Utilizing Various Statistical Models on Time Series Temperature Data

Mason Chen
Stanford Online High School
San Jose, CA 95131
mason05@ohs.stanford.edu

Charles Chen
STEAMS
San Jose, CA 95131
charles.chen.training@gmail.com

Abstract

This STEM paper will study the Time Series data file Raleigh Temps.jmp in the JMP sample Library, which contains maximum monthly temperatures measured in degrees Fahrenheit from 1980 to 1990. The objective of this paper is to forecast the Raleigh Temperature for 1990 to 2010. Among the four STEM components, the Science studied is Climatology or the study of the Earth's Weather patterns; Technology is used to predict the Raleigh Max Monthly Temperature for the next twenty years; Engineering focuses on comparing Different Statistical Models on the Time Series Data, and mathematical tools like Statistics are applied. Most traditional and modern Data mining platforms can visualize the data distribution, Normal Quantile, Normal Mixture, Outliers very well, detect process capability and process stability, build Multiple Regression Model and identify month as the main factor, detect Clusters or Principal Components through Eigen Analysis, and use Neural Network or Partition Trees to build the Transfer Function Profiler. The major findings among these non-Time Series Platforms are that there is month to month cyclic behavior within each year but no year-to-year trend pattern is detected. However, there are several limitations among these non-time series platforms. It cannot decompose the seasonal component (cyclic month-month) from the trend component (year to year), cannot determine the relative strength of the "seasonal" and "trend" components, cannot determine the optimal smoothing setting if the curve is highly modulated, and cannot forecast or predict future points. JMP Time Series and Forecast platform were further used on the same time series data file. The Time Series decomposition statistics were utilized to separate the seasonal component from the trend component, and the smoothing technique cleaned the Forecasting Error. Seasonal lag was detected and confirmed by the Autocorrelation Function (ACF) plot and Variogram plot. The Time Series Forecast Platform can help predict Raleigh Temperature for 1990 to 2010 based on the 1980 to 1990 data. Authors are also continuing this Time Series STEM project on the following areas – Decomposition and Smoothing Statistics, Non-Seasonal and Seasonal ARIMA Models, and Forecasting and Prediction Interval Statistics. Time Series Analysis is not just popularly used in Finance Forecasting but also powerful for predicting any future uncertainty from the Time Series data.

Keywords
Time Series, Forecast, Statistics, Climatology, Data Mining

1. Introduction

1.1 Climatology
Climate science investigates the structure and dynamics of earth's climate system. It seeks to understand how global, regional and local climates are maintained as well as the processes by which they change over time Arnold (2011), Bindoff (2013). Climatology employs observations and theory from a variety of domains, including meteorology, oceanography, physics, chemistry and more. These resources also inform the development of computer models of the climate system, which are a mainstay of climate research today. Climatologists seek to understand three main aspects of climate. The first aspect is the weather patterns that govern normal conditions in different regions throughout the

world. Secondly, climate scientists try to understand the relationship between different aspects of weather such as temperature and sunlight. The third aspect of climate that climatologists investigate is the way that weather changes over time. This paper will also address how to select the most appropriate statistical model to analyze the historic Raleigh temperature data to forecast the monthly maximum Raleigh temperature for the next 20 years.

1.2 Time Series and Forecast

Time Series Analysis and Forecasting modeling were utilized on the Raleigh Temperature data. Climatology research has used Time Series and Forecasting model such as ARIMA to forecast the weather temperature to study the global warming trend Baillie (2002). In this paper, we will compare several non-Time Series Statistical modeling to Time Series modeling on the Raleigh Temperature data.

2. Data Collection and Sampling Plan

2.1 Raleigh Data and Sampling Plan

The data source for this paper is from the JMP Sample Library, Raleigh Temps.jmp which has collected the monthly maximum temperature data from 1980-1990 as partially shown in Figure 1. The intent of this Raleigh project is to study the global warming trend to forecast how much temperatures will have risen in 1990-2010. Stratified sampling is used by splitting the data to each month and taking the maximum temperature as the sampled data.

	Month	Year	Temperature	Month/Year
1	1	1980	48.92	01/1980
2	2	1980	48.02	02/1980
3	3	1980	56.84	03/1980
4	4	1980	75.74	04/1980
5	5	1980	81.86	05/1980
6	6	1980	86.54	06/1980
7	7	1980	89.24	07/1980
8	8	1980	90.86	08/1980
9	9	1980	84.2	09/1980
10	10	1980	69.62	10/1980
11	11	1980	60.98	11/1980
12	12	1980	52.16	12/1980
13	1	1981	45.86	01/1981
14	2	1981	56.84	02/1981
15	3	1981	59.18	03/1981
16	4	1981	74.48	04/1981
17	5	1981	75.74	05/1981
18	6	1981	88.7	06/1981

Figure 1. Raleigh Temp.jmp data file

3. Non-Time Series Statistical Analysis

Conduct several JMP 16 Statistical Analytical Platforms on the "Raleigh Temp" time series data as following: (1) Visualization and Normality Analysis, (2) Simple and Multiple Linear Regression, (3) Univariate and Multivariate Statistical Process Control, (4) Process Capability, (5) Multivariate Statistics and Principal Component Analysis, and (6) Data Mining Analysis. The objectives are to explore the capability and limitations of conducting these non-Time Series Analytical platforms on the Time Series data.

