Nathan Kim

Predicting El Niño Indicators using Recursion, Data Trees and Evolutionary Algorithms

**Rationale and Background:**

El Niño, one of the major meteorological phenomena of today’s world, is a period in which ocean temperatures in the Equatorial Pacific are unusually warm (Tropical Ocean Atmosphere Project) and is more formally defined as the occurrence of a “sea surface temperature departure from normal (for the 1971-2000 base period) in the Niño 3.4 region greater than or equal in magnitude to 0.5 degrees C (0.9 degrees Fahrenheit), averaged over three consecutive months” (NOAA News Online). It is characterized by “unusually high atmospheric sea level pressures… in the western tropical Pacific and Indian Ocean regions, and unusually low sea level pressures… in the southeastern tropical Pacific” (Tropical Ocean Atmosphere Project). El Niño causes increased rainfall across the southern tier of the US and in Peru, which has caused destructive flooding, and drought in the West Pacific, sometimes associated with devastating brush fires in Australia. Unfortunately, these damages linked to El Niño can cause billions of dollars in damages (Pielke and Landsea 2030), making methods of predicting El Niño so important.

One of the most important facets of El Niño predictions is the use of indices. An index is “a number scale in which all the individual factors needed to describe a complicated phenomenon are boiled down to a single number” (Barnston) and is easy to track and analyze over long periods of time. Due to this, indices are common and widespread ways to summarize El Niño status. There are many different El Niño indices used in practice today, including the Southern Oscillation Index (SOI), sea surface temperatures (SSTs) in different regions, and the Multivariate ENSO Index (MEI). The SOI is a “a standardized index based on the observed sea level pressure differences between Tahiti and Darwin, Australia” (National Centers for Environment Information: Southern Oscillation Index). Negative values of the SOI indicate a higher air pressure in Darwin than in Tahiti and often indicate El Niño episodes (Australian Government Bureau of Meteorology). SSTs of regions in the Equatorial Pacific Ocean are also commonly used as El Niño indices. One such region which is extremely important is the Niño 3.4, given by the region ranging from 120° W to 170° W and 5° N to 5° S (NOAA News Online). Recent average temperature anomalies in this region are given by the Oceanic Niño Index, or ONI (National Centers for Environment Information: Equatorial Pacific Sea Surface Temperatures). The Japan Meteorological Index, or JMI, is a similar index. It measures recent average temperature anomalies in the region given by 90° W to 150°W and 4° N to 4° S (Álvarez *et al*.). Unlike the previous three indices, the MEI uses many different variables in its calculations (as the name suggests). Specifically, it incorporates “sea-level pressure, zonal and meridional components of the surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of the sky” (Wolter).

There are many different methods which can be used to develop El Niño predictions, one of which is the use of an evolutionary algorithm. Evolutionary algorithms have many uses, one of which is the development of a predictive model. In this use, evolutionary algorithms usually follow this basic structure (What is an Evolutionary Algorithm?):

1) Generate many random models

2) Test each model for effectiveness

3) Create new models using characteristics of the better models of step 2

4) Apply mutations by randomly changing characteristics of present models or by adding random new models

5) Delete models with low effectiveness

6) Repeat steps 2 through 5 until a desired effectiveness is reached

As can be seen, the evolutionary algorithm is named as such due to its similarity with Charles Darwin’s theory of evolution (Jones 1).

Previous work has been done demonstrating the effectiveness of an El Niño prediction model which made use of an evolutionary algorithm (Álvarez *et al*.). However, this work is quite limited. First, the model only looked at the Japan Meteorological Agency (JMA) Index, the sea surface temperature of a region in the tropical Pacific. Other El Niño prediction models use many other factors and indices, as noted previously. Furthermore, the study only produced a model but did not test it for accuracy. These issues could be improved through the incorporation of additional and diverse factors into the model. In fact, by using different indices, insights regarding the relative importance of each factor in predicting El Niño may arise. This could be done by implementing different models incorporating different factors and observing which models work best. Also, the models could be tested for accuracy by looking at how well they would have predicted past El Niños. This could be achieved by using previous data which excludes the last several years to develop the models, and then by comparing the models’ predictions during these several years to the actual data.

**Hypotheses:**

El Niño indicators can be predicted to some degree using a data tree involving an evolutionary algorithm. In addition, these predictions can be made more accurate by incorporating more factors, some of which will work better than others.

**Questions:**

* Can a predictive model of El Niño be developed with the use of data trees and an evolutionary algorithm?
* Can this model be improved by selecting different factors to use over others? (In other words, do some factors produce more accurate predictions than others?) If so, which factors are these?

**Engineering Goals:**

The main goal of this project is to build a predictive model of El Niño using indices, data trees, recursion, and an evolutionary algorithm. Another goal is to find out which indices or other factors are more predictive of El Niños than others, potentially improving the model.

**Expected Outcomes:**

Based on prior research, it was expected that a model of El Niño could be built using methods stated beforehand. Furthermore, it was also expected that some indices would demonstrate better effectiveness than others. By building a model that combined these “good” indices, the effectiveness of the model was expected to improve.

**Materials:**

* A computer to run Python (Windows 7 used)
* Any version of Python 3 (3.4.3 used)
* Optional: Python interface (IDLE used)
* Data: Values of indices, such as monthly SOI values from 1991 to 2015 (<https://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soidatafiles/MonthlySOI1887-1989Base.txt> )

**Procedure:**

Let f(t) = the value of an index at a month t. Thus f(t-k) equals the value of that index k months before the time t.

The models were represented using a data structure known as a tree. The root of the tree is an empty node, and all other nodes consisted of an operator and a value. The set of all operators was +, -, \*, /, and ^. Values could either be constants, such as 0.5, 1, or 2.97, or previous SOI values such as f(t-60).

