Notebook: EDA for TX state weather data.

Jun 27th, 2018

Overview

This is a R Markdown notebook for exploratory data analysis on a subset of the weather data collected across airports (weather stations) in the United States. The data comes from NCDC NOAA Local Climatological Data (LCD); the codes reproduced here explore a subset of 47 different station locations in Texas state, collected in the years 2015-2017. While most of the observations were made hourly, slight irregularities exist between time intervals, mostly due to different types of forecast bundled together. Here, for simplicity the few disparities across different points in the same hour are averaged. The response variable of interest here is hourly wind speed (in MPH).

Packages required

```
library(tidyverse)
library(GGally)
library(lubridate)
library(reshape2)
library(sp)
library(spacetime)
library(maps)
library(maptools)
library(geosphere) #Compute distances on WGS84 ellipsoid (long/lat)
library(gstat) #Spatial analysis (kriging) tools
library(mgcv) #GAM fitting
```

In addition, a number of user-defined functions are used in this report. Their codes can be sourced from a separate file called `000_utils.R`. User-defined functions have suffix `_fn`; global user variables have prefix `g_`. In the same vein, variables that may be of temporary use only or to be repeatedly changed have prefix `c_`.

Loading the data

Currently, the entire subset of data used here is stored as a binary file storing R data frame object. The data frame is read from a CSV file, which is a merged product of individual CSV files for each weather station provided by NCDC website.

Preprocessing / Preparatory Work

Some preprocessing to the raw data format is necessary. Important information to be added to the table below includes:

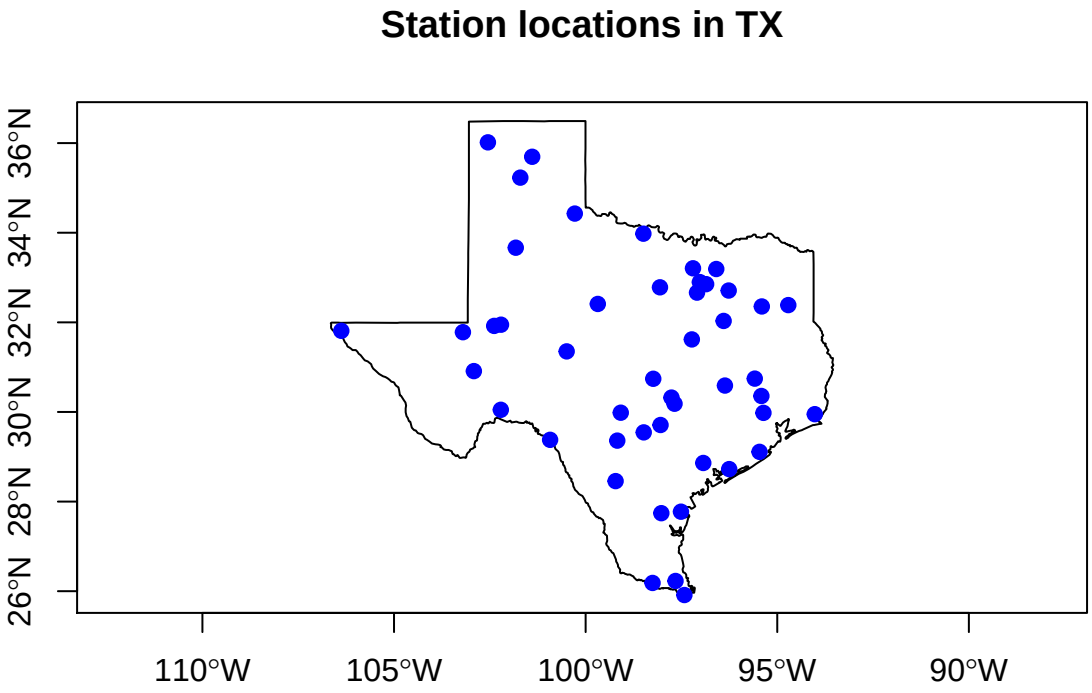
- Hour of the day
- Day of the year (indexed from 1 to 365/6).

