Modeling and Forecasting Hourly Electric Load by Multiple Linear Regression with Interactions

Tao Hong[1, 2], Min Gui[1], Mesut E. Baran[3], *Senior Member, IEEE*, and H. Lee Willis[1], *Fellow, IEEE*

***Abstract* — Short-term electric load modeling and forecasting has been intensively studied during the past 50 years. With the emerging development of smart grid technologies, demand side management (DSM) starts to attract the attention of electric utilities again. To perform a decent DSM, beyond *when* and *how much* the demand will be, the utilities are facing another question: *why* is the electricity being consumed? In other words, what are the factors driving the fluctuation of the electric load at a particular time period? Understanding this issue can also be beneficial for the electric load forecasting with the purpose of energy purchase. This paper proposes a modern treatment of a classic technique, multiple linear regression, to model the hourly demand and investigate the causality of the consumption of electric energy. Various interactions are discovered, discussed, tested, and interpreted in this paper. The proposed approach has been used to generate the 3-year hourly energy demand forecast for a US utility.**

***Index Terms* — Load forecasting, load management, load modeling, multiple linear regression.**

I. Introduction

DURING the past several decades, the US electric utility industry has been trying to model the demand for various reasons. One of the primary reasons is load forecasting, of which the results can be further used in T&D planning, operations, demand side management (DSM), and energy purchase in the competitive energy markets. According to different forecast horizons and resolutions, load forecasts can be roughly grouped into long-term, mid-term, and short-term ones [1]. In most long-term load forecasts with T&D planning purposes, the questions of interest are *when*, *where*, and *how much* the load will be [2, 3]. In most short-term and mid-term load forecasts with energy purchase and DSM purposes, the core questions to be answered for these purposes are *when* and *how much* the demand will be. Nowadays, with the development of smart grid technologies, DSM becomes an emerging emphasis in the electric utility industry. Effective DSM requires utilities to understand the patterns of the consumption of electricity, and to perform peak load reduction accordingly. This emerging need leads to another question beyond *when* and *how much*: *why* is the electricity being consumed? A proper answer would not only be beneficial to the traditional forecasting process, but also help the utilities decide whose load should be reduced or cut during the peak load hours, e.g., residential customers, office buildings, or industrial customers, etc.

A large variety of pioneer research and practices have been devoted to modeling and forecasting the short-term electric load in *when* and *how much* level. Lots of these methods in the open literature can be roughly classified into two broad categories: artificial intelligence (AI) and statistical approaches. Among all the AI-based approaches, artificial neural networks (ANN) may have received the most attention by both academia and industry [4, 5]. Not only in short term load forecasting, ANN has also been tried in several other utility applications, such as distribution system failure forecast [6], with varying degrees of success. As a data-driven approach, ANN has been favored by several major vendors in the power industry due to its flexibility and the advantage of handling the nonlinearity of the data. Although empirically the ANN models can produce fairly acceptable forecast, as a black-box approach, ANN, together with most other AI-based approaches, such as fuzzy logic [7, 8], have not been completely convincing. The lack of interpretability and over-fitting issue are still challenging problems in the research field of applying ANN to electric load modeling and forecasting.

On the other hand, statistical approaches, such as similar day, time series, and regression,, are favored by the practitioners due to both accuracy and interpretability. Similar day approach, as one of the earliest methods for short term load forecasting, is still being used by many utilities due to the simplicity of implementation as well as acceptable results. The basic idea of this approach is to pick up the historical days with similar characteristics to the forecasted day as the major reference when performing the short term load forecasting [9]. Time series approaches have been extensively used by utilities in short term forecast [10]. Due to the involvement of autoregressive and/or moving average terms, time series models have the advantage on the short term accuracy, while their interpretability is not as straightforward as that of the regression models.

Regression, or more specifically, multiple linear regression is the most widely applied statistical technique in short-term and mid-term electric load forecasting [11-13]. In the aspect of load modeling, the regression-based models have also been applied to a wide field in power systems such as load monitoring at distribution substations [14]. The idea of multiple linear regression is to model the electric load as a

1. Quanta Technology, LLC, Raleigh, NC 27607 USA

2. Operations Research Graduate Program, North Carolina State Univ., Raleigh, NC 27695

3. Electric & Computer Engineering Department, North Carolina State Univ., Raleigh, NC 27695

978-1-4244-6551-4/10/$26.00 ©2010 IEEE

linear function of several independent quantitative variables, e.g., temperature, and independent dummy variables, e.g., time of the day and day of the week. However, due to the quality of data sources and limited capability of computers, the power of regression methods could not be fully demonstrated in the old days when it started to be deployed in the utilities. Modern techniques allow utilities to record and store historical data of load and weather in sufficiently high quality in terms of both resolution and correctness. Meanwhile, modern computing environment allows engineers and researchers to perform computational intensive analysis over large scale data. For instance, in the case of applying the regression analysis to the dataset containing 4 years of hourly load and temperature, it takes less than 5 minutes to update the parameters of the proposed model in a computer with a 3.5GB RAM and 2.2GHz CPU.

