A novel method for mapping spatiotemporal structure of mobility patterns during the COVID-19 pandemic
Short Paper Track

Evgeny Noi[1]
Department of Geography, UC Santa Barbara
noi@ucsb.edu

Alexander Rudolph
Department of Statistics and Applied Probability, UC Santa Barbara
arudolph@umail.ucsb.edu

Somayeh Dodge
Department of Geography, UC Santa Barbara
sdodge@ucsb.edu

Abstract

Many classic exploratory data analysis tools in quantitative geography, designed to measure global and local spatial autocorrelation (e.g. Moran's I statistic), have become standard in modern GIS software. However, there has been little development in amending these tools for visualization and analysis of patterns captured in spatiotemporal data. We design and implement a new open-source Python library, VASA, that simplifies analytical pipelines in assessing spatiotemporal structure of data and enables enhanced visual display of the patterns. Using daily county-level social distancing metrics during 2020 obtained from two different sources (SafeGraph and Cuebiq), we demonstrate the functionality of the developed tool for a swift exploratory spatial data analysis and comparison of trends over larger administrative units.

1 Introduction

Many political, social, and economic processes are geographical in nature and are, therefore, subject to first order non-stationary effects (spatial heterogeneity) and second order stationary effects (spatial autocorrelation), which may be scale-specific [8]. These effects invalidate two assumptions of classic inferential (confirmatory) statistics, namely independence and identical distribution, making usage of traditional statistical tools problematic. In classical statistical modeling, these violations may lead to instability in parameter estimation and biasedness. To account for these spatial effects, different methods from spatial statistics were devised, including tools for exploratory spatial data analysis, such as local indicators of spatial association (LISA) [3]. ESDA, an extension of conventional exploratory data analysis (EDA), combines various graphic and numerical statistical techniques designed to generate hypothesis and detect spatial patterns in data.

Currently, there are many software implementations designed to gauge spatial associations in areal data [7]. Recent progress in computational methods and advances in locational

[1] Corresponding author

data acquisition technologies have made calculation of local multivariate spatial association [4, 5] and spatio-temporal pattern mining [10, 9] more manageable. However, the graphic display for many of the outlined tools has not changed within the last 10 years and is not ideally suited for spatial-temporal analysis on multivariate data sets. In fact, many current research investigations rely on static two-dimensional choropleth maps, space-time cubes, and Moran's scatter plots. In this study, we develop three visualization techniques that help identify spatiotemporal structure of the data. These techniques which are developed as a Python package are based on conventional LISA and are ideally suited for analysis of spatiotemporal areal data. The developed Python package, named VASA, will be accessible at `https://github.com/move-ucsb/VASA`. To demonstrate the applicability of the designed visualizations, the variability in spatiotemporal structure of human mobility patterns during COVID-19 pandemic in the United States is assessed.

2 Methods

VASA offers three novel multivariate visualizations: A *stacked recency and consistency map*, a *line-path scatter plot*, and a *categorical strip (dot) plot*. All three techniques use LISA as the base and utilize local Moran's I and permuted p-values. The techniques are best suited for analysis of areal data at two levels of analysis: the object-level and the summary-level. The object-level of analysis receives the data at the finest available scale (e.g. county, census blocks, etc.), whereas the summary-level (e.g. state) refers to the less granular spatial units that contain object-level units. The stacked recency and consistency map allows to ascertain the spatiotemporal structure of data at both object- and summary-level. The categorical strip plot allows for comparison of trends at the summary-level. The line-path visualization is better suited for a fine-detail analysis of individual object-level trajectories within a specified summary-level. The developed visualizations can be made interactive and web-based, allowing user interaction to enhance user exploration performance[2].