A separate data frame object that stores station-particular information, i.e., WGS84 coordinates, station name, and elevation (in meters), will be also useful for future analysis. Note that, against regular geographic conventions, longitude comes in front of latitude, following the orders of Euler's angle notation.

```
TX_df <- TX_df %>%
  mutate(HOURLYWindSpeed = as.integer(HOURLYWindSpeed),
         ## type conversion
         Hour = lubridate::hour(DATE),
         ## hour of the day
         Day = lubridate::yday(DATE))
## Warning in evalq(as.integer(HOURLYWindSpeed), <environment>): NAs
## introduced by coercion
         ## day of the year

# Encode station-unique info a data frame
stationInfo_df <- data.frame(
  long = unique(TX_df$LONGITUDE), ## Longitudes
  lat = unique(TX_df$LATITUDE), ## Latitudes
  name = unique(TX_df$STATION_NAME), ## Station names
  elevation = unique(TX_df$ELEVATION) ## Elevation in meters
)
```

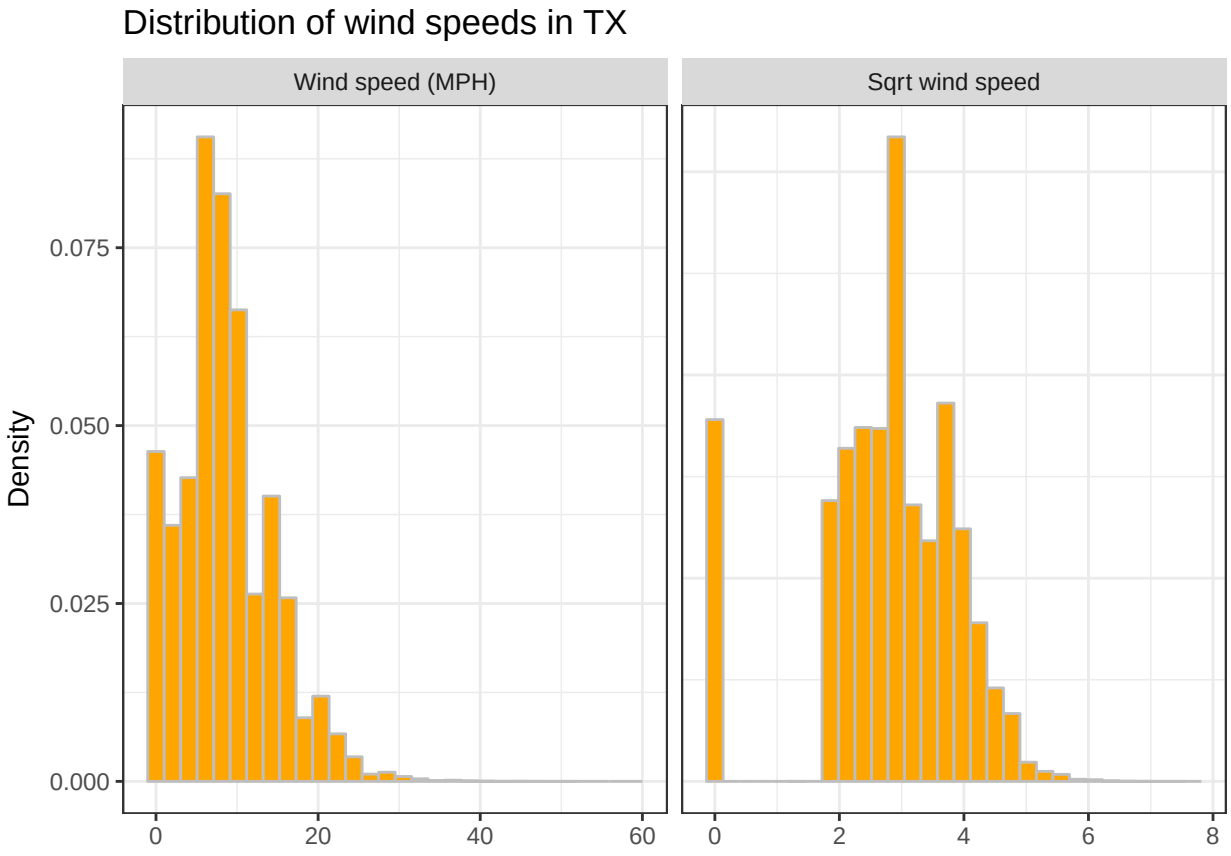
The locations of each 47 different stations are overlayed onto the map of Texas state below.



