Proceedings Paper:
Adepeju, M (2017) Testing the adequacy of a single-value Monte Carlo simulation for space-time interaction of crime. In: Gervasi, O, Murgante, B, Misra, S, Borruso, G, Torre, CM, Rocha, AMAC, Taniar, D, Apduhan, BO, Stankova, E and Cuzzocrea, A, (eds.) Lecture Notes in Computer Science. 17th International Conference on Computer Science and Its Applications (ICCSA 2017), 03-06 Jul 2017, Trieste, Italy. Springer Nature, pp. 779-786. ISBN: 978-3-319-62407-5. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-319-62407-5_60

Testing the adequacy of a single-value Monte Carlo simulation for space-time interaction analysis of crime

Monsuru Adepeju (Orcid ID: 0000-0002-9006-4934)

School of Geography, University of Leeds, LS21 1HB, United Kingdom

M.O.Adepeju@leeds.ac.uk

Abstract. The goal of this study is to determine the number of iterations (r) required in a Monte Carlo based space-time interaction analysis of crime data sets, in order to test the adequacy of using a single value of 999 iterations. A case study of burglary crime data sets is presented in which Knox test is used for the analysis of space-time interactions. The outcomes of this analysis demonstrate that the use of a single value, such as 999, does not always represent the most appropriate number of iterations especially when multiple ST neighbourhood sizes are involved. This analysis opens further research opportunities into determining the best strategy to defining the expected distribution in a space-time interaction analysis of crime.

Keywords: ST neighbourhoods, Monte Carlo simulation, crime, Knox test

1 Introduction

The use of a Monte Carlo (MC) simulation in space-time interaction analysis using the Knox test (Knox, 1964) usually involves 999 iterations in order to generate the expected distribution under the assumption of no space-time interactions. Despite the potentials of varying reliabilities relating to the underlying normal distribution for different pairs of spatial and temporal thresholds, the same number of iterations is usually employed in crime applications (Johnson et al., 2007). This study therefore aims to test the adequacy of the generally adopted 999 iterations at a chosen reliability level for different spatial and temporal thresholds.

One way to test the reliability of a MC simulation for a normal distribution is to specify a desired percentage error for the computed mean value of the

random variables (in this case, the Knox statistic), while the iterations are continuously repeated (Driel and Shin, 2004). Thus, the number of iterations needed (r) to attain the specified error can be determined by monitoring the convergence of r in relation to the actual number of iterations being performed. To the best of the author's knowledge, this type of analysis has not been carried out for space-time interaction analysis in relation to a crime data set. Therefore, the major goal of this study is to address this research gap by determining the number of iterations (r) required for different spatio-temporal (ST) neighbourhoods of crime, and subsequently, examine the adequacy of using a single value of 999 iterations in the context of the generated results.

2 Space-time interaction analysis with the Knox test

The Knox test is the most commonly used technique for the analysis of the spatio-temporal interactions of crime data sets (Johnson et al., 2007). The Knox test measures whether there are disproportionate instances of observed pairs of events within a defined spatio-temporal neighbourhood than would be expected if the events had occurred randomly. Therefore, the hypothetical random occurrences represent the expected distribution, which is generally modelled as a normal distribution.

Mathematically, the Knox statistic is a product of two 'closeness matrices'. The first matrix (X_{ij}) describes the closeness of all pairs of events in space, while the second matrix (Y_{ij}) describes the closeness of all the pairs of events in time. The closeness is defined by specifying a spatial neighbourhood (δ) and a temporal neighbourhood (τ), within which event j is considered close to event i in space and time dimensions, respectively. Technically, each neighbourhood is the intersection of two distance thresholds; $[\delta_1, \delta_2]$ and $[\tau_1, \tau_2]$, where $\delta_2 > \delta_1$, and $\tau_2 > \tau_1$ (see Fig. 1 for an illustration).

