[AM-03-058] Hot Spots and Getis-Ord Gi* Analysis

Abstract

A common goal in spatial analysis is the identification of regions containing unusually high or low values. These areas may be called hot spots if the values are high and cold spots if the values are low. These hot/cold spots indicate where the effects of spatial heterogeneity are greatest. Point density, heat, and choropleth maps all highlight these areas in one way or another. However, due to the limitations of subjective symbolization, statistical methods of hot spot detection are common. Some, like Moran's I, simply identify the pattern for the entire study area. Local methods display the location and magnitude of individual high and low clusters. Getis-Ord Gi* analysis is the local method most associated with the term hot spots and it is the focus of the second half of the article. Getis-Ord Gi* combines the logic of a probability map with moving windows, kernels and/or adjacency weights. The result is an output surface showing neighborhoods with means significantly above or below the global mean. A primary concern is the correct parameterization, especially the correct conceptualization of spatial relationships. Spatiotemporal variants, limitations, and future directions of hot spot analysis are briefly discussed.

Keywords: choropleth maps, clustering, heatmap, hot spot analysis, spatial statistics

Author & citation

Lester, K. (2024). Hot Spots and Getis-Ord Gi* Analysis. The Geographic Information Science & Technologies Body of Knowledge. John P. Wilson (Ed.).
DOI:10.22224/gistbok/2024.1.23

Explanation

1. Visual Identification of Hot and Cold Spots
2. Statistical Identification of Hot and Cold Spots
3. The Basics of Getis-Ord Gi* Hot Spot Analysis
4. Conceptualization of Spatial Relationships
5. Emerging Hot Spot Analysis
6. Future Directions / Current Frontiers

1. Visual Identification of Hot and Cold Spots

When presented with a thematic map, most people's attention is drawn to areas of unusually high or low values. These areas may be called hot spots if the values are unusually high and cold spots if the values are unusually low. However, visual inspection is just one of the methods used to identify these clusters. This article briefly considers the role of visual inspection, global measures of autocorrelation, and local cluster detection before turning focus to the local Getis-Ord Gi* method.

The most basic form of hot spot analysis is simple visual inspection of the data to determine any areas of unusually high or low values (Murray, 2020). In some cases, a simple point or event map may be sufficient to identify areas of unusually high incidence,

as in the case of John Snow's famous cholera map of London's West Side (Figure 1). By marking the locations of known cholera incidents, emergent patterns provided a better understanding of the spread of disease and improved targeting of intervention.

