"Food flows between counties in the United States": Visualization and Analysis

Brian Slobotsky[a]
December 17, 2019

Abstract

Origin-destination flow relationships are common in spatial data analysis. The dynamic nature of the data presents a challenge to researchers who wish to visualize and analyze the data efficiently and effectively. In this report, the Food Flow Model will be used as a case study, and visualization and analysis will be conducted in Python using appropriate spatial data science methods. The report will show examples of how a spatial flow dataset can be visualized as interactive maps, how origin-destination flow relationships can be used to construct a contiguity matrix, and how to discover key spatial information about global and local spatial autocorrelation. The findings of this case study can be used as a guide for future spatial research with flow relationships in order to improve the geo-visualization and analysis.

1. Introduction

The study of flow relationships is common in geographic information science research. In reality, location is often dynamic, and understanding the movement of a location or the trajectory of a process is a necessary research exploration. Analyses of commute patterns in transportation studies, migration across borders over time, propagation of disease, supply chain distribution networks, and stormwater drainage systems are just a few examples of research topics which must take consideration of flow relationships. Therefore, it is imperative to discuss methods for effectively visualizing and statistically exploring spatial flow data (Symanzik 2014).

The paper used as a case study in this report is entitled “Food flows between counties in the United States” and was authored by Lin et al and published in Environmental Research Letters in July 2019. In the paper, Lin et al developed the Food Flow Model based on spatially downscaled production, consumption, and network estimates from eight data sources. The first was the Commodity Flow Survey from the US Census Bureau, which tracked over 4.5 million shipments and collected data on state or metropolitan area of origin and destination, type of commodity, mode of transport, and weight of shipment (US Census Bureau 2016). The second was the Freight Analysis Framework (FAF) from Oak Ridge National Laboratory, a derivative of the Commodity Flow Survey, which developed 132 “FAF Zones” and quantified the magnitude of food flows between FAF Zone pairs (Oak Ridge National Laboratory 2015). The FAF data also contained breakdowns by commodity type, based on the standard classification of transported goods (SCTG). The first two sources provided a constraint for the Food Flow Model values with basis in published literature. The US Census Bureau Economic Census and US Department of Agriculture Census of Agriculture were used to estimate production values and determine strength of links between counties. Input—Output Accounts Data and Personal Income figures from the US Bureau of Economic Analysis were used to model consumer habits in the model. Finally, county-to-county distance and network impedance from Oak Ridge National Laboratory and Port Trade from the US Census Bureau were considered in the estimate for network relationships. Regression analysis, cost optimization, and machine learning were used to synthesize the eight data sources into a single dataset. The result of the Food Flow Model is a data product with 161,394 county pairs covering 3,132 counties in the United States, with SCTG categorizations quantified in addition to a total flow value, all measured in kilograms (Lin et al 2019).

2. Literature Review

It has been widely documented that visualization and exploratory data analysis are necessary methods for the synthesis and presentation of geographic data (MacEachren and Kraak 2001). The complexity of space in a geographic sense is a major challenge which presents multiple avenues for possible visual and analytical exploration (Andrienko et al 2007)(Adrienko et al 2010). Existing literature notes that dealing with dynamic spatial-temporal data is a persistent challenge (Çöltekin et al 2017). Methods for exploratory spatial-temporal data analysis (ESTDA) have been documented. Equations have been established for augmented Moran's I calculations for spatial autocorrelation which incorporate temporal elements, as well as directional local indicators of spatial autocorrelation (LISA). Bivariate LISA analysis presents an opportunity to make conclusions about dynamic spatial processes. Interpreting differences in a series of individual global and local autocorrelation analyses are another ESTDA approach (Rey 2014). In these cases, origin to destination flow data can be conceptualized as data with two temporal states. Flow data is one of the four main types of spatial data often analyzed in exploratory spatial data analysis (Symanzik 2014). One of the main issues in origin-destination flow data is that n locations produce n^2 interactions, which can result in an extremely large dataset and a cluttered visualization given a large n. A solution could be to set a threshold or filter to reduce the visual complexity and data size (Fischer and Wang 2011). A spatial interaction matrix can be used to better understand data patterns (Bailey and Gatrell 1995). Spatial weight matrices can be used to measures of contiguity besides adjacency. The distance of economic relationships (Conley and Ligon 2002), social influences (Páez et al 2008), interaction between governments (Brueckner 1998), migration flows (Baicker 2005), and commuting flows (Rincke 2010) have all been used as adjacency alternatives in contiguity matrices.

3. Visualizing the Food Flow Model data

3.1. Nature and limitations of the data

The paper's dataset is published as a .csv file with each row corresponding to a single flow between two county pairs. There are seven columns for the seven SCTG category values, a column for the total value, and two columns for the origin and destination county (represented by their US Census Bureau FIPS codes). Therefore, the data is ten columns and 161,394 rows.

	sctg_1	sctg_2	sctg_3	sctg_4	sctg_5	sctg_6	sctg_7	total	ori	des
0	0.000000e+00	0.000000	0.0	4.763500e+04	149048.644244	0.0	40.262859	1.967239e+05	1001	1001
1	2.030100e+07	0.000000	0.0	1.106758e+07	0.000000	0.0	0.000000	3.136858e+07	1001	1007
2	2.646100e+00	208558.088186	0.0	6.272792e+07	219073.870005	0.0	21525.787037	6.317708e+07	1001	1021
3	0.000000e+00	0.000000	0.0	8.156041e+06	0.000000	0.0	0.000000	8.156041e+06	1001	1037
4	0.000000e+00	0.000000	0.0	0.000000e+00	0.000000	0.0	3400.448878	3.400449e+03	1001	1047

Figure 1: The first five rows of the Food Flow Model dataset

The data did not require any tidying besides adding leading zeroes to some FIPS codes to ensure all were exactly five digits long. The data was not suitable for visualization in Python, as the data contained nothing spatial beyond county FIPS codes. Two additional data sources were utilized to overcome this limitation, a table of FIPS codes from the US Census Bureau to translate FIPS values to the names of counties and county equivalents, and a shapefile of US counties read into Python as a GeoDataFrame from the *geopandas* module. The Food Flow Model very often had multiple flow relationships per FIPS code, and each flow had two counties as endpoints, which presented a challenge to visualize. Another limitation was that the SCTG categories were too broad to deduce interesting results from. For example, the SCTG code 6 corresponds to "Milled Grain Products and Preparations, and Bakery Products." Each SCTG category has more specific subcategories but they were not included in the dataset.