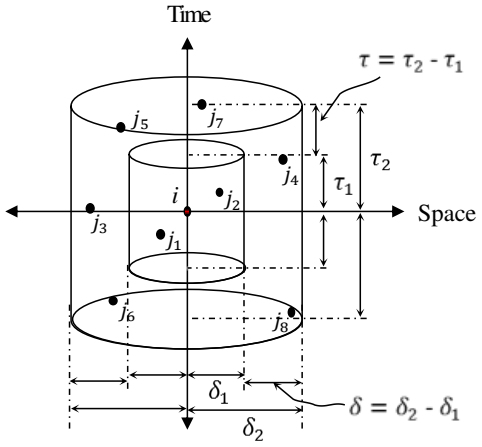

Fig. 1. An illustration of the spatio-temporal neighbourhood around a point i (in a Knox test). The event i is the reference, while events $j_1, j_2, \ldots\ldots, j_8$ are examined for 'closeness' to i. Events $j_3, j_4, \ldots\ldots j_8$ fall within the spatial neighbourhood $\delta = \delta_2 - \delta_1$ and temporal neighbourhood $\tau = \tau_2 - \tau_1$, and are therefore considered close to i in space and time (Diagram from: Adepeju, 2017a).

For each pair of spatial and temporal neighbourhood, the closeness is evaluated for every point (i) across the entire study area and finally added together in order to derive the Knox statistic as follows:

$$n_{\delta,\tau} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n-1} X_{ij}Y_{ij} \tag{1}$$

$$X_{ij} = \begin{cases} 1, & \text{if event } j \text{ is within } \delta \text{ of } i \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1, & \text{if event } j \text{ is within } \tau \text{ of } i \\ 0, & \text{otherwise} \end{cases}$$

The $n_{\delta,\tau}$ is referred to as the *observed,* with which the expected statistics $e_{\delta,\tau}$ are compared to estimate the critical value (p) through using the formula:

$$p = \frac{1 + \sum_{v=1}^{r} I(n_{\delta,\tau} \geq e_{\delta,\tau})}{r + 1} \tag{2}$$

Where r is the number of iterations generated, $e_{\delta,\tau}$ is the equivalent list of expected statistics, and $I(.)$ is the indication function.

An expected statistic is calculated via a new replica of the original data set, which is generated by randomising the time attribute of the events, while the spatial locations are kept constant. This process is the MC simulation (also called the iteration process). If each $e_{\delta,\tau}$ is considered a random variable, a plot of all $e_{\delta,\tau}$'s should assume a normal distribution defined by a mean value, and standard errors which can be evaluated at varying confidence levels. Theoretically, the mean value of the 'obtained' normal distribution is close to the mean value of the hypothetical normal distribution within an error bound. Hence, it is assumed that this hypothetical normal distribution can only be attained if the iteration is run infinitely. However, both distributions can be compared in order to examine the reliability of the obtained normal distribution.

3 Determining the number of iterations (r) needed to attain a specified error bound.

It is possible to determine the minimum number of iterations (r) needed to provide a desired degree of reliability for the expected distribution ($e_{\delta,\tau}$), as described in section 2. This process was used in Driel and Shin (2004) so as to estimate the r required in a precision analysis of military weapon effectiveness. Firstly, it is argued that r needs to be large enough to obtain sufficient granularity in the cumulative density function of the $e_{\delta,\tau}$. For example, if $r = 30$, it would not be possible to obtain a 1% rank. This is because an analyst needs at least 100 iterations. Using the Driels and Shin (2004) approach therefore requires a continuous generation of large number of replicas in which a plot of r against the number of replicas, at a specified maximum acceptable percentage error of the mean value can be used to monitor the convergence of r.

Table 1 shows the values of the confidence coefficient z_c for different confidence levels of a normally distributed random variable. The ranges for a given z_c are usually expressed in the form of an upper (U) and lower bound (L), whereby:

$$U = \mu_x + z_c \sigma_x \tag{3}$$

$$L = \mu_x - z_c \sigma_x \tag{4}$$

Where μ_x the population is mean and σ_x is the population standard deviation of the random variables x.

Table 1. Values of z_c for different confidence levels for a normally distributed random variable.

