UTILIZING TIME SERIES ANALYSIS TO FORECAST LONG-TERM ELECTRICAL CONSUMPTION

Danny Robert Modlin

A Thesis Submitted to
University of North Carolina Wilmington in Partial Fulfillment
Of the Requirements for the Degree of
Master of Science

Department of Mathematics and Statistics

University of North Carolina Wilmington

2006

Approved by

Advisory Committee

Chair

Accepted by

Dean, Graduate School

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Developing an accurate forecast model for the amount of power consumed will include such factors as time of day, day of the year, the weather, among many others. Based upon these given factors, current models use a neural network approach to forecast in the very near future. For the purpose of business operations, this model should be accurate for predicting power usage at least six months into the future. Using regression with time series analysis, the goal is to build a model that reflects systematic movements in the data and predict them so errors would be more or less random and minimized.

ACKNOWLEDGMENTS

There are several individuals without whom I could not have completed this thesis. First of all, I would like to thank all the mathematics and statistics professors at UNCW for their guidance and support. I would like to especially thank Dr. Frierson and Dr. TenHuisen for serving on my thesis committee and Dr. Blum for being my thesis advisor. I would also like to express my gratitude to Mike Settlage from Progress Energy. His assistance in this thesis is greatly appreciated.

I would also like to thank my parents, Bobby and Mary, and my siblings, Loria and Mike. You continually pushed me to do my best and to strive to attain the highest level that I could. Thank you greatly.

1 INTRODUCTION

Power consumption is a topic that everyone can easily discuss, especially on the day that the power bill arrives, as people have strong opinions as to what factors affect the consumption of power. Some, in efforts to lower their bills, try to predict what their power usage the next month will be and act accordingly. Similarly, power providers perform this task as well.

Progress Energy is a Fortune 250 energy company serving customers in North Carolina, South Carolina, Georgia, and Florida. Approximately 75% of the power generated by Progress Energy is consumed by industrial users or by residential users. The remaining 25% is used in the generation of electricity. Accurate forecasting of demand will allow for a more reliable and more cost efficient delivery of energy. However, building a model to minimize forecast errors is a challenging task since there are numerous factors that drive power consumption.

Given historical weather and power consumption data within the state of Georgia (Figure 1) the goal is to build a forecast model that accurately predicts the power demand for any time in the future. The emphasized regions are those for which Progress Energy provides power, with the blue dots being locations of weather stations, usually associated with television stations.

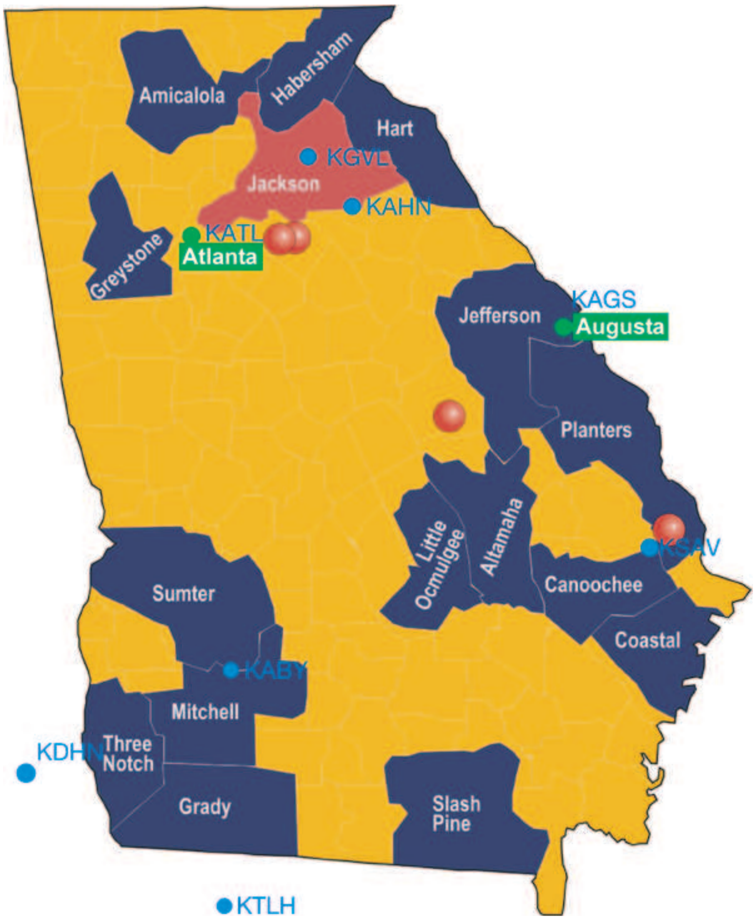

Figure 1: Regions Served by Progress Energy in Georgia

2 TIME SERIES BACKGROUND

"Today, time is what it was to Isaac Newton, a smoothly flowing stream bearing the phenomenal world along at a uniform pace." [7] With relative ease, people can plot points in time and measure distances between them with great accuracy. Whether it is with candles, clocks, or calendars, mankind has concerned itself with measuring the passage of time. On the contrary, the study of time series as a science is of recent origin. Until 1925, a time series was regarded as being generated deterministically. Any departures from trends, cycles, or other patterns of behavior observed in nature were regarded as errors similar to errors in observation. In 1927, George Yule embarked upon new ground with an idea that dictated much of the subsequent work in time series analysis. While working with sun-spot numbers, which do not fluctuate simply by chance, Yule was struck by irregularities in the series, both in amplitude and in the distances between successive peaks and troughs. His illustration, which lead to the theory of stochastic processes, was:

> If we have a rigid pendulum swinging under gravity through a small arc, its motion is well known to be harmonic, that is to say, it can be represented by a sine or cosine wave, and the amplitudes are constant, as are the periods of the swing. But if a small boy now pelts the pendulum with peas, the motion is disturbed. The pendulum will still swing, but with irregular amplitudes and intervals. Instead of leading to behavior in which any difference between theory and observation is attributable to an evanescent error, the peas provide a series of shocks which are incorporated into the future motion of the system. [7]

Definition 1 *A time series is a set of observations ordered in time.*

Definition 2 *When a variable is defined at all points in time, we say that the time series is continuous. If the set of time points is countable then the time series is discrete.*

Examples of a continuous time series include the temperature at a given location, the price of a commodity on the open market, or the position of a projectile moving through space. Some are created by collection over an established period, such as rainfall, industrial production, or total passenger miles recorded by an airline. Others are defined only at discrete points in time such as annual crop yields of a harvest or monthly salaries. In some of these cases the choice of when data is collected may be limited. For example, information is released only at specific times as with the monthly publication of many government statistics. Under other situations, the recording times may be freely chosen. For example, surveys of political opinion may be carried out as often as funding allows, or members of a medical staff may check a patient's pulse every hour. Finally, a continuous time series can be made into a discrete time series. For example, temperature and barometric pressure, or the alpha rhythm of the brain on an encephalograph, may be digitized and the continuous record converted into a time series reported at regular intervals. All applications of statistical analysis involve using a discrete time series that assumes that the data is recorded at regular points or over regular intervals of time.

According to Kendall and Ord, there are several major types of investigation in time series analysis.

1. Describe its behavior in a concise way.
2. Explain the behavior of a series in terms of other variables.
3. Forecast the behavior of the series in the future.
4. Control a system, either by generating warning signals of future unknown events or by examining what would happen if we alter either the inputs to the system or its parameters.