This paper applies a multiple linear regression method to identify and quantify the factors affecting the electric load. The affecting factors are obtained through an iterative process of plotting, interpreting, and testing. With various interactions (cross effects), the resulting model offers the insights of seasonal patterns and the variations within each of the three types of seasonal components, e.g., year, week, and day. The proposed modeling approach has been used to generate the 3-year electric load forecast which is deployed in the energy purchase process for a US utility company.

The rest of this paper is organized as following: Section II reviews the fundamentals of multiple linear regression and the diagnostic statistics used in this paper. Section III identifies the trend of the load and the relationship between the temperature and the load through the analysis of characteristics of the load. Section IV studies the interactions of multiple seasonal components through the analysis of several plots and tests of the trial models. The paper is concluded in Section V with discussion of the findings and future work.

II. Multiple Linear Regression and Diagnostic Statistics

A. Multiple Linear Regression

In the area of electric energy demand modeling, the purpose of multiple linear regression is to model the relationship between several independent variables (e.g., temperature and time of the day) and a dependent variable (electric load) as a linear function. For instance, a multiple linear regression with two independent variables can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e, \tag{1}$$

where Y is the dependent variable, X_1 and X_2 are independent variables, β's are parameters to estimate, and e is the error term. In multiple linear regression, the error term represents a set of random variables independent and identically distributed with a Gaussian distribution having zero mean.

In this paper, the independent variables can be classified into three categories:

1) Quantitative variables: temperature, temperature square, etc.
2) Dummy variables: hour of the day, month of the year, etc.
3) Interactions: multiplications among quantitative variables and dummy variables.

The boundary among the above three categories are sometimes ambiguous. Month, a dummy variable with 12 levels can be taken as 12 binary quantitative variables, one for each month. Square of temperature, which we classify as a quantitative variable, can also be taken as an interaction between temperature and itself. An interaction between hour of the day (24 levels) and day of the week (7 levels) can also be taken as a new dummy variable with 24 x 7 = 168 levels. In this paper, the first two categories are called main effects, while the cross effects are defined as the multiplication of two or more different main effects. When a dummy variable interacts with a quantitative variable, this quantitative variable can be taken out from the set of main effects without affecting the diagnostic statistics. When two dummy variables interact together, both of them can be taken out from the set of main effects. The details of selecting these variables will be introduced in the next section.

B. Diagnostic Statistics

Goodness-of-fit statistics are the diagnostic statistics calculated using the training sample (the sample that was used to fit the model), while accuracy statistics are the diagnostic statistics calculated using a holdout sample (the sample that was not used in modeling). The winning model should be assessed using accuracy statistics, while the goodness-of-fit statistics tend to give an optimistic estimate of implementation accuracy. The following goodness-of-fit statistics and accuracy statistics are used in this paper: adjusted R-square, mean of the absolute error (MAE), standard deviation of the absolute error (STDAE), mean absolute percentage error (MAPE), and standard deviation of the absolute percentage error (STDAPE).

Four years (2005 – 2008) of electric load and hourly temperature history from a US utility is used in this study. Cross validation is implemented by taking any three years of data as the training sample and the left one as the hold-out sample. There is no major different conclusion drawn for the variable selection but only minor quantitative difference in parameter estimation. For the simplicity of discussion, the rest of this paper is based on using 2005 – 2007 data as the training sample and 2008 as the hold-out sample.

III. Trend and Temperature Effect

A. Overall Characteristics of the Load

The analysis in this paper starts at a yearly resolution. The four years of system level load (in KW) and temperature (in F) are plotted in Fig. 1 and Fig. 2 respectively. As mentioned above, the first three years (2005 – 2007) of load and temperature data are used for modeling and discussion, while the load in 2008 is used as the hold-out sample to test the model given the actual temperature in 2008. The following observations can be obtained from the two figures:

1) There is an overall increasing trend year by year. This trend may be due to the increase of temperature and/or human activities. Since it is hard to draw the conclusion from Fig. 2 that the temperature has an increasing or decreasing trend, we can infer that human activities are the major cause of the increasing trend in the electricity consumption over the years. The human activities may include the local economic development, which results in more customers, and/or change of end use behaviors.

2) There is one peak load period in winter and another peak load period in summer for every year, which seems to be a seasonal pattern. Comparing Fig 1 to Fig. 2, the winter peak load occurs during the valley temperature period in winter, and the summer peak load occurs during the peak temperature period in summer, which provides the evidence that the seasonality of the temperature leads to the seasonality of the electricity consumption. Yet, it is not clear at the moment whether there is a seasonality of human activities that affect the load in the monthly resolution.

