Adrian Seemangal

Models and Methods for Spatial Data Science

Friends from Afar: Assessing Virtual Space as a Metric of Physically Spatial Processes

The famous Tobler's First Law (nearer things are more related than further things) has been the foundation of spatial analysis since its inception. For most of human history, behavior, resources, and connections have been instrumental in the structure of one's life. As man progressed, innovation naturally tended to seek to decrease the importance of space. Taming horses made the world smaller, as did ships, as did trains, as did planes, as did automobiles , as did the telephone. Then, a seemingly drastic change to the notion of "space" and "closeness" occurred with the conception and development of the internet, which challenged many geographical ideas because of the "death of distance" (Wang et al. , 2018). Relatively suddenly, it was just as easy to communicate with someone hundreds of miles away as it was to communicate with a neighbor.

With the rise of social media, there are easy metrics to track one's connectedness, or number of relationships, in virtual space. Whereas physical connectedness may be defined by how many people one knows on a first-name basis, which may be a difficult metric to measure, at best being an approximation, now one's virtual friends and followers are displayed on the home page of their social media profiles, and serve as a new metric of closeness in virtual space (Han et al., 2018).

Question

In light of the importance that virtual connectivity has in modern-day society, the study seeks to determine if virtual closeness can be a metric of what are typically considered highly-spatial processes.

Focus Area

The study focuses on the east coast of the United States–it consists of all states with a coastline, plus some that are intrinsically considered "east coast" states, both culturally and geographically. These states include Florida, Georgia, South Carolina, North Carolina, Maryland, District of Columbia, Virginia, Delaware, New Jersey, Pennsylvania, New York, Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, and Maine.

Armed Conflict Location & Event Data Project (ACLED)

The Armed Conflict & Event Data Project (ACLED) is a real-time database on political violence, armed conflict, and protest data globally. It collects disaggregated events, meaning the data is totally discrete, oftentimes differentiated by location, type of violence, or time. This makes it useful for spatial analysis because it is highly granular. Furthermore, it focuses on group activity, not the activity of individuals. Luckily, the United States is not known for internal armed conflict or political violence, but the protections on freedom of speech, and the right to protest, do mean that there are regular protests happening in all parts of the country. This makes ACLED a useful source for monitoring them, historically, and in near real-time (ACLED, 2017).

Facebook Social Connectedness Index (SCI)

The Facebook Social Connectedness Index (SCI) is an index that measures connectivity between two areas based on how many Facebook friends the areas share. The index ranges from 1, to 1 billion. It is unitless, making it a good metric to add to any analysis dealing with personal relationships between two areas. It is calculated by dividing the number of Facebook friends between two areas by the number of users in the origin multiplied by the number of users in the destination. It can be thought of as a metric that measures the probability of two Facebook users in two locations being Facebook friends. The smallest unit of the SCI is from county to county, noted by five-digit FIPS (Federal Information Processing Standards) codes for each location. There is also country to country, and US county to country (Data for Good, n.d.).

Specific Inquiry

The study will determine if areas with the highest overall SCI values also experienced the highest volume of protests. The idea is that a high SCI value may indicate more of a secondhand, but non media-driven awareness of issues like social injustice or political changes, and may also lead to more protests in that area. People may also flock to areas with higher SCI for protests as they may be gathering points due to high degree of connectivity. The study will determine if the protest events are clustered, it will determine the mean center of the protest events, and then it will compare these results to clusters of high SCI values, and see if they tend to match.

Methods

Python Packages

Data Munging, Analysis, and Acquisition, and Conversion

The numpy package was used to generate random seeds needed for point pattern analysis**.** The pandas and geopandas packages were used for data munging and various types of

aggregations like merging, grouping, concatenation, and spatial joins. The datetime package was used to convert the date to a format that could be used easier in temporal analysis. The cenpy package was used to communicate with the Census API and download spatial data at the county level for all states in the area of interest.Shapely was used to convert coordinate pairs into point and polygon geometries.

