COMP – 4442 11/22/2020 Paul O'Leary

Executive Summary:

This report attempts to answer the question, "Can future rising sea levels be successfully modeled by SARIMA for time series forecasting of existing data?" or on a more personal level, "When will my family's Delaware coastal home be threatened by rising sea levels?" According to research published by the scientific magazine Nature Communications (Nature Communications, 29 October 2019), 200 million people will live below high tide level by the year 2100. This is a conservative estimate and could be much worse depending on how conditions progress. Accurate sea level predictions can help inform those that will be affected, allowing as much time as possible to address the situation.

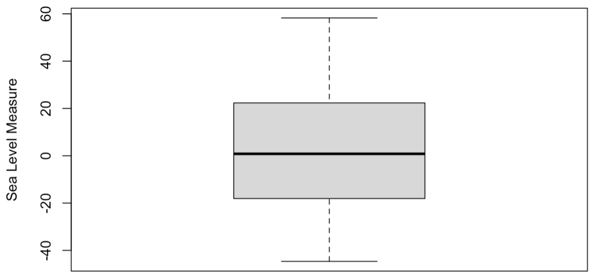
Univariate sea level data from NASA's MEaSUREs program was used to create a prediction model, and that model was compared to test data. The SARIMA forecasting exhibited great potential. However, due to some shortcomings with the data, the outcome predicted by this model fell short of actual sea level changes. Those shortcomings include the single variable nature of this analysis – numerous other factors may affect sea levels, and addressing those factors was beyond the scope of this analysis and issues with the data set chosen.

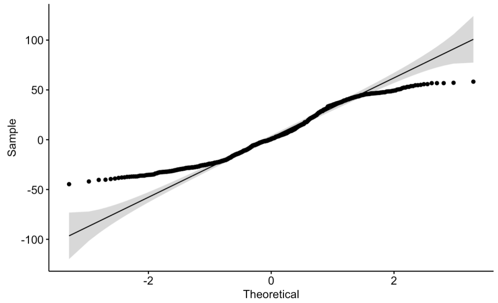
Based on the limited predictions of this model, the researcher's home will be well above sea level for more than 1200 years. However, given that the model predicted slower sea level rise than actually recorded, the humans that live close to sea level should be concerned.

Data source:

The Global Mean Sea Level (GMSL) data was computed at the NASA Goddard Space Flight Center under the auspices of the NASA MEaSUREs (Making Earth System Data Records for Use in Research Environments) program. The GMSL was generated by measuring sea surface heights from a number of satellites to a common terrestrial reference frame. More information regarding the data source is in the References section.

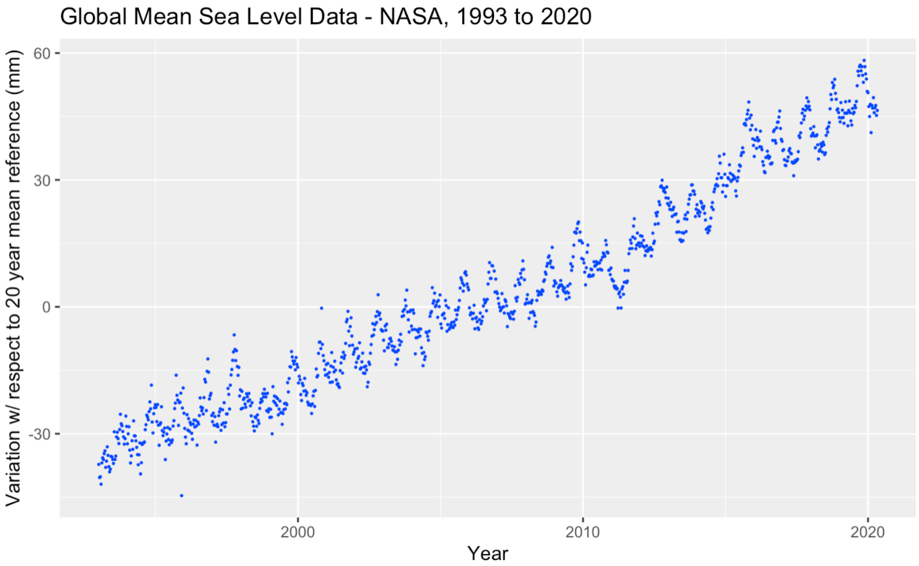
The data consists of 36 or 37 samples per year from 1993 to the present. Measurements are adjusted with respect to a 20-year mean reference. From the data, the date and the raw GMSL were used in this analysis. Basic data analysis revealed no missing or bad records in the data set.





