference between the actual measurement and its estimation.

12 out of 48 tests were negative (failure of hypothesis that the means of the two populations –measured data and estimated data are same), indicating an accuracy of just 25% for the estimation process, for the ideal case (difference between measured data and estimated data being zero).

But if some difference between the measured load data and its corresponding estimation is acceptable, then the success rate of these‘t’ tests would increase considerably.

Table 3. ‘t’ test results for residential class

Lower limit of difference between means	Upper Limit of difference between means	Result (P=PASS, F=FAIL)
0.1055	0.1713	F
-0.0024	0.06	P
0.0057	0.0611	F
0.0607	0.1105	F
0.0069	0.0483	F
-0.0504	-0.0094	F
0.0087	0.0461	F
-0.037	-0.0036	F
0.0025	0.0343	F
-0.0167	0.0139	P
-0.0327	-0.0003	F
0.0162	0.0532	F
-0.0873	-0.0349	F

Table 3. (Continued)

Lower limit of difference between means	Upper Limit of difference between means	Result (P=PASS, F=FAIL)
0.0097	0.0891	F
-0.3043	-0.1985	F
-0.0347	0.0911	P
-0.2798	-0.1682	F
-0.0548	0.0376	P
-0.1244	-0.0368	F
-0.0485	0.0465	P
0.0415	0.1429	F
-0.0914	0.0096	P
0.1388	0.2524	F
0.017	0.1304	F
-0.1218	-0.0072	F
-0.0085	0.1095	P
-0.3319	-0.2029	F
-0.2022	-0.0654	F
0.0861	0.2377	F
-0.052	0.1064	P
-0.2142	-0.0366	F
0.3172	0.5028	F
-0.3069	-0.1121	F
-0.2441	-0.0565	F
0.0777	0.2811	F
-0.3204	-0.1076	F
0.5184	0.7212	F
0.0107	0.2101	F
0.053	0.2288	F
0.0489	0.2043	F
-0.215	-0.075	F
-0.057	0.0616	P
0.1984	0.3144	F
-0.3129	-0.2059	F
-0.1668	-0.0632	F
-0.0074	0.0872	P
-0.0657	0.0201	P
-0.0091	0.0643	P
-0.1287	-0.0615	F

Chapter 3

Harmonics based Time-series Model

3.1 General Multiple Regression Procedure

Multiple linear regression is a means to express the idea that a response variable ‘ y ’, varies with a set of independent variables $x_1, x_2 ....., x_m$ .The variability that the response variable y exhibits has two components: a systematic part and a random part. The systematic variation of y can be modeled as a function of x variables. The model that relates y to

$x_1, x_2....., x_m$ is called the *regression equation*. The random part accounts for the fact that the model does not exactly describe the behavior of the response variable.

Multiple regression fits a response variable y as a function of regressor variables and parameters. The general linear regression model can be seen as :

$$y = \beta_0 + \beta_1.x_1 + \ldots + \beta_m.x_m + \varepsilon \quad (1)$$

Where,

y =response variable

$\beta_0, \beta_1, ..., \beta_m$ are unknown parameters

$x_1, x_2....., x_m$ are the regressor or independent variables

ε is a random error part

Least Squares is a technique which is used to estimate the unknown parameters based on a set of observed values of these variables. The aim is to find the estimates of the parameters $\beta_0, \beta_1, ..., \beta_m$ that can minimize the sum of the squared difference between the actual values of the response variable y and the values of y that are predicted by the model (1).

The estimates of the unknown parameters $\beta_0, \beta_1, ..., \beta_m$ are called the least-squares estimates and the quantity that is minimized to find these estimates is called the 'Error sum of squares'. The whole process can be laid down in terms of the following 5 steps:

1) Identify a list of probable predictors $x's$, which could be used to build a model to predict the dependent variable.

2) Use 'step-wise regression' to identify the important independent variables out of that list. 'step-wise' regression is described as follows:

Stepwise Regression: Used to identify the important independent variables out of many given variables, to be used to construct the model.

The user first identifies the variables $x_1, x_2, ..., x_k$

First the computer fits all the possible one-variable forms of the form

$E(y) = \beta_0 + \beta_1 . x_i$

For each model the test of

Ho: β_1=0

Ha: β_1 ~= 0

Is carried out and the variable which produces the largest 't' value is considered as the best one variable predictor of y and is included in the model.

Now the test is done on the model

$E(y) = \beta_0 + \beta_1.x_1 + \beta_2.x_i$ with k-1 option for x_i

‘t’tests are again performed again and the variable which produces the largest ‘t’ value is retained

the better software packages go back and check if the ‘t’ value for β_1 has changed ,if yes ,again the search is made.

Such a search is made until all the x’s with significant‘t’ values are identified

3) Based on them model hypothesized after the stepwise regression, subject it to the least squares process in SAS, and obtain the estimates of β parameters . R^2 goodness of fit test should be used to check how good the model is, in predicting the dependent variable. value of R^2 indicates what percent variation in the dependent variable ‘y’ is explained by the model. R^2 is a sample statistic ,so a more formal ,statistical test of hypothesis is used to check the correctness of model

4) Perform the ANOVA ‘F’ test, to check the adequacy of the correctness of the model.

Testing the utility of a model –F test

Ho: $\beta_0 = \beta_1 = \beta_2 = ... = \beta_k = 0$

Ha: atleast one of the parameters differs from zero

Rejection region: $F > F_{\alpha}$

Degrees of freedom, numerator = k, denominator = (n - k+1)

Test statistic: F = $\dfrac{\text{mean square for model}}{\text{mean square for error}}$

$$= \frac{[SS(\mathrm{mod}\,el)/k]}{[SSE/(n-(k+1))]}$$

If Ho is accepted, then hypothesize another model, else conduct t test on those β which seems more relevant (usually the higher order terms)

5) Check if certain terms in the proposed model are required or not, example the 2nd order terms, which contribute curvature to the model, by performing individual‘t’ tests for each important β parameters.

Test of individual parameter coefficient in the model

One-tailed test

(for 2nd order curvature terms

/ negative or positive curvature)

Ho: β_i=0

Ha : β_i<0 (or β_i>0)

Test statistic t = β_i / s (β_i)

Rejection region

$t > t_\alpha$

(or $t < t_\alpha$)

Two tailed test

Ho: β_i=0

Ha: β_i ~= 0

Rejection Region

$|t| > t_{\alpha/2}$

Where,

n =number of observations,

k =number of independent variables in the model

3.2 Proposed Harmonic-Auto correlative Load Modeling technique

Energy meters can transmit data to intermediate controllers at very short time intervals. For some wireless systems, this is a one-way communication for such a transmission of data, but there is two-way communication between the controller and the utility. So, using the on demand reading of every constant time interval, we can estimate the average real-time power at time *t*

$$P_t = \frac{kWh_t - kWh_{t-\Delta t}}{\Delta t}$$

Where,

Δt is the time period

kWh_t is the meter reading at time t

Since, actual AMR data was not available to develop the model, actual power data from ‘Pacific Gas & Electric Company’ was used to develop and test the modeling technique. Another point worth noting is that only one variable (power ‘P’) would be used to build the model and predict the future consumption.

3.2.1 Principle used in the modeling

Observing the trend of changes occurring in a day and seasonal changes, we propose (hypothesize) a model using harmonic components based on the concept of Fourier series. The power consumption would be broken down into its harmonic parts.

Based on the yearly data available, following are the plots of the load data for the month of January for the residential class and the power consumption at 3 PM throughout the year. The plot clearly shows a distorted sinusoidal variation, which can be modeled as a sum of harmonic components.

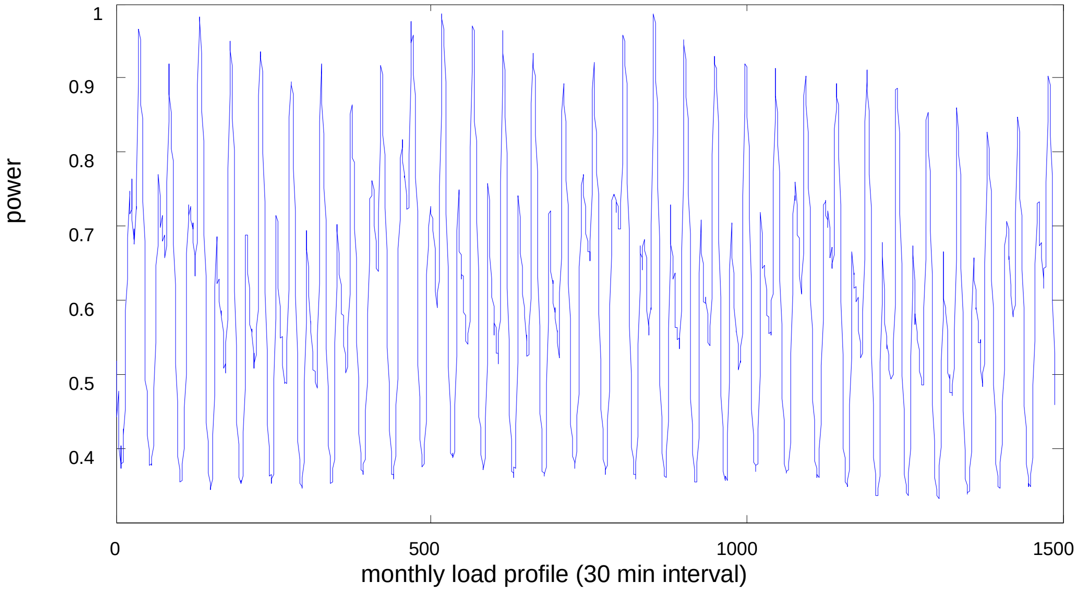


Figure 3. Monthly load data for residential class (January)

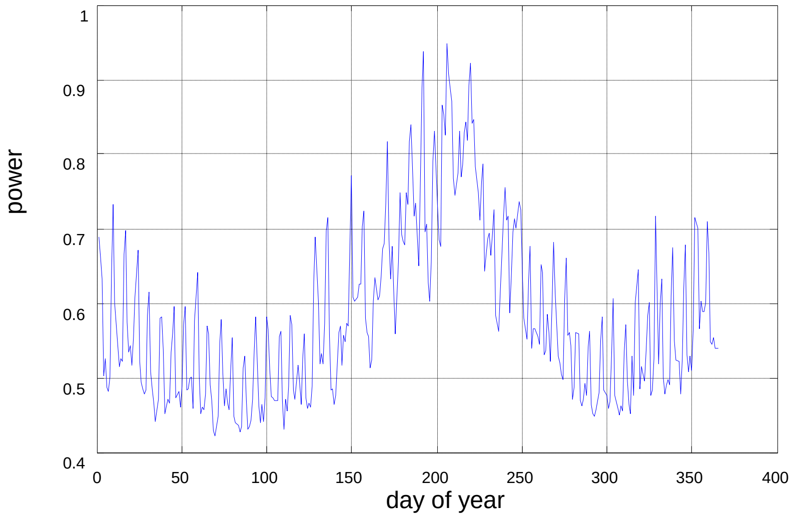


Figure 4. Yearly load data at 3 PM (residential class)

The pattern shows the increase in energy consumption in the summer months (July, August and September) and winter months (December and January). Hence, this leads to the inclusion of seasonal harmonics components in the model.

Similarly the load consumption of any random day shows a distorted sinusoidal variation, which can be modeled by harmonics. Figure 4. shows the load profile for the first day of January of the available data of residential class.

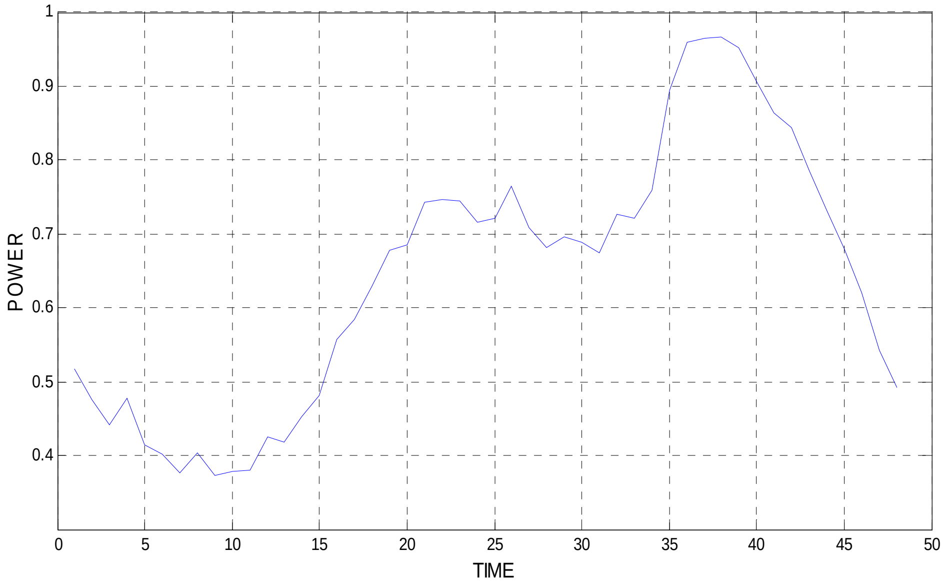


Figure 5. Load profile of January -1 (residential class)

Fourier series: An infinite series whose terms are constants multiplied by sine and cosine functions and that can, if uniformly convergent, approximate a wide variety of functions.

Since, the daily variation can be seen as a distorted sinusoid so it can be modeled using harmonics. Same for the yearly variation.

We hypothesize the model with 5 harmonics for seasonal (yearly variation) and 5 harmonics for daily variation. The reason for hypothesizing the model with 5 harmonics is as follows:

The available yearly data was fitted with the time series model for various numbers of harmonics. The results of the SAS regression output for 3, 5, 7 and 10 harmonics is attached in appendix A. The R-square value for the four settings of number of harmonics is 99.09, 99.15, 99.19 and 99.22. When the number of harmonics is doubled the change in R-square value is insignificant. Also, the value of the ‘P’ values shows that the excess seasonal harmonics have no contribution to the model. Hence hypothesizing the model with 5 harmonics seems to be an optimum value to start with. Depending on the SAS output and the ‘P’ values obtained and the R-square value of the fitted model, the number of harmonics to be used in the model can again be varied.

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i \cos(i2\pi t/48*365) + \sum_{i=1}^{5} \beta_i \sin(i2\pi t/48*365) + \sum_{j=1}^{5} \beta_j \cos(j2\pi t/48)$$
$$+ \sum_{j=1}^{5} \beta_j \sin(j2\pi t/48) + \varepsilon$$

y = power (response variable in the regressive model)

β_0 = slope of the regressive model

β_i's = unknown parameters for the harmonic components, representing the yearly (seasonal) variation.

β_j's = unknown parameters for the harmonic components, representing the daily variation

ε =uncorrelated errors.

To check if the errors in the model are correlated or not, the ‘Durbin-Watson’ test is used,

The Durbin-Watson test is a test for first-order serial correlation in the residuals of a time series regression. A value of 2.0 for the Durbin-Watson statistic indicates that there is no serial correlation.

This result is biased toward the finding that there is no serial correlation if lagged values of the regressors are in the regression. Formally, the statistic is:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^{\,2}}$$

Where the series of e_t are the residuals from a regression.

Multiple regression procedure was performed in SAS,

From the output: D (Durbin-Watson) =.0325

Since D =2 implies no correlation.

The more closer the value to 0 implies stronger positive correlation

The more closer the value to 4 implies stronger negative correlation

So in this case there is a very strong evidence of positive correlation.

Hence, 'AUTOREG' procedure is used in SAS with autoregressive model for the correlated errors a second order model is hypothesized:

$R_t = \phi_1.R_{t-1} + \phi_2 R_{t-2} + \varepsilon$

ϕ_1=First order lag

ϕ_2= Second order lag

ε=uncorrelated errors.

As seen from the output R square= .9913 for the auto correlated model, as compared to R square =.8518 of the model with uncorrelated errors.

3.3 Choice of components

The 'P' values obtained dictate which components are to be included in the model .The 'P' value (denoted by 'Pr' in the SAS output) means the probability of getting a‘t’ value greater than the threshold .Hence, with 95% level of confidence, any variable with ‘P’ value > .05 would fail to make a place in the model.

Following is the SAS output for predicting the load consumption for January 22 (Residential), based on the first 3 weeks of data of the same month. ‘s’ and ‘c’ are the seasonal components whereas ‘sd’ and ‘cd’ are the daily components. As seen clearly, the seasonal components seem to not have any regressive influence on the load consumption, which is intuitive as the data belongs to the same month, where seasonal variation would be minimum. Hence only daily-variation components would be used to predict the consumption for January 22. Including seasonal components would make the prediction results poor.

