The College Mathematics Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucmj20

An Investigation of Metropolitan Crime Distribution

Lindsey Bell & Paul Hill

To cite this article: Lindsey Bell & Paul Hill (2022) An Investigation of Metropolitan Crime Distribution, The College Mathematics Journal, 53:5, 364-371, DOI: 10.1080/07468342.2022.2125263

To link to this article: https://doi.org/10.1080/07468342.2022.2125263

Published online: 14 Nov 2022.

Submit your article to this journal

Article views: 268

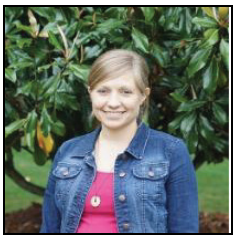
View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ucmj20

An Investigation of Metropolitan Crime Distribution

Lindsey Bell and Paul Hill



Lindsey Bell (lbell2@coastal.edu) received her Ph.D. in Biostatistics from Florida State University. She is currently associate professor of statistics at Coastal Carolina University. Besides academic pursuits, Lindsey enjoys playing with her young children and all things outdoors.

Paul Hill (pchill@coastal.edu) earned his Ph.D. in Statistics from Florida State University. He is currently an assistant professor at Coastal Carolina University. Paul has a profound passion for teaching and supporting students in their academic pursuits. Outside of pedagogical interests, Paul enjoys watching and playing soccer as well as spending time with family and friends.

Have you ever researched crime data for your neighborhood? There are a variety of websites that offer crime map filters that display the distribution of various crimes by location. The distribution of crimes by location is of importance to many. From the inquiry of a potential homeowner to the deployment of resources for law enforcement, the dispersion or spatial arrangement of crimes can provide insights into a community. Several factors such as the degree of urbanization, unemployment and other socioeconomic demographics may influence the distribution of crimes in a particular location.

In the following discussion, we examine the distribution of crimes in the District of Columbia (DC). In particular, we evaluate whether the spatial arrangement of crimes in the third district of DC is random, uniformly dispersed or clustered. We will consider three methods that may be applied to study the allocation of crimes.

Spatial randomness

We begin our study into the distribution of the crime locations by first defining complete spatial randomness and related vocabulary. Points of interest are called *events*. In our case, an event refers to the location of a crime. A *mark* is a particular characteristic about the event. Type of crime and time of day are possible marks to describe the event. A *window* is the area of interest within the two dimensional space. The third district of DC serves as our window in this study and marks are ignored for simplicity. The set of observed events (locations of the crimes) and its window (third district of DC) is referred to as our *spatial point pattern*.

In the context of this study, *Complete Spatial Randomness* assumes:

- Crimes are equally likely to occur anywhere in DC;
- Crime locations are independent of each other;

doi.org/10.1080/07468342.2022.2125263

 © THE MATHEMATICAL ASSOCIATION OF AMERICA

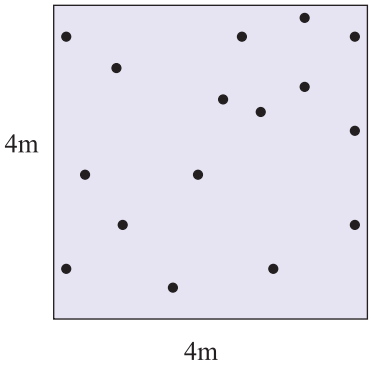


Figure 1. 4m × 4m window.

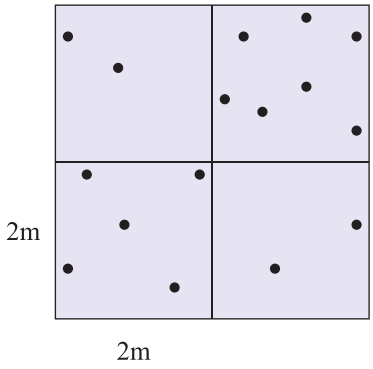


Figure 2. Four 2m × 2m quadrats.

- The number of crimes in a region has a Poisson distribution; and
- The rate of the Poisson distribution is ρ, the density of crimes in DC.