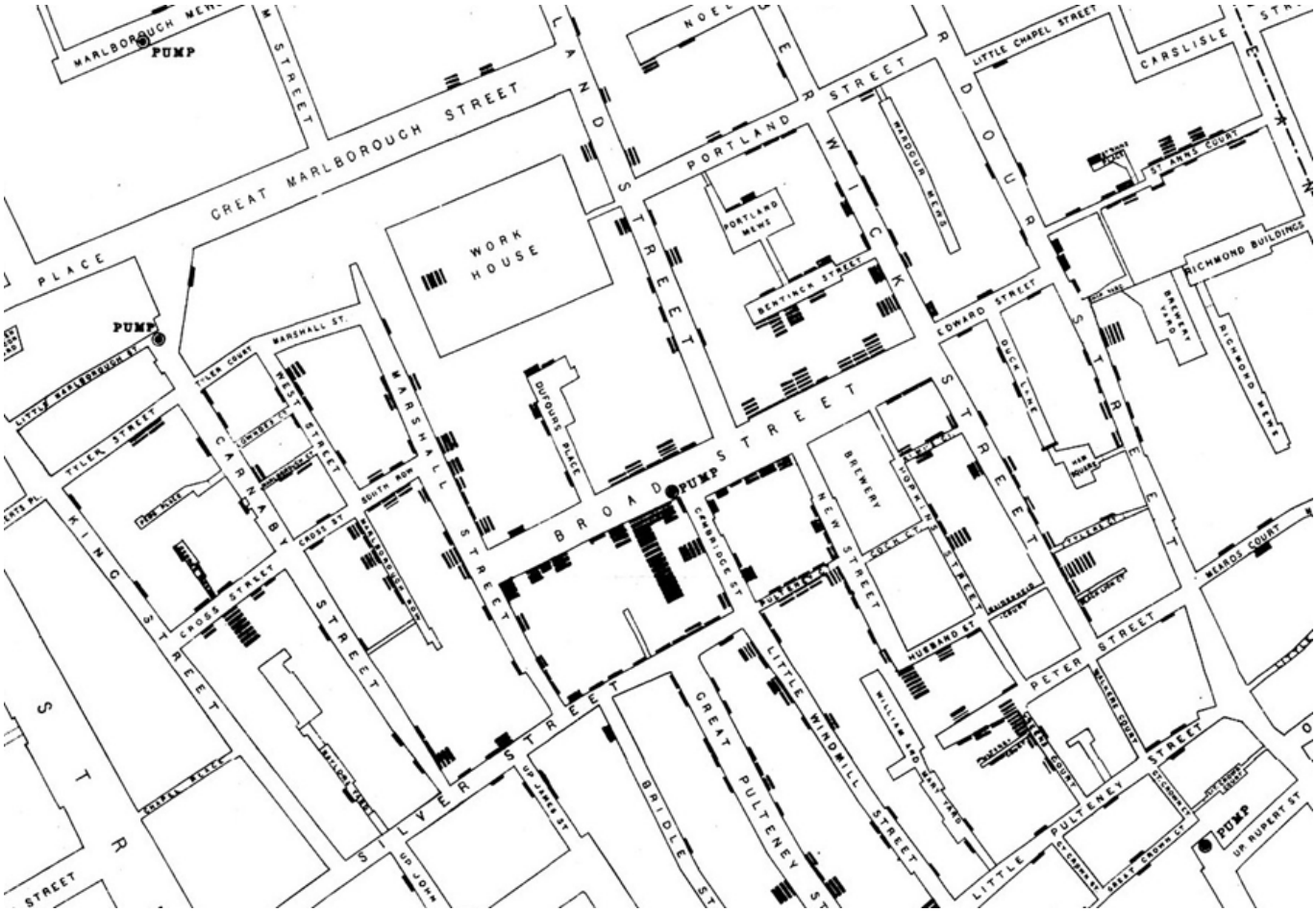


Figure 1. A close-up of John Snow's map of London from On the Mode of Communication of Cholera. Buildings with the highest frequency of disease are adjacent to the contaminated water pump. Source: Snow, 1854.

Researchers may use heat maps to investigate larger sets of point data when visual discernment of individual observations are impossible. Heat maps are the general term often given to the output of kernel density estimation (KDE). KDE uses a moving window approach to create a smoothed surface of local point density. Thus, areas of unusually high local density can be identified as hot spots (Oliveira & de Oliveira, 2017).

Hot and cold spots in areal or polygon data may be identified visually using choropleth maps (Cardozo-Stolberg et al., 2023). Choropleth maps assign specific colors to areal units based on the attribute value of interest and a pre-determined classification scheme. Choropleth maps are easy to make and intuitive to interpret, but they have some drawbacks as tools for hot spot identification. For example, areal units are rarely uniform and map readers often read large units as more important or more unique than they truly are (Schiewe, 2019). Small areas may be overlooked in a heterogeneous environment.

Critically, the chosen classification scheme can dramatically change the perception of the map. Schiewe (2019) provided 260 participants with choropleth maps using different

classification schemes (equal interval, quantile, etc.) and no legends. When asked to identify the map with larger average values, 93.1% selected the map with more dark colors. Simple human bias towards dark or bright colors can interfere with interpretation.

Consider the example of a normally distributed dataset and five categories of classification. In a quantile scheme, 20% of the observations will fall into the highest category and 20% into the lowest category. However, in an equal interval scheme, only about 7% of the observations would fall into each of the extremes. Should the analyst decide to base their conclusions on the classification of cases in a choropleth map, conclusions could vary widely for the same data.

2. Statistical Identification of Hot and Cold Spots

While visual inspection may be sufficient for some applications, statistical methods of hot spot detection provide more specificity and assign significance to observations. These tools fall within the general realm of clustering detection (i.e. k-means, scan statistics, regionalization, etc.). In general, analytical approaches to cluster detection rely on both attribute similarity and spatial proximity or adjacency (Murray, 2020). More specifically, hot spot analysis is most associated with methods incorporating both spatial autocorrelation and probability testing.