1. Distribution of wind speed

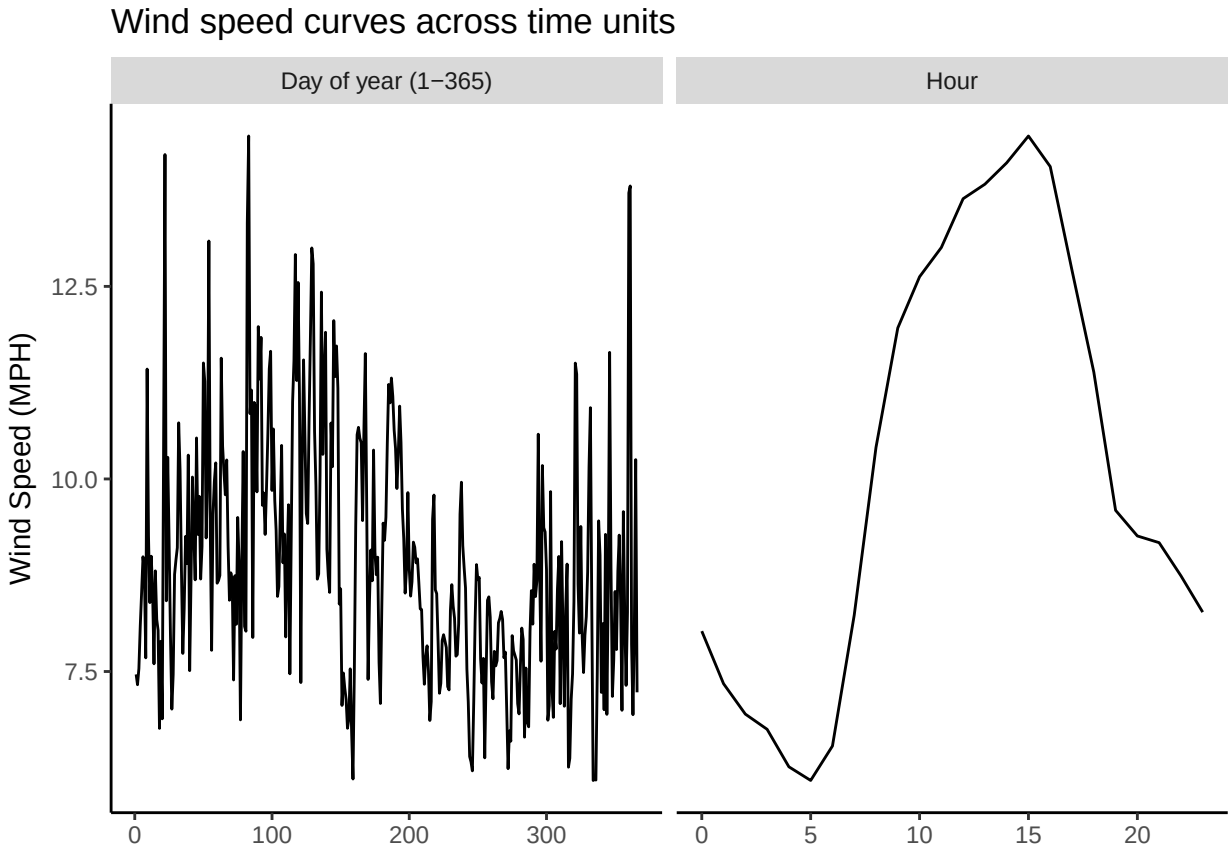
It is not hard to expect that the distribution of wind speeds is heavily skewed by extreme large values and have a wide range. Square root transformation of the distribution looks much more close to being symmetric around the center, but now the many zero values at the lower tail end heavily affect the shape of the distribution. Future modelling attempts may or may not need to account for the transformation of response depending on their distributional assumptions.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 64142 rows containing non-finite values (stat_bin).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 64142 rows containing non-finite values (stat_bin).
```

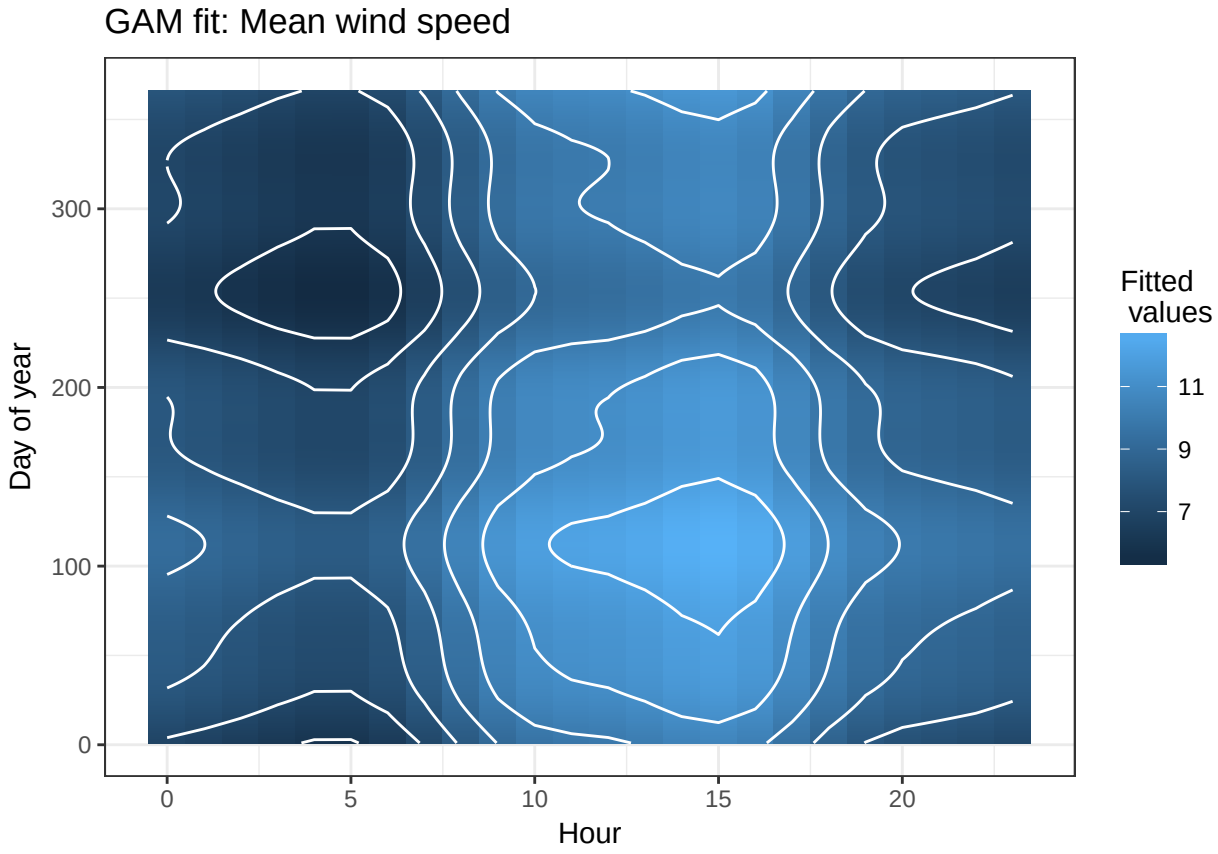


2. Diurnal patterns: Mean trend of intraday wind speed curves