The point density is estimated by $\widehat{\rho}$, the number of events in a window relative to the area of the window. Consider the case where there are 16 events present in a 4m × 4m window as seen in Figure 1. Then we have

$$\widehat{\rho} = \frac{\text{Number of crimes in the window}}{\text{Area of the window}} = \frac{16 \text{ events}}{4\text{m} \times 4\text{m}} = 1 \text{ event/m}^2.$$

Given that $\widehat{\rho} = 1$ event/m^2 we expect that a 2m × 2m = 4m^2 region would contain 1 event/m^2 × 4m^2 = 4 events. We can estimate the expected number of crimes in some region i with area A more generally with

$$\widehat{N}_i = \widehat{\rho} \times A.$$

Therefore, under complete spatial randomness N_i, the number of crimes in a region i of area A and given point density ρ, is

$$N_i \sim \text{Poisson}(\rho A).$$

This seems appropriate as a Poisson distribution is typically used in establishing the probability of a given number of events occurring in a fixed interval of time or space. These events occur with a constant mean rate and the occurrence of each event is independent of the previous event. This provides

- $E[N_i] = \rho A$,
- $P(X = k|\rho, A) = \frac{(\rho A)^k e^{-\rho A}}{k!}$.

For example, the probability of observing 5 events in an area of 4m^2 where $\widehat{\rho} =$ 1 event/m^2 is

$$P(X = 5|\rho = 1, A = 4) = \frac{[1(4)]^5 \, e^{-1(4)}}{5!} \approx 0.1563.$$

This indicates that there is approximately a 15.63% chance of observing 5 events in an area of 4m^2.

Now that we have established the probability distribution for the number of events observed in a given window, we can begin to look at various methods to determine the spatial arrangement of these events.

Three methods to test for spatial randomness

The quadrat method. A *quadrat* is a subset of the window which frames a unit of area so that the distribution of items in that area can be studied. In this study, the area of DC is divided into rectangular quadrats. The null hypothesis under the model of complete spatial randomness as defined previously, is that the crimes are equally likely to occur anywhere in the window. The estimated density of points in the window, $\widehat{\rho}$, is used to estimate the number of expected crimes, $\widehat{N}_i$, in a particular quadrat.

Consider the hypotheses,

H_o: The data are spatially random

H_a: The data are not spatially random.

Given that $N_i \sim \text{Poisson}(\rho A)$, a goodness of fit test statistic for complete spatial randomness in Q quadrats is of a Chi-squared distribution:

$$\chi^2 = \sum_{n=1}^{Q} \frac{(N_i - \widehat{N}_i)^2}{\widehat{N}_i}, \quad \text{with } (Q - 1) \text{ degrees of freedom.}$$

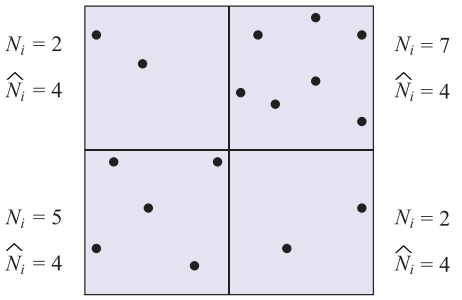


Figure 3. Spatially random events in 2m × 2m quadrats.

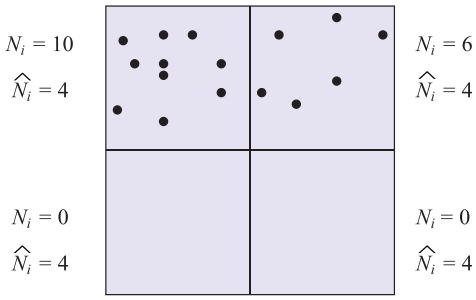


Figure 4. Clustered events in 2m × 2m quadrats.

Figure 3 shows the observed number of crimes N_i as well as the expected number of crimes $\widehat{N}_i$ for each quadrat in our 4m × 4m window. The test statistic can be calculated as follows:

$$\chi^2 = \frac{(2-4)^2}{4} + \frac{(7-4)^2}{4} + \frac{(2-4)^2}{4} + \frac{(5-4)^2}{4} = 4.5.$$

This results in an approximate p-value of 0.2123 providing little evidence that the data are not spatially random.

Now, consider Figure 4 where the 16 events are distributed across the window in a more clustered manner. The test statistic in this case is calculated to be:

$$\chi^2 = \frac{(10-4)^2}{4} + \frac{(6-4)^2}{4} + \frac{(0-4)^2}{4} + \frac{(0-4)^2}{4} = 18.$$

© THE MATHEMATICAL ASSOCIATION OF AMERICA

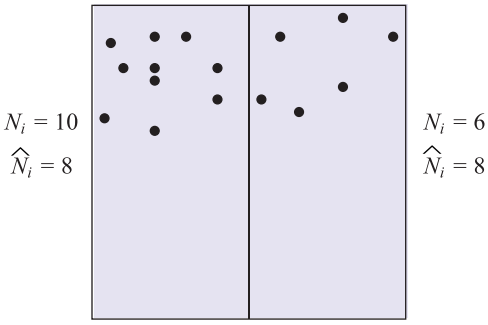


Figure 5. Clustered events partitioned into two 2m × 4m quadrats.

