Mathematics and Society
University of Isfahan
http://math-sci.ui.ac.ir

Journal of Mathematics and Society
ISSN (print): 2345-6493, ISSN (on-line): 2345-6507
Vol. 9 No. 3 (2024), pp. 107-124.

http://www.ui.ac.ir

INVESTIGATING POPULATION BALANCE IN THE STATE OF GEORGIA USING SPATIAL CLUSTERING

ALI ABOLHASSANI* AND SOMAYYEH TARI

Abstract. Identification of population ratio disruption in the population structure is one of the challenges that every country faces. Population aging is a kind of demographic abnormality that lack of attention causes population problems. A timely warning about the aging of society can be useful for planning about having children on the one hand and providing suitable facilities for the elderly on the other hand. For example, the capital of Japan is a good example of an urban environment suitable for the elderly.

One of the anomaly detection tools is spatial clustering using scan statistics. In the last three decades, the scan statistic method has been a very important and active field in statistical research. Identifying areas on geographic maps, where the concentration of points (elderly, sick, criminals, certain animal species, etc.) is significant, is important in many fields such as epidemiology, politics, criminology, zoology, and so on. With the help of scan statistic method, spatial clusters can be identified. In this article, we introduce the scan statistic method based on Poisson distribution. Using simulation, we investigate the efficiency of this method in identifying spatial clusters. Based on the results obtained from the simulation, the Poisson scan statistic method is a suitable method for detecting anomalies in the count spatial data. As an application of spatial clustering, we consider the population structure of the state of Georgia and identify areas where the elderly population is significantly high. These areas should be prioritized in the implementation of population reform programs.

Keywords: Spatial clustering, Scan statistic, Monte-Carlo hypothesis testing, Poisson distribution, Likelihood ratio

Article Type: Research Paper.

Communicated by Afshin Parvardeh.

*Corresponding author.

Received: 12-06-2023, Accepted: 27-04-2024, Published Online: 26-11-2024.

This work has been financially supported by Azarbaijan Shahid Madani University under the grant number 906/1402.

Cite this article: A. Abolhassani and S. Tari, Investigating population balance in the state of Georgia using spatial clustering, *Journal of Mathematics and Society*, **9** no. 3 (2024) 107–124.

http://dx.doi.org/10.22108/msci.2024.138014.1585 .

1. Introduction

The outbreak of diseases such as COVID-19 has significantly impacted communities, with some cities experiencing widespread transmission while others have lower prevalence rates. Identifying regions with a high density of infected individuals that hold statistical significance is crucial. Timely alerts to governmental and health authorities about the existence of such areas can effectively prevent human casualties and alleviate the burden of additional healthcare costs on the government. Pinpointing these areas may even yield insights into the causative factors of the disease. For instance, during the cholera epidemic, Dr. John Snow mapped the geographical locations of individuals who succumbed to the disease in London and discerned a correlation between mortality due to cholera and contamination of drinking water sources, as most individuals affected were in proximity to these water sources [30]. This paper presents the map of cholera in London in Figure 1 which is generated using R software. Additionally, interested readers can refer to [20] to view the original map designed by Snow.

While mortality rates may be higher around water sources, such concentrations may occur randomly. Hence, investigating the significance of these clusters is imperative. When population density with a specific characteristic (such as being infected, elderly, or obese) in an area is statistically significant, we can infer spatial clustering has occurred. One statistical tool for detecting these areas is the scan statistic method, introduced by Kulldorff [18].

Identifying spatial clusters is not only vital in epidemiology but also applicable across various research domains: astronomy [19], [5], [9], [16], image analysis [12], [34], [4], criminology [22], [29], [14], [8], [10], ecology [31], [23], geography [24], [6], [21], [28], pattern recognition [13], biology [11], [25], epidemiology [27], [17], [7], [33], public transportation [32], [3], [15]. A comprehensive review of the scan statistic method is provided in [2].

