Project 4. Polygon & Point Data Analytics

Abstract

This report will detail the process of developing custom functions for some common polygon analytics to be applied to the provided campus building shapefiles and an account of performing common GIS oriented tasks with R. Difficulties will be shared as applicable.

1. Cleaning up the Buildings Data Shapefile

The provided buildings shapefile, as with most data, was in dire need of some cleaning at the start of this analysis. There were invalid geometries, empty polygons, buildings that were long gone, and buildings that didn't seem to be correct. Fortunately, sf provides a handful of tools for locating these issues quickly. "st_is_valid", "st_is_empty", and R's built in "is.na" provided all the tools needed to clean up the dataset. Pre- and post-cleanup can be seen in Figure 1a and 1b.

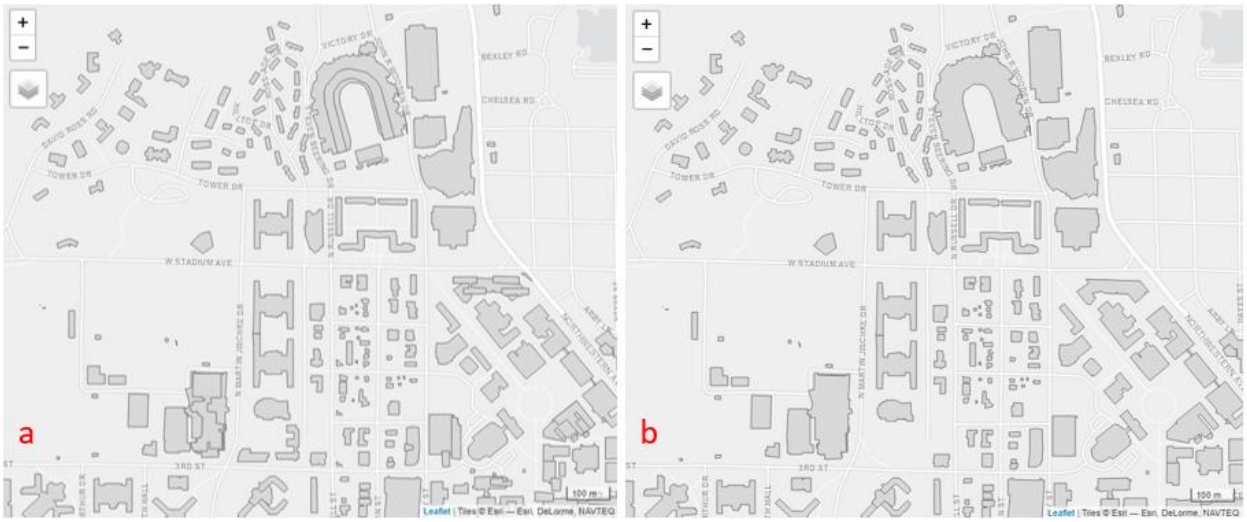

Figure 1a, 1b: Pre-cleaned and post-cleaned building shapefiles respectively. Map is zoomed in for clarity.

2. Polygon Analytics

Development of custom functions to perform polygon analytics, like perimeter, area, and, centroid calculations, was surprisingly easy in some ways but provided some unexpected challenges. Despite many attempts at working out how to develop "vectorizable" functions to use on the building shapefile the final code that was settled on for this analysis used for loops. Despite the performance degradation from operating element by element, the analytics were still very quick. Larger datasets would require this code base to be re-written. The

formulas used for perimeter, area, and centroids can be seen in the equations provided in lecture.

The part that presented the most difficulty in my computation was navigating the geometries. The geometries are presented as list items and it takes a fair number of brackets to drill down to the matrix containing the x and y components for each polygon. After a fair amount of trial and error to determine the correct bracketing in R, the code was correctly accessing each Polygon. Originally, a check was put in place to deal with multipolygons, but after further cleaning (as presented in section 1) the multipolygons were no longer an issue and thus the check was abandoned in the centroids equations.

As for the accuracy of the custom-built functions when compared with R's built in functionality for determining the perimeters, lengths, and centroids, the custom functions fared quite well. After rounding both sets of results to three decimal places and running summary against my solutions minus R's solutions, the min, max, 2nd/3rd quartile, and median all evaluated to zero pointing to my solutions being spot on.