Table 4. ANOVA (Analysis of variance) table from SAS

Variable	DF	Standard Estimate	Approx Error	t Value	Pr > |t|
Intercept	1	0.6151	0.004165	147.66	<.0001
s1	1	-0.0117	0.005906	-1.98	0.0483
s2	1	0.003039	0.005904	0.51	0.6069
s3	1	0.0185	0.005902	3.14	0.0017
s4	1	0.002236	0.005899	0.38	0.7047
s5	1	0.006772	0.005894	1.15	0.2509
c1	1	-0.003158	0.005875	-0.54	0.5910
c2	1	0.0148	0.005873	2.52	0.0119

Table 4 (contd.) ANOVA (Analysis of variance) table from SAS

c3	1	-0.0161	0.005871	-2.75	0.0061
c4	1	0.008577	0.005868	1.46	0.1441
c5	1	0.0000412	0.005863	0.01	0.9944
sd1	1	-0.1590	0.005683	-27.97	<.0001
sd2	1	-0.1484	0.004969	-29.87	<.0001
sd3	1	0.0191	0.003921	4.87	<.0001
sd4	1	0.0256	0.002926	8.76	<.0001
sd5	1	-0.0178	0.002176	-8.16	<.0001
cd1	1	-0.0402	0.005656	-7.11	<.0001
cd2	1	-0.0467	0.004952	-9.44	<.0001
cd3	1	-0.0230	0.003914	-5.86	<.0001
cd4	1	0.004495	0.002924	1.54	0.1246
cd5	1	-0.002855	0.002175	-1.31	0.1897

RRMSE (as a measure of goodness of estimator):

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{48}[P(actual) - P(predicted)]^2}{\sum_{i=1}^{48} P(actual)^2}}$$

Hence, the above method seems to fit a considerably accurate model to the data and then the forecast obtained is also accurate.

3.4 Prediction using the harmonic model

On the same lines, a model is hypothesized with the monthly data of January, May and September, using only the harmonic components for daily variation, not the seasonal variation (as the 'p' values of the seasonal components make them useless in the predicting model)

Following results show the actual and forecasted power for January 22, May 22 and September 22 (Residential) of the data. Figure 5 shows the pattern of load consumption observed for the month of January, pertaining to the residential class of customers. Initially neglecting this trend, prediction for this class is based irrespective of the type of day's data, used for prediction. For the commercial class, based on the available data it was observed there is significant difference between the power consumption for weekdays and weekends/public holidays (Figure 6). Hence, for predicting the consumption of a weekday, historical data corresponding to weekdays is used. Forecast of February 2, June 8 and

October 1 is shown below along with the actual data. The RRMSE for each prediction is also shown

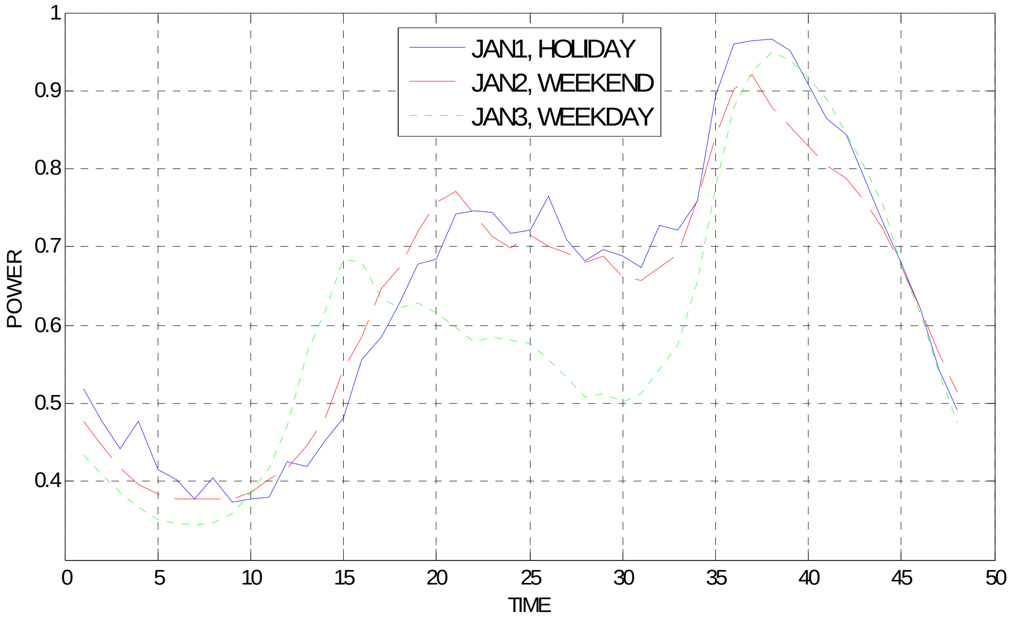


Figure 6. Consumption pattern of residential class, for various type of day

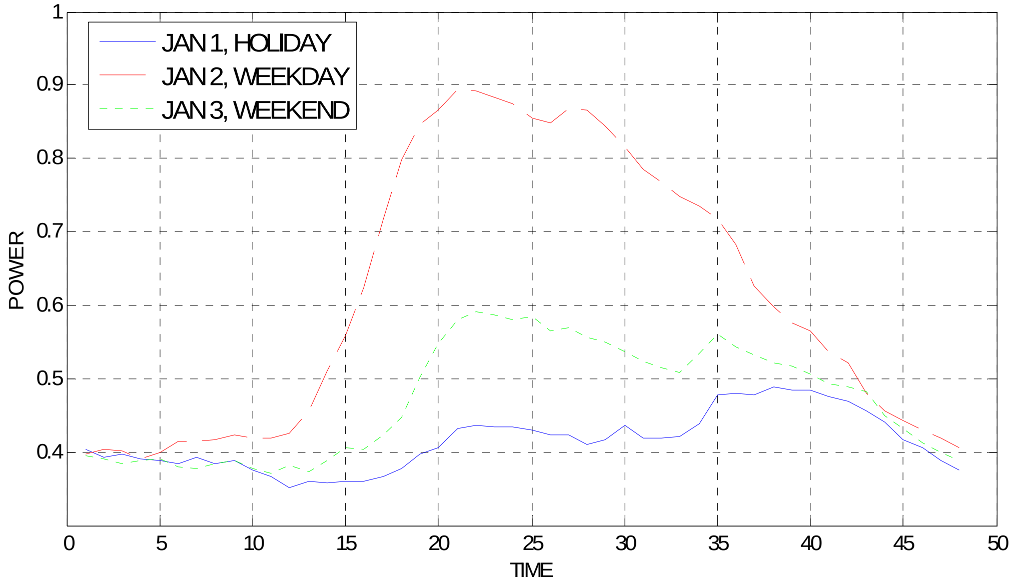


Figure 7. Consumption pattern of commercial class, for various type of day

3.4.1 Prediction Results and corresponding RRMSE values

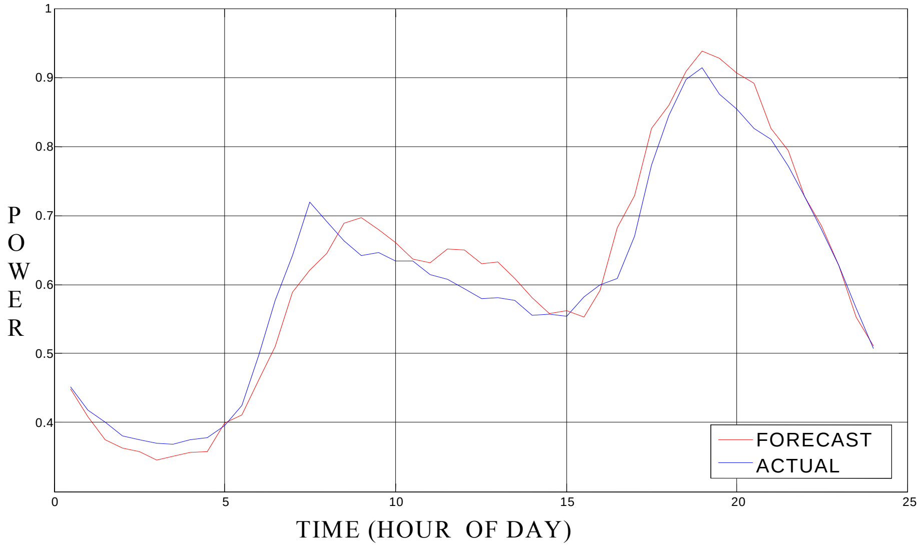


Figure 8. Forecast of January 22 (RRMSE =.0561),Residential class

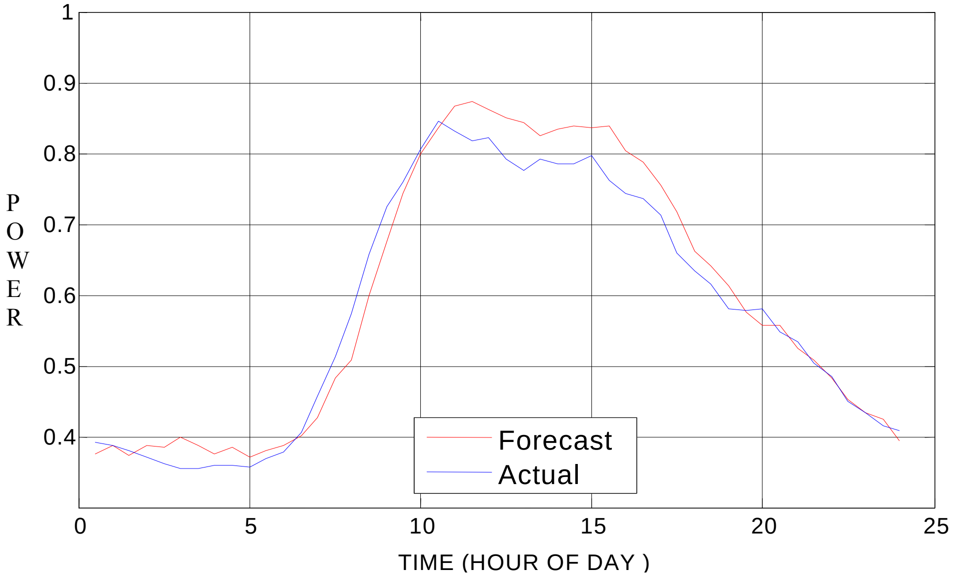


Figure 9. Forecast of Feb 2 (RRMSE=.0585), Commercial class

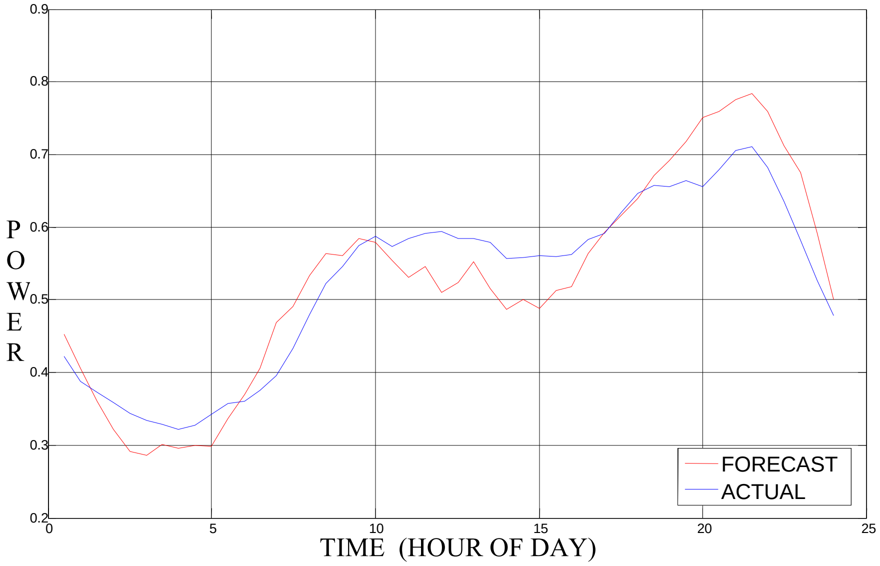


Figure 10. Forecast of May 22 (RRMSE =.0852),Residential class

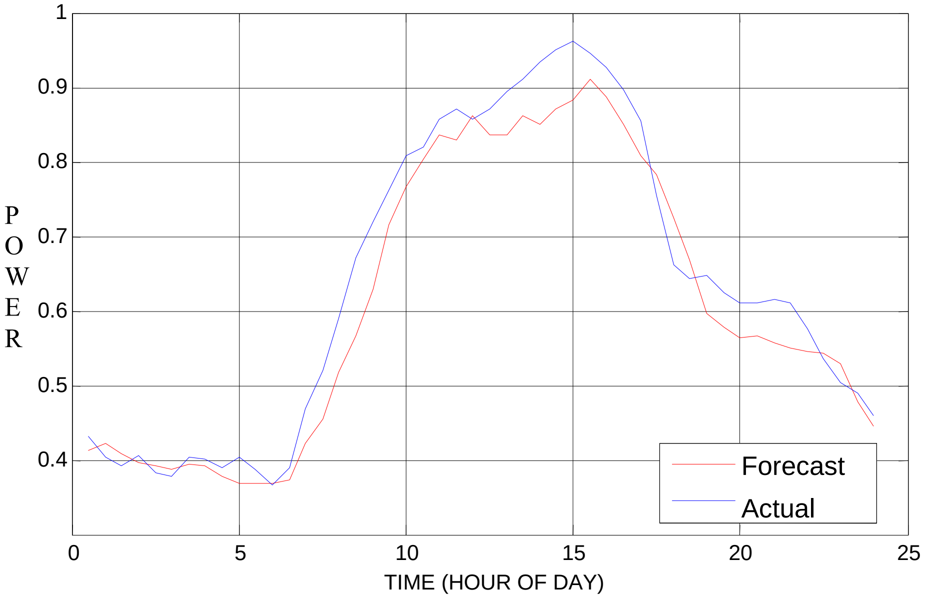


Figure 11. Forecast of June 8 (RRMSE=.0673), Commerical class

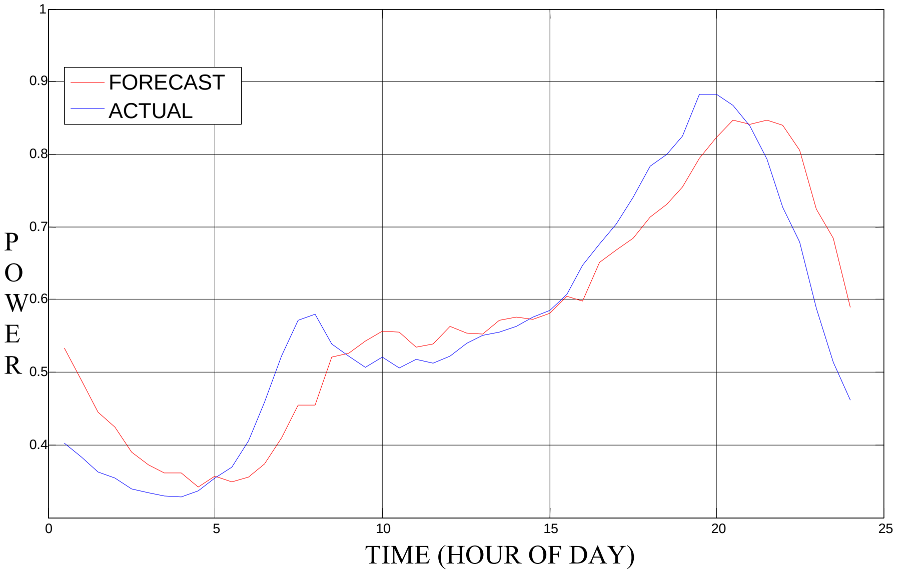


Figure 12. Forecast of September 22 (RRMSE =.11), Residential class

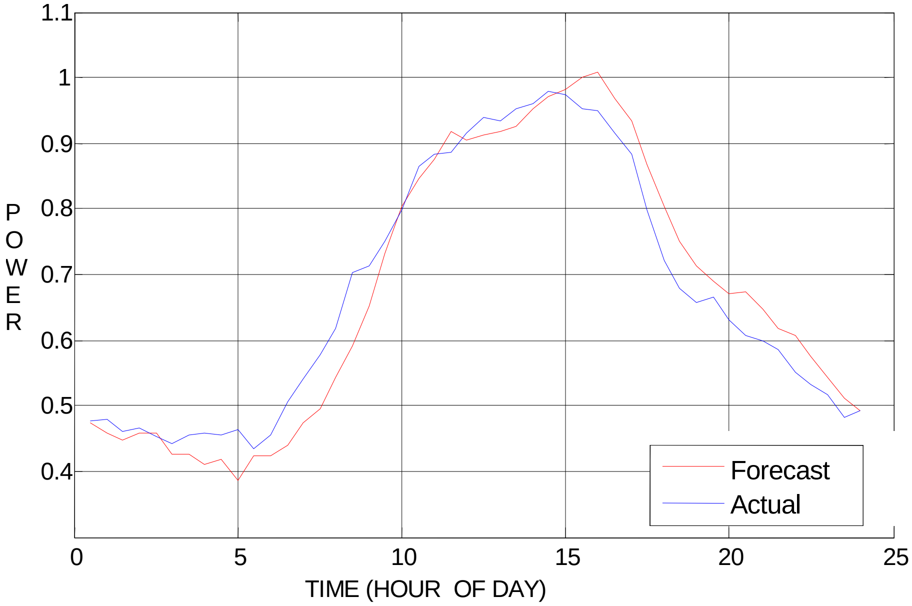


Figure 13. Forecast of Oct 1(RRMSE=. 0660), Commercial class

3.5 Prediction of weekdays, based on historical weekdays data

3.5.1 Residential class

For the residential class, a weekday and weekend/public holidays load consumption was observed and it was found that the consumption on weekends and public holidays followed a pattern different from that on weekdays, primarily during the middle of a day as shown in Figure 13.