Confidence level (C.L) %	99.75	99	98	96	95.5	95	90	80	68	50
z_c	3	2.58	2.33	2.05	2	1.96	1.65	1.28	1	0.6745

Given a confidence level of 95% for example, the confidence interval of the mean is therefore as follows:

$$(L, U)_{0.95} = \mu_x \pm 1.96\sigma_x \tag{5}$$

This is stated as: we are 95% confident that the true mean is within (L, U) of a sample of the mean of x. The general form of equation 5 can then be written as:

$$(L, U)_{C.L} = \mu_x \pm z_c \sigma_x \tag{6}$$

If the simulation is run for a finite number of iterations (r), the sample mean $\bar{x}$ and the standard error S_x are thus estimates of the population statistics.

$$(L, U) = \bar{x} \pm z_c (S_x/\sqrt{r}) \tag{7}$$

By considering the confidence interval as representing twice this maximum error, we have:

$$error_{max} = \bar{x} \pm z_c (S_x/\sqrt{r}) \tag{8}$$

Hence, the percentage error of the mean becomes:

$$E = \frac{100 \times z_c S_x}{\bar{x}\sqrt{r}} \tag{9}$$

By solving the equation 9 for r, we have:

$$r = \left[\frac{100 \times z_c S_x}{\bar{x}E}\right]^2 \tag{10}$$

r in equation 5.10 is the number of iterations needed to be carried out for a given error bound E and a confidence interval whose coefficient is z_c . Let's imagine we want a confidence level of 95% at an error percentage of 1% of the mean, z_c and E will be 1.96 and 1, respectively, $\bar{x}$ and S_x can be estimated from the generated random samples. The value of r can thus be calculated continuously as the replicas increase. Furthermore, the required value of r can be taken as the point where r stabilises or converges in a plot of r against number of replicas.

4 Case study: data sets and experimental parameters

This test is demonstrated using the burglary crime data set of the San Francisco area of the United States for the year 2015. The two most prominent sub-categories of burglary are used. They are, 'burglary-in-residence' (2,990 records) and 'burglary-of-shops' (1,166 records). The choice of these data sets is based on a previous finding that demonstrates that sub-categories of burglary crimes possess distinct spatio-temporal interactions (Adepeju, 2017b). Besides, the 3-D visual exploration of both data sets illustrates that 'burglary-in-residence' is denser than 'burglary-of-shops' spatio-temporally; indicating a potential for two distinct spatial and temporal interactions (Fig. 2).

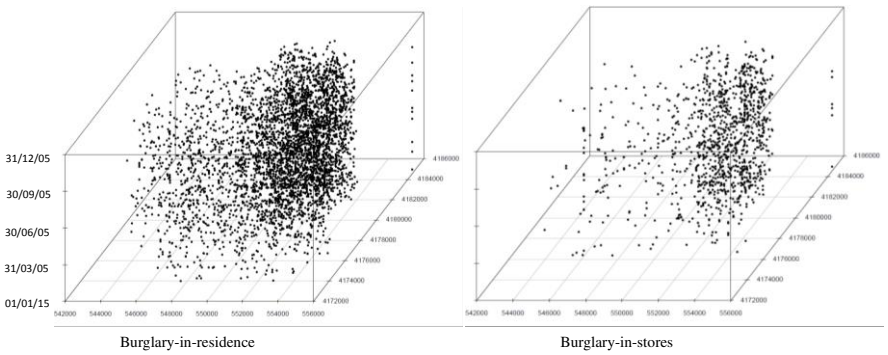

Fig. 2. A 3-D scatterplot of the case study data sets

In order to ensure a robust analysis, three common levels (sizes) of spatial and temporal neighbourhoods in both dimensions are considered. These include small, medium and large levels. Thus, the following lists are can be defined:

- Spatial neighbourhoods, δ = [0-200m], [301-500m], [701-900m]
- Temporal neighbourhoods, τ = [0-2days], [7-21days], [30-60days]

The three levels are as demarcated with the brackets. Based on the two lists, corresponding spatio-temporal neighbourhood are formed by pairing each spatial neighbourhood with a temporal neighbourhood. Hence, a total of nine ST neighbourhoods are formed. This covers a range of levels commonly used in crime analysis.

