Design of a Weather-Normalization Forecasting Model

Final Report

Abram Gross
Yafeng Peng
Jedidiah Shirey
5/12/2014

Table of Contents

1.0 Executive Summary

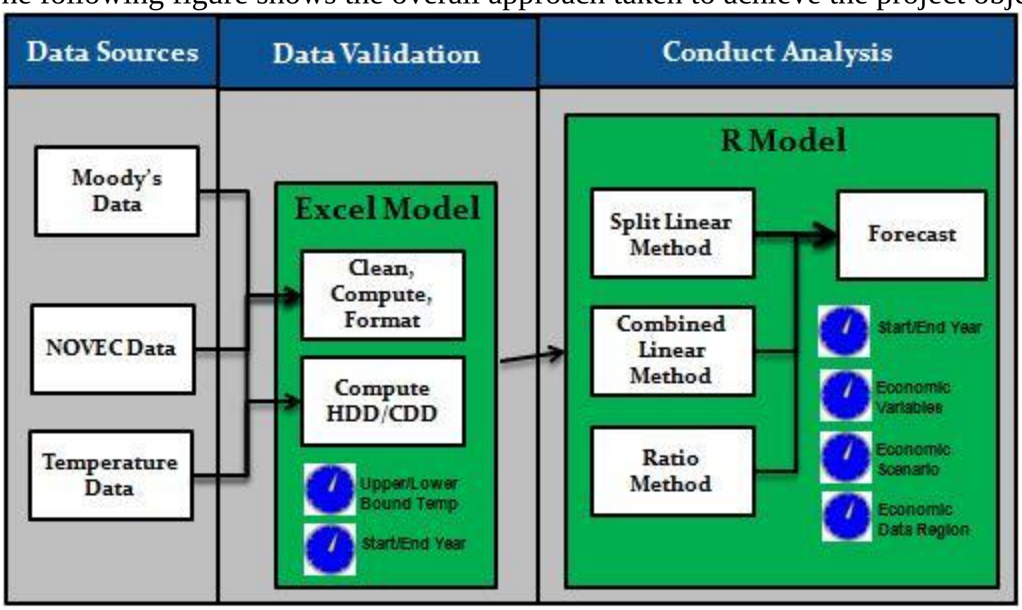
Northern Virginia Electric Cooperative (NOVEC) is an electricity distributor servicing parts of six Northern Virginia counties. In order to provide power to their customers, NOVEC purchases power in two ways: long-term bulk purchases and as-needed spot purchases. Bulk purchases occur up to five years in advance and are sized to meet expected power demand during that time period. In the event bulk purchases are insufficient to meet demand, spot purchases provide the power to cover the difference. Temperature fluctuations, mainly during the summer months, are a significant contributor to increased power demand in excess of the bulk purchase amount.

In order to purchase an appropriate amount of power through bulk purchases, NOVEC has developed a forecasting model that forecasts future power purchases over a 30-year horizon. NOVEC makes bulk power purchases based on the first 5 years of the forecast.

Based on recent warming trends, NOVEC believes that the current model may no longer be the best available and that a new weather-normalization method may better reflect weather trends. Improving the accuracy of the forecast would limit the amount of power that NOVEC has to buy beyond the bulk amount, thus decreasing costs. NOVEC requests analytical support to develop a new weather-normalization methodology to improve the existing forecasting model or to determine that the existing modeling approach offers better forecasts.

The purpose of this project is to develop a new weather normalization methodology to improve NOVEC's forecasting model by more accurately modeling future power demand. The model will take into account historical data as inputs: customer and power purchase total by month starting from 1983, hourly weather data starting from 1963, and Moody's Washington, D.C. metro economic data starting from the 1970s and projecting 30 years forward under varied scenarios. The end product of the project is a forecast of monthly power demand for the next 30 years and a forecasting model that will give NOVEC the ability to perform additional analysis.

The following figure shows the overall approach taken to achieve the project objectives.



The economic variables, power purchases, customer base, and temperature data was evaluated in the Data Validation step using Excel with macros developed for preprocessing and exploratory analysis. Where needed, records were re-formatted and gaps in the data were filled using linear interpolation. Heating Degree Days (HDD) and Cooling Degree Days (CDD) were calculated at hourly resolution using temperature observations; these variables are used as measures seasonal impact to power demand. A recorded temperature outside of a defined neutral zone, 55°F to 65°F between which temperatures are assumed to have insignificant impact on power consumption, is aggregated up to monthly resolution. The model developed in Excel uses an interface to permit changes to the neutral zone lower and upper bounds as well as transform all data records for testing a variety of general linear regression models. Additionally, split linear regression modules are provided with application capabilities though functionality was included to export processed data to files and launch an R model which utilizes the data in Excel to forecast the power demand. The R script developed allows for more powerful analysis beyond the capacity of those developed in Excel.

The methodology utilized is as follows: the economic variables and customer total are fit using a linear regression to the historic power demand. Residential services are assessed separately from non-residential. Based on this relationship and the computed historical HDD and CDD, the base power load and the seasonal power load are determined. In order to forecast future power demand, the forecasted economic variables provided by Moody's Analytics report are utilized to predict the future customer base. In turn, the customer base informs the size of the base power load based on the historical relationship under the assumption that imputed monthly rates will sufficiently model average consumption for future customers. In order to determine the total power demand, HDD and CDD are forecasted to determine the seasonal power load. The seasonal power load and the base load are combined to form the forecasted monthly power load. Three different methods were utilized to forecast the HDD and CDD: Holt-Winters method, ARIMA method, and BAT method. Each of these methods was utilized in each of the three modeling approaches listed above: Combined Linear Regression, Split Linear Regression, and Customer Ratio Method.

Each of the three models produced a different 30-year power demand forecast. Based on the statistical analysis of the different forecasts, the Split Linear Regression model using the Holt-Winters method produces the most accurate forecast. We recommend that NOVEC utilize the capabilities provided by the Excel and R models to supplement their current forecasting methods. Additional alternatives that can be studied using the capabilities provided by the models are varying the economic scenarios, varying the range of input years for temperature data and power demand, varying values for determining the HDD and CDD, and varying the economic variables used to determine the customer base.

2.0 Introduction

2.1 Background

Northern Virginia Electric Cooperative (NOVEC) is an electricity distributor headquartered in Manassas, Virginia. NOVEC provides power to nearly 150,000 customers across six counties – Prince William, Stafford, Loudon, Fairfax, Fauquier, and Clarke. NOVEC's service territory constitutes a fraction of each of these six counties, wherein it is required to provide power to meet any customer demand. In order to meet power demand, NOVEC purchases power from PJM Interconnection, a regional power supplier, in two ways: long-term bulk purchases and spot purchases. Bulk purchases occur up to five years in advance and are meant to satisfy estimated demand over this time period. In the event bulk purchases are insufficient to meet any demand over this timeframe, spot purchases provide the power to cover the difference and flexibility to materialize hours or days before delivery. Temperature fluctuations, mainly during the summer months, are a significant contributor to increased power demand in excess of the bulk purchase amount. Bulk purchases offer economies of scale and are more cost efficient than spot purchases which constitute a higher premium for accommodating unscheduled orders on short notice.

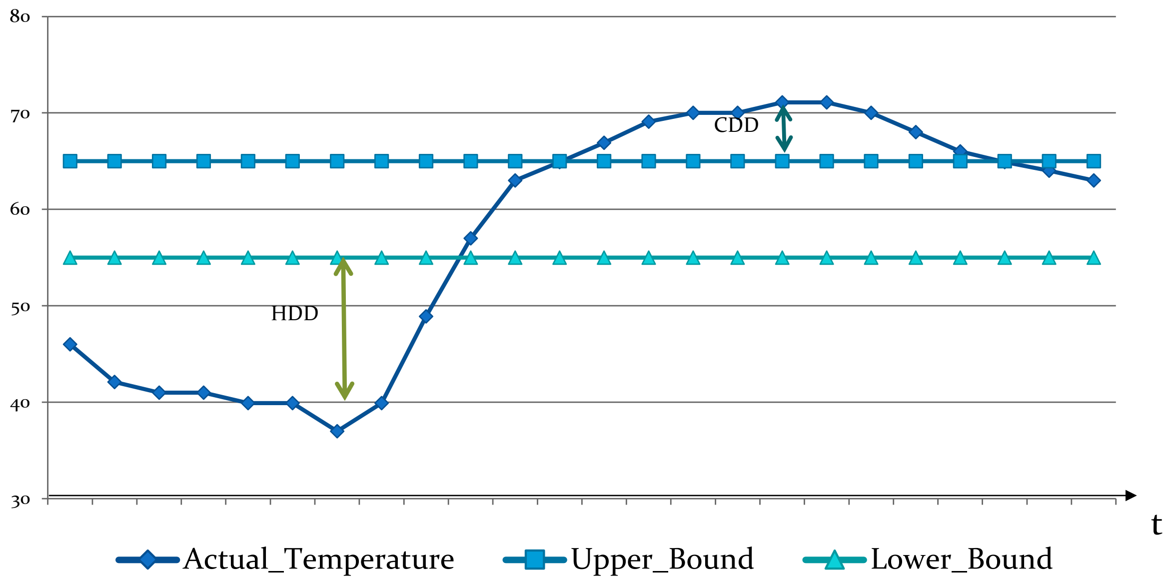
In order to minimize the amount of ad hoc purchases without overcompensating for their avoidance with excessive bulk purchases, NOVEC has developed a forecasting model that estimates future power purchases over a 30-year horizon. While bulk purchases do not necessitate forward planning for 30 years, existing statutes do require this length of forecast. NOVEC leverages forecast model insights to inform the magnitude (kilowatt-hours) and length of bulk power purchases from PJM. Economic metrics included in the model seek to characterize the basic load by capturing economic growth or decline in the Northern Virginia area. The basic load is the power requirement based solely on the size and typical consumption of customers, the number of which changes with time. To reasonably determine the size and growth rate of customers, local weather data is collected and used to remove the effects of weather on historical power purchases. This is known as weather normalization.

A major component of the weather normalization involves the calculation of heating degree days (HDD) and cooling degree days (CDD). Heating and cooling degree days provide a method of approximating the amount of power needed to heat or cool a building. HDD and CDD are calculated for each month. The equations for calculating HDD and CDD are:

$$HDD = \sum_{i=1}^{N} (T_b - \bar{T}_i)^+$$

$$CDD = \sum_{i=1}^{N} (\bar{T}_i - T_b)^+$$

Where N is the number of days per month, T_b is the base temperature coinciding with a non-heating/cooling temperature, and $\bar{T}_i$ is the average temperature per day. T_b, the base temperature, is selected based on regional factors and conditions. A base temperature is selected for HDD and CDD; the temperatures in between the base temperatures are considered to be non-heating or cooling – no heating or cooling needs to be done at those temperatures.



The figure above is a representation of HDD and CDD and the neutral area. As the actual temperature changes during a given period of time, it can either be below the lower bound, above the upper bound, or in between the two boundaries. Below the lower bound, power is typically used to provide heat. Above the upper bound, power is used to cool. The assumption in using HDD and CDD is that an insignificant amount of power is used to heat or cool in the neutral region between the boundaries.

Normalizing for the weather allows NOVEC to determine changes to their service-base, which provides necessary insight for capacity planning (such as infrastructure), in addition to deriving reasonable estimates of future power consumption. Rather than predicting the weather patterns for the future, the model uses the long-run average value for a given time period as determined from historical data. Historical power consumption since 1983, weather data since 1963, and economic forecasts at the state and county level are all inputs to the model. The output of the model is a monthly power demand forecast over a 30-year horizon.

2.2 Problem Statement

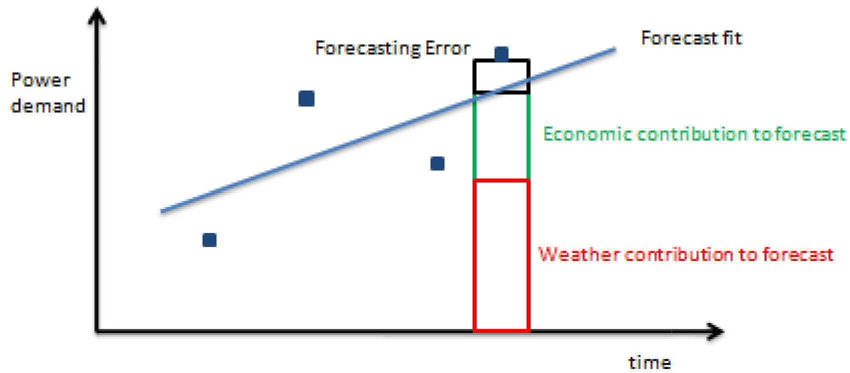
In order to predict future power demand, the model performs weather normalization for 50 years of hourly weather data and evaluates economic data provided by Moody's economic forecast. Each of these factors can be evaluated at the state, county, or Washington, D.C. metropolitan area. Accordingly, each data set must be weighted to correspond to the impact it would have on NOVEC's service area and thus power demand. For instance, Prince William County data would be more heavily weighted than those for Clarke County since NOVEC's territory in Prince William is much larger than in Clarke County; therefore, economic factors impact the basic load differently.

In accounting for these variables, NOVEC believes that the current model may no longer be the best available and that a new weather-normalization method may better reflect recent changes in weather trends. Improving the accuracy of the forecast would limit the amount of power that NOVEC has to buy beyond the bulk amount, thus decreasing costs. NOVEC requests analytical support to develop a new weather-normalization forecasting model or to determine that the existing model is the best available.

3.0 Scope

The purpose of this project is to develop a new weather normalization methodology to improve NOVEC's forecasting model by more accurately predicting future power demand. However, in order to develop a methodology to normalize for weather, the economic factors contributing to changes in power demand must also be accounted for in the analysis.

Notional Forecasting Methodology



The figure above gives a notional representation of the weather-normalization forecasting method. Over time, the power demand has increased. The forecast, which is fit to historic power demand, is made up of some combination of weather impact, economic impact, and forecasting error. In accomplishing the goal of changing the weather-normalization methodology, the weather's contribution to this model must change. As the weather contribution changes, either the economic contribution or the forecasting error must also change. Thus, in order to effectively develop a new weather normalization method, the economic factors must also be addressed.

3.1 Objectives

Our objective is to develop a model that will output a 30-year power demand forecast. The model will take into account historical data as inputs: customer and power purchase totals by month starting from 1983 and hourly weather data starting from 1963. These data sets provide us with a plethora of data that will necessitate extensive evaluation. The weather data contains over 400,000 records detailing hourly measurements of temperature, dew point, humidity, wind speed, and precipitation. Furthermore, historical power purchases provide over 6,000 data entries on total customer demand. In these historical data, some records are blank or contain errors, a problem that will have to be mitigated by this project through data validation. Additionally, this analysis will leverage Moody's state, county, and Washington, D.C. metro economic data starting from the 1970s. In particular, per sponsor guidance, data relating to employment, housing stocks, and GDP will be used to predict the growth or decline of NOVEC's customer base, though other metrics are available for analysis. Moody's economic data includes projections of economic variables across varied scenarios, only one of which is currently used to inform NOVEC's forecasts. Testing the model under additional scenarios offers a means to conduct sensitivity analysis and inform the sponsor's decisions with some measure of risk related to modeling assumptions.

3.2 System Requirements

NOVEC needs to gauge 30-year power requirements at a monthly resolution to inform bulk purchase negotiations. Historic and projected total power purchases must maintain an ability to characterize customer growth by type, residential or non-residential. In order to more accurately depict growth, NOVEC needs to be able to strip out the effects of weather; this is the ultimate purpose of the study and dictates a requirement to develop a methodology that will more accurately remove weather-effects. This will provide a better interpretation of the base load exerted by a dynamic customer base as well as reasonable estimates to how this base is changing.

Results of this study must also be able to synchronize with NOVEC's existing forecast model. To accomplish this, insights must be summarized within the context of two variables, heating- and cooling-degree days, which quantify cold and hot, respectively, temperature's impact on observed load. To assess the quality of the methodology to strip out weather-effects, the sponsor also requires an ability to report the error associated with output.

