1397 Location Intelligence in Supply Chains 1 (LI 1)

Assignment 1: Short literature review on exploratory point pattern analysis

Maciej Kilijański, h12018188

Spatial data, as defined by Fischer, Leitner, and Staufer-Steinnocher (2001), refers to data where the absolute or relative positioning of entities is explicitly considered. This type of data is essential in many fields, as it allows for the analysis of spatial relationships, patterns, and processes. In their work, Fischer et al. (2001) categorize spatial data into four main types: spatial point patterns, area data, field data, and spatial interaction data. Each category reflects a different way of capturing spatial variation - whether those would be specific event locations (spatial point patterns), attributes distributed over defined areas (area data), continuous variation across space (field data), or interactions between spatial units (spatial interaction data).

For the purposes of this literature review, we will focus exclusively on spatial point patterns. Spatial point patterns refer to the distribution of events at specific locations within a study area. Understanding these patterns is critical for identifying clustering, regularity, or randomness in spatial phenomena, such as the spread of diseases, distribution of services, or the positioning of retail outlets. By grounding this review in the framework provided by Fischer et al. (2001), we aim to provide a comprehensive summary of the key methods and insights derived from the analysis of spatial point patterns in geographic research.

Spatial point pattern analysis is concerned with two kinds of information: the points and the region these are located in. Considering the region as well might be crucial in order to understand some external factors impacting the distribution of the points analyzed, hence at the very beginning of an analysis one should learn some characteristics about the geography analyzed (Fischer, Leitner and Staufer-Steinnocher, 2001). Each of the analyses can have multiple orders of properties, depending on the level of locality of the hypothesis. The baseline paper for this literature review determines two: first being the intensity of the points, allowing to determine their location pattern and the second one being relationships between the points themselves, such as spatial dependance.

Fishter, Leitner and Staufer-Steinnocher (2001) start describing steps for a spatial point pattern analysis with a concept called Complete Spatial Randomness (CSR). In hindsight, it is necessary to first reject the null hypothesis of points being randomly distributed in the analyzed space in order to carry on with any further analyses. The authors determine two possible scenarios for points not being randomly distributed:

- Clustering – when points are distributed close to each other and create statistically sound hotspots with higher-than-average density and cold spots, where the density is lower.
- Regularity – when points are regularly distributed across the space, not creating any sort of clusters. Distances between all points tend to be very similar.

The authors then explain four techniques one can utilize to explore the Spatial Points Patterns.

The simplest ones being so-called Quadrat Methods. In hindsight, one should arbitrarily, upon previously conducting necessary research of the region all points are located in, come up with a fitting grid size and then divide the region into parts of such size. The next step would be to count points that happen to exist in each of the cells. This then gives all necessary data to plot a histogram of points density. Such plot can then be compared to a completely randomly generated test sample, leading to CSR's null hypothesis rejection. However, this method is rather prone to human error, as it relies heavily on the grid size used.

A second, more sophisticated way of dealing with finding clusters, described by Fishter, Leitner and Staufer-Steinnocher (2001) are the Kernel Estimators. It helps with smoothing the results of previously described Quadrat Methods. Here, the choice of smoothing constant is crucial, as it impacts the densities received. The larger the smoothing constant, the flatter the results will be, with lowering the constant increasing spikes in results. Here, once again we must rely on the expertise on the region of a researcher performing the analysis.

Third, there are the distance methods which utilize the exact locations of events to provide insights into spatial dependencies without the arbitrary partitioning of the study region. These methods focus on distances between events, offering a more direct means of identifying clusters or regularity. There are two key measures used:

- Event-to-event nearest neighbor distances (W) – the distance between a randomly selected event and the nearest event to it.
- Point-to-event nearest neighbor distances (X) – the distance from a randomly chosen point in the region to the nearest observed event.

After calculating these distances, a cumulative probability distribution is computed. It can then be plotted and checked against the random distribution, which will result in a function with a slope of 45 degrees. If the resulting function gains value quickly, with low distances between events, later slowing down, that suggest clustering. If the function climbs at the end of the distances, this suggests regularity, with very similar distances between all points measured.

