Project 2. Spatial Analysis of Twitter Data

Abstract

This report details the experiences, successes, and difficulties in conducting this project. This report will aim to summarize the results of writing csv data to a shapefile, plotting the data to a map, and manipulating the data for marginally more interesting graphical representations.

1. Section 1: I/O Operations

Much in the same way as Project 1, the data was read in from the provided csv file. Unlike Project 1 though, the data did not stay in csv format for long. The task at hand was to write the csv to an SP and SF object followed by writing both to a shapefile and one other geo enabled format. To accomplish what needed to be done for this portion of the project, the SP, SF, and rgdal libraries were used.

The SP object was straightforward to initialize. The coordinates were assigned to a single variable by using cbind on the longitude and latitude values followed by a projection string being defined. Each of these were passed to SP's SpatialPointsDataFrame function along with the csv data in order to create the SP object. The SF object proved to be even simpler to create but required comparable information. The st_as_sf function from SF was used here with a CRS of 4326 established.

When writing these files to their respective folders, I set up a chunk of code to remove the shapefile and all associated dependencies from the folder. This seemed clunky at the time and later proved to be so when it was discovered that there was in fact an overwrite option in the writeOGR and st_write functions for SP and SF objects respectively when writing these objects to a file. In order to involve one other geo enabled format, I wrote the files to a geoJSON format given its wide applicability in web applications.

2. Section 2: Plotting and Data Manipulation

Relatively few issues had occurred up to this point in the project. Working with the mapping libraries proved to be somewhat difficult at times but not all together insurmountable. It was difficult at first to understand exactly what the tmap syntax was but after some tinkering it was discovered that a tm_shape function needed to be called with my SF object of choice followed by a map element (like tm_dots) needed to be added to the data. The initial attempt at plotting isn't shown as it was not very useful.

In order to tease out more information about what was occurring spatially, the tweets were plotted by month, day of week, and hour. These plots can be seen in **Figure 1a, 1b, and 1c**.

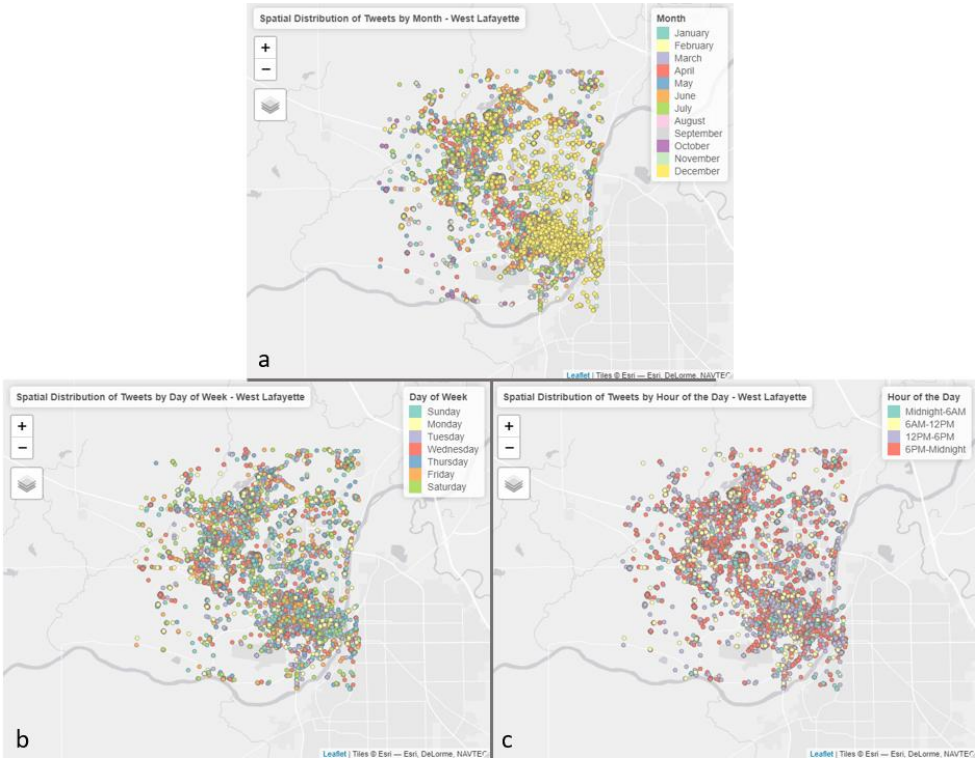

Figure 1a,b,c: West Lafayette tweets plotted by Month, Day of Week, and Hour of the Day respectively. Hours have been aggregated into 4 time periods for simplicity.