Global measures of spatial autocorrelation may be considered a form of hot spot analysis. For example, the Moran's I statistic uses a correlation matrix to assess the similarity of observations based on both attribute scores and spatial lag from a spatial weights matrix (Weeks, 2023). A high, significant result indicates that similar values (high or low) are occurring in proximity to each other.

The Getis-Ord G statistic goes a step further by indicating whether these clusters are in the low or high end of the data (hot or cold spots). The Getis-Ord statistic is a modified two-sample t-test of means to examine the presence of spatial heterogeneity. If the results are significant, then the mean of values within a defined neighborhood around each point are different from the regions outside of these neighborhoods (Rogerson, 2024). The positive or negative test statistic indicates whether the local mean is higher (hot) or lower (cold) than the outlying areas.

Global measures of spatial autocorrelation have limited utility because they do not provide specific information on where these hot or cold spots occur. Thus, the need for local methods of hot spot detection (Ord and Getis, 1995). In general, these methods compare the observed distribution of events to an expected distribution, usually based on randomness. While not a necessary component, most methods also use a moving window approach. Moving windows operationalize the concept of spatial autocorrelation by incorporating information about nearby events within each case used in the analysis.

Both spatial scan statistics and local Moran's I analysis could be considered variations of local hot spot analysis (Xie, Shekar, & Li, 2022). These methods are explored in more depth elsewhere in the Body of Knowledge. However, one method has become more synonymous with hot spot analysis than any other: local Getis-Ord Gi* analysis. For many years this tool was called "Hot Spot Analysis (Getis-Ord Gi*)" in the Esri ArcGIS ecosystem. Newer versions of the tool are simply called "Optimized Hot Spot Analysis." Thus, it is has become quite common to see the term "hot spot analysis" used without any mention of the Getis-Ord Gi* statistic underlying the method, especially among researchers outside of geography and

the spatial sciences. The rest of this entry explores the process and considerations specific to the local Getis-Ord Gi* method.

3. The Basics of Getis-Ord Gi* Hot Spot Analysis

The overall goal of Getis-Ord Gi* analysis is to determine if local neighborhoods are significantly different than the global study area, based on both distance and attribute values. Getis-Ord Gi* is part of a suite of Getis-Ord statistics which drew from earlier global clustering measures, especially the Ripley's K function (Getis & Ord, 1992).

The simplest way to understand the process is to start with a probability map. Probability maps compute a z-score for each case in the dataset based on an attribute of interest. Each z-score is compared to the normal distribution to derive a p-value for each unit. The resulting map shows which units are significantly different from the global mean and may indicate if this difference is above or below average (i.e. hot or cold). However, this method does not include any information about distance, and does not fit well into the taxonomy of spatial clustering methods (Murray, 2024).

Hot spot analysis follows the same mechanics but introduces the idea of neighborhood. Each unit is represented by a neighborhood value, defined by some conceptualization of spatial relationships set by the analyst. Then, the neighborhood value is compared to the global mean to determine if the neighborhood value is statistically different. If a unit's neighborhood is significantly different from the study area, then the unit is classified as a hot spot or cold spot, based on the z-score. Figure 2 illustrates this process.

