Workshop 3_Timeseries Models

Shalimar G. Moreno

1/31/2020

Objectives

The primary objectives of this time series analysis is to fit an ARIMA model to our dataset by:
1. Creating a time series object.
2. Decomposing a time series.
3. Testing for stationarity and detecting autocorrelation.
4. Choosing the order of an ARIMA model.
5. Using explanatory series to improve models.

Methods

Site Information

The Everglades has a unique landscape that has been shaped by water flow creating varying habitats for plant and animal communities across its spatial extent. As a subtropical wetland ecosystem, primary productivity changes in response to fluctuating water regimes. Long-term eddy covariance study sites have been placed along a hydrological gradient as part of the Everglades Flux Tower Network within the Everglades National Park. The data used in this data set was collected over 2 years from a mangrove scrub area at TS/Ph-7 site (Fig 1.)

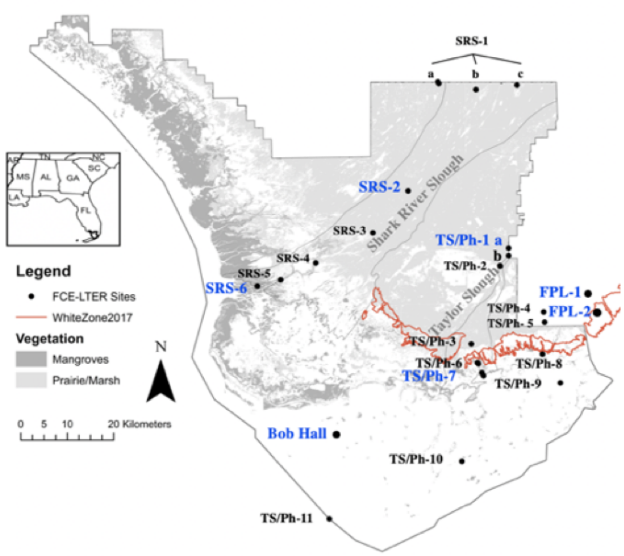

Fig 1. The Everglades Flux Tower Network displaying the long-term eddy covariance study sites.

Statistical Analysis

The data obtained from an eddy covariance tower at the Florida Coastal Everglades Long-term Ecological Research site TS/Ph-7 includes net ecosystem exchange (NEE; g C m2 day-1), total daily photosynthetically active radiation (PAR; W m-2 day-1), air temperature (tair; C), water temperature minimum and maximum (water.tmax and water.min; C) and water salinity minimum, maximum and mean (salinity.min, salinity.max, salinity.mean; ppt). To analyze what variables are affecting the NEE of our data set we will analyze the data using a time series model: autoregressive integrated model average (ARIMA). In our ARIMA model we specify our three parameters: p, d and q.

p: The number of lag observations included in the model (lag order)
d: The number of times that the raw observations are differenced (the degree of differencing)
q: The size of the moving average window (moving average)

In order to apply a time series model our data must meet two assumptions including:
1. Data must be stationary – its mean, variance and autocovariance are time invariant
2. the properties of the series doesn't depend on the time when it was captures. A white noise series and series with cyclic behavior can also be considered

To account for non-stationary we look to see if there is a trend or seasonal component. To check that our dataset met our assumptions we ran the augmented Dickey-Fuller (ADF) test, a formal statistical test for stationarity. The null hypothesis assumes that the series is non-stationary.

A second test is the auto correlation functions (ACF), a visual tool in determining whether a series is stationary. ACF plots display correlation between a series and its lags. If the series is correlated with its lags then we can assume there are some trend or seasonal components and the statistical properties are not constant over time.

Once we understand our data and the components that may be affecting it: trend, seasonal, cyclical or irregular components, we can then fit our model.

```
arima.nee1 <- auto.arima(nee, trace=TRUE)
```

To test if we trust our model we can plot our residuals using tsdisplay(). Once we are confident we trust it and it has met our assumptions we plot our model auto.arima and we test the fit using AIC.