Hence weekday's data was used to predict the load consumption of a weekday.

Also 3 PM consumption for a month showed the following pattern (to account for a very slight seasonal variation component) Figure 14. As seen from the scale of the ‘y’ axis, there is insignificant seasonal variation on the load consumption, which is expected as only January data is being considered right now. The SAS output is shown below which shows that the seasonal components (Pr > .05) are insignificant in predicting the load consumption.

```
                          Using 5 Harmonics                              2

                        The AUTOREG Procedure

                Estimates of Autoregressive Parameters

                                    Standard
          Lag     Coefficient          Error    t Value

            1       -0.815935       0.029701     -27.47
            2       -0.063503       0.029701      -2.14


                        Yule-Walker Estimates

   SSE                 0.15885665    DFE                  1129
   MSE                  0.0001407    Root MSE          0.01186
   SBC                 -6807.3384    AIC            -6923.4713
   Regress R-Square        0.9023    Total R-Square     0.9959
   Durbin-Watson           2.0101


                                 Standard                 Approx
Variable        DF     Estimate     Error    t Value    Pr > |t|

Intercept        1       0.5892  0.002882     204.45      <.0001
s1               1     0.008827  0.004095       2.16      0.0313
s2               1     0.005942  0.004083       1.46      0.1459
s3               1     0.009414  0.004062       2.32      0.0206
s4               1     0.001070  0.004033       0.27      0.7908
s5               1    -0.007750  0.003997      -1.94      0.0528
c1               1     0.002320  0.004047       0.57      0.5667
c2               1     0.001343  0.004035       0.33      0.7394
c3               1     0.001876  0.004015       0.47      0.6404
c4               1    -0.004816  0.003987      -1.21      0.2274
c5               1    -0.001885  0.003952      -0.48      0.6336
```

```
sd1                1       -0.0956     0.002767      -34.57      <.0001
sd2                1       -0.0114     0.001710       -6.67      <.0001
sd3                1        0.0275     0.001207       22.82      <.0001
sd4                1       -0.0149     0.000930      -16.02      <.0001
sd5                1      0.003979     0.000759        5.24      <.0001
cd1                1       -0.2306     0.002751      -83.81      <.0001
cd2                1        0.0507     0.001706       29.74      <.0001
cd3                1     -0.003031     0.001205       -2.52      0.0120
cd4                1       -0.0187     0.000929      -20.18      <.0001
cd5                1      0.005638     0.000758        7.43      <.0001
```

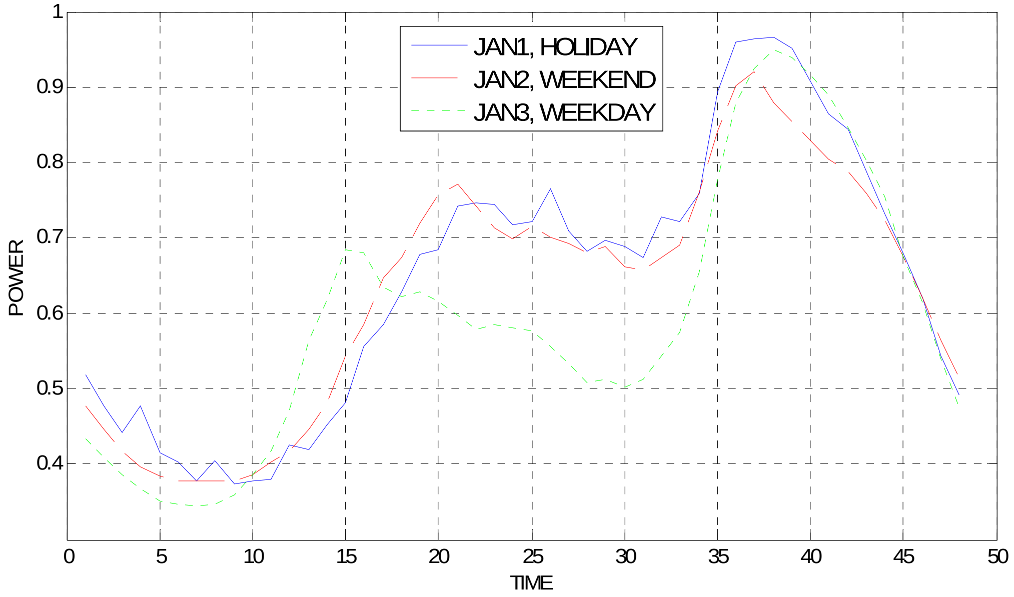


Figure 14. Consumption pattern of residential class, for various type of day

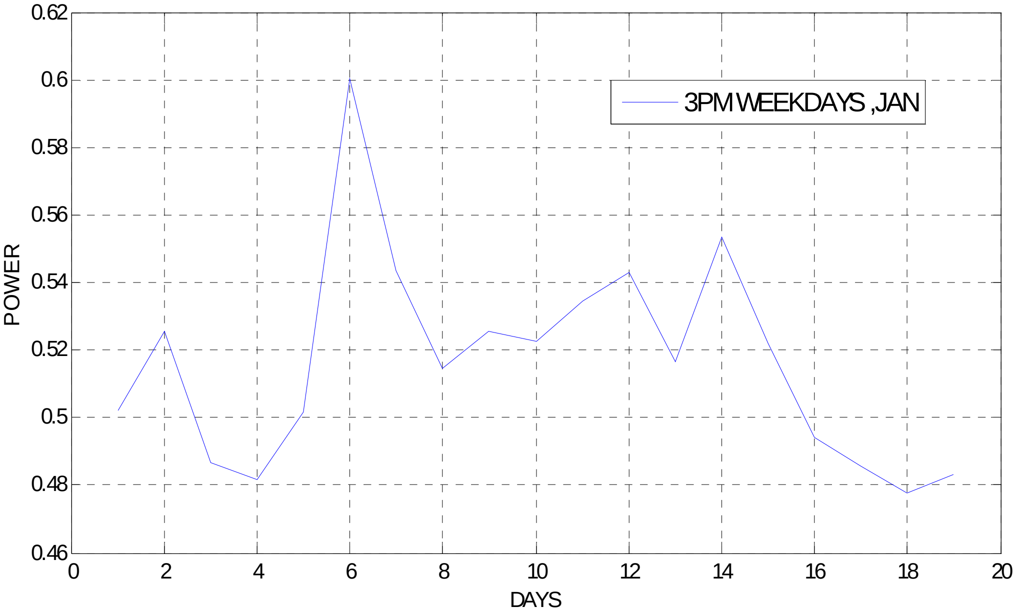


Figure 15. Consumption at 3PM during the weekdays of January

Following is the prediction result for Feb1-Feb5, Residential class (Figure 15-19):

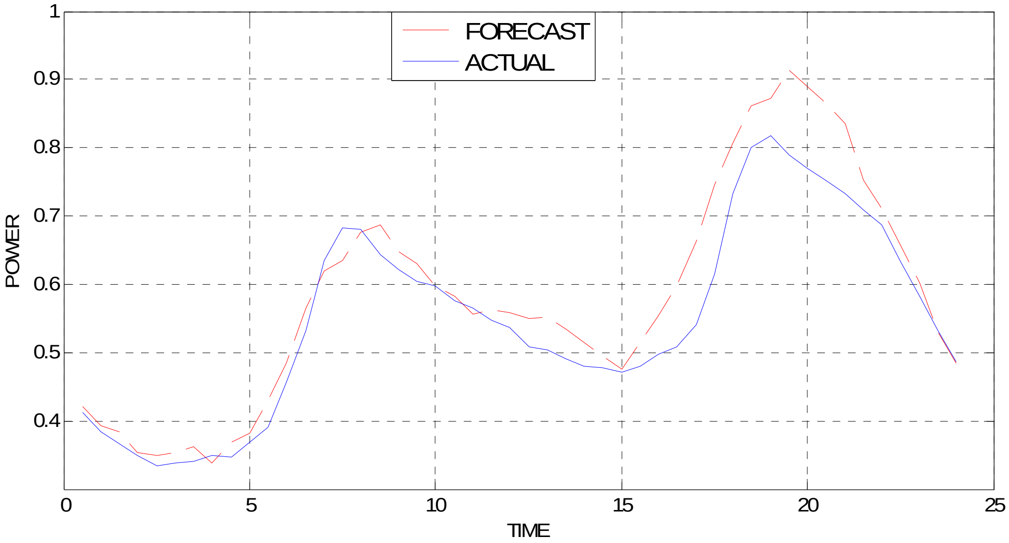


Figure 16. Forecast of February 1 (RRMSE=.0558), Residential class.

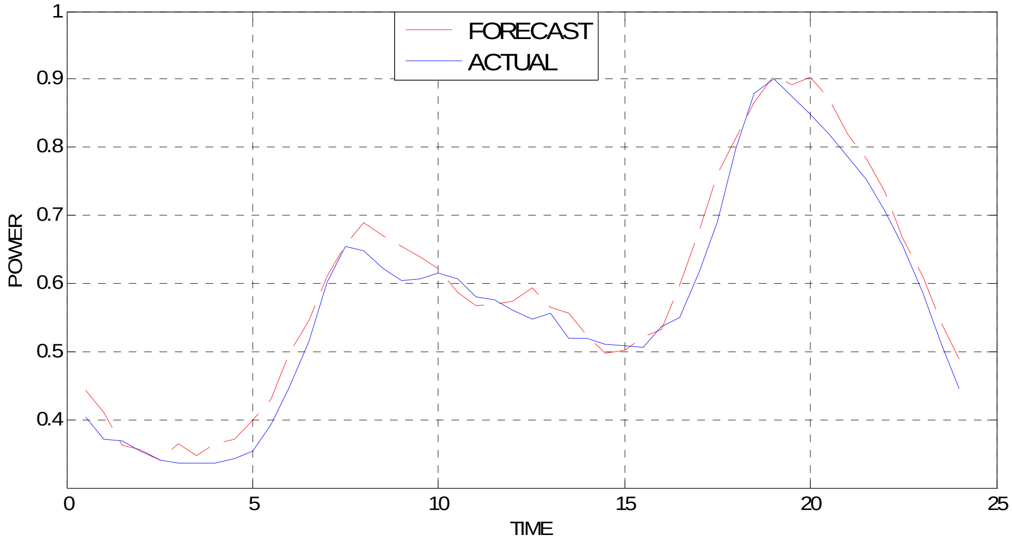


Figure 17. Forecast of February 2 (RRMSE=.0892), Residential class.

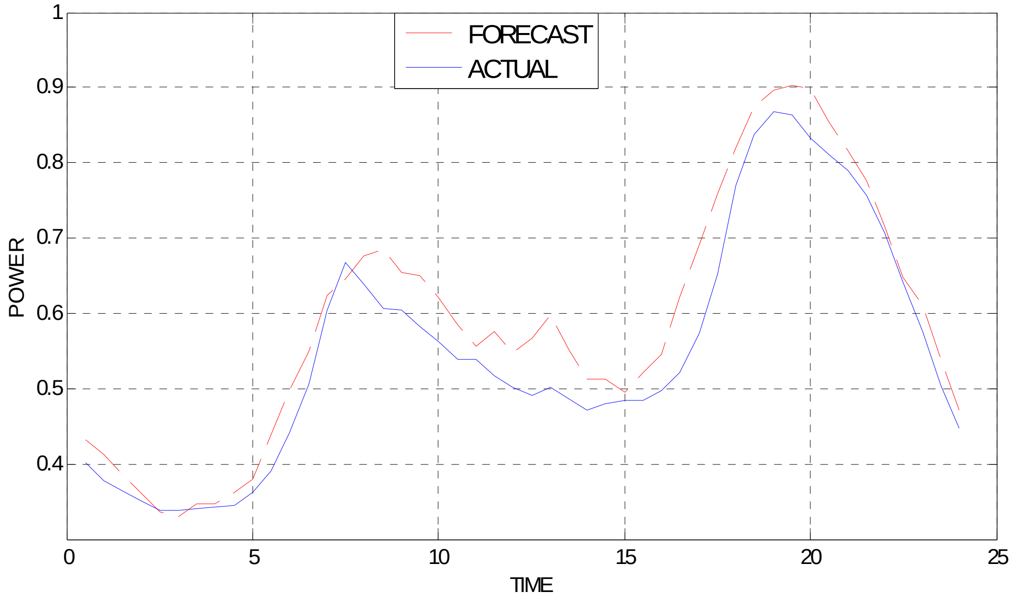


Figure 18. Forecast of February 3 (RRMSE=.0856), Residential class.

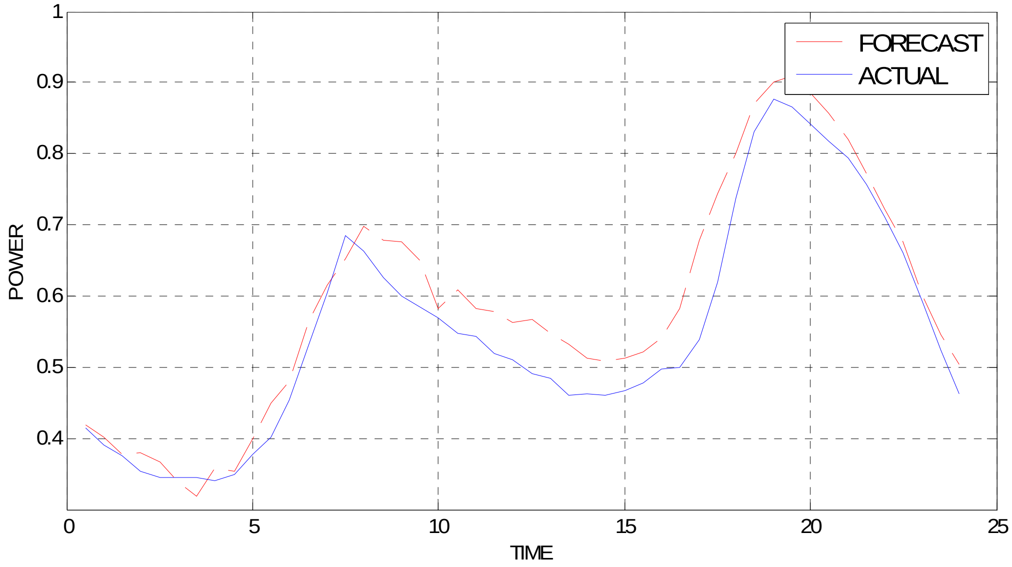


Figure 19. Forecast of February 4 (RRMSE=.0900), Residential class.

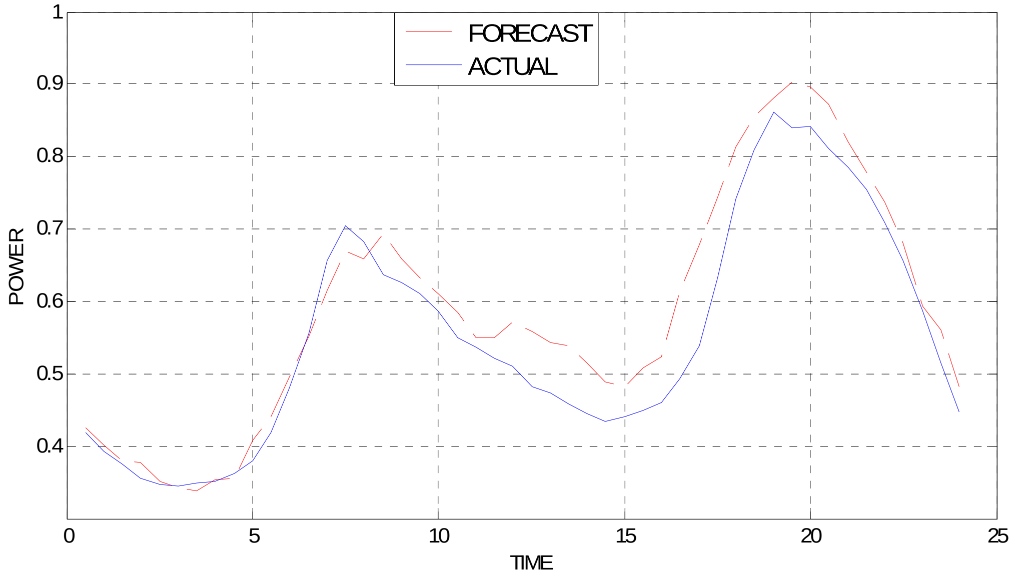


Figure 20. Forecast of February 5 (RRMSE=.0918), Residential class.