3) The summer peak is higher than the winter peak for every year, which tells that electricity is primarily used for cooling in summer. Also it indicates that some electricity is used for warming in the winter.

4) This is one valley in spring and the other valley in fall for every year, which is the result of less need of A/C comparing with winter and summer.

With the above observations, we define a quantitative variable (Trend) to capture the increasing trend by assigning the natural number to each hour in the natural order. For instance, the Trend variable of the first hour in 2005 is 1, the second hour in 2005 is 2, and so forth.

B. Temperature

To further study the correlation between the demand and the temperature, we draw the scatter plot of the load (vertical axis) and the temperature (horizontal axis) in Fig. 3, where we can observe that the upper boundary of load-temperature plot shows nonsymmetrical V-shape, and the lower boundary of load-temperature plot shows nonsymmetrical U-shape

Since the plot shows overall nonsymmetrical shape, a piecewise linear or piecewise quadratic function can be used to model the relationship between load and temperature (suggestion: explain the quadratic term doesn't violate the definition of multiple linear regression). With some experiments, a piecewise nonlinear function is preferred in this dataset with a cut-off at 65F. Therefore, we define a dummy variable (TMPID) by assigning 1 to the temperatures smaller than 65, and 2 to the temperatures greater than or equal to 65. Meanwhile temperature (TMP) and the square of temperature (TMP2) are to be considered as quantative variables in the model.

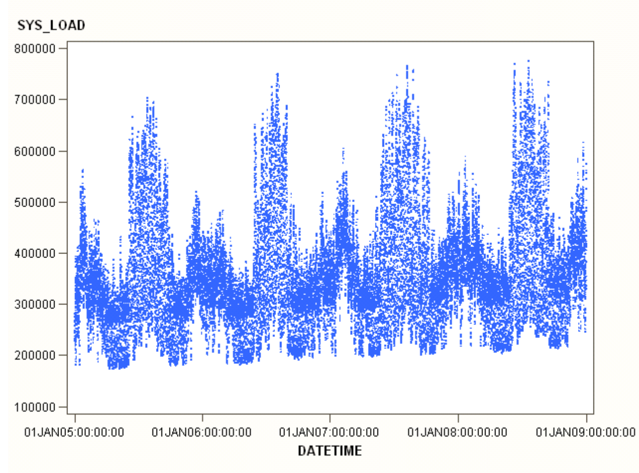


Fig. 1. Load plot (2005 - 2008)

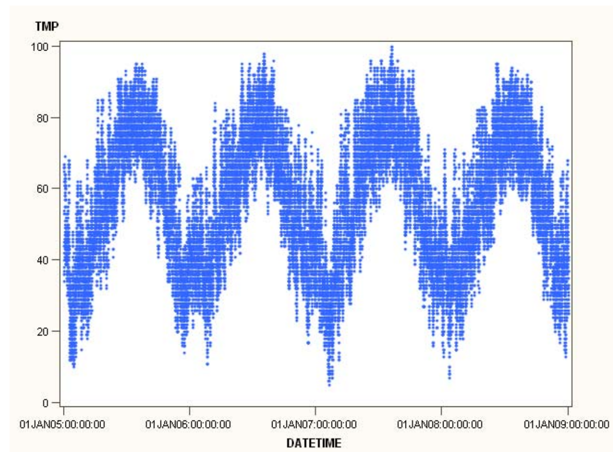


Fig. 2. Temperature plot (2005 - 2008)

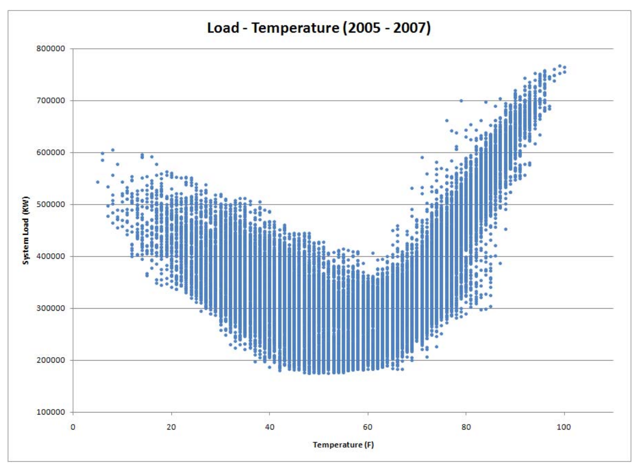


Fig. 3. Load-temperature scatter plot (3 years)

IV. SEASONALITY

As we have observed in Fig.1 and Fig. 2, seasonality of the load may be driven by seasonality of the temperature and seasonality of the human activities. There are three major seasonal blocks in out multi-year hourly data: year, week, and day. In this section, effects of temperature and human activities will be discussed for each seasonal block.