A box plot showed no apparent outliers. The QQ Plot looks a bit questionable because the data is not stationary, which will be addressed later.

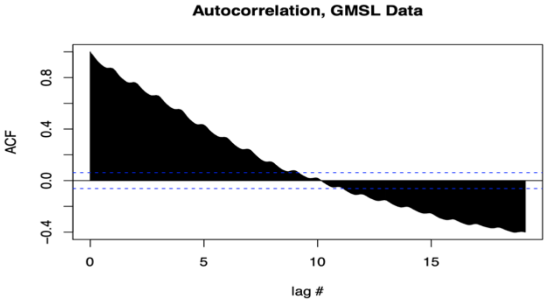
Time Series Data:



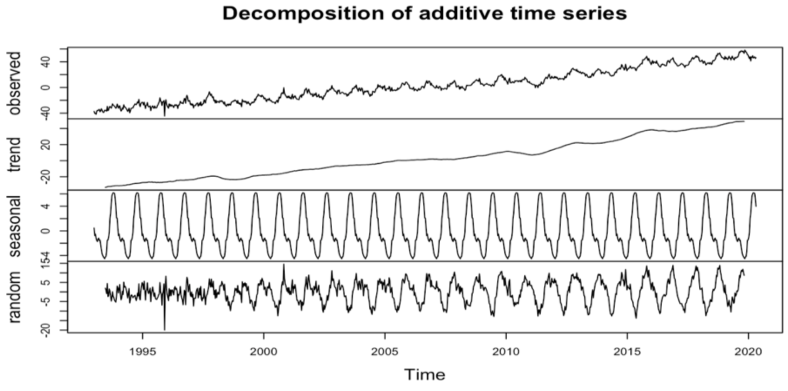
The plot above shows the raw GMSL data imported from the NASA file. The plot shows typical time series data. Time Series data is simply the measure of some data value recorded over time, such as stock prices, smog levels or sea levels. Generally, time is the independent variable. Often, researchers want to attempt to predict the future based on what the data was observed to be in the past. Certain aspects of time series data need to be addressed before any analysis can proceed.

Time series data can be additive or multiplicative. Additive data will show a generally linear trend, and the amplitude of a seasonal cycle, if it exists, will be consistent. For example, for this GMSL data, the difference between a peak and a valley in 1993 will be roughly the same as that difference for 2018. This GMSL data is additive.

Stationary data will have a mean and a variance that remain constant over time. The mean for the GMSL data will be increasing because the sea level is increasing over time. Therefore, the GMSL data is not stationary. There are visual and quantitative methods to check for stationary data. The serial correlation plot calculates the correlation with the observations from previous time steps, called lags. Almost all the lags for this data are outside the 95% confidence interval, meaning the data is not stationary. The Augmented Dickey-Fuller Test can also be used.



Finally, time series data may show seasonality. Seasonal data will show a cyclical repeating pattern, the most familiar example being average temperatures over the course of a year for a location that has seasons. The GMSL data appears to have a cyclical seasonality. The data can be decomposed into components, to show a trend, the seasonality and the leftover random or "noise" component of the data. In this case, the GMSL data shows clear seasonality.

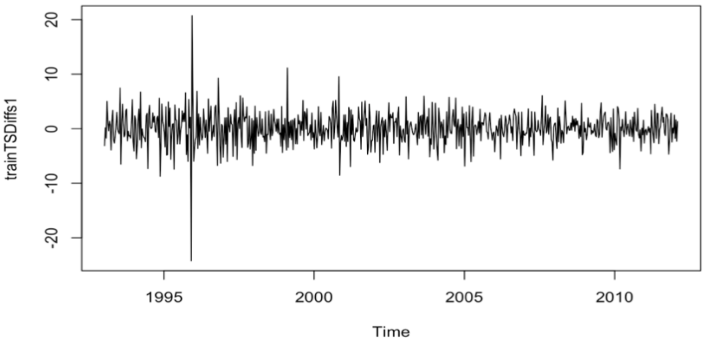


SARIMA Modeling:

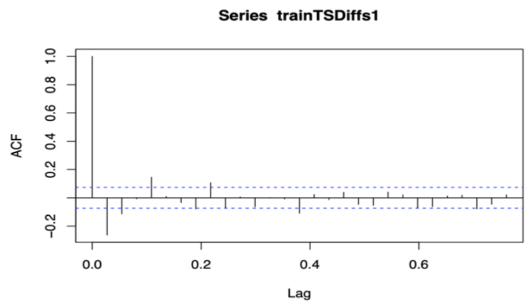
In order to address the research question, the GMSL data will be split into train (70%) and test (30%) sets. This results in 704 measurements in the train set and 302 measurements in the test set.