3.5.2 Commercial class

A weekday and weekend load consumption was observed and it was found that the consumption on weekends and public holidays followed a pattern different from that on weekdays, primarily during the middle of a day as shown in Figure 6

Weekday's data of January is used to predict the first week of February (weekdays) as shown (Figure 20-24) :

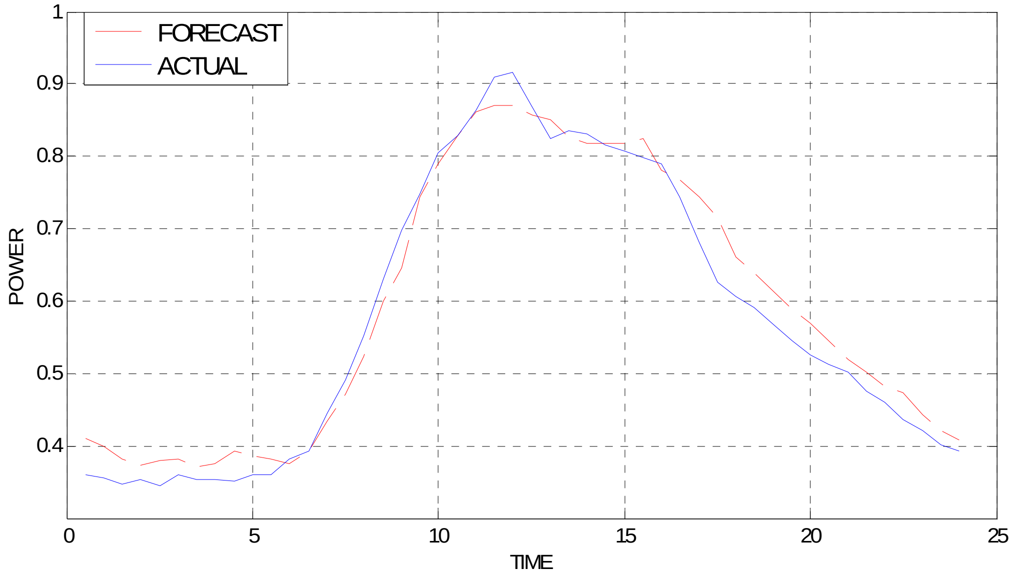


Figure 21. Forecast of February 2 (RRMSE=.0461), Commercial class

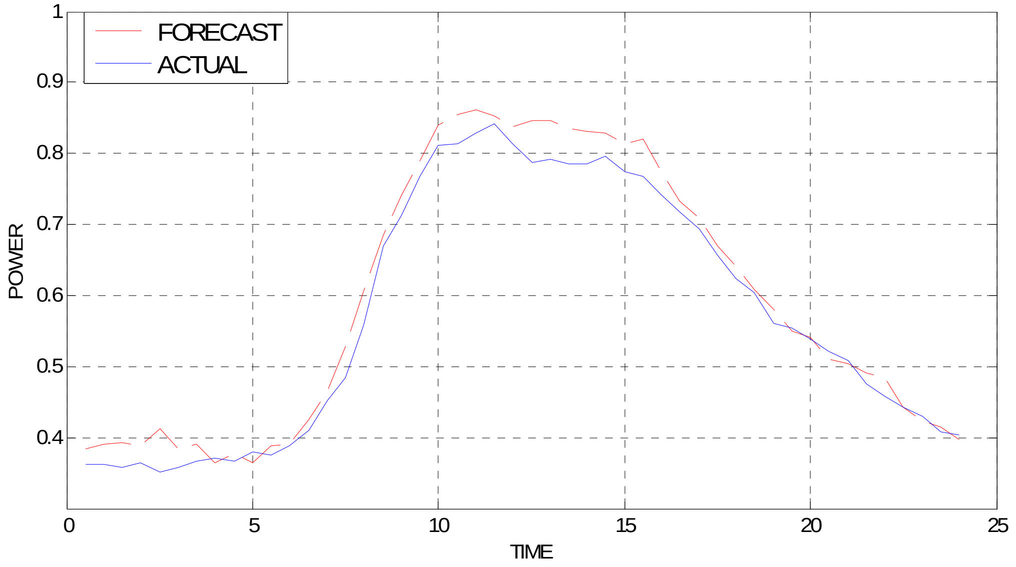


Figure 22. Forecast of February 3 (RRMSE=.0499), Commercial class

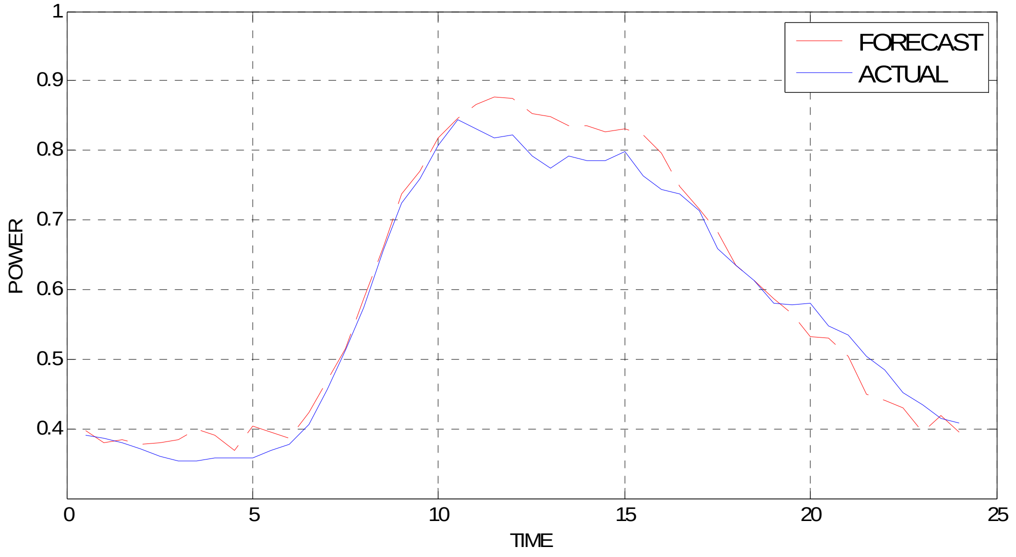


Figure 23. Forecast of February 4 (RRMSE=.0337), Commercial class

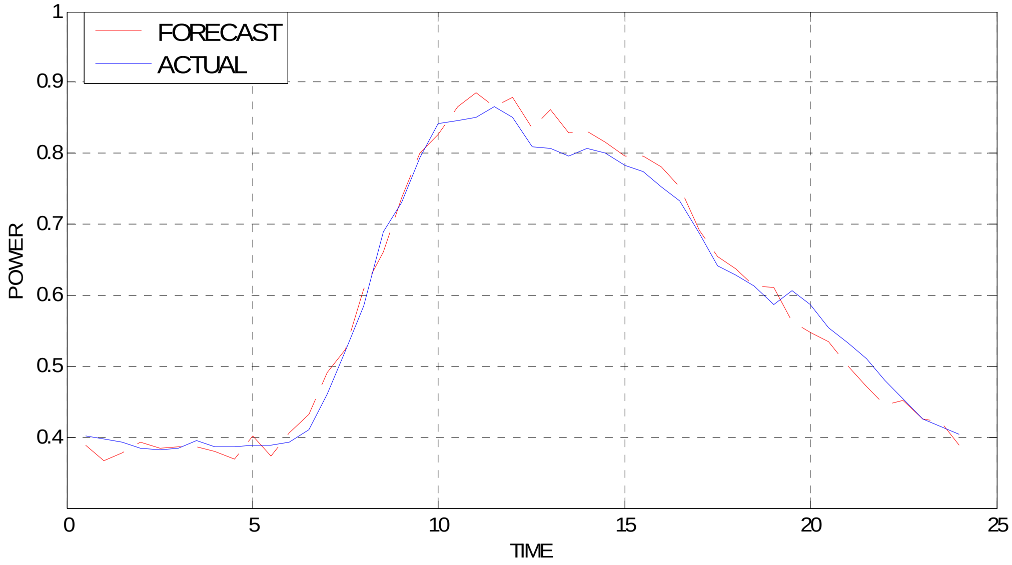


Figure 24. Forecast of February 5 (RRMSE=.0663), Commercial class

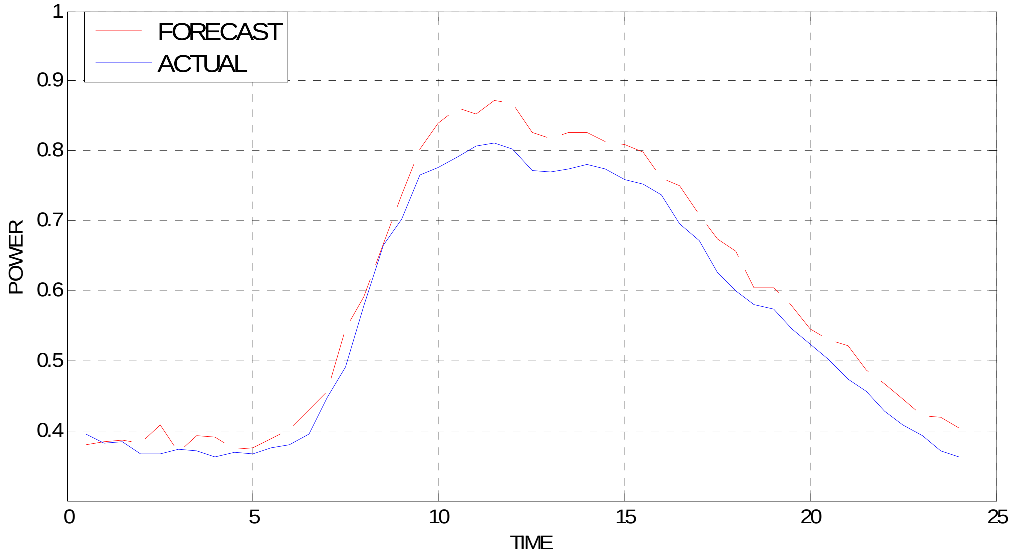


Figure 25. Forecast of February 6 (RRMSE=.0566), Commercial class

Chapter 4

Proposed Approach

4.1 Clustering Analysis

All the Load modeling methods reported in literature (chapter 1) and the method explained in chapter 3 cannot conduct overall analysis on the recorded data. The models obtained by implementing those techniques only have the ability to make curve-fitting for several groups of data. As it is unfeasible to install the load measurement units in each substation and its unnecessary to construct a model for each load nodal point, ‘load clustering’ technology provides an effective approach on handling the above mentioned problem and increasing the credibility of the modeling technique.

Cluster analysis is a technique used for combining observations into various groups such that:

a) Each group or cluster is homogenous with respect to certain characteristics, implying that the various observations in each cluster are similar to each other.

b) Each cluster should be different from other groups with respect to the same characteristics, that is, observations of one group should be different from the observations of other groups.

Zalewski [18] used fuzzy inference approach to cluster various substations and then used fuzzy regression models to predict load consumption of the substation clusters. Wang and Li [19] explained a fuzzy approach to choose cluster centers (cluster means). [5] and [20] explains the k-means method of clustering, the most widely used method used in clustering analysis.

Based on the similarities or distances (dissimilarities), objects are grouped together into groups (no assumptions on the number of groups). All the customers can be classified into several clusters, using a clustering technique, and in case of partial available AMI field data from the customers, data for the remaining customers in the clusters can be estimated using their historical data and real-time AMI data of those in their clusters.

4.1.1 Distance and similarity coefficients

All clustering algorithms require some type of a measure to assess the similarity of a pair of observations or clusters. Similarity measures can be distance measures, association coefficients or correlation coefficients. Distance measures are the most commonly used measures used in clustering algorithms, which are further divided into Euclidean and Statistical distances. For two p-dimensional observations ‘x’ and ‘y’ (p=number of variables), these distances are defined as follows:

For $x=[x_1,x_2...,x_p]$ and $y=[y_1,y_2...,y_p]$

Euclidean distance: $d(x,y)=\sqrt{(x-y)^{'}(x-y)}$

Statistical distance: $d(x,y)=\sqrt{(x-y)'.A.(x-y)}$

$A=S^{-1}$, S contains sample variances

Without prior knowledge about population variances, statistical distance cannot be computed. Hence, Euclidean distance is offered preferred for clustering.

4.2 K-means method (Non-Hierarchical clustering method)

K-means method of cluster detection is one of the most widely used ‘disjoint cluster analysis’ technique [5] [6] [20]. In k-mean clustering, the cluster centers are derived from the means of observations assigned to each cluster when the algorithm is run to complete convergence. In the k-means model, each iteration reduces the variation within the clusters and maximizes the difference between the distinct clusters until convergence is achieved. A set of points called ‘cluster seeds’ is selected as the first guess of the means of the clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seed are then replaced by the means of temporary clusters, and the process is repeated until no changes occur in the clusters. The procedure to group data through this method is summarized as follows:

1) Partition the items into K initial clusters –While no perfect way to determine the number of clusters exist, macro FASTCLUS in SAS uses some statistics (Cubic Clustering Criterion, pseudo F statistic and pseudo T^2 statistic) to determine the optimum number of clusters.[21] These statistics are plotted against number of clusters and the place where a jump occurs is selected as a good number of clusters.

a) The Cubic Clustering Criterion (CCC) was developed by SAS as a comparative measure of the deviation of the clusters from the distribution expected if data points were obtained from a uniform distribution. The criterion is calculated as

$$CCC = \ln[\frac{1-E(R^2)}{1-R^2}] * K$$

where $E(R^2)$ is the expected R^2, R^2 is the observed R^2, and K is the variance-stabilizing transformation. R^2 is explained in section 4.4. Larger positive values of the CCC indicate a better solution, as it shows a larger difference from a uniform (no clusters) distribution.

b) The pseudo-*F* statistic is intended to capture the 'tightness' of clusters, and is in essence a ratio of the mean sum of squares between groups to the mean sum of squares within group The value reported is obtained from SAS PROC FASTCLUS and is calculated as

$$Pseudo - F = \frac{(T - P_G)/(G-1)}{P_G/(n-G)}$$

where G is the number of clusters, T is the total sum of squares, and P_G is the within-group sum of squares. Larger numbers of the pseudo-*F* usually indicate a better clustering solution

c) The pseudo-T^2 statistic is a derived by transforming the ratio of Je(2)/Je(1) [23]. Je(2) is the sum of squared errors within clusters when the data are divided into two clusters and Je(1) is the sum of squared errors when only one cluster is present. The hypothesis of one cluster is rejected is smaller than a specified critical value.

2) Proceed through the list of items (number of customers) assigning an item to the cluster whose centroid is nearest.

3) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

4) Repeat above step, till no more assignments take place

The above process can be explained via an example. Suppose we measure two variables X1 and X2 for four observations A, B, C and D (Table 5).

Table 5. Data for illustrative example

Item	Coordinates of Centroids	
	$\bar{x}_1$	$\bar{x}_2$
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Suppose we want to find two clusters out of these four items such that items within a cluster are closer to one another than items in different cluster. We arbitrarily partition the items into two clusters (AB) and (CD) and compute the coordinates ($\bar{x}_1$ and $\bar{x}_2$) of the cluster centroids. Thus at step 1 we have (Table 6.)

Table 6. Calculation of cluster centroids (Step 1)

Cluster	Coordinates of Centroids	
	$\bar{x}_1$	$\bar{x}_2$
(AB)	[5+(-1)]/2=2	[3+1]/2=2
(CD)	[1+(-3)]/2=-1	[-2+(-2)]/2=-2

At step 2 Euclidean distances of each item from the group centroids are computed and each item is reassigned to the nearest group. If an item is moved from the initial configuration, the cluster centroids (mean) must be updated. The squared distances are:

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

Since A is closer to cluster (AB) than (CD), it is not reassigned. For B,

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

Consequently B is reassigned to cluster (CD), giving cluster BCD. The updated centroids are:

Table 7. Re-calculation of cluster centroids

Cluster	Coordinates of Centroids	
	$\bar{x}_1$	$\bar{x}_2$
(A)	-5	3
(BCD)	-1	-1

Again each item is checked for reassignment. Computing the squared distances.

Table 8. Check for re-assignment of items

Cluster	Squared distances to group Centroids			
	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

Hence, we see that each item is currently assigned to the cluster with the nearest centroid

The residential load data from five different utilities is used as a database whose data is to be clustered using the k-means algorithm. Customers are generated by perturbing the load profiles with a known value of standard deviation. An idea about the goodness of clustering can be made easily since it is known beforehand that which customer is derived from which base profile (actual data) and also by visually looking at the pattern of load consumption of the five base profiles. Following steps can be used to cluster the available load data :

Step 1: The database is generated by perturbing the actual load data with a known value of standard deviation. This database is imported into the SAS Enterprise Miner. Macro FASTCLUS is used to implement the above explained ‘k-means algorithm’ in SAS Enterprise Miner.

Step 2: FASTCLUS performs disjoint cluster analysis on the basis of distances computed from one or more variables. In this case the variables are times at which load data is available, 240 variables in this case as data is the weekday data for two weeks ,i.e. 10days, each day having 24 variables.

Step 3: The observations are divided into clusters such that every observation belongs to only **one cluster**

Clustering of load profiles

- Five residential profiles (from different utilities) are used. 50 customers’ load data generated from these profiles is used to implement the clustering algorithm.

- Customers are generated from the profiles by using a deviation of $3.\sigma$ =20% , where σ = standard deviation used to simulate customers associated with a profile.

- For clustering purposes, two weeks data (weekdays) was used (Figure 25): Following standard deviations were observed for a month's data belonging to the five profiles. Since considerable deviation was observed, it justifies using a month's weekday data to cluster the customers belonging to these profiles.