3.2 Visualization

In this section I describe my methods in Python for constructing visualizations from the data, as well as some advantages and some disadvantages for each approach. There are two main visualizations: an interactive version of the map of all flows provided in the paper, and an interactive map where a user input can filter the data to a single county and a single commodity.

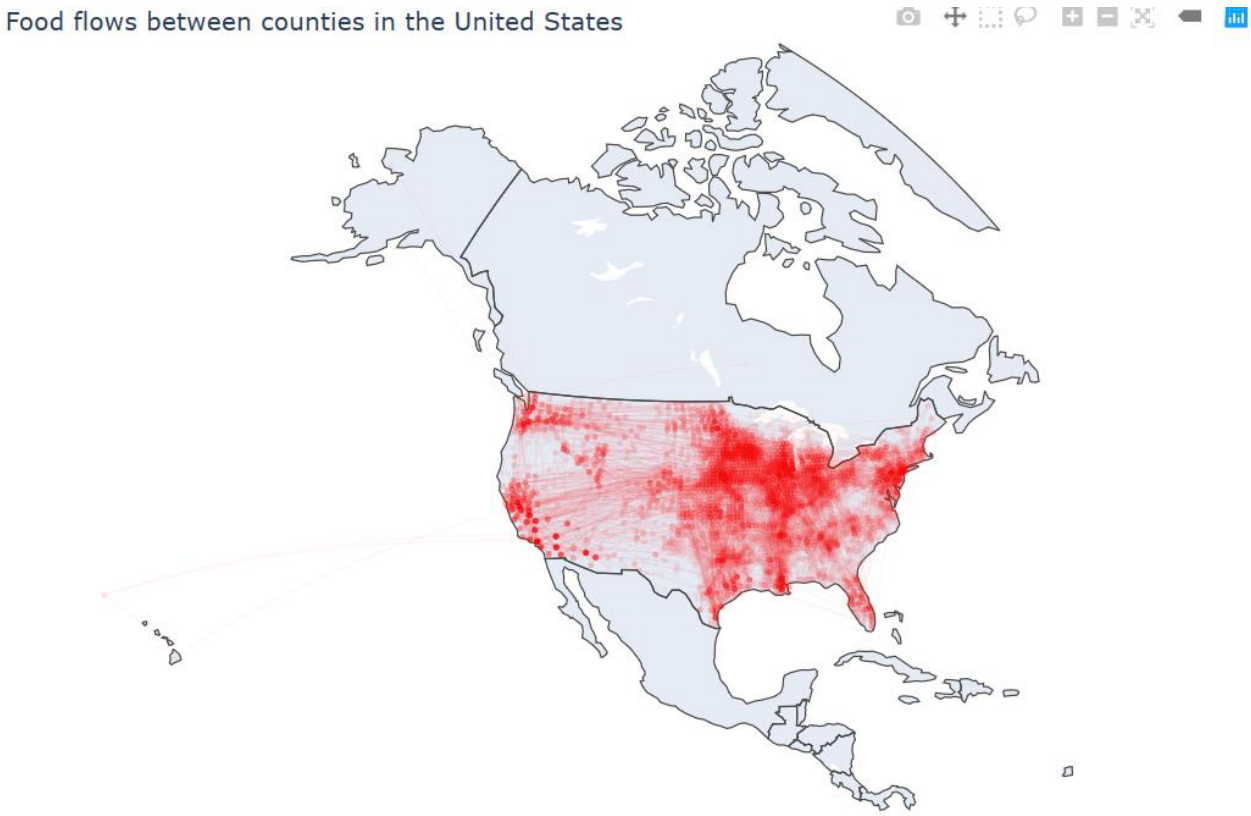

Figure 2: Screenshot of the interactive food flow map from *plotly*

I decided the best way to visualize the flows between counties was by connecting the county centroids with a line, as seen in Figure 2. After extracting centroid coordinates from county polygon geometry, I joined each set of coordinates to the corresponding FIPS code for both the origin and destination columns. I used the *plotly* module to construct lines between origin and destination centroids and kept the centroids as point markers. The lines vary in width and opacity

by the logarithmic proportion of the total value of the county pair to the overall largest value between any two county pairs. The markers vary in the same way by size. Wider, more opaque lines and larger markers denote larger volumes of food flows between county pairs. The addition of some data classification improves the understanding of trends in the data over space. The size of centroid markers provides useful information when the view extent of the map is smaller and the entire United States cannot be seen for comparison. One of the biggest limitations with *plotly* is the inability to convey flow direction with lines. Another limitation is the performance of the interactive display, which is too slow to update with thousands of lines on the map. In my code, I keep only the top 10,000 flow relationships measured by total volume to alleviate this issue. An additional functionality that could improve the visualization would be the ability to see a line highlighted upon hovering, showing information on the origin, destination, and value.

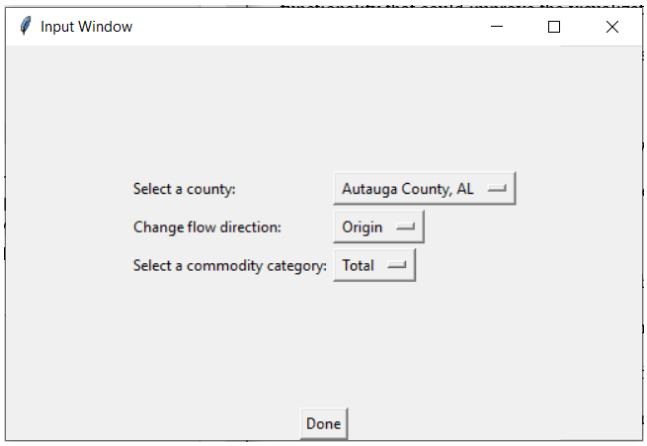

Figure 3: Screenshot of the pop-up input window

With help from the *tkinter* and *folium* modules, I was able to develop a script to generate a user-desired visualization. Running the script creates a pop-up window (shown in Figure 3) that asks for a county input from a list of all 3,142 counties, a commodity input from a list of the seven SCTG category names and "total," and a flow direction (whether the county is the origin or destination of the flows). The inputs are assigned to variables and narrow the GeoDataFrame to

the given specifications. Additional columns are merged into the GeoDataFrame to provide contextual statistics. The code will stop and print an error if the combination is invalid; otherwise, the GeoDataFrame is converted to a .geojson file. The .geojson file is used as a choropleth Leaflet map of counties, with color based on the commodity or total value between the county pairs. A message is printed showing the final selection and a map is automatically loaded in a new browser tab as an HTML file. The script has a runtime of a few seconds.