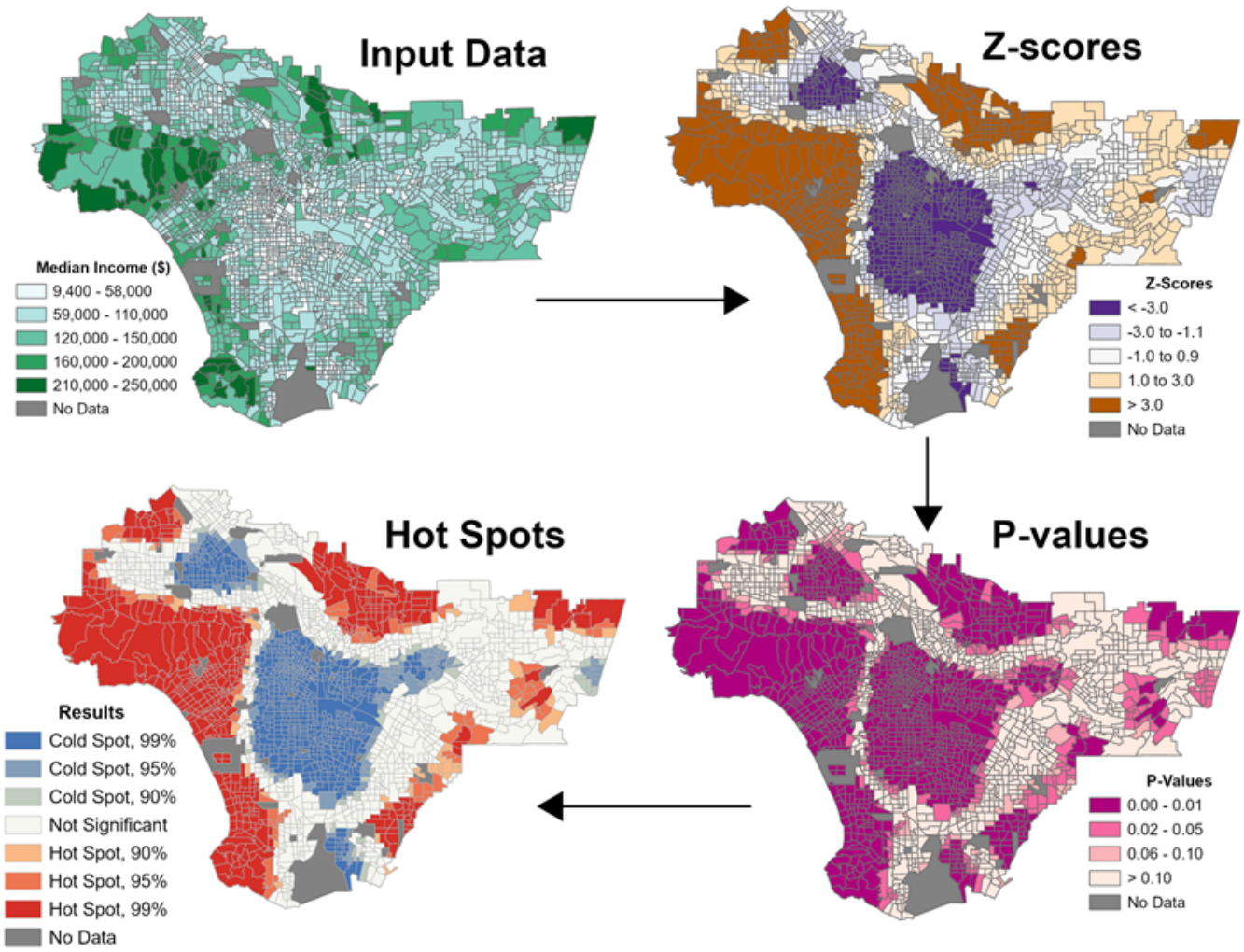


Figure 2. The process of hot spot analysis. Data: American Community Surveys 2018-2022. Source: Author

Figure 3 shows two different visualizations of median income in the Los Angeles metro area. In both, median income values have been transformed into z-scores and p-values. The difference lies in the conceptualization of space. The probability map on the left shows significant tracts on an individual case-by-case basis. Each tract in the map on the right is classified based on whether the mean value of its neighborhood (in this case, every tract within a 5 km buffer) is significantly different from the global mean value. The inclusion of neighborhood smooths the data and clearly highlights regions of high and low values.

