

Evaluating El Niño's Effect in the Mid-Michigan Area and Puerto Rico using Multiple Linear Regression

Emily Baker
Kelan Hlavaty
Dana Kell
Sarah Nowak
Central Michigan University
Mt. Pleasant, MI 48858

Maida Cabezudo
University of Puerto Rico - Cayey
Cayey, PR 00736

July 30, 2007

Abstract

El Niño is a phenomenon that has been investigated for over 100 years and is still a topic of research today. Our project uses data collections of sea surface temperatures in the El Niño 3.4 region off the West Coast of South America to evaluate air temperatures in the Mid-Michigan area and Puerto Rico. The air temperature data is a collection of monthly averages dating back to 1950. In this talk we will discuss background information on the El Niño phenomenon and linear regression theory and present two competing linear models to explain the air temperatures in Mid-Michigan based on sea surface temperature and one model that analyzes El Niño's effect in Puerto Rico.

Linear regression theory was chosen to evaluate the data sets because the underlying theory has been very well developed, is very versatile in what can be done with it and is not overly difficult to understand. In order to use linear regression, three basic assumptions must be met: constant variance, independence and normality. To satisfy the assumption of constant variance the residual plots obtained from our models must exhibit a random scatter and give no evidence of any pattern such as fanning. To satisfy the assumption of independence, a Durbin-Watson Test must be performed in order to evaluate the autocorrelation, and transformations must be performed to adjust the autocorrelation as close as possible to zero. In order to satisfy the last

assumption, normality, four tests (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling and Cramer-von Mises) can be performed to test for normality in the distributions. Once these assumptions are satisfied, we have obtained a workable model.

Our first model integrates the sea surface temperature data to create a cumulative sea surface temperature variable, which is then used to predict temperatures in Michigan. The second model examines each season individually, essentially creating four models to better explain the variability of the four seasons. Finally, the Puerto Rico model examines the air temperature data versus sea surface temperature to connect the effects of El Niño to this island. The issue of global change is also addressed in each model for these two geographical areas.

1 Introduction

To begin our research, we came to the decision that we wanted to find a model that explained temperature variations over time due to different El Niño factors. We chose to use temperature data from the Central Michigan area and Puerto Rico while focusing on sea surface temperature as a predictor of air temperature. The data for air temperature was a series of monthly averages dating back to 1950 and through the present. The sea surface temperature data was monthly averages of the sea surface temperature off the west coast of South America, the El Niño 3.4 region, dating back to 1950 and to the present.

After constructing various models, our team came up with two competing Mid-Michigan models and one model for Puerto Rico. The first Michigan model uses integration to look at the air temperature in the Mid-Michigan area based on cumulative sea surface temperatures. The second Michigan model divides the months in four groups of three months; four seasons. This allows for the variability of the months to be better explained by grouping the months with similar temperatures. The Puerto Rico model uses the same sea surface temperature data to model air temperature in Puerto Rico. The models can be used to predict future temperatures in Central Michigan and Puerto Rico during future El Niño periods. Although similar research has been done on this topic previously, not much has been done that considers the strength and duration of an El Niño period rather than just acknowledging the presence of El Niño in Central Michigan and Puerto Rico.

This paper will discuss various aspects of the research process starting with background information on El Niño and linear regression theory. Next, model information and results will be presented for the Michigan Cumulative Sea Surface Temperature Model, the Michigan Four Seasons Model and the Puerto Rico Model followed by a brief summary and discussion of potential future research topics.

2 Background Information

El Niño has been a topic of research for various years. Although much research has been done, not a lot is known about the phenomenon as a whole. By definition, El Niño is an abnormal warming of sea surface temperatures and a sea-saw pattern of reversing surface air pressure between the eastern and western tropical Pacific. It occurs irregularly, making it extremely difficult to predict. However, it is known that El Niño periods always begin around Christmas time and last several months with varying strength. The strongest El Niño period to date occurred in 1997-1998.

The effects of El Niño vary with different parts in the world. El Niño can increase rainfall in one part of the world, leading to landslides and mudslides, while causing severe drought in another part of the world simultaneously.

Similarly, El Niño may cause severe cyclones and storms in one area while preventing these storms in a different area.