In this instance, the p-value is now approximately 0.0004 which provides strong evidence that the data are not spatially random.

The Quadrat method does have some drawbacks. Its ability to detect complete spatial randomness is heavily dependent on the size and shape of the quadrats used to partition the window.

Consider the distribution of events in Figure 4 partitioned now into two 2m × 4m rectangular quadrats (instead of four quadrats). This is displayed in Figure 5. In this case, the expected number of events, $\widehat{N}_i$, is now 1 event/m^2 × 8m^2 = 8 events per quadrat.

The χ^2 test statistic is now calculated to be:

$$\chi^2 = \frac{(10-8)^2}{8} + \frac{(6-8)^2}{8} = 1.$$

This results in an approximate p-value of 0.3173 indicating little evidence of clustering. This contradicts the previous conclusion when four 2m × 2m quadrats were used for the identical distribution of crimes in Figure 4.

Variance-mean ratio (VMR). The VMR is a measure of clustering that is generally used when the underlying distribution can be assumed to be Poisson or Exponential. Recall that under spatial randomness, the number of events in quadrat i has the distribution $N_i \sim \text{Poisson}(\rho A)$. Then we may observe

$$E[N_i] = Var[N_i] = \rho A \text{ and } VMR = \frac{Var[N_i]}{E[N_i]} = 1.$$

- If E[N_i] = Var[N_i], then the empirical VMR should be close to 1, indicating that the events are spatially random.
- If E[N_i] >> Var[N_i], then the empirical VMR should be close to 0, indicating that the events are distributed uniformly.
- If E[N_i] << Var[N_i], then the empirical VMR should be larger than 1, indicating that the events are clustered.

Consider the distribution of the 16 crimes in a spatially random manner (Figure 6). The mean is 4 events per quadrat with a variance of approximately 4.5 events^2. This produces a VMR = $\frac{4.5}{4}$ = 1.125 events ≈ 1, which supports the conclusion that the distribution of events is spatially random.

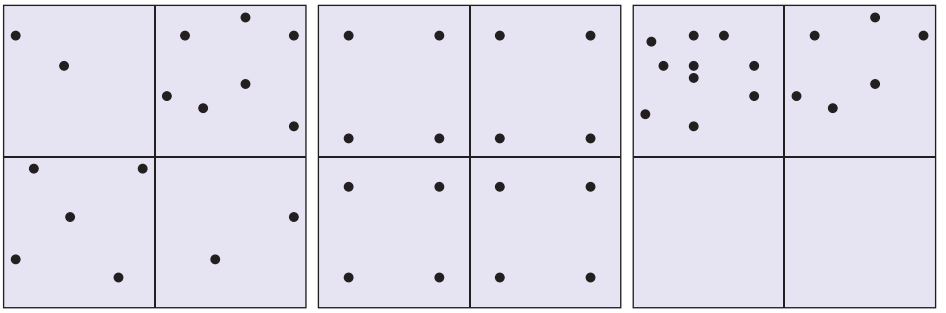


Figure 6. Random events. **Figure 7.** Uniform events. **Figure 8.** Clustered events.

Suppose the 16 crimes were now uniformly distributed across the 4 quadrats as shown in Figure 7. In this case, the mean is clearly 4 events but the variance is now equal to 0 events^2. Hence, the VMR $= 0$ events, indicating that the events are distributed uniformly.

Once again, consider the distribution of crimes as shown in Figure 9 where the crimes observed are clustered. In this scenario, the mean is again 4 events but with an approximate variance of 18 events^2. Therefore, the VMR $= \frac{18}{4} = 4.5$ events $>> 1$, supporting that the events are clustered.

It should be noted that the VMR method does have similar disadvantages as the Quadrat Method. The size and shape of the partitions used in subdividing the window can significantly affect its ability to ascertain complete spatial randomness among events. Therefore, we turn our attention to a third approach for studying spatial distribution.