Data Visualization

The matplotlib, seaborn, and contextily packages were used to create plots and maps. Data visualization was an integral piece in all parts of the project, not just to communicate results. Generally speaking, most of the visualizations were spatial, being plotted from GeoDataFrames.

Spatial Analysis

Packages in the Pysal suite were used for spatial analysis. ESDA was used to create Local Moran's I statistics and plots. Pointpats was used to conduct point pattern analysis and determine if the data were clustered, and also to determine the center of the point data.

ACLED Usage

Preprocessing and Munging

ACLED does have a REST API to programmatically interface with their data, but connection issues made it difficult to actually implement this step. Instead, the ACLED Data Export Tool was used to select events and locations of interest (in this case, protests in the United States for the entire year of 2020). The resulting CSV file contained many (31) columns, including, but not limited to event date, event type, latitude, longitude, source, and a notes

section column to describe the protest event. For this study, the columns of interest were data ID, event type, state, latitude, longitude, and timestamp. Then, the data had to be filtered down to only states of interest. There was one mislabeled event, which was removed from the data. The next step was to create a jointplot to simultaneously visualize the protest events and get an idea of their distribution.

Point Pattern Analysis

A CSV file containing the geometries of every county in the area of interest was loaded into a geopandas GeoDataFrame. Then, these geometries were dissolved into one large polygon that represented the entirety of the area of interest using shapely's cascaded union method. This serves as the window for the point pattern analysis. Then, the longitude and latitude columns of the protest data were turned into a numpy array from which a point pattern was created. A summary of the point pattern was created. After that, multiple distance functions (G, F, J, K, L) were calculated on the point pattern using the respective modules of the pointpats.distance_statistics module, all of which were visualized to determine if the points were clustered or not. Furthermore simulation envelopes were calculated at 1000 simulations of complete spatial randomness generation from a poisson point process for the G and F functions. K and L envelopes proved to be too computationally intensive to be run in this study. The mean center of the points was also determined using the mean center module of the pointpats.centrography package (Rey et al., n.d.).

Facebook Social Connectedness Index Usage

Preprocessing and Munging

The Facebook Social Connectedness Index (SCI) was downloaded as a CSV file from the Humanitarian Data Exchange. It consists of three columns: user location, friend location, and the scaled SCI value between the two locations. The locations are denoted by FIPS codes. There is an SCI value from one location to every other location in the area of interest. The SCI was divided by states based on the first two digits of the user FIPS code. The same county geometry CSV file was read into a DataFrame. Then, the FIPS codes were converted to strings, and each DataFrame was merged with their corresponding SCI value. Now, the SCI value is attached to a geometry and spatial analysis can be performed. The SCI values were converted from strings to integers so that aggregation steps could be performed. All the state DataFrames, containing their SCI values, were concatenated into one large DataFrame. Then, the total SCI value (meaning, the absolute connectedness) for each was summed for each county and merged onto the DataFrame.

Exploratory Steps

The data was visualized with GeoPandas to view the overall connectedness of each county and see if there may be an uneven distribution of connectedness.

Hot Spot Analysis

Queen weights were derived for the area of interest using the Queen module from the libpysal.weights package. Then, local Moran's I statistics were calculated based on the total SCI of each county, and cluster maps and Local Moran scatter plots were generated.

Integration

The dispersion and center of the points was compared to hot spots in SCI to see if places with a high SCI value also had a high volume of protests.