A. Month of the Year

In general, summer has high temperature, while winter has low temperature. Hence, the season and the pattern of temperature are correlated. To study whether the seasonality of the load at monthly resolution is caused by weather, or human activities in different seasons/months, or both, we draw the 12 scatter plots of load and temperature, one for each month, in Fig. 3. And we have the following observations:

1) The pieces of the 12 plots can be put together to construct the load-temperature plots shown in Fig. 4. For example, plots for June through September contribute to the right part of Fig. 3; plots for December through March contribute to the left part; and the rest plots contribute to the bottom part.

2) The levels of the plots appear to vary in different months. For instance, the level of load plot in March is higher than that of April. Namely the same temperature tends to result in higher load in March than that in April, which can be explained by different human activities in those two months, e.g., changing mode of the A/C, or both human activities and temperature, e.g., taking a shower in March will lead to more electricity for water heater than in April.

The above observations suggest that we can use the same method described in Section III to model the relationship between load and temperature for each month individually. Therefore, we consider the month of the year as a dummy variable (Month), and we also consider the interaction terms TMP*TMPID*Month and TMP2*TMPID*Month.

B. Hour of the Day

Due to sunrise and sunset, time of the day and the temperature are correlated. To study the seasonality of the load at daily resolution, we draw the 24 scatter plots of load and temperature, one for each hour of the day, in Fig. 4. And we have the following observations:

1) The pieces of the 24 plots can be put together to construct the load-temperature plots shown in Fig. 2. Each of them has similar shape to, but thinner than the one shown in Fig. 2, which suggests that the piecewise quadratic function can be considered.

2) The (vertical) levels of the plots appear to be different. For instance, given the same temperature the load at 21:00 is higher than that at 23:00, which suggests that different functions can be used for different hours.

3) The horizontal positions of the plots appear to be different. For instance, the load plot at 18:00 seems to be shifted to the left from that at 21:00. This also suggests using different functions to model the load-temperature relationship for different hours. On the other hand, it also indicates that the temperature during the previous hour may affect the load during the current hour. This is also commonly known as heat build-up effect.

Therefore, we consider the hour of the day as a dummy variable (Hour), and we also consider the interaction terms TMP*TMPID*Hour and TMP2*TMPID*Hour. To model the heat build-up effect, we consider the interaction term DTMP*TMPID*Hour, where DTMP represents the difference between the current temperature and previous hour temperature (DTMP = current temperature – previous hour temperature).

C. Day of the Week

Since there is no statistically significant differences among the weather conditions over the days in a week, the electric load in the weekly block are mainly affected by the human activities. Human activities are different in weekdays and weekends. To capture the differences, we use a dummy variable to represent the days of the week. And consider the cross effect between this dummy variable and Hour in the model. There are two classic ways to assign the values to this dummy variable in the literature. One (D1) is to define three day types: Weekdays, Saturdays, and Sundays [12]. The other one (D2) is to define four day types: Mondays, other weekdays, Saturdays, and Sundays [7]. We have tried both of these two classifications in our model.

D. Summary and Results

We build the model using the variables discussed above. A trial-and-error approach is applied. Table I shows the variable setup for 9 representative models. An "X" sign indicates that the corresponding variable is selected in the model. For instance, in M3, there are four types of variables: Trend, TMP*TMPID, TMP2*TMPID, and Month. As we mentioned earlier, the data from the year 2005 to 2007 is used as training sample and the data from the year 2008 is used as testing sample. The diagnostic statistics are listed in TABLEII from which we can see that when the complexity of the models increases, the performance of the models increases. Among these nine models, improvements of goodness-of-fit always results in improvements of accuracy. Also, when the MAPE decreases, the standard deviation of the APE decreases too.

With a trend and a piecewise linear and quadratic term of temperature, M2 gives hourly forecasts for the year 2008 with a MAPE of 12.785%. With considering cross effects between temperature terms and dummy variable month of the year in M4 the forecast MAPE is brought down to 11.475%. The addition of cross effects between hour of the day and temperature terms in M6 greatly improved the model performance and reduced the forecast MAPE to 5.215%. The consideration of DTMP in M7 further reduced the forecast MAPE by around 0.1%. Finally the cross effect between day of the week and hour of the day in M9 resulted in a forecast MAPE of 4.558%, which is approximately 0.5% improvement in the forecast MAPE comparing with M7.