Although not required, a newly developed forecast model developed in conjunction with the weather normalization routine would be evaluated for enduring use at NOVEC. An ideal model for such consideration would need to be robust to changes in temperature and economic trends.

Based on these factors, the following requirements were derived:

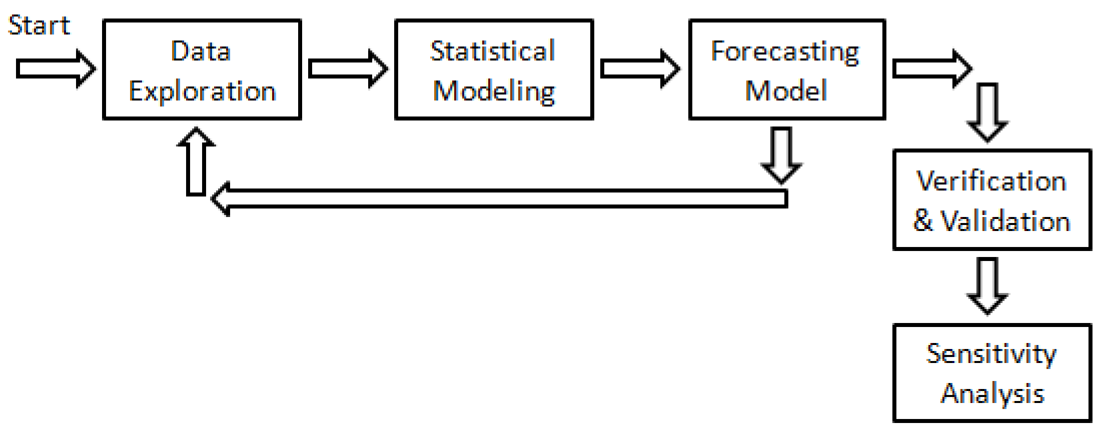
1.0 The project shall deliver a weather-normalization forecasting model (WNFM).
- 1.1 The WNFM shall accept data inputs.
  - 1.1.1 The WNFM shall accept as an input at least 51 years of historical weather data.
  - 1.1.2 The WNFM shall accept as an input at least 31 years of historical power demand data.
  - 1.1.3 The WNFM shall accept as an input Moody's economic data and economic forecast.
- 1.2 The WNFM shall output a weather-normalized power demand forecast.
  - 1.2.1 The WNFM shall output a heating degree day variable.
  - 1.2.2 The WNFM shall output a cooling degree day variable.
  - 1.2.3 The WNFM shall output a monthly power demand forecast for a 30 year time horizon.

2.0 The project shall deliver an error report that evaluates the accuracy of the WNFM.

3.0 The project shall deliver documentation for the WNFM.
- 3.1 The WNFM documentation shall include detailed description of the modeling process.
- 3.2 The WNFM documentation shall include detailed description of how to use the model.

4.0 Technical Approach

An overarching approach to accomplish the study's intent comprises a general sequencing of objectives. The flow chart below shows the high-level steps to complete the weather-normalization forecasting model. After Data Exploration and Statistical Modeling, the Forecasting Model will be constructed. At that point, our project will utilize an iterative methodology in order to modify the dynamics between weather-normalization and economic parameterization procedures. This will allow us to increase the accuracy of the forecasting model as well as observe the relationship between input data and end results. Concluding model development, Verification and Validation will be conducted with input from the sponsor. Assuming that we have time, Sensitivity Analysis will also be performed by varying the model parameters.



Each of the study phases introduced above is discussed in more detail below.

1) Data Exploration

- Identify and amend data gaps and inconsistencies.
- Determine diminished correlation of weather over long periods of time. Consider removing or lowering weighting of older weather data from the 1960s; entire data set is averaged in current model.
- Evaluate trends and empirical distributions in weather and economic data by plotting histograms, time series plots, and x-y scatter plots over varied timeframes.
- Utilize smoothing technique that accounts for seasonality of weather in addition to overall economic and meteorological trends.
- Investigate whether variable transformations are needed.

2) Statistical Modeling

- Determine best combination of explanatory variables to predict monthly power purchases; selected by statistical significance at 95% significance level.
- Provide 95% confidence intervals for independent variable parameters as well as for predicted values.
- Aggregate hourly weather data into monthly data to correspond to power load data.
- Select model based on goodness of fit test.

3) Forecasting Model

- Determine different options for weighting economic factors; current method uses service area in proportion to county size.
- Incorporate Moody's Economic projections using results from statistical model.

4) Verification and Validation

- Verify model consistency; ensure model is implemented as designed.
- Validate with NOVEC's power demand data from 2011-2012; serves as basis for comparison to current weather normalization methodology.

5) Sensitivity Analysis

- Vary weather parameters; test for impact of change in trends.
- Vary economic variable weights.

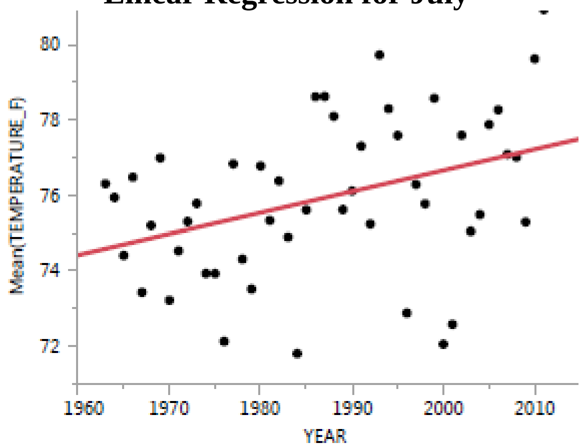
The goal is to improve their current modeling capability by quantifying a relationship between total monthly power purchases, temperature, and relevant economic factors relatable to sales growth which will then be used to inform the 30-year monthly forecast model.

5.0 Model and Architecture

5.1 Data Exploration

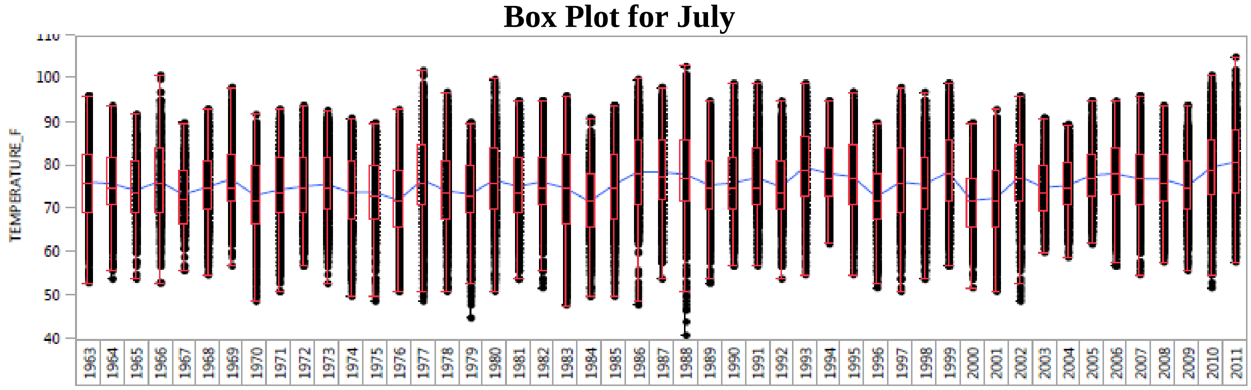
The first step in the project is to evaluate the input data, primarily the historical weather data. Already organized in an Excel spreadsheet, Excel and JUMP were utilized to evaluate the weather data. The goal of this step was to organize the data in an accessible format and begin evaluating the changes in temperature since 1963. Two statistical tests were conducted using JMP in order to evaluate the change in temperature for each month since 1963. First, linear regression was utilized to determine the trend in temperature per month. The results for the month of July are shown below.

Linear Regression for July

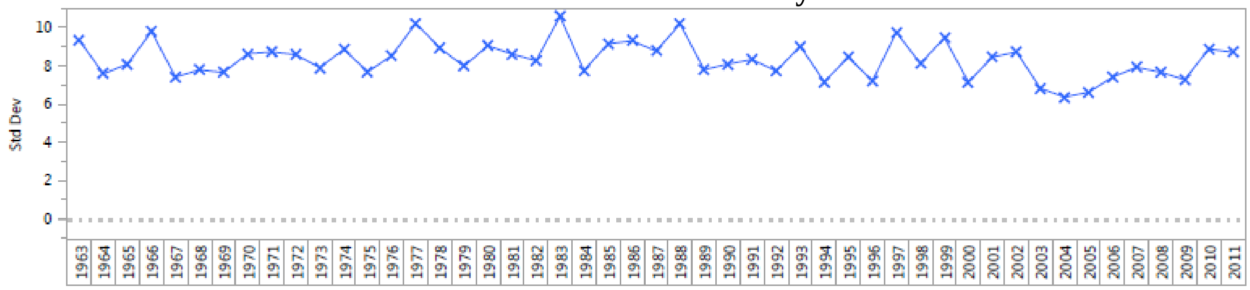


Although the slope of the line for each month varies, the linear fit for each month shows that the average temperature has increased since 1963. Linear regressions for each month can be found in Appendix B. Another factor of interest is whether the variability in temperatures is increasing. This was evaluated for each month using a box plot.

Box Plot for July



Standard Deviation for July



Evaluating the box plots and standard deviations for each month since 1963 showed that the variance in temperature has not changed significantly. Box plots for each month can be found in Appendix C. Thus, our initial data exploration revealed that temperatures are increasing, verifying NOVEC's need for a new weather-normalization methodology. However, statistical analysis shows that the variation is not increasing, meaning that the model does not need to account for increasing variance in temperature.

5.2 Model

5.2.1 Model Overview

The Weather Normalization Forecasting Model (WNFM) evaluates the data described above and outputs the 30-year monthly forecast. The model is constructed in Excel and R and follows a seven-step process that is described below.

1. Find the relationship between customer base and economic variables.
2. Find the relationship between average customer usage and HDD and CDD.
    a. Different regression analysis is performed to find the base usage, usage due to HDD and/or CDD.
    b. We have also adopted NOVEC's practice to convert a non-residential customer to an equivalent amount of residential customer as a supplementary approach to predict average non-residential usage in addition to the regression method.
3. Using various time series methods to find the relationship between HDD and CDD towards time.
4. Based on step 3, we have predicted future HDD and CDD and they are used together to predict the change in average customer usage from step 2.
5. Based on step 4 and the result from step 1, we can predict the total usage as contributed by base load as well as the weather from both residential customer as well as non-residential customer.
6. Combine steps 1 and 2 to find the predicated demand due to customer behavior change and economic development.
7. A linear regression model containing the economic variables as well as HDD and CDD is used to forecast the total load as a comparison.

Parameters that may be changed for each model run include the boundaries for CDD and HDD, the dates to define the historic domain for regression modeling as well as weather data, economic variables to be included, and the economic scenario providing varied projections of future economic variable values. For this study the only economic scenario assessed was the base case, all economic variables provided were included for all model runs, and neutral zone boundaries for calculating CDD and HDD were held at 55 to 65 degrees per NOVEC's existing modeling construct.

Further analysis may be conducted as each one of these variables can be changed in the GUI provided in Excel. The customer base is forecasted based on a linear regression. The HDD and CDD were forecasted using three different methods: Holt-Winters method, ARIMA method, and BAT method.

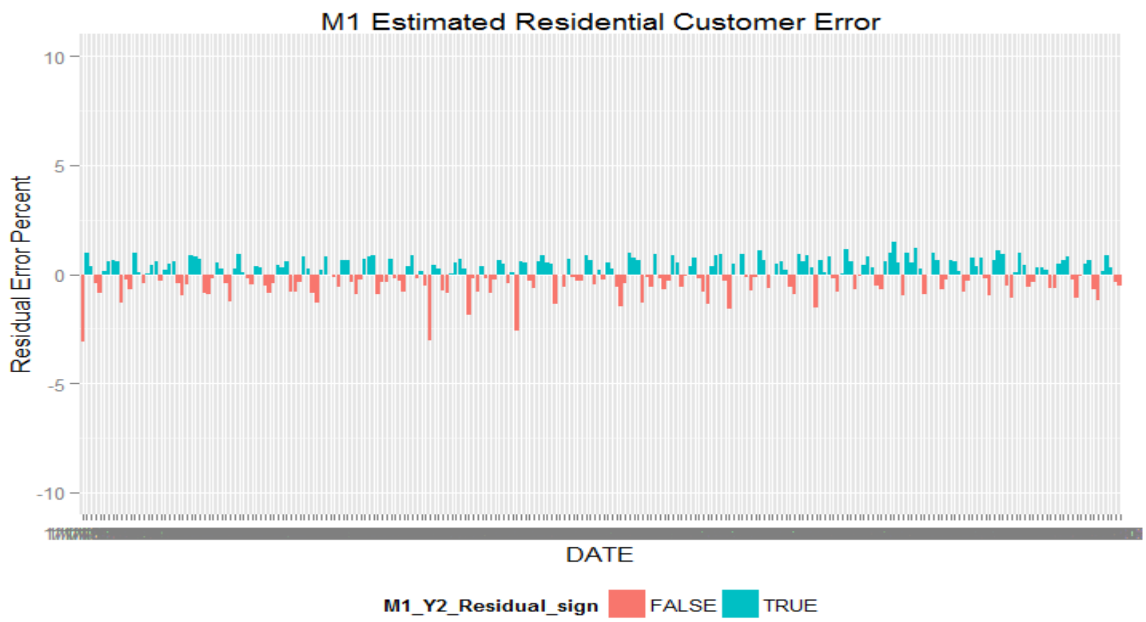
5.2.2 Assumptions and Limitations

- Assumptions
  - Neutral zone between HDD/CDD has no impact on power consumption.
    - 55 and 65 degrees are the lower and upper bounds utilized in the model.
  - Economic variables currently utilized provide proper indicators for power demand:
    - Employment: Total Non-Agricultural
    - Gross Metro Product: Total
    - Housing Completions: Total
    - Households
    - Employment (Household Survey): Total Employed
    - Employment (Household Survey): Unemployment Rate
    - Population: Total
- Limitations
  - Due to time constraints and after consulting with NOVEC, it was determined that this project will not attempt to develop a deep understanding of NOVEC's current forecast model. This could hinder adopting the WNFM into NOVEC's existing model. Also, this limitation could skew comparisons of forecast accuracy.
  - Due to time constraints, less time was spent evaluating the economic regression model to determine customer base. This has the potential to cause inconsistent forecast comparisons between the WNFM and NOVEC's current model output.
  - Only one set of economic data was used in the model. Although state, county, and Washington, D.C. metro data is available, the model only uses the metro data.

6.0 Results and Sensitivity Analysis

6.1 Estimating Customer Base

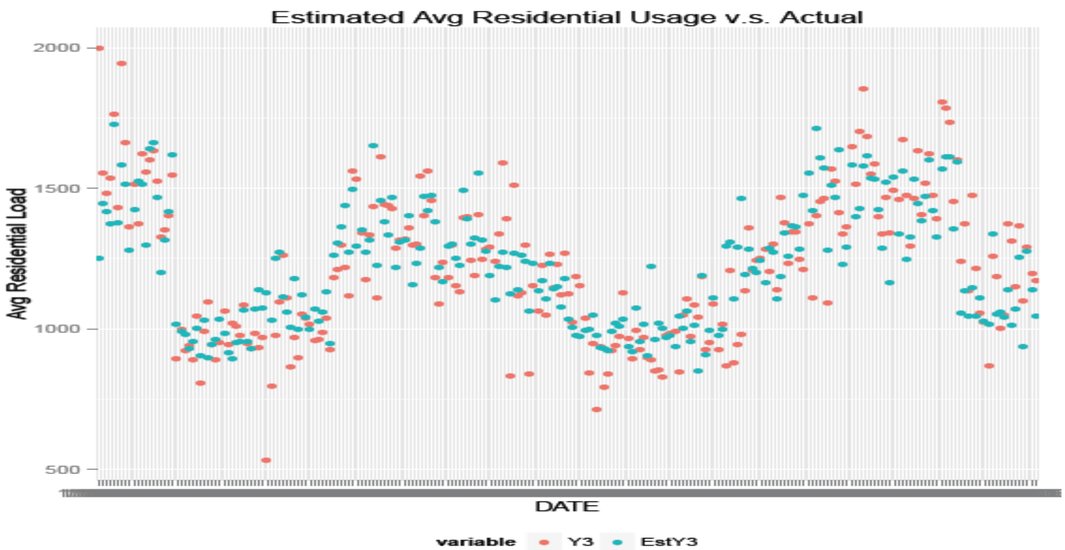
A linear regression model is used to predict the customer base, either residential or non-residential, as a function of the economic variables NOVEC has been using. The adjusted R-square of 0.99 was found for both customer types, indicating that the linear regression model almost explains all the variations in the customer base. As shown in the chart below, the maximum delta between estimated residential customer count and actual residential customer count is about 3% and the error typically fluctuates between +/- 1%. Since the focus of the project is to identify the weather impact and NOVEC already has an economic model they are comfortable with, we decided to utilize the linear regression model.