Regardless of which investigation type is used during time series analysis, it is useful to consider the typical series as a potential mixture of four components:

- a trend, or long-term movement;
- fluctuations about the trend of greater or less regularity;

- a seasonal component;
- a residual, irregular, or random effect.

It is convenient to represent the series as a sum of these four components, and one of the objectives may be to break the series down into its components for individual study. It may be reasonable to suppose that trends are due to permanent forces operating uniformly in more or less the same direction, that short-term fluctuations about these long movements are due to a different set of causes, and that there is in both some disturbance attributable to random events, giving rise to the residual. The effects of all the causes are additive and one must be ready to discard the model if it fails to fit the data. [7]

3 DETRENDING THE DATA SET

3.1 Background

Definition 3 *Trend in a time series is a slow, gradual change in some property of the series over the whole interval under investigation.*

Trend is sometimes loosely defined as a long term change in the mean, but can also refer to change in other statistical properties. Applying traditional time series analysis, a time series can be decomposed into trend, seasonal or periodic components, and irregular fluctuations, and in our data, the various parts were studied separately. Modern analysis techniques frequently treat the series without such routine decomposition, but separate consideration of trend is still often required.

Definition 4 *Detrending is the statistical or mathematical operation of removing trend from the series.*

Detrending is often applied to remove a long-term feature thought to obscure other relationships of interest. Detrending is also sometimes used as a preprocessing step to prepare time series for analysis by other methods. Meko [8] lists four approaches to detrending: first differencing, curve-fitting, digital filtering, and fitting piecewise polynomials.

First differencing is simply taking the difference of the value of the series at times t and $t-1$. First differencing is not suitable for time series whose level itself has importance, as the differenced series essentially is just change in level from one observation to the next, regardless of the level itself.

Curve-fitting is best if the time series changes in level gradually over time or the trend is some simple function of time itself. The simplest and most widely used function of time used in detrending is the least-squares-fit straight line. Simple linear regression is used to fit the model

$$x_t = a + bt + e_t \tag{1}$$

where x_t is the original time series at time t, a, b are regression coefficients, and e_t are the regression residuals. The trend is then described by

$$g_t = \hat{a} + \hat{b}t \tag{2}$$

where g_t is the trend, $\hat{a}$ is the estimated regression constant, and $\hat{b}$ is the estimated regression coefficient. The advantage of the straight-line method is simplicity, though a line may be unrealistic in some cases. It is then that other functions of t might be better depending on the type of data.

Digital filtering is another procedure for dealing with trend by describing the trend as a filtered version of the original series. The original series is converted to a smooth trend line by weighting the individual observations

$$g_t = \sum_{r=-q}^{s} a_r x_{t+r} \tag{3}$$

where x_t is the original series, a_r is a set of filter weights (summing to 1.0), and g_t is the smooth trend line. The weights are often symmetric, with $s = q$ and $a_j = a_{-j}$. If the weights are symmetric and all equal, the filter is a simple moving average, which generally is not recommended for measuring trend.

Fitting piecewise polynomials is an alternative to fitting a single curve to the entire time series by fitting polynomials of time to different parts of the time series. Polynomials that are used in this manner are piecewise polynomials. For example, the cubic smoothing spline is a piecewise polynomial in time with the following properties:

- The polynomial is cubic.
- A separate polynomial is fit to every sequence of three points in the series.
- The first and second derivatives are continuous at each point.
- The "spline parameter" specifies the flexibility and depends on the relative importance given to "smoothness" of the fitted curve, and "closeness of fit," or how close the fitted curve passes to the individual data points. [8]

De Boor provides more information on the spline curve. [3]

Once a trend line has been fit to the data, the question remains of how to remove the trend. If a method has identified a trend line, two options are available. First is to subtract the prediction from the original data, giving a new time series of residuals from the trend. This difference is attractive because of its simplicity, and for giving a convenient breakdown of the variance. The residual series is in the same units as the original series, and the total sum of squares of the original data can be expressed as the trend sum-of-squares plus the residual sum-of-squares. Second, a ratio is attractive for some kinds of data because the ratio is dimensionless, and the ratio operation tends to remove trend in variance that might accompany trend in mean. Ratio-detrending generally is feasible for non-negative time series only, and runs the risk of explosion of the detrended series to very high values if the fitted trend line approaches zero.[8]

3.2 Application

In the data set provided, the aspect of time is represented by two variables: forecast date and forecast hour. These variables must be combined so that a regression can be performed. Deciding that midnight of January 1, 1998 is the zero hour, time is incremented for each hour that has passed. Once a year the value of the power consumed is recorded as zero, due to daylight savings. To correct this, code was composed to analyze the two hours prior to the zero reading and calculate a comparable value to replace the zero. (When the clock is turned one hour ahead, there is no value recorded for 2am that morning.)

If one thinks about how power consumption may change overall, it is intuitive to conclude that the amount of power consumed will increase over time. This is due to the influx of people to the area, the establishment of new industry, and the advances in technology that depend on power for operation. Figure 2 shows the power consumption plotted against the number of hours since "zero hour." The graph supports a subtle increase over time.

Using SAS, a linear regression of power against hours passed was calculated. The function reported was:

$$Y_n = 43.06280 + 0.00021863h + \epsilon \tag{4}$$

Using Equation 4, a predicted data set was created. Once complete, the forecasted values were subtracted from the actual time series. The resulting residuals are a detrended version of the time series and are stationary.

Definition 5 *A time series is classified as stationary when fluctuations are centered around a constant mean.*

With the long term growth removed, the new data set was used to locate cycles dependent within the series. This new series is depicted by the graph in Figure 3.