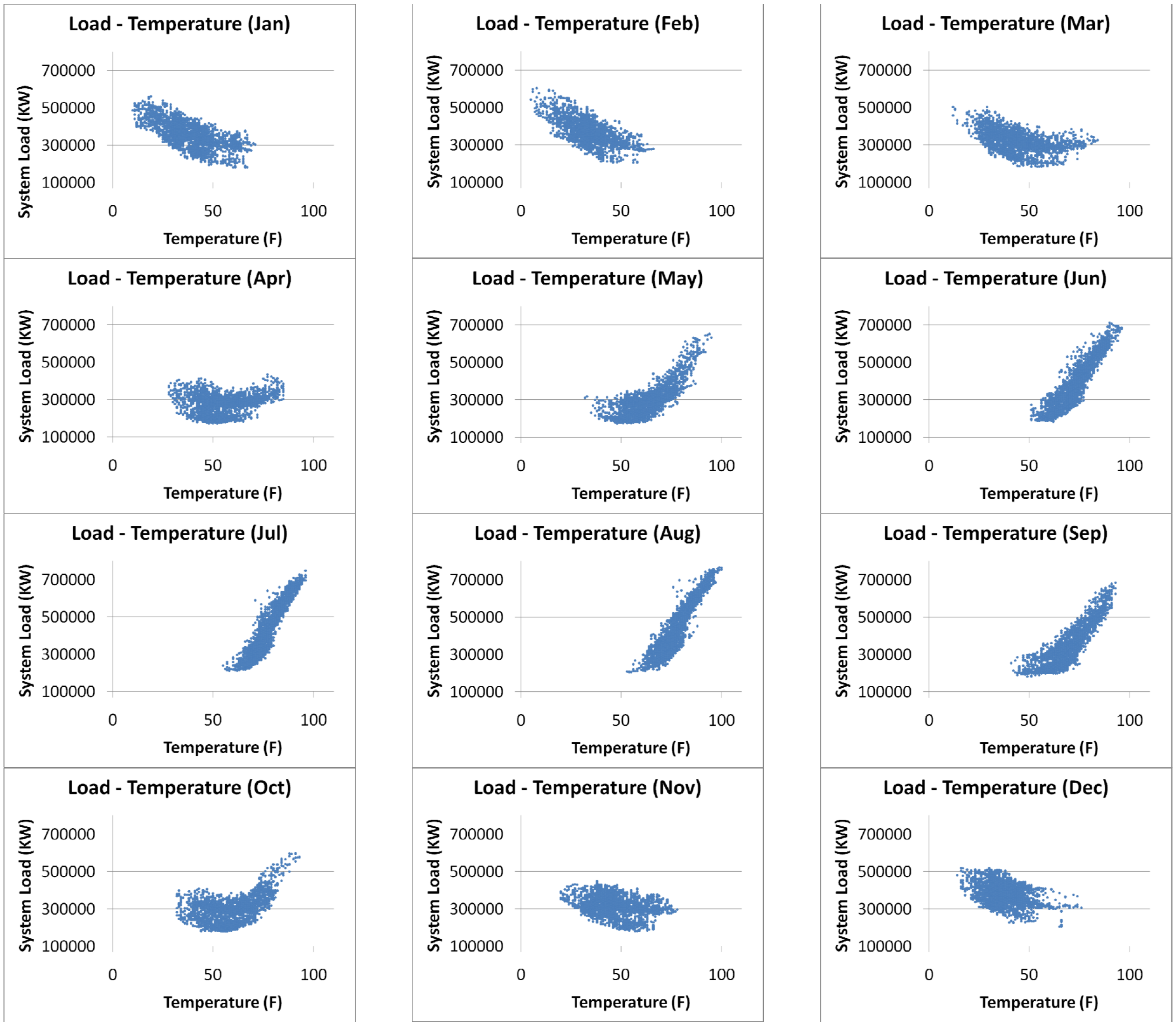
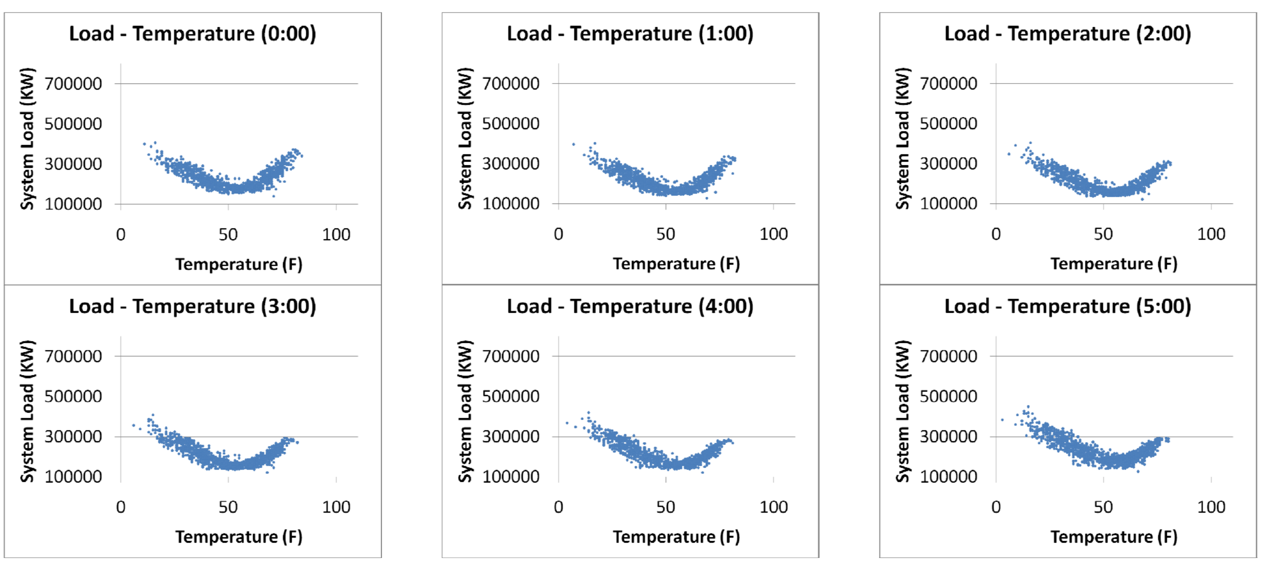