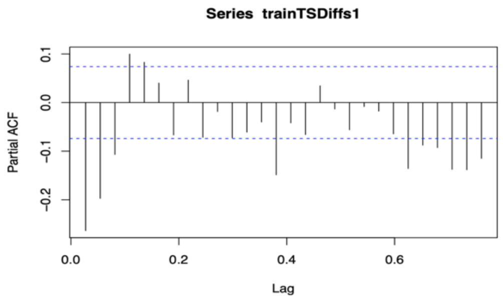
SARIMA is an extension of ARIMA, which is a combination of Autoregression and Moving Average forecasting methods – Seasonal AutoRegressive Integrated Moving Average. The autoregression component – AR(p) – maps the time series onto itself, with current values depending on previous values with some lag. The moving average component – MA(q) – creates a series of means of different subsets of the data. The order of integration – I(d) – is the number of differencings necessary to make the data stationary. The seasonality portion of the model – S(P, D, Q)m – adds a P and a Q for the seasonal component of the data. The m represents the length of each season, and the D is the number of differences to remove seasonality from the data set. Putting that all together - SARIMA(p, d, q)(P, D, Q)m.

Attempts were made to estimate the p, d, q and P, D, Q components of the train set of the GSML data. First the data was differenced to remove the seasonal component. Differencing transforms the data to have a mean of 0, with the result shown here:



As issues with the model were first encountered, multiple differencings and other differencing techniques, including differences of logs of the data, were attempted, with no improvement to the model. ACF and PACF plots were analyzed to estimate the q and p components, respectively.





In order to estimate a value for q from the ACF graph, q represents the lag after which the remaining lags are less significant. In a similar manner, the p can be estimated from the PACF graph. Similar steps can be taken to estimate the seasonal components for the model estimator.

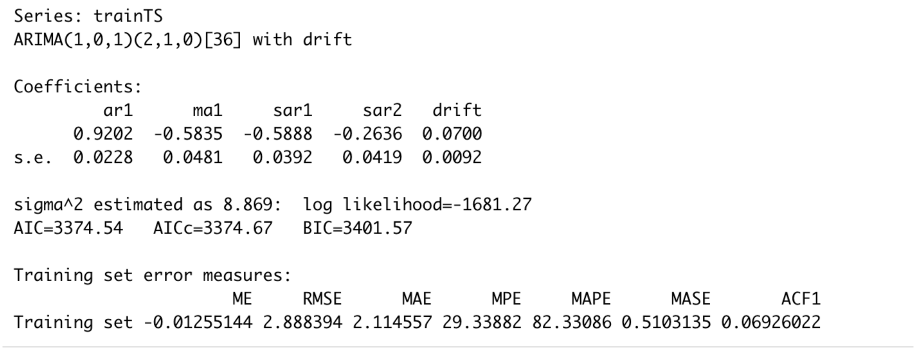
Many manual, iterative attempts were made to determine the components for the SARIMA process, without determining a model that was an acceptable predictor. However, in the "Forecast" package for the R programming environment, a systematic method exists to determine the best possible model given the data - "auto.arima".

The predicted results:

The auto.arima function takes the train data set, parameters for seasonality, stationarity, drift and fitting method. The command used in R is shown here:

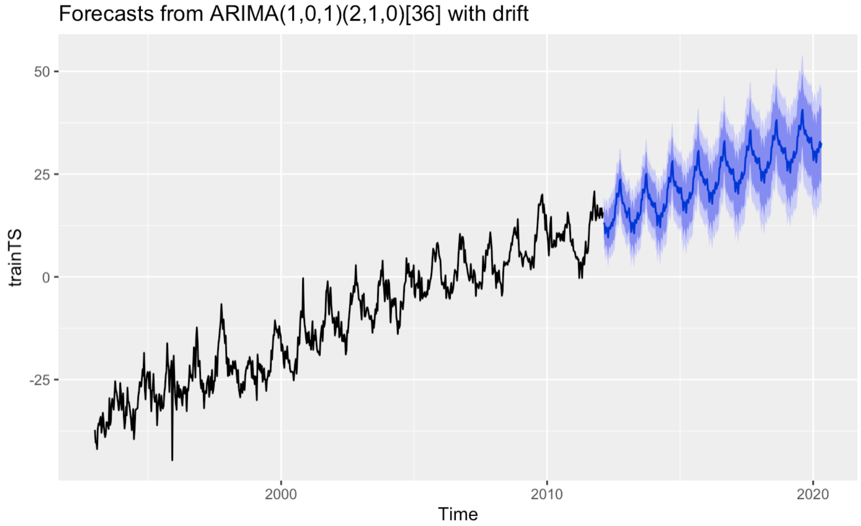
```
auto.arima(trainTS, allowdrift = TRUE, stationary = FALSE, seasonal = TRUE, method = "ML")
```

The auto.arima process steps through possible values for p, d, q and P, D, Q, m. Starting values, as well as maximum values, for p, d, and q are parameters. The function chooses the model that minimizes the AIC, BIC and AICc. The summary of the model selected is shown here:

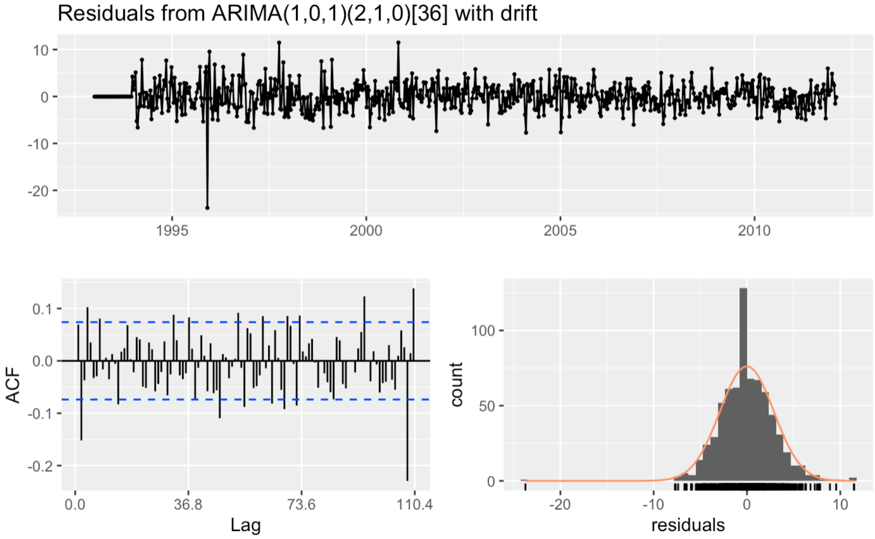


```
Series: trainTS
ARIMA(1,0,1)(2,1,0)[36] with drift

Coefficients:
         ar1      ma1     sar1     sar2   drift
      0.9202  -0.5835  -0.5888  -0.2636  0.0700
s.e.  0.0228   0.0481   0.0392   0.0419  0.0092

sigma^2 estimated as 8.869:  log likelihood=-1681.27
AIC=3374.54   AICc=3374.67   BIC=3401.57

Training set error measures:
                      ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set -0.01255144 2.888394 2.114557 29.33882 82.33086 0.5103135 0.06926022
```

A visual of the prediction of sea levels through 2020 based on this model found by auto.arima was encouraging.