Table 9. Customer Standard Deviation

Profile	**Standard Deviation**
Profile 1	9.02%
Profile 2	6.33%
Profile 3	5.65%
Profile 4	8.62%
Profile 5	2.57%

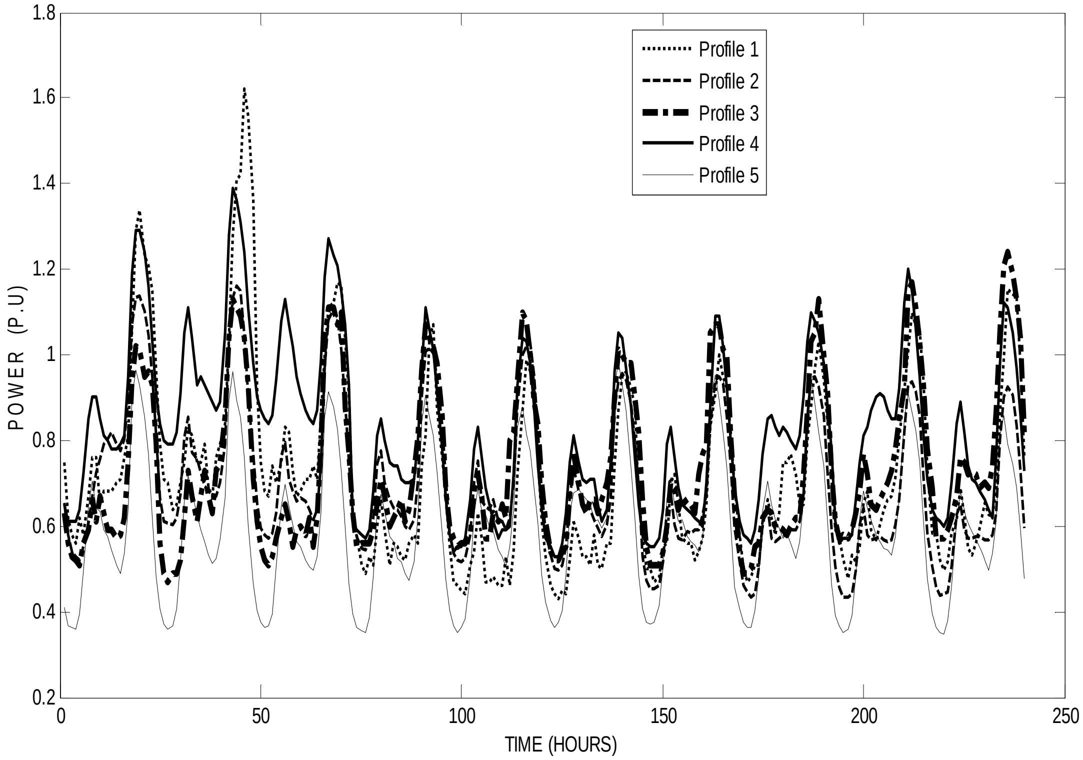


Figure 26. Two weeks of weekdays data for the five base profiles

4.3 Clustering Results

Figure 26. shows a day's plot of consumption patterns of the five profiles used in the clustering analysis. As can be visually observed, profiles profile 5 and profile 4 follow a distinct pattern, very different from each other. Profiles 1,2 and 3 follow a somewhat similar pattern. These observations are verified from the results obtained from SAS Enterprise miner. In a practical situation there would be hundreds of such profiles which can't be manually clustered, there it becomes imperative to use a reliable clustering algorithm.

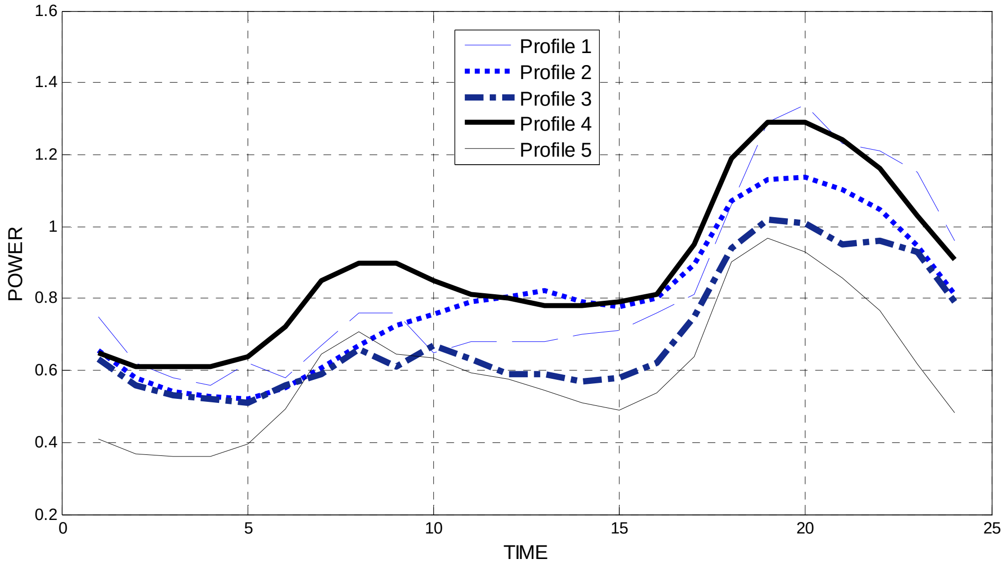


Figure 27. Consumption pattern of the five residential profiles.

The following parameters needed to be specified as input to the SAS macro FASTCLUS (which performs k-means clustering on a given data set), else default values are used.

RADIUS option specifies the minimum distance between an observation in consideration for potential seed and the existing seeds. If the observation does not meet this criterion, it cannot be selected as a seed. A too large value of RADIUS may result in number of seeds being less than the desired clusters. In this case, since per-unit values (small scale) of power consumed are being considered, a value of zero is used for RADIUS.

REPLACE option governs how the seeds could be replaced after the initial selection. REPLACE=full uses two available criterion in SAS to determine replacement of seeds [6]

MAXCLUSTER option specifies the number of clusters.

MAXITER option specifies the maximum number of iterations. The iterations are continued until the change in the cluster centroids of two successive iterations is less than the convergence value specified by the researcher. Here a default value of .001 is used as the threshold convergence value

4.3.1 Case 1: Customers generated from profiles with 10% Deviation (Tightly correlated profiles)

Customers are generated from the profiles by using a deviation of $3.\sigma=10\%$, where $\sigma=$ standard deviation used to simulate customers associated with a profile. Following is the SAS clustering output:

```
                        The FASTCLUS Procedure
              Replace=FULL  Radius=0  Maxclusters=3 Maxiter=1

                 Criterion Based on Final Seeds =   0.0601


                           Cluster Summary

                                      Maximum Distance
                          RMS Std          from Seed     Radius     Nearest
Cluster     Frequency   Deviation      to Observation   Exceeded    Cluster
-----------------------------------------------------------------------------
   1              30      0.0765              1.3235                      3
   2              10      0.0203              0.3261                      1
   3              10      0.0277              0.4322                      1



                        Statistics for Variables

      Variable      Total STD    Within STD      R-Square     RSQ/(1-RSQ)
      OVER-ALL        0.11075       0.06197      0.699717        2.330195
```

Cluster 1: Profile 1 (10) + Profile 2 (10) + Profile 3 (10)

Cluster 2: Profile 5 (10)

Cluster 3: Profile 4 (10)

Which is in accordance to what could be predicted in 4.2.2

4.3.2 Case 2: Customers generated from profiles with 20% Deviation

Customers are generated from the profiles by using a deviation of $3.\sigma$ =20% , where σ = standard deviation used to simulate customers associated with a profile. Following is the SAS clustering output:

```
                              The FASTCLUS Procedure
               Replace=FULL  Radius=0  Maxclusters=3 Maxiter=1

                  Criterion Based on Final Seeds =   0.0737


                                 Cluster Summary

                                      Maximum Distance
                         RMS Std             from Seed      Radius     Nearest
Cluster      Frequency   Deviation        to Observation    Exceeded    Cluster
-----------------------------------------------------------------------------
   1              10       0.0569               0.9104                        2
   2              30       0.0883               1.5835                        1
   3              10       0.0424               0.7203                        2


                              Statistics for Variables

      Variable     Total STD    Within STD       R-Square     RSQ/(1-RSQ)
      OVER-ALL       0.11875       0.07597       0.607469        1.547568
```

Cluster 1: Profile 4 (10)

Cluster 2: Profile 1 (10) + Profile 2 (10) + Profile 3 (10)

Cluster 3: Profile 5 (10)

As compared to the previous case, where clusters are same in terms of constituents but as would be seen in the next section, that clusters obtained in case 1 are more homogenous and well separated than in case 2.

Figure 27, 28 and 29 show the two weeks of data for the three clusters in cases 1 and 2.

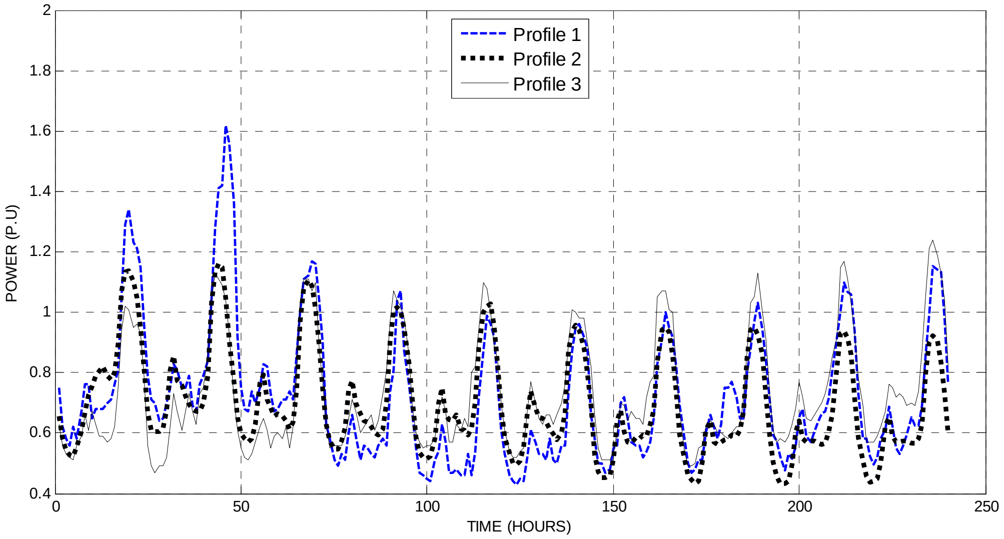


Figure 28. Two weeks of weekday's data for the three clustered base profiles

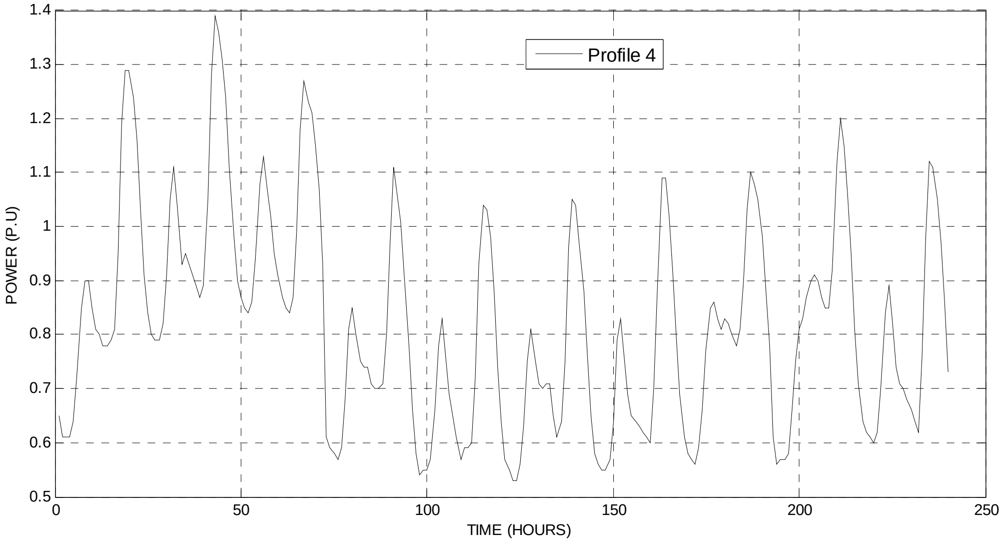


Figure 29. Two weeks of weekday’s data for the base profile-4

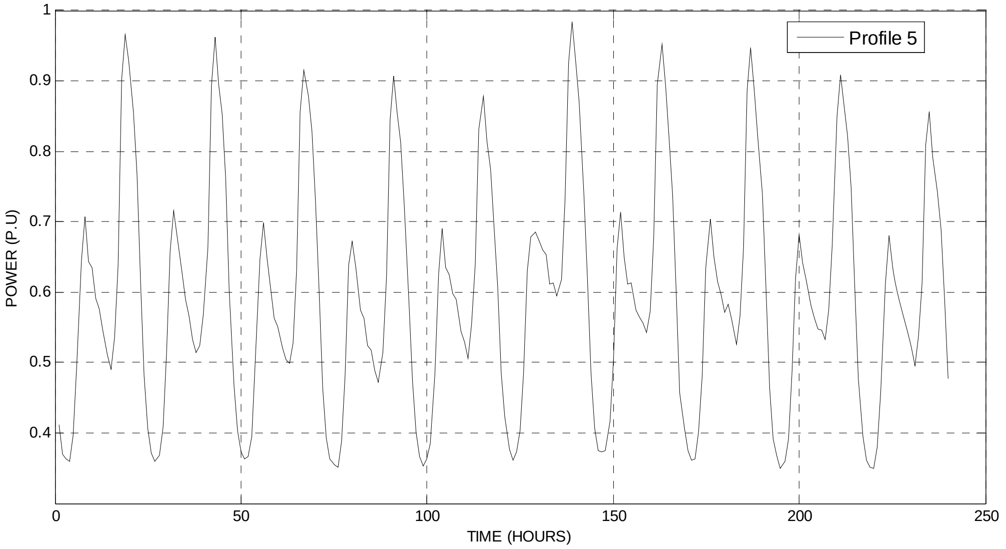


Figure 30. Two weeks of weekday’s data for the base profile-5

4.4 Performance of clustering

Although visual inspection of data and the clustering results can give a good idea as to how good is the clustering, i.e. how far apart the clusters are and also some idea about the homogeneity of the clusters. Still, some statistics are required which can quantify the goodness of clustering. Following statistics computed from the SAS output gives a measure of effectiveness of clustering:

a) **Overall R-square** : It is the ratio of SS_b to SS_t , where,

SS_b = sum of the squares of distances between clusters, is a measure of the extent to which the groups (clusters) are different from each other

$$SS_b = \sum_{j=1}^{p} \sum_{g=1}^{G} n_g.(\bar{x}_{jg} - \bar{x}_j)^2$$

Where,

G =number of clusters,

n_g=number of observations in group g

$\bar{x}_{jg}$=mean of ‘j’th variable in the group ‘g’

$\overline{x}_j$ =mean of 'j'th variable in the total data set

p = number of variables

$$SS_{wg} = \sum_{j=1}^{n_g} (x_{jg} - \overline{x}_j)$$

Where ,

x_{jg} = 'j'th observation of the group 'g'

SS_{wg} =sum of squares (within) for the group 'j'

SS_t = total sum of squares of all clusters (within + between), which is a constant for a given input data set.

SS_w = sum of the squares of distances within clusters, from the cluster centroid, which is a measure of the extent to which the groups (clusters) are homogenous.

$SS_t = SS_b + SS_w$

Hence for a given data set, the greater the differences between groups the more homogeneous each group is and vice-versa.

R-square values ranges from 0-1, the values of 0 indicating no differences between clusters and 1 indicating maximum difference between clusters.

b) **Ratio of Within RMS-STD to Total RMS-STD**: The relative value of within RMS-STD (Root Mean Square standard deviation) to total RMS-STD s a good measure of the homogeneity of the clusters .It should be as low to indicate high homogeneity of clusters.

$$RMSSTD = \sqrt{\frac{\sum_{j=1}^{p} \hat{s}_j^{\,2}}{p}}$$

$\hat{s}_j^{\,2}$=variance of the jth variable

A low value suggests the clusters are homogenous.

c) **RMS-STD of the clusters**: The RMS Standard Deviation of the clusters formed through the process gives an idea of how homogenous the clusters are. A low value suggests good homogeneity and also that cluster with the minimum value out of all the clusters is the most homogenous cluster out of the all.

d) **Centroid Distance between nearest clusters**: The overall distance between two nearest clusters (considering all variables) is also a measure of the goodness of clustering.

From Table 6 and 7, it can be inferred that although clustering results are better for Case 1, but since these were based on different assumptions on standard deviations, in practical situation data to be clustered would be available for clustering, in other words data need not be ‘generated’, it would be available with the utility.

Table 10. Customers generated from profiles with 10% Deviation

Performance Characteristic	**Value**
R-square	.70
Ratio of Within RMS-STD to Total RMS-STD	.5992
RMS-STD of the clusters	.0765, .0203 and .0277
Centroid Distance between nearest clusters	Around 2

Table 11. Customers generated from profiles with 20% Deviation.