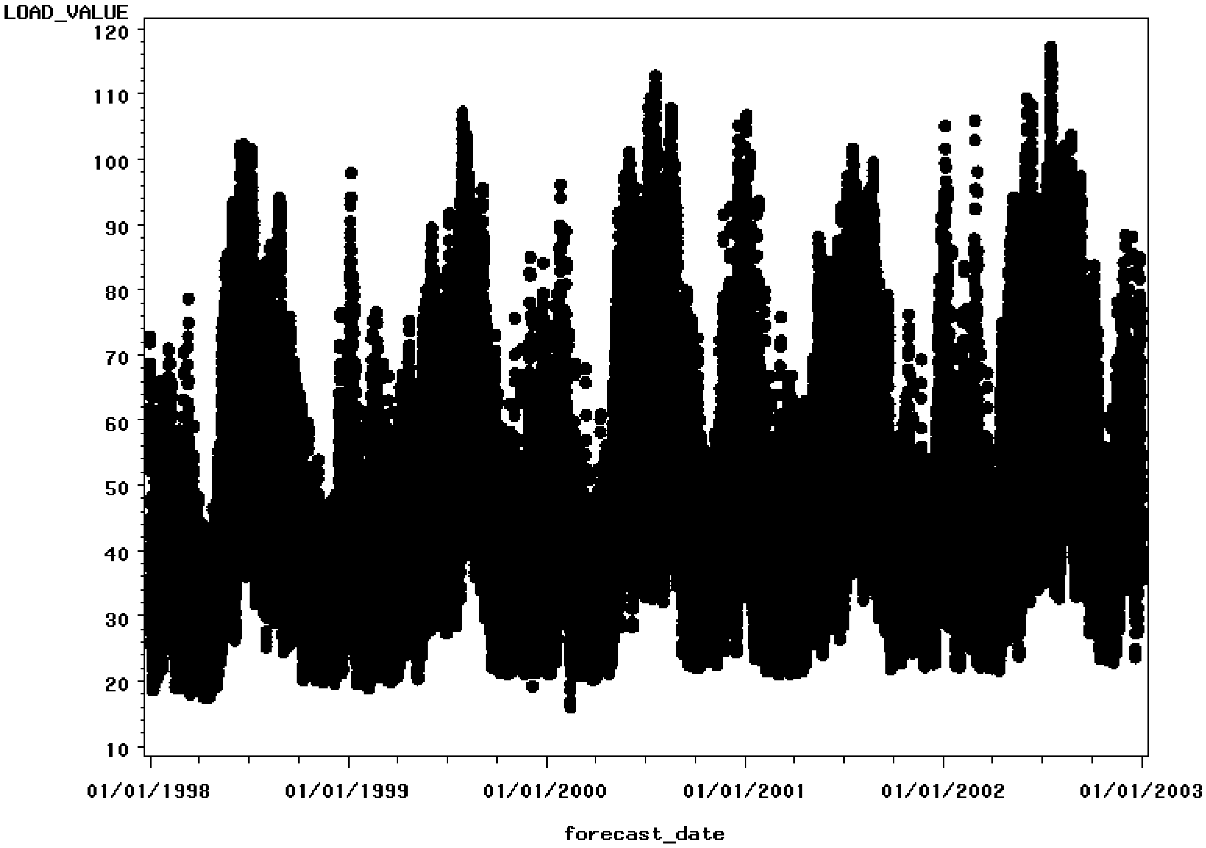

Figure 2: Scatterplot of Original Data Set

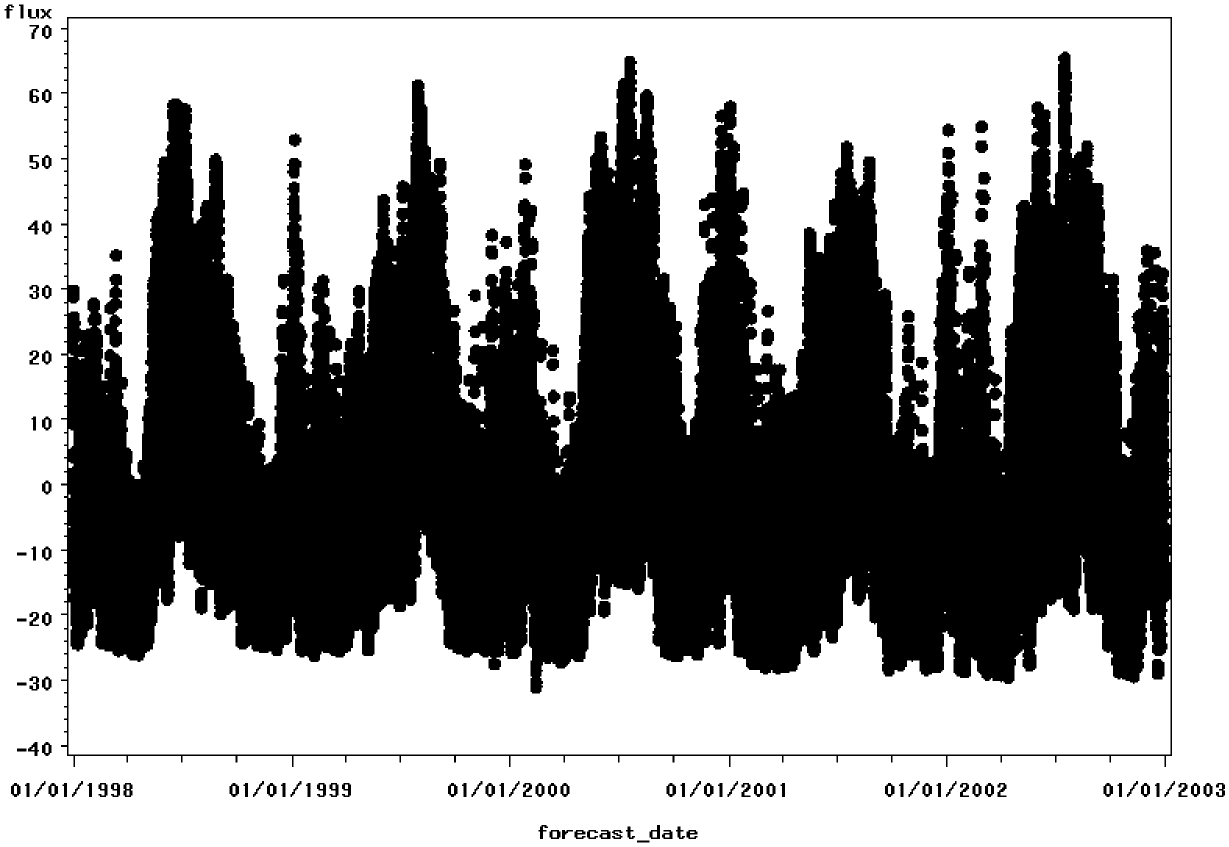

Figure 3: Scatterplot of Detrended Data Set

4 SEARCHING FOR CYCLES

4.1 Background

Definition 6 *The spectrum is the distribution of variance of a time series as a function of frequency.*

The spectrum is studied because many natural events contain variability that is frequency-dependent, and understanding this frequency dependence may yield information about the underlying physical structure of the time series.

SAS performs a spectral analysis through the Spectra Procedure. This procedure performs spectral and cross-spectral analysis of time series, aiding in the search for cyclical patterns in the data. SAS produces estimates of the spectral densities using a discrete Fourier transform to obtain the periodogram. The Fourier transform decomposition of the series Y_t is

$$Y_t = \frac{a_0}{2} + \sum_{k=1}^{m} [a_k \cos(\omega_k t) + b_k \sin(\omega_k t)] \tag{5}$$

where t is the time, $t = 1, 2, \ldots, n$; n is the number of observations in the time series; m is the number of frequencies in the Fourier decomposition: $m = \frac{n}{2}$ if n is even or $m = \frac{n-1}{2}$ if n is odd; a_0 is the mean term: $a_0 = 2\bar{x}$; a_k are the cosine coefficients; b_k are the sine coefficients; and ω_k are the Fourier frequencies: $\omega_k = \frac{2\pi k}{n}$. The amplitude periodogram J_k is defined as:

$$J_k = \frac{n}{2}(a_k^2 + b_k^2).$$

However, the periodogram is a volatile and inconsistent estimator of the spectrum. Thus, SAS provides the spectral density estimate by smoothing the periodogram. This smoothing reduces the variance of the estimator but introduces bias. SAS uses a weight function, similar to a moving average technique, for the smoothing effect. SAS allows you to specify any number of weight constants. However, if you do not specify any weight system, the standard periodogram, as previously described, is produced. The Spectra Procedure also can test whether or not the data is white noise. The procedure prints two test statistics: Fisher's Kappa statistic [2] [5] and Bartlett's Kolmogorov-Smirnov statistic[1] [5] [4]. Once the white noise test determines that the data is not just reacting to white noise, the frequencies of the spectrum then can be read without issue.