6.2 Estimating Customer Average Usage

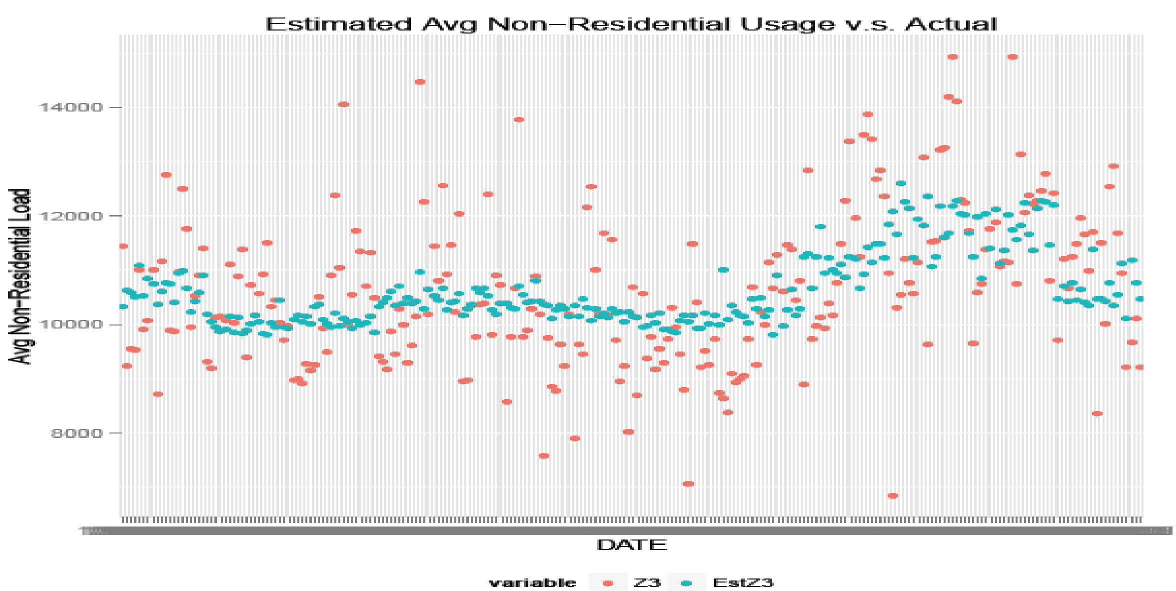
6.2.1 Estimating Average Residential Customer Usage

The average residential customer usage was correlated with the economic variables and the weather variables (HDD and CDD). The regression analysis found that the weather model alone has an adjusted R-square of 0.66, which suggests a decent fit. The complete model that contains both economic variables and weather variables only has a modest improvement at 0.67. This suggests that economic variables are not contributing to residential customer's behavior as measured by the average usage. Hence, it was determine that the more straightforward approach would be utilized to correlate the average residential customer load with the two weather variables. An overlay of the estimated average residential customer usage and the actual is provided below.

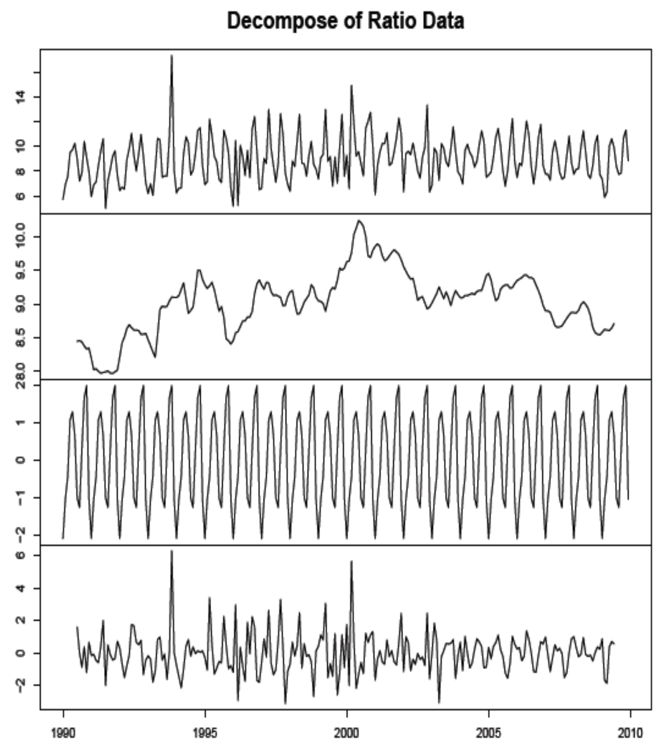


6.2.2 Estimating Average Non-residential Customer Usage

The same approach as outlined in 6.2.1 is adopted for analyzing the behavior of an average non-residential customer. However, it was found that while weather has a moderate explanation of the behavior change of the average residential customer, it does not adequately explain the average non-residential customer. The combined regression model of both economic variables and weather variables only has an adjusted R-square of 0.39. Furthermore, if we split the combined model into the economic part and the weather part, the team found that the economic part has an adjusted R-square at 0.17 while the weather part has an adjusted R-square of 0.21. This suggests that both economy and weather contributes to the behavior of an average non-residential customer and neither of them are a good model. The chart below shows how good a fit it is between the actual average non-residential usage and the predicted usage using the weather variables only.



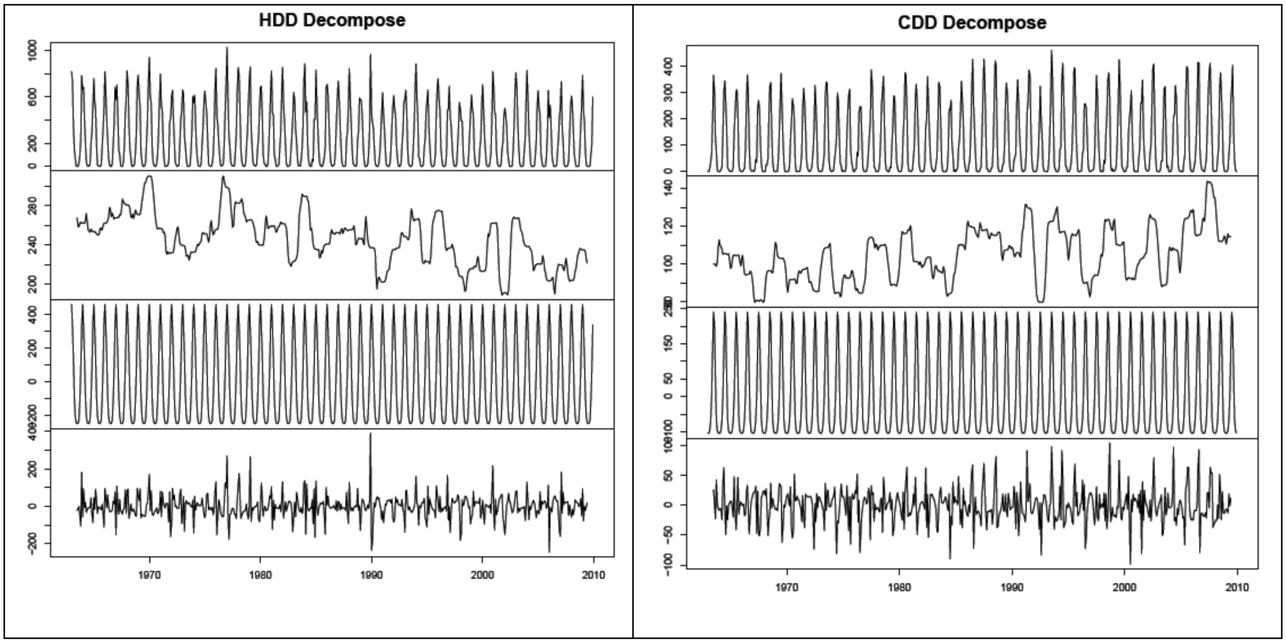
Alternative approaches to estimating average non-residential load were studied, and one method surfaced through our meetings with NOVEC. NOVEC's current model converts a non-residential customer to a residential customer based on a fixed ratio and then, based on the forecast made for a residential customer, to derive the intended usage from a non-residential customer in the same time period. The ratio of residential to non-residential customers from the historical data was calculated. The Holt-Winters method is used to decompose the ratio data into trend, seasonality, and error and it is found that the ratio does have a seasonal impact. Using a constant mean value, the conversion can be improved by using a forecast that factors both the trend and the seasonality. The team will forecast the average non-residential customer usage using both the regression method as well as the ratio method.

<table>
<tr><th>Remarks</th><th>Plot of the Ratio Data</th></tr>
<tr><td>
1. Row 1 is a plot of the actual ratio data at a monthly resolution. The insight is that while the mean of the ratio does not vary a lot from 1990 and 2009, there seems to be a seasonal impact.
2. Row 2 is the impact from trend, and the trend line suggest that the first 10 years see a modest increase while the last 10 years see a corresponding drop which causes the overall impact from trend to be zero. Will the change in the last 10 years be due to explanatory variables, like improvement in technology that helps non-residential customer save power? This is an interesting question for further study.
3. Row 3 is the seasonal impact. Since there are 12 months, the season is set at 12.
4. Row 4 is the residual, and the residual has a mean of zero and almost constant variance which indicates a good fit.
</td><td>


</td></tr>
</table>

6.3 Estimating Weather Variables – HDD & CDD

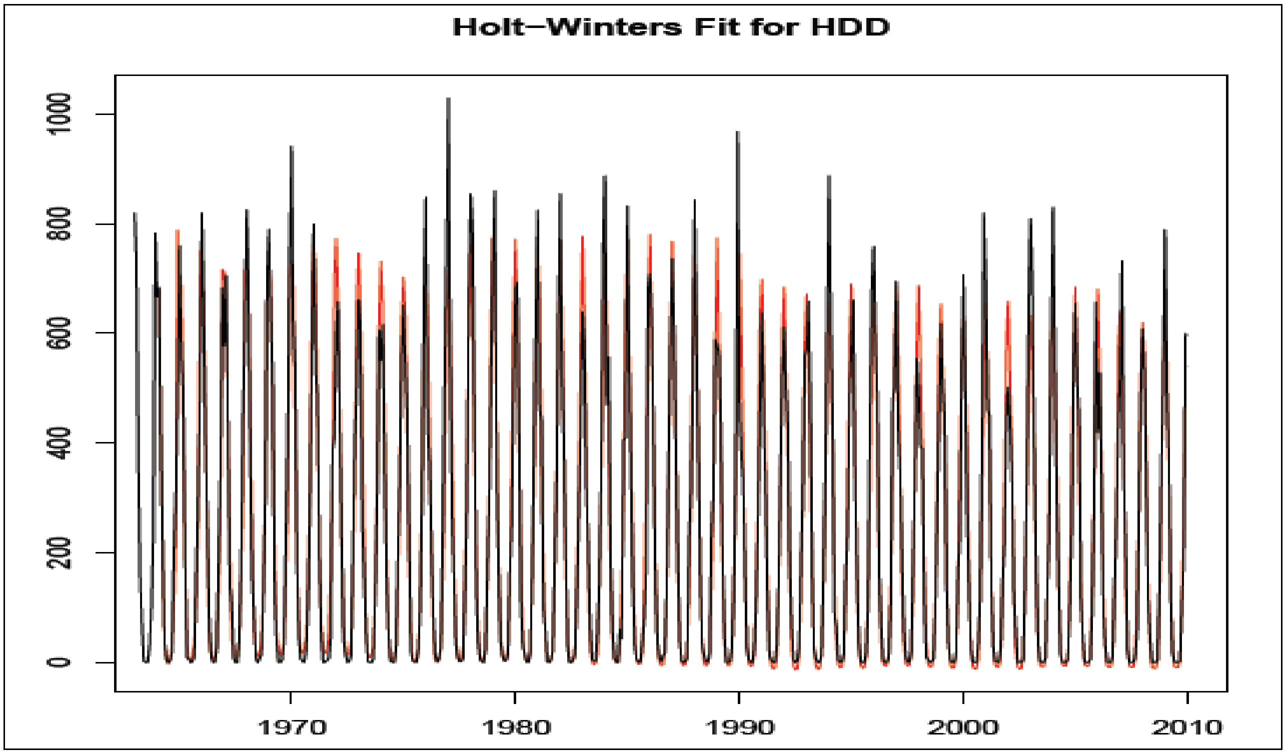
The team attempted to 3 different methods, namely Holt-Winters, ARIMA and BAT, to understand how the weather variables are contributing to average residential and non-residential customer usage. Before we go into detail, the first step is to look at the data and see if there are any hidden insights.



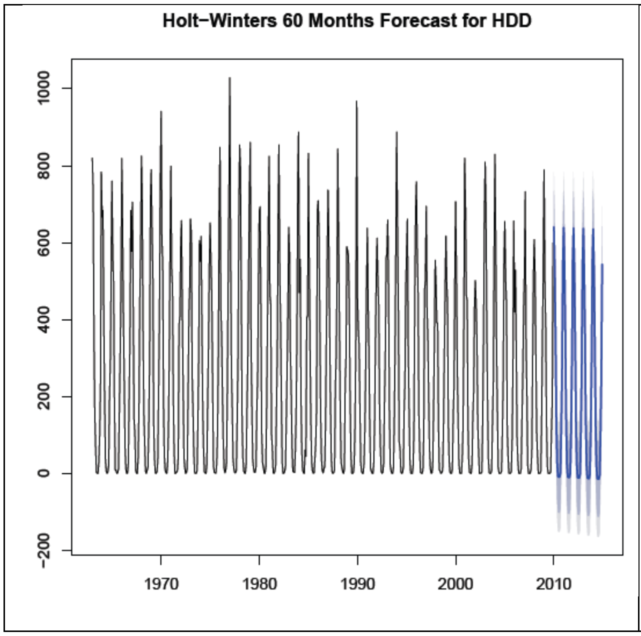
Row 1 is a plot of the actual data. As one can see, the HDD and CDD clearly follow a certain seasonal pattern. Since there are 12 months in a year, the seasonality can be modeled at a monthly level for better granularity. Row 2 is the trend of the data. As one can see, the HDD is trending down while the CDD is trending up. This indicates that while the year to year weather data may fluctuate, the trend should be modeled in any forecasting model to capture the global warming impact. Row 3 shows how the seasonality is affecting the impact from HDD and CDD. Row 4 shows the residual which has a mean of about 0 and almost constant variance, suggesting it is a good fit.