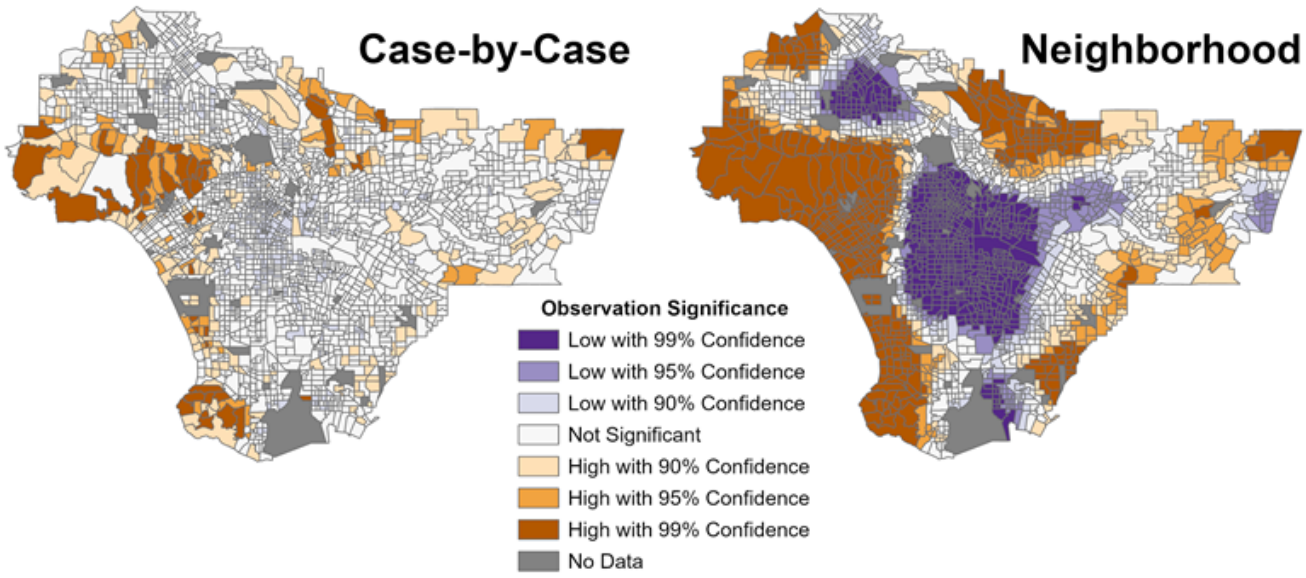


Figure 3: Case-by-case probability map results compared to z-scores calculated by fixed distance (5km) neighborhoods. Data: American Community Surveys 2018-2022. Source: Author.

4. Conceptualization of Spatial Relationships

The most difficult decision in clustering methods generally is the correct specification of spatial relationships within the dataset. Unfortunately, the true form of the relationships is rarely known prior to analysis. Moreover, misspecification of neighborhood type is likely to introduce Type I error and over-assigning significance when it is not appropriate (Rogerson, 2024). Therefore, model fit, as far as a high degree of significant results, is not a proper way to determine neighborhood.

The maps created using different neighborhood specifications may result in wildly dissimilar patterns. Figure 4 illustrates three common neighborhood types and the resulting maps.

Most programs provide a variety of neighborhood types. Fixed distance will create a circular buffer of a specified radius around each case. Nearest neighbors selects the specified number of the closest units. Inverse distance is a linear weighting function based on distance ranging from 1 (closest) to 0 (outside the specified threshold value). Inverse distance squared creates a steeper slope. Other common specifications include contiguity (edges), contiguity (edges and corners), and custom weight matrices.