All the frequencies ω_k lie in $[0, \pi]$. According to Hamilton [6], suppose that the data was generated by the following,

$$Y_t = a_k \cos(-\omega_k t) + b_k \sin(-\omega_k t), \tag{6}$$

where $-\omega_k < 0$ represents some particular negative frequency and a_k and b_k are the coefficients. Since $\cos(-\omega_k t) = \cos(\omega_k t)$ and $\sin(-\omega_k t) = -\sin(\omega_k t)$, the previous equation can be equivalently written

$$Y_t = a_k \cos(\omega_k t) - b_k \sin(\omega_k t). \tag{7}$$

Thus there is no way of using observed data on Y_t to decide whether the data was generated by a cycle with frequency $-\omega_k$ as in Equation 6 or by a cycle with frequency $+\omega_k$ as in Equation 7. It is simply convention that positive frequencies are chosen. Suppose the data were generated from a periodic function with frequency $\omega_k > \pi$, say, $\omega_k = \frac{3\pi}{2}$:

$$Y_t = a_k \cos(\frac{3\pi}{2} t) - b_k \sin(\frac{3\pi}{2} t). \tag{8}$$

Properties of the sine and cosine function imply that the previous equation may be rewritten

$$Y_t = a_k \cos(\frac{-\pi}{2} t) - b_k \sin(\frac{-\pi}{2} t). \tag{9}$$

Using the previous argument, a representation with cycles of frequency $(\frac{3\pi}{2})$ is indistinguishable from one with cycles of frequency $(\frac{\pi}{2})$ [6].

4.2 Application

Now that the data has been made stationary by removing the long-term growth, the isolation of cycles embedded within the time series can proceed. Step one was to check the white noise test to ensure that the data is not reacting solely as white noise. With that complete, the series was divided into more than 30,000 frequencies between 0 and π. For each of these frequencies, SAS returned the period, the estimated spectral density, and the amplitudes of the sine and cosine components. Figures 4 and 5 show plots of the periods $p = 1/freq$ against the spectral density. As is seen, there are specific periods in the series that appear as spikes in the density. Among these are 12 hours, 24 hours, half a year, and one year. Figure 6 shows a scatterplot of daily consumption of power over each of the 365 days of the year. The two troughs in the plot are spring and autumn while the crests are summer and winter. Although the lengths of these seasons are not equal, one can agree that their occurrence is approximately six months apart.

Focusing only on cycles with a spectral density of 40,000 or greater, a data step procedure was created (see Appendix A for SAS code) to reconstruct the data set using only the isolated frequencies seen in Table 1. With the reconstructed values, a new residual is found by subtracting the residual from detrending and the spectral component. This new residual is another time series that will be modeled as a function of weather.