Algorithm 1 Recency and consistency mapping algorithm.

set spatial (e.g. *county*) and temporal (e.g. *week*) unit for aggregation;
for *each aggregate spatial (i) and temporal unit (j)* **do**
  calculate local Moran's I (I_{ij}) and permuted p-value;
  recode statistically significant values ($\alpha = 0.05$) of I_{ij} according to Moran's scatterplot quadrants (i.e. hotspots and coldspots);
  calculate cumulative sum of classifications for coldspots (S_c) and hotspots (S_h).;
  retain the week number the county was classified as either coldspot (W_c) or hotspot (W_h).;
end
map county polygons to centroids and display as a marker;
map retained hotspots to red color and coldspots to blue color;
map the cumulative sums (S_c, S_h) to the size of the marker;
map retained week numbers (W_c, W_h) to the intensity of marker color;

[2] See `https://move-ucsb.github.io/covid19-mobility-vis/` for some examples.

2.1 RECO: the stacked recency and consistency map

The stacked recency and consistency map (RECO) improves traditional graphic representation of local Moran's I using several visual variables (e.g. color hue and value, shape, size, texture, transparency) [6] to better illustrate spatial and temporal structure of data. The mapping algorithm is described in Algorithm 1. An example of the generated RECO map is illustrated in Figure 1, and explained later in the Case Study section. Note that the RECO map is a snapshot and static representation of the accumulative patterns up to the last data frame. To represent changes over time dynamically, the RECO can be animated for various timelines.

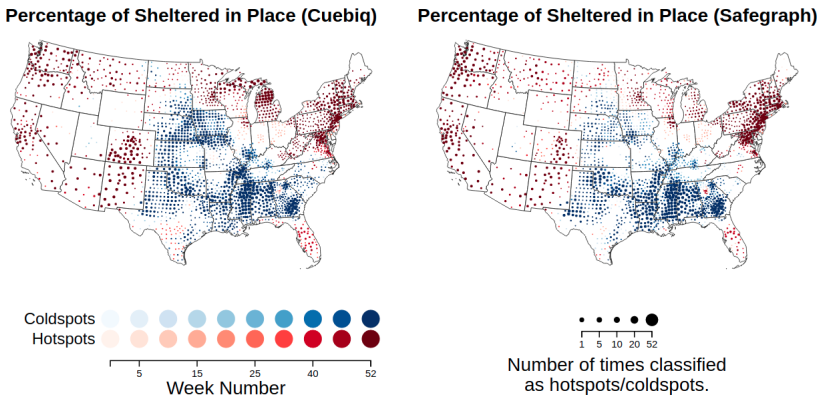

Figure 1 The stacked recency and consistency map of mobility. The color represents hotspots (red) and coldspots (blue). The hue denotes the week number, and not the intensity of hotspot/coldspot. For cases when the county was a hotspot/coldspot for more than one week, the hue intensity denotes the hue for the most recent week. The size of the marker corresponds to the number of weeks the county was classified as either a hotspot/coldspot.

Algorithm 2 The line-path scatter plot algorithm.

set spatial (e.g. *county*) and temporal (e.g. *week*) unit for aggregation;
for *each aggregate spatial (i) and temporal unit (j)* **do**
  calculate local Moran's I (I_{ij}) and permuted p-value;
  recode statistically significant values ($\alpha = 0.05$) of I_{ij} according to Moran's scatterplot quadrants (i.e. hotspots and coldspots);
  calculate cumulative sum of classifications for coldspots (S_c) and hotspots (S_h).;
  retain the week number the county was classified as either coldspot (W_c) or hotspot (W_h).;
end
plot the cumulative sums (S_c, S_h) against the week number;
map retained hotspots to red color and coldspots to blue color;
mark the cumulative sum on the last week number of classification with a circle;

2.2 Line-path scatter plots

To provide an analyst with a better view of trends for individual counties (or, potentially, any other spatial units) the line-path scatter plots (Algorithm 2) is developed. This visualization