3 Linear Regression Theory

Linear Regression was chosen to evaluate the data sets because the underlying theory has been very well developed, is not overly difficult to understand and is very versatile in what can be done with it.

3.1 Assumptions

Constant variance, independence and normality of the residuals must be satisfied in order to use linear regression. If one or more of these assumptions is not met, then transformations must be performed on the data in some manner to ensure linearity and satisfy all the assumptions. It is not a guarantee that making one or more transformations will ensure that the assumptions are satisfied. The following equation represents these assumptions:

$$\epsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

Where:

- ϵ = the residuals (actual-predicted)
- iid= independent, identically distributed
- $N(0, \sigma^2)$ = normally distributed random variables

3.2 First Assumption (Constant Variance)

Definition (Variance) *Variance is the sum of the difference between the actual and predicted values squared, divided by the difference between the number of observations and parameters. Variance is denoted by the following equation:*

$$\sigma^2 = \Sigma(Y_i - \hat{Y})^2 / (n - p)$$

Note that because Y is dependent upon X, there is a possibility that as X changes, the variance will also change. If this occurs, then the variance is considered to be non-constant and the assumption is not satisfied. One way to examine this condition is to observe the residual plots of the data.

Definition (Residual) *How far above or below the actual value fall from the predicted value. See figure 3.2.*

In order to satisfy the assumption of constant variance the residual plot needs to exhibit random scatter or a gun shot affect. A model may contain both high and low variances (thus being non-constant) if it appears to fan. In the context of the El Nio models, fanning B was an issue when all the months were group together because the early months of January and February are much more variable than the later months of August and September.

3.3 Second Assumption (Independence)

Definition (Independence) A random variable is independent when the probability of the first event occurring does not have an impact on the probability of the second even occurring.

Time series data often violates this assumption. Due to the fact that a given day's temperature is often related to the days preceding it, the temperature data used was autocorrelated and not independent.

Defintion (First Order Autocorrelation) When X_i is related to X_{i-1}

Definition (Durbin-Watson Test) The Durbin-Watson Test tests for independence by checking for autocorrelation between the residuals. It does so by developing test statistic, based of the following formula, and then compares the value to an interval that is based on the number of observations, independent variables and significance level.

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

Where e_1, e_2, \dots, e_n are time-ordered residuals.

Durbin-Watson Hypothese:

$$H_o: \rho = 0 \text{ (No Autocorrelation Present)}$$

$$H_1: \rho > 0 \text{ (Autocorrelation Present)}$$

H_o is rejected if $d < d_{L,\alpha}$

H_o is not rejected if $d > d_{U,\alpha}$

The test is inconclusive if $d_{L,\alpha} \geq d \leq d_{U,\alpha}$

When H_o is rejected, a value of ρ (correlation coefficient) is assigned to the lag of the term as follows:

$$X_t^* = X_t - \rho X_{t-1}$$

The reason for this transformation is as follows:

$$Y_t = \beta_0 + \beta_1 x + \epsilon_t$$

Where Y_t is the equation of the model and ϵ_t represents the error.

$$\epsilon_t = \rho(\epsilon_{t-1}) + \delta_t$$

Where $\rho(\epsilon_{t-1})$ is the part that is autocorrelated and δ_t = the normal distribution of $\delta \stackrel{iid}{\sim} N(0, \sigma^2)$.

$$Y_t^* = Y_t - (\rho Y_{t-1})$$

$$Y_t^* = (\beta_0 + \beta_1 x_t + \epsilon_t) - \rho(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1})$$

$$Y_t^* = \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \epsilon_t - \rho \epsilon_{t-1}$$

$$\textbf{Final Equation } Y_t^* = \beta_0^* + \beta_1^* x_t + \delta_t$$

Where $\beta_0^* = \beta_0(1 - \rho)$ and $x_t^* = x_t - \rho x_{t-1}$. This will take care of any dependence between the terms.