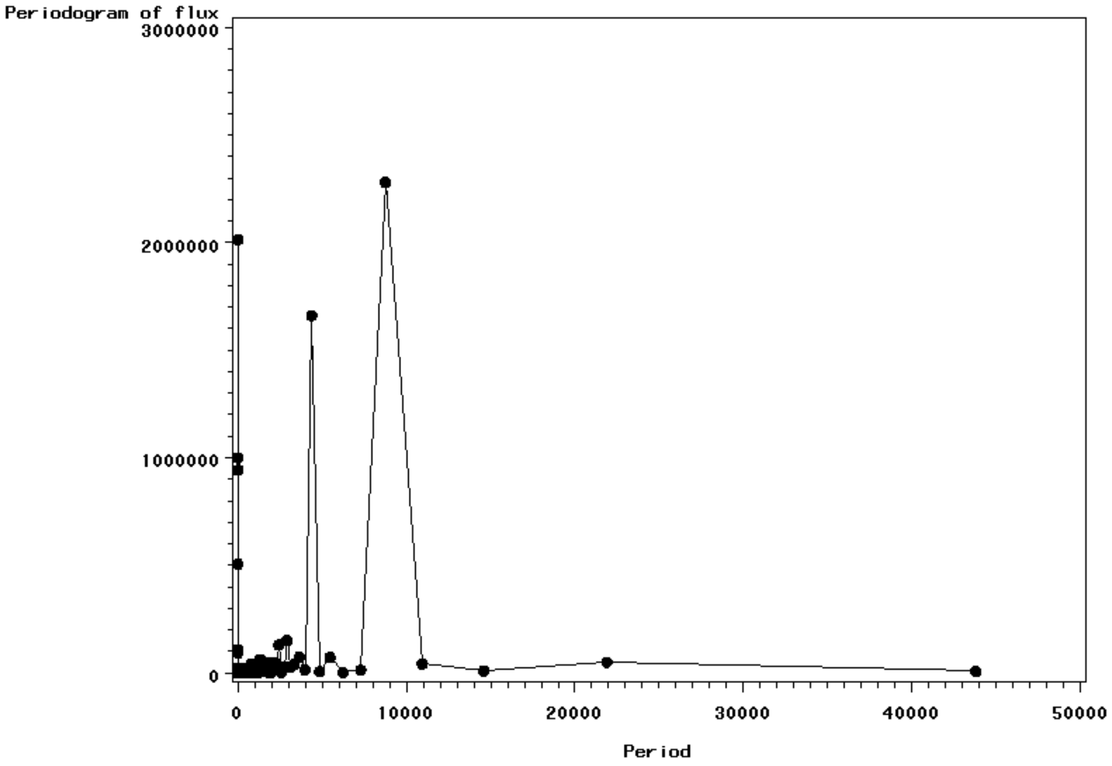

Figure 4: Graph of Spectral Densities by Period

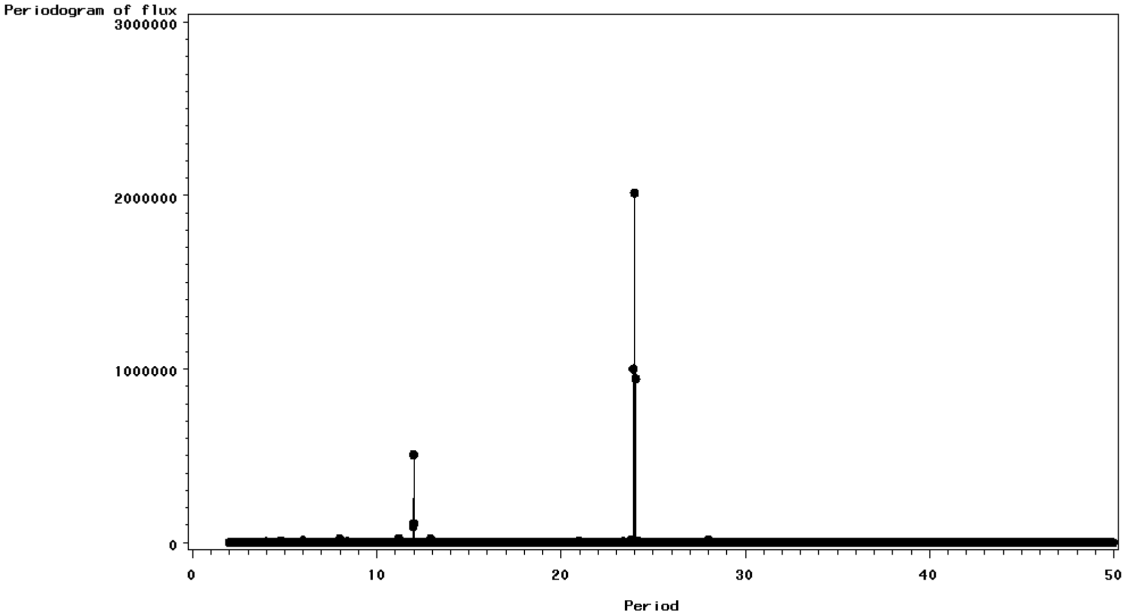

Figure 5: Close-up of Spectral Densities for Periods Under 50 Hours

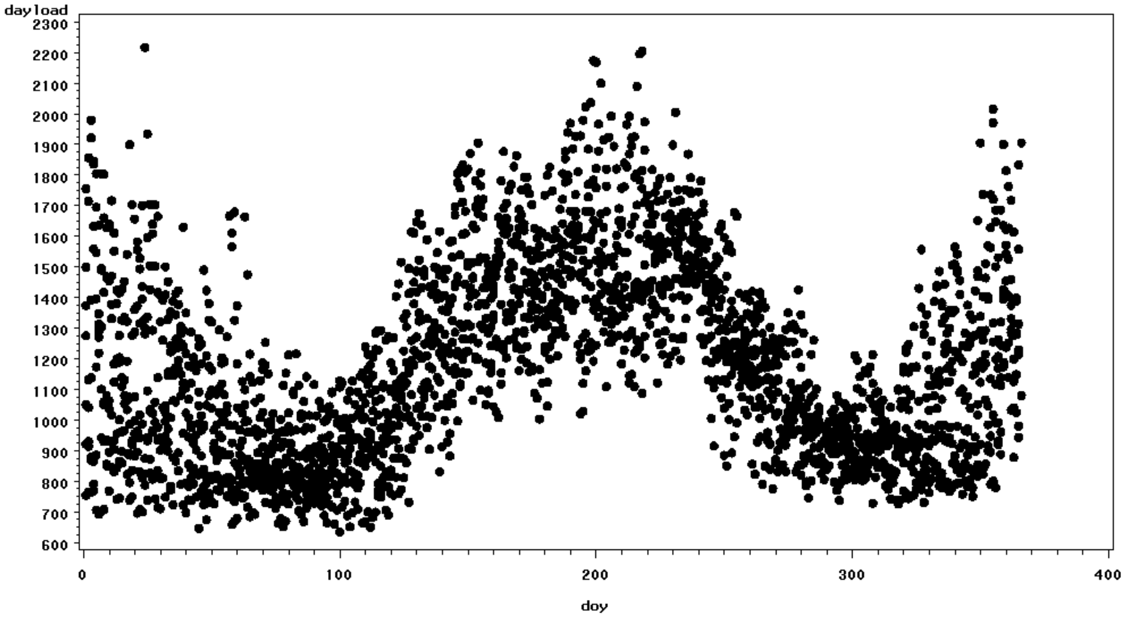

Figure 6: Scatterplot of Consumption by Day of Year

Table 1: Frequencies of Importance

Frequency	Period	Cosine Amplitude	Sine Amplitude	Periodogram
0	.	-0.000169656	0	0.0006306961
0.0002867463	21912	1.5222202595	0.1288622346	51137.348985
0.0005734926	10956	-0.86039011	1.1138405072	43405.739723
0.0007168658	8764.8	-9.608029266	-3.432680537	2280984.9031
0.0011469853	5478	0.8615582144	-1.608903267	72985.635141
0.0014337316	4382.4	8.0657461951	3.2863785631	1662168.6303
0.0017204779	3652	1.8480853251	0.032476618	74861.780469
0.0018638511	3371.0769231	-0.554283825	1.2547913611	41232.509794
0.0021505974	2921.6	2.5751262875	0.5415848805	151731.60686
0.0025807169	2434.6666667	-0.744250166	-2.326741726	130762.82624
0.0028674632	2191.2	-0.743058439	-1.243866086	46000.709235
0.0030108363	2086.8571429	0.7951312846	-1.170930697	43896.582568
0.003584329	1752.96	1.4782380514	0.3704391866	50888.71207
0.0044445679	1413.6774194	0.8866382143	-1.179194607	47694.256162
0.0045879411	1369.5	-1.511224466	0.6643509353	59713.739946
0.0047313142	1328	-1.699685513	0.124467834	63641.71673
0.0053048069	1184.4324324	-1.341112678	-0.480806761	44476.065087
0.0078855237	796.8	1.4009691202	-0.242484107	44295.393247
0.261082522	24.065897858	4.0192379214	5.1870022084	943514.62297
0.2617993878	24	-4.155658272	-8.64183758	2014827.0783
0.2625162536	23.934462043	1.1051758925	6.6693276689	1001407.8784
0.5228819098	12.016451878	-1.82280099	-1.309402936	110373.80071
0.5235987756	12	-2.96705318	-3.788770122	507442.11927
0.5243156414	11.983593109	-0.924773803	-1.859841681	94533.137516

5 MODELING THE RESIDUALS

5.1 Weather Variables

A plethora of data was made available for the project, including hourly weather data from eight weather stations in and around the state of Georgia. Spanning a time frame of January 1, 2001 to July 1, 2005, the weather data consisted of the following eleven variables: temperature, dew point, humidity, heat index, wind chill, wind direction, wind speed, wet bulb temperature, cloud cover, number of sunshine minutes, and probability of precipitation.

Some variables had to be eliminated for various reasons. Due to an inability to forecast and issues with multicolinearity, wet bulb temperature was eliminated. Cloud cover and sunshine minutes are minutes in an hour that there is cloud cover or sunshine, respectively. Through inspection, the sum of these two variables, within each hour, did not equal sixty. Therefore, a clear understanding on how cloud cover is measured could not be made. Due to these reasons, along with difficulty in predicting, this variable was also eliminated from use. In the data provided, probability of precipitation was primarily missing. At other times, the variable was at 0 or 100; it appears that this variable actually recorded whether or not it actually did rain on a given day. Heat index and wind chill are terms that people hear frequently during summer and winter; however, they are just combinations and interactions of other variables in the list. Use of these variables will be retained but an investigation of interactions present within their formulas was performed.

5.2 Application

With the overall linear trend and the internal cycles removed, the original time series has been reduced into a list of residuals which will now be explained using the weather variables. Indicator variables were created for hour of day (midnight-11pm), day of the week (Sunday-Saturday), and month of the year (January-December). All of these indicator variables are binary in nature.

With upwards of fifteen weather variables, scatterplots of the residual value against each weather variable were created. The plot of the residual against temperature can be seen in Figure 7. The plots of heat index and wind chill also had similar displays. It can be seen that residuals have at least a quadratic relationship with temperature.

Weather, especially temperature, comes in trends. At the same time, the aspects of weather that a person reacts to is different. For example, there are those that make their decisions based solely on the high or low temperature of the day. There are some who listen to only the temperature of the hours immediately around the current time. There are those that look to the average temperature of the current day and those of previous days. To correctly adjust for this, variables were created that reflected these ideas.

Beginning with the binary variables then proceeding to weather, the model was created by systematically isolating those variables that were deemed significant with an alpha level of 0.05. For the weather variables that seemed to follow more of a non-linear relationship, model fits were achieved by continually adding terms of higher power until a non-significant value was achieved. SAS's regression procedure indicated that the quartic term of temperature was the last power of significance. All of these variables that were significant for the model were kept and the search continued as the adjusted r-squared value increased.

Systematically progressing through the other weather variables, all variables and interactions whose p-value was below 0.05 were kept. When the heat index variable was added, its p-value was low enough to keep in the model and the r-squared value increased. Since the plot of heat index against the residuals was very similar to temperature, power terms of the heat index variable were added until the p-value surpassed 0.05. The same was true for wind chill.

The coefficients for the model are presented in Table 2. With this model, the final error was calculated and utilized in testing the overall validity of the model.

