2-2018

New Approaches to Model Simulated Spatio-Temporal Moran's Index

Nhan Bu

Jennifer Lorio

Norou Diawara

Kumar Das

Lance Waller

Journal of Probability and Statistical Science
16(1), 11-24, Feb. 2018

New Approaches to Model Simulated Spatio-Temporal Moran's Index

Nhan Bu Jennifer Lorio Norou Diawara — *Old Dominion University*
Kumer Das — *Lamar University*
Lance Waller — *Emory University*

ABSTRACT The Moran's index is a statistic that measures spatial autocorrelation; it quantifies the degree of dispersion (or clustering) of objects in space. However, when investigating data over a general area, a single global Moran statistic may not give a sufficient summary of the spread, behavior, features or latent surfaces shared by neighboring areas; rather, by partitioning the area and taking the Moran statistic of each divided subareas, we can discover patterns of the local neighbors not otherwise apparent. In this paper, we present a simulation experiment where the local Moran values are computed and a time variable is added to a spatial Poisson point process. Changes in the Moran statistics over the neighboring areas are investigated and ideas on how to perform the analysis are proposed.

Keywords Extreme value distribution; Moran's index; Simulated processes; Spatio-temporal model.

1. Introduction

In the era of big data, we rely on modeling correlation between features of data to make inference. One such correlation in spatial data is the Moran's Index. As first described by Moran [16], when given a set of variates (x, y) (defined on some two-dimensional discrete area) we may want to investigate whether there is any evidence that spatial autocorrelation is present overall or in neighboring clusters based on selected features. Applications of such spatial statistics can be found in many areas, for example, in agricultural research, specific plots of land may influence in several aspects the production of nearby plots. Defining random variables with spatial components as described in Vaillant *et al*. [21] can further advance the understanding of

Received July 2017, revised October 2017, in final form December 2017.
Nhan Bui, Jennifer Lorio, and Norou Diawara (corresponding author; email: ndiawara@odu.edu) are affiliated to the Department of Mathematics and Statistics at the Old Dominion University, Norfolk, VA 23592, USA. Kumer Das is affiliated to the Department of Mathematics at Lamar University, Beaumont, TX 77710, USA. Lance Waller is affiliated to the Department of Biostatistics and Bioinformatics at Emory University, Atlanta, GA 30322,.

 JPSS: ISSN 1726-3328

the correlation. Because local trends in data may not be shared globally, many authors such as Baddeley [2] and Anselin [1] have partitioned the global Moran value into a sum of local indices of spatial association (LISA). However, the impact of time was not incorporated. Spatio-temporal autocorrelation was first introduced by Cliff & Ord [5] and the concept has been explored by many others over the years (see for example, [14, 12, 23]). We develop a computational method to understand the local trends in the spatio-temporal environment. Logically, we can expect observations that are close to likely be more similar in characteristic than those that are far apart. Hence the hypothesis test of the spatial arrangement of these feature variates are of interest, i.e., we ask "Are the spatio-temporal arrangements random or distinctly clustered?"

To answer that question, we refine the use of the Moran's Index by taking a general area, subdivide it into smaller regions, and measure the spatial dependencies of points generated under a Poisson point process over a discrete time sequence. This is a novel approach for which the analysis of the temporal evolution of spatial patterns in the spread could be summarized.

The rest of the paper is organized as follows. Section 2 provides an in depth overview of the Moran's Index. Section 3 presents the time sequence Poisson process algorithm and the simulation study completed on a partitioned area domain. Section 4 provides an overview model fit of the spatio-temporal Moran's statistic under the Generalized Pareto Distribution (extreme value modeling). And lastly, we end with a conclusion.