It is also easy to expect that long-term seasonality exists for wind curves. Plotting the run-sequence of daily averages of wind speed identify changes in both maximum and minimum wind speeds and the amount of variations. Less certain is the "average" wind curve acorss each hour within a day. Plotting hourly averages of wind speed across different days captures some recurrent patterns, but the variance across different wind speed curves is very high (not shown here).

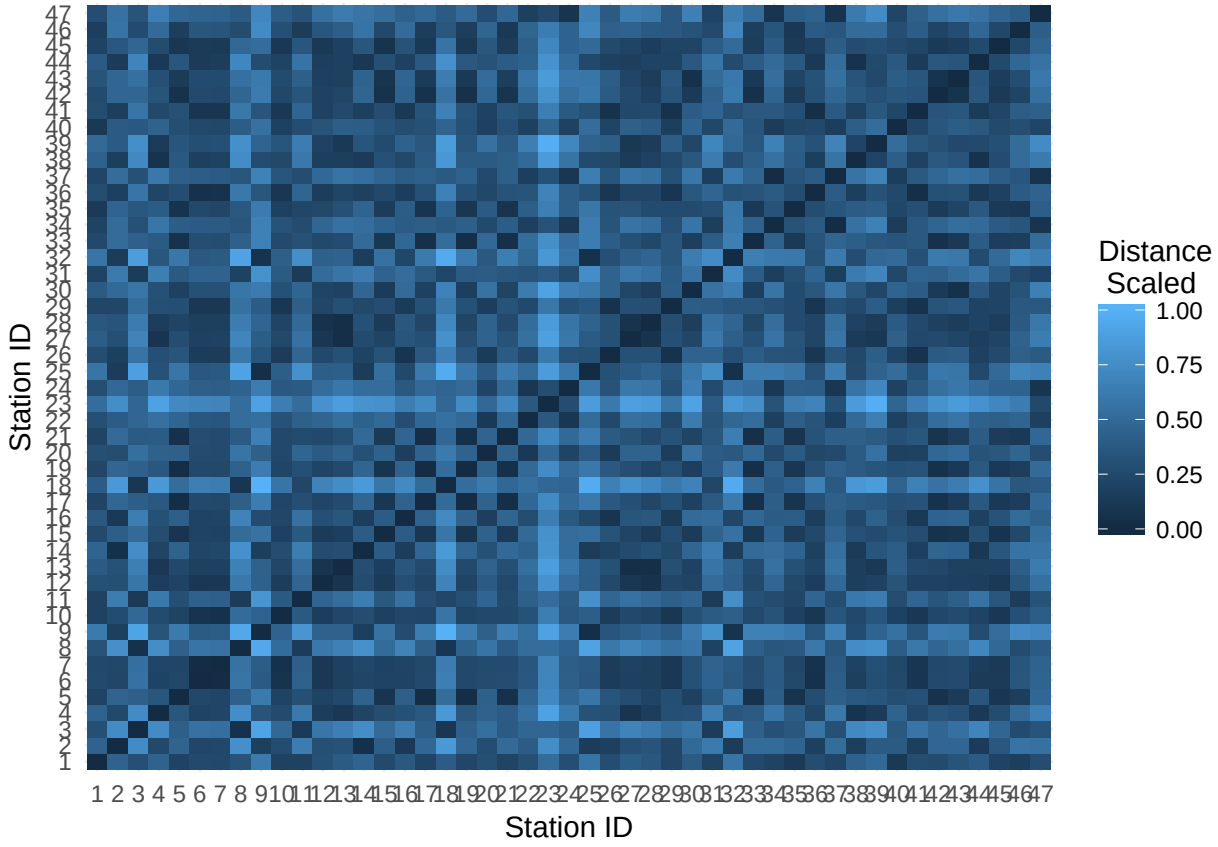


Since statistical modelling for wind speed predictions attempt at moddeling short-term variations over a narrow time window, ranging from few hours to a day, obtaining information about the mean trend of an "average" intraday wind curve can be of interest. Temporal or spatiotemporal predictive models then can subtract out the mean trend from the original series and estimate the latent residual process. A GAM model is fitted to the wind speed against two variables: t, the hour of the day, and d, the day of the year, indexed as an integer from 1 to 365(6). A heatmap