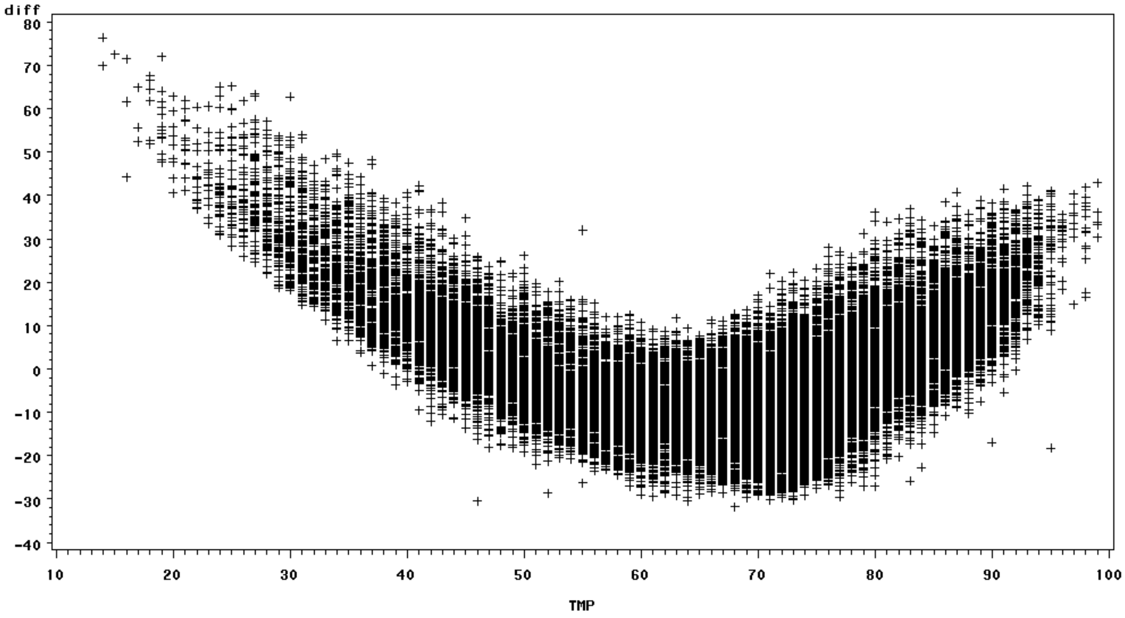

Figure 7: Scatterplot of Current Error Against Temperature

Table 2: Regression Coefficients for Residual Model

Variable	Coefficient	Variable	Coefficient
am4	-1.59848	pm1	1.26592
am5	-1.35444	pm2	0.62367
am6	2.41209	pm4	-1.17938
am8	-1.61240	pm5	-1.16056
am9	-1.77832	pm8	1.08644
am10	-1.39349	pm9	2.20576
noon	0.77613	pm10	1.27616
January	41.92404	July	30.46393
February	46.91118	August	31.01620
March	45.91622	September	36.51017
April	42.12628	October	45.39062
May	38.76690	November	49.01250
June	32.09803	December	42.90409
Wednesday	0.85000	Friday	-0.80523
Thursday	0.34551	Temperature	2.75096
Temperature^2	-0.06478	Temperature^3	0.00062309
Temperature^4	-0.00000176	Avg.Temp.Today	4.97022
Avg.Temp.Today^2	-0.08329	Avg.Temp.Today^3	0.00055849
Avg.Temp.Yest.	0.60479	Avg.Temp.2DaysAgo	0.28699
High Temp	-1.60502	High Temp^2	0.01276
Low Temp	-0.20230	Low Temp^2	0.00153
HeatIndex	-4.24547	HeatIndex^2	0.05976
HeatIndex^3	-0.00026842	Avg.HeatIndex	-0.58120
Avg.HeatIndexYest.	-0.55708	Avg.HeatIndex2DaysAgo	-0.14067
WindChill	-0.32577	Avg.WindChill	-0.41562
Avg.Humidity	0.04873	Avg.HumidityYest.	0.02391

6 TESTING THE MODEL

To judge the validity of the model, the aspect of Mean Absolute Percent Error, MAPE, will be utilized. This number is found by completing the following:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|\text{ error of prediction } i|}{\text{actual power load value}} \tag{10}$$

where n is the number of predictions made. Throughout the process of creating the model, calculations of the MAPE were taken to monitor the improvement. Table 3 shows the improvement of MAPE during all three stages.

To create this model, data from Mitchell Region, located in the southwestern part of the state, was used. The weather data is from KABY, which has a weather station on the border of the Mitchell and Sumter Regions. All but two years of data was used in the creation of the model. These two years were reserved to test the model. With this model forecasting further into the future than a single month, the predictions of each hour were summed and comparisons were made by day and week. Table 4 shows the quartiles and the mean of the MAPE. Figures 8 and 9 show the scatterplot of the final error compared by day and week.

The model was then applied to the data from Sumter. Since these two areas are close in proximity and share similar weather patterns, the hope was that the model will perform equally. When the model was subjected to the data from the Sumter Region, the MAPE increased due to the model being dependent on the average power consumed and the linear growth of the region. Any differences between Sumter and Mitchell in these two concepts will dramatically affect the model.

Table 3: MAPE In-sampling Test Values

Stage of Model	MAPE
Removal of Trend	0.3187867
Spectral Analysis	0.1513085
Regression on Residuals	0.0998232

Table 4: MAPE Forecast Test Values

	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean
By Day	0.0006	0.0466	0.0928	0.1535	0.4352	0.1081
By Week	0.0012	0.0485	0.0925	0.1372	0.3263	0.0992