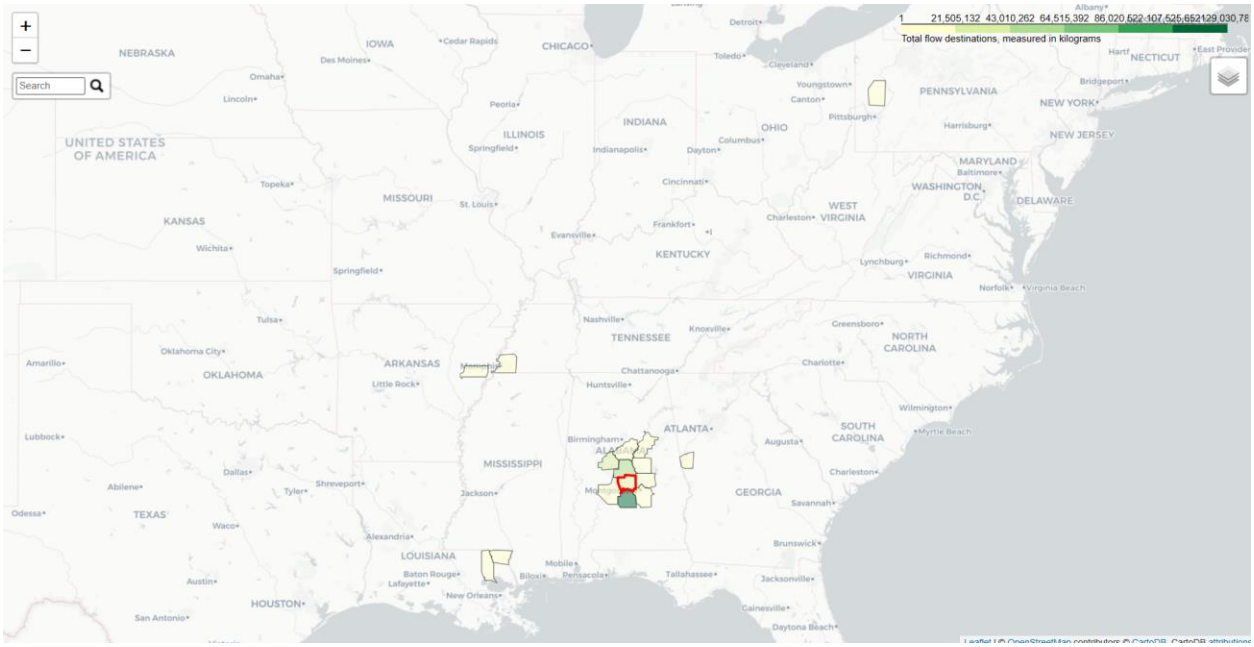

Figure 4: Screenshot of the user-defined Leaflet map, showing total flow destinations for Autauga County, Alabama.

The Leaflet map output has additional features, such as a tooltip which displays relevant and contextual information upon hovering, like the county name, value of the flow in kilograms, number of origin and destination counties, largest origin and destination counties (shown in Figure 5). The map also contains a layer with a red polygon outline on the selected county, layer control to turn layers on and off, and a search bar to look for a specific county. This interactive map was also limited by the size of the data, as counties with zero values were intentionally

omitted in order to improve the script performance. As a result, not every county can be searched in the search bar. Also, working with Leaflet via *folium* limits the possibilities for visualization. A loaded Leaflet map cannot be updated after it has been displayed and click events cannot be recorded as user inputs with *folium*. The accessibility of Leaflet, and plethora of functionalities make *folium* an advantageous method for visualizing spatial data.

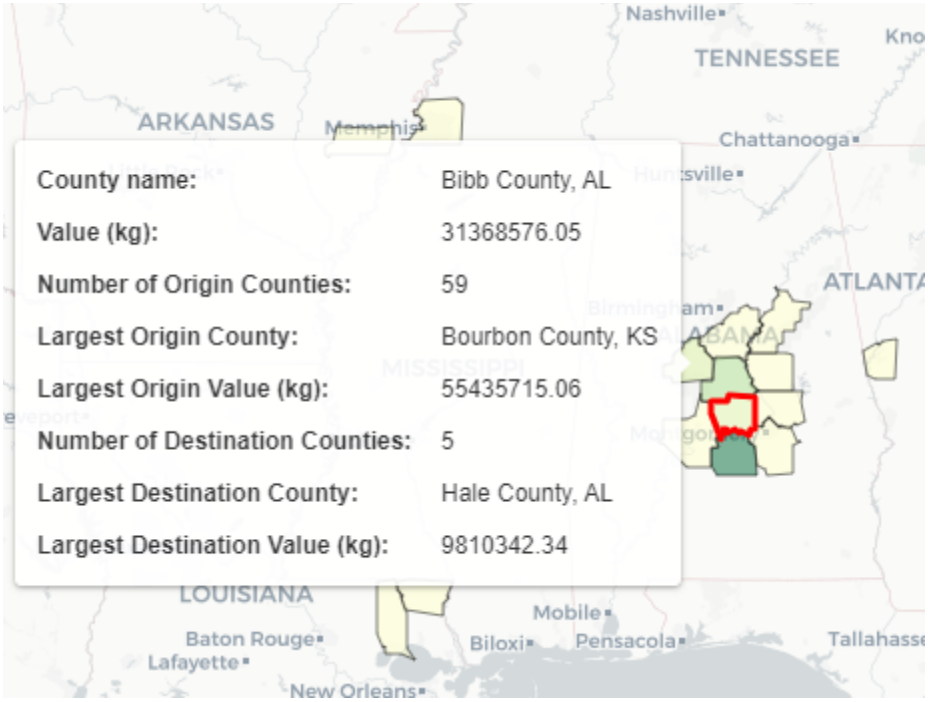

Figure 5: Hovering over Bibb County, Alabama, on the Autauga County destination map.

4. Analytical methods employed

There were two main aspects of the data I explored through exploratory spatial data analysis. First, investigating the relationships between county pairs by constructing two spatial weights matrices based on flow relationships. Second, determining whether the custom spatial weights matrices showed similar spatial autocorrelation results as adjacency-based matrices, and examining the spatial patterns of clusters and outliers.

To begin, I explored the paper's model by constructing contiguity matrices based on flow linkages, one for each flow direction. Essentially, the origin matrix will treat all destinations from a single origin county as neighbors of the origin county, and the weights of the matrix are the total value of flows between the appropriate county pairs. It is important to note that flow-based contiguity is not spatial in terms of adjacency, and that further discussion of spatial autocorrelation among flow relationships can be interpreted as similar values between county pairs. The cardinality of the constructed matrices, essentially the count of a county's neighbors, was analyzed to determine how the origin and destination matrices differ numerically and spatially.