Fig. 4. Load-temperature scatter plots (12 plots, one for each month)



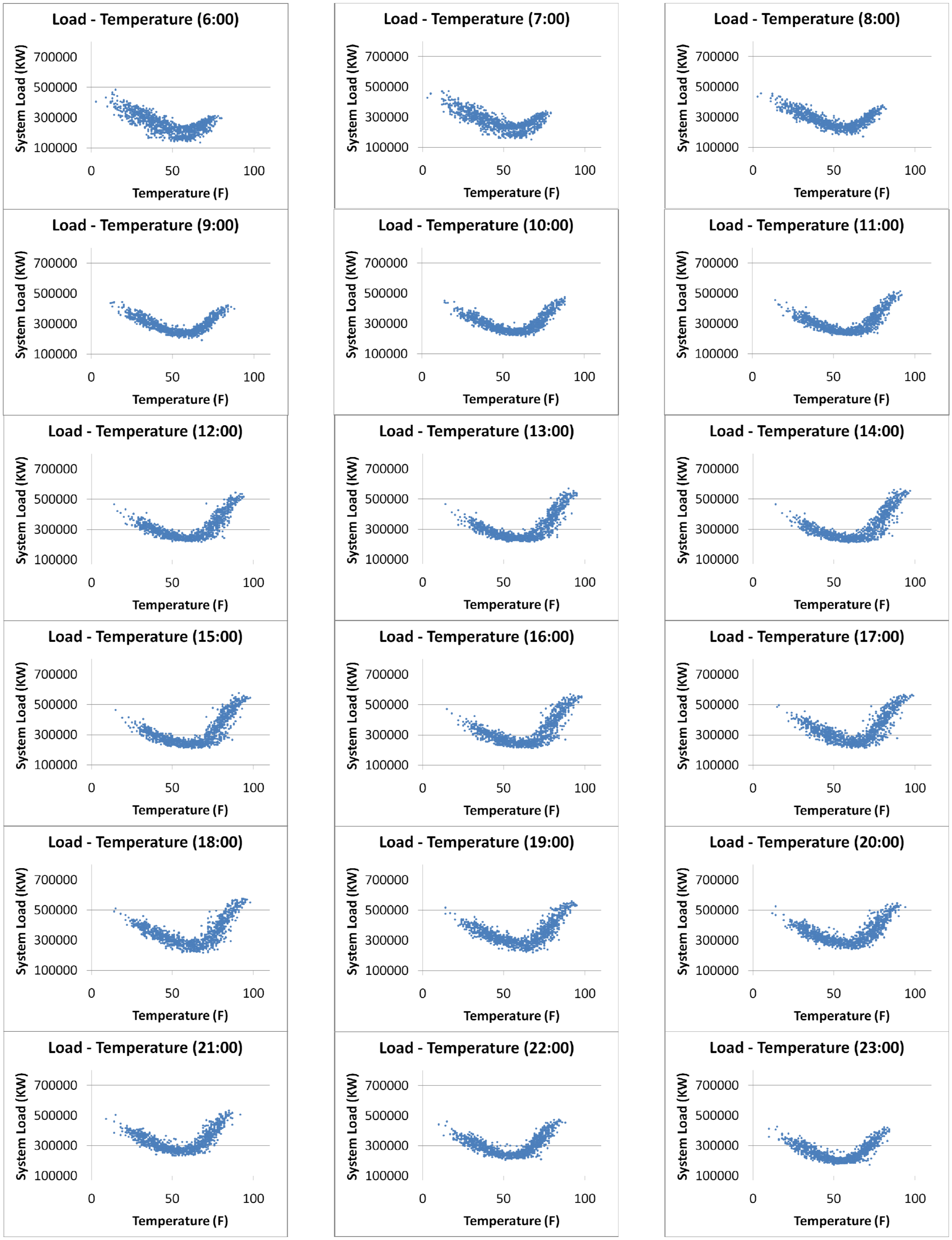


Fig. 5. Load-temperature scatter plots (24 plots, one for each hour)

TABLE I. MODEL SETUP

Test#	Trend	TMP TMPID	TMP2 TMPID	Month	TMP TMPID Month	TMP2 TMPID Month	Hour	TMP TMPID Hour	TMP2 TMPID Hour	DTMP TMPID Hour	D1 Hour	D2 Hour
M1	X											
M2	X	X	X									
M3	X	X	X	X								
M4	X			X	X	X						
M5	X			X	X	X	X					
M6	X			X	X	X	X	X	X			
M7	X			X	X	X	X	X	X	X		
M8	X			X	X	X		X	X	X	X	
M9	X			X	X	X		X	X	X		X

TABLE II. DIAGNOSTIC STATISTICS

		Goodness-of-fit Statistics			Accuracy Statistics			
Strategy #	DF	Adj. R-Square	MAPE	STDAPE	MAPE	STDAPE	MAE	STDAE
M1	1	0.0188	23.139	19.131	22.552	18.537	80.854	65.446
M2	5	0.7506	13.127	11.192	12.785	10.302	45.115	31.812
M3	16	0.7742	12.538	10.267	12.244	9.131	43.473	29.520
M4	59	0.7977	11.615	10.041	11.475	9.286	40.560	29.678
M5	82	0.9425	5.326	4.565	5.464	4.482	20.872	18.326
M6	174	0.9500	4.921	4.360	5.215	4.307	19.851	17.478
M7	222	0.9531	4.823	4.244	5.107	4.200	19.430	16.881
M8	270	0.9611	4.228	3.574	4.562	3.620	17.711	16.056
M9	294	0.9615	4.199	3.547	4.558	3.595	17.690	15.946

V. CONCLUSION

This paper models the hourly electric load by multiple linear regression. Several variables involved in the model are discussed through various plots. The proposed regression model has been used to generate the 3-year hourly demand forecast, which has been in place of the existing mid-term load forecast of a US utility.

The proposed approach has good interpretability of the behavior of the electricity consumption in the service territory, which helps the power engineers understand the system load profile when performing DSM. What's more, a load forecast with satisfying accuracy also provides a good reference for the energy purchasing purpose.