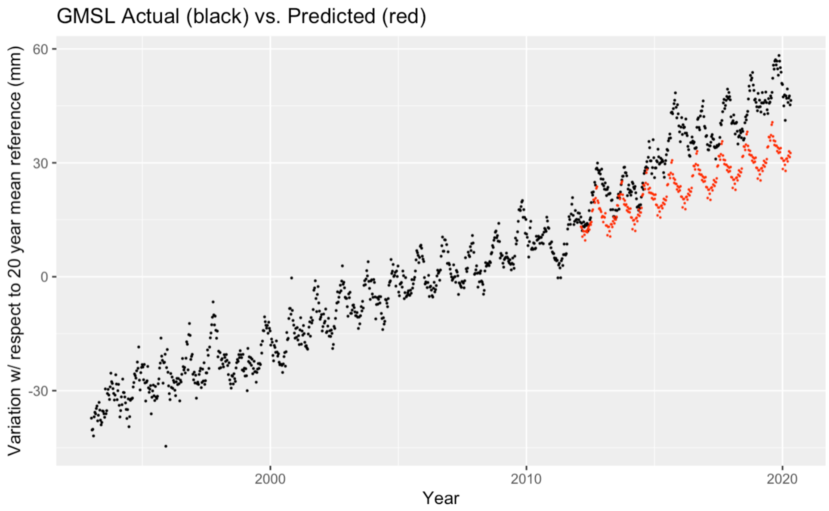


The model shows what appears to be a continuing trend upwards, and it maintains an obvious seasonality. The residuals plots are shown here.

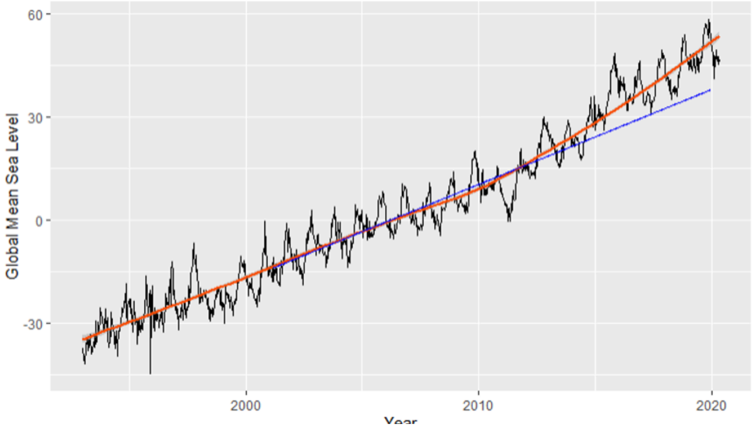


The residuals and the count vs. residuals plot for the modeled prediction are unremarkable. However, the ACF vs. Lag plot suggest that the model fit could be improved.

The most revealing plot is the predicted data overlayed with the actual data from the test set.



The model appears to have substantially under-predicted the increase in GMSL from the test set. A small amount of further analysis revealed that by pure chance, the data begins to increase at a slightly faster rate at about the point that the data was split into train and test sets.



In the graph above, the blue line very closely mimics the trend predicted by the auto.arima model for the test set. “Goodness of Fit” test results were, unsurprisingly, not stellar. The Ljung-Box test results in a very small p-value, which suggests that the model shows a lack of fit. An R-Squared test (76%), Mean Square Error (157.5) and Mean Absolute Percentage Error (29.6%) also showed less than good results.

Conclusions

The SARIMA process for predicting time series data can be very powerful. The model created maintained the trend and the seasonality of the train set. However, in order for modeling techniques to work properly, the data should be thoroughly examined for unexpected behavior. Predictions are only as good as the data used to create the model. Vagaries in the data chosen may have been overlooked as focus was directed to successfully modeling with the SARIMA process.

As for this researcher's family home near the Delaware coast, there will be plenty of time to enjoy coastal living. Based on the model created, the home will be safe until about the year 3296. However, rising sea levels will remain a concern for a large percentage of the global populace.

References:

Astaraky, Davood. (2015). Time Series Analysis in R – Decomposing Time Series. Retrieved from https://rpubs.com/davoodastaraky/TSA1

Brownlee, Jason. (2019). A Gentle Introduction to SARIMA for Time Series Forecasting in Python. Retrieved from https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

Coghlan, Avril. A Little Book of R For Time Series. Retrieved from https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/index.html

GSFC. 2017. Global Mean Sea Level Trend from Integrated Multi-Mission Ocean Altimeters TOPEX/Poseidon, Jason-1, OSTM/Jason-2 Version 4.2 Ver. 4.2 PO.DAAC, CA, USA. Dataset accessed 2020-08-08 at http://dx.doi.org/10.5067/GMSLM-TJ42

Kulp, Scott A. and Benjamin H. Strauss. "New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding." Nature Communications. 29 October 2019. https://www.nature.com/articles/s41467-019-12808-z.pdf

Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program. (2020, February 18). Retrieved October, 2020, from https://earthdata.nasa.gov/esds/competitiv-programs/measures

Piexeiro, Marco. (2019). The Complete Guide to Time Series Analysis and Forecasting. Retrieved from https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775