Next, spatial autocorrelation analysis was employed in order to better understand spatial patterns of the total flow value variable for origin counties and destination counties. At a global level, the flow-based contiguity model was compared to an adjacency-based contiguity model, with a k-nearest neighbor conceptualization of contiguity. Both models were tested for global spatial autocorrelation in the single origin and single destination cases. Lastly, local spatial autocorrelation was mapped for all four combinations of matrix conceptualizations and flow directions. The goal of the constructed matrices, cardinality analysis, and global and local spatial autocorrelation is to better understand spatial patterns of the Food Flow Model dataset.

5. Results

Numerically, both matrices have very similar cardinality distribution, shown in Figure 6. A few counties have hundreds of origins or destinations in the Food Flow Model, while most have only dozens of origins or destinations. The mean origin matrix cardinality and mean destination matrix cardinality are both about 51.4, as a result of each flow having one origin and one destination, making the mean the total number of flows divided by the total number of counties.

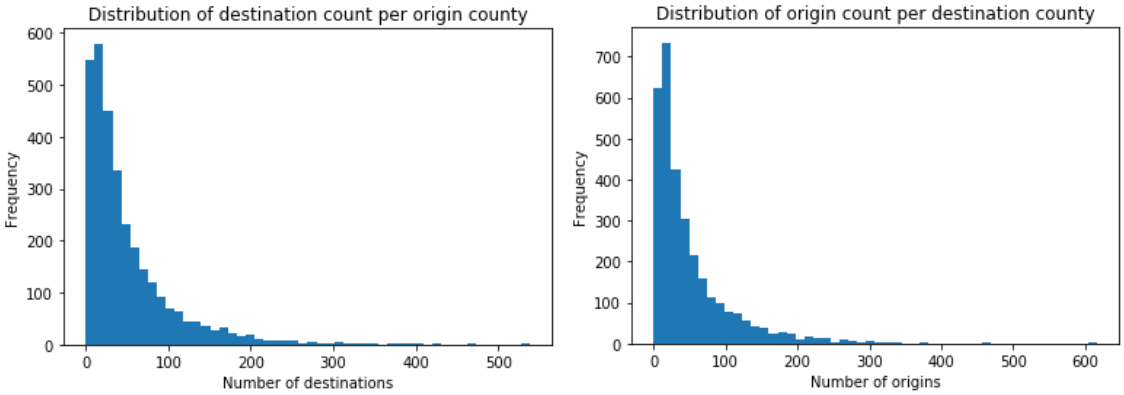

Figure 6: Histogram of cardinalities for both matrices

The mean is larger than the median cardinality for both matrices, which is 32 in the origin matrix and 29 in the destination matrix. The standard deviation of cardinality values is about 57.7 in the origin matrix and approximately 59.4 in the destination matrix, which is consistent with how the maximum value for the destination matrix is greater than the maximum value for the origin matrix, meaning the spread of values is larger in the destination matrix.