When presented with the opportunity to plot the twitter activity of 3 individuals, I decided to focus on the top 3 by number of tweets. This was to determine whether these tweets were truly from bots or if they were mobile in any way. The result was somewhat surprising, these users were quite mobile! The results can be seen in **Figure 2**.

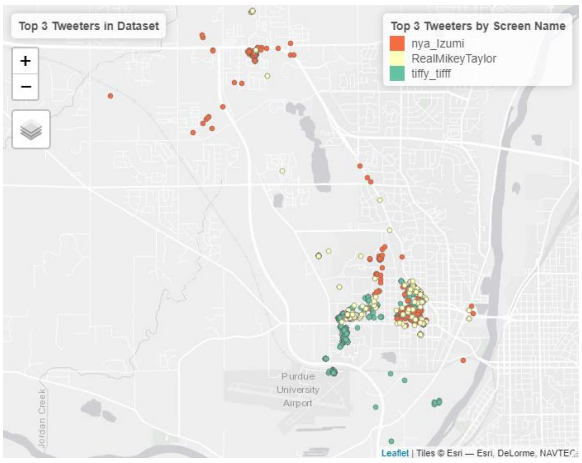

Figure 2: Top 3 users by tweet frequency.

3. Section 3: Hotspots & Spatial Temporal Distribution of the Top Three.

This section was particularly interesting as it provided the opportunity to leverage some deeper features in tmap. In particular, in order to show some of the hotspots on campus, the clustering method in tm_dots was used alongside of the dots themselves. This showed significant activity on campus and less so as you moved to the outer edges of campus. The result was sensible due to the daily activity on Purdue's campus, the result can be seen in **Figure 3**.

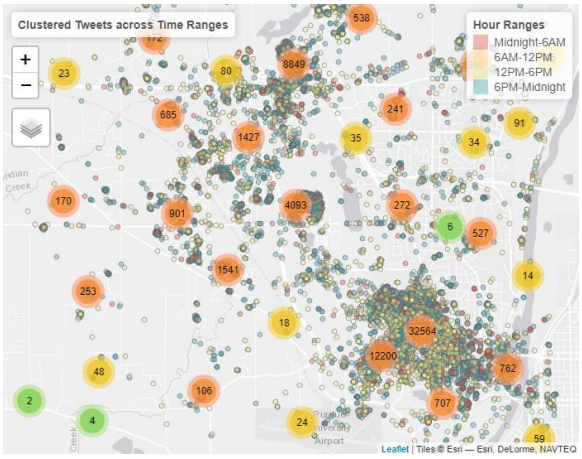

Figure 3: Hotspots across campus using tmap's clustering.

Finally to describe the spatio-temporal distribution of the data, the lattice package was put to use to show these top 3 users across time with their distance traveled in meters in **Figure 4**. Interestingly, there are large gaps in these users twitter usage.

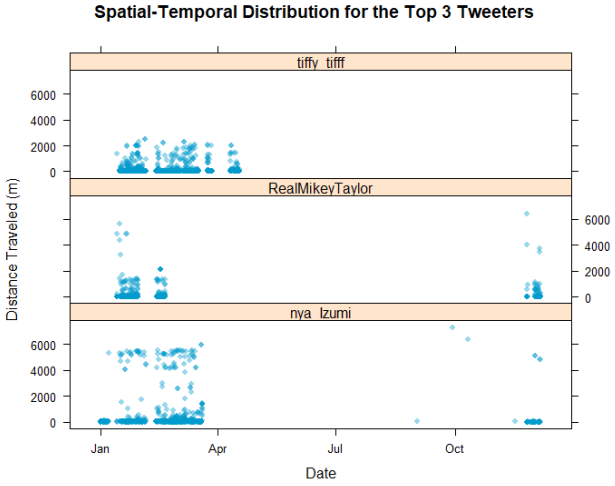

Figure 4: Travel distance patterns across time for the top 3 most frequent tweeters.

4. Section 4: Conclusions

This project proved to be challenging in many respects, but it has provided excellent experience for working with the packages used going forward. There was some work that went into the section 1.6 extra credit. The resulting visualization can be seen in the appendix below. Further refinement of this project could be had by employing some of the clustering methods discussed in class.

Appendix

Figure 5: Extra Credit - Tweet locations averaged for the top 3 users and colored by time range.