Results

Protest Insights

	data_id	event_type	state	latitude	longitude	timestamp	geometry	date
2	8679807	Protests	North Carolina	35.6010	-82.5540	1638920199	POINT (-82.55400 35.60100)	2021-12-07 23:36:39
6	8681115	Protests	District of Columbia	38.9381	-77.0451	1638920207	POINT (-77.04510 38.93810)	2021-12-07 23:36:47
7	8681184	Protests	Pennsylvania	40.4406	-79.9958	1638920208	POINT (-79.99580 40.44060)	2021-12-07 23:36:48
8	8681227	Protests	North Carolina	35.9940	-78.8986	1638920208	POINT (-78.89860 35.99400)	2021-12-07 23:36:48
9	8681256	Protests	Virginia	37.5388	-77.4336	1638920208	POINT (-77.43360 37.53880)	2021-12-07 23:36:48
...	...	...	...	...	...	...	...	...
21529	7616952	Protests	North Carolina	36.0999	-80.2442	1612546518	POINT (-80.24420 36.09990)	2021-02-05 17:35:18
21531	7617464	Protests	New York	43.1547	-77.6155	1612546519	POINT (-77.61550 43.15470)	2021-02-05 17:35:19
21534	8265777	Protests	Florida	26.7097	-80.0642	1624483201	POINT (-80.06420 26.70970)	2021-06-23 21:20:01
21537	7616799	Protests	New York	43.1547	-77.6155	1612546518	POINT (-77.61550 43.15470)	2021-02-05 17:35:18
21539	7617409	Protests	New York	40.7834	-73.9663	1612546519	POINT (-73.96630 40.78340)	2021-02-05 17:35:19

8388 rows × 8 columns

Table 1: A look at the filtered protest data.

