ACMS 60786: Applied Linear Models

Final Exam

Rules

Collaboration in any form is **not allowed** in the final. **The exam is due on Monday, December 16th before noon.**

Final Exam

Welcome to the final exam. The exam comprises of two parts, one data analysis and one simulation study.

- The data analysis is divided in two subparts

    1. A **report**, 6 pages maximum for the data analysis part, same format and rules as the two midterms. The structure of the report (and the grading guidelines) are the same as in midtem 2, plus any additional analysis you deem appropriate with model selection, collinearity and use of neural networks.
    2. A **file** with one variable, called Y, with the predictions, and **your code**. The format of the file name will have to be lastname_firstname.Rdata (or whatever data format you are using). You will also have to submit your code in the format lastname_firstname.ipynb (or .R) to the TA. If the code is not reproducible, it will count as zero. Email these file to the TA (jstenema@nd.edu).

    So summarizing you will need 1) To submit a report 2) Send the prediction and the code to the TA. Both items must be received before the deadline.

    There are two data files on Canvas, one for training and one for prediction. Both come from MERRA2, the same dataset as in midterm 2. There are two data sets (more later), both focused on **winter data**. The training set focuses on Winter 2017, i.e. from December 1st, 2016 to February 28th, 2017, whereas the prediction set on Winter 2018, i.e. from December 1st, 2017 to February 28th, 2018. Both data sets comprise of a total of 90 days. The domain is a 182×101

(longitude×latitude) grid comprising the whole United States and Mexico.

In particular, I am providing the variables below. All but snowfall are the same as in midterm 2, albeit some with a different name.

- **TLML, the variable we want to predict, temperature in degrees Fahrenheit. Same as in midterm 2.**
- ALBEDO, CLDHGH, CLDMID and CLDLOW, same as in midterm 2.
- SWGDN is the surface incoming shorwave flux, in Watts per meter square. Same as in midterm 2
- PRECTOT, the total precipitation, in Kg per meter square per second. Same as in midterm 2.
- QLML, the specific humidity, unitless and a ratio betrween 0 and 1. Same as in midterm 2
- SPEED, wind speed in meters per second, an example of the wind fields for the first day over out grid is plotted in Figure 1. Same as in midterm 2.
- PRECSNO, the snowfall rate, in Kg per meter square per second.
- lon (lat) is a vector of length 182 (101) which has the values of longitude and latitude in the grid. These variables are not for building a model, but for making plots.

As mentioned before, there are two data set: a **training set** and a **prediction set**. In the training set, you will have to build your model for the temperature in South Bend, which you can find at longitude 119 and latitude 84, so that to extract it you'd have to enter Y=data_training$TLML[119,84,]. Your goal is to find the **best and simplest model**. You can do whatever you want to find it: use some or all the variables in South Bend, temperature at other locations, etc. Interpretation (both statistical and physical) is critical, so a choice of a model without a thorough explanation will be penalized.

Once you are satisfied with your model, you will have to load the **prediction set**. You will notice that I have removed the temperature in South Bend (if you try to load it, you'd find NaNs). Come up with your predictions, and send them along with your code.

- The simulation study has a 4 pages maximum. You will have to answer a single question: which one is the best model selection criterion for a linear regression model? Design your own simulation study to make your point. You have full independence on choice of the setting, think carefully about how would you choose a criterion against another one.

Good luck!

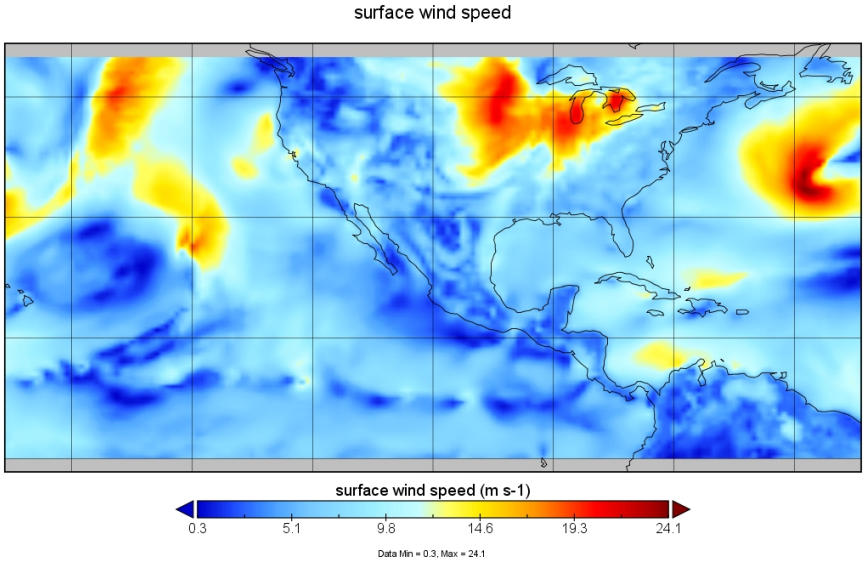

Figure 1: Wind speed (reference height 2 meters) for December, 1st 2016. Units are meters per second