Performance Characteristic	**Value**
R-square	.6074
Ratio of Within RMS-STD to Total RMS-STD	.6394
RMS-STD of the clusters	.0569, .0883 and .0424
Centroid Distance between nearest clusters	2-2.5

The overall R-square of .7 and .607 are large suggesting that the clusters are quite homogenous and well separated, but the result in case 1 shows that clustering is better than case 2 (since the profiles were generated from 10% deviation or the customer profiles were tightly correlated). Ratio of within RMS-STD to total RMS-STD (.5992 and .6394) is low indicating the resulting clusters are quite homogenous. Again the result in case 2 is toward the higher side than case 1. The RMS-STD of the clusters is also low giving another measure of good homogeneity. Since the distance computed between the centroids of the clusters is high indicating that the clusters are well separated.

4.5 Case Study

For load monitoring studies, usually all the loads on a line section are lumped together for implementing the State Estimation technique. Consider a line section on a distribution feeder with ‘x’ customers. On a specific day and time, usually all the customers data is not known. In that case, a load estimation technique helps to estimate the missing customer data. Based on the historical data of all the customers, available with the utility, a clustering algorithm can be used to find groups or ‘clusters’ of customers. There would be some real time data (AMI) available within each cluster and some of the data may be unavailable. So within each cluster the Load Estimation technique can be used to predict the missing customer’s data, using the available data within that cluster.

Consider a 33 node sample feeder (Figure 27); there are some customers between node 2 and node 3.

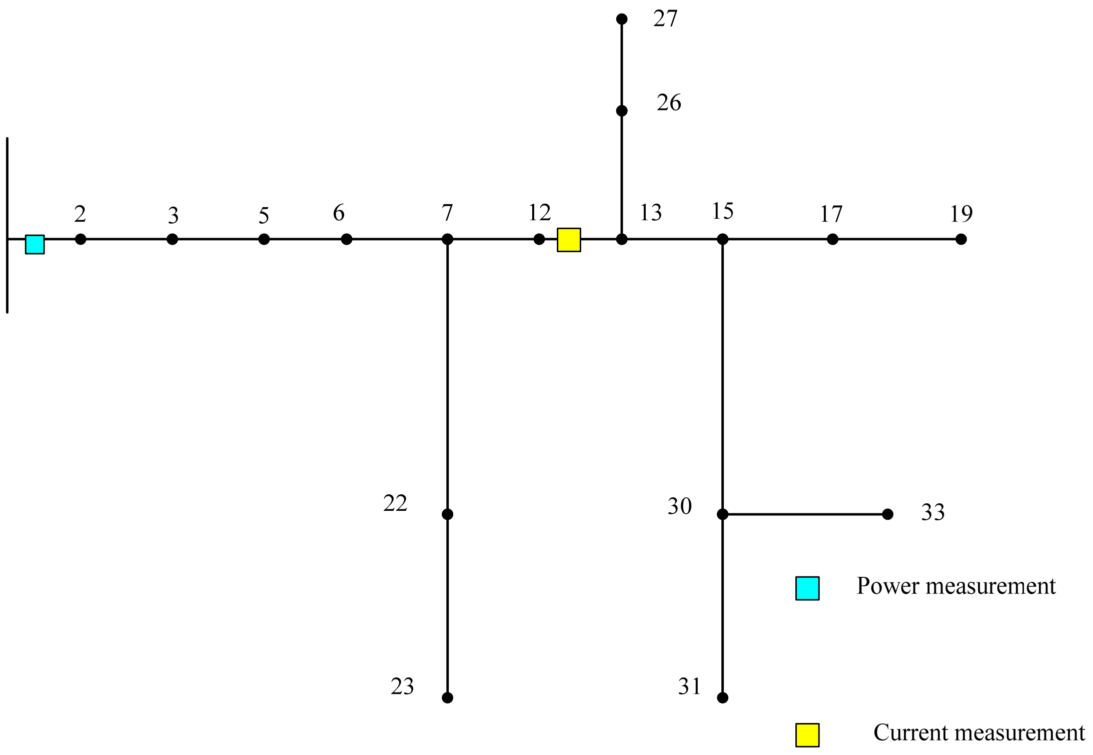


Figure 31. Sample feeder

Some of these customers’ data (real time power consumption) would be not available due to a failed meter or any other reason. The previously explained ‘clustering technique’ would be implemented to cluster these customers between into various homogenous groups. In a particular cluster, there may be some customers with missing data. The load modeling technique proposed in Chapter 3 would be modified to estimate the missing data of the affected customers from their historical data and other customers in their cluster.

The load estimation technique explained in chapter 3 can be used, along with available measurements as predictor variables. For a given cluster, AMI data belonging to that cluster can be modeled as another predictor variable, to estimate the load for those customers with missing real-time data.

The model from chapter 3 is:

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i \cos(i2\pi t/48*365) + \sum_{i=1}^{5} \beta_i \sin(i2\pi t/48*365) + \sum_{j=1}^{5} \beta_j \cos(j2\pi t/48)$$

$$+\sum_{j=1}^{5} \beta_j \sin(j2\pi t/48) + R_t \tag{1}$$

Where,

$R_t = \phi_1 . R_{t-1} + \phi_2 R_{t-2} + \varepsilon$

ϕ_1=First order lag

ϕ_2= Second order lag

ε =uncorrelated errors.

y = power (response variable in the regressive model)

β_0 = slope of the regressive model

$\beta_i 's$ = unknown parameters for the harmonic components, representing the yearly (seasonal) variation.

$\beta_j 's$ = unknown parameters for the harmonic components, representing the daily variation

ε =uncorrelated errors.

Consider a case where just one AMI data is used as another predictor variable (assuming just one customer's real-time data is available for each cluster), that data can be added to the model in (1) as follows,

$$
\begin{aligned}
y = \beta_0 + \sum_{i=1}^{5} \beta_i \cos(i2\pi t/48*365) + \sum_{i=1}^{5} \beta_i \sin(i2\pi t/48*365) + \sum_{j=1}^{5} \beta_j \cos(j2\pi t/48) \\
+ \sum_{j=1}^{5} \beta_j \sin(j2\pi t/48) + \beta_{k0} P_t + R_t
\end{aligned}
\tag{2}
$$

Here, at time'‘t’ prediction is made for the missing data, using real time AMI data at time ‘t’, P_t of a related customer (belonging to the same homogenous group).

In the model shown in (2),

β_{k0} is the unknown coefficient relating the dependence on the AMI data at time ‘t’.

Customers are generated from 5 base residential profiles. Case 4.4.1 shows the clustering and estimation results for customers generated by perturbing the base profiles with a perturbation of 10% ($3\sigma = 10\%$). Case 4.4.2 shows the clustering and estimation results for customers generated by perturbing the base profiles with a perturbation of 20% ($3\sigma = 20\%$). The customers are generated from the 5 base profiles as follows,

4.4.1 Case 1: Customers generated from profiles with 10% Standard Deviation

Figure 28 shows a day's trend for the 4 customers generated from base profile 1 and Figure 29 shows the general pattern of the base profiles.

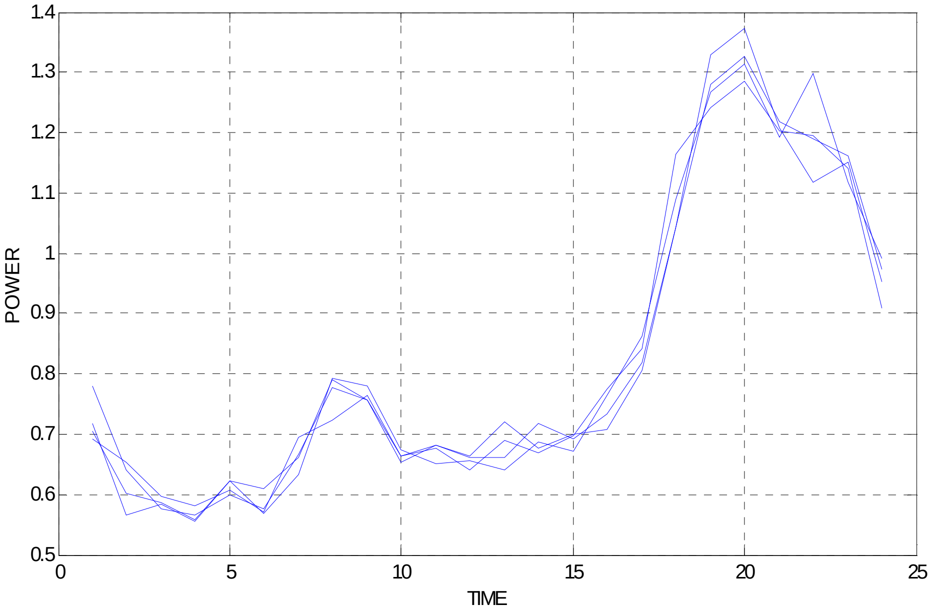


Figure 32. 4 customers generated from base profile 1 ($3\sigma = 10\%$)

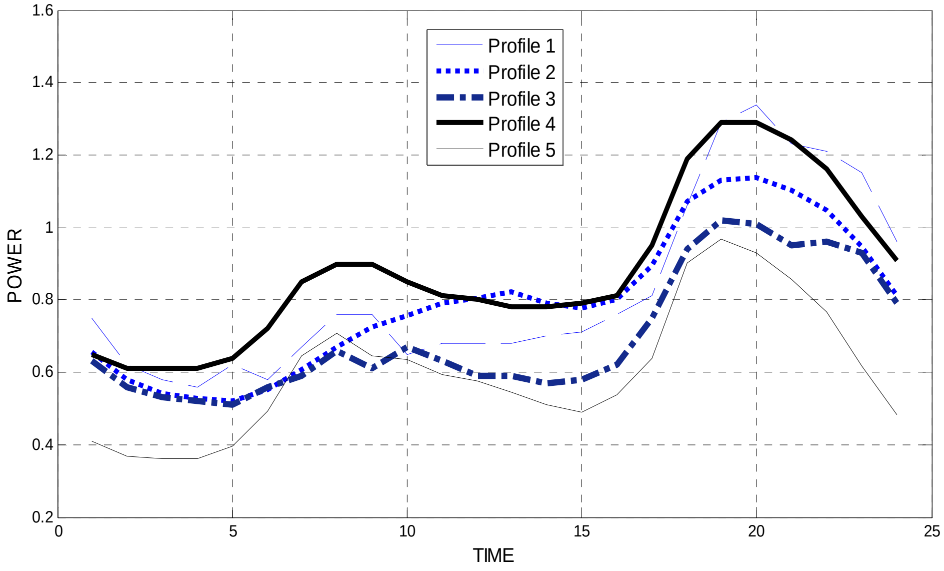


Figure 33. A sample day variation of the 5 base profiles.

The clustering results by using the k-means method in SAS Enterprise Miner is as follows:

```
                         The FASTCLUS Procedure
              Replace=FULL  Radius=0  Maxclusters=3 Maxiter=1

                 Criterion Based on Final Seeds =    0.0577


                              Cluster Summary

                                       Maximum Distance
                          RMS Std             from Seed      Radius      Nearest
Cluster      Frequency   Deviation       to Observation    Exceeded      Cluster
--------------------------------------------------------------------------------
   1             10        0.0805                1.3262                        3
   2              3        0.0207                0.2622                        1
   3              6        0.0282                0.4210                        1
```

Cluster 1: Profile 1 (4) + Profile 2 (3) + Profile 3 (3)

Cluster 2: Profile 5 (3)

Cluster 3: Profile 4 (6)

Now based on the modified time series model, prediction for customer 1 (assume its meter went bad) is done based on its historical data and real time data from customer 2 (available meter data), both are from the same cluster. Figures 30-33 show the results. ‘Forecast with meter’ implies using the modified time series model (taking into account the real-time data of a related meter) and ‘Forecast without meter’ is the prediction based on historical data only.

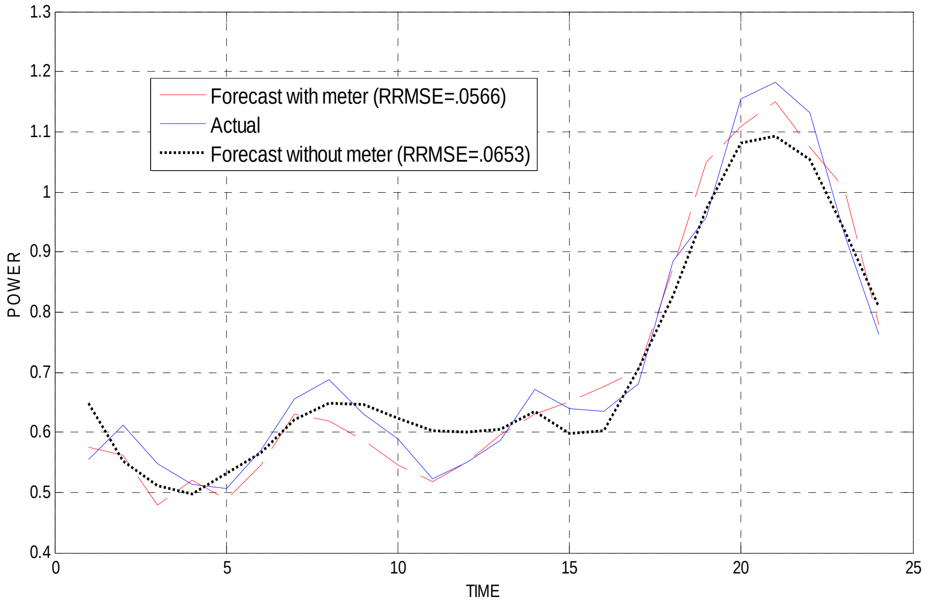


Figure 34. Forecast of February 15（$3\sigma = 10\%$）

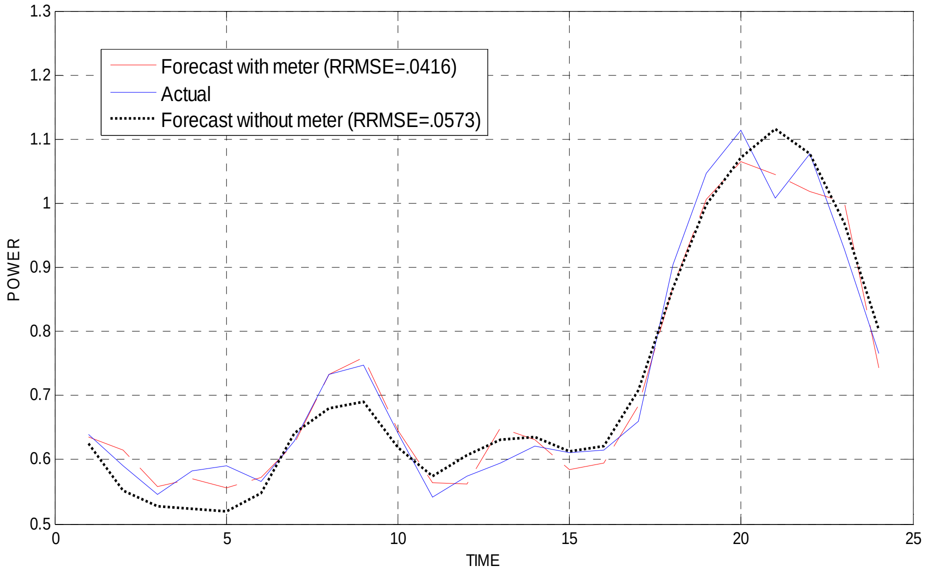


Figure 35. Forecast of February 16 ($3\sigma = 10\%$)

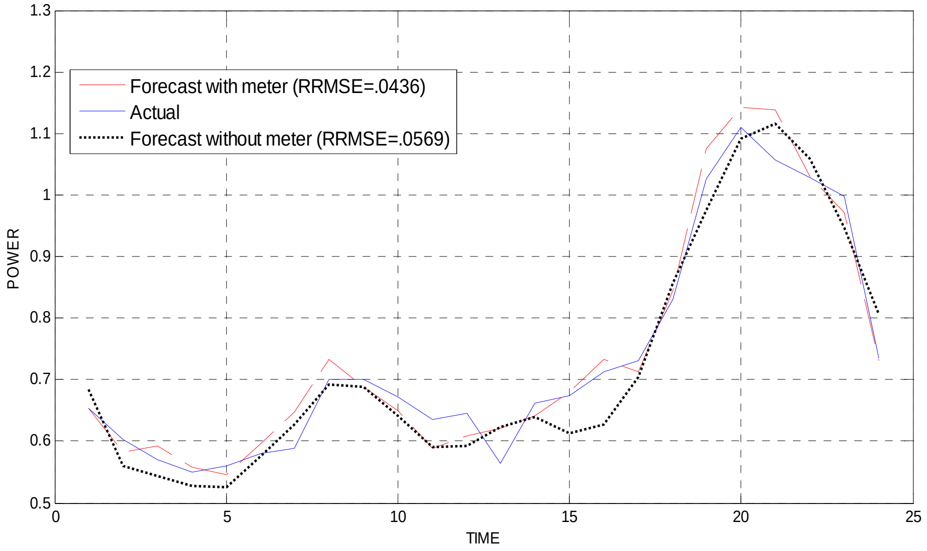


Figure 36. Forecast of February 17 ($3\sigma = 10\%$)

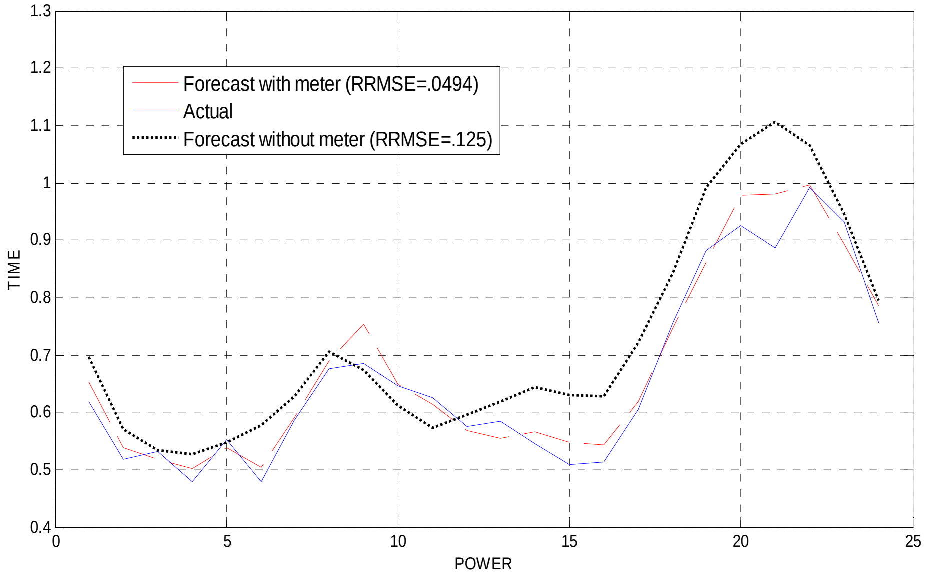


Figure 37. Forecast of February 18 ($3\sigma = 10\%$)

As seen from Figures 30-33, RRMSE for the predictions ‘Forecast with meter’ (using real-time data of a related meter for prediction purpose) is lower than just based on historical data (Forecast without meter).