2. Main Results

Kulldorff and Nagarwalla [18] introduced the Likelihood Ratio Test (LRT) method for spatial cluster detection. In this method, a two-dimensional window scans the study area map, and the window size changes during scanning. After scanning the map, spatial clusters of *points* or *cases* are examined based on a statistical hypothesis test introduced in equation 2.1. It is worth mentioning that each member of the population with a specific attribute is called a *point* or *case* and the term *control* is applied to members of the population that lack that specific attribute.

Let's assume that the study map has I cities or cells, and for the ith city, $i = 1, \ldots, I$, the number of cases and the population are denoted by x_i and n_i, respectively. Also, suppose the number of cases in each city follows a Poisson distribution, i.e., $X_i \sim Pois(n_i\theta_i)$. Each sub-area Z of the map can be considered as a candidate cluster. Thus, the class of candidates will have $2^{|I|} - 1$ members, where $|I|$ is the cardinality of the set I.

http://dx.doi.org/10.22108/msci.2024.138014.1585

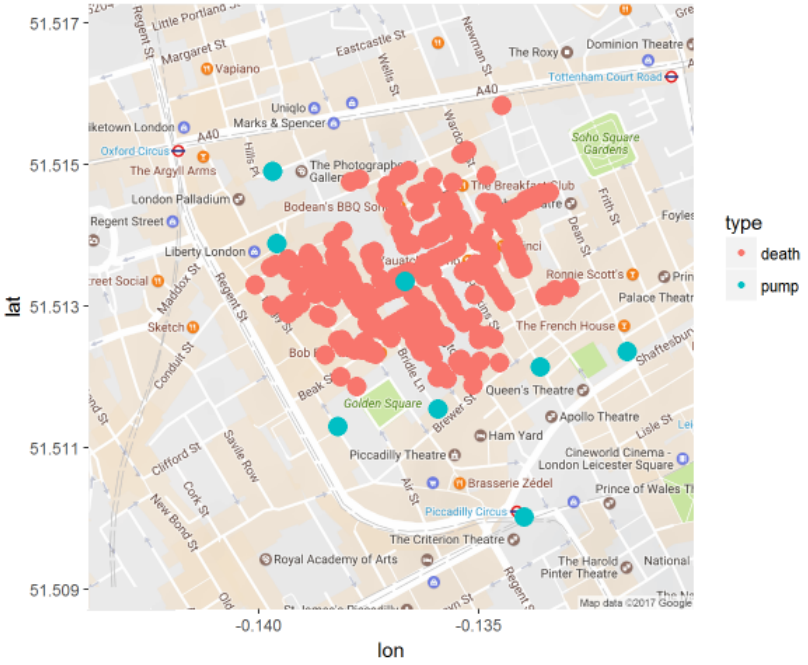

FIGURE 1. The geographic location of water pumps (blue) and cholera patient fatalities (red) on the map of London in 1856.

The class containing all possible candidates is denoted by $\mathcal{Z}$. The total number of cases in Z is $X_Z = \sum_{i\in Z} X_i$. Similarly, $X_{\bar{Z}} = \sum_{i\in \bar{Z}} X_i$ represents the total number of cases in $\bar{Z}$, the complement of Z. The total population in areas Z and $\bar{Z}$ are $n_Z = \sum_{i\in Z} n_i$ and $n_{\bar{Z}} = \sum_{i\in \bar{Z}} n_i$, respectively. Thus, the total population on the map is $n = n_Z + n_{\bar{Z}}$. Suppose $\theta_i = \theta_Z$ for each area $i \in Z$, and similarly, $\theta_i = \theta_0$ for each area $i \in \bar{Z}$. To find the cluster location on the map, a hypothesis test

$$H_0 : \theta_Z = \theta_0, \forall Z \in \mathcal{Z} \qquad v.s. \qquad H_1 : \exists Z \in \mathcal{Z}, \theta_Z > \theta_0 \tag{2.1}$$

is conducted.