2. Moran's Index Spatial Autocorrelation

2.1 Global Spatial Autocorrelation

In practice, the Moran's Index, I, is typically used to evaluate the clusters in the spatial arrangement of a given variable. In other words, it is a measurement method for quantifying the degree of clustering or dispersion. The global Moran Index based on a sample of n observations is defined as:

$$I = \frac{n}{\sum_{i \neq j} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j}{\sum_{i=1}^{n} x_i^2},$$

where x_i and x_j are the actual values of spatial characteristic or feature indications i and j, and w_{ij} is the weight between features i and j. Recent advances in Geographic Information System (GIS) tools offer better measures of location and can be used to make integral reference to effect or environment. There are many choices on how the weights are defined. For example, in taxology (scientific identification, naming, and classification of living things), $w_{ij} = 1$ if species i and j belong to the same group as presented in [17]. Another choice of w_{ij} had been proposed by Gittleman and Kot [8] in their phylogenetic inertia effects study where they describe the evolutionary biology for phylogenetic inertia under the assumption that species (trees) cannot be treated as independent points for statistical analysis, but rather they share

characteristics at distinct distances and time. They defined the weights as:

$$w_{ij} = \frac{1}{(d_{ij})^{\alpha}},$$

where d_{ij} are the distances measured on trees and α is the correction factor. In general, the range of I is from -1 to $+1$. Values closer to $+1$ indicate clustering and values closer to -1 indicate dispersion. Waller and Gotway [22] noted that the range $[-1, 1]$ can actually be misleading as the choice of the weights can affect the range.

The work by Deng *et al.* [7] further illustrates the usefulness of the Moran's Index where time is added to the statistic. In their paper, six main indicators were adopted to describe the physical characteristics of river systems in the Taihu Basin region of China. The spatial-temporal evolution of the distribution pattern of this river system was analyzed with data from the 1960s to 2000s. The feature or characteristic indicators of these rivers include river density, river frequency, water surface ratio, river development coefficient, river systems complexity, river systems stability, river bifurcation ratio, river length ratio, river sinuosity, main river area length ratio, and box dimension. The Moran's Index was then calculated to measure the spatio-temporal evolution of the river systems with the conclusion that global distribution patterns of river length and box dimension were statistically significant, and hence spatially clustered. Moran values were computed and averaged by decades from the 1960s, 1980s, and 2000s.

While the Moran's Index captures information on spatial autocorrelation of the location and covariate (features), more generalized modifications to the statistic have been utilized in case studies by Vaillant *et al.* [21], where analysis was conducted on the spread of sugar cane yellow leaf virus. In particular, they modeled the propagation of infection with a focus on the spatial spread of disease over time. For a time interval partitioned into periods and each pair of observation dates (t_{i-1}, t_i), $i = 1, \ldots, I-1$, for some fixed positive integer I, they defined the Moran's index based on a nearest neighbor scheme:

$$M_i = \sum_{(x,y)\in D} w_{x,y} \mathbf{1}_{[0,t_{i-1}]}(T_x) \mathbf{1}_{(t_{i-1},t_i]} T_y,$$

where D denotes the discrete set of plant locations, T_x the date (time variable) of virus detection for plant x, and $\mathbf{1}_{[0,t_{i-1}]}(T_x)$ is an indicator of whether time T_x falls in the interval $[0, t_{i-1}]$ and $w_{x,y}$ denotes the distance between points in the same interval. Similarly, $\mathbf{1}_{[t_{i-1},t]}(T_y)$ is an indicator of whether time T_y falls in the interval $[t_{i-1}, t]$.

2.2 Localized Spatio-Temporal Autocorrelation

Since the relationship between measurable features needs to be validated over time and space, subdividing the area into neighboring areas will allow us to explore pathways where statistical measures of correlations can be effectively interpreted, in a computationally tractable algorithm. A Poisson point process in the domain area is a great starting point to build a model

for Moran's values. In this paper, we utilize a similar definition of the Moran's Index and while the values are not between -1 and 1, it is essentially the same in spirit as in [6].