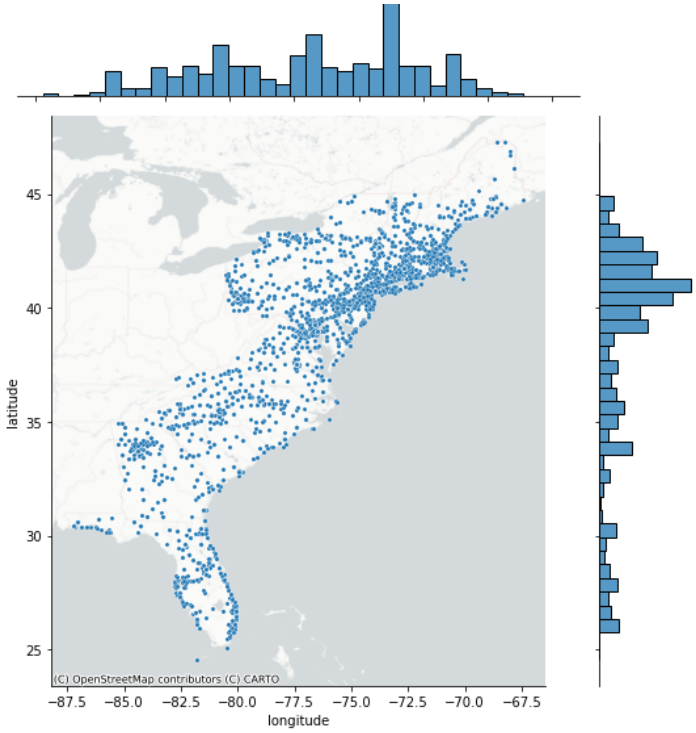

Figure 1: Joint plot of points, showing some possible clustering

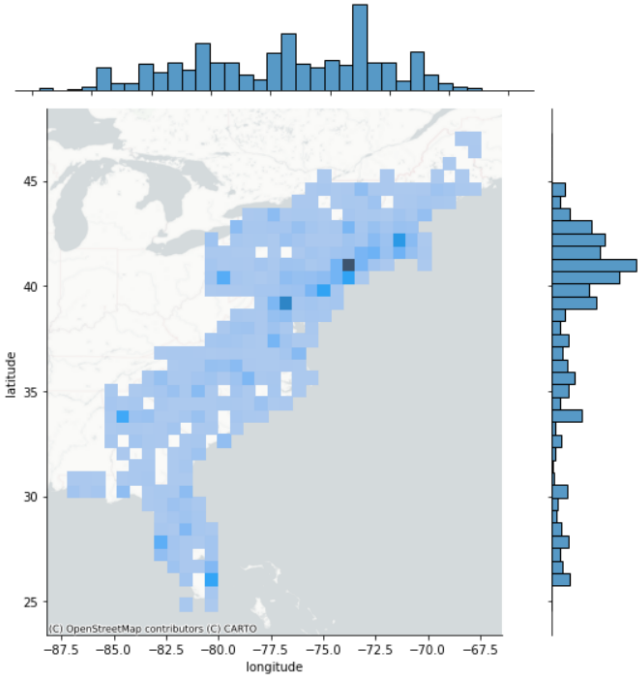

Figure 2: Protest Data aggregated by histogram, shows possible clusters.

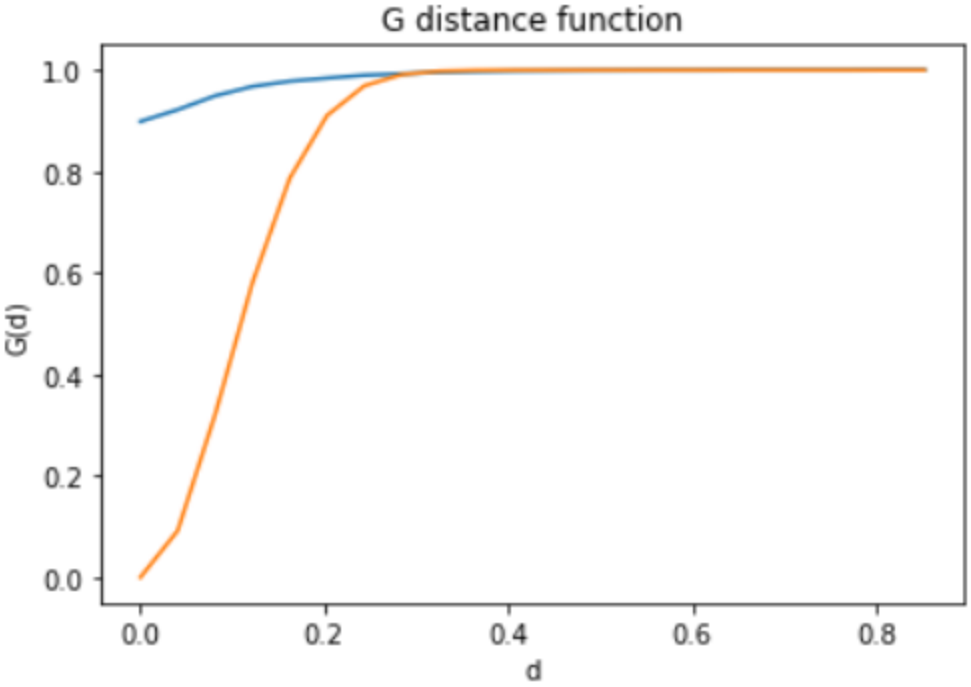

Figure 3: G distance function, blue line is above orange line, showing clustering.