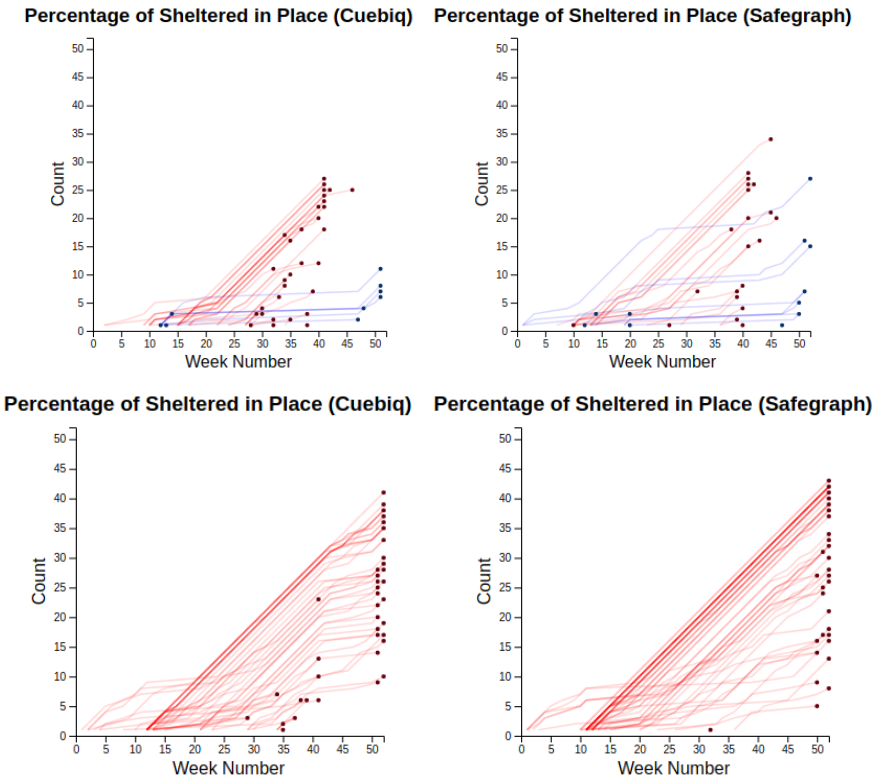

Figure 2 Line-path of social distancing metrics for counties in Florida (top row) and counties in California (bottom row). Each line traces the cumulative number of times a county is identified as a hotspot or coldspot.

uses a line for each county to represent the cumulative number of times it has been identified as a hotspot or coldspot. Compared to RECO, this representation has the advantage of being able to show county behavior over time (e.g. at each week), instead of just the final time point. Figure 2 illustrates an example of the line-path scatter plot (see Case Study).

2.3 The categorical strip (dot) plot

The primary purpose of the categorical strip plot is to combine the LISA clusters with a categorical locational variable, such as U.S. states. Unlike RECO, this visualization provides an aggregate aspatial view of the data to ease the comparison between different variables or data sources. The algorithm that creates visualization is detailed in Algorithm 3. An example of the categorical strip (dot) plot is provided in Figure 3, and further explained in Case Study.

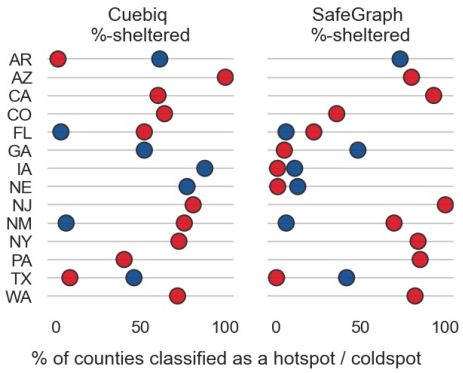

Figure 3 Strip (dot) categorical plot for mobility metrics.