3.4 Third Assumption (Normality)

Definition (Normal Distribution) *A distribution is considered normal when a histogram of the residuals is bell-shaped and centered around the mean. See Figure 3.4.*

There are four tests (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling and Cramer-von Mises) that test for normality in distributions. They are all Goodness of Fit Tests that compare the distribution to the theoretical bell curve. The only way that these tests differ is in the amount of weight they assign to each value according to where it falls on the bell curve (i.e. $\sigma_1, \sigma_2, \sigma_n$) They are evaluated based on the following hypotheses:

$$H_o: \beta_p = 0 \text{ (Data is normally distributed)}$$

$$H_1: \beta_p \neq 0 \text{ (Data is not normally distributed)}$$

In order for each test to pass, the null hypothesis cannot be rejected, or in other words, the test statistic must be greater than the alpha level.

3.5 El Niño Models' Hypotheses

$H_o: \beta_p = 0$ (Sea surface temperature is not a significant indicator of air temperature in neither the Mid-Michigan nor Puerto Rico areas.)

$H_1: \beta_p \neq 0$ (Sea surface temperature is a significant indicator of air temperature in the Mid-Michigan and Puerto Rico areas.)

The null hypothesis was tested at a significance level of $\alpha = .05$. This means that if the p-value (probability of the event occurring) is less than .05, then the null hypothesis is rejected and the results are considered statistically significant, or in other words, cannot be attributed to pure chance alone.

4 Cumulative Sea Surface Temperature Model

The first model that we developed to explain air temperatures in Mid-Michigan is the Cumulative Sea Surface Temperature Model (or the Cumulative SST Model). The cumulative sea surface temperature variable was found by taking the index values from the data set, modeling these values as a graph, and finding the area under the curve. We wrote a program for SAS that took every scenario of the graph into account to accumulate the sea surface temperature based on a point's previous value. This variable was labeled A. The equation for this model, including A, is as follows:

$$\text{Airtemp} = \beta_0 + \beta_1 i + \beta_2 i^2 + \beta_3 A + \beta_4 A^2 + \beta_5 \text{Feb} + \beta_6 \text{Mar} + \beta_7 \text{Apr} + \beta_8 \text{May} + \beta_9 \text{Jun} + \beta_{10} \text{Jul} + \beta_{11} \text{Aug} + \beta_{12} \text{Sep} + \beta_{13} \text{Oct} + \beta_{14} \text{Nov} + \beta_{15} \text{Dec}.$$

β_0 is the intercept, or January, i is the date, and the months are all indicator variables.

Several transformations were performed on the variables to ensure that there was linearity in the model. One of these was adding the autocorrelation coefficient, ρ , to create A_2 , i_2 , and t_2 , where t is the air temperature. For this model, $\rho = 0.15$. We squared two terms, i_2^2 and A_2^2 , to help linearity and diminish p-values. Finally, we raised t_2 to the 1.25 power to help the non-constant variance in the residual plot.

In the SAS output, we looked at the normality tests, the autocorrelation, the p-values, the residual plot, and the R^2 value to determine if the Cumulative SST Model was acceptable. The three normality tests we used were the Shapiro-Wilk test, the Cramer-von Mises test, and the Anderson-Darling test, all of which are above alpha. The values are 0.4142, 0.2500, and 0.2500 respectively. The fourth test that was present in the SAS output was the Kolmogorov-Smirnov test, however, we did not use this test because it is for sample sizes above 2000 (it does not apply to our model). The autocorrelation was adjusted with ρ to be 0, therefore the values are not dependent on the lag and the assumption of independence is satisfied.

The p-values for all of the variables (the intercept, i_2 and i_2^2 , A_2 and A_2^2 , and all of the months) are all below alpha so they are significant. None of the variables had to be removed. Note that i , the date, is significant. This is representative that there has been global change in Mid-Michigan from 1950 to the present. In further expanding on this, we found **Figure 4a**. From 1950 until 1973, there has been a global cool down and from 1973 to the present there has been global warming. Although this data shows global change, keep in mind that this is a small time period (only 57 years) and it is a small geographical region, therefore nothing conclusive can be drawn from these results to apply to a grander scale (i.e. statewide, nationally).