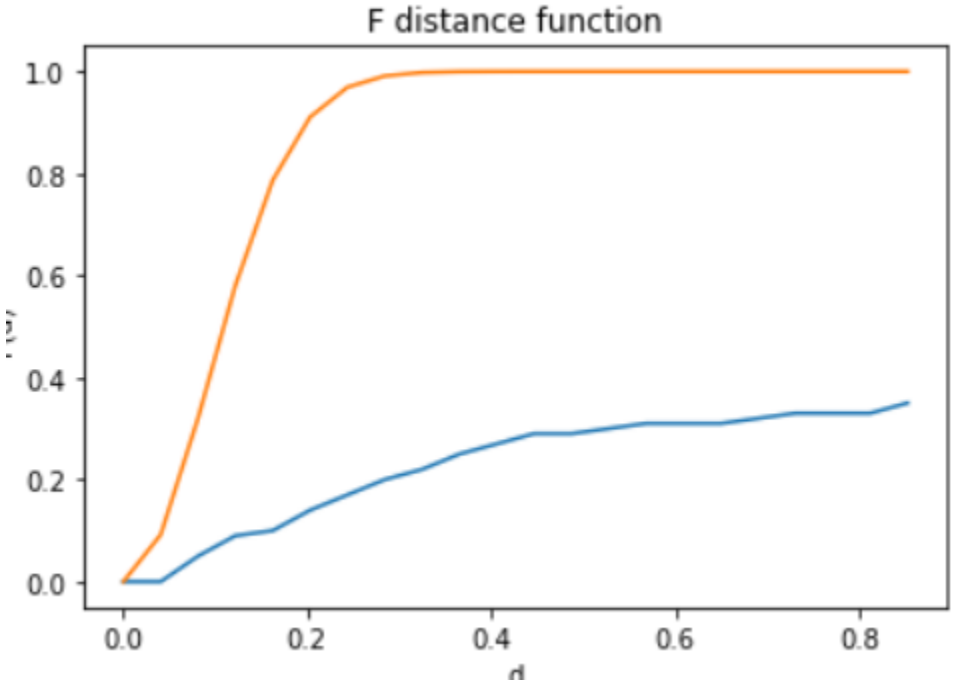

Figure 4: F distance function. Blue line is BELOW the orange line, showing clustering.

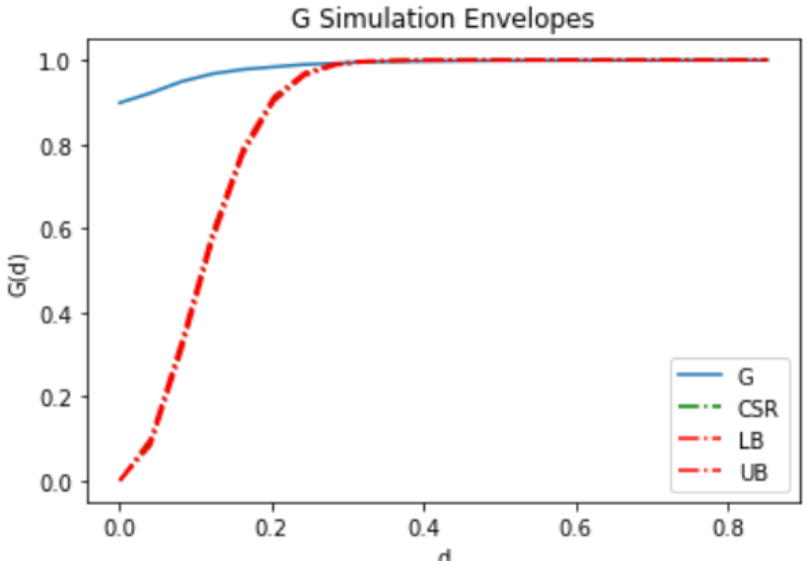

Figure 5: G Simulation envelope.