Consider a space $\mathbf{S}$ and let $\mathbf{B}$ be a (measurable) subset of $\mathbf{S}$, i.e. $\mathbf{B} \subset \mathbf{S}$. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a sample of random points in $\mathbf{B}$. Set $N(\mathbf{B})$ as the total count of the sample points in $\mathbf{B}$ where occurrences of events have been noticed. Then $N(\mathbf{B})$ is a measure and is called a point process. That is,

$$N(\mathbf{B}) = \sum_x \mathbf{1}_{\mathbf{B}}(x), \text{ where } \mathbf{1}_{\mathbf{B}}(x) = \begin{cases} 1, & \text{if } x \in \mathbf{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Now suppose random samples of points $x_1,\ x_2, \cdots, x_n$ are uniformly distributed on a bounded region $\mathbf{B}$. Then, the density is simply

$$f(x) = \begin{cases} \frac{1}{|B|}, & \text{if } x \in \mathbf{B}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $|B|$ is the area (Lebesgue measure) of the bounded region $\mathbf{B}$ and is referred to as the intensity $\lambda(\mathbf{B})$. That is, $E(N(\mathbf{B})) = \lambda(\mathbf{B})$, as in [11]. Intuitively, if we generate an infinite number of points based on the density, we would expect the spread to converge to the whole area of $\mathbf{B}$ itself over time. We will verify that expectation. Instead of a global measure of spread, suppose we partition $\mathbf{B}$ into subareas $\mathbf{B}_j$, $\mathbf{B}_j \subseteq \mathbf{B}$, $j = 1, 2, \ldots, m$ for some fixed $m \in \mathbb{N}$. Then the probability of x being in the subset $\mathbf{B}_j$ is defined as

$$P(x \in \mathbf{B_j}) = \int_{\mathbf{B_j}} f(x)dx = \frac{\lambda(\mathbf{B_j} \cap \mathbf{B})}{\lambda(\mathbf{B})}.$$

The total count of x's in $\mathbf{B_j}$ is denoted $N(\mathbf{B_j})$ and it follows that

$$N(\mathbf{B_j}) \sim Binomial\left(n,\, p = \frac{\lambda(\mathbf{B_j} \cap \mathbf{B})}{\lambda(\mathbf{B})}\right).$$

Note that since $\mathbf{B_j}$ are partitions of $\mathbf{B}$ for each $j = 1, \ldots, m$, then it follows that the $N(\mathbf{B_j})$'s are mutually independent. The number of points relative to the intensity may vary: "rare" in some locations or "dense" in others, but it follows the Poisson point process. A special case of a point process in $\mathbf{B}$ with independent events, with associated intensity $\lambda > 0$ the process is defined as:

1. $N(\mathbf{B})$ follows a Poisson distribution with mean $\lambda|B|$.
2. $P(N(\mathbf{B}) = n) = e^{-\lambda|B|}\lambda|B|^n/(n!)$, $\forall n \in \mathbb{N}$.
3. For $\mathbf{B_j} \in \mathbb{R}^2$, $N(\mathbf{B_j})$'s are mutually independent.

The Poisson point process assumes that the events are equally likely to occur anywhere in $\mathbf{B}$ and the events do not interact with each other neither avoiding spread. We will use such nonhomogeneous time sequence Poisson process to generate points within the area domain of interest. With this new approach of describing spatio-temporal spread, we propose to measure the Moran's statistics under simulation experiment.

3. Simulation Study

With the novelty in the approach, we propose a simulation under a Poisson point process adding time with subarea clusters. For the simulation, the area will be a square of 4×4 dimension, partitioned into unit squares. We include time at indicators $i = 0, 1, 2, \ldots, I$, and add that component to the domain subareas D_{ij}, $j = 1, 2, \ldots, m$, where m is the total number of subareas. We take on a similar approach as Vaillant *et al.* [21] where we treat our observed field as a spatial grid with regular spacing between rows. That is, we have grids D_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, m_i$ where m_i is the number of points generated in the time interval $[i-1, i]$. As described in [13], we will define the Moran's statistic in the grid D_{ij} as:

$$M_i^j = \sum_{u,u' \in D_{ij}} w_{u,u'} \mathbf{1}_{(t_{i-1},t_i]} (T_u, T_{u'}) ,$$

where

1. T_u is the time at which location u is observed (e.g. becomes infected).
2. $w_{u,u'}$ represents a spatial weight between any two distinct event locations which could be a function, e.g,
   (a) the inverse distance between the two points.
   (b) the inverse distance squared between the two points.
   (c) an estimate of the autocorrelation/semivariance statistic.