3.1 Visualization and Normality Analysis

To visualize data distribution and conduct Normality Test, JMP Distribution platform was used. In Fig.2, the histogram plot has shown a clear Normal Mixture 2 Goodness Fit (Bimodal Distribution). This is an interesting finding on why among 12 months' data, there are two modes observed (one mode is peaked at 80-85F, one mode is 60-65F). There is a weak mode separation around 70-75F. The Box Plot has shown no Outliers, Mean ~ Median, and near Symmetric. The Normal Quantile Plot (Normality Test) showed a similar information on the data distribution and normality test violation.

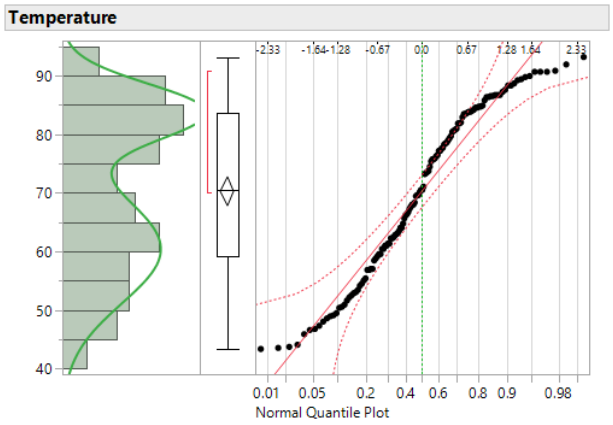

Figure 2. JMP Distribution platform analysis

To further validate any outlier risk, JMP Outlier Explorer platform was conducted as shown in Fig.3. Since the temperature data distribution is not Normal (Bimodal distribution), we should use the non-parametric outlier detection method. Quantile Range Outlier method [Benford , 1938; Penny, 1996] was used and the set the outlier criteria: (1) Tail Quantile 0.25 to specify the Inter-quantile Range = Q3 (83.66) - Q1 (59.18), and (2) Q multiple factor = 1.5 to set the outlier detection threshold is 1.5*IQR distance from Q1 or Q3. The lower outlier detection threshold is 22.46 and the upper outlier threshold is 120.38. There is no outlier detected in 1980-1990. Therefore, we could keep all the data in the following data analysis without worrying the outlier distortion risk. Since no significant outlier and skewness risk, we will use the Robust Statistics [Huber,2009] parametric analytics across several JMP platforms such as Linear Regression [Rousseeuw, 1987], Process Capability and SPC Control Chart.

Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile 0.25

Q 1.5

☐ Restrict search to integers

☐ Show only columns with outliers

Select columns and choose an action.

Identify Outliers in Table

Clear Outliers in Table

Column	25% Quantile	75% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
Temperature	59.18	83.66	22.46	120.38	0	

Figure 3. Outlier Range Outliers Analysis.

3.2 Simple and Multiple Linear Regression

To further analyze the time factor (Month) on the temperature impact, both the simple linear regression (month/year as X) and multiple linear regression (month, year as two Xs) were conducted Belsley (1980), Hocking (1985). Simple Linear Regression was conducted by JMP Graph Builder Platform as shown in Fig.4. Both Slope and R-Square are near zero (random noise pattern). Simple Linear Regression model has confirmed little long term year-year trend. Temperature at Raleigh has not been increased significantly from 1980-1990. Though, the regression model could not quantify the month-month cyclic pattern.

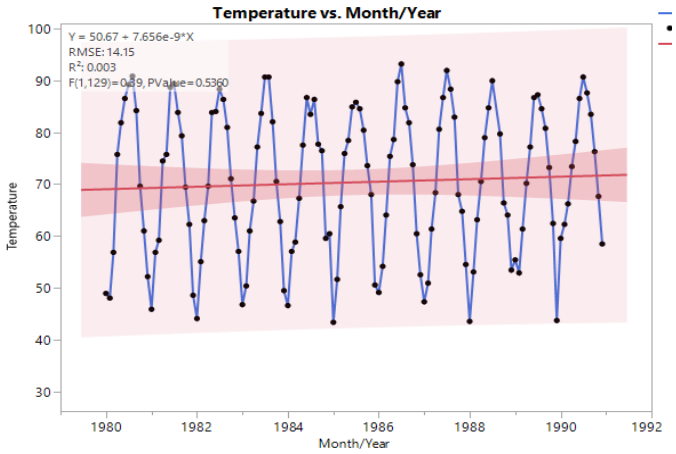

Figure 4. Simple Regression Analysis.

Multiple Linear Regression was conducted through JMP Fit Model platform by splitting the "Month/Year" factor to "Month" and "Year" factors as shown in Fig.5 "Actual by Predicted Plot". A decent goodness of fit as R-Square is 0.91. Though, there are some clustering behavior across the temperature range. We will further investigate this cluster pattern in section 3.6.