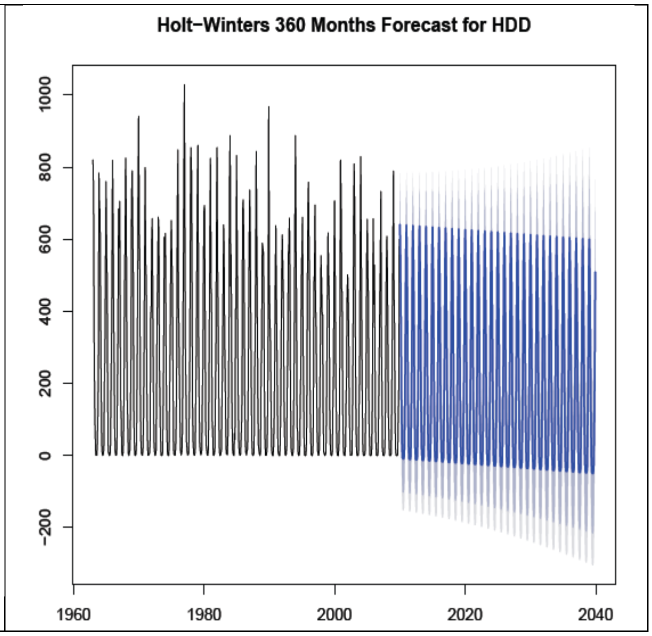
6.3.1 Holt-Winters Method

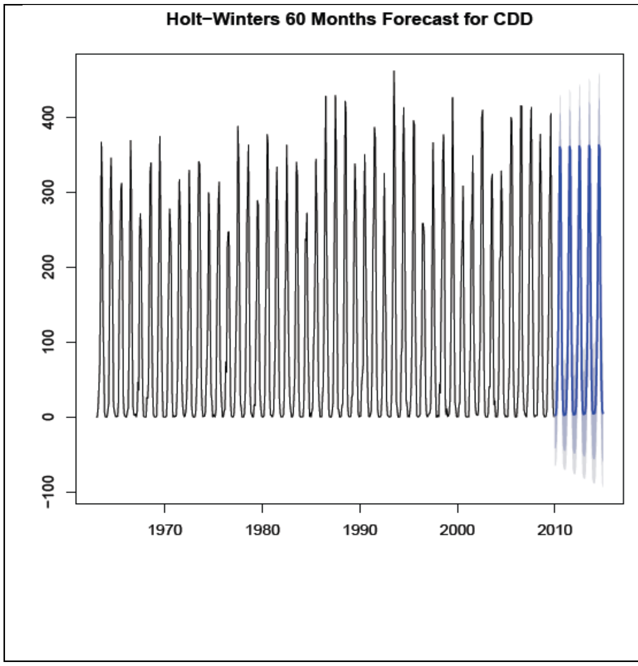
The Holt-Winters (HW) method is first tested to predict HDD and CDD. As you can tell from the chart below, HW method produces a good fit between the actual data and the observed data.

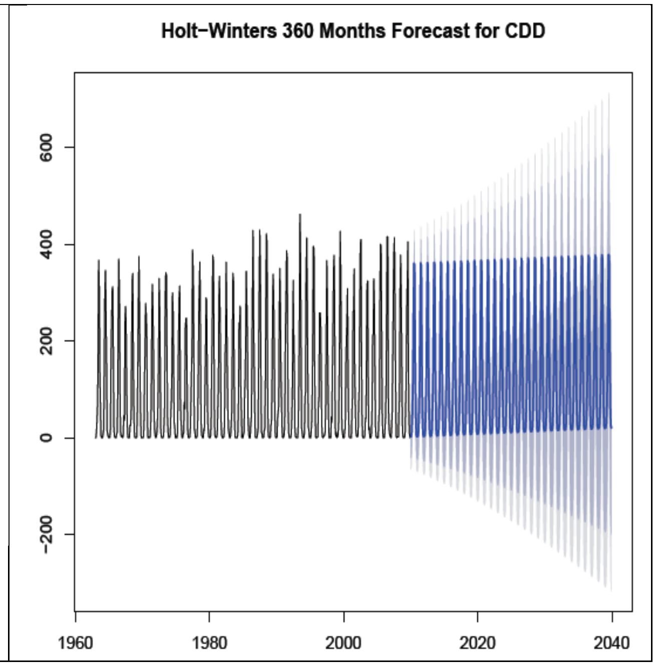


The forecasted 5-year and 30-year forecast for HDD and CDD are plotted below. The forecast indicates that the HDD is slowly decreasing while the CDD is slowly increasing. The shaded area indicate the upper (lower) 80/95 percentile.



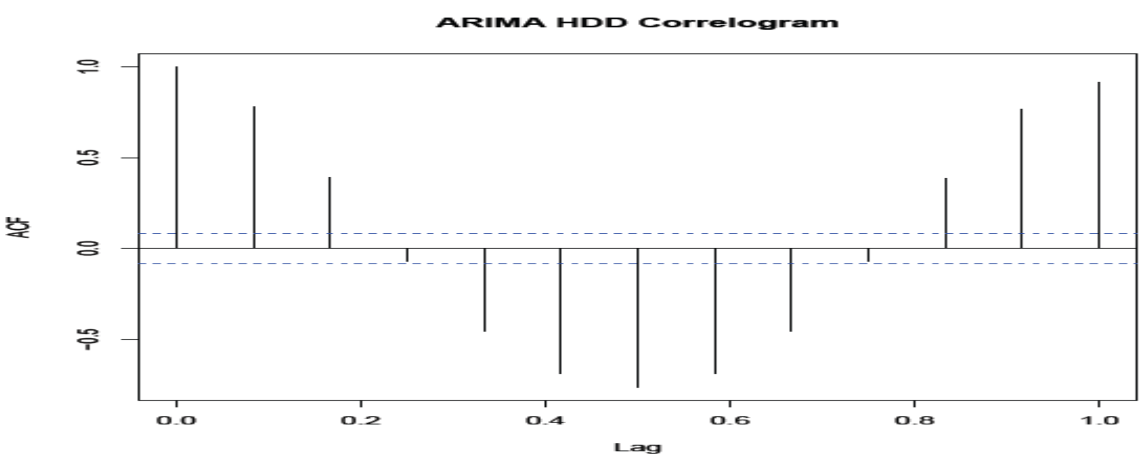






6.3.2 ARIMA Method

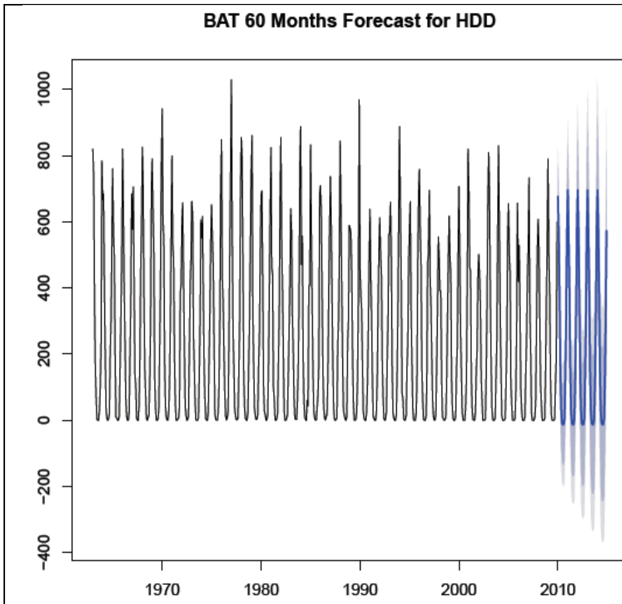
The ARIMA model is good at tracking the correlations. Applying the HDD and CDD data directly does not yield a good fit as the correlogram violates the control limit.

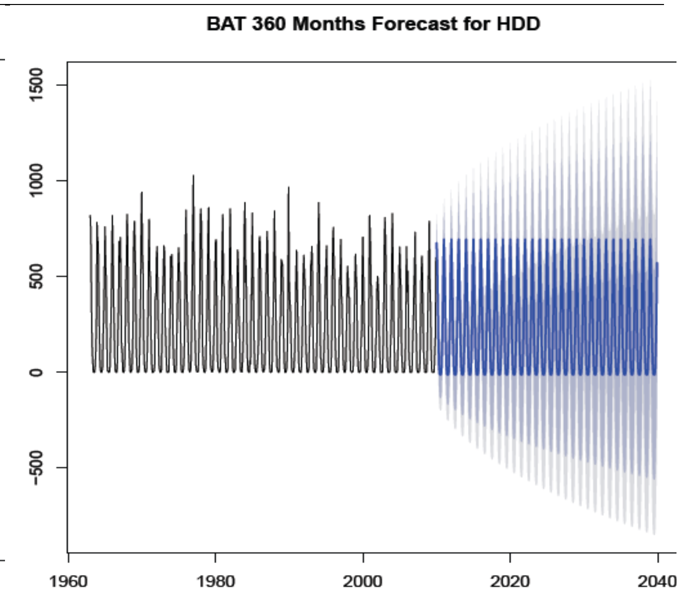


Due to the lack of a good fit, the ARIMA model was not utilized as a forecasting methodology. An area of further study is to calculate the changes of daily HDD and CDD and uses that to predict future HDD and CDD changes.

6.3.3 BAT Method

The BAT method is basically a superset of the Holt-Winters method which allows users to set up more than one seasonal impact. The predictions from BAT model is fairly similar to the ones made by Holt-Winters method as indicated by the plot below.

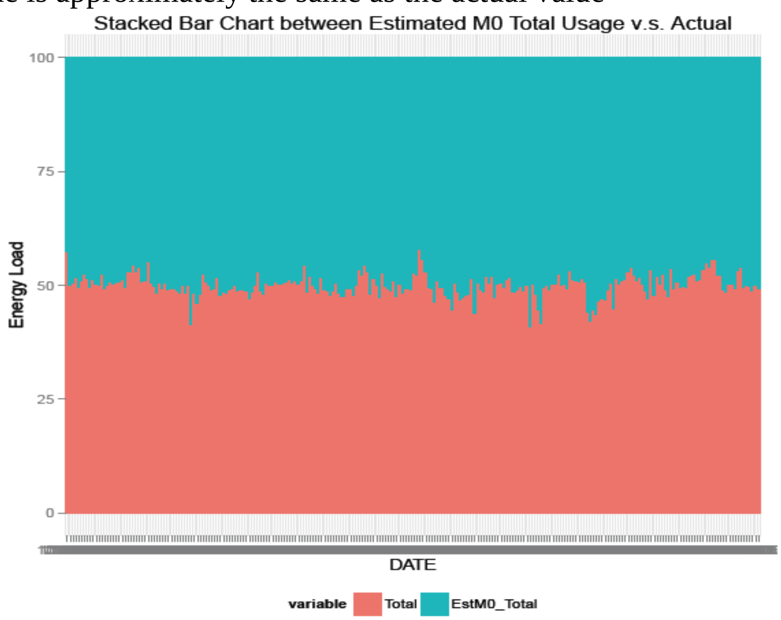




6.4 Split Linear Method

Usage should be a function of economic contributions, weather contributions, etc. The adjusted R square for the regression model stands at 0.925 which indicates that the model is reasonably well at explaining the total observed load as a function of economical variables and the weather impact from HDD and CDD. However, a closer look at the t-value for each variable indicates that not all of them are statistically significant which indicates that the model be over-fit.

Below is a stacked bar chart between predicted and actual load. If the regression model is reasonably good, we would expect the stack chart to fluctuates at the 50 percentile indicating that the predicated value is approximately the same as the actual value



7.0 Evaluation

7.1 Split Models with HDD/CDD Trends

Initial insights prompted further analysis into adjusting the ratio methodology while omitting further assessment of the combined linear regression, ARIMA, and BAT modeling approaches. Regression models were adjusted to include a first order interaction term between CDD and HDD, resulting in average load estimates in accordance with the following equation:

$$y_i = \beta_0 + \beta_1(CDD) + \beta_2(HDD) + \beta_3(CDD)(HDD)$$

Regression statistics from each of residential and non-residential customers determine the seasonal effects, as before, using the above equation. The results constitute the "Split Regression Model" approach using CDD and HDD forecasts via the HW methodology. Thus, total load is calculated as follows:

$$\begin{aligned}&\#RES * [\beta_0 + \beta_1(CDD_{HW}) + \beta_2(HDD_{HW}) + \beta_3(CDD_{HW})(HDD_{HW})]_{RES}\\ &\qquad + \ \#NonRES * [\beta_0 + \beta_1(CDD_{HW}) + \beta_2(HDD_{HW})\beta_3(CDD_{HW})(HDD_{HW})]_{NonRES}\end{aligned}$$

7.2 Average Load Trends

The non-residential consumption relative to the residential load was analyzed as a time series trend. This trend was forecasted using HW and applied to the residential average load which, in turn, is a function of forecasted CDD and HDD.

$$Avg. Load\ Ratio = \frac{Avg\ NonRES\ Load}{Avg\ RES\ Load}$$

This approach allows the non-residential customer base to be expressed in terms of residential service equivalency; a resulting total load is then computed as follows:

$$\begin{aligned}&[\#RES + \#NonRES * AvgLoad_{HW}]\\ &\qquad * [\beta_0 + \beta_1(CDD_{HW}) + \beta_2(HDD_{HW}) + \beta_3(CDD_{HW})(HDD_{HW})]_{RES}\end{aligned}$$

7.3 Split CDD and HDD Trends

Subtracting the base load out of the average load for each of residential and non-residential service types allows the assessment of relative seasonal demand between each service type. To relate the relative impact on non-residential vice residential service type the ratio of relatable split model regression coefficients are first evaluated:

$$CDDratio = \frac{[\beta_1]_{NonRES}}{[\beta_1]_{RES}} \qquad\qquad HDDratio = \frac{[\beta_2]_{NonRES}}{[\beta_2]_{RES}}$$

From the residential service regression model, we define the impact of CDD and HDD on residential average loading. Under the assumption that CDD and HDD equally contribute to the first-order interaction term, the influence is halved between them resulting in the below:

$$CDDimpact = \left[\beta_1(CDD_{HW}) + \frac{\beta_3(CDD_{HW})(HDD_{HW})}{2}\right]_{RES}$$

$$HDDimpact = \left[\beta_2(HDD_{HW}) + \frac{\beta_3(CDD_{HW})(HDD_{HW})}{2}\right]_{RES}$$

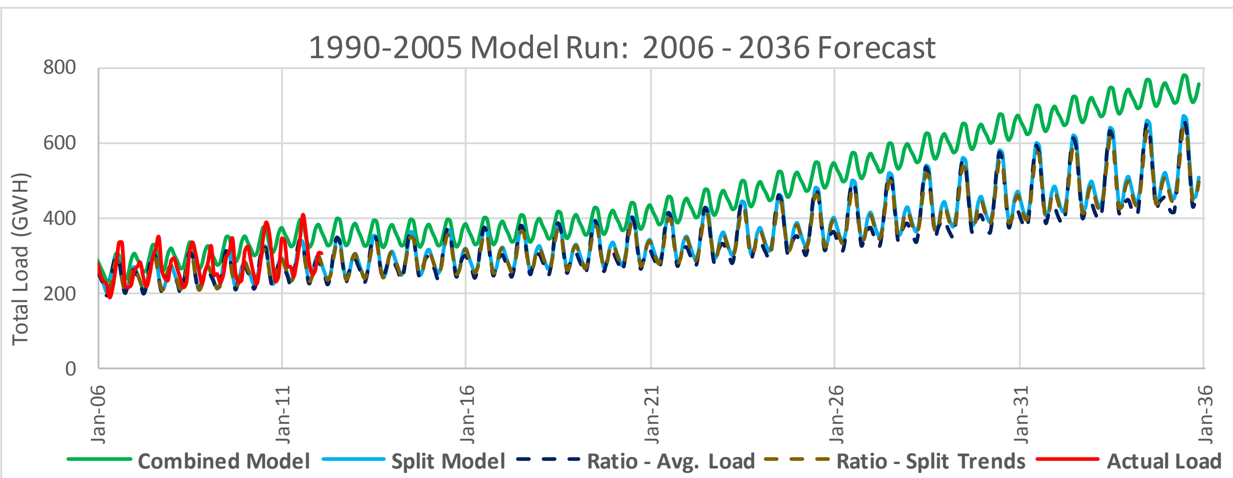
The total forecasted monthly demand is therefore:

$$\#RES * [\beta_0 + CDDimpact + HDDimpact]_{RES} + \#NonRES * [\beta_0 + CDDratio * CDDimpact + HDDratio * HDDimpact]$$

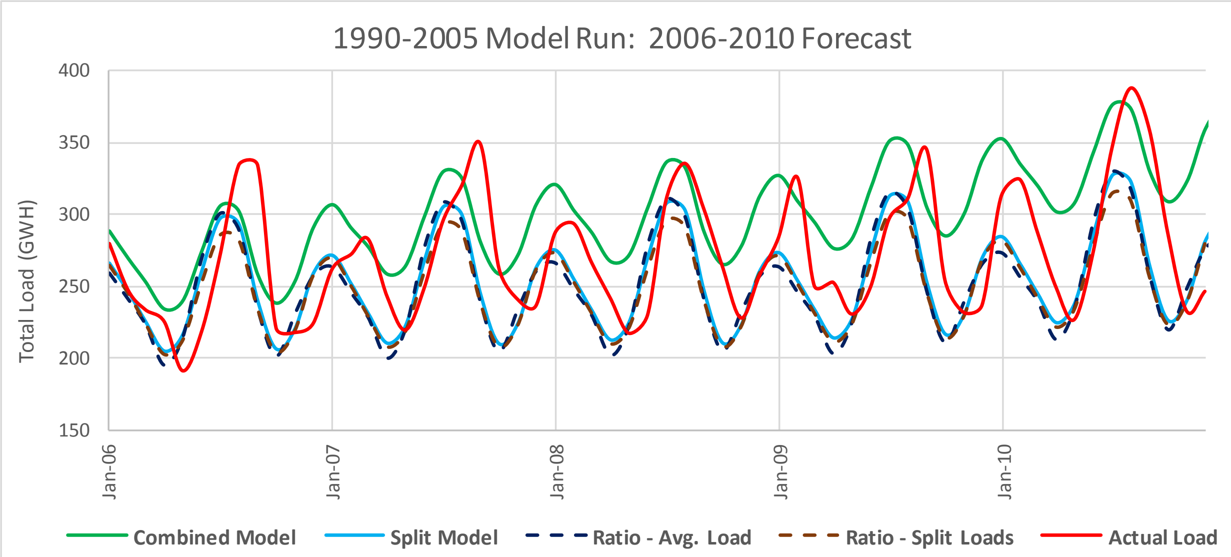
7.4 Model Results

To assess the best candidate weather normalization routine each of the developed model were tested by forecasting against historic observations on monthly power consumption. Merit was given to a modeling construct that balances accuracy and robustness. To accomplish this, predictions for cumulative load for the first five years was compared to actual data.