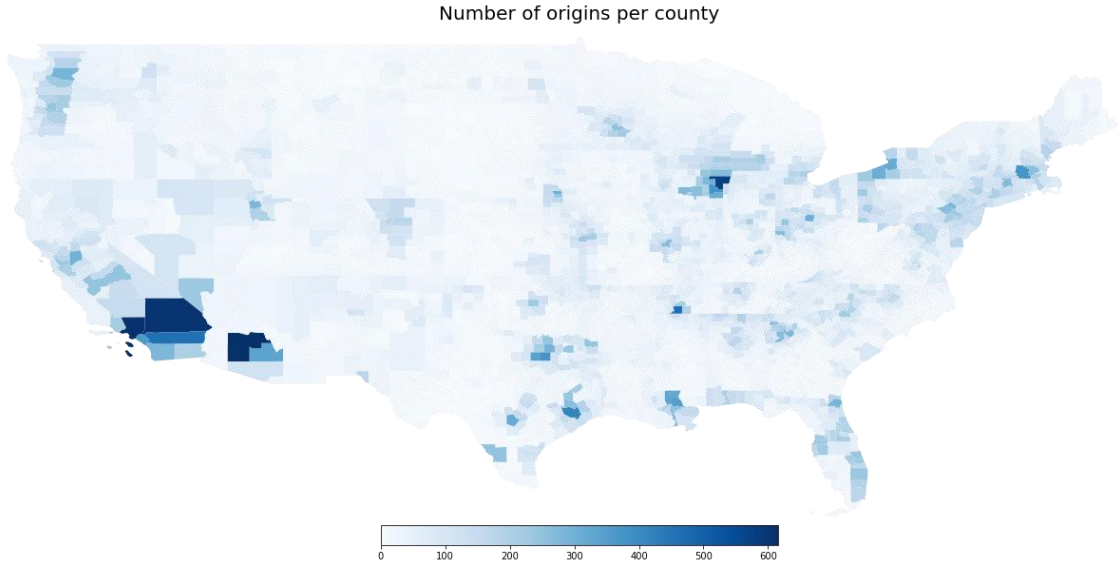

Figure 7[b]: Matrix cardinalities mapped for both matrices

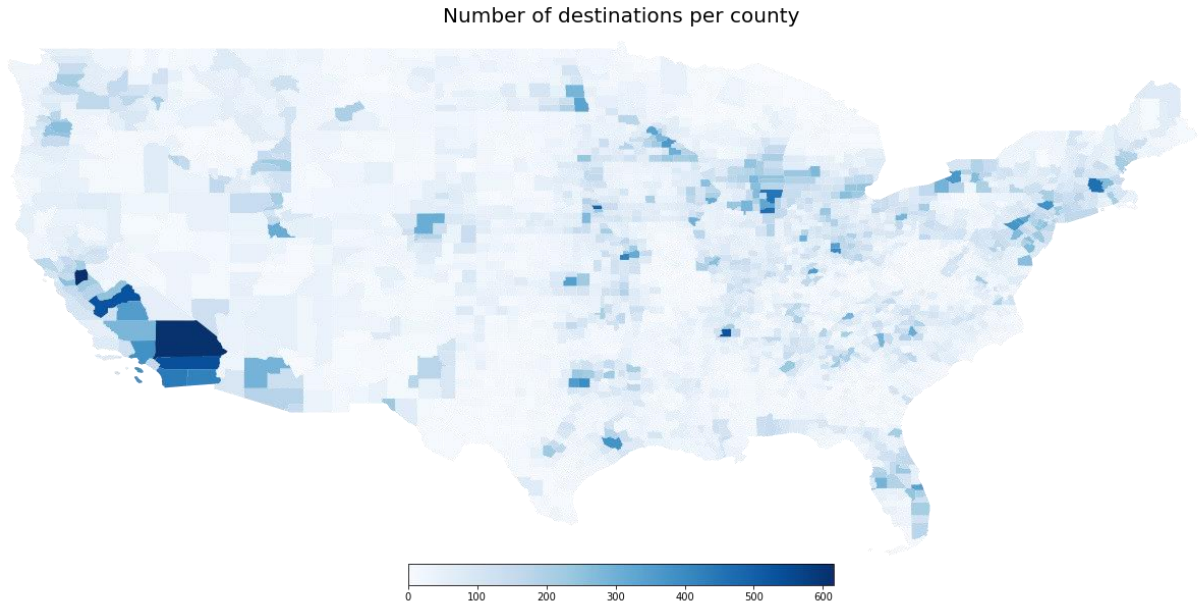

Figure 7 (continued)

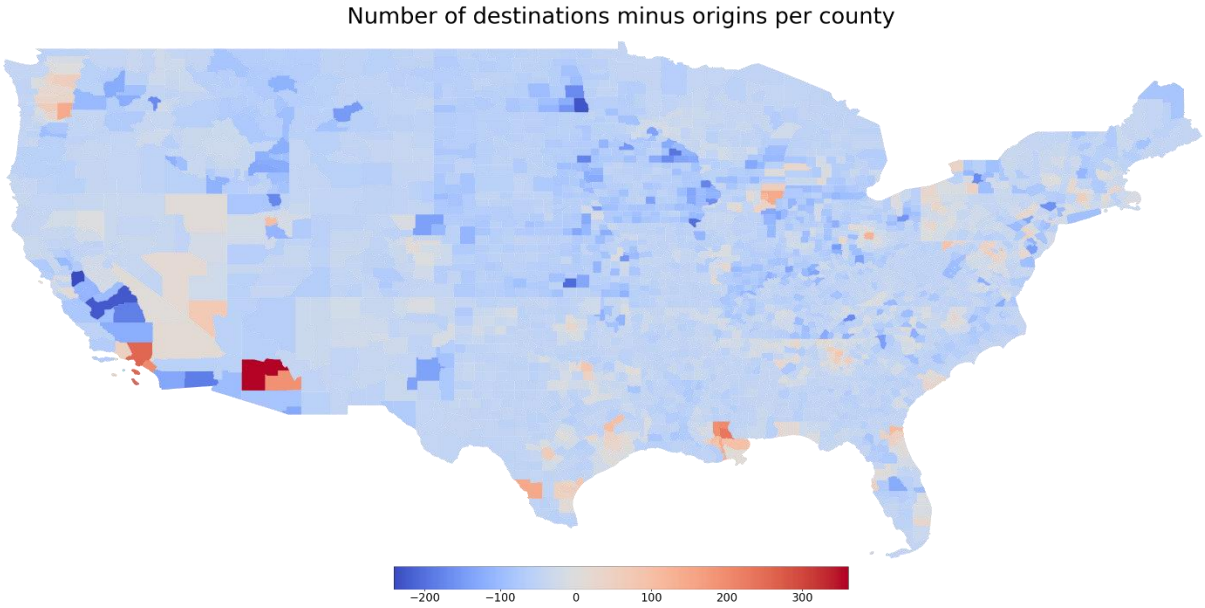

Figure 8[b]: Difference of matrix cardinalities by county

Viewing the spatial distribution of matrix cardinalities, the maps appear to be very similar. A few trends are apparent when considering the difference between origin and destination counts per county, shown in Figure 8. One general trend is that counties in California's Central Valley region and counties in Midwest states like Iowa, Minnesota, and North Dakota have higher

numbers of destination counties relative to their number of origin counties (shown in dark blue in Figure 8). These counties can be inferred as food suppliers of the Food Flow Model. Another trend is that populated areas, such as Los Angeles County CA, Maricopa County AZ, and Cook County IL, have higher counts of origin counties relative to their count of destination counties (shown in dark red in Figure 8). Those counties hold large population centers like Los Angeles, Phoenix, and Chicago, respectively, and are therefore associated with high levels of consumer demand in the Food Flow Model.

Next is an exploration of spatial patterns in the model values, unlike the previous analysis of county relationships in the model. Four Moran's I tests checked for global spatial autocorrelation in the total flow value variable. A k value of 30 was used in the k-nearest neighbor contiguity models because that is approximately equal to the median number of cardinalities for both flow-based matrices.

k-nearest neighbor contiguity model		
Flow direction	Moran's I value	p-value
Single origin	0. 40268384670963375	0.001
Single destination	0.34682965125584575	0.001
Flow-based contiguity model		
Flow direction	Moran's I value	p-value
Single origin	0.0995876761791736	0.271
Single destination	0.06914841408133425	0.326

Table 1: Moran's I test results

The large p-values associated with the two tests with flow relationship contiguity signifies that the constructed matrices do not have statistically significant spatial autocorrelation. Again, this

should not be mistaken for adjacency-based spatial autocorrelation. Rather, the lack of spatial autocorrelation implies counties do not have similar total flow values as their origin and destination counties. For example, a county with significant food production (and subsequently high volume of food flowing out of the county) will tend not to produce food commodities for destination counties with the same level of food production. The k-nearest neighbors adjacency-based model shows statistically significant positive spatial autocorrelation for both flow directions, meaning that similar total flow values cluster together across the United States, in both flow direction cases. The I value for the single origin direction is larger than the I value for the single destination direction, meaning there is more clustering in the single origin direction. This is consistent with the previous discussion that certain regions of the United States are net-producers while specific counties, dispersed across the country, are net-consumers.