plot of intraday wind curves against the hour and the day is produced below. The idea and the practice, including the use of package `mgcv`, have been already described in *Tupper, Matteson, and Anderson, 2015*.



3. Distances and Correlations

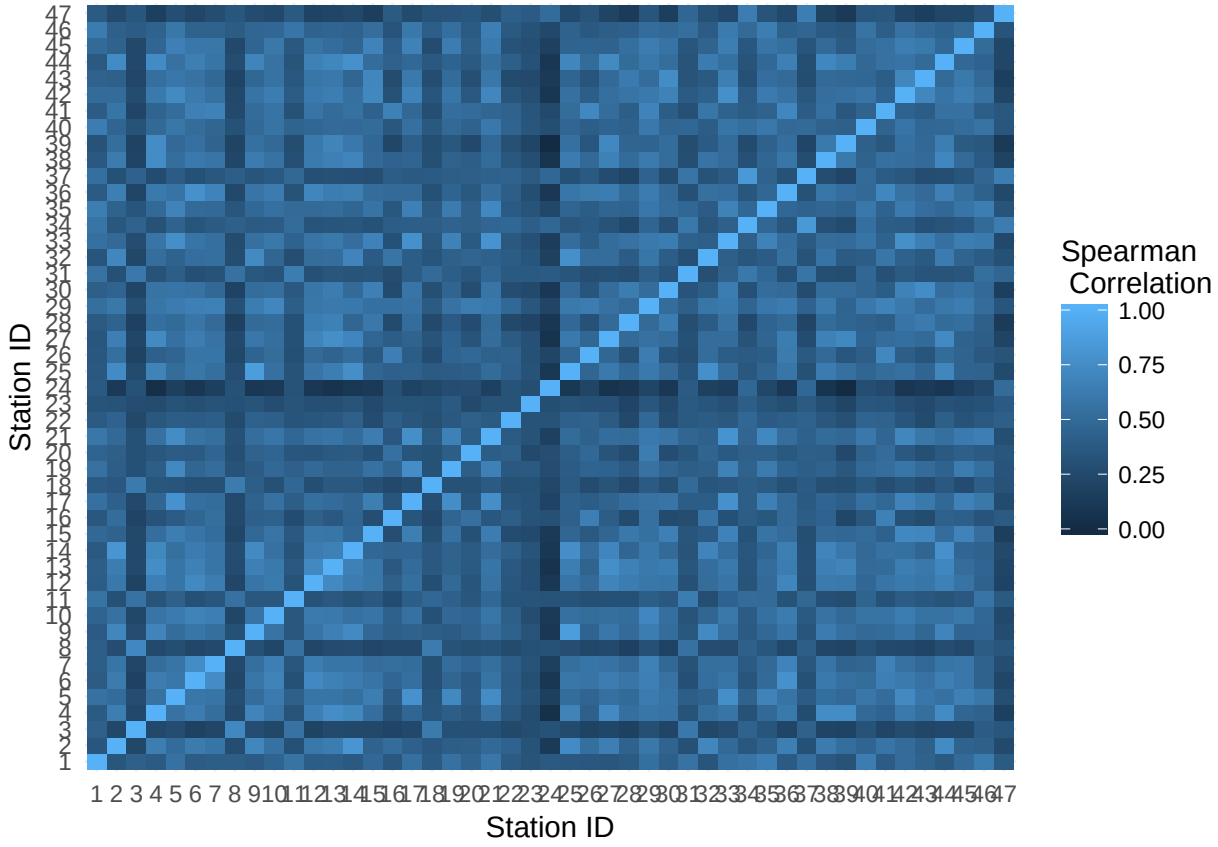
Another interesting aspect of spatial dataset is the relationship between distances and correlations between response of interest. First, a distance matrix between 47 sations is produced. Distances are normalized to take values between 0 and 1.



Station ID 23 is on average the farthest removed from any of the other airports. It is El Paso International Airport, close to the border to Mexico and located in the Western tip of Panhandle region; it is also one of the higher stations, located 1194.2 meters above ground.

Now, correlation matrix is more tricky. Below there are raw codes that include some data wrangling to identify locations and time periods for which there exist missing values. The missing values occur in Kerrville Municipal Airport for three hours in a leap year: Feb. 29th, 2016.

```
## Adding missing grouping variables: `Day`
```



It is notable that the information visualized in the distance matrix and the correlation matrix do not necessarily align. For example, El Paso International Airport do not seem to be an outlier from this plot. On the other hand, ID 24: Fort Stockton County Airport, and ID 47: Winkler County Airport, are on average the most uncorrelated with any of the other airports, despite their not being especially identifiable on the distance matrix. Their elevations also do not seem to be sufficient explaining features for such uncorrelatedness. One possible explanation can be particular geographic features of their locations; Panhandle region of Texas is known to have much wind, and the abundance of valleys and peaks may introduce much more variance into wind patterns relative to other stations.