Each member of $\mathcal{Z}$ can be considered a potential cluster. Since the actual cluster location is unknown, Z is treated as a parameter. The likelihood function is

$$L(Z) = L(Z, \theta_0, \theta_Z) = \Big[\Pi_{i\in Z} \frac{e^{-n_i\theta_Z}(n_i\theta_Z)^{x_i}}{x_i!}\Big]\Big[\Pi_{j\notin Z} \frac{e^{-n_j\theta_0}(n_j\theta_0)^{x_j}}{x_j!}\Big]. \tag{2.2}$$

Hence, the likelihood ratio is

$$\lambda(Z) = \frac{\sup_{\theta_Z>\theta_0} L(Z, \theta_Z, \theta_0)}{\sup_{\theta_Z=\theta_0} L(Z, \theta_Z, \theta_0)}. \tag{2.3}$$

The scan statistic is defined as $\lambda = sup_{Z\in\mathcal{Z}}\lambda(Z)$.

To find the spatial cluster, $\lambda(Z)$ is calculated for each $Z \in \mathcal{Z}$. The next step is to determine the region Z that maximizes $\lambda(Z)$, known as the Most Likely Cluster (MLC). Since the distribution of

the scan statistic is unknown, hypothesis testing is performed using the Monte Carlo method. This method is described as follows:

To perform the test using the Monte Carlo method, the number of cases under the null hypothesis in hypothesis testing 2.1 is first generated for each city. The value of $\lambda(Z)$ is then calculated based on the simulated data to determine the MLC. This process is repeated 499 times, and the scan statistic is calculated in each iteration. Let the computed values be $\lambda_i, i = 1, 2, \ldots, 499$. These values are sorted in ascending order, i.e., $\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(499)}$. The value of λ_R for the actual data is placed among the sorted values. The p-value is then calculated as $\frac{1+\sum_{i=1}^{499} I(\lambda_i > \lambda_R)}{500}$, where $I()$ is the indicator function. It should be noted that the detection of circular and non-circular cluster are discussed in [1].

Five criteria are introduced to measure the efficiency of spatial clustering methods: Recall, Precision, F1 score, Biasness, and Power:

$$\text{Recall} = \frac{\#(\text{Identified Cluster} \cap \text{True Cluster})}{\#(\text{True Cluster})}$$

$$\text{Precision} = \frac{\#(\text{Identified Cluster} \cap \text{True Cluster})}{\#(\text{Identified Cluster})}$$

where $\#(A)$ denotes the number of members of set A. The F1 score is the harmonic mean of Precision and Recall and is defined as:

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Biasness can also be a measure for comparing clustering methods. Prates et al. [26] discussed Biasness in clustering methods. It is noteworthy that the closer these metrics are to one, the better the method has identified the true cluster with fewer errors. Therefore, one is the optimal value for these metrics.

Another criterion for assessing the effectiveness of clustering methods is the power of the test. It is the probability of correctly rejecting the null hypothesis. If a method used for clustering has lower power, it means that it is less likely to detect a cluster in a map containing a cluster.

3. Simulation, real application, and conclusions

In this section, we delve into simulation studies considering various scenarios. Initially, we consider a map similar to Figure 2. Each cell of the map includes the number of patients, the total number of individuals in that cell, and the geographical coordinates of its center. We utilize a Poisson distribution to simulate the number of patients in each cell. In the first scenario, the number of patients inside the purple region (cluster location) is generated from a $Pois(24)$. Outside this region, the number of patients is generated from a $Pois(20)$. We perform cluster detection according to the method described in Section 2 and compute the bias, power, precision, recall, and F1 score. We repeat the simulation of maps under the null hypothesis 499 times. In each iteration, we calculate the introduced metrics.

http://dx.doi.org/10.22108/msci.2024.138014.1585

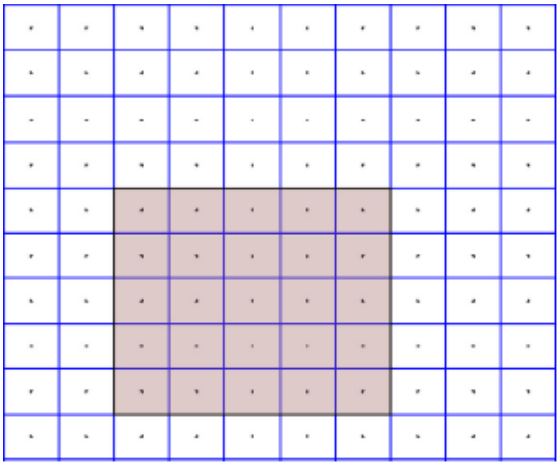

FIGURE 2. A map with 100 cells and a spatial cluster (purple area).