Algorithm 3 Categorical strip (dot) plot algorithm.

set spatial (e.g. *county*) and temporal (e.g. *week*) unit for aggregation;
for *each aggregate spatial (i) and temporal unit (j)* **do**
  calculate local Moran's I (I_{ij}) and permuted p-value;
  recode statistically significant values ($\alpha = 0.05$) of I_{ij} according to Moran's scatterplot quadrants (i.e. hotspots and coldspots);
  calculate cumulative percentage of classifications for coldspots (P_c) and hotspots (P_h).;
end
aggregate cumulative percentage (P_c, P_h) at the state level into P_{sc}, P_{sh};
map aggregated percentages (P_{sc}, P_{sh}) on the x-axis;
map states on the y-axis;
map retained hotspots to red color and coldspots to blue color;

3 Case study: investigating spatiotemporal structure of mobility during COVID-19 pandemic

3.1 Data

In order to demonstrate the applicability of the developed visualization techniques, two social distancing mobility metrics are used. These indices include *% sheltered in place, weekly rolling average*[3] obtained from Cuebiq and *% sheltered, weekly average* obtained from SafeGraph. The choice of these two sources is intentional: Cuebiq [1] provides data by subscription, while SafeGraph [2] provides data free of charge through an open data consortium. Both data sources are widely used in COVID-19 related research. More importantly, they provide a full coverage across the U.S. and throughout 2020 which make them easily comparable. The data are collected daily[4] at the county level for all of the United States for a total of 12 months of observation in 2020. Figure 4 below plots the time series of 7-days rolling national average for the observation period. Both metrics illustrate sharp increase in the percentage of people staying home at the start of the pandemic and initiation of governmental cordoning policies in March 2020, followed by a steady decline over the summer. However, some differences are observed in the captured patterns using these indices.

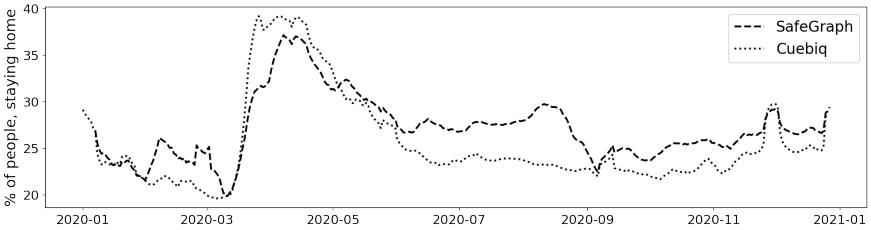

Figure 4 Mobility metrics from Cuebiq and SafeGraph used for demonstration purposes

[3] *% sheltered* indicator measures the percentage of population, staying at home, where the location of *home* is identified computationally based on proprietary algorithms of Cuebiq and SafeGraph.
[4] We aggregate data by week to make computations more manageable for demonstration purposes.

3.2 Experimental results

Figure 1 illustrates the recency and consistency map of the produced spatial clusters, assessed via LISA. In this figure, Cuebiq is depicted on the left and SafeGraph is plotted on the right. Cluster 'recency' (i.e. how recent a cluster appeared towards the end of 2020) is mapped to the color gradient. That is, the hotspots/coldspots occurring in the beginning of 2020 are denoted in lighter shades of red/blue. Conversely, more recent clusters are visualized in darker shades. Cluster consistency (i.e. how frequently a cluster appeared throughout 2020), defined as the number of weeks the county was consistently ranked as either a hotspot or coldspot, is mapped to the centroid marker size. That is, the higher consistency values correspond to proportionally larger marker sizes. While the location of coldspots in both sources is centered in the South for both data sets, with the core in Northern Texas, Mississippi, Alabama, Louisiana and Georgia, a higher cluster consistency (i.e. bigger markers) is observed in Cuebiq data for Nebraska, Iowa and Kansas. The location of hotspots is similar in the Northeastern states, Great Lakes region, and in Western United States for both data sources, but the recency and consistency is higher (i.e. darker and bigger markers) for Cuebiq data in Wisconsin and Michigan. In both sources, the states with some of the least recent hotspots are situated in Indiana and Ohio. It is necessary to remark that here, the hotspots represent a lower mobility in the area (higher percentage of sheltered-in-place population).