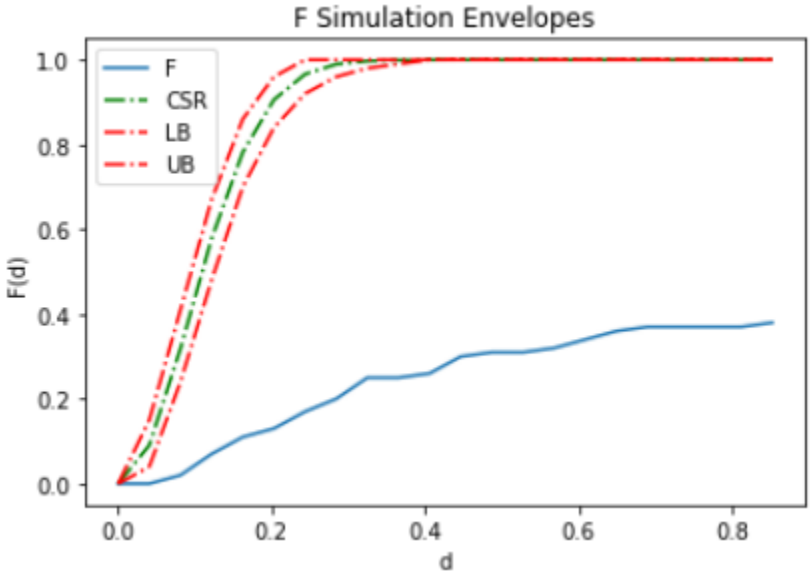

Figure 6: F simulation envelope

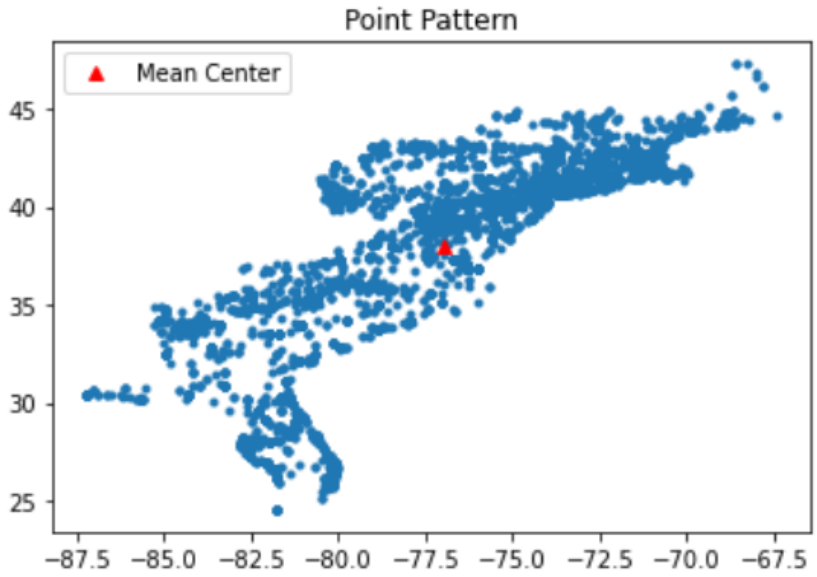

Figure 7: Mean center of point pattern

The point pattern of protests for the area of interest was proven to be clustered under every distance function. When simulation envelopes were generated for the G and F function, they also showed that the points were clustered. In other words, the distribution of the points was not completely spatially random. The mean center of the points was near the border of southern Maryland and Virginia.