In this study, the percentage error of 1% is chosen for the mean value, and a confidence interval of 95%. Both the mean value ($\bar{x}$) and the standard deviation (S_x) are calculated after each iteration step. The values, $\bar{x}$ and S_x are then substituted into equation 10, where E = 1 and z_c= 1.96.

5 Results and discussions

Fig. 3 shows the plot of the number of iterations needed (r) for the specified error parameters against the actual number of iterations (replicas) carried out. These plots enable the convergence of r to be monitored.

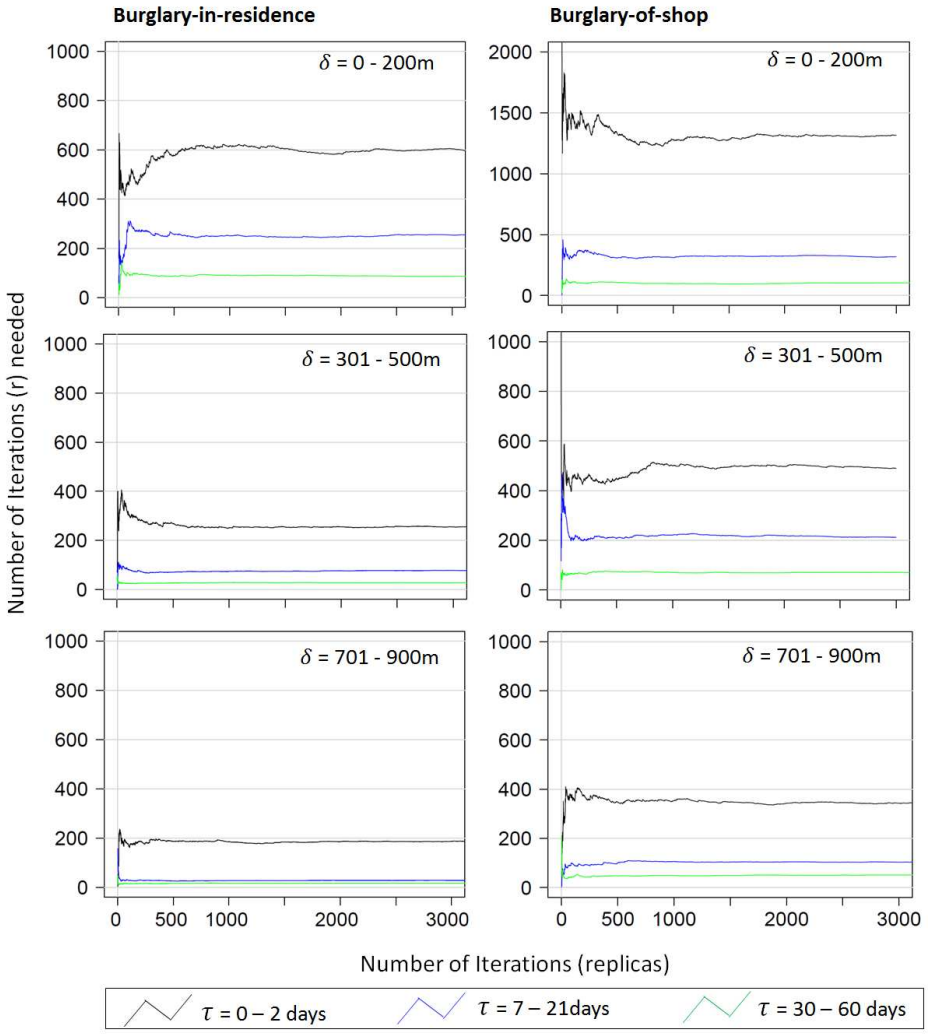

Fig. 3. The number of iterations needed (r) against the number of iterations (replicas).