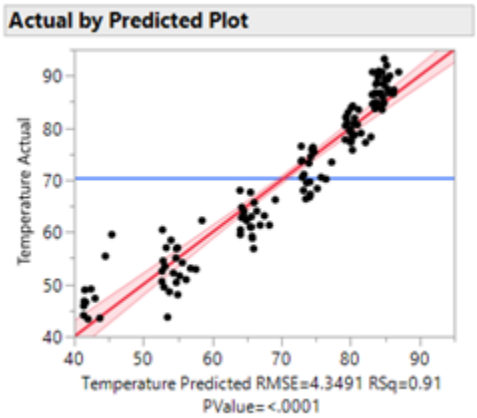

Figure 5. Multiple Linear Regression Analysis.

In previous Simple Linear Regression analysis, the model is near random and no slope on the factor "Month/Year". In the Multiple Regression Profiler analysis Sobol (1993) as shown in Fig.6, there is a clear quadratic term of "month" factor but little on the "Year" factor. Also, no significant interaction effect between "month" and "year". The quadratic "month" term can be explained by the month-month cyclic term within each year. No interaction term can be interpreted that the cyclic monthly temperature pattern has been repeated well each year during 1980-1990.

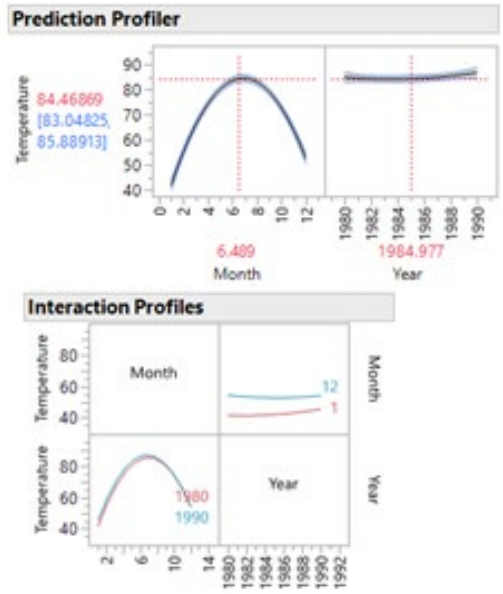

Figure 6. Multiple Regression Profiler Analysis

3.3 Univariate and Multivariate Statistical Process Control

To further investigate the temperature data in the time domain, univariate and multivariate Statistical Process Control (SPC) chart AIAG (2005), Nelson (1984 & 1985), Wheeler (2004) were used to detect the temperature data stability. Xbar-Control chart, as shown in Fig.7, has shown that the temperature data is not stable within 12 months. The test has detected two alarms: #1 is out of control limits, #6 is Freak II (4 out of 5 beyond 1 sigma from the center Green line. However, current Xbar chart could not detect the cyclic oscillation mode still.

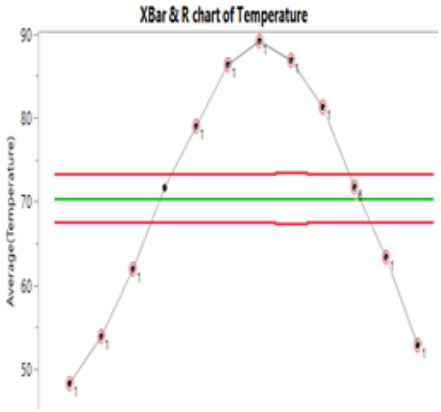

Fig.7 Xbar Control Chart

To further understand the dependency of the "Temp" response and "Month, Year" factors, Model Driven Multivariate T-Square Control Chart Mason (2002), Tracy (1992), Kourti (1996), Nomikos (1995] was shown in Fig.8. From the left T-Square chart, the biggest outlier was happened on 12/1989. The right T-Square Contribution Proportion Plot has shown the variance component analysis among three components (Temp, Month, Year). As expected, temperature has the biggest variance component. More interestingly is that "Month" factor has more contribution than the "Year" factor.