SCI Insights

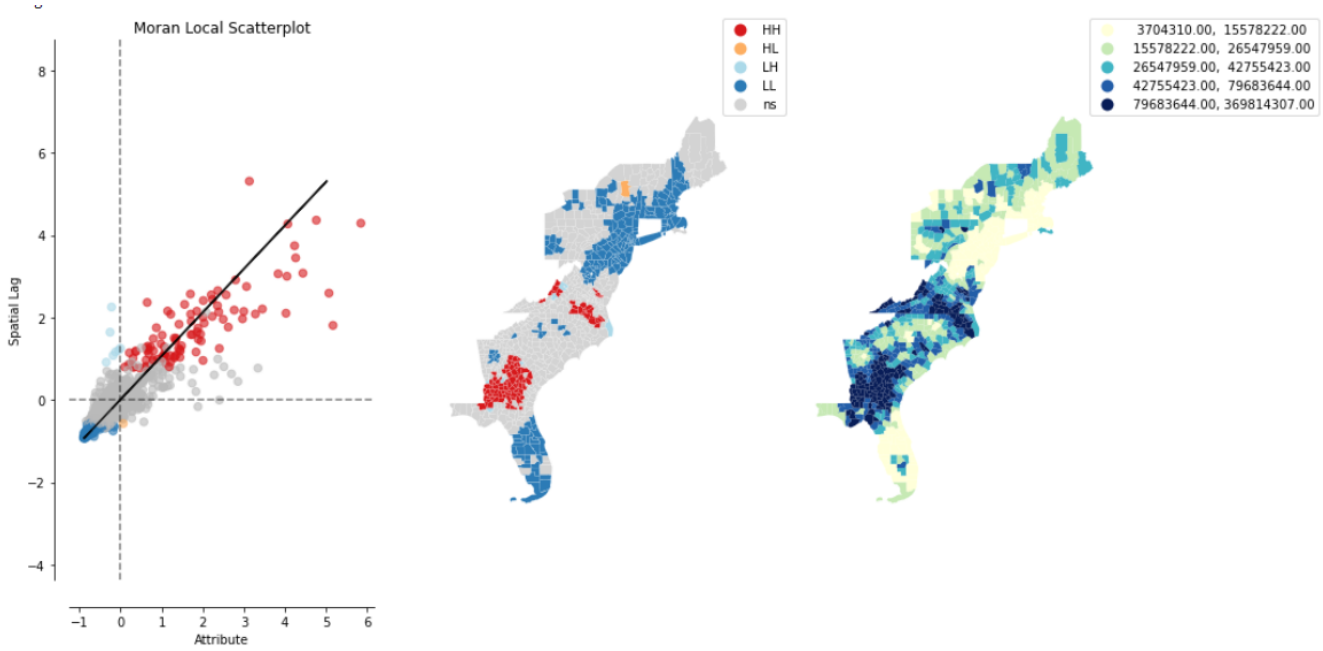

Figure 8: Moran's scatterplot, Moran's cluster map, spatial lag for SCI values in study area

The level of overall SCI was clustered, with numerous hot spots in the south, and even more cold spots in the northeast.

Discussion and Conclusion

In terms of overall SCI values, hot spots were concentrated more so in southern states, like Georgia, whereas most of the northeast is either not significant or has cold spots. Conversely, though, there were more protests concentrated at in the northeast, as shown by the jointplot graphs. If high SCI values were a metric of numerous protests, one would also expect the mean center to be much further south than it was, or would expect there to be many hot spots in the northeast for high SCI values, but there is only a singular county with a relatively high SCI

value in the northeast. With this in mind, there is no conclusive evidence that high SCI is related to high volumes of protest, or the clustering of protests.

Next Steps

Next steps would involve making the methodology more rigorous by focusing more on SCI between two places rather than overall SCI values. Then, a more careful, temporal analysis would be conducted to see how places with high connectivity act at the same time of year. Another option would be to identify a notable event in the past and see which areas protested at the same time, or did not protest at the same time. Regressing social connectivity on counts of protests for two counties may also be an option. Once the methodology is fine-tuned, it would be interesting to turn the focus to a global scale, seeing how events within the US mirror events elsewhere, and vice versa.

References

Armed Conflict Location & Event Data Project . (2017). *Armed conflict location & event data project (ACLED)*. Frequently Asked Question (FAQ): Internal. Retrieved December 16, 2021, from https://acleddata.com/wp-content/uploads/dlm_uploads/2017/10/ACLED_FAQs_2017_October.pdf

Facebook data for good social connectedness index methodology. Data For Good Home. (n.d.). Retrieved December 16, 2021, from https://dataforgood.facebook.com/dfg/docs/methodology-social-connectedness-index

Geographic Data Science with python. Point Pattern Analysis - Geographic Data Science with Python. (n.d.). Retrieved December 16, 2021, from https://geographicdata.science/book/notebooks/08_point_pattern_analysis.html

Han, S. Y., Tsou, M.-H., & Clarke, K. C. (2017). Revisiting the death of geography in the era of big data: The friction of distance in Cyberspace and real space. *International Journal of Digital Earth*, *11*(5), 451–469. https://doi.org/10.1080/17538947.2017.1330366

Wang, Z., Ye, X., Lee, J., Chang, X., Liu, H., & Li, Q. (2018). A spatial econometric modeling of online social interactions using microblogs. *Computers, Environment and Urban Systems*, *70*, 53–58. https://doi.org/10.1016/j.compenvurbsys.2018.02.001