Ripley's K function. Ripley's K function analyzes the density of events around each individual event at varying distances. The function places a distance band around each event, where $K(r)$ is the average density of points at each distance r, divided by the average density of points in the entire study window. If the density of points is high in a particular band, then this leads to the conclusion that localized clustering is occurring. One benefit of utilizing the $K(r)$ function is that it provides insight at different distances around an event.

In this instance we will construct a circle of radius r, around a point (event), s. The number of events that fall within this circle is counted. This is repeated for all events s in the window and the results are totaled. The distance r, is then incremented by a fixed amount and the above calculation is repeated. The function $K(r)$ can be plotted against the distances, r, as seen in Figure 11. Ripley's K function is formally defined as

$$K(r) = \frac{A}{N^2} \sum_{j \neq i} I\left[s_j \in C(s_i, r)\right]$$

where,

- $N =$ number of events in the window
- $A =$ area of the window
- $s_j =$ event of interest

 © THE MATHEMATICAL ASSOCIATION OF AMERICA

- $C(s_i, r)$ is a circle of radius, r, centered at s_i
- $I = \begin{cases} 1 & s_j \in C(s_i, r) \\ 0 & \text{otherwise.} \end{cases}$

Application of methods to DC crimes

Now that we have discussed three tools to examine the distribution of spatial data, let's return to our original problem. What can we say about the distribution of crimes in our nation's capital? Are these events spatially random? Clustered? Uniformly distributed? Figure 9 shows location of crimes in DC's 3rd district during the year of 2017 and was obtained from mpdc.dc.gov. Different colors represent the mark of crime type. It is clear that the region does not have an overall rectangular form as applied in the previous explanations. For simplicity, we will focus only on a rectangular subset of the 3rd district as shown. Application of the quadrat method, variance to mean ratio, and Ripley's K function are all performed on this subset.

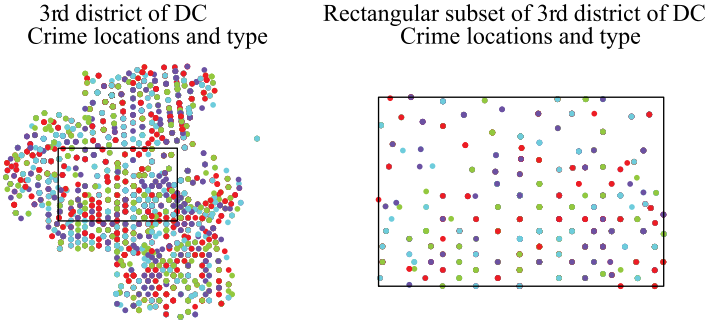


Figure 9. Crimes in DC's 3rd district.

The quadrat method. To apply the quadrat method, we first need to divide the window into equally sized quadrats. A 5×5 division of the window is used and illustrated in Figure 10. Each quadrat contains the observed counts, the expected counts under the assumption of complete spatial randomness, and the Pearson residuals. The Pearson residuals indicate each quadrat's contribution to the test statistic. They are a measure of the discrepancy between the observed data and the expected count under spatial randomness within a quadrat. With a test statistic of $\chi^2 = 876.08$, $df = 24$, and p-value< 0.0001, we conclude that strong evidence exists for clustering of crime locations.

Variance to mean ratio. By using the same 5×5 quadrats, we can take advantage of the observed counts in each quadrat to calculate the mean and variance of the number of events among the quadrats. As a result, we have

$$\bar{x} = 58.26, \qquad s^2 = 2108.44 \qquad VMR = 36.2.$$

$VMR >> 1$, indicating that the variability in number of events among the quadrats is much greater than the mean. That is, some quadrats have high crime counts while

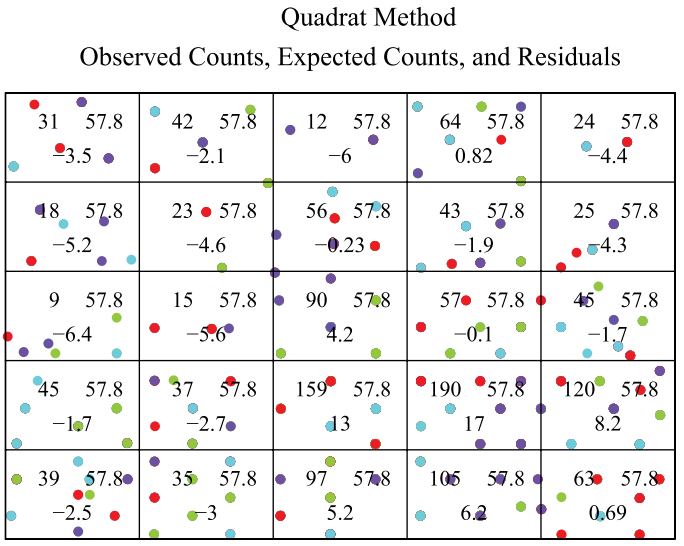


Figure 10. Results of quadrat test.

others have low crime counts leading to the conclusion of spatial clustering in crime. This agrees with our conclusion from the quadrat method.