Results

We plotted our data using the partial autocorrelation plots (PACF) which displays the correlation between a variable and its lags and it shows a few spikes at certain lags. Plotting our residuals shows that there is a small error range, mostly centered around 0, and so we can trust our ARIMA fitted model. Our ACF/PACF plots that display correlations between a series and its lags and correlation between a variable and its lags show if there are significant lags outside of our 95% significance boundaries. Our data at first had significant lags and so to fit the model better we change MA to the lags (Fig 2).

Once we adjusted the MA and had no significant lags we can use the Ljung-Box test for testing if our residuals from the ARIMA model have no autocorrelations and they didn't with a p-value 0.2325>0.05. After fitting an ARIMA model that met our assumptions, we can then forecast and make projections of our data.

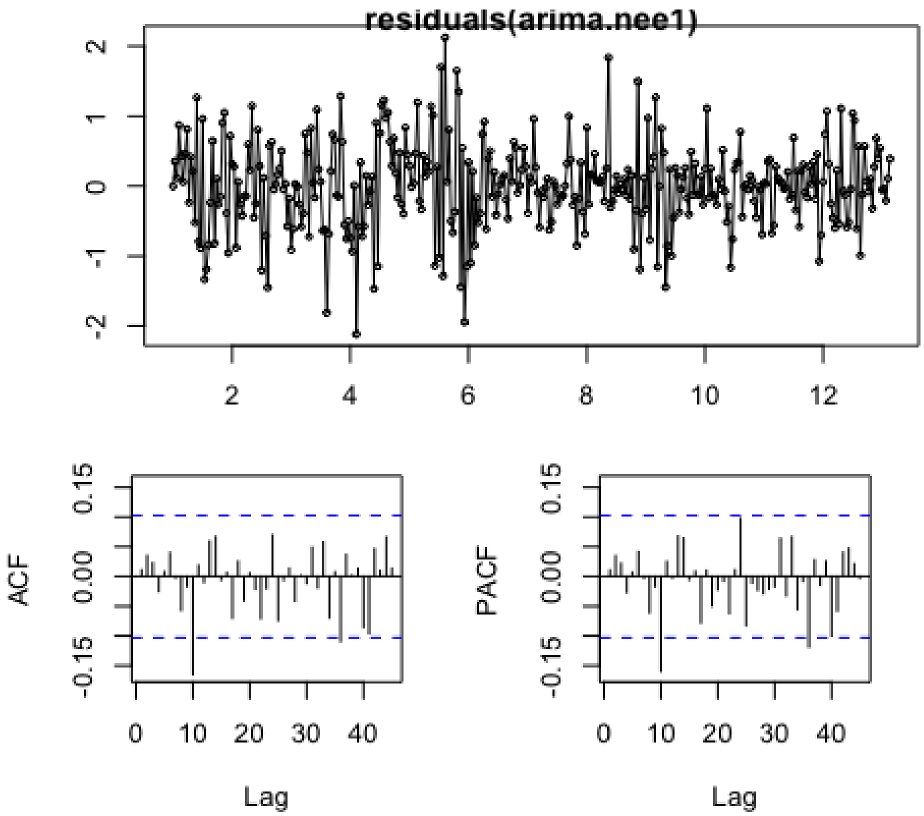

Fig. 2 Residuals from ARIMA plotted to show lags in ACF/PACF at 10 and 35.

To test if we have a good model fit for our data we use the Akaike Information Criteria (AIC), this quantifies the goodness of fit of the model and the simplicity of the model, by estimating the amount of information lost by a model. The lower the AIC, the better fit for the data. We then tested for independence by the Ljung-Box, which tests the "overall" randomness based on a number of lags, to the residuals of our fitted ARIMA model.

We fitted multiple models to forecast NEE values and see which was the best fit for our data. Table 1 shows that air temperature has the lowest AIC and therefore fits our data with the best model.

ARIMA Model	df	AIC
Nee	18	704.7663
Salinity	9	700.7734
Air temperature	8	660.8001

Table 1. List of the Akaike Information Criterion results from our three different ARIMA models.

Discussion

Our ARIMA model allows us to forecast and make projections of our data. We plotted NEE over time and forecasted to predict additional NEE values. We then used the salinity and air temperature data to test if it could create a better model to forecast addition NEE values and we found that air temperature fits the data the best.