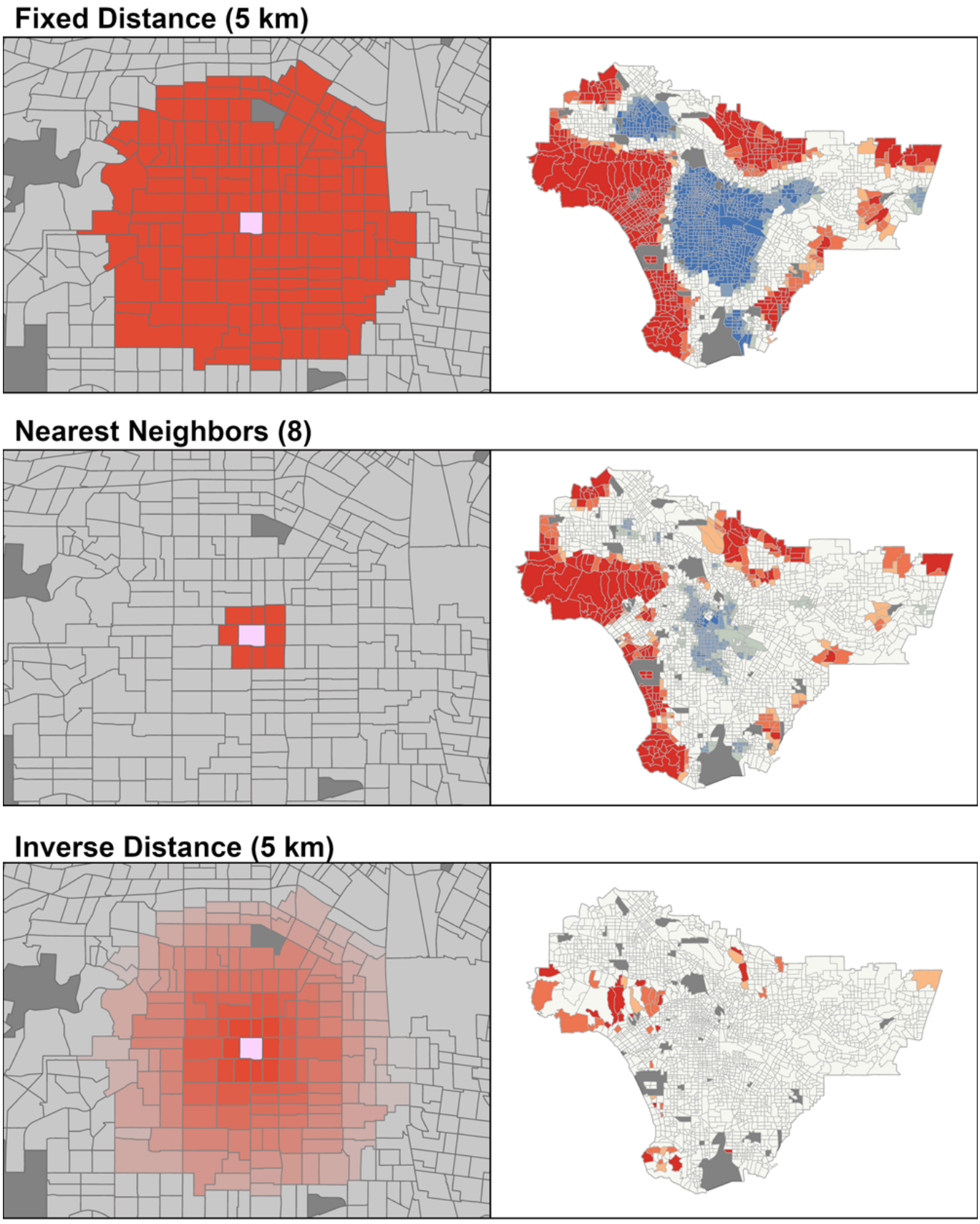


Figure 4. Three common neighborhood types and the resulting hot spot maps. The pink tract is the target unit. Pink and dark orange tracts are assigned a weight of one, while gray tracts are excluded from the calculations. Partial weights are assigned only in the inverse distance condition. Data: American Community Surveys 2018-2022. Source: Author.

As with all statistics, the best decisions are based on theory and the specific context of the data. However, there are a few rules of thumb. First, the recommended default neighborhood is fixed distance (Grekousis, 2020). When in doubt, try fixed distance first. Second, every unit should have at least eight neighbors (Grekousis, 2020). Most GIS include an algorithm to calculate the ideal fixed distance neighborhood so that all units have at least one neighbor and most units have at least eight. The output layer contains a column indicating the number of cases used to compute the values at each unit. This column should be examined before settling on a final conceptualization of distance.

If the Los Angeles median income is analyzed using a 1 km fixed distance neighborhood, the mean number of neighbors is 3.8 and 501 tracts have only one neighbor. The results are likely undergeneralized. In contrast, when the fixed distance band is set at 10 km, the average number of neighbors is 252.9 and 53 tracts have 500 neighbors or more. The default distance set by the program's algorithm is 4.954 km, resulting in a mean of 71.9 neighbors and only 10 cases with fewer than eight neighbors, all larger mountainous tracts near the edge of the study area. Researchers should use caution when interpreting edge data, as with all analysis methods using moving windows, filters, and kernels.

5. Emerging Hot Spot Analysis

Hot spot analysis can be expanded to include changes over time. Emerging hot spot analysis considers whether spatiotemporal neighborhoods vary significantly from the global dataset in both time and space. As the name suggests, spatiotemporal relationships require specifying both conceptualization of spatial relationships and neighborhood time steps or temporal scale. Trends over time are evaluated with the Mann-Kendall trend test, and each spatial unit is assigned a classification (Esri, 2024).