4.4.2 Case 2: Customers generated from profiles with 20% Standard Deviation

Figure 34 shows a day’s trend for the 4 customers generated from base profile 1

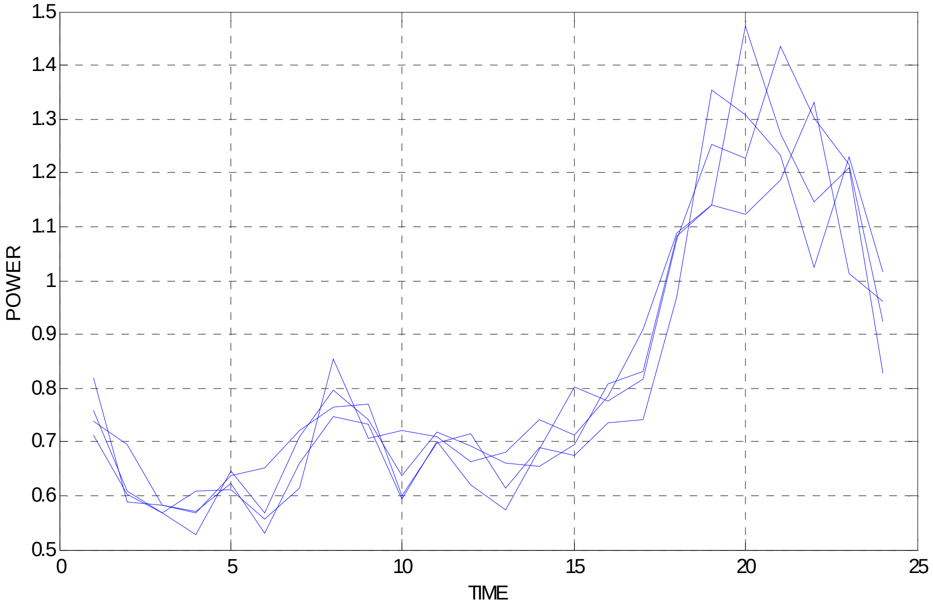


Figure 38. 4 customers generated from base profile 1 ($3\sigma = 20\%$)

The clustering results by using the k-means method in SAS Enterprise Miner is as follows:

```
                    The FASTCLUS Procedure
           Replace=FULL  Radius=0  Maxclusters=3 Maxiter=1

              Criterion Based on Final Seeds =   0.0715


                         Cluster Summary

                                       Maximum Distance
                             RMS Std          from Seed      Radius     Nearest
Cluster     Frequency      Deviation     to Observation    Exceeded     Cluster
-------------------------------------------------------------------------------

   1                6         0.0561             0.8201                     3
   2                3         0.0434             0.5646                     3
   3               10         0.0929             1.4882                     1
```

Cluster 3: Profile 1 (4) + Profile 2 (3) + Profile 3 (3)

Cluster 2: Profile 5 (3)

Cluster 1: Profile 4 (6)

Now based on the modified time series model, prediction for customer 1 (assume its meter went bad) is done based on its historical data and real time data from customer 2 (available meter data), both are from the same cluster. Figures 35-38 show the results. 'Forecast with meter' implies using the modified time series model (taking into account the real-time data of a related meter) and 'Forecast without meter' is the prediction based on historical data only.

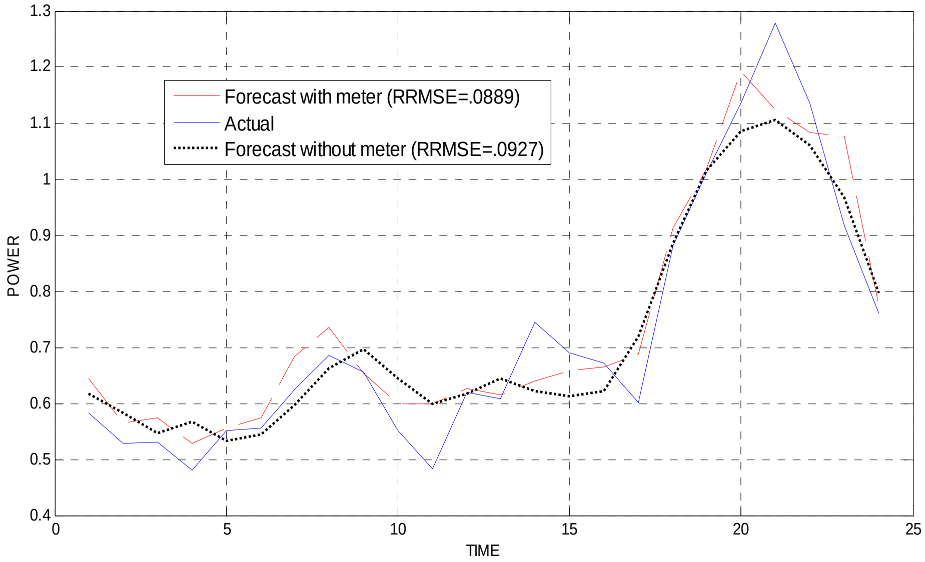


Figure 39. Forecast of February 15 ($3\sigma = 20\%$)

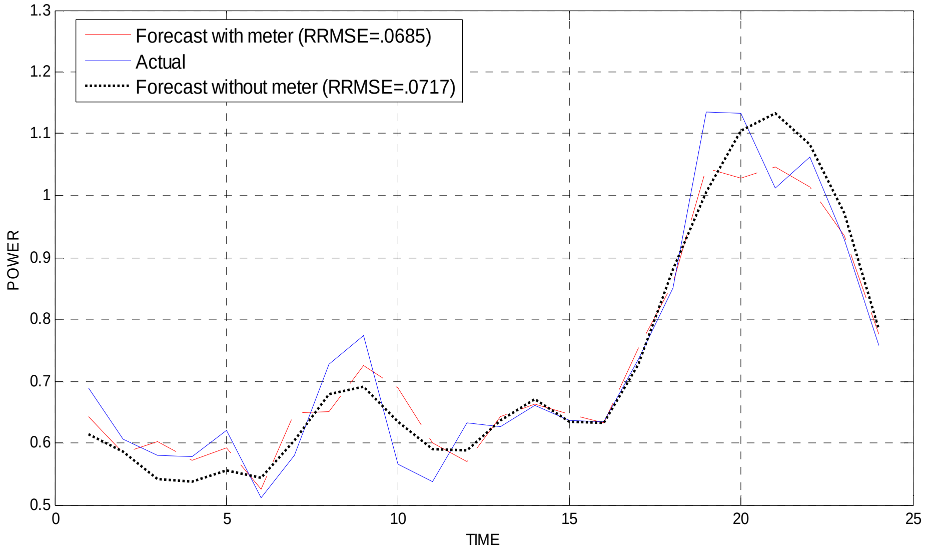


Figure 40. Forecast of February 16 ($3\sigma = 20\%$)

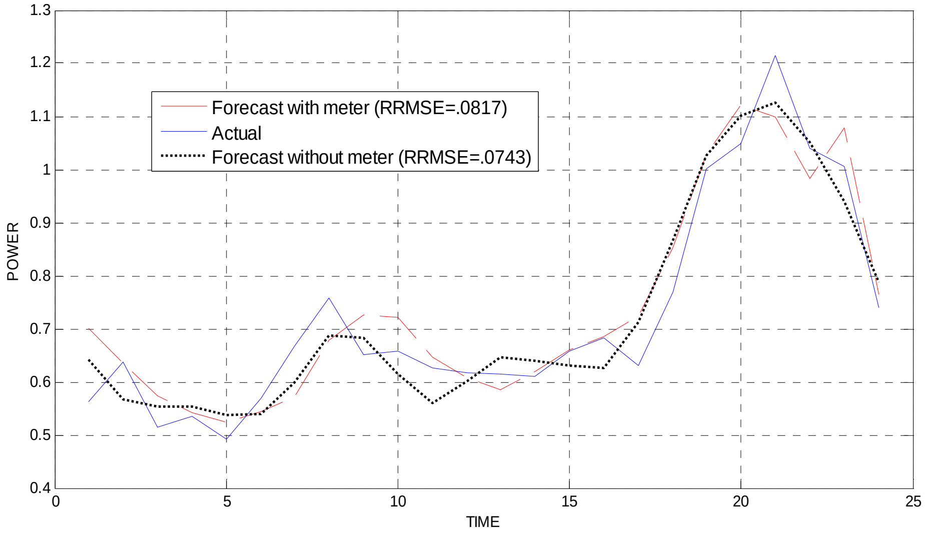


Figure 41. Forecast of February 17 ($3\sigma = 20\%$)

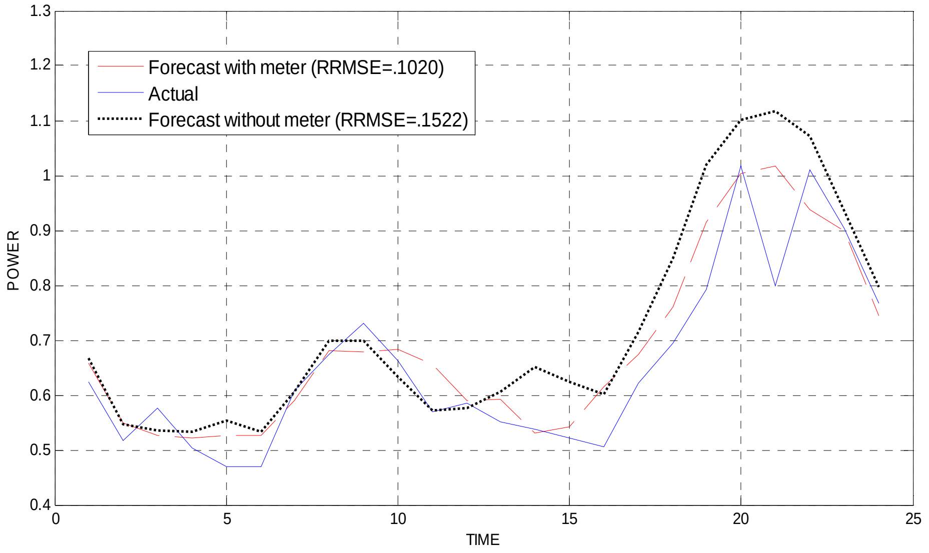


Figure 42. Forecast of February 18 ($3\sigma = 20\%$)

As seen from Figures 35, 36 and 38 RRMSE for the predictions 'Forecast with meter' (using real-time data of a related meter for prediction purpose) is lower than just based on historical data (Forecast without meter).

4.4.3 Case 3: Customers generated with 20% Standard Deviation (Predictor variable from different base profile)

In this case, the predictor meter in the model,

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i \cos(i2\pi t / 48*365) + \sum_{i=1}^{5} \beta_i \sin(i2\pi t / 48*365) + \sum_{j=1}^{5} \beta_j \cos(j2\pi t / 48)$$

$$+ \sum_{j=1}^{5} \beta_j \sin(j2\pi t / 48) + \beta_{k0} P_t + R_t$$

is from the same cluster (as the meter being estimated) but from a different base profile. The meter being predicted is from profile 1 and the predictor from base profile 2. Following results are obtained,

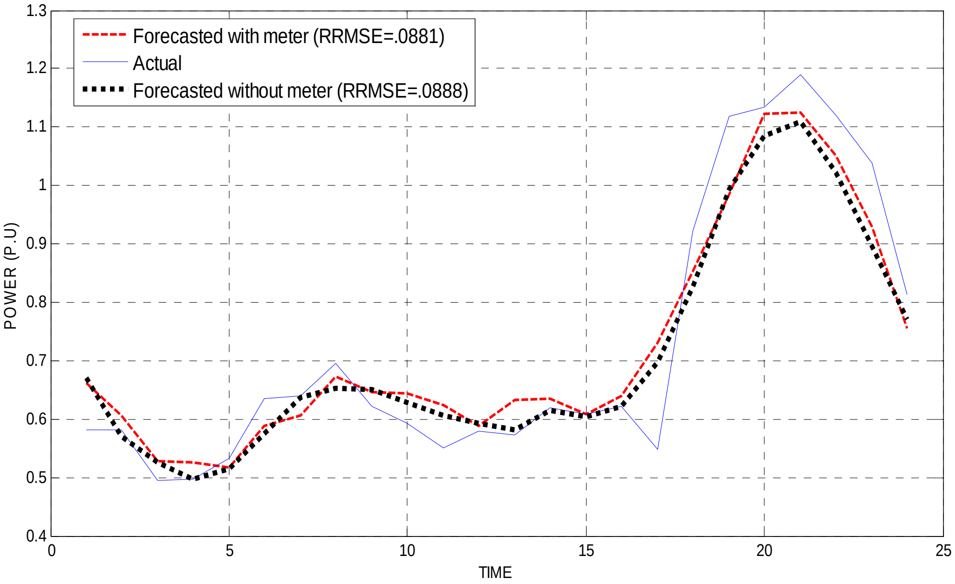


Figure 43. Case 3, Forecast of February 15 ($3\sigma = 20\%$)

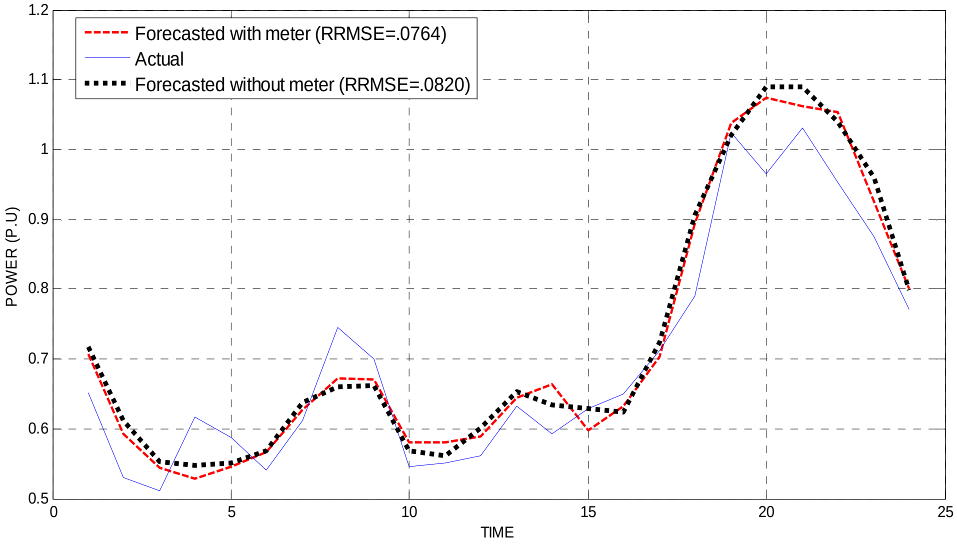


Figure 44. Case 3, Forecast of February 16 ($3\sigma = 20\%$)