The figure below shows such output using 1990-2005 data to parameterize a forecast for 2006-2036.



An illustration of the first 5 years only is depicted in the next figure.



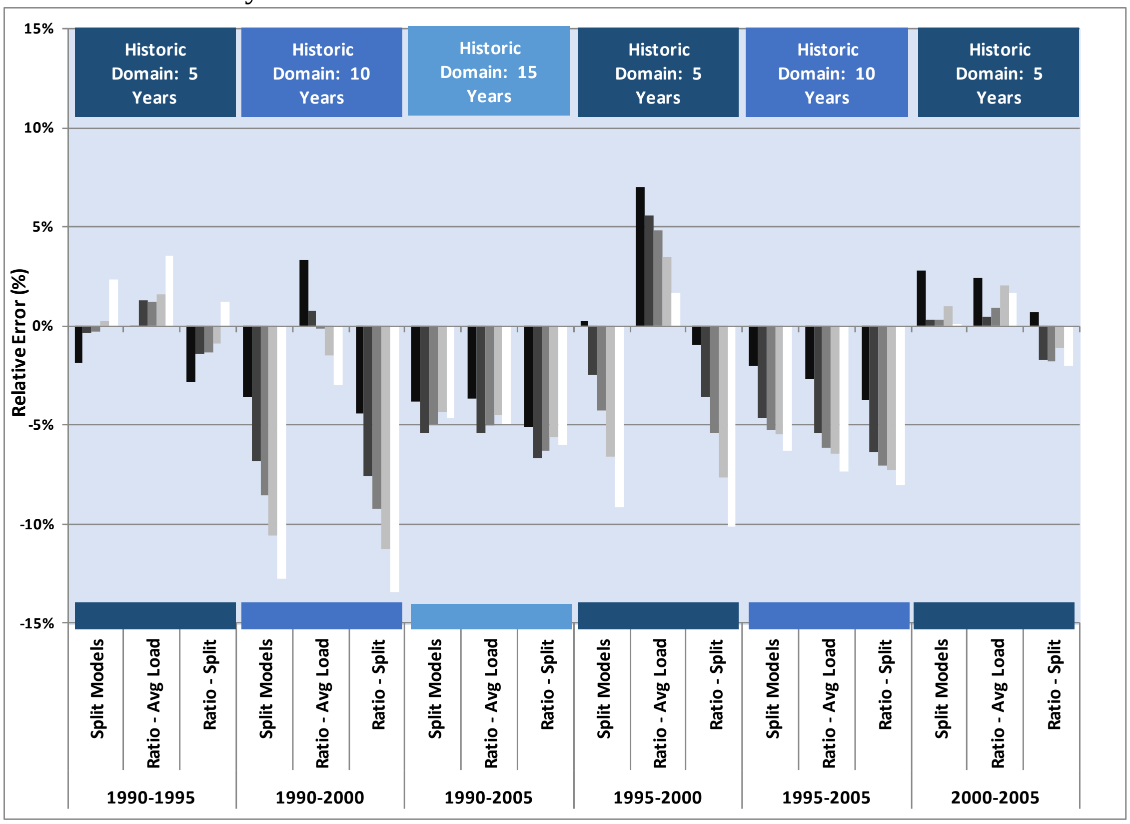
In order to test the robustness of each model, the models were tested across varied intervals of time characterized by noticeable changes in econ data records. The figure below illustrates the sensitivity of the combined regression model to such discontinuities in trends.

(Data Domain)	YEARLY %-ERROR in CUMULATIVE LOAD	FORECAST HORIZON				
	Modeling Approach	1	2	3	4	5
1990-1995	Split Models	-2%	0%	0%	0%	2%
	Ratio - Avg Load	0%	1%	1%	2%	4%
	Ratio - Split	-3%	-1%	-1%	-1%	1%
1990-2000	Split Models	-4%	-7%	-9%	-11%	-13%
	Ratio - Avg Load	3%	1%	0%	-1%	-3%
	Ratio - Split	-4%	-8%	-9%	-11%	-13%
1990-2005	Split Models	-4%	-5%	-5%	-4%	-5%
	Ratio - Avg Load	-4%	-5%	-5%	-5%	-5%
	Ratio - Split	-5%	-7%	-6%	-6%	-6%
1995-2000	Split Models	0%	-2%	-4%	-7%	-9%
	Ratio - Avg Load	7%	6%	5%	3%	2%
	Ratio - Split	-1%	-4%	-5%	-8%	-10%
1995-2005	Split Models	-2%	-5%	-5%	-6%	-6%
	Ratio - Avg Load	-3%	-5%	-6%	-6%	-7%
	Ratio - Split	-4%	-6%	-7%	-7%	-8%
2000-2005	Split Models	3%	0%	0%	1%	0%
	Ratio - Avg Load	2%	0%	1%	2%	2%
	Ratio - Split	1%	-2%	-2%	-1%	-2%

Reference the table above, a forecast starting in 2006 would benefit from fewer years used to initialize forecasts; as is seen by comparing the 1990-2005, 1995-2005, 2000-2005 forecast error. While we make no assertion for the true meaning it is likely that a change in a running trend occurred which makes a more recent depiction of the new evolution of monthly consumption more revealing. A more expansive dataset would allow further testing of the sensitivity to time though a deeper look into the economic variables that best predict number of customer for each service may also improve forecasts by aligning the historic domain to one that matches expectations for the future economic outlook.

8.0 Recommendations

The battery of excursions lead to characterization of the split model methodology, using a hierarchy of sub-modules similar to the construct NOVEC currently employs, as the best candidate for implementation. The resulting accuracy for relative error of cumulative power purchases for each of 5 years on back-tested data is illustrated below.



Observations also lend themselves to the characterization that the different methodologies tested provide information that may be beneficial to holistically inform bulk purchases. For instance, an average of a forecasted demand may diversify the risk across parameters that each method may possess certain sensitivities towards.

Thus, based on the results of the WNFM, we recommend that NOVEC utilize the Split Regression Model with the Holt-Winters HDD/CDD forecasting methodology to augment their current forecasting model. The capabilities provided by the WNFM will inform NOVEC's power purchases and give them the ability to perform additional analysis. Although this project is limited to temperature data from 1963-2011, HDD and CDD boundaries of 65 and 55 degrees respectively, economic data from 1990-2011 using the seven economic variables described above, the baseline economic scenario, and customer totals and power consumption from 1990-2011, the WNFM will give NOVEC the ability to change each one of these limitations and generate multiple forecasts depending on their needs.

9.0 Future Work

Future work on this problem can be divided into two categories: 1) Evaluation using the WNFM and 2) Additional work to modify the WNFM. As described above, there are numerous parameters and inputs that can be changed in the model. These include changing the temperature inputs year, changing the upper and lower boundaries for calculating HDD and CDD, adding additional economic factors to evaluate, including economic scenarios beyond the baseline scenario, and incorporating economic data from other geographic areas. As noted above, we recommend that utilize the WNFM to perform this analysis. Future projects could also focus on these parameters. Secondly, additional work can be done on improving the methodologies used by the model. This project focused on the seasonal load and did not develop as robust a methodology for determining the economic contribution to power demand. Future work focused on the economic aspect of the model could significantly improve forecasting.

References

[1] www.novec.com

[2] NOVEC 25th Anniversary History Film http://www.youtube.com/watch?v=2qfeOKnPPGg

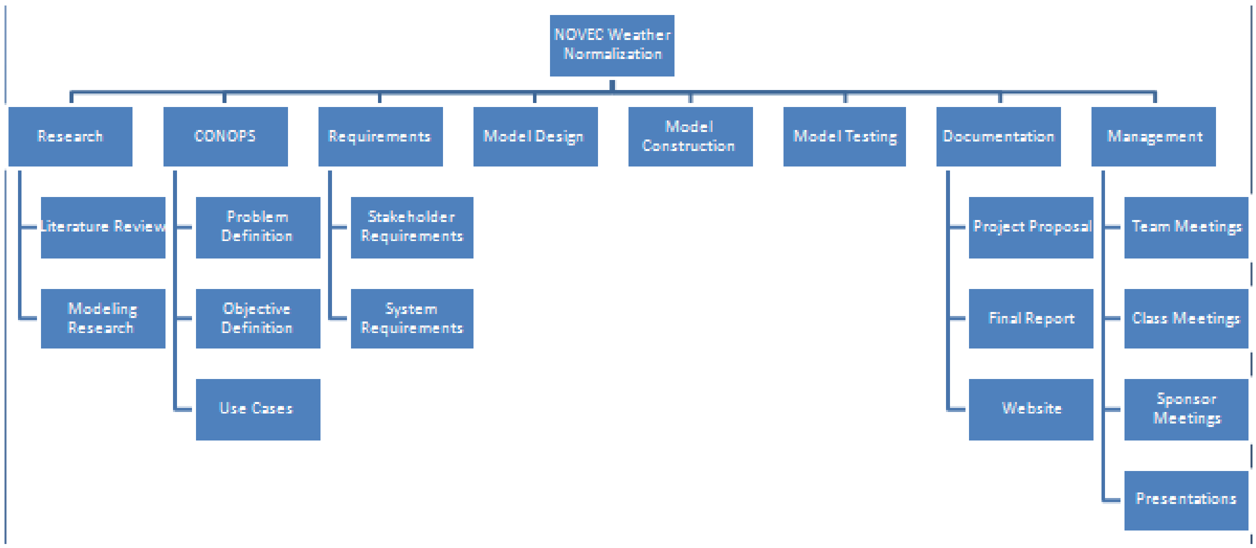
[3] Kozera, Lohr, McInerney, and Pane, “Improving Load Management Control for NOVEC”. http://seor.gmu.edu/projects/SEOR-Spring12/NOVECLoadManagement/team.html

[4] Didier Thevenard, “Methods for Estimating Heating And Cooling Degree-Days to Any Base Transactions;2011, Vol. 117 Issue 1, p884.

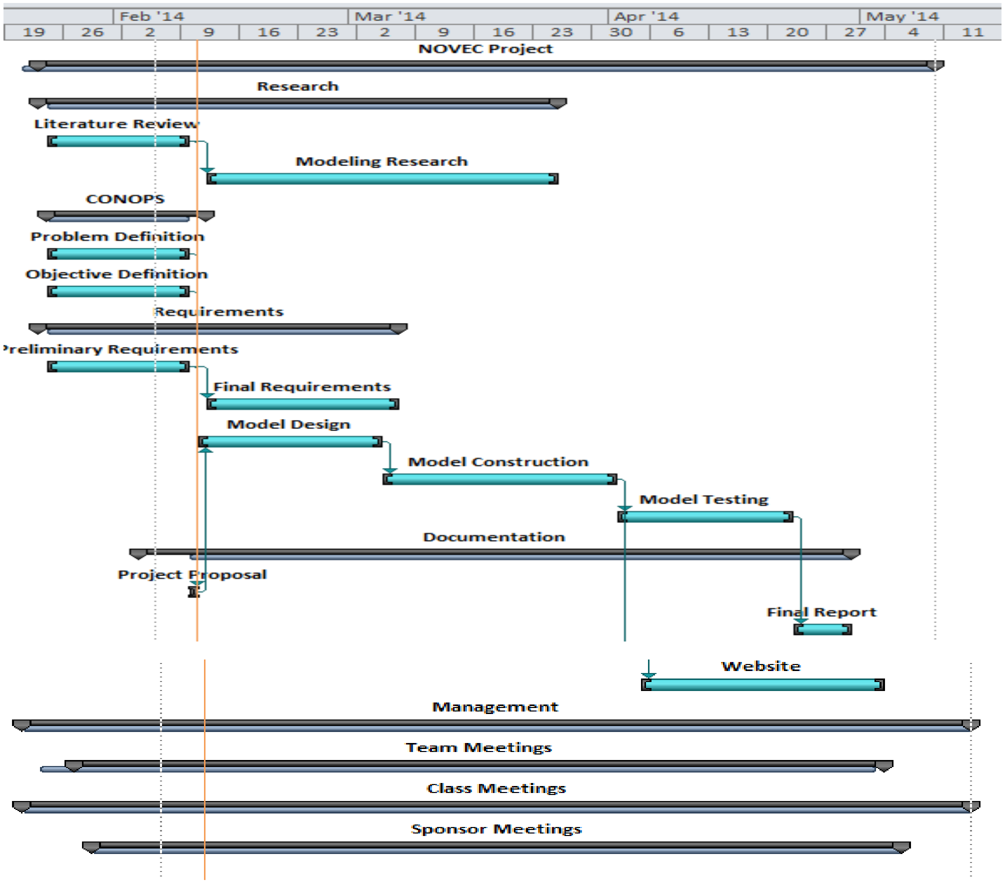
Appendix A: Project Management

Our project plan includes a Work Breakdown Structure (WBS), a project schedule, and earned value graphs. The earned value graphs are based on timesheets based on the WBS.