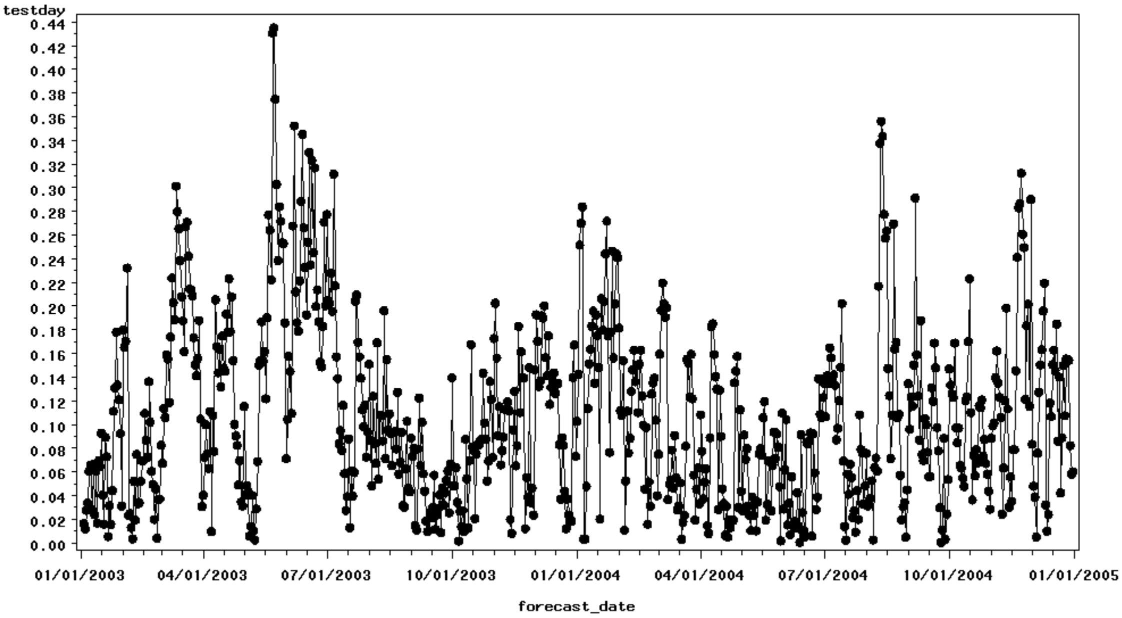

Figure 8: Scatterplot of Final Error Against Day

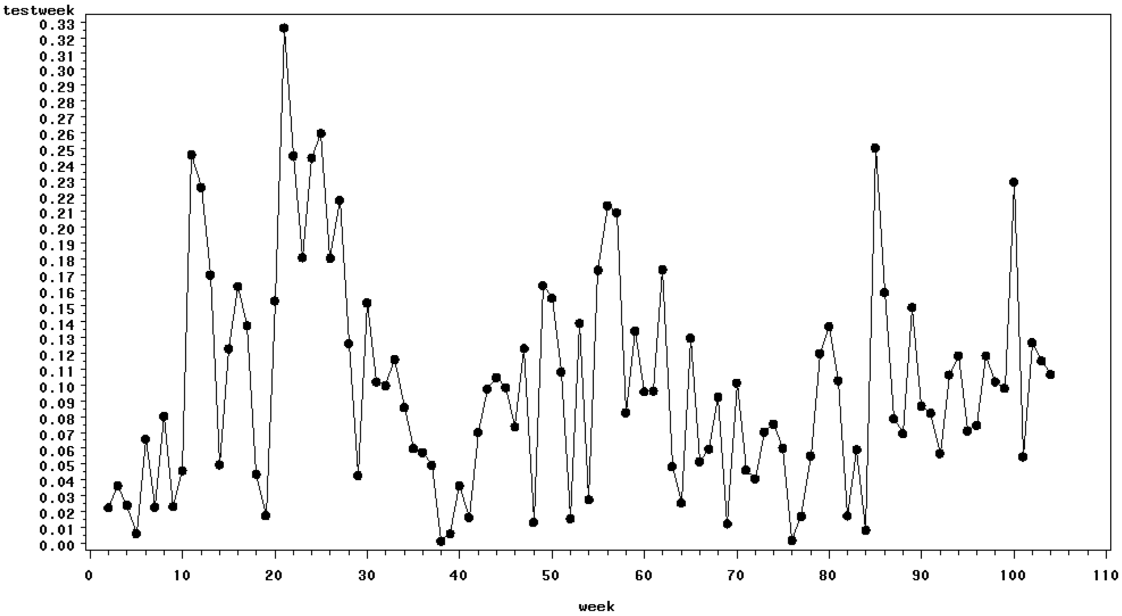

Figure 9: Scatterplot of Final Error Against Week

7 CONCLUSION

Currently, Progress Energy uses an “hour ahead” model and a “day ahead” model to forecast consumption of power. Both of these models forecast all hours from the current time until seven days into the future.

The “Day-Ahead” model is a regression that utilizes the following variables:

- Calendar days (Sunday through Saturday)
- Calendar month (January through December)
- Indicator variable for Holidays (New Year’s Day, Fourth of July, Labor Day, Thanksgiving, Friday after Thanksgiving, Christmas Eve, Christmas Day, Days Christmas Lights are in use)
- Overall growth due to time
- Weather variables (Temperature, Heat Index, Wind Chill, Cloud Cover, Max-Min daily temperature, “Parameter Bucket”, Hot Degree and Cold Degree Days, Yesterday’s average temperature, Day before yesterday’s average temperature, Average temperature so far for today)

The “Hour-Ahead” model consists of 26 Neural Networks where one node represents each hour to predict and the final two nodes represent daily energy consumed and the daily peak of consumption. The daily energy model consists of:

- Calendar days (Sunday through Saturday)
- Calendar months (Jan-Mar, May-Dec)
- Sunrise and sunset
- Time trend

- Constant (Intercept)
- Average daily temperature
- Average daily temperature squared
- Hot and cold day buildup
- Average wind chill and heat index

The daily peak model was composed of four nodes. Two are named “hot” nodes and the other two are “cold” nodes. The “hot” nodes utilize:

- Daily max temperature
- Cold Degree Days
- Max heat index
- Heat index buckets
- Hot buildup
- Day type (Weekday= Tues, Wed, Thurs; Weekend= Sat, Sun)

The “cold” nodes contain:

- Daily low temperature
- Hot Degree Days
- Min wind chill
- Wind chill buckets
- Cold buildup
- Day type (weekday and weekend)

Each Neural Network that is used to forecast a particular hour is composed of three nodes. The first node is exactly like the daily energy model. Nodes 2 and 3 utilize the daily energy predicted, the daily peak predicted, and includes three hours of lagged actual temperature and actual power load. In its raw format, these models perform at an MAPE of around 6%. After the predictions are created, they are passed to human eyes and are corrected based on previous predictions or similar style days. After this process, the MAPE reduces to just under 4%. [10]

As one can see, the new model is not as complex as the current design but gives an accurate prediction for up to two years with a MAPE of approximately 10%. The model closely mimics the "Day-Ahead" model currently used with a few exceptions. When the spectral analysis was performed, the holidays were taken into account. There are not too many holidays that do not occur on the same day each year. Also, many families light Christmas lights around the same time period every year, which again was taken into account.

Unfortunately, the model seems to work best for Mitchell, the region whose data created the model. However, using this procedure, a model for each region can be created to predict power consumption. A suggestion to improve upon this fact would be to include aspects of the counties into the model: the percentage of industry in the county, its population, its location, etc. Placing items such as these into the model would enhance the versatility of the model and allow one model to work for all the counties.

8 APPENDIX A - SAS CODE

Code used to isolate and prep data from Mitchell County and from KABY weather station.

```
data trial.mitchellmod trial.mitchelltest;
    set outdata.Mitchell;
    retain time 0;
    powerb1= lag(load_value);
    powerb2= lag(powerb1);

    if load_value eq 0 then do;
        if powerb1 gt powerb2 then do;
            load_value = powerb1 + (powerb1-powerb2)*0.5;
        end;
        if powerb1 lt powerb2 then do;
            load_value = powerb1 - (powerb2-powerb1)*0.5;
        end;
    end;
    if year(forecast_date) le 2002 then output trial.mitchellmod;
    if year(forecast_date) gt 2002 then output trial.mitchelltest;
    time=time+1;
    keep load_value forecast_date forecast_hour time;
run; quit;
```

Code used to detrend the data and create our first testing variable for MAPE.

```
proc reg data=trial.mitchellmod;
    model load_value= time;
run; quit;


data trial.detrendmit;
```

```
    set trial.mitchellmod;
    predict=43.06280+.00021863*time;
    flux=load_value-predict;
    test1=abs(flux)/load_value;
run; quit;
```

Code used to perform Spectral Analysis on Time Series.

```
proc spectra data=trial.detrendmit out=trial.specdemit coeff;
    var flux;
run; quit;


data trial.newspec;
    set trial.specdemit;
    c=1;
    if ( _n_ =1 ) then do; c=0.5; output; end;
    if (p_01 gt 40000) then output;
run; quit;


proc gplot data=trial.specdemit;
    *where period lt 15000;
    plot p_01 * period;
        symbol i=join;
run; quit;
```

Code used to reconstruct data set using only the specified frequencies.

```
data trial.reconstruct;
    time=0;
    fluxB=0;
run; quit;
```

```
%macro reconstruct;
    %do q=1 %to 43824;
        data trial.t;
            set trial.newspec end=finish;
            retain fluxB 0;
            time=&q;
            *if _n_ eq 1 or finish then c=0.5;
            *else c=1.0;
            fluxB= fluxB + c*(cos_01*COS(freq*time) + sin_01*SIN(freq*time));
            keep time fluxB;
            if finish then output;
        run;
        quit;

        data trial.reconstruct;
            set trial.reconstruct trial.t;
        run;
        quit;
    %end;
%mend;

%reconstruct;
```