We will use (a) in the simulation. Many options for implementing the process are available and we use R with the `spatstat` package. We begin with an observed (4×4 units) area that is sub divided into 16 smaller areas (i.e. a grid). Next, within these subareas, we create disks with radius of half unit. These disks will allow us to generate sample points within the subareas as well as prevent any overlap of points between the subareas. We choose the Poisson point process to randomly generate points in the 16 subareas. Lastly, our Moran statistic is calculated as the sum of the inverse distances between each point that are generated within each of the disks over time.

The steps are as follows: in R, we (i) introduce a perturbation (using a Poisson point process) at each local subarea based on some initial location with constant rate λ at an initial time, (ii) find the Moran statistic within each subarea, (iii) generate new points with different Poisson process rates defined for each time period as $\lambda_i = i\lambda, i = 1 \ldots, I$ where I is the number of time intervals (iv) calculate the Moran statistic. We continue steps (iii) through (iv) until all time periods are covered.

Algorithm: Iterative local Moran statistics

Procedure:

(i) Define the first time and subareas and include their centers.

(ii) Iterate local points from a Poisson process in that subarea.

(iii) Compute the Moran statistic within that time and subarea combination.

Repeat the generation of the Poisson points at the next time period within each subarea. Compute the Moran statistic within each subarea.

Stop when all times and subareas are reached.

End

Figure 1 (a) shows our observed area with 16 subareas and associated disks. To keep it simple, the disks do not overlap and the point process is subsampled and confined to these disks. Another benefit of this model is the tractability of data handling from the output. Figure 1 (b) below shows two points generated within a certain disk. This point generation process is repeated across all 16 disks.

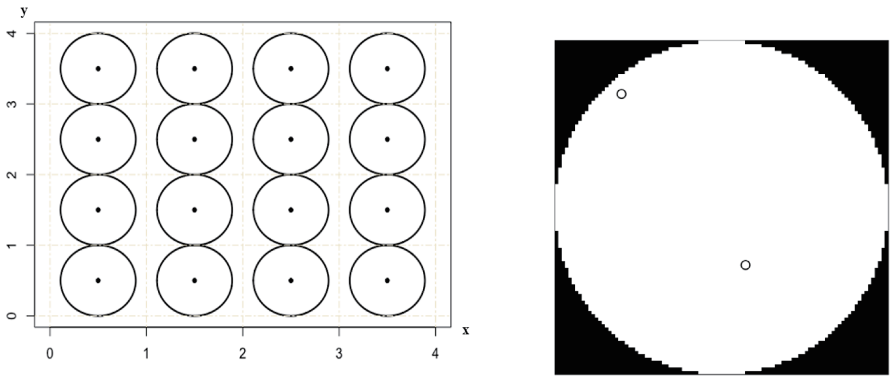

Figure 1 Area plot and (b) points generated in one disk.

The Moran values generated within each disk across the 6 time periods are represented in Table 1 for $\lambda = 2$. The table shows that there is no extra point generated in disks 3 and 14 at the first time period and this is denoted with "NA". The algorithm only considers two or more points when computing the Moran statistic, due to the way we defined our weights (inverse distance).

The table also displays the global Moran values at each time period in the last line. The values are quite large relative to the measures of the subareas. In addition, for time period 1, disks 3 and 14 did not generate any points; the global Moran does not sufficiently portray a good description of the correlation and leads to exaggerated large values. The local Moran's statistics based on subareas provide a better description of the data according to the number of points generated within the disk.

Another benefit of the proposed approach is the tractability of data handling from the output. The spatial distribution of the generated points over time is displayed in Figure 2. This is not a surprising result since time is a function of the intensity $\lambda|B|$ where B is our observed area. In the simulation, we kept intensity fixed for each time period and allowed for time interaction, i.e. letting it vary across (sequential) time points. Thus, the density plot shows that for

sufficiently large time intervals, the amount of points generated will start to spread in the entire area as expected in the description of the model in the previous section.

