# Homework 5: More Irish Wind

36-467/667, Fall 2020

Due at 6 pm on Thursday, 8 October 2020

AGENDA: Practice extracting periodic components; practice with spatio-temporal prediction.

We continue to work with the data set `wind` from the library `gstat`.

1. *Ignoring inconvenient facts* (3) Four rows of the data set represent leap days. What are the dates of those leap days, and what are their row numbers? Create a new data set which removes those four rows, and use that data in all the rest of this homework assignment.

2. *Annual patterns*

   (a) (2) What is the average wind speed at Dublin on January 1st? (You should be averaging 18 measurements.)

   (b) (2) What are the average wind speeds at all 12 stations on January 1st? You should get a vector of length 12, each element of which is an average of 18 measurements.

   (c) (7) Find the average wind speed at each day in the calendar year, at each of the 12 stations. Your answer should have a 365 rows and at least 12 columns. (You may find it convenient to add two columns for the month and the day of the month.) Display your results by plotting these twelve annual trends. (It is ideal if you can fit all 12 annual trends into the same plot.)

   (d) (5) Describe, in words, the two annual patterns at Dublin and at Valentina.

3. *Seasonally-adjusted values*

   (a) (3) Subtract the average wind speed at Dublin on January 1st from the actual wind speeds at Dublin on January 1st, and display the results as a time series. (The new series should have 18 points.)

   (b) (4) Subtract the average wind speed at each station, on January 1st, from the actual wind speeds on January 1st. Display the results as time series. (You should have 12 series, each of 18 points.) If possible, show all the series in the same figure.

*R hint:* The `scale()` function can be used to subtract a vector from each row of a data frame, without scaling. (There are other ways to approach this as well.)

(c) (5) Create a new data frame where the appropriate trend has been subtracted from the observation for each station. What are the summary statistics for the 12 new, de-trended time series?

4. *Covariance and seasons* In this problem, use the de-trended data you prepared in problem 3c. Assume that each of the time series is stationary. When you are asked to compare to the previous homework, you can refer to the solutions, rather than to your own work, if you prefer.

   (a) (3) Report the autocovariance function for Dublin, out to a lag of 800 days. How does this compare to the autocovariance you computed in the previous homework?

   (b) (3) Report the cross-covariance function between Dublin and Shannon, out to a lag of 800 days. How does this compare to the cross-covariance you computed in the previous homework?

   (c) (5) Assuming the data is now stationary, use the data from the first nine years to estimate a model which predicts Dublin on day $t$ from Dublin on day $t-1$ and from Shannon on day $t$. This model should have an intercept and two slopes — what are they?

   (d) (5) Use this model to predict the seasonally-adjusted wind speed in Dublin for each day of 1975. Plot the predictions and the actual values as two time series (in the same plot). What is the root-mean-squared error?

   (e) (5) Using the same model, predict the *un*-adjusted wind speed in Dublin for each day in 1975. Plot the predictions and the actual values as two time series (in the same plot). What is the root-mean-squared error? Is the value you get for the RMS error surprising, or something you should anticipate, given your results in previous problems?

5. *Spatial kriging* In this problem, use the seasonally-adjusted data from problem 3c. *General hint:* The slides for lecture 9 (linear prediction over space) will be helpful; read all of them before plunging in to this problem.

   (a) (5) Using `lm`, and the first nine years of data, linearly regress the wind speed at Dublin on day $t$ on the wind speed at the other 11 stations. What are the coefficients? (There should be 12 of them; use a table or figure.)

   (b) (5) What are the correlation coefficients between the twelve stations, based on the first nine years of the data? Your answer should be a $12 \times 12$ matrix, where the $i, j$ entry is the correlation between $X(i,t)$ and $X(j,t)$.

(c) (5) Plot the correlation coefficients you just found against the distance, in kilometers, between stations.

(d) (6) Fit an exponential function to these correlations. What is the correlation length (in kilometers)?

(e) (6) Using the estimated spatial correlation function, *not* the sample covariances, find the coefficients of the best linear predictor of the wind speed at Dublin on day $t$ from the wind speed at the other stations on day $t$. What are the 11 slopes and the intercept? (Report your answer in the form of a table.) Compare these values to those in problem 5a, and explain why they are the same (if they *should* be the same) or explain the differences (if they should be different). *Hint:* Make sure Dublin isn't used to predict itself.

(f) (4) Use your model from problem 5e to give predictions for the wind speed in Dublin on each day of 1975. Plot these predictions as a time series, along with the actual (seasonally-adjusted) speeds. What is the RMS error?

(g) (5) Using the estimated spatial correlation function an assumption of spatial stationarity, find a prediction for the wind speed at Belfast for every day of 1975. Plot this as a time series.

(h) (1) Explain why you cannot find the RMS error for this last set of predictions.

6. (1) How much time did you spend on this problem set?

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. All plots and tables are generated using code embedded in the R Markdown and automatically re-calculated from the data. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.