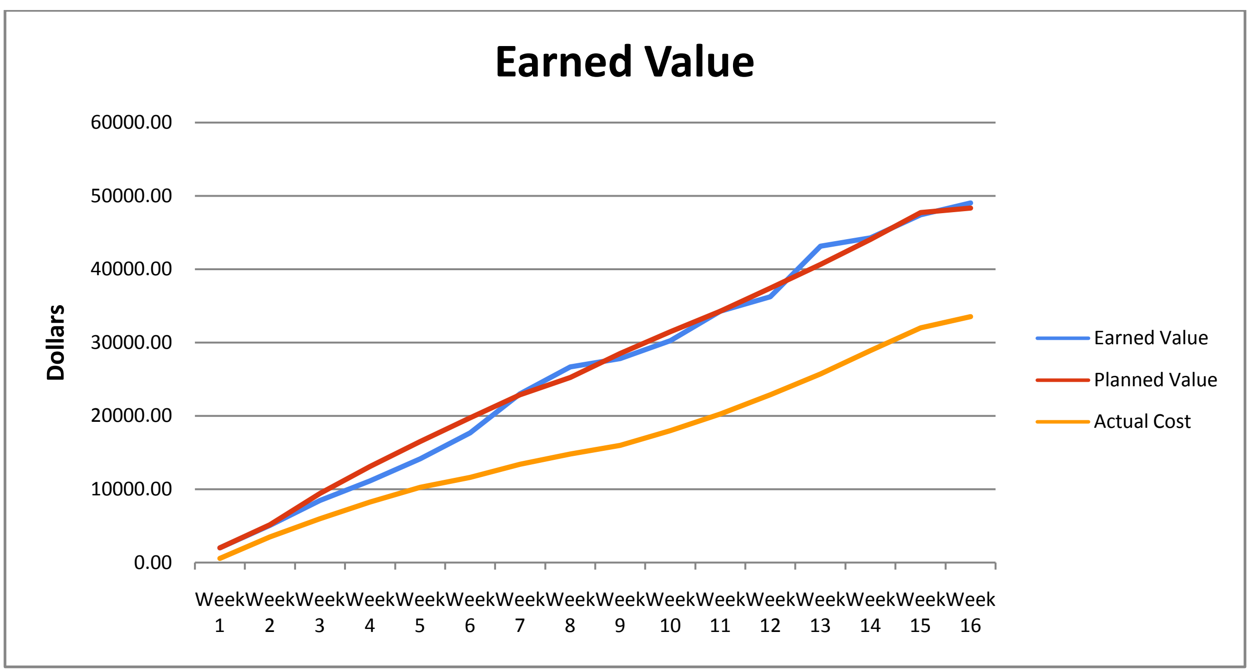
NOVEC Project WBS



NOVEC Project Schedule



NOVEC Project Earned Value Chart

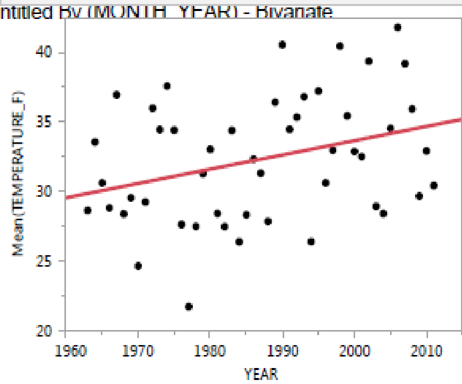


Appendix B: Historical Temperature Mean

Linear regression of temperature per month since 1963.

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=JAN

Untitled By (MONTH YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -171.8781 + 0.1027598*YEAR

Summary of Fit

RSquare	0.106965
RSquare Adj	0.087964
Root Mean Square Error	4.287472
Mean of Response	32.30575
Observations (or Sum Wgts)	49

Analysis of Variance

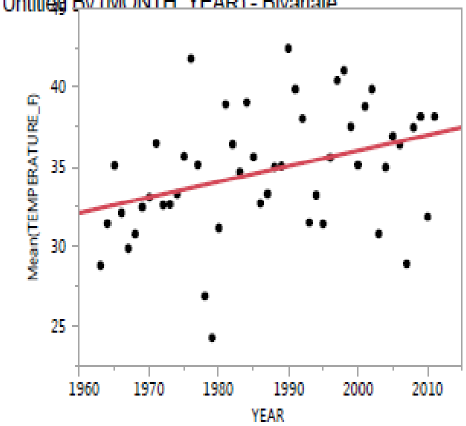
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	103.48394	103.484	5.6295
Error	47	863.97340	18.382	Prob > F
C. Total	48	967.45734		0.0218 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-171.8781	86.05916	-2.00	0.0516
YEAR	0.1027598	0.04331	2.37	0.0218 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=FEB

Untitled By (MONTH YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -158.8531 + 0.0974388*YEAR

Summary of Fit

RSquare	0.12634
RSquare Adj	0.107752
Root Mean Square Error	3.699956
Mean of Response	34.75783
Observations (or Sum Wgts)	49

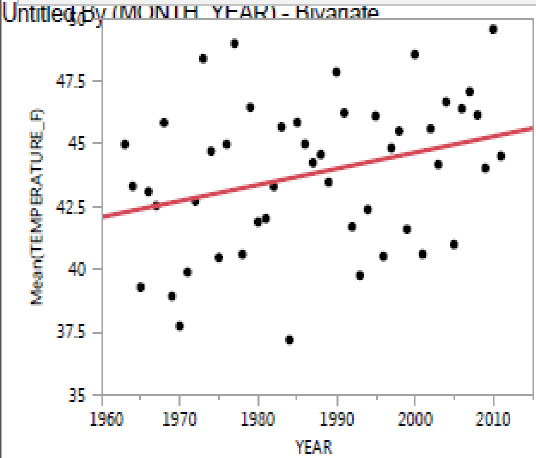
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	93.04437	93.0444	6.7967
Error	47	643.41455	13.6897	Prob > F
C. Total	48	736.45892		0.0122 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-158.8531	74.26639	-2.14	0.0377 *
YEAR	0.0974388	0.037375	2.61	0.0122 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=MAR



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -83.83776 + 0.0642463*YEAR

Summary of Fit

RSquare	0.095344
RSquare Adj	0.076096
Root Mean Square Error	2.857637
Mean of Response	43.81964
Observations (or Sum Wgts)	49

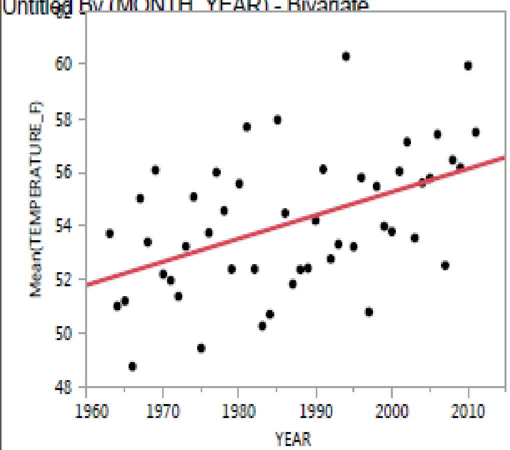
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	40.45035	40.4504	4.9535
Error	47	383.80629	8.1661	Prob > F
C. Total	48	424.25664		0.0309 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-83.83776	57.35918	-1.46	0.1505
YEAR	0.0642463	0.028866	2.23	0.0309 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=APR



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -117.8638 + 0.0865618*YEAR

Summary of Fit

RSquare	0.232168
RSquare Adj	0.215831
Root Mean Square Error	2.273118
Mean of Response	54.13442
Observations (or Sum Wgts)	49

Analysis of Variance

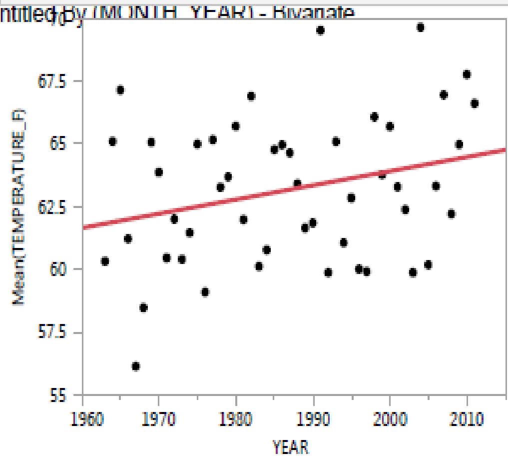
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	73.43079	73.4308	14.2113
Error	47	242.85217	5.1671	Prob > F
C. Total	48	316.28296		0.0005 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-117.8638	45.62658	-2.58	0.0130 *
YEAR	0.0865618	0.022962	3.77	0.0005 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=MAY

Untitled By (MONTH_YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -49.00032 + 0.0564572*YEAR

Summary of Fit

RSquare	0.07647
RSquare Adj	0.05682
Root Mean Square Error	2.833109
Mean of Response	63.18015
Observations (or Sum Wgts)	49

Analysis of Variance

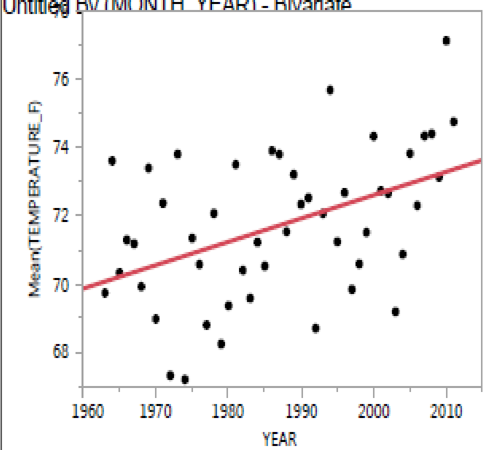
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	31.23668	31.2367	3.8917
Error	47	377.24591	8.0265	Prob > F
C. Total	48	408.48259		0.0544

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-49.00032	56.86685	-0.86	0.3932
YEAR	0.0564572	0.028619	1.97	0.0544

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=JUN

Untitled By (MONTH_YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -63.87755 + 0.068241*YEAR

Summary of Fit

RSquare	0.203693
RSquare Adj	0.18675
Root Mean Square Error	1.948328
Mean of Response	71.71741
Observations (or Sum Wgts)	49

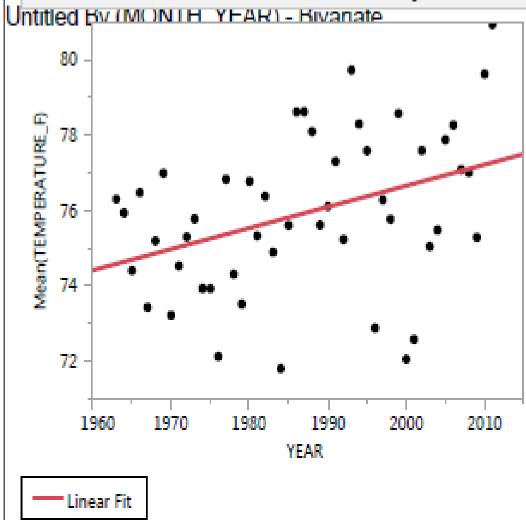
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	45.63704	45.6370	12.0225
Error	47	178.41111	3.7960	Prob > F
C. Total	48	224.04815		0.0011 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-63.87755	39.1073	-1.63	0.1091
YEAR	0.068241	0.019681	3.47	0.0011 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=JUL



Linear Fit

Mean(TEMPERATURE_F) = -35.3627 + 0.056011*YEAR

Summary of Fit

RSquare	0.144297
RSquare Adj	0.12609
Root Mean Square Error	1.969567
Mean of Response	75.9311
Observations (or Sum Wgts)	49

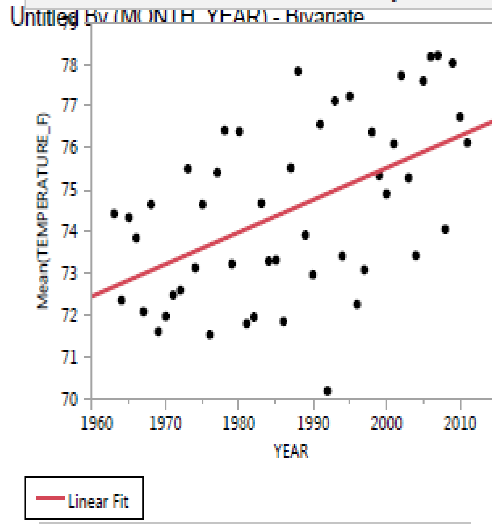
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	30.74485	30.7448	7.9256
Error	47	182.32221	3.8792	Prob > F
C. Total	48	213.06706		0.0071 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-35.3627	39.53363	-0.89	0.3756
YEAR	0.056011	0.019896	2.82	0.0071 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=AUG



Linear Fit

Mean(TEMPERATURE_F) = -77.83216 + 0.0766774*YEAR

Summary of Fit

RSquare	0.268
RSquare Adj	0.252426
Root Mean Square Error	1.829867
Mean of Response	74.52593
Observations (or Sum Wgts)	49

Analysis of Variance

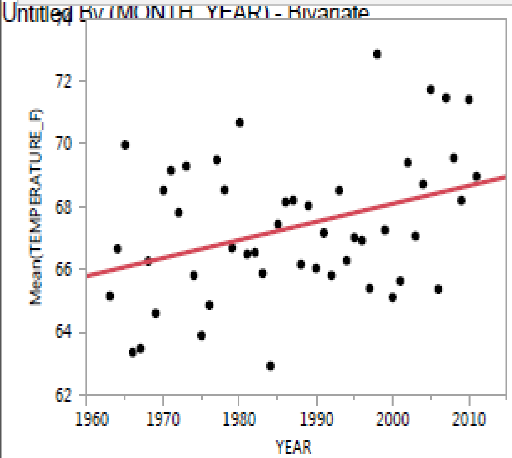
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	57.61842	57.6184	17.2077
Error	47	157.37540	3.3484	Prob > F
C. Total	48	214.99382		0.0001 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-77.83216	36.72953	-2.12	0.0394 *
YEAR	0.0766774	0.018484	4.15	0.0001 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=SEP

Untitled By (MONTH YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -46.99631 + 0.0575444*YEAR

Summary of Fit

RSquare	0.132432
RSquare Adj	0.113973
Root Mean Square Error	2.126779
Mean of Response	67.3444
Observations (or Sum Wgts)	49

Analysis of Variance

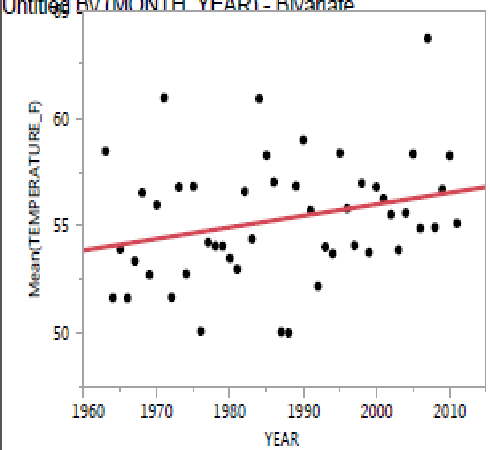
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	32.45130	32.4513	7.1744
Error	47	212.58990	4.5232	Prob > F
C. Total	48	245.04120		0.0102 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-46.99631	42.68922	-1.10	0.2765
YEAR	0.0575444	0.021484	2.68	0.0102 *

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=OCT

Untitled By (MONTH YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -51.07798 + 0.0535343*YEAR

Summary of Fit

RSquare	0.071848
RSquare Adj	0.0521
Root Mean Square Error	2.778424
Mean of Response	55.29467
Observations (or Sum Wgts)	49

Analysis of Variance

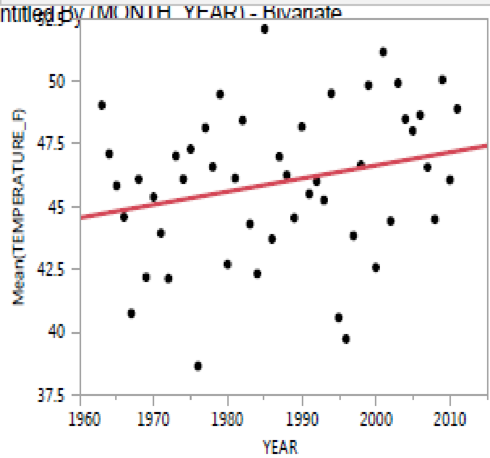
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	28.08602	28.0860	3.6383
Error	47	362.82295	7.7196	Prob > F
C. Total	48	390.90898		0.0626

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-51.07798	55.76918	-0.92	0.3644
YEAR	0.0535343	0.028066	1.91	0.0626

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=NOV

Untitled By (MONTH_YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -57.75592 + 0.0522*YEAR

Summary of Fit

RSquare	0.061554
RSquare Adj	0.041587
Root Mean Square Error	2.94315
Mean of Response	45.96543
Observations (or Sum Wgts)	49