Code used to create binary time variables and create adjusted weather variables for the regression portion.

```
data trial.avgtmpmod trial.avgtmptest;
    set outdata.avgtmp;
    if year(forecast_date) le 2002 then output trial.avgtmpmod;
    if year(forecast_date) gt 2002 then output trial.avgtmptest;
run; quit;
```

```
data trial.hilowmod trial.hilowtest;
    set outdata.highlow;
    if year(forecast_date) le 2002 then output trial.hilowmod;
    if year(forecast_date) gt 2002 then output trial.hilowtest;
run; quit;

data trial.weathermod;
    merge trial.avgtmpmod trial.hilowmod;
    by forecast_date;
run; quit;

data trial.powermod;
    merge trial.detrendmit trial.reconstruct;
    where time gt 0 AND time ne 43824;
    by time;
run; quit;

data trial.powermod;
    set trial.powermod;
    diff= flux- fluxB;
    test2= abs(diff)/load_value;
run; quit;

data trial.model;
    merge trial.powermod trial.weathermod;
    by time;
run; quit;

data trial.model;
```

```
set trial.model;
dow=weekday(forecast_date);
month=month(forecast_date);
if forecast_hour eq 0 then do;
    mid=1; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 1 then do;
    mid=0; am1=1; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 2 then do;
    mid=0; am1=0; am2=1; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 3 then do;
    mid=0; am1=0; am2=0; am3=1; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 4 then do;
    mid=0; am1=0; am2=0; am3=0; am4=1; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 5 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=1; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 6 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=1;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 7 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=1; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 8 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=1; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 9 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=1; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 10 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=1; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 11 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=1; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 12 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=1;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 13 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=1; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 14 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=1; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 15 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=1; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 16 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=1; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 17 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=1; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 18 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=1; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 19 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=1;
    pm8=0; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 20 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=1; pm9=0; pm10=0; pm11=0;
end;
if forecast_hour eq 21 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=1; pm10=0; pm11=0;
end;
if forecast_hour eq 22 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=1; pm11=0;
end;
if forecast_hour eq 23 then do;
    mid=0; am1=0; am2=0; am3=0; am4=0; am5=0; am6=0;
    am7=0; am8=0; am9=0; am10=0; am11=0; noon=0;
    pm1=0; pm2=0; pm3=0; pm4=0; pm5=0; pm6=0; pm7=0;
    pm8=0; pm9=0; pm10=0; pm11=1;
end;
if dow eq 1 then do;
        sunday=1; monday=0; tuesday=0; wednesday=0;
        thursday=0; friday=0; saturday=0;
        weekday=0; weekend=1;
end;
if dow eq 2 then do;
        sunday=0; monday=1; tuesday=0; wednesday=0;
        thursday=0; friday=0; saturday=0;
        weekday=1; weekend=0;
end;
if dow eq 3 then do;
        sunday=0; monday=0; tuesday=1; wednesday=0;
        thursday=0; friday=0; saturday=0;
        weekday=1; weekend=0;
end;
if dow eq 4 then do;
        sunday=0; monday=0; tuesday=0; wednesday=1;
        thursday=0; friday=0; saturday=0;
        weekday=1; weekend=0;
end;
if dow eq 5 then do;
        sunday=0; monday=0; tuesday=0; wednesday=0;
        thursday=1; friday=0; saturday=0;
        weekday=1; weekend=0;
end;
if dow eq 6 then do;
        sunday=0; monday=0; tuesday=0; wednesday=0;
        thursday=0; friday=1; saturday=0;
        weekday=1; weekend=0;
end;
if dow eq 7 then do;
        sunday=0; monday=0; tuesday=0; wednesday=0;
        thursday=0; friday=0; saturday=1;
        weekday=0; weekend=1;
end;

if month eq 1 then do;
        jan=1; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 2 then do;
        jan=0; feb=1; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 3 then do;
        jan=0; feb=0; mar=1; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 4 then do;
        jan=0; feb=0; mar=0; apr=1; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 5 then do;
        jan=0; feb=0; mar=0; apr=0; may=1; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 6 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=1; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 7 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=1;
        aug=0; sep=0; oct=0; nov=0; dec=0;
```

```
end;
if month eq 8 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=1; sep=0; oct=0; nov=0; dec=0;
end;
if month eq 9 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=1; oct=0; nov=0; dec=0;
end;
if month eq 10 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=1; nov=0; dec=0;
end;
if month eq 11 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=1; dec=0;
end;
if month eq 12 then do;
        jan=0; feb=0; mar=0; apr=0; may=0; jun=0; jul=0;
        aug=0; sep=0; oct=0; nov=0; dec=1;
end;

tmp2= tmp*tmp; tmp3= tmp*tmp2; tmp4=tmp*tmp3;
tmpavg2=tmpavg*tmpavg; tmpavg3=tmpavg*tmpavg2;
tmpavg4=tmpavg*tmpavg3; avgtpb1=lag24(tmpavg);
avgtpb2=lag24(avgtpb1); avgtpb3=lag24(avgtpb2);
heat2= heat_index*heat_index; heat3= heat_index*heat2;
tmpback1= lag(tmp); tmpback2= lag2(tmp);
tmpback3= lag3(tmp); highb1= lag24(high);
highb2= lag24(highb1); highb3= lag24(highb2);
```

```
    lowb1= lag24(low); lowb2= lag24(lowb1); lowb3= lag24(lowb2);
    high2 = high*high; high3= high*high2; high4 = high*high3;
    low2= low*low; low3= low*low2; low4= low*low3;
    hiavgb1= lag24(hiavg); hiavgb2= lag24(hiavgb1);
    hiavgb3= lag24(hiavgb2); humavgb1= lag24(humavg);
    humavgb2= lag24(humavgb1); humavgb3= lag24(humavgb2);
    windavgb1= lag24(windavg); windavgb2= lag24(windavgb1);
    windavgb3= lag24(windavgb2); wind2= wind_chill*wind_chill;
    wind3= wind_chill*wind2; wind4= wind_chill*wind3;
run; quit;
```

REFERENCES

[1] Bartlett, M.S. 1966, *An Introduction to Stochastic Processes*, Second Edition, Cambridge: Cambridge University Press.

[2] Davis, H.T. 1941, *The Analysis of Economic Time Series*, Bloomington, IN: Principia Press.

[3] de Boor, C. 1978, *A Practical Guide to Splines*, New York: Springer-Verlag.

[4] Durbin, J. 1967, "Tests of Serial Independence Based on the Cumulated Periodogram," *Bulletin of International Statistics Institute*, 42, 1039-1049.

[5] Fuller, W.A. 1976, *Introduction to Statistical Time Series*, New York: John Wiley and Sons, Inc.

[6] Hamilton, J.D. 1994, *Time Series Analysis*, New Jersey: Princeton University Press.

[7] Kendall, K. and Ord, J.K. 1990, *Time Series*, Third Edition, London: Hodder and Stoughton Limited.

[8] Meko, D.M. Applied Time Series Analysis. University of Arizona. 11 October 2005. (http://www.ltrr.arizona.edu/ dmeko/geos585a.html).

[9] SAS Institute Inc. 1993, *SAS/ETS*® *User's Guide, Version 6*, Second Edition, Cary, NC: SAS Institute Inc.

[10] Settlage, M. "UNCW Thesis Assistance." E-mail to the author. 12 January 2006.