With considering cross effects among the explanatory variables, the performance of the model has been significantly improved with a MAPE of 4.558%. Moreover, the proposed model still can be further fine tuned to obtain more accurate forecasting results. For instance, the model can be improved when holidays are considered specifically. Time series models can be applied to model the residuals of the proposed model when doing the short-term load forecasting.

Another way to improve the forecast accuracy is to enhance the quality of the data. For example, the historical loads in this paper are obtained by summing up the actual measures at the low side of the transformers and the estimated losses. With enhanced losses estimation methodologies, the load data can be more accurate, which will potentially result in a better forecast [15].

VI. REFERENCES

[1] H. L. Willis, "Spatial Electric Load Forecasting", 2nd ed., New York: Marcel Dekker, Inc., 2002, pp. 1-35.

[2] T. Hong, S. M. Hsiang, L. Xu, "Human Machine Co-construct Intelligence on Horizon Year Load in Long Term Spatial Load Forecasting", *IEEE Power and Energy Society General Meeting*, Calgary, Alberta, Canada, July 26-30, 2009.

[3] T. Hong, "Spatial Load Forecasting Using Human Machine Co-construct Intelligence Framework". Master thesis, Operations Research Graduate Program, North Carolina State University, Dec 2008.

[4] H. S. Hippert, C. E. Pedreira, "Estimating temperature profiles for short-term load forecasting: neural networks compared to linear models," *IEE Proceedings- Generation, Transmission and Distribution*, vol.151, no.4, pp. 543-547, 11 July 2004.

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

[6] M. Gui, A. Pahwa; S. Das, "Analysis of Animal-Related Outages in Overhead Distribution Systems With Wavelet Decomposition and Immune Systems-Based Neural Networks," *IEEE Transactions on Power Systems*, vol.24, no.4, pp.1765-1771, Nov. 2009.