For a deeper, multi-scale look at spatial patterns, the K-Function is elaborated on by Fischer, Leitner, and Staufer-Steinnocher (2001). It allows us to go beyond nearest neighbor distances, offering a full analysis of spatial dependence over a range of distances. Essentially, the K-Function examines how the number of events within a given distance from any event compares to what we should expect under a CSR.

What makes the K-Function especially useful is that it gives us a detailed view of spatial relationships across distances, not just at a single scale. Unlike quadrat or distance methods, it can handle various complexities in the data. It also stands out as the only method suggested, that does not require any human judgement to be executed.

The authors then give a glossary on how to test for CSR using these methods, with the K-Function standing out as the most logical to me. Upon generating the random sample of points, one should calculate the K-Function for these and create upper and lower bound for a Confidence Interval. Then, one plots the actual K-function against these bounds and checks the result. If the actual result is between the bounds, we cannot reject the CSR’s null hypothesis. The result being higher than the upper bound suggest clustering and a lower one suggests regularity.

There are some new, exciting ways of discovering clustering patterns in a dataset. One of such was described in a recent article by Kalinová (2021), where she explored these with use of Artificial Intelligence. When we compare this study to Fischer, Leitner, and Staufer-Steinnocher’s (2001) discuss various methods, which are easier to grasp and less computing-power-demanding, for identifying spatial patterns, while Kalinová (2021) uses neural networks for clustering, which similarly helps to reveal hidden patterns in data, if such would arise with inclusion of let’s say the area data.

Having rejected the CSR and having seen potential clustering of points one can proceed and conduct a Hotspot Analysis. A good procedure of such was described by Bambrick (2016). Both Moran's I and Getis-Ord General G will result in p and z-values being returned for each region (polygon or grid) considered. As with all statistical problems, a low p-value signals the significance of measurement. The z-value signals either a hot or a cold spot. In other words, high z-value signals a cluster and low z-value lack of thereof.

Practical applications of methods described by Bambrick (2016) and Fischer, Leitner, and Staufer-Steinnocher (2001) are plenty. For instance, determining areas with need for sea traffic safety improvements. In the article "Improving Near Miss Detection in Maritime Traffic in the Northern Baltic Sea from AIS Data," Du et al. (2021) employed both clustering and hotspot analysis to enhance the detection of near-miss events in maritime traffic. Clustering was used to categorize ship behaviors and identify risk patterns, especially vessel encounters. The clustering process involved analyzing Automatic Identification System (AIS) data, which provides real-time tracking information for vessels around the world, such as location, speed, and course.

Hotspot analysis was then applied to detect areas of high collision risk – "hotspot" zones. These areas were identified by analyzing the spatial distribution of near-miss events, focusing on high-traffic zones where collision risks are more frequent. The study used the analysis to highlight critical regions in the Northern Baltic Sea where maritime safety measures should to be intensified. For instance, hotspots were found in the Gulf of Finland, around major ports, and in other busy maritime traffic areas where navigational complexity and traffic density increase the likelihood of accidents.

Another use of the Hotspot analysis (Bambrick, 2016) can be found in an article by Aati et al. (2024) which identifies areas with a high frequency of road traffic accidents (RTAs) in densely populated cities. Hotspot analysis was employed using geographic information system (GIS) tools, specifically ArcGIS, to spatially cluster accident data and identify locations with a higher risk for traffic incidents. The Getis-Ord General G statistic was the primary tool used for this analysis, which allowed the researchers to map areas with statistically significant concentrations of accidents, and distinguish hotspots (areas with high accident rates) from cold spots (areas with lower accident rates).

Hotspot analysis was also employed in Baker (2010) in order to identify regions of high pollutant concentrations related to air mass transport patterns. Hotspot analysis was integrated with cluster analysis of four-day trajectories to assess how various air mass patterns contributed to elevated pollution levels in the UK, focusing on Birmingham and Harwell (the regions), two distinct monitoring sites.