The results related to this scenario are presented in Table 1, in the first column from the right. As observed, the test power is low, around 0.19. This means that in this scenario, the presence of a cluster is only alerted by the scan statistic in 19% of the repetitions. On the other hand, bias, recall, precision, and F1 score deviate from the optimal value of 1, indicating that the identified clusters track the true cluster with more errors.

In the second scenario, the parameter inside the cluster is set to 30, and the parameter outside the cluster is set to 20. Similar to the previous stage, we perform cluster detection and calculate the five introduced metrics. We expect these metrics to be closer to the optimal value compared to the previous scenario. The results are shown in Table 1, in the middle column. Not only has the test power increased from 0.19 to 0.93, but also the values of other metrics have become closer to the optimal value. In other words, the presence of a cluster is alerted in 93% of the repetitions, and the actual cluster location is more accurately tracked.

In the final scenario, we consider the parameter value of 60 for the Poisson distribution inside the cluster and 20 for the distribution outside the cluster. The results of cluster detection are presented in Table 1, in the leftmost column. In this scenario, in all iterations, a cluster presence alert is issued (test power equals 1), and the identified cluster exactly matches the true cluster.

To work on real data, we use the `georgia` dataset available in the `GISTools` package in the `R` software. This dataset includes 14 variables, including the geographical longitude and latitude of each region, its population, the elderly population in each region, and so on. The map of the state of Georgia, which consists of 159 regions is in Figure 3.

TABLE 1. Simulation results for three different scenarios

	$Pois(20)$ vs $Pois(60)$	$Pois(20)$ vs $Pois(30)$	$Pois(20)$ vs $Pois(24)$
Bias	1.01	1.05	1.1752
Recall	1	1	0.81
Precision	1	1	1.9
Power	1	0.93	0.19
F1	1	1	0.83

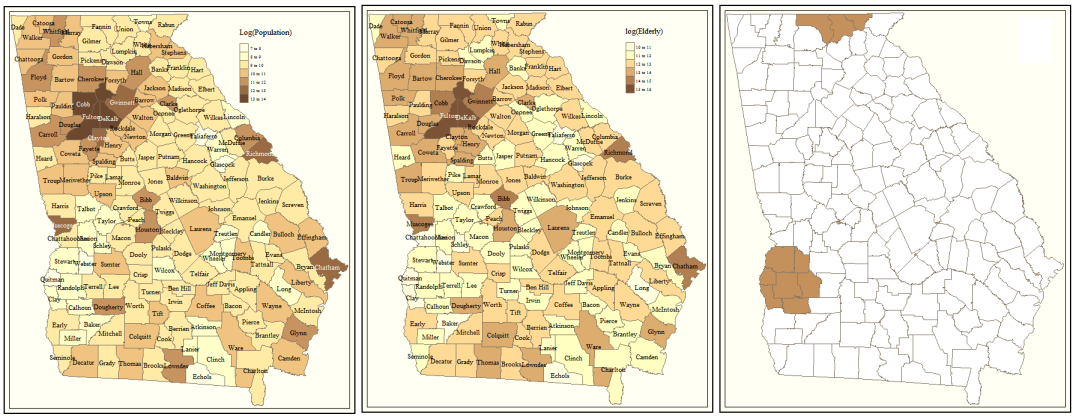

FIGURE 3. The map of the state of Georgia. Left figure: Color coded based on population count, middle figure: Color coded based on the elderly rate, right figure: Location of identified clusters.