Each path directly from the root to a node represented a model. For example, if the nodes in such a path were "" (the root), "+f(t-1)", "\*2", and "+f(t-3)", in that order, then the model was given by f(t) = f(t-1) \* 2 + f (t-3). Note that the operator of the second leaf is disregarded.

The program, written in Python 3.4.3 for Windows 7, did the following:

1. The program first took in the values of the given index. To develop the models, only values up to 2010 were used as training data.
2. The program then generated a tree containing 100 random models with very low depth (no more than 2).
3. The program tested each model against the training data and calculated the accuracy of each one using an average of RSS (residual sum of squares), RMSE (root-mean-square deviation), and MSE (mean squared error). The accuracy of the most accurate model, the accuracy of the least accurate model, and the median accuracy were recorded.
4. The program deleted the branches of the 50 worst models (i.e., those which have the lowest accuracy).
5. To replace these 50 models, the program duplicated the best 35 models and created 15 new models by attaching leaves to pre-existing nodes. The duplication was similar to reproduction, and the creation was similar to mutation.
6. Next, the program performed approximately 10 crossovers. A crossover consisted of swapping an entire subtree with another subtree, while still keeping these subtrees and the rest of the tree intact. This process preserved the number of leaves in the tree and thus the total number of models as well.
7. The program performed approximately 10 mutations. A mutation consisted of replacing random nodes in the tree having sufficiently high depth with new, randomly generated nodes. Again, this process also preserved the number of leaves in the tree and the total number of models. Note that these mutations differed from those described in step 4.
8. Steps 3 through 7 were repeated 99 more times in order to simulate 99 more generations (100 generations total).
9. Finally, the most accurate model and its accuracy were recorded. Furthermore, this model was compared against the data from 2011 to 2015, and its accuracy at predicting the index itself during that five-year stretch was also recorded.

(Note: Source code will be attached to the final project.)

In order to test the program, various sets of input data using different indices were fed and the program was run each time.

**Risks and Safety Precautions:**

There were no major safety risks involved with this project. One very unlikely, but still possible, risk was the chance of fire, burning, short circuiting, or any other physical problems caused by hardware failure. To take precautionary measures against this, the computer was carefully monitored for signs of these problems. Also, a flawed program can have memory issues, not only crashing the program but also potentially deleting the program and altering other important files. But the chances of these memory issues were lowered by conducting smaller amounts of trials at a time and by setting low bounds for various parameters, such as number of generations and size of trees.

**Data Analysis:**

As stated in step #9 of the program in the procedure, accuracy of any model can be found by comparing its predictions to real-life data. This accuracy can then be compared against those of other models, whether created in this project or not. This would indicate how effective, if at all, the model in question was at predicting facets of El Niño. Furthermore, multiple models using the genetic algorithm archetype can be created, each of which involves a different index or indices. Then the accuracy of each of these models can be compared in order to find out which indices were most predictive of El Niño.

**Results and Conclusions:**

(To be completed once data analysis is actually done)

**References:**

Barnston, Anthony. "Why Are There so Many ENSO Indexes, Instead of Just One?" *NOAA Climate.gov*. National Oceanic and Atmosphere Association, 9 Jan. 2015. Web. 7 Dec. 2015.

"Definitions of El Nino, La Nina, and ENSO." *Tropical Ocean Atmosphere Project*. Pacific Marine Environmental Laboratory, n.d. Web. 27 Oct. 2015.

"EL NIÑO/SOUTHERN OSCILLATION (ENSO) DIAGNOSTIC DISCUSSION." *National Weather Service Climate Prediction Center*. National Oceanic and Atmosphere Association, n.d. Web. 7 Dec. 2015.

"Equatorial Pacific Sea Surface Temperatures." *National Centers for Environmental Information*. National Oceanic and Atmosphere Association, n.d. Web. 7 Dec. 2015.

Jones, Gareth. "Genetic and Evolutionary Algorithms." *Encyclopedia of Computational Chemistry* (n.d.): n. pag. *Wiley Online Library*. John Wiley and Sons, Ltd., 15 Apr. 2002. Web. 7 Dec. 2015.

Álvarez, Alberto, Pedro Vélez, Alejandro Orfila, Guillermo Vizoso, and Joaquín Tintoré. "Evolutionary Computation for Climate and Ocean Forecasting: “El Niño Forecasting”." *Proceedings of the Second International Conference on Euro* (1999): n. pag. *DIGITAL.CSIC*. Web. 26 Oct. 2015.

"NORTH AMERICAN COUNTRIES REACH CONSENSUS ON EL NIÑO DEFINITION." *El Nino Definition*. NOAA News Online, 23 Feb. 2005. Web. 26 Oct. 2015.

Pielke, Roger A., Jr., and Christopher N. Landsea. "La Niña, El Niño, and Atlantic Hurricane Damages in the United States." *Bulletin of the American Meteorological Society* (1999): 2027-033. *Hurricane Research Division*. Atlantic Oceanographic & Meteorological Laboratory of NOAA, 7 Apr. 1999. Web. 26 Oct. 2015.

"Southern Oscillation Index." *Australian Government Bureau of Meteorology*. Australian Government, n.d. Web. 7 Dec. 2015.

"Southern Oscillation Index (SOI)." *National Centers for Environmental Information*. National Oceanic and Atmosphere Association, n.d. Web. 7 Dec. 2015.

"What Is an Evolutionary Algorithm." *What Is an Evolutionary Algorithm?* Department of Computer Sciences at VU University Amsterdam, n.d. Web. 26 Oct. 2015.

Wolter, Klaus. "Multivariate ENSO Index (MEI)." *Earth System Research Laboratory Physical Sciences Division*. National Oceanic and Atmosphere Association, n.d. Web. 7 Dec. 2015.