Analysis of Variance

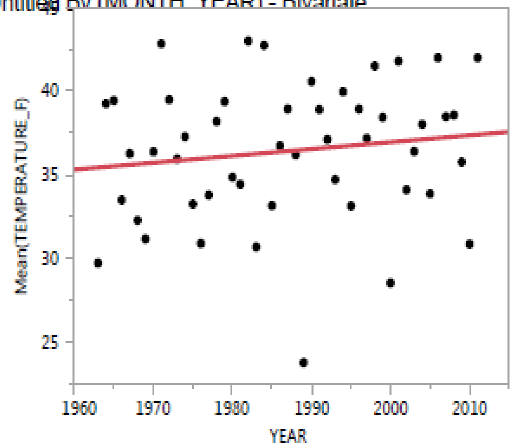
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	26.70340	26.7034	3.0828
Error	47	407.12021	8.6621	Prob > F
C. Total	48	433.82361		0.0856

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-57.75592	59.07561	-0.98	0.3332
YEAR	0.0522	0.02973	1.76	0.0856

Bivariate Fit of Mean(TEMPERATURE_F) By YEAR MONTH=DEC

Untitled By (MONTH_YEAR) - Bivariate



Linear Fit

Linear Fit

Mean(TEMPERATURE_F) = -45.03388 + 0.0409908*YEAR

Summary of Fit

RSquare	0.020071
RSquare Adj	-0.00078
Root Mean Square Error	4.135849
Mean of Response	36.4148
Observations (or Sum Wgts)	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	16.46639	16.4664	0.9627
Error	47	803.94660	17.1052	Prob > F
C. Total	48	820.41299		0.3315

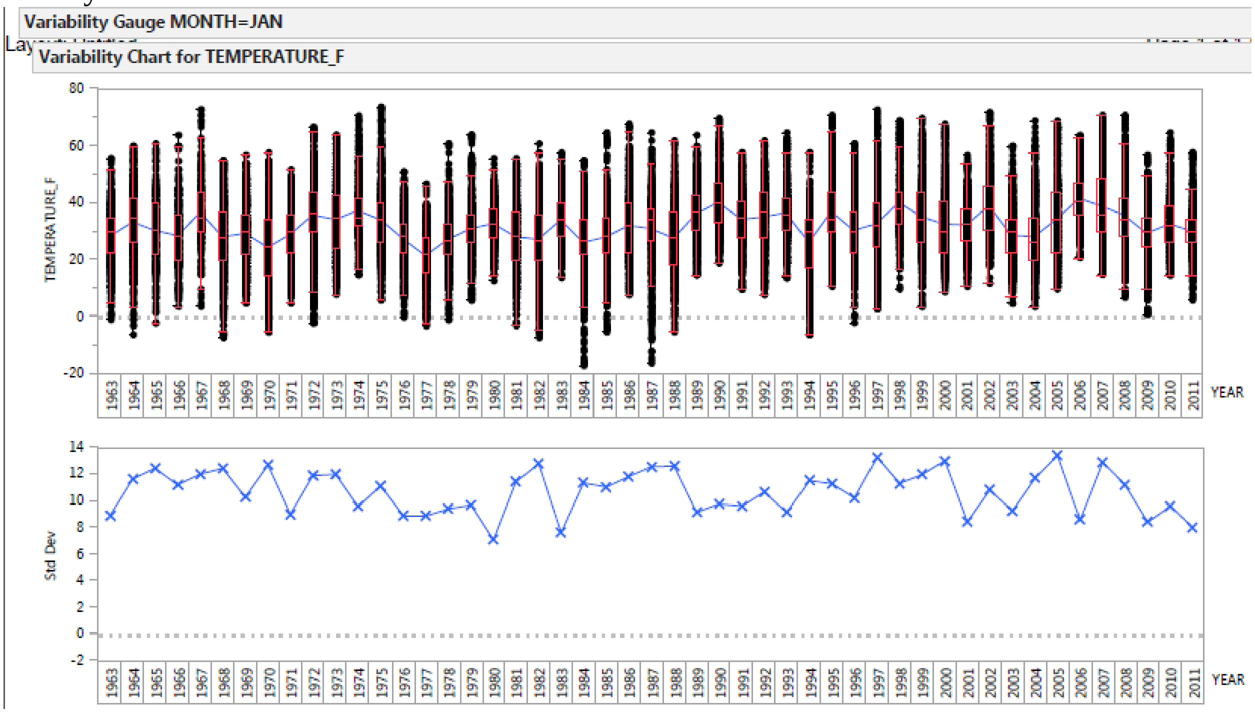
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-45.03388	83.01575	-0.54	0.5901
YEAR	0.0409908	0.041778	0.98	0.3315

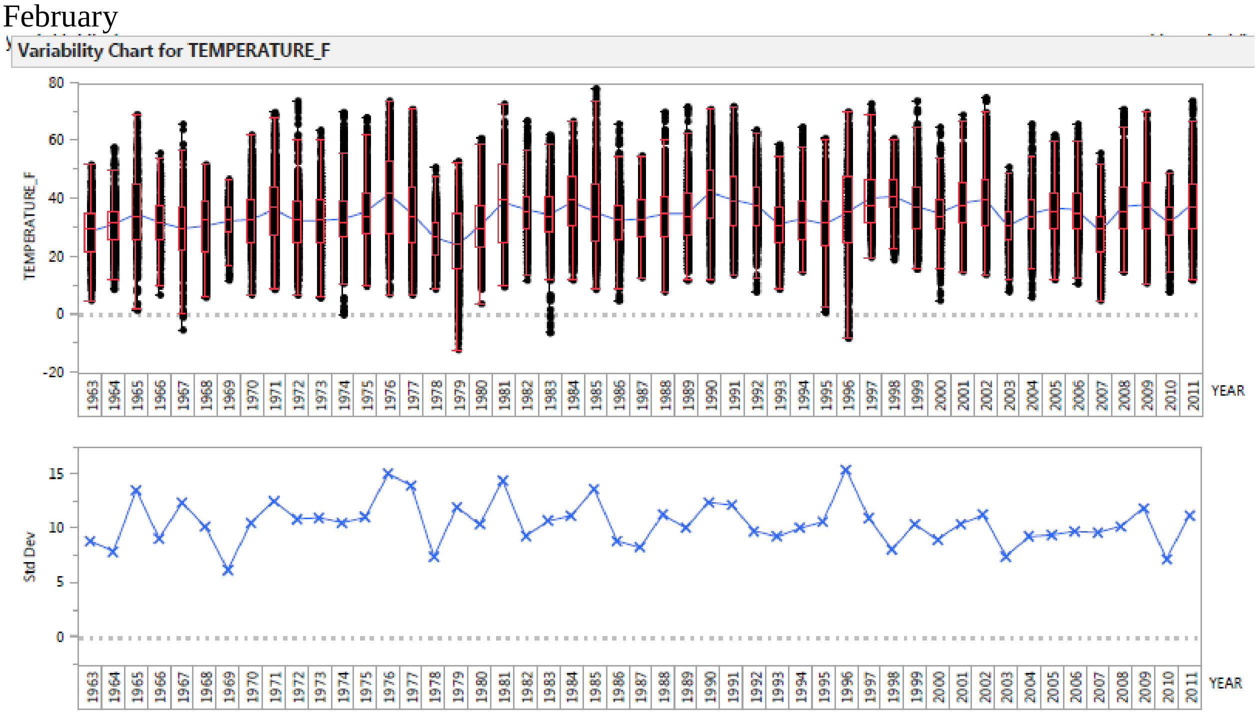
Appendix C: Historical Temperature Variance

Box plots showing weather variability per month since 1963.