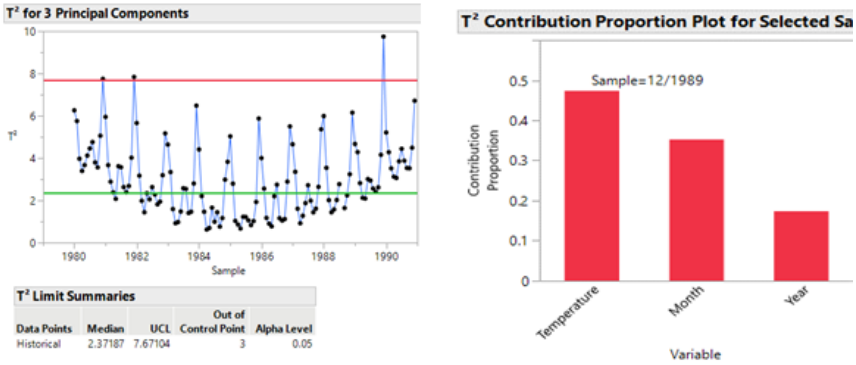

Figure 8. Model Driven Multivariate Control Chart Analysis

3.4 Process Capability Analysis

To further study the temperature data range, by setting the temperature spec from 45F to 85F. The Ppk Process Index Bissel (1990) has been shown in Fig.9 JMP Goal Plot. The vertical Y axis is presenting the process dispersion and the horizontal X axis is for meeting the target. Any point in Green Zone means process is capable; process is marginally capable in yellow zone; and process is not capable in the red zone. Fig.9 has shown Fall quarter temperature is the most capable on dispersion and target, and Spring season is marginal (dispersion marginal but meeting the target). Both Winter and Summer are not capable (off the target). Winter has the worst scenario not just off the target but also wider dispersion.

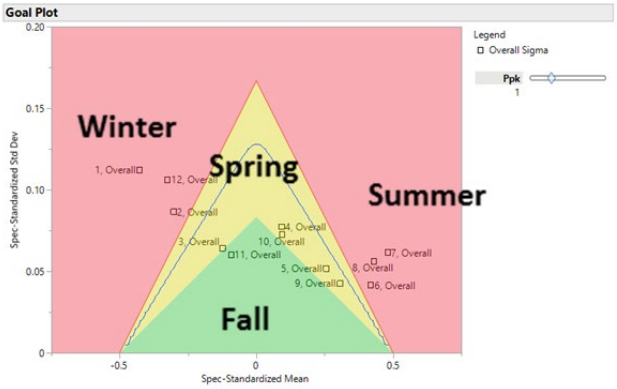

Fig. 9 Goal Plot Analysis

3.5 Multivariate Correlation and Principal Component Analysis

For the next 3.5 and 3.6 sections, we would conduct the modern Data Mining analysis. Section 3.5 would present Multivariate Correlation [Mardia,1979; Anderson, 1958], and Principal Component Analysis. Fig. 10 has shown the scatterplot and multivariate correlation color map. “Temp” response and “Month” factor have been observed a high correlation and quadratic cyclic pattern while the “Year” factor has no involvement at all. The first Multivariate Correlation analysis did not provide us any insight.

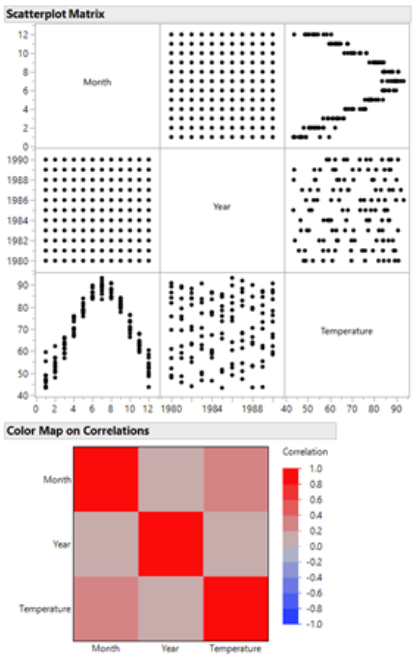

Fig.10 Multivariate Correlation Analysis

Principal Components Eigen Analysis Jackson (2003), Nelson (1996), Candes (2009) has been further conducted in Fig.11. The top two eigenvalues would accumulate 74% of the total variance. The middle Eigenvector table has shown that the first Principal Component was consisted mainly of the "Temp" and "Month" factor. The second Principal Component (orthogonal to the 1st Principal Component) was from "Year" Factor. This Eigen analysis has further confirmed that "Temp" cyclic response was due to the "Month" factor, not related to the "Year" factor. The bottom Biplot has displayed the "Temp" response and (month, year) factors on the Eigenvector 2D plot. "Temp" and "Month" variables are oriented to the first Principal Component X axis, and "Year" factor are close to the second Principal Component Y axis.

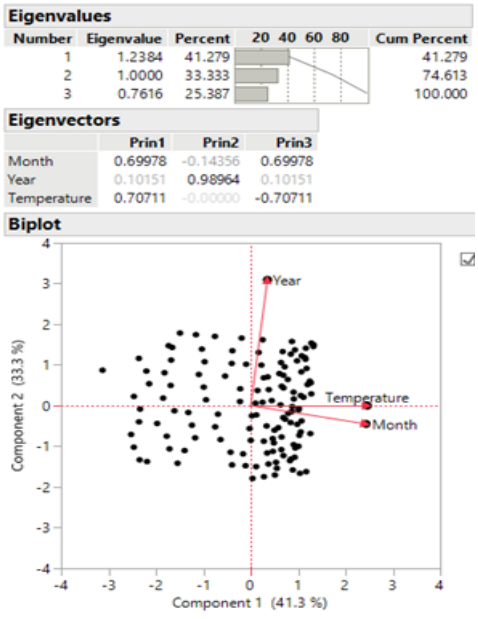

Figure. 11 Principal Component Analysis

3.6 Data Mining Analysis

In this section, we would conduct three modern Data Mining Analysis Hand (2001), Hastie (2009) and see what they can offer anything further. The first one is the Partition Tree (CART Classification) analysis as shown in Fig.12. The data set was split at nodes based on the classification algorithm (purity). Partition model has shown very decent RSquare at 0.958 after 23 splits. "Month" factor has over 99% contribution and "Year" factor less than 1%. The Partition Profiler has shown similar Quadratic pattern between the "Temp" response and the "Month" factor but at more discrete scale. The discrete scale is because the partition algorithm treated the "month" factor as "Ordinal" data not "Continuous" data.