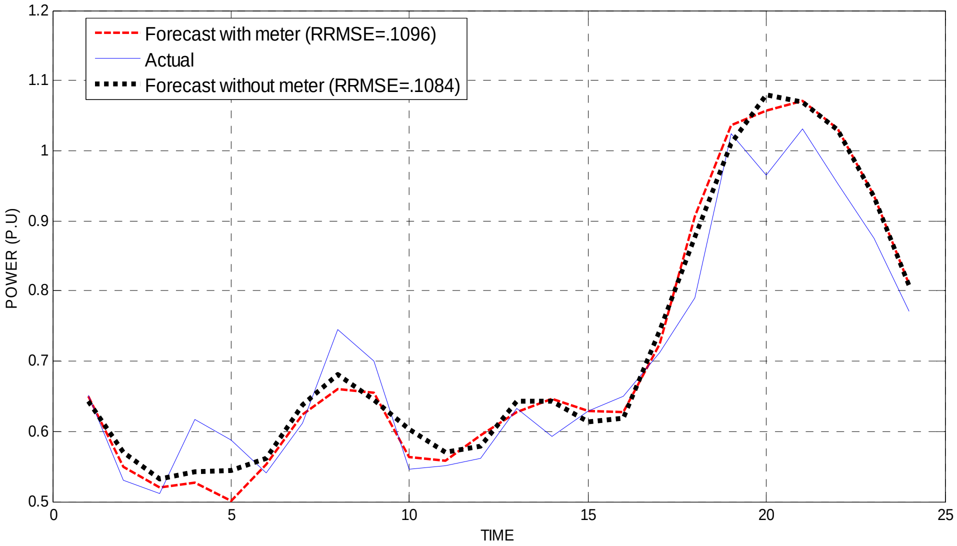


Figure 45. Case 3, Forecast of February 17 ($3\sigma = 20\%$)

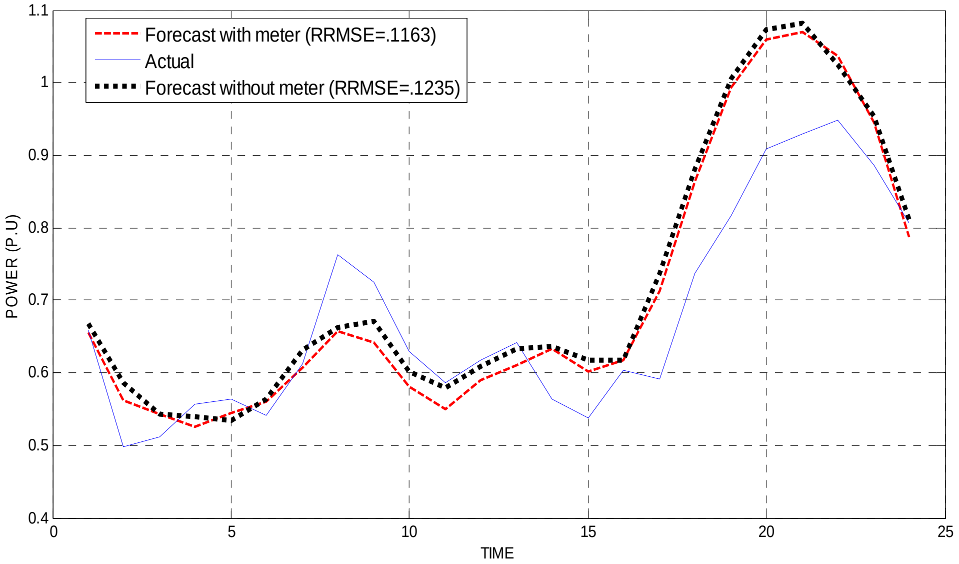


Figure 46. Case 3, Forecast of February 18 ($3\sigma = 20\%$)

As seen from Figures 42, 43 and 45 RRMSE for the predictions ‘Forecast with meter’ (using real-time data of a related meter for prediction purpose) is lower than just based on historical data (Forecast without meter).

Hence, as seen from these results, using real-time data of customers from the same homogenous cluster along with the available historical data helps to increase the accuracy of the forecasts.

Chapter 5

Conclusion and future work

5.1 Conclusion

On the distribution feeder, measurements for all the loads are not available all the times (meters are not installed at all the customer sites or due to some meter failures). A load estimation technique is required which can estimate the missing data about the customers. A new approach to model and predict the power consumption using an auto-regressive model, based on harmonic decomposition of the power consumption is proposed in this thesis. To

.

Furthermore, application of a clustering algorithm is presented, which is used to cluster or group customers based on similar consumption pattern. A non-hierarchical clustering technique ‘k-means clustering’ is implemented in SAS enterprise miner to group various consumer profiles into homogenous groups.

In a particular cluster, there may be some customers with missing data. The load modeling technique proposed in Chapter 3 was modified to estimate the missing data of the affected customers from their historical data and real-time data of other customers in their cluster. The load estimation technique explained in chapter 3 was used, along with available measurements as other predictor variables in the time series model. For a given cluster, AMI data belonging to that cluster can be modeled as another predictor variable, to estimate the load for those customers with missing real-time data.

As seen from the results in chapter 4, using real-time data of customers from the same homogenous cluster along with the available historical data helps to increase the accuracy of the estimates of power of meters with missing data.

5.2 Future work

The only data used in this work [22] was the power of the customers. Significant accuracy was achieved in the estimation of power values even without considering the real-time temperature of the location of consumers. With historical and real-time data about the temperature available, the time-series model could be modified and accuracy of results is expected to increase.

State Estimation is the main tool used for monitoring of distribution feeders and any load modeling technique helps in providing the pseudo-measurements to fill the voids created by bad meters or absence of meters due to any reason. The results of the proposed load estimation technique should be used as input to state estimation technique as pseudo-measurements and results should be accordingly assessed.

References

1. Haibin Wang; Schulz, N.N.; ‘A load modeling algorithm for distribution system state estimation’, *Transmission and Distribution Conference and Exposition, 2001 IEEE/PES.*

2. Ghosh, A.K.; Lubkeman, D.L.; Jones, R.H.; ‘Load Modeling for Distribution circuit state estimation’, *Power Delivery, IEEE Transactions* on Volume 12, Issue 2, April 1997

3. William Mendenhall, Terry Sincich. ‘*A second course in statistics: regression analysis*’, Pearson Education, 2003.

4. Rudolf J. Freund and Ramon C. Littell, ‘*SAS System for Regression, Third Edition*’, SAS Institute ,Cary ,NC.

5. Subhash Sharma, ‘*Applied Multivariate techniques’,* Wiley 1996.

6. SAS online manual , *support.sas.com/onlinedoc/913/docMainpage.jsp*

7. George Fernandez, ‘*Data Mining using SAS applications’,* CRC Press, 2003.

8. Mesut Baran; AA Freeman; Frank Hanson; Valleau Ayers, ‘Load Estimation for load monitoring at distribution substations’, *IEEE Transactions on power systems*, vol 20,no 1,February 2005.

9. Charytoniuk, W.; Chen, M.S.; Kotas, P.; Van Olinda, P.; 'Demand Forecasting in Power Distribution System using Non-Parametric Probability Density Estimation' *Power Systems, IEEE Transactions* onVolume 14, Issue 4, Nov. 1999 Page

10. Sargent, A.; Broadwater, R.P.; Thompson, J.C.; Nazarko, J.; 'Estimation of Diversity an KWHR-to-Peak KW Factors from Load research Data', *Power Systems, IEEE Transactions* on Volume 9, Issue 3, Aug. 1994

11. Chen, C.S.; Hwang, J.C.; Tzeng, Y.M.; Huang, C.W.; Cho, M.Y.; 'Determination of Customer Load Characteristics by Load Survey System at Taipower' *Power Delivery, IEEE Transactions* on Volume 11, Issue 3, July 1996

12. Atish K Ghosh,D.L Lubkeman,M.J Downey, R.H. Jones; 'Distribution Circuit State Estimation Using a Probabilistic Approach ', *IEEE Trans. On Power Systems*, Vol.12 ,No 1,Feb 1997,pp 45-51

13. Borozan,V and Rajakovic, N, *Load Estimation for Distribution Systems with Minimum Information.*

14. Westinghouse Distribution System, *Electric Utility Engineering Reference Book*,vol 3,First Edition,Fifth Printing,1965.

15. Lee, Y.C and Etezadi-Amoli,M, An Improved Modeling Technique for Distribution feeders with Incomplete Information, *IEEE transactions on Power Delivery*, Vol 8, No 4, October 1993, pp 1966-1972.

16. Rural Electrification Administration, *REA Bulletin 45-2 Demand Tables,* June 1963

17. Mesa, V.N , Pacific Gas and Electric Company, *Distribution feeder Calculations: a Load, Voltage and Line Loss Analysis of Distribution Circuits.*

18. Turan Gonen, *Electric Power Distribution System Engineering,* Mcgraw-Hill Book Company,1986

19. Zalewski W, Application of fuzzy inference to electric load clustering, *Power India conference,* 2006

20. Jin Wang; Xinran Li and Cailing Li, Load characteristics clustering based on improved FCM method, *International Conference on Industrial Technology*, 2008

21. Richard Johnson, *Applied Multivariate Statistical Analysis* , Prentice-Hall Inc 1992

22. Pacific Gas & Electric (PG&E) load profiles available at http:// www.pge.com/nots/rates/006fo4_class_load_prof.shtml

23. Duda R.O and Hart P.E, *Pattern classification and scene analysis,* New York, Wiley 1973

APPENDIX I

A) SAS output for fitting the time-series model on the yearly data with sets of 3 harmonics.

```
                          Yule-Walker Estimates

     SSE                  4.45212709    DFE                  17505
     MSE                   0.0002543    Root MSE           0.01595
     SBC                  -95156.085    AIC             -95272.652
     Regress R-Square         0.4210    Total R-Square      0.9909
     Durbin-Watson            1.9206


                                          Standard                 Approx
  Variable        DF     Estimate            Error    t Value    Pr > |t|

  Intercept        1       0.5708         0.001525     374.27      <.0001
  s1               1      -0.0226         0.002157     -10.47      <.0001
  s2               1       0.0251         0.002157      11.64      <.0001
  s3               1      -0.0170         0.002157      -7.86      <.0001
  c1               1    -0.003823         0.002156      -1.77      0.0762
  c2               1       0.0480         0.002156      22.24      <.0001
  c3               1     0.005355         0.002156       2.48      0.0130
  sd1              1      -0.1659         0.002086     -79.51      <.0001
  sd2              1      -0.1180         0.001693     -69.70      <.0001
  sd3              1     0.000637         0.001126       0.57      0.5716
  cd1              1      -0.0278         0.002085     -13.35      <.0001
  cd2              1      -0.0328         0.001693     -19.35      <.0001
  cd3              1      -0.0152         0.001126     -13.48      <.0001
```

B) SAS output for fitting the time-series model on the yearly data with sets of 5 harmonics.

```
                          Yule-Walker Estimates

     SSE                  4.18006471    DFE                  17497
     MSE                   0.0002389    Root MSE           0.01546
     SBC                  -96182.713    AIC             -96361.449
     Regress R-Square         0.4653    Total R-Square      0.9915
     Durbin-Watson            1.9577
```

```
                                 Standard                 Approx
Variable        DF     Estimate     Error    t Value    Pr > |t|

Intercept        1       0.5708  0.001667     342.39      <.0001
s1               1      -0.0226  0.002358      -9.58      <.0001
s2               1       0.0251  0.002358      10.64      <.0001
s3               1      -0.0170  0.002358      -7.19      <.0001
s4               1     0.001439  0.002358       0.61      0.5417
s5               1    -0.007191  0.002358      -3.05      0.0023
c1               1    -0.003810  0.002357      -1.62      0.1060
c2               1       0.0480  0.002357      20.36      <.0001
c3               1     0.005368  0.002357       2.28      0.0228
c4               1    -0.000014  0.002357      -0.01      0.9953
c5               1     0.006272  0.002357       2.66      0.0078
sd1              1      -0.1659  0.002096     -79.11      <.0001
sd2              1      -0.1180  0.001501     -78.62      <.0001
sd3              1     0.000637  0.000980       0.65      0.5159
sd4              1       0.0120  0.000652      18.37      <.0001
sd5              1      -0.0102  0.000456     -22.40      <.0001
cd1              1      -0.0278  0.002096     -13.28      <.0001
cd2              1      -0.0328  0.001501     -21.83      <.0001
cd3              1      -0.0152  0.000980     -15.49      <.0001
cd4              1    -0.007045  0.000652     -10.81      <.0001
cd5              1    -0.007013  0.000456     -15.38      <.0001
```

C) SAS output for fitting the time-series model on the yearly data with sets of 7 harmonics.

```
                        Yule-Walker Estimates

   SSE                 3.98212391    DFE                   17489
   MSE                  0.0002277    Root MSE            0.01509
   SBC                 -96954.445    AIC              -97195.349
   Regress R-Square        0.4895    Total R-Square       0.9919
   Durbin-Watson           2.0092


                                 Standard                 Approx
Variable        DF     Estimate     Error    t Value    Pr > |t|

Intercept        1       0.5708  0.001729     330.04      <.0001
s1               1      -0.0226  0.002447      -9.23      <.0001
s2               1       0.0251  0.002447      10.26      <.0001
s3               1      -0.0170  0.002447      -6.93      <.0001
s4               1     0.001439  0.002447       0.59      0.5564
s5               1    -0.007191  0.002446      -2.94      0.0033
```

```
s6               1      0.000120      0.002446        0.05       0.9609
s7               1      0.001257      0.002446        0.51       0.6072
c1               1     -0.003805      0.002445       -1.56       0.1197
c2               1        0.0480      0.002445       19.62       <.0001
c3               1      0.005373      0.002445        2.20       0.0280
c4               1     -9.277E-6      0.002445       -0.00       0.9970
c5               1      0.006277      0.002445        2.57       0.0103
c6               1     -0.000778      0.002445       -0.32       0.7502
c7               1      0.005307      0.002445        2.17       0.0300
sd1              1       -0.1659      0.002097      -79.08       <.0001
sd2              1       -0.1180      0.001439      -82.05       <.0001
sd3              1      0.000636      0.000931        0.68       0.4942
sd4              1        0.0120      0.000622       19.24       <.0001
sd5              1       -0.0102      0.000438      -23.31       <.0001
sd6              1     -0.005306      0.000324      -16.38       <.0001
sd7              1      0.005089      0.000250       20.39       <.0001
cd1              1       -0.0278      0.002097      -13.27       <.0001
cd2              1       -0.0328      0.001438      -22.78       <.0001
cd3              1       -0.0152      0.000931      -16.31       <.0001
cd4              1     -0.007046      0.000622      -11.32       <.0001
cd5              1     -0.007013      0.000438      -16.01       <.0001
cd6              1     -0.003646      0.000324      -11.25       <.0001
cd7              1     -0.001948      0.000250       -7.80       <.0001
```

D) SAS output for fitting the time-series model on the yearly data with sets of 10 harmonics.

```
                         Yule-Walker Estimates

   SSE                  3.83266524    DFE                    17477
   MSE                   0.0002193    Root MSE             0.01481
   SBC                  -97507.387    AIC               -97841.544
   Regress R-Square         0.5038    Total R-Square        0.9922
   Durbin-Watson            2.0524


                                       Standard                  Approx
Variable        DF      Estimate          Error     t Value    Pr > |t|

Intercept        1        0.5708      0.001731      329.76       <.0001
s1               1       -0.0226      0.002449       -9.22       <.0001
s2               1        0.0251      0.002449       10.25       <.0001
s3               1       -0.0170      0.002449       -6.92       <.0001
s4               1      0.001439      0.002449        0.59       0.5568
s5               1     -0.007191      0.002449       -2.94       0.0033
```

```
                                       Standard                 Approx
Variable        DF      Estimate          Error    t Value    Pr > |t|

s6               1      0.000120       0.002449       0.05      0.9610
s7               1      0.001257       0.002448       0.51      0.6076
s8               1     -0.001493       0.002448      -0.61      0.5419
s9               1      0.004631       0.002448       1.89      0.0586
s10              1     -0.005089       0.002448      -2.08      0.0377
c1               1     -0.003796       0.002447      -1.55      0.1209
c2               1        0.0480       0.002447      19.61      <.0001
c3               1      0.005382       0.002447       2.20      0.0279
c4               1     -1.815E-7       0.002447      -0.00      0.9999
c5               1      0.006286       0.002447       2.57      0.0102
c6               1     -0.000769       0.002447      -0.31      0.7533
c7               1      0.005316       0.002447       2.17      0.0298
c8               1     -0.001851       0.002447      -0.76      0.4493
c9               1     -0.004492       0.002447      -1.84      0.0664
c10              1     -0.002499       0.002447      -1.02      0.3071
sd1              1       -0.1659       0.002090     -79.36      <.0001
sd2              1       -0.1180       0.001423     -82.94      <.0001
sd3              1      0.000636       0.000917       0.69      0.4875
sd4              1        0.0120       0.000611      19.59      <.0001
sd5              1       -0.0102       0.000430     -23.76      <.0001
sd6              1     -0.005306       0.000318     -16.71      <.0001
sd7              1      0.005089       0.000245      20.80      <.0001
sd8              1      0.003152       0.000195      16.16      <.0001
sd9              1     -0.002912       0.000160     -18.19      <.0001
sd10             1     -0.000141       0.000135      -1.05      0.2957
cd1              1       -0.0278       0.002089     -13.31      <.0001
cd2              1       -0.0328       0.001423     -23.02      <.0001
cd3              1       -0.0152       0.000917     -16.56      <.0001
cd4              1     -0.007045       0.000611     -11.53      <.0001
cd5              1     -0.007013       0.000430     -16.31      <.0001
cd6              1     -0.003645       0.000318     -11.48      <.0001
cd7              1     -0.001948       0.000245      -7.96      <.0001
cd8              1     -0.000503       0.000195      -2.58      0.0100
cd9              1      0.001242       0.000160       7.76      <.0001
cd10             1      0.000404       0.000135       3.00      0.0027
```