Table 1 Moran values for $\lambda = 2$

	Time					
disk	1	2	3	4	5	6
1	3.53	7.75	8.90	114.25	116.28	13.92
2	2.29	2.66	50.59	117.16	266.95	43.08
3	NA	5.80	2.19	278.04	65.05	60.36
4	23.04	7.70	14.01	212.57	24.99	63.85
5	19.26	NA	29.60	55.70	245.33	77.13
6	6.28	10.46	192.15	93.76	192.52	54.88
7	2.22	0.00	47.04	24.61	25.14	130.91
8	11.95	20.35	43.11	32.57	185.78	129.88
9	8.41	46.86	46.84	47.26	67.79	83.96
10	11.29	16.86	46.44	91.11	170.05	112.64
11	38.81	7.80	68.91	256.41	125.90	207.29
12	3.30	91.13	25.43	62.48	253.22	237.19
13	29.81	30.73	21.67	71.18	44.23	118.23
14	NA	17.25	26.71	98.46	107.99	173.16
15	2.19	211.91	79.45	44.79	415.01	251.05
16	12.63	20.09	72.53	34.14	61.49	229.93
Global Moran	148.61	450.36	1017.681	1086.424	1887.994	3736.684

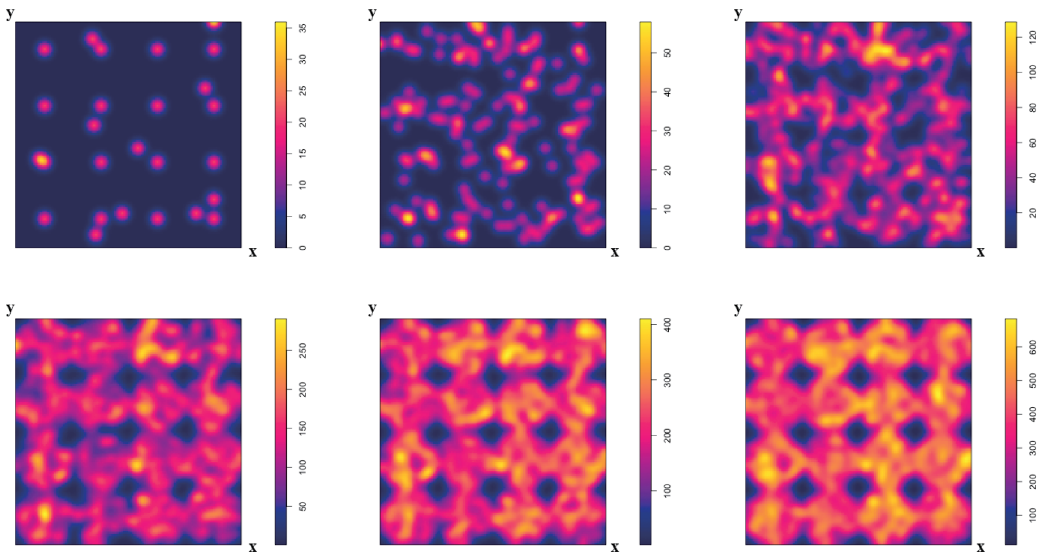

Figure 2 Density over time

4. Model Analysis of Moran's Values

From the above, the local Moran's values provide a better summary of the spread after partitioning the spatial area into subareas and looking at behaviors associated with spread. We

further explore the Moran's values, as comparisons of the Moran's statistics will offer more insight about the nature of the autocorrelation. One method used is the mixed linear model. As a generalization of the standard linear model, this method allows us to account for data which exhibit correlation and nonconstant variability as noticed in the previous section. The results show that time is significant. While the model captures the profile of the most important characteristic, time, the data has large variations associated with the Moran's values. Thus, we should be careful with methods that have an underlying normality assumption.