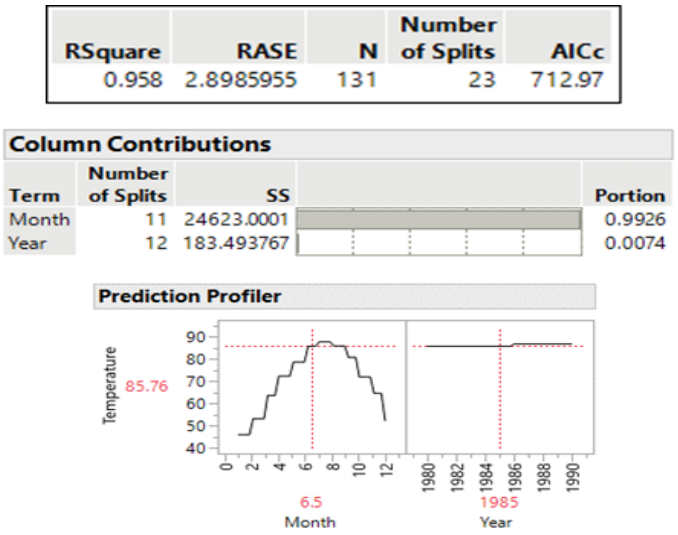

Figure.12 Partition Tree Analysis

The 2nd Data Mining platform used is Clustering Hartigan (1981), Milligan (1980). In previous 3.1 Histogram analysis, we observed Bimodal distribution and in 3.2 Multiple Regression analysis, we observed more than three clusters. In this Clustering analysis, we used the Normal Mixture Cluster algorithm as shown in Fig.13. Normal Mixture Clustering algorithm can separate clusters with certain overlapping than the other clustering algorithms. This time in Normal Mixture Clustering Analysis, we have observed 3 clusters. The Blue cluster is well separated from the other two clusters. The Red cluster has significant overlapping with the Green Cluster. The Principal Component 2D domain has shown the similar orientations among three variables as seen earlier. Further in the bottom parallel plot Inselberg (1985), Wegman (1990), we may visualize these three clusters more effectively. "Tear" factor has little contribution among three clusters. The first cluster is for the Jan.-April when temperature is lower. The second cluster is for the May-Sep. when temperature is high. The third cluster is for the Oct.-Dec. when temperature is low again. This cluster analysis has exposed one concern on whether "month" factor is Continuous, Ordinal and Nominal. We have coded "Month" factor from 1-12 for Jan.-Dec. Though, the month "Jan." is right after the month "Dec". But month "1" is not after month "12" since they do not know "Month" factor is repeat after 12 months. Either the "Continuous" or "Ordinal" data type can not handle this Seasonal Factor. This "Seasonal" factor can not be taken care in most Non-Time Series data analysis.

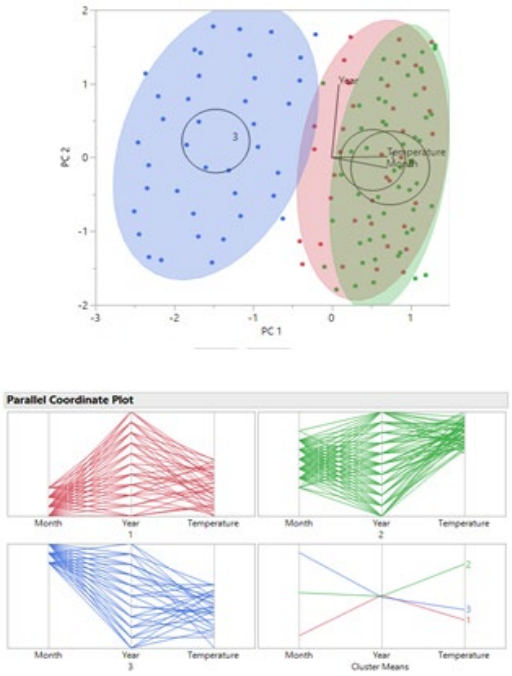

Fig.13 Normal Mixture Cluster Analysis

The third Data Mining analysis is "Neural Network" as shown in Fig.14. Modern Neural Network has adopted very powerful transformation through the activation function. The RSquare for both Training and Validation sets are over 95%. The Neural Network profiler has shown similar quadratic pattern on the "Month" factor and flat pattern on the "Year" factor. The bottom Neural diagram has shown the one middle layer perceptron transformation (Black Box) between the factor layer and the response layer.