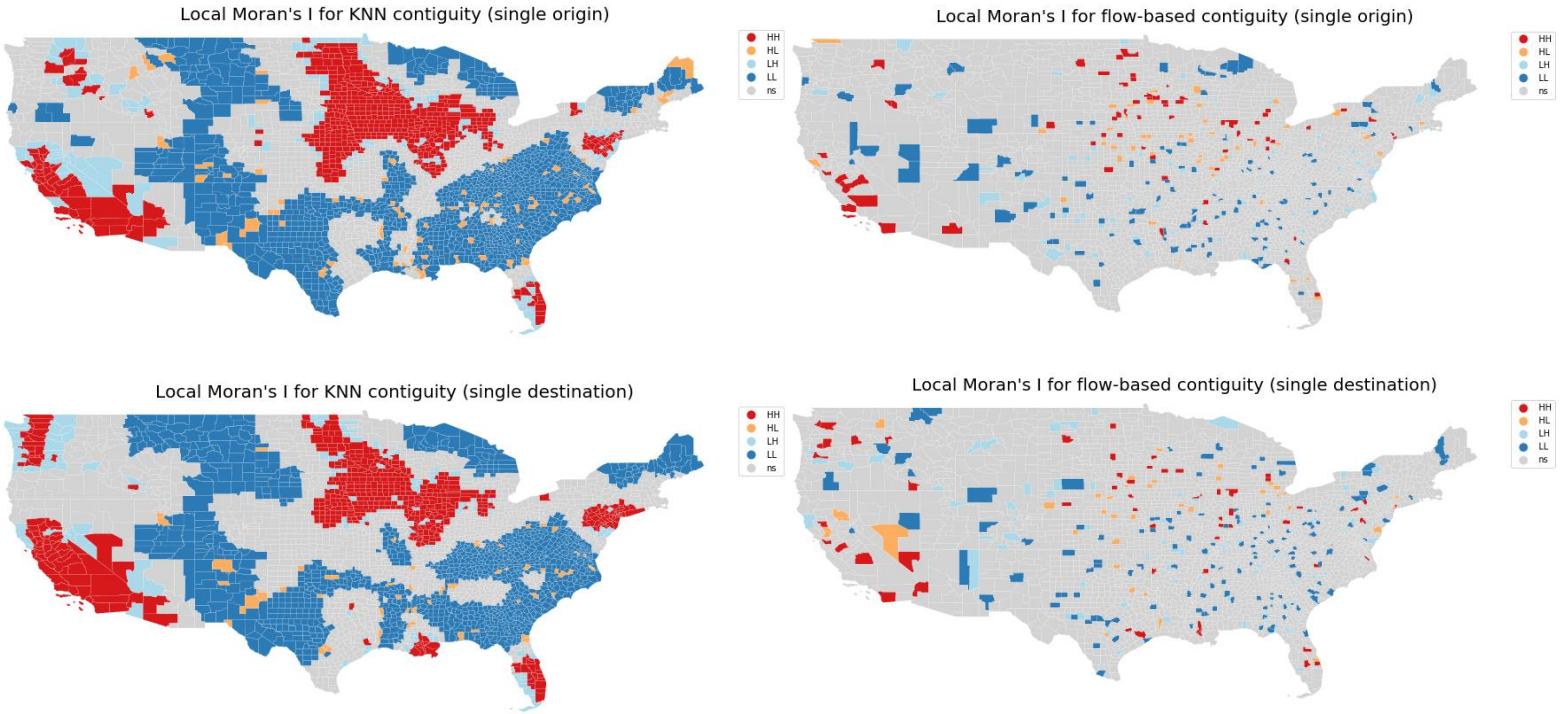

Figure 9[b]: Maps of the four tests of local spatial autocorrelation

Local spatial autocorrelation downscales the previous Moran's test results to the county level by showing clusters of similar values and outliers of different values. There are vast clusters of similar values in the k-nearest neighbor maps, shown in Figure 9, and the two flow direction maps are noticeably quite similar. Some exceptions are high value clusters in Kansas in the single origin map, and high value clusters in western Washington and southern Louisiana in the single destination map. The flow-based maps appear to show "dispersed clusters" since contiguity is not defined by adjacency. This allows for a clearer view of high and low value clusters and outliers in the Food Flow Model. The flow-based maps display trends like the cardinality maps seen previously. Midwestern and California counties contain clusters of high values in the origin model. Populated counties such as San Diego County CA (San Diego), Clark County NV (Las Vegas), King County WA (Seattle), Cook County IL (Chicago), and Jackson County MO (Kansas City) are all in high value clusters in the destination model.

6. Conclusions

A few conclusions about the Food Flow Model can be drawn from the previous exploratory spatial data analysis. First, origin and destination counts per county have similar distributions numerically, but not spatially. The number of flow relationships per county had similar mean, median, and standard deviation values in both single origin and single destination scenarios but displayed different trends when mapped. A measure of total flow value has positive spatial autocorrelation with a k-nearest neighbor conceptualization of contiguity. In other words, the total flow variable displays similar values among spatially adjacent counties, regardless of flow direction. Conversely, there is no statistically significant "spatial" autocorrelation with flow relationships as a conceptualization of contiguity. Counties do not have similar total flow values compared to their origin and destination counties. From these two spatial autocorrelation results,

it is apparent that food travels beyond nearby counties in the Food Flow Model and there are often inter-regional flow relationships. The overarching spatial pattern found was that there are higher cardinalities and local clusters of high values in Californian and Midwest origin counties relative to Californian and Midwest destination counties. The same is true when comparing destinations to origins in counties with large cities. Essentially, California and the Midwest produce large quantities of food which travels to many counties while cities consume large quantities of food travelling from many counties. These spatial trends speak to the nature of the multiple sources of underlying food supply and food demand data. By investigating the Food Flow Model, knowledge about the spatial nature of the original data sources was gained.

7. Discussion

The Food Flow Model presented a simple case study of visualizing and analyzing flow data. Deconstructing flows into destinations of a single origin and origins of a single destination provides the opportunity for two analyses of the same data, where one can compare the similarities and differences in results. The limitation is that this approach can be confusing to follow and difficult to draw definite conclusions in real terms. Visuals can help alleviate the confusion; the interactive Leaflet map used in this report is one way to can make the concept of single origin and single destination flow relationships more easily understood. However, one limitation in terms of visualization is that Python modules designed to handle spatial data are often ill-prepared for displaying flow data. Constructing custom spatial weights matrices based on flow relationships allows for simple origin count per destination and destination count per origin spatial analysis. The custom spatial weights matrices can be also used to analyze spatial autocorrelation among flow relationships. Flow-based matrices are a simple approach to

investigating the complex and dynamic relationships between units of study and investigating clusters of similar values across space.

The processes outlined in this paper are meant to be reproducible for other datasets but are not meant to be considered exhaustive. Future studies can analyze the Food Flow Model or similar origin-destination datasets and use other statistical tests outlined in previous literature, like bivariate LISA analysis and temporally adjusted Moran's I tests. Furthermore, future work can be done to incorporate flow-friendly visualization and analysis tools in existing Python modules.