Even though t_2 was raised to the 1.25 power to fix the non-constant variance in the residual plot, slight fanning of the type B is still present. **See Figure**

4b. This a drawback of this model: there is still a small fan in the residual plot. Despite this downside, one of the major advantages of the Cumulative SST Model is found in the residual plot: the R^2 value, which equals .9604. This means that the model explains 96.04 percent of the variability in temperature in the actual data, which is extremely high.

We are accepting the Cumulative Sea Surface Temperature Model as a significant predictor of air temperature in Mid-Michigan based on cumulative sea surface temperature because the p-values are all below alpha, the normality tests are above alpha, the autocorrelation is 0, and the R^2 is extremely high.

5 Michigan Four Seasons Model

The second model we developed groups the months into four groups of three months with similar temperatures in order to account for the fact that Michigan has four distinctly different seasons. We used the same temperature data for the Central Michigan area that was used in the Cumulative SST model, however, we used the actual index values for sea surface temperature. All p values were less than $\alpha=0.05$ for those variables included in the equations for each model. Only a rho transformation was used to correct the autocorrelation in each season's model. All values for the normality tests that applied to the model were greater than $\alpha=0.05$. The residuals plots did not have any fanning issues and the R^2 values were acceptable.

5.1 Seasonal Model Summaries

The winter season model included temperature values for the months of January, February and March. The equation for the final model was as follows: $T_2 = \beta_0 + \beta_1 \text{Date}_2 + \beta_2 \text{Date}_2^2 + \beta_3 \text{SSTemp}_2 + \beta_4 \text{SSTemp}_2^2 + \beta_5 \text{Feb} + \beta_6 \text{Mar}$ where β_0 is the intercept, January, Date is the global change component, SSTemp is the index value for sea surface temperature in the El Niño 3.4 region and February and March are indicator variables. In order to correct the autocorrelation to zero, we used $\rho=0.123$ and our R^2 value came out to be 0.6533.

See Figure 5.1a

The spring season model included temperature values for the months of April, May and June. The equation for the final model was as follows: $T_2 = \beta_0 + \beta_1 \text{May} + \beta_2 \text{Jun}$ where β_0 is the intercept, April, and May and June are indicator variables. In order to obtain an autocorrelation of zero, a ρ value of 0.008 was used, yielding an R^2 value of 0.8926.

See Figure 5.1b

The summer season model included temperature values for the months of July, August and September. The equation for the final model was as follows: $T_2 = \beta_0 + \beta_1 \text{Aug} + \beta_2 \text{Sep}$ where β_0 is the intercept, July, and August and September are indicator variables. The autocorrelation was equal to zero when a ρ value of 0.215 was used, leaving us with an R^2 value of 0.8279.

See Figure 5.1c

The fall season model included temperature values for the months of October, November and December. The equation for the final model was as follows: $T_2 = \beta_0 + \beta_1 \text{Nov} + \beta_2 \text{Dec}$ where β_0 is the intercept, October, and November and December are indicator variables. The autocorrelation was adjusted to zero using a ρ value of 0.144, resulting in an R^2 value of 0.8911.

See Figure 5.1d

5.2 Explanation of Models

The model for the winter season is the only model that includes the sea surface temperature variable and the global change component in the equation. This means that evidence of global change and the effects of El Niño are only noticeable in the winter months in Central Michigan.

5.3 Comparison of the Michigan Models

Both the Michigan Cumulative Sea Surface Temperature Model and the Michigan Four Seasons Model use the same monthly temperature data for the Central Michigan data and the same sea surface temperature data from the El Niño 3.4 region, however the sea surface temperature data is used in two different ways. The Cumulative SST Model uses accumulation and integration of the sea surface temperature index values while the Four Seasons Model uses the straight index values.

With two competing models, it is important to evaluate the pros and cons of each model. The biggest pro of the Cumulative SST Model is the fact that it is one encompassing model whereas the Four Seasons Model is four models total. However, since the months with similar temperatures are grouped together in the Four Seasons Model, the diagnostics are better than those of the Cumulative SST Model and it acknowledges that predictors such as sea surface temperature have different influences in different seasons and cannot be generalized over all the months. The residual plots for the Four Seasons Model exhibit constant variance whereas the residual plot for the Cumulative SST Model shows evidence of slight fanning, which is undesirable. The R^2 value for the Cumulative SST Model is extremely high (.9604) as opposed to a range of 0.6533-0.8926 for the Four Seasons Model, however, an unusual transformation had to be performed on the air temperature variable in the Cumulative SST Model in order to obtain a working model.