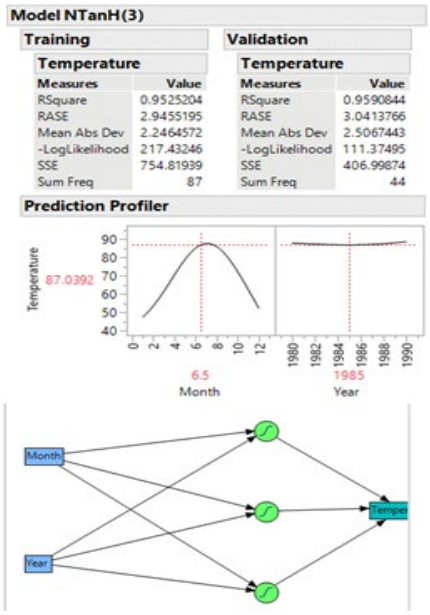

Figure 14. Neural Network Analysis

3.7 Non-Time Series Analytics Results

We can summarize the analytics results from 3.1 to 3.6 on their pros and cons Below:

What they can do:

- Can visualize data distribution, Normal Quantile, Normal Mixture, Outliers
- Can detect process capability and process stability very well
- Can build Multiple Regression Model and identify Month as the biggest Factor
- Can detect Clusters or Principal Components through Eigen Analysis
- Can use Neural Network or Partition Trees to build the Transfer Function Profiler

Major Findings: there is month-month cyclic behavior within each year, there is no year-year trend pattern detected.

What they can NOT do:

- Can not decompose the Seasonal component (cyclic month-month) to Trend Component (year-year)
- Can not handle any "Seasonal" factor like month as Continuous, Ordinal or Nominal data type
- Can not determine the relative strength of the "Seasonal" and "trend" components
- Can not determine the optimal smoothing setting if curve is highly modulated (White Noise)
- Can not Forecast or Predict the future points

4. Time Series Analysis

To address the 3.7 "Ca not do" concerns, this section will introduce the Time Series Analysis on the time series data set Raleigh Temp.jmp.

4.1 Introduce Time Series Decomposition

First, authors would introduce basic Time Series statistics: Decomposition. Time series data can be decomposed to several components such as "Trend", "Seasonal" and Cyclic", "Random" Box (2006), Hyndman (2018). Trend Component: a trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as "changing direction," when it might go from an increasing trend to a decreasing trend.

Seasonal: A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. Cyclic: A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to economic conditions, and are often related to the "business cycle." In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitudes of cycles tend to be more variable than the magnitudes of seasonal patterns. Fig.15 has shown several Time Series Decomposition patterns.

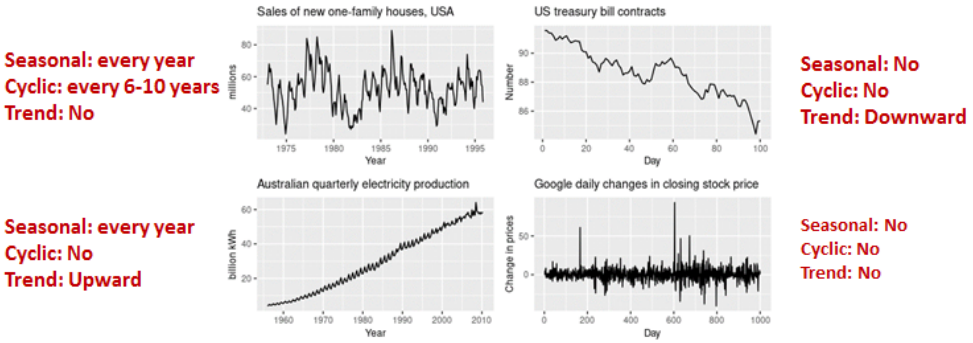

Figure 15. Time Series Decomposition Patterns

4.2 Time Series Decomposition Analysis

In Fig.16, Raleigh Temperature data has been decomposed through JMP Time Series platform. Autocorrelation values and ACF plot have been displayed on the left portion. ACF plot has detected seasonal lags at (6,12,18,24….). Though only (12,24…) are positive correlated lags and (6,18…) are negatively correlated lags. This could be interpreted that each the seasonal "month" factor is "12" months within each year for positive Autocorrelation. The negative group (6,18…) has indicated that the temperature level were opposite if exactly 6 months away such as Summer Vs, Winter. The Variogram plot on the right can provide more accurate information by calculating the standard deviation at every

lag level for the entire data distribution. Lag 12 has shown the lowest standard deviation in the Variogram plot which means the dispersion among the temperature data every 12 months showing the lowest variance level. Interestingly, the Variogram plot has shown the largest peak at lag 6 which is related to the negative autocorrelation behavior. The Autocorrelation and Variogram plots have clearly described the Seasonal pattern which was also shown in Sections 3.1-3.6. Though, this decomposition and seasonal lag information would be further utilized in Time Series Forecasting which is not available in previous Section 3.1-3.6 tools.