[a] University of Maryland Department of Geographical Sciences; email: bslobots@umd.edu

[b] Maps in Figures 7, 8, and 9 have versions which include Alaska and Hawaii available in the Appendix

8. References

Papers which were assigned during the course are colored green.

Andrienko, Gennady, Natalia Andrienko, Piotr Jankowski, Daniel Keim, M-J. Kraak, Alan MacEachren, and Stefan Wrobel. "Geovisual analytics for spatial decision support: Setting the research agenda." *International Journal of Geographical Information Science* 21, no. 8 (2007): 839-857.

Andrienko, Gennady, Natalia Andrienko, Urska Demsar, Doris Dransch, Jason Dykes, Sara Irina Fabrikant, Mikael Jern, Menno-Jan Kraak, Heidrun Schumann, and Christian Tominski. "Space, time and visual analytics." *International Journal of Geographical Information Science* 24, no. 10 (2010): 1577-1600.

Baicker, Katherine. "The spillover effects of state spending." *Journal of public economics* 89, no. 2-3 (2005): 529-544.

Bailey, Trevor C., and Anthony C. Gatrell. *Interactive spatial data analysis*. Vol. 413. Essex: Longman Scientific & Technical, 1995.

Brueckner, Jan K. "Testing for strategic interaction among local governments: The case of growth controls." *Journal of Urban Economics* 44, no. 3 (1998): 438-467.

Conley, Timothy G., and Ethan Ligon. "Economic distance and cross-country spillovers." *Journal of Economic Growth* 7.2 (2002): 157-187.

Çöltekin, Arzu, Susanne Bleisch, Gennady Andrienko, and Jason Dykes. "Persistent challenges in geovisualization–a community perspective." *International Journal of Cartography* 3, no. sup1 (2017): 115-139.

Fischer, Manfred M., and Jinfeng Wang. "Models and Methods for Spatial Interaction Data." In *Spatial Data Analysis*, pp. 47-59. Springer, Berlin, Heidelberg, 2011.

MacEachren, Alan M., and Menno-Jan Kraak. "Research challenges in geovisualization." *Cartography and Geographic Information Science* 28, no. 1 (2001): 3-12.

Lin, Xiaowen, Paul J. Ruess, Landon Marston, and Megan Konar. "Food flows between counties in the United States." *Environmental Research Letters* 14, no. 8 (2019): 084011.

Oak Ridge National Laboratory. “Freight Analysis Framework Version 4: User’s Guide for Release 4.0” *Center for Transportation Analysis.* (2015)

Páez, Antonio, Darren M. Scott, and Erik Volz. "Weight matrices for social influence analysis: An investigation of measurement errors and their effect on model identification and estimation quality." *Social Networks* 30, no. 4 (2008): 309-317.

Rey, Sergio J. "Spatial dynamics and space-time data analysis." *Handbook of Regional Science* (2014): 1365-1383.

Rincke, Johannes. "A commuting-based refinement of the contiguity matrix for spatial models, and an application to local police expenditures." *Regional Science and Urban Economics* 40, no. 5 (2010): 324-330.

Symanzik, Jürgen. "Exploratory spatial data analysis." *Handbook of regional science* (2014): 1295-1310.

United States Census Bureau. “2012 Commodity Flow Survey (CFS) Public Use Microdata (PUM) File Data Users Guide.” (2016)