6 Puerto Rico Model Results

For this model, we used Puerto Rico's monthly average temperature from January 1956 to December 2006 and the sea surface temperature from the El Niño 3.4 region. In the Puerto Rico Model there is an intercept [JanuaryDecember], the date, the cumulative sea surface temperature variable, and all the months. In order to get an acceptable model, we had to make several transformations to some of the variables to make sure that the assumptions of linear regression were not violated. The transformations we used were the rho transformation and the lag transformation. The rho that was used to fix our model was 0.5654. Also we added a squared term and a cubed term to the date variable.

The equation for the Puerto Rico Model is:

$$y = \beta_0 + \beta_1 \text{datelag} + \beta_2 \text{datelag}^2 + \beta_3 \text{datelag}^3 + \beta_4 \text{Alag} + \beta_5 \text{Feb} + \beta_6 \text{Mar} + \beta_7 \text{Apr} + \beta_8 \text{May} + \beta_9 \text{Jun} + \beta_{10} \text{Jul} + \beta_{11} \text{Aug} + \beta_{12} \text{Sep} + \beta_{13} \text{Oct} + \beta_{14} \text{Nov}$$

Where;

β_0 is the intercept [JanDec],

$\beta_1 \text{datelag}$ is the transformation in date [global warming component],

$\beta_2 \text{datelag}^2$ is the square term for the datelag,

$\beta_3 \text{datelag}^3$ is the cube term for the datelag,

$\beta_4 \text{Alag}$ is the transformation for the cumulative sea surface temperature [El Niño component], and

$\beta_5 \text{Feb}$ thru $\beta_{14} \text{Nov}$ are the month variables.

The R square for the model is 0.7827, which means that the model explains 78 percent of the variability of the months. The p-values or the probability for the event to occur are all less than alpha (.05). The autocorrelation is -0.001, which is near 0, so the variables are not correlated. The normality tests for this model that we are considering are the Shapiro-Wilk, Cramer-von Mises, and Anderson-Darling, which are all above alpha. The residual versus the predicted values graph shows a random scatter pattern and the residual versus the months graph does not show fanning (**See Figures 6a and 6b**). Because of this, we accept the model as a good predictor of air temperature for Puerto Rico.

If we consider the equation for this model and change the betas with the parameter estimate values that were displayed in the output and plot them versus different variables we can see interesting graphs (recall that the equation predicts air temperature).

See Figure 6c.

With this graph we can explain the effect of global change in Puerto Rico. According to this graph, Puerto Rico shows an increasing of temperature and

then at some point it starts to cool down, however, at the end it starts to slightly increase again. Also, with this graph we can see that the effect of global change differs depending on the geographical region being studied (refer to the Cumulative Sea Surface Temperature graph for Mid-Michigan).

See Figure 6d.

With this graph we can trace how the temperature varies from month to month during the year. Notice how the temperature increases until June and then decreases in July to increase again in August.

Now, when we back transformed the variables we can see that our model can predict the temperature. The predicted values are almost the same as the actual temperatures recorded.

To summarize, the model shows evidence of the presence of El Niño because the sea surface temperature component is present. Also, there is evidence of global change in the island.

7 Conclusion/Future Research

Throughout this summer we have discovered various models that can be used to represent the relationship between Mid-Michigan air temperature versus sea surface temperature and Puerto Rico air temperature versus sea surface temperature. We have found evidence that global change exists and are able to predict air temperature with the encompassing Cumulative SST Model for Mid-Michigan. Likewise, the Four Seasons Model, which divides each year into four seasons of three months each, also shows global change in Mid-Michigan. However, this Four Seasons Model demonstrates that global change and El Niño are only noticeable in the winter months of January, February, and March. The final model, Puerto Rico, reveals that global change and El Niño are present in Puerto Rico.

Looking into the future, we are planning on exploring precipitation in relationship with sea surface temperature and how El Niño affects it. We would also like to investigate the effects that El Niño has on the Great Lakes.