A k-means clustering algorithm was applied to group air mass trajectory sets into clusters, which represented distinct synoptic transport patterns. These clusters were then used to investigate pollution hotspots by associating specific air mass pathways with pollutant concentrations. Each cluster type reflected different geographic origins of air masses and their associated pollution loads, particularly in terms of ozone, NOx, PM10, and other pollutants. Hotspot analysis revealed that specific air masses, such as the slow-easterly air mass passing over continental Europe, were consistently associated with the highest pollution levels, forming distinct pollution hotspots in both Birmingham and Harwell. In contrast, air masses arriving from the Atlantic Ocean (e.g., south-westerly and strong-westerly clusters) resulted in lower pollutant concentrations, thus acting as "cold spots" in terms of air quality.

The method can also be used for crime measuring and prevention, as shown by Kalinic and Krisp (2018). Hotspot analysis, in this context, was used to complement kernel density estimation (KDE), which showed where crime clusters existed, but couldn't determine whether these clusters were statistically meaningful.

The authors aggregated crime incidents and detected areas with intense clustering of high crime counts (hot spots) and low crime counts (cold spots). The analysis resulted in the identification of zones with significant clustering of high crime values, highlighted by dark red and bright red hexagons that corresponded to 99% and 95% confidence levels, respectively. What is interesting, the analysis did not identify any statistically significant cold spots.

One can also try to identify areas of high tourist concentration within specific scenic spots, as presented by Liao (2020). The core of the analysis was based on DBSCAN algorithm (another frequently used clustering algorithm), which was used to cluster tourist stay points from trajectory data collected through an app, allowing to identify specific locations where tourists stayed the longest.

Once the clustering was completed, Getis-Ord General G was applied to further analyze and map these stay points, helping visualize the intensity and significance of tourist concentrations. This analysis quantified the level of clustering by assigning high z-values to the clusters, determining which tourist attractions were the most popular based on visitor density.

The applications of spatial point pattern analysis and hotspot detection highlight how important these techniques are for understanding complex spatial patterns across scientific fields. The framework provided by Fischer, Leitner, and Staufer-Steinnocher (2001) described the reasoning behind such analysis. As covered in this analysis the Hotspot analysis (Bambrick, 2016) can be used in various applications such as maritime safety for near-miss detection (Du *et al.*, 2021), monitoring long-range pollutant pathways (Baker, 2010), mapping urban crime (Kalinic and Krisp, 2018), or managing traffic and tourist activity (Liao, 2020), providing insights for better decision-making in a variety of situations.

Bibliography

Aati, K. *et al.* (2024) ‘Analysis of Road Traffic Accidents in Dense Cities: Geotech Transport and ArcGIS’, *Transportation Engineering*, 16, p. 100256. Available at: https://doi.org/10.1016/j.treng.2024.100256.

Baker, J. (2010) ‘A cluster analysis of long range air transport pathways and associated pollutant concentrations within the UK’, *Atmospheric Environment*, 44(4), pp. 563–571. Available at: https://doi.org/10.1016/j.atmosenv.2009.10.030.

Bambrick, G. (2016) ‘What is Hotspot Analysis?’, *Geospatiality*, 21 January. Available at: https://glenbambrick.com/2016/01/21/what-is-hotspot-analysis/ (Accessed: 14 October 2024).

Du, L. *et al.* (2021) ‘Improving Near Miss Detection in Maritime Traffic in the Northern Baltic Sea from AIS Data’, *Journal of Marine Science and Engineering*, 9(2), p. 180. Available at: https://doi.org/10.3390/jmse9020180.

Fischer, M., Leitner, M. and Staufer-Steinnocher, P. (2001) ‘Spatial point pattern analysis: Some useful tools for analysing locational data’.

Kalinic, M. and Krisp, J.M. (2018) ‘Kernel density estimation (KDE) vs. hot-spot analysis– detecting criminal hot spots in the City of San Francisco’, *Lund, Sweden* [Preprint].

Kalinová, E. (2021) ‘Artificial Intelligence for Cluster Analysis: Case Study of Transport Companies in Czech Republic’, *Journal of Risk and Financial Management*, 14(9), p. 411. Available at: https://doi.org/10.3390/jrfm14090411.

Liao, Y. (2020) ‘Hot Spot Analysis of Tourist Attractions Based on Stay Point Spatial Clustering’, *Journal of Information Processing Systems*, 16(4), pp. 750–759. Available at: https://doi.org/10.3745/JIPS.04.0177.