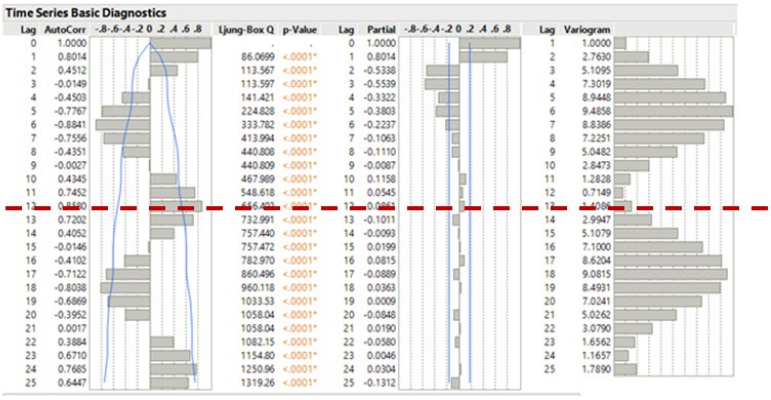

Fig.16 Time Series Decomposition Analysis.

Another powerful Time Series Decomposition is to remove the Seasonal Component (season-adjusted, Shiskin, (1967) as shown in Fig.17. Once seasonal lag =12 was identified in previous Autocorrelation analysis, another JMP function can remove the seasonal component by searching the optimal Cosine Transfer function. JMP platform would remove the seasonal component and conduct the Time Series Analysis in the bottom chart. The new transformed data distribution has little seasonal pattern, more like random (white noise) pattern. This Cosine transformation has further indicated that the seasonal lag is at 12 only. For our Raleigh case study, 12 months seasonal lag is very obvious. For other more complicated data, using both AutoCorrelation/Variogram and Remove Seasonal Component methods may be necessary to validate the seasonal lag which is critical for the Time Series Forecasting part in next Section 4.3. Sometimes, the data may have double seasonal components and may need to use "Twice- Remove Seasonal Component" method (Seasonal ARIMA) which is not covered in this paper.

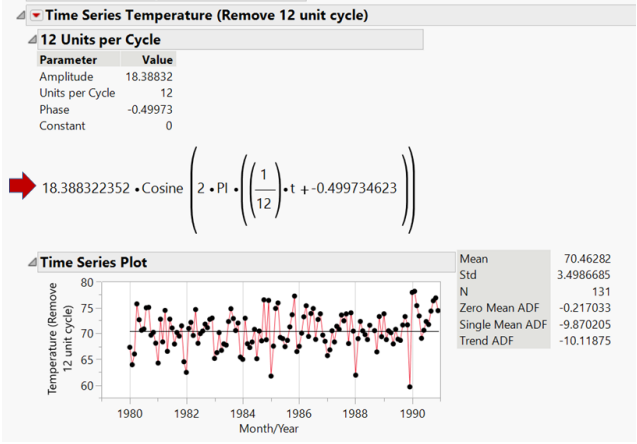

Figure.17 Remove Seasonal Component Analysis

4.3 Time Series Forecast

In previous 3.1-3.6 Non-Time Series Analysis, there is little forecasting power to predict the future points. To predict the future points (point estimation of the mean and prediction interval of the error), we need to utilize the previous decomposition algorithm. The forecasting formula should include the long term trend component, carry the seasonal component momentum, and the uncertain error level. Time series analysis has used the decomposition technique to

quantify the seasonal and trend components. Use the smoothing technique to smooth out the noise to enhance the signal of "Seasonal" and "Trend" components. Time series has also used the more adv. smoothing techniques such as "moving average", "exponential smoothing" and "state space smoothing" Hyndman (2008) to estimate the error term for forecasting purpose. These smoothing techniques would not be addressed here (future work). In Fig.18, JMP Time Series Forecast platform has shown the forecasting of the Raleigh Temperature for 1990-2010. The yearly mean is almost the same since no long term trend component detected in time series analysis. Seasonal lag=12 has been assigned in Forecasting analysis. The forecasting analysis has duplicated the seasonal pattern. The Model summary has provided the goodness of fit criteria like Likelihood, AIC, BIC Burnham (2004 & 2011). The 95% prediction interval was wider for the time period far away from the last point in Dec. 1990. This is understood that it was more uncertain to predict any future points far away from today like Stock market. One thing noticed is that the prediction interval is not symmetric (wider in the higher temperature range). Authors have no right answer for this observation. This paper used the maximum temperature to report the raw data. There may be some intrinsic reason why upper prediction interval has observed spikes and wider range than the lower prediction interval. Also, the upper prediction interval may be , meaningless since the interval upper limit is beyond 120F most years after 1990. Instead of following up this abnormal upper prediction interval observation, authors may suggest finding more powerful Time Series Model like Seasonal ARIMA.