[7] K. Song, Y. Baek, D. H. Hong, G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Transactions on Power Systems*, vol.20, no.1, pp. 96-101, Feb. 2005.

[8] R. H. Liang,; C. C. Cheng, "Combined regression-fuzzy approach for short-term load forecasting," *IEE Proceedings- Generation, Transmission and Distribution*, vol.147, no.4, pp.261-266, Jul 2000.

[9] R. Weron, "Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach", John Wiley & Sons Ltd, England, 2006, pp79.

[10] S. J. Huang; K. R. Shih, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," *IEEE Transactions on Power Systems*, vol.18, no.2, pp. 673-679, May 2003

[11] A. D. Papalexopoulos, T. C. Hesterberg, "A regression-based approach to short-term system load forecasting ," *IEEE Transactions on Power Systems*, vol.5, no.4, pp.1535-1547, Nov 1990.

[12] B. Krogh, E. S. de Llinas, D. Lesser, "Design and Implementation of An on-Line Load Forecasting Algorithm," *IEEE Transactions on Power Apparatus and Systems*, vol.PAS-101, no.9, pp.3284-3289, Sept. 1982.

[13] G. T. Heinemann, D. A. Nordmian, E. C. Plant, "The Relationship Between Summer Weather and Summer Loads - A Regression Analysis," *IEEE Transactions on Power Apparatus and Systems* , vol.PAS-85, no.11, pp.1144-1154, Nov. 1966.

[14] M. E. Baran, L.A.A. Freeman, F. Hanson, V. Ayers, "Load estimation for load monitoring at distribution substations," *IEEE Transactions on Power Systems*, vol.20, no.1, pp. 164-170, Feb. 2005.

[15] T. Hong, J. J. Burke, "Calculating Line Losses in Smart Grid: A New Rule of Thumb", accepted by *IEEE PES Transmission and Distribution Conference and Exposition*, New Orleans, Louisiana, Apr 19-22, 2010.

VII. Biographies

Tao Hong (S'06) is a Senior Engineer in Quanta Technology. His major areas of expertise are in operations research and its related applications in power distribution engineering and T&D planning. He has applied various statistics and optimization techniques to T&D loss studies, development of long term and short term load forecasting algorithms and tools. His work has been applied to many US utilities. He received his Bachelor of Engineering degree in Automation from Tsinghua University, Beijing, a Master of Science degree in Electrical Engineering, and a Master of Science degree with co-majors in Operation Research and Industrial Engineering from North Carolina State University, where he is currently pursuing a PhD in Operations Research. Mr. Tao Hong is a winner of the poster competition in SAS' 12th Annual Data Mining Conference in 2009 for his work "*Behavior Mining of Electric Load Consumption: A Regression Approach*".

Min Gui (S'07) received the Ph.D. degree in Electrical Engineering from Kansas State University, Manhattan, USA, in 2009. Her employment experience includes HuNan XiangNengXuJi High Technique Company in China and the Electrical and Computer Engineering Department, Kansas State University. Since June, 2009, she has been working as a Senior Engineer at Quanta Technology, Raleigh, NC. Her special fields of interest includes energy modeling, time series forecasting, optimization and artificial intelligence. Ms. Gui received the second prize in the student poster-paper session and contest at the IEEE Power Engineering Society General Meeting in Tampa, FL, in June 2007.

Mesut E. Baran (S'87–M'88–SM'05) received the Ph.D. degree from the University of California, Berkeley, in 1988. He is currently a Professor with North Carolina State University, Raleigh. His current research interests include distribution and transmission system analysis and control, integration of renewable energy resources, and utility applications of power electronics-based devices.

H. Lee Willis, PE, (F'92) is a Senior Vice President and Executive Advisor in Quanta Technology. He has more than 35 years of electric T&D systems planning and engineering experience. He has been transmission planning manager at a major investor-owned utility, an executive with a major equipment supplier, and a senior consultant who has directly performed or supervised over 400 system planning and asset management projects for utilities around the world. Lee pioneered many of the modern planning and asset management methods now considered industry best practice, including spatial simulation load forecasting for T&D planning, load-reach voltage planning of feeder system layouts, and options-based asset management prioritization. He is a Fellow of the IEEE and a past chairman of its Power System Planning and Implementation committee. From 1999 to 2005 he served on the US National Research Council, which advises the US Congress and Dept. of Commerce on the nation's civilian technology needs. Lee has published more than 240 papers and articles on power systems, forecasting and planning, and utility practices during his career, including 78 papers in peer-reviewed and refereed technical journals, one theme cover article in the IEEE-wide *Proceedings of the IEEE,* and five cover articles for the *IEEE Computer Applications in Power* journal. He is the author of eight books on power system planning including the *Power Distribution Planning Reference Book* and *Spatial Electric Load Forecasting*.