9. Appendix

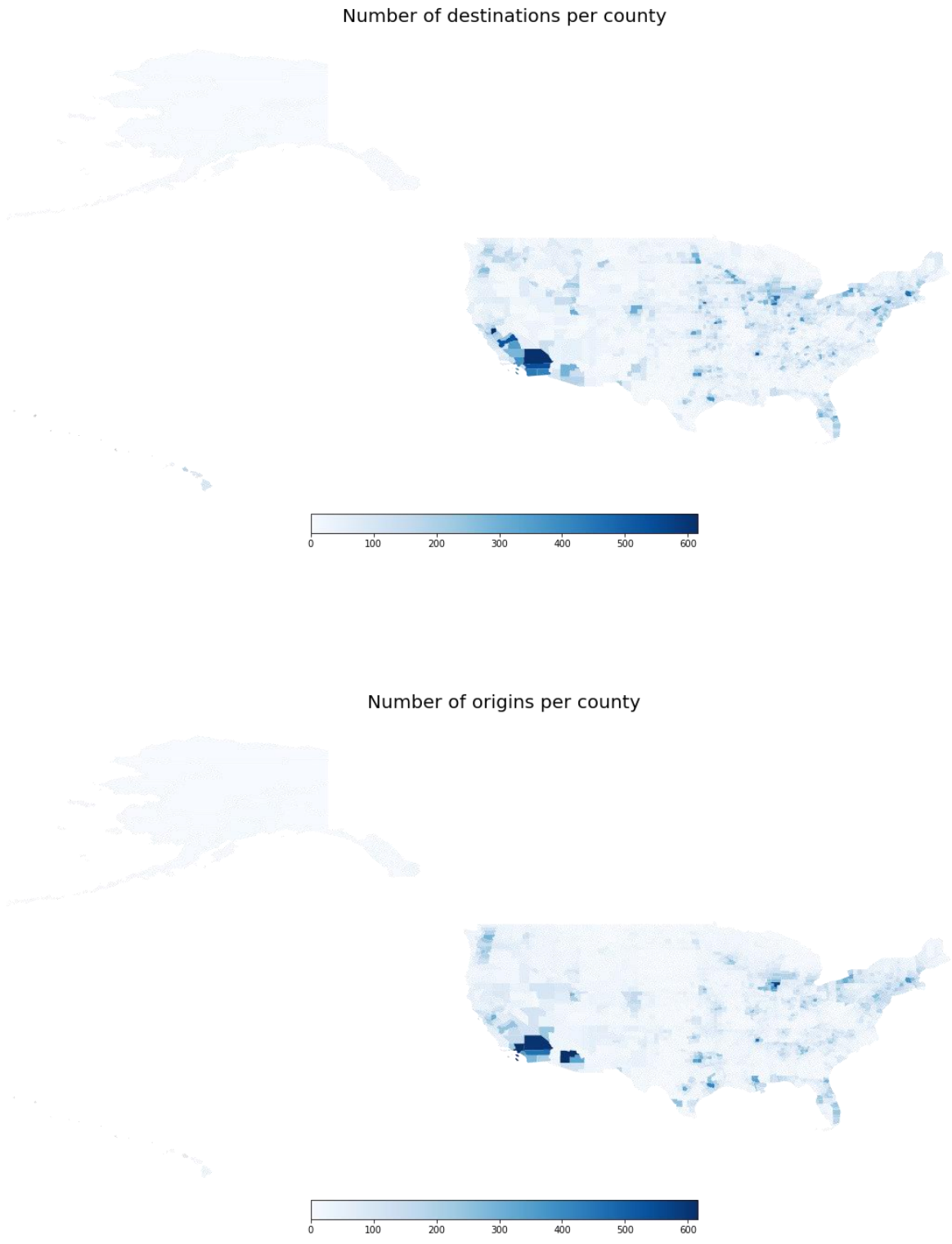

Figure 10: Matrix cardinalities mapped for both matrices, with Alaska and Hawaii

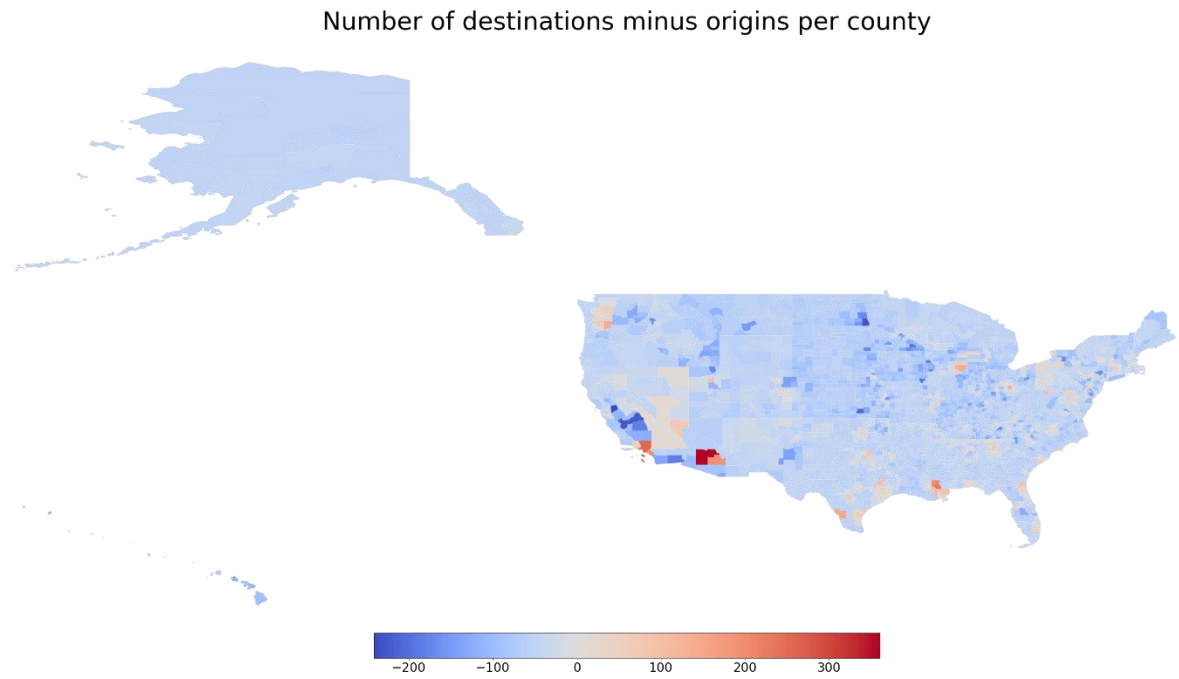

Figure 11: Difference of matrix cardinalities by county, with Alaska and Hawaii

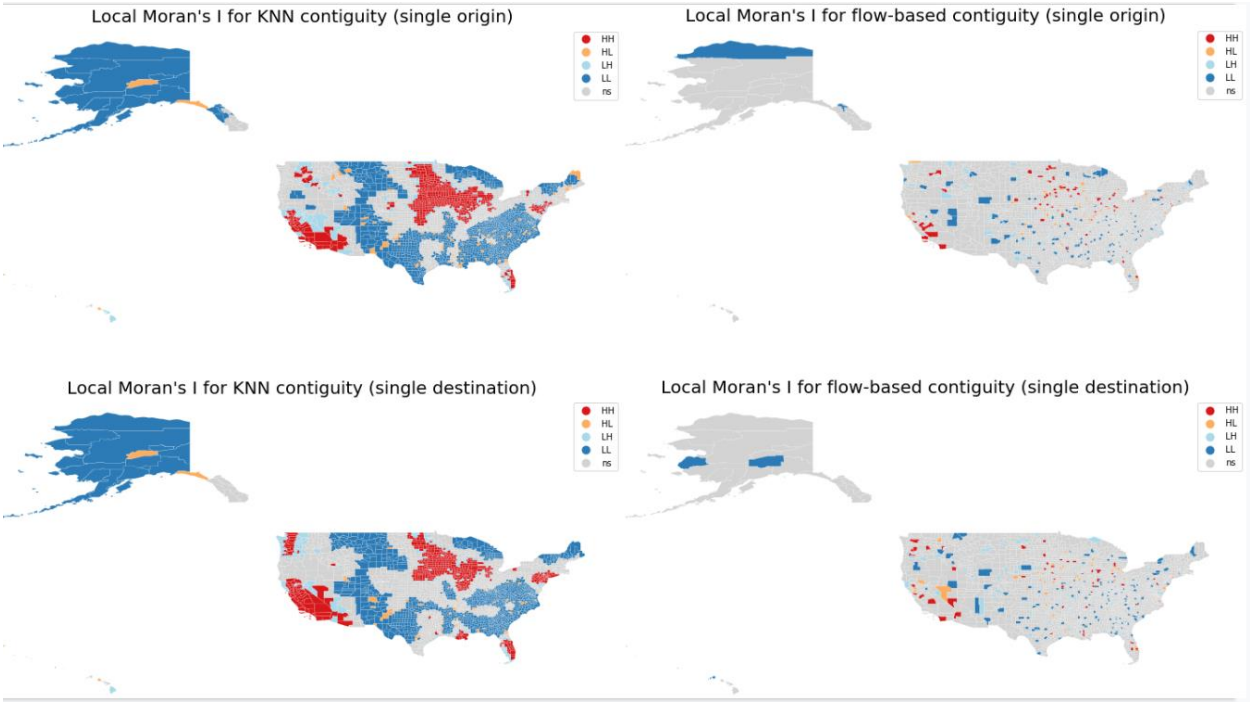

Figure 12: Maps of the four tests of local spatial autocorrelation, with Alaska and Hawaii