Ripley's K function. The results from applying Ripley's K function are shown in Figure 11. The dotted line represents the expected value of $K(r)$ under the assumption of spatial randomness. The solid line represents the observed value of $K(r)$ in the data. Because the density of observed events within a circle of any size radius is greater than the expected under spatial randomness, we are confident in the clustering of crimes.

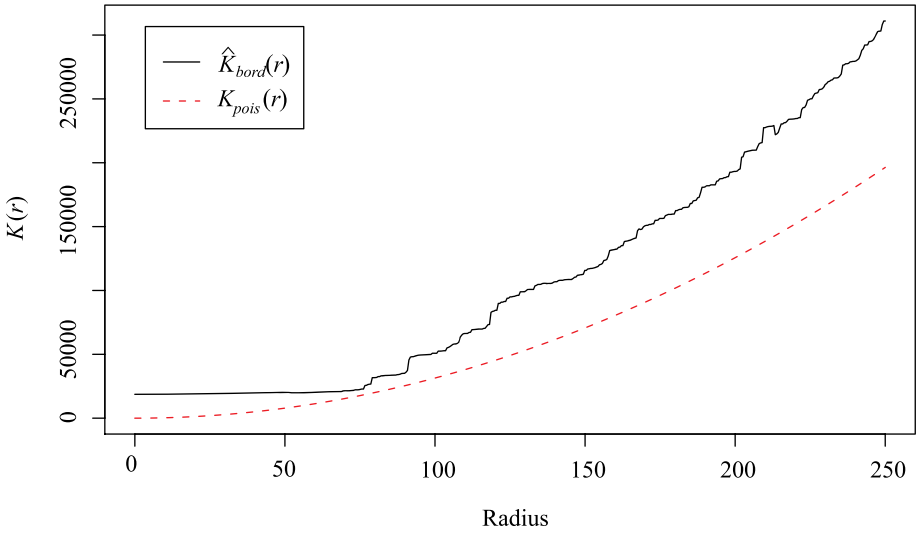


Figure 11. Results of Ripley's K function.

© THE MATHEMATICAL ASSOCIATION OF AMERICA

Discussion

The ideas and vocabulary of spatial statistics have been introduced. Three methods for testing spatial randomness were described and applied to a subset of crime data from our nation's capital. We invite the reader to explore other settings that might provide interesting spatial contexts. Some recent topics of interest include the locations of disease spread, voter turnout, and the effects of global warming on species migration. The `spatstat` and `rspatial` packages of the open-source software `R` can be used to implement these techniques with your own questions of interest.

Extensions and deeper exploration into the methods presented is also encouraged. For example, Ripley's K function can be used in a bivariate setting. Perhaps we would like to include the mark of crime type and explore the spatial distribution of traffic violations as compared to all other types of crimes. The dependency of the quadrat method and VMR on quadrat selection was noted. In our example, the window was arbitrarily divided into a 5×5 grid in order to define the quadrats. What would be the effects of increasing or decreasing the number of quadrats? Is there a way to determine an optimal number of quadrats? How can these methods be adapted for application to non-rectangular regions and in what other ways can the use of marks be incorporated?

The spatial distribution of data is an area of study that is not often seen by undergraduate students. However, we have shown that it can be accessible and applied to a wide array of important topics. We invite students and mentors alike to investigate this rich topic.

Summary. The distribution of spatial data is fascinating and holds applications across a variety of disciplines. We introduce the ideas and vocabulary related to spatial data. The quadrat method, variance-to-mean ratio, and Ripley's K function are described for assessing the distribution in spatial data. The techniques are applied to open data on crimes in Washington DC. The reader is left with ideas for advanced exploration.

References

[1] Dixon, P. M. (2002). Ripley's K function. *Encyclopedia of Environmetrics*, Vol. 3. Chichester: Wiley, pp. 1796–1803.

[2] Baddeley, A., Rubak, E., Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton, FL: Chapman & Hall/CRC.