Emerging hot spot analysis introduces a new array of hot/cold spot typologies. This array of hot spots may emphasize consistency (persistent, historical), recency (new, consecutive), magnitude (intensifying, diminishing), or variety (sporadic, oscillating). Every type of cluster has both a hot and cold variant.

Emerging hot spot analysis is not the only technique available for assessing unusually high or low values in spatiotemporal data. Other methods include spatiotemporal versions of KDE, kriging, and scan statistics (Butt at al., 2020). However, the limitations of non-temporal hot spot analysis (correct specification of neighborhood, multiple testing issues, background autocorrelation, etc.) may be magnified by adding another dimension requiring parameterization.

6. Future Directions/Current Frontiers

Hot and cold spots are central to the spatial analysis. Many methods exist, including visual, global statistics, and local approaches. The Getis-Ord Gi* analysis is a popular and versatile method that is often referred to as simply "hot spot analysis" among the rapidly expanding circle of non-spatial subject experts incorporating spatial approaches in their work.

There are some important considerations and limitations to the Getis-Ord Gi* approach. Most critically, definition of neighborhood is often ambiguous. If researchers go fishing for the neighborhood that looks "right" they introduce the modifiable areal unit problem. The

chosen neighborhood may not be appropriate for the data, introducing error and too many significant observations. Hot spot analysis cannot address the small numbers problem or other sources of noise in the data, though it may hide these issues if results are interpreted incorrectly. Rogerson (2024) suggests incorporating Getis's LOSH (local spatial heterogeneity) statistic as a representation of variance in addition to the traditional hot and cold spots based on means. Additionally, there are ongoing conversations about the best way to address the confounding potential of background autocorrelation (Yang, Liu, & Deng, 2023).

Despite these limitations, Getis-Ord Gi* hot spot analysis creates striking visualizations of spatial patterns that are more rigorous than simple choropleth maps and intuitively understood by non-experts. It is an important component of the spatial analysis toolbox in addition to simple choropleth mapping and other global and local explorations of spatial autocorrelation.

References

Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., & Sherazi, H. H. R. (2020). Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review. IEEE Access, 8, 166553–166574.

Cardozo-Stolberg, S., Szydziak, E., Angus, L. D. G., Blake, R., Marsden, D., Venkatesh, S., & Mon, N. O. (2023). Geospatial Mapping to Target Injury Prevention for Older Adult Ground-Level Falls: A Feasibility Study. Journal of Trauma Nursing, 30(6), 340–345.

Esri. (2024). Emerging hot spot analysis (space time pattern mining).

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189-206.

Grekousis, G. (2020). Spatial analysis methods and practice: Describe - explore - explain through GIS. Cambridge: Cambridge University Press.

Murray, A. T. (2020). Spatial Clustering, Detection and Analysis of. In A. L. Kobayashi (Ed.), International Encyclopedia of Human Geography (Second edition.). Elsevier.

Oliveira, U. C. & de Oliveira, P. S. (2017). Heat Maps Applied to Environmental Management: An Analysis of Hot Spots in Acaraú River Basin, Ceará, 2010-2015. Espaço Aberto, 7(1), 87–99.

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis, 27(4), 286-306.

Rogerson, P. A. (2024). Inspired by Art. Journal of Geographical Systems, 26(2), 235–248.

Schiewe, J. (2019). Empirical Studies on the Visual Perception of Spatial Patterns in Choropleth Maps. Kartographische Nachrichten, 69(3), 217–228.

Snow, J. (1855). On the Mode of Communication of Cholera (2nd ed.). London, U.K.: J. Churchill.

Weeks, J. R. (2024). The importance of Arthur Getis to spatial demography. Journal of Geographical Systems, 26(2), 201–208.

Xie, Y., Shekhar, S., & Li, Y. (2023). Statistically-robust clustering techniques for mapping spatial hotspots: A survey. ACM Computing Surveys, 55(2), 1-38.

Yang, J., Liu, Q., & Deng, M. (2023). Spatial hotspot detection in the presence of global spatial autocorrelation. International Journal of Geographical Information Science: IJGIS, 37(8), 1787-1817.