We intend to examine which regions have a higher population density of the elderly and whether this density is statistically significant. Using scan statistic method, two clusters are identified in the western and northern regions of Georgia. In the identified areas, the elderly population is higher than in other areas. More precisely, these areas are relatively older compared to other areas. Therefore, it is recommended that these areas be prioritized in population planning, and appropriate measures be taken in terms of urban infrastructure for the elderly and providing conditions for fertility for young people.

The final conclusion of the article is as follows: population aging is a demographic anomaly that can lead to population-related problems if not addressed. Timely warnings about the aging of society can be useful for planning both in terms of fertility and providing suitable facilities for the elderly. Currently, Iran is also experiencing the transition phase of the age structure of the population from youth to elderly, and it seems that by the beginning of the fifteenth century (Solar Hijri), we are facing an increase in the elderly population in the country. Therefore, it is necessary to issue timely warnings

http://dx.doi.org/10.22108/msci.2024.138014.1585

to control population aging and provide appropriate urban and welfare facilities for this segment of society through thoughtful measures.

References

[1] A. Abolhassani, *Some new methods in spatial clustering*, Doctoral dissertation in statistics, Isfahan University of Technology 2020, [In Persian].

[2] A. Abolhassani and M. O. Prates, An up-to-date review of scan statistics, *Stat. Surv.*, **15** (2021) 111–153.

[3] A. Abolhassani, M. O. Prates, F. Castellares and S. Mahmoodi, Zero-inflated Bell scan: A more flexible spatial scan statistic, *Spat. Stat.*, **36** (2020) 18 pp.

[4] A. Abolhassani, M. O. Prates and S. Mahmoodi, Irregular shaped small nodule detection using a robust scan statistic, *Statistics in Biosciences*, **15** (2023) 141–162.

[5] K. L. Adelberger, C. C. Steidel, M. Pettini, A. E. Shapley, N. A. Reddy and D. K. Erb, The spatial clustering of star-forming galaxies at redshifts $1.4 \leq z \leq 3.5$, *The Astrophysical Journal*, **619** (2005) 619–697.

[6] L. Anselin, Local indicators of spatial association—LISA, *Geographical Analysis*, **27** (1995) 93–115.

[7] L. H. Duczmal, G. J. P. Moreira, D. Burgarelli, R. H. Takahashi, F. C. Magalhães and E. C. Bodevan, Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town, *International Journal of Health Geographics*, **10** (2011) 1–4.

[8] J. Eck, S. Chainey, J. Cameron and R. Wilson, *Mapping crime: Understanding hotspots*, 2005.

[9] M. D. Gladders and H. K. Yee, A new method for galaxy cluster detection. I. The algorithm, *The Astronomical Journal*, **120** (2000) 2148–2162.

[10] T. H. Grubesic, On the application of fuzzy clustering for crime hot spot detection, *Journal of Quantitative Criminology*, **22** (2006) 77–105.

[11] A. Gutteridge, G. J. Bartlett and J. M. Thornton, Using a neural network and spatial clustering to predict the location of active sites in enzymes, *Journal of Molecular Biology*, **330** (2003) 719–734.

[12] R. Haralick and I. H. Dinstein, A spatial clustering procedure for multi-image data,*IEEE Trans. Circuits and Systems*, **22** (1975) 440–450.

[13] R. M. Haralick and G. L. Kelly, Pattern recognition with measurement space and spatial clustering for multiple images, *Proceedings of the IEEE*, **57** (1969) 654–665.

[14] K. D. Harries, *Mapping crime: Principle and practice*, US Department of Justice, Office of Justice Programs, National Institute of Justice, Crime Mapping Research Center, 1999.

[15] L. Huang, D. G. Stinchcomb, L. W. Pickle, J. Dill and D. Berrigan, Identifying clusters of active transportation using spatial scan statistics, *Am. J. Prev. Med.*, **37** (2009) 157–166.