Figure 2 illustrates individual trajectories for counties in Florida and California. These two states are selected for illustration purposes because they had different non-pharmaceutical policies. Top row shows the spatiotemporal paths of counties in Florida. As we can see, there are counties that remained hotspots (higher percentage staying home) and coldspots (lower percentage of people staying home) during the observation period. Both SafeGraph and Cuebiq show consistent hotspot behavior for counties starting from week 10 to week 40, when the governor opened the Florida's economy. On the other hand, only a few counties were consistent coldspots. California (bottom row Figure 2) is a completely different story: none of the counties were classified as coldspots, and a lot more counties were consistent hotspots starting from week 10, when the stay-at-home order was issued. These group of counties remained consistent until the end of 2020, which can be seen from the slope of the lines, where the slope of 1 indicates that the county was classified as either a hotspot or a coldspot each week for each particular segment.

Figure 3 illustrates the resulting summaries as a series of scatter plots, where each state is marked on the y-axis, and there is a separate column for each data source[5]). The location of each marker on the horizontal axis denotes the percentage of coldspots/hotspots (from the total number of counties within a state). The blue markers indicate coldspots and the red markers indicate hotspots. It is not uncommon for some states to include counties of both cluster types. In such cases, two markers (blue and red) per line are visualized. The primary goal of this visualization is to summarize and communicate the differences in spatial structure across various mobility indices or data sources for different states. The visualization can be read vertically and horizontally. Vertically, it allows to assess the overall split between the hotspots and coldspots, and to identify the states which are primarily hotspots, coldspots or a combination of both. Horizontally, it allows a comparison in three dimensions: state-level, source-level, and indicator-level.

[5] For demonstration purposes and due to limitation on the size of the submission only a subset of states is selected

4 Conclusion

This paper demonstrated the applicability of three developed visualization techniques for assessing spatiotemporal structure of mobility during COVID-19. When combined, these techniques provide an effective visual analytic tool to identify micro and macro patterns in data at different scales by identifying consistent spatial clusters with statistically significant observed spatial dependence. By adding a temporal component, the presented mapping techniques will help identify primary interest areas for further more detailed and fine-grained analysis. Furthermore, these tools provide enough flexibility to switch between the level of detail and efficiently track individual trajectories of various administrative units. These package can be further extended to allow aggregation and analysis of various geographic processes on irregular lattice. In the course of the next few months, user studies will be conducted to improve the interface and assess the effectiveness of the tool.

5 Acknowledgements

The authors gratefully acknowledge the support from the National Science Foundation Award BCS # 2043202 “CAREER: Modeling Movement and Behavior Responses to Environmental Disruptions”.

References

1 Cuebiq Data for Good, 2020. Accessed: 2020-01-19. URL: `https://www.cuebiq.com/about/data-for-good/`.

2 US consumer activity during COVID-19 pandemic: the impact of coronavirus (COVID-19) on foot traffic., 2020. Accessed: 2020-01-19. URL: `https://www.safegraph.com/dashboard/covid19-commerce-patterns`.

3 Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

4 Luc Anselin. A local indicator of multivariate spatial association: extending geary’s c. *Geographical Analysis*, 51(2):133–150, 2019.

5 Luc Anselin and Xun Li. Operational local join count statistics for cluster detection. *Journal of geographical systems*, 21(2):189–210, 2019.

6 J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, 1967/1983.

7 Roger S Bivand and David WS Wong. Comparing implementations of global and local indicators of spatial association. *Test*, 27(3):716–748, 2018.

8 Urška Demšar, Paul Harris, Chris Brunsdon, A Stewart Fotheringham, and Sean McLoone. Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, 103(1):106–128, 2013.

9 Changfeng Jing, Yanli Zhu, Mingyi Du, and Xintao Liu. Visualizing spatiotemporal patterns of city service demand through a space-time exploratory approach. *Transactions in GIS*.

10 Youngok Kang, Nahye Cho, and Serin Son. Spatiotemporal characteristics of elderly population’s traffic accidents in seoul using space-time cube and space-time kernel density estimation. *PLoS one*, 13(5):e0196845, 2018.