3. Tweet Data in Light of the Building Polygons

Upon bringing the tweet data into this reports analysis, it was established as an sf object and transformed to the same coordinate system provided by the buildings shapefile, that is the Transverse Mercator projection with the NAD83 Datum. Once this was done, the data could be accurately compared. Without this step, the tweet data could not be compared with the buildings shapefile data. In order to match the tweet data and the buildings data, a buffered buildings dataset was created with a 30ft buffer using sf's st_buffer function. This does introduce potential overlap in the buildings, leading to double counted tweets. This uncertainty is acceptable as it is uncertain when people are outside a particular building whether or not they should go with the overlapping buildings in question. Quantifying the uncertainty in the tweet count numbers would be an important step if this work were to be taken to publication. Figure 2 shows the tweet counts by building.

Figure 2: Tweet count by main campus buildings.

The next step in the analysis required the association of tweets with their closest building. This was quickly accomplished with the st_nearest_feature function. Buildings near Stewart Center were selected for demonstration as can be seen in Figure 3.

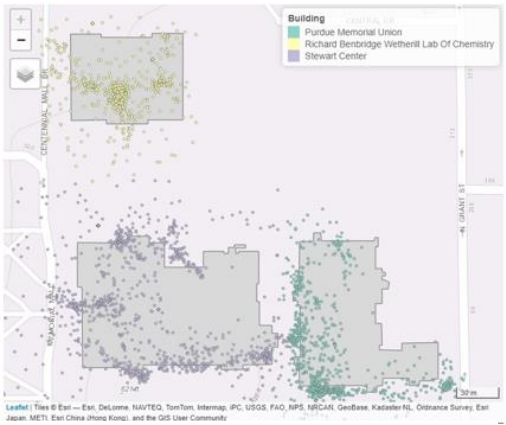

Figure 3: Tweets by building.

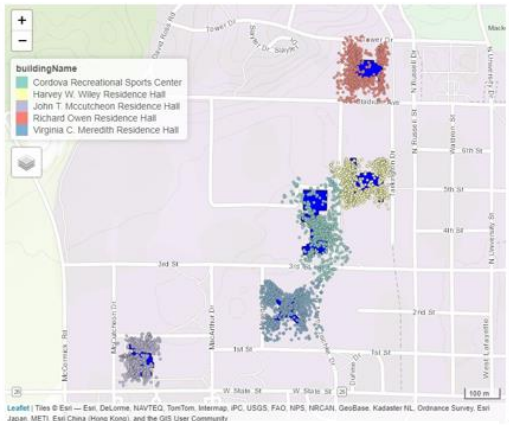

Figure 4: Top 5 most popular buildings by tweet frequency.

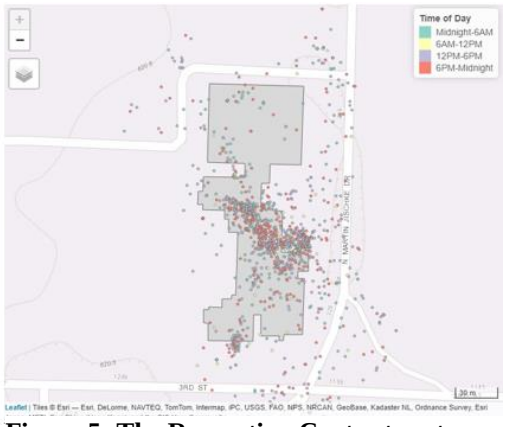

Figure 5: The Recreation Center tweets characterized by time of day.

Assigning the tweets to each building provided a fairly straightforward and simple way to determine the top buildings. The dataframe containing tweet frequency for each building was sorted and the top 5 buildings were taken. The same buildings were selected from the tweet data and were co-mapped with tmap, coloring by the building name as can be seen in Figure 4. It was determined that the busiest building was the Recreation Center. The results for the top 5 busiest buildings can be seen in Table 1.

Finally, similar to what was performed in Project 2, the hours for each tweet were categorized into categories and mapped. The recreation center seems to see the most tweets in the 6PM to Midnight time frame as can be seen in Figure 5, which makes intuitive sense.

Table 1: Top 5 buildings by tweets associated with that building.

Building	Tweet Count
Cordova Recreational Sports Center	2,389
Richard Owen Residence Hall	2,197
John T. McCutcheon Residence Hall	1,655
Virginia C. Meredith Residence Hall	1,486
Harvey W. Wiley Residence Hall	1,463

4. Concluding Remarks

This project provided an excellent opportunity to get much more familiar with R's GIS capabilities. Given more time in the future, I may revisit this project to improve my code performance and leverage the vectorized operations available in R to perform computations. It proved to be true that data preparation takes the longest in data analysis. Thanks to Professor Shan for the provided datasets.