Each plot shows the results generated for the selected temporal thresholds at each spatial threshold. For example, the top-most left and the top-most right plots are the results of the three temporal thresholds at the spatial threshold of 200m, for the 'burglary-in-residence' and 'burglary-in-shops' crimes, respectively. The general pattern across all the plots is that the number of iterations needed (r) reduces as the sizes of the spatial and temporal thresholds increase. That is, in each crime sub-category, the highest value of

r is obtained at the smallest spatiotemporal neighbourhood (i.e. intersection of δ = 0-200m and τ = 0-2 days), while the lowest r is obtained at the largest ST neighbourhood (i.e. δ = 701-900m and τ = 30-60days. Technically, at large ST neighbourhoods, the mean values become relatively large in comparison to the standard error, thereby allowing r values to converge faster. Whereas, there is higher variabilities at smaller ST neighbourhoods because of the relatively small values of mean in comparison the standard error, thereby requiring larger replicas to converge.

The results of 'burglary-in-residence' shows that the number of iterations (r) needed at the τ = 2days across all spatial thresholds is multiple times larger than all of the other temporal thresholds. The largest value of $r = 590$ is obtained at the spatial threshold of 0-200m; a value which stabilises after only around 500 iterations. At the other spatial thresholds, r stabilises much faster; converging even before 250 iterations. In this case, the values of r are generally between 20 and 300. These are relatively small numbers compared to the commonly used 999 iterations. A similar result is obtained for 'burglary-of-shops', except with a slightly higher value of r for each corresponding ST neighbourhoods. Thus, this indicates that the 'burglary-of-shops' crime possesses a relatively higher variability in the ST distribution compared to the 'burglary-in-residence' crime. This is apparent in the 3D scatterplots (Fig. 2), in which 'burglary-of-shops' appears sparser compared to the findings for the 'burglary-in-residence' crime. The result of 'burglary-of-shops' at the smallest spatial and temporal neighbourhoods also shows that r could exceed 999. Additionally, this demonstrates that if the percentage error E value is reduced or the confidence level increased, the value of r can be much greater than 999. In summary, it is observed that the ST distribution of a data set, as well as the ST neighbourhood sizes used, influence the reliability of the expected distribution in a MC based space-time interaction analysis of crime data sets. Thus, the use of a single value, such as 999, may not represent the most appropriate number of iterations in the case of multiple ST neighbourhood sizes.

6 Conclusion and recommendations

This study has examined the number of iterations required for different sizes of spatio-temporal neighbourhoods of crime data sets. The aim was to test how reliable the practice of using a single value, such as 999 in a MC simulation process. The result obtained shows that given some specified errors, different spatio-temporal neighbourhoods require different numbers of iterations in order to generate reasonable expected (normal) distribution. This is in contrary to the general practice in which a uniform (single) value, particularly 999, is often used. Hence, it is argued that this is generally a practice used for convenience and to ensure the uniformity of precision in the reported critical values. In the future, the author would like to investigate how the findings of this study could be employed to achieve a more reliable result using a Knox test.

It is therefore recommended that this type of analysis should first be carried out in any spatial and temporal point pattern analysis. It will help to establish the reliability of the MC simulation process.

References

1. Adepeju, M.: Investigating the Repeat and Near-Repeat Patterns in Sub-categories of burglary crime. An abstract submitted to the 2017 International Conference on GeoComputation (2017b)

2. Adepeju, M.: Modelling of Sparse Spatio-Temporal Point Process (STPP) – an application in Predictive Policing. A PhD thesis submitted to the University College London (2017a).

3. Driels, M.R. and Shin, Y.S.: Determining the number of iterations for Monte Carlo simulations of weapon effectiveness (2004).

4. Johnson, S.D., Bernasco, W., Bowers, K.J., Elffers, H., Ratcliffe, J., Rengert, G., Townsley, M.: Space-time patterns of risk: a cross national assessment of residential burglary victimization. J Quant Criminol, 23, 201–219 (2007)

5. Knox, G.: Epidemiology of Childhood Leukaemia in Northumberland and Durham. Brit. J. of Prev. and Social Med, 18, 17-24 (1964).