Laura E. Barnes
Julianne Quinn SYS 4021 / SYS 6021

# Project 2
# Modeling and Simulating Univariate Time Series

*Assignment:* Design and test models for use in modeling air pollution data. The data set for this assignment is on the class web page under AirQualityUCI.csv. You will use this dataset to build models of ambient daily maximum nitrogen dioxide ($NO_2$) concentrations. To complete this assignment, answer the questions and follow the instructions given below. **Submit your assignment before 11:59 pm on December 7, 2020**.

*Assignment Objective:* In this assignment, you will demonstrate your ability to use time series analysis to model air quality data.

*Data:* The data in AirQualityUCI.csv provides hourly observations of ambient pollutant concentrations in an Italian city from March 2004-February 2005. You will first need to impute missing values (currently -200). You will then need to aggregate the hourly data to daily maxima using the 'aggregate' function in R. The code for both of these steps is shown in the sample Project2.Rmd file.

*Instructions:*

1. You may discuss this assignment with other students in the class. However, **work with your group** and reference any contributions from others. You must pledge your submission.
2. Use the assignment page on the Collab site for submission.
3. Perform the analyses described below using the techniques described in class to visualize and model time series. For this assignment, you will turn in a detailed R Markdown notebook in both **Rmd** and **PDF** formats with your analysis and responses to the questions below.

*Assignment:* Utilize the time series methods learned in class to develop univariate models of daily maximum ambient nitrogen dioxide ($NO_2$) concentrations.

1. **Building Univariate Time Series Models:** Build a time series model of daily maximum nitrogen dioxide ($NO_2$) concentrations using all but the last 7 days of observations. Make sure to address the following items: (50 points)
   (a) How you discovered and modeled any seasonal components, if applicable. (5 points)
   (b) How you discovered and modeled any trends, if applicable. (5 points)
   (c) How you determined autoregressive and moving average components, if applicable. (10 points)

(d) How you assessed alternative models (e.g. adjusted $R^2$, AIC, diagnostics, etc.). Assessments should discuss diagnostics and at least one metric. Show and discuss diagnostics of the residuals' homoscedasticity, Gaussianity, and independence. What problems, if any, remain in the diagnostics of the selected model? (20 points)

(e) Forecast the next 7 days of $NO_2$ concentrations using your selected model. Plot the forecasts vs. true values. What is the MSE of the 7-day forecast? (10 points)

2. **Simulating Univariate Time Series Models:** Simulate a year of synthetic observations of daily maximum nitrogen dioxide ($NO_2$) concentrations from your selected model. Set the seed so you will get the same results each time. You will need to consider the sum of the linear models of the trend + seasonality, and the residual models. Assess and compare the model's performance with respect to: (50 points)

   (a) Ability to reproduce appearance of time series. Plot observations and simulations and visually compare their characteristics. (10 points)

   (b) Ability to reproduce observed trends. You can assess this by building a linear model of the trend + seasonality of the simulations and comparing the coefficient estimates with the linear model of the trend + seasonality of the observations. What is the percent difference in the coefficient on time? (10 points)

   (c) Ability to reproduce seasonality of the time series. Analysis can be visual, simply comparing the periodogram of the observations and simulations. (10 points)

   (d) Ability to reproduce observed mean and variance of the time series (Hint: Use the functions 'mean(ts)' and 'var(ts)' where ts is a time series, and find the percent difference between observations and simulations) (10 points)

   (e) Ability to reproduce the autocorrelation of the time series. Analysis can be visual, simply comparing the ACF and PACF of the observations and simulations. (10 points)

**Formatting** Up to 5 points will be deducted for poor formatting. Two points will be deducated if text does not wrap neatly, two points if discussion of figures/models is not near them in the report, and one point if all group members' names are not listed on the assignment.