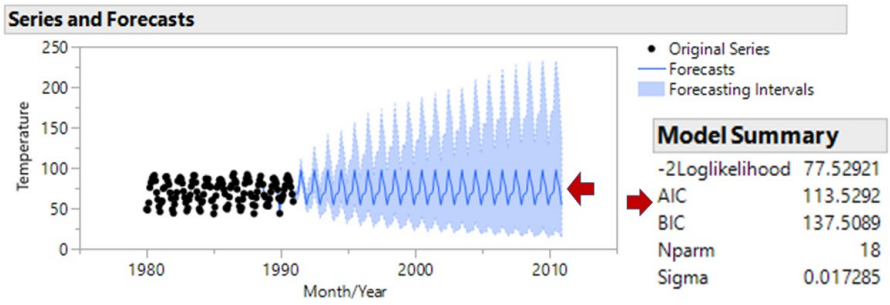

Figure.18 Time Series Forecasting Analysis

5 Conclusions

The STEM approach is adopted on analyzing the Time Series data file regarding the Raleigh Temperature records from 1980 to 1990. Neither Traditional Engineering and Statistics platforms nor Modern Data Mining platforms could not decompose the Time Series Components. Time Series Analysis can decompose the Seasonal and Trend Components, smooth out the Error Component for enhancing Forecasting capability, and help predict Raleigh Temperature for 1990 to 2010 based on the 1980 to 1990 data.

Future Work

Authors are continuing current Raleigh Temperature project to learn more about Advanced Time Series Techniques such as Decomposition and Smoothing Statistics, Non-Seasonal and Seasonal ARIMA Models, Forecasting and Prediction Interval Statistics, and more.

Acknowledgements

The authors would like to thank JMP Advisor Patrick Giuliano and IEOM STEM Co-Chairs Dr. Ali and Dr. Reimer.

References

AIAG (2005). *Statistical Process Control.* 2nd ed. Troy, MI: Automotive Industry Action Group.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.

Arnold, Denis G. (ed.), 2011, *The Ethics of Global Climate Change*, New York: Cambridge University Press. doi:10.1017/CBO9780511732294

Baillie, RT, Chung, SK, 2002, "Modeling and forecasting from trend-stationary long memory models with applications to climatology", *International Journal of Forecasting*. Volume 18, Issue 2, Pages 215-226

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Benford, F. (1938). "The law of anomalous numbers." *Proceedings of the American philosophical society*, 551–572.
Bindoff, Nathaniel L., Peter A. Stott, et al., 2013, "Detection and Attribution of Climate Change: from Global to Regional", in Stocker et al. 2013: 867–952
Bissell, A. F. (1990). "How Reliable is Your Capability Index?" *Applied Statistics* 30:331–340.
Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
Burnham, K. P., and Anderson, D. R. (2004). "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods and Research* 33:261–304.
Burnham, K. P., Andersen, D. R., and Huyvaert, K. P. (2011). "AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons." *Behavioral Ecology and Sociobiology* 65:23–35.
Candes, E. J., Li, X., Ma, Y., and Wright, J. (2009). "Robust Principal Component Analysis?" *Journal of the ACM*, 58:1–37.
Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
Hartigan, J. A. (1981). "Consistency of Single Linkage for High–Density Clusters." *Journal of the American Statistical Association* 76:388–394.
Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.
Hocking, R. R. (1985). *The Analysis of Linear Models*. Monterey, CA: Brooks/Cole.
Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. New York: John Wiley & Sons.
Hyndman, R.J., Athanasopoulos, G. 2018, *Forecasting: Principles and Practice*, 2nd
Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer-Verlag.
Inselberg, A. (1985). "The Plane with Parallel Coordinates." *Visual Computing* 1: 69–91.
Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley & Sons
Kourti, T., and MacGregor, J. F. (1996). "Multivariate SPC Methods for Process and Product Monitoring." *Journal of Quality Technology* 28:409–428.
Mardia, K. V., Kent, J. T., and Bibby J. M. (1979). *Multivariate Analysis*. New York: Academic Press.
Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications.* Philadelphia: SIAM.
Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325–342.
Nelson, L. (1984). "The Shewhart Control Chart—Tests for Special Causes." *Journal of Quality Technology* 15:237–239.
Nelson, L. (1985). "Interpreting Shewhart X Control Charts." *Journal of Quality Technology* 17:114–116.
Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations." *Chemometrics and Intelligent Laboratory Systems* 35:45–65.
Nomikos, P., and MacGregor, J. F., (1995). "Multivariate SPC Charts for Monitoring Bach Processes." *Technometrics* 37:41–59.
Penny, K. I. (1996). "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance." *Journal of the Royal Statistical Society*, Series C 45:73–81.
Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.
Shiskin, J., Young, A. H., and Musgrave, J. C. (1967). *The X-11 Variant of the Census Method II Seasonal Adjustment Program. Technical Report 15*, US Department of Commerce, Bureau of the Census.
Sobol, I. M. (1993). "Sensitivity Estimates for Nonlinear Mathematical Models." *MMCE* 1.4:407–414.
Tracy, N. D., Young, J. C., and Mason, R. R. (1992). "Multivariate Control Charts for Individual Observations." *Journal of Quality Technology* 24:88–95.
Wheeler, D. J. (2004). *Advanced Topics in Statistical Process Control.* 2nd ed. Knoxville, TN: SPC Press.