January

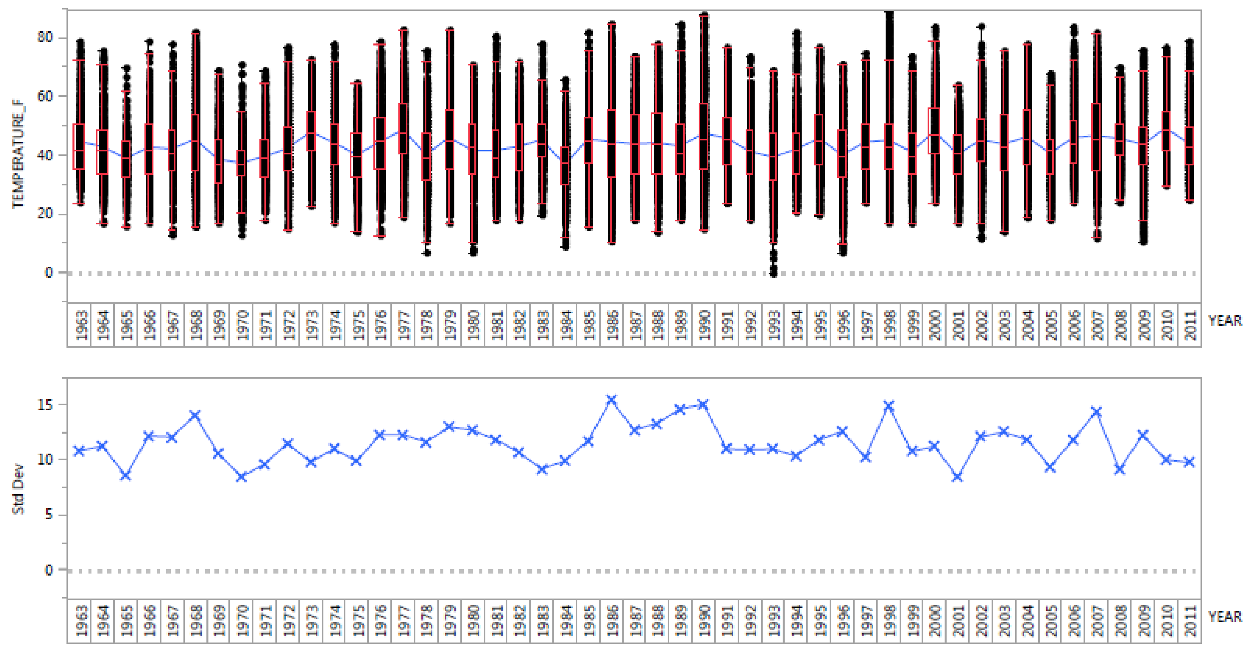


February



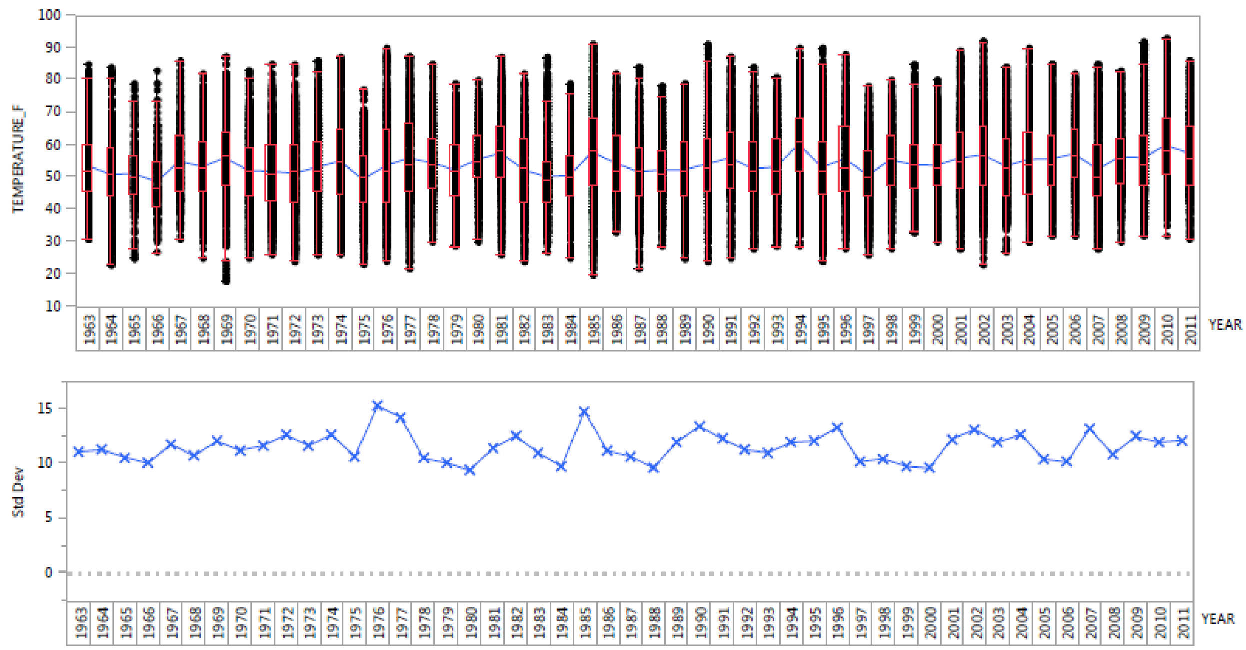
March

Variability Chart for TEMPERATURE_F



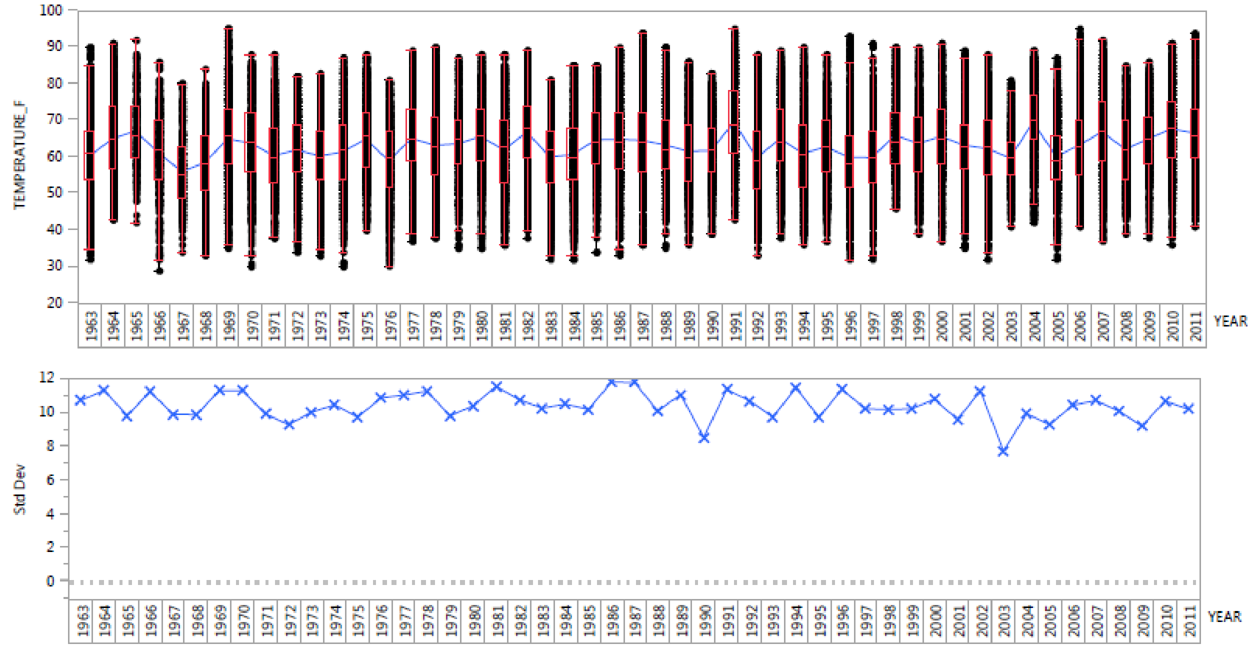
April

Variability Chart for TEMPERATURE_F



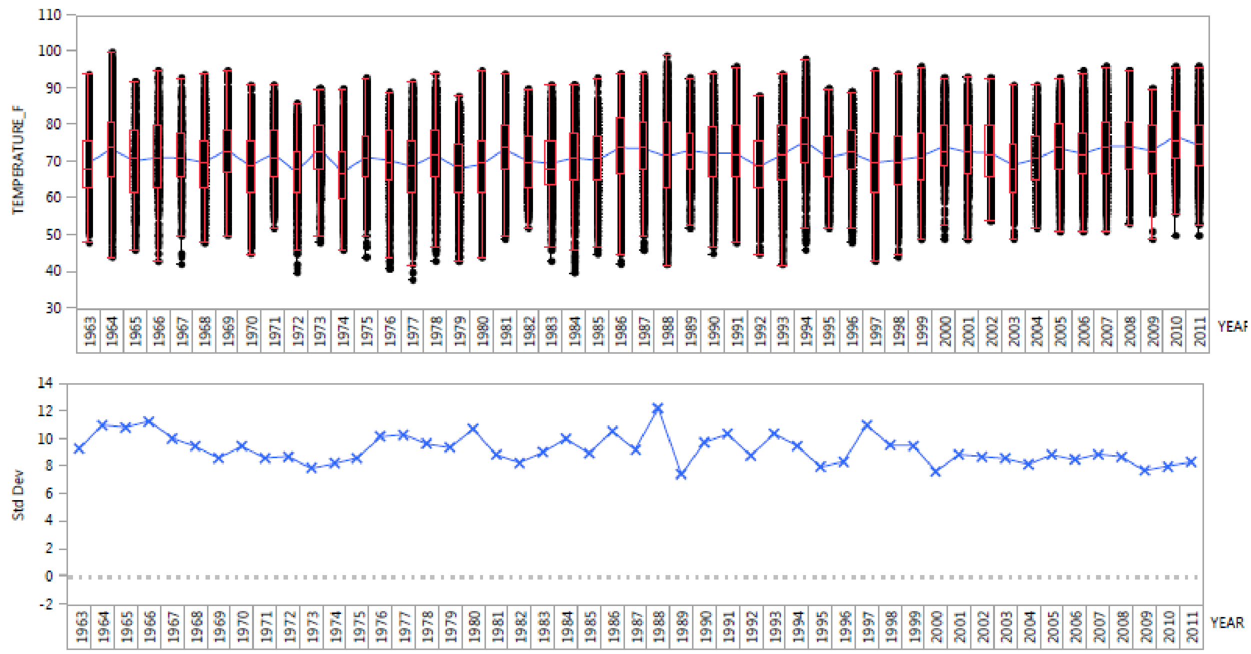
May

Variability Chart for TEMPERATURE_F



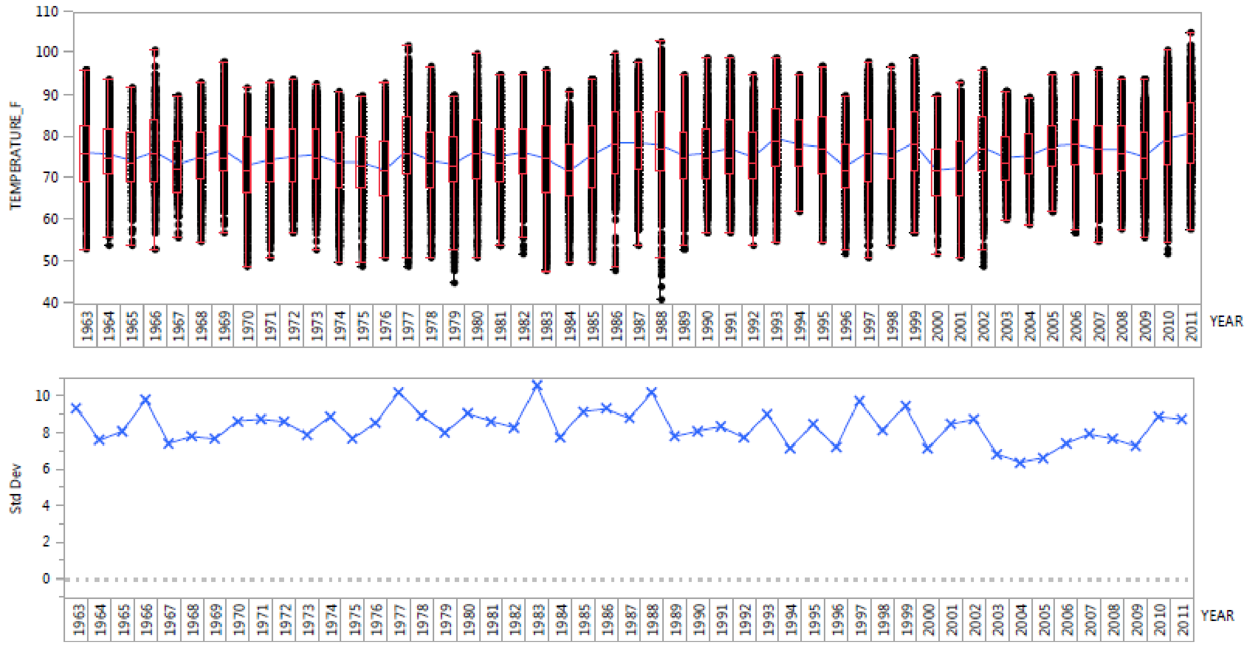
June

Variability Chart for TEMPERATURE_F



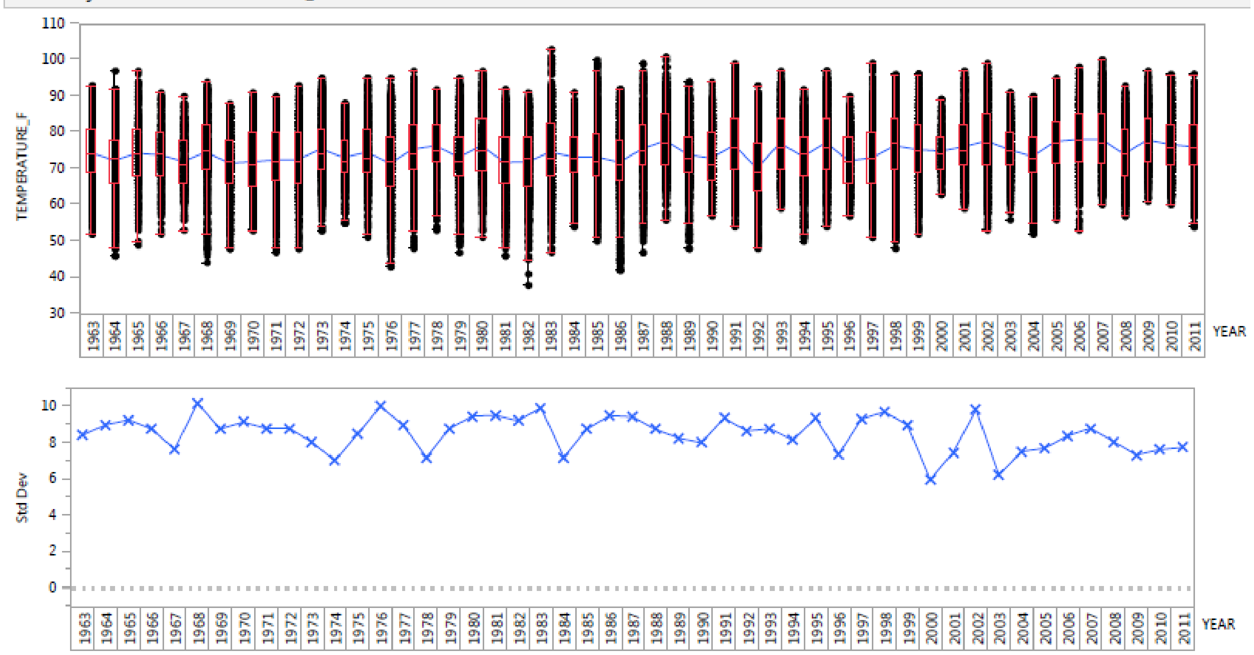
July

Variability Chart for TEMPERATURE_F

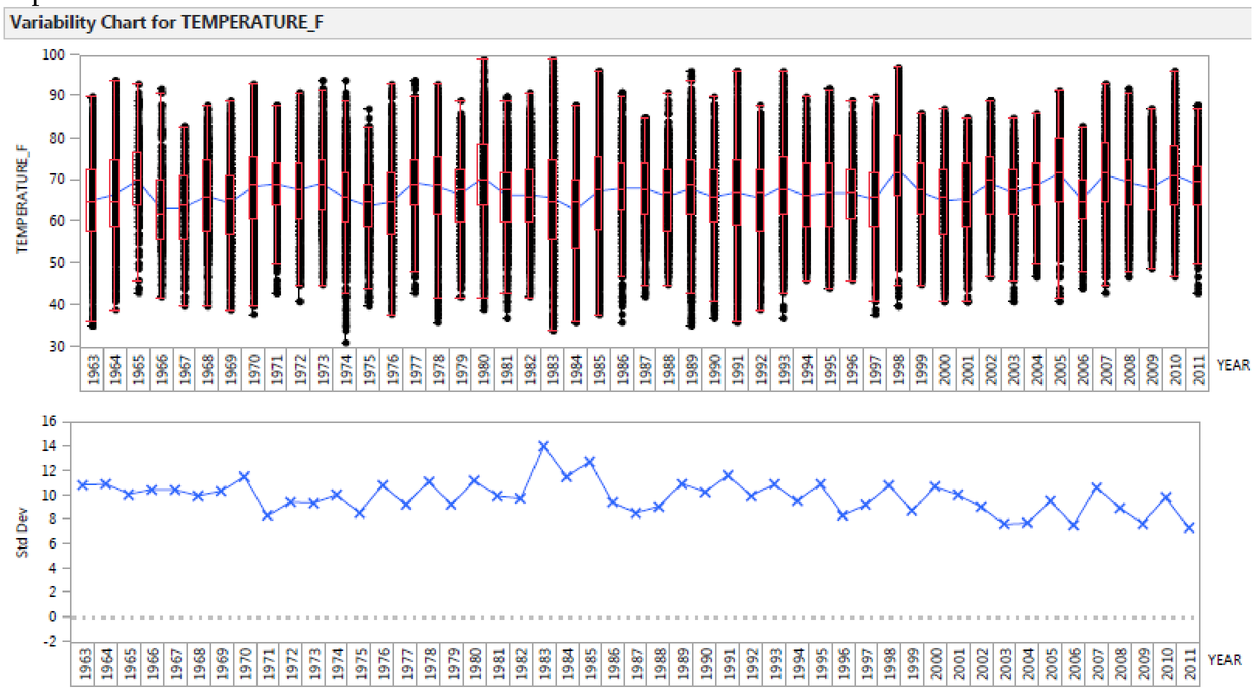


August

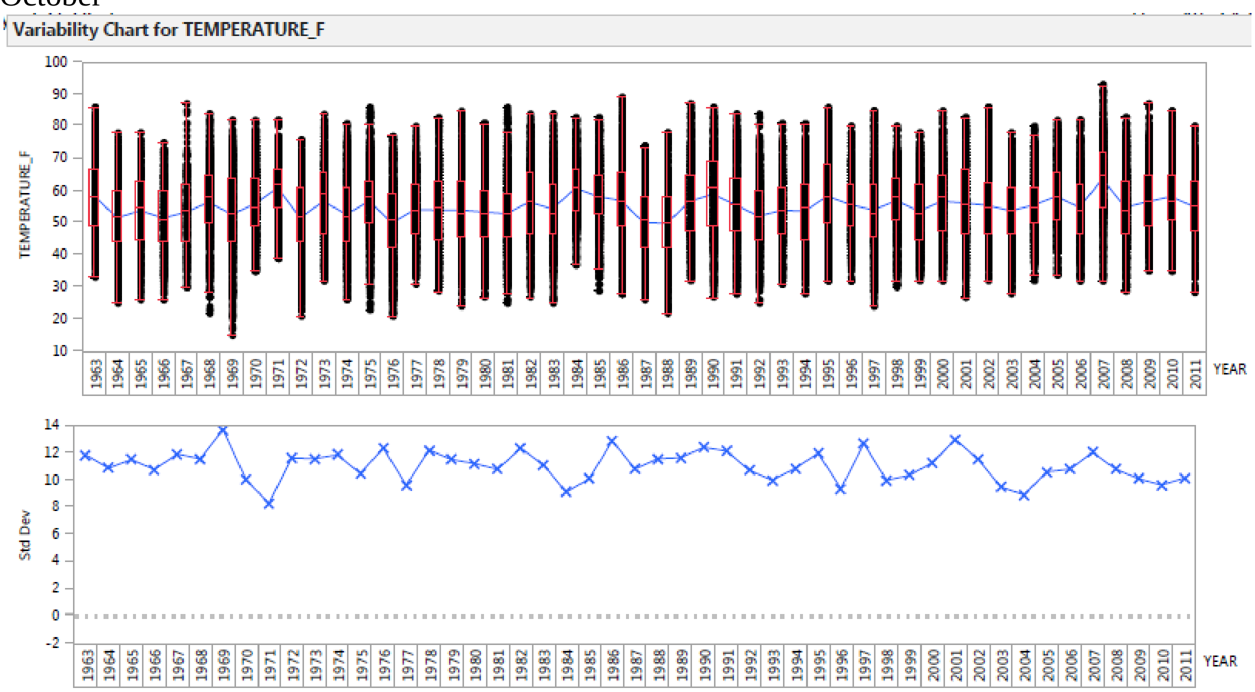
Variability Chart for TEMPERATURE_F



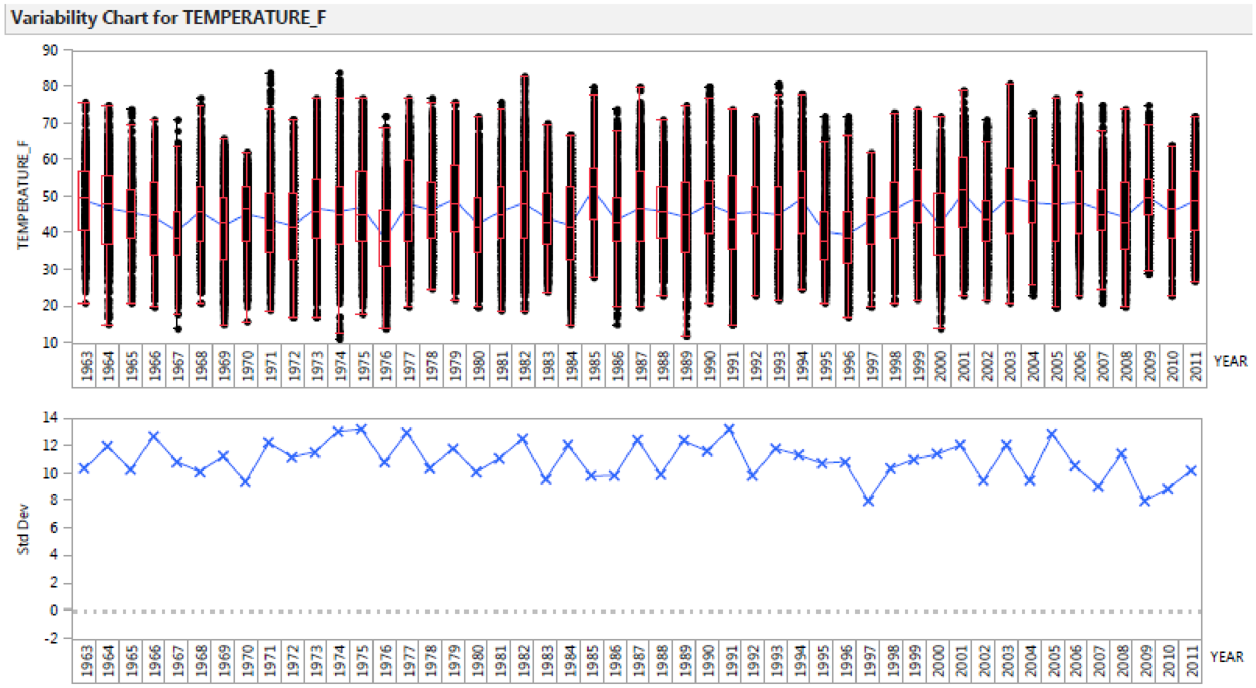
September



October



November



December