[16] R. S. Kim, J. V. Kepner, M. Postman, M. A. Strauss, N. A. Bahcall, J. E. Gunn, R. H. Lupton, J. Annis, R. C. Nichol, F. J. Castander and J. Brinkmann, Detecting clusters of galaxies in the sloan digital sky survey. i. monte carlo comparison of cluster detection algorithms, *Astron. J.*, **123** (2002) 20–36.

[17] M. Kulldorff, A spatial scan statistic, *Comm. Statist. Theory Methods*, **26** (1997) 1481–1496

[18] M. Kulldorff and N. Nagarwalla, Spatial disease clusters: detection and inference, *Stat. Med.*, **14** (1995) 799–810.

[19] H. J. Mo and S. D. White, An analytic model for the spatial clustering of dark matter haloes, *Mon. Not. R. Astron. Soc.*, **282** (1996) 347–361.

[20] M. Mohammadzadeh, *Spatial statistics and its applications,* Tarbiat Modares university, 2019. [In Persian]

[21] A. T. Murray and V. Estivill-Castro, Cluster discovery techniques for exploratory spatial data analysis, *Int. J. Geogr. Inf. Sci.*, **12** (1998) 431–443.

[22] A. T. Murray, T. H. Grubesic and R. Wei, Spatially significant cluster detection, *Spat. Stat.*, **10** (2014) 103–116.

[23] N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. Da Fonseca and J. Kent, Biodiversity hotspots for conservation priorities, *Nature*, **403(6772)** (2000) 853–858.

[24] J. K. Ord and A. Getis, Local spatial autocorrelation statistics: distributional issues and an application, *Geographical Analysis*, **27** (1995) 286–306.

[25] D. Pellin and C. Di Serio, A novel scan statistics approach for clustering identification and comparison in binary genomic data, *BMC Bioinformatics*, **17** (2016) 61–71.

[26] M. O. Prates, M. Kulldorff and R. M. Assunção, Relative risk estimates from spatial and space-time scan statistics: are they biased?, *Stat. Med.*, **33** (2014) 2634–2644.

[27] C. J. Ribeiro, A. D. Dos Santos, S. V. Lima, E. R. da Silva, B. V. Ribeiro, A. M. Duque, M. V. Peixoto, P. L. Dos Santos, I. M. de Oliveira, M. W. Lipscomb and K. C. de Araújo, Space-time risk cluster of visceral leishmaniasis in Brazilian endemic region with high social vulnerability: an ecological time series study, *PLoS Neglected Tropical Diseases*, **15** (2021) 1–20.

[28] P. Rogerson and I. Yamada, *Statistical detection and surveillance of geographic clusters, CRC Press*, 2008.

[29] L. W. Sherman and D. Weisburd, General deterrent effects of police patrol in crime hot spots: A randomized, controlled trial, *Justice Quarterly*, **12** (1995) 625–648.

[30] J. Snow, *On the mode of communication of cholera*, John Churchill, 1849.

[31] T. J. Stohlgren, D. Binkley, G. W. Chong, M. A. Kalkhan, L. D. Schell, K. A. Bull, Y. Otsuki, G. Newman, M. Bashkin and Y. Son, Exotic plant species invade hot spots of native plant diversity, *Ecological Monographs*, **69** (1999) 25–46.

[32] Y. Tanoue, D. Yoneoka, T. Kawashima, S. Uryu, S. Nomura, A. Eguchi, K. Makiyama and K. Matsuura, Public transportation network scan for rapid surveillance, *Biostatistics and Epidemiology*, **7** (2022) 1–15.

[33] S. C. Wieland, J. S. Brownstein, B. Berger and K. D. Mandl, Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes, *Proceedings of the National Academy of Sciences*, **104** (2007) 9404–9409.

[34] L. Zhang and Z. Zhu, Spatial multiresolution cluster detection method, (2012), arXiv preprint arXiv.

Ali Abolhassani

Department of Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran

Email: ali.abolhassani@azaruniv.ac.ir

Somayyeh Tari

Department of Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran

Email: s_tari@azaruniv.ac.ir

 http://dx.doi.org/10.22108/msci.2024.138014.1585