We simulate larger data of Moran's values using Monte Carlo inference technique about the mean of the Moran's statistics hoping to capture the distribution of the Moran's statistics at values of $\lambda = 2$. The algorithm ran at $10,000$ iterations produced the histogram displayed in Figure 3.

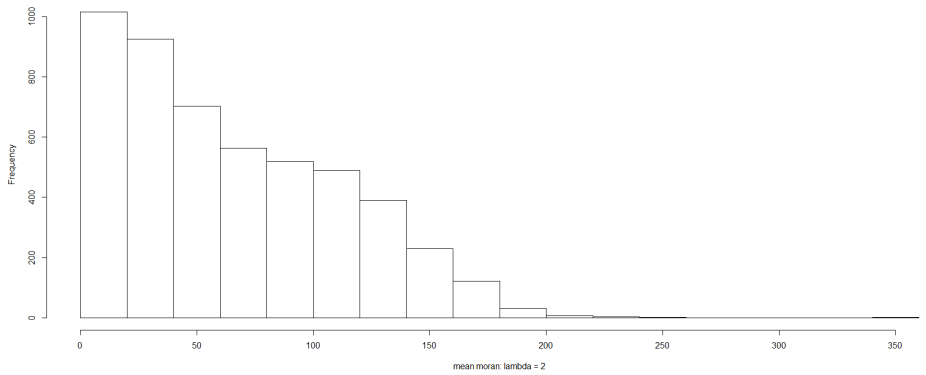

Figure 3 Histograms of Moran's statistics

Figure 3 shows that the distribution of Moran's values is skewed to the right, showing that estimation techniques of Moran that are based on normality assumptions may be questionable. This is due to the varying nature of the dependency of the intensity. However, if we did not allow for time interaction, then the well established methods associated with Moran's indices are sufficient for the analysis.

4.1 Hypothesis Test: Randomness

Spatial autocorrelation with glm is however not well understood (see [8-9]). To further conceptualize the idea of normality violations, we implement the Von Neumann rank test to assess whether or not our sample of points come from an underlying normal population. Using R for the calculations, the Von-Neumann test is constructed as in [19]:

Description: Tests if a sample is sampled randomly from an underlying normal population

Assumptions: Data are at least measured on an ordinal scale. And let $X_1, \ldots, X_n$ be a sequence of random variables with observations $x_1, \ldots, x_n$.

Hypothesis: H_0 : Sequence $X_1, \ldots, X_n$ is randomly generated vs H_1 : Sequence is not randomly generated.

Test Statistic: $Z = \left(1 - \frac{V}{2}\right)\sqrt{(N-2)/(N^2-1)}$, where $V = \frac{\sum_{i-1}^{N-1}\left(X_{i+1}-X_i\right)^2}{\sum_{i-1}^{N}\left(X_i-\bar{X}\right)^2}$.

We modify the test to account for the fact that our data is two dimensional. Using the standard Euclidean distance formula as "observations" of X_i's within disk j, we set

$$X_i = \sqrt{(x_1 - d_1)^2 + (x_2 - d_2)^2},$$

where (d_1, d_2) denotes the center of the disk and (x_1, x_2) a randomly generated point. That is, use the distance from center of the disks as a ranking mechanism. If our sample was indeed normally distributed, it would then be reasonable to see more points generated near the center of the disk or clustering at some area of the disk.

The results in Table 2 show non-significant p-values. In fact there is significant evidence that our points are randomly generated within each disk. This verifies that the normality assumption should not be used, randomness is present in some form, and we should look towards alternative models to analyze our data. These constants led to the use of extreme value distribution.

Table 2 p-value results of H_0 : sequence is randomly generated

Disc	1	2	3	4	5	6	7	8	9	10	11	12	13
p-value	.65	0.98	.23	.25	.24	.90	.76	.11	.68	.64	.22	.30	.89

Disc	14	15	16
p-value	.14	.94	.11

4.2 Fitting Extreme Values

One main concern is the nature of spread over time. The Moran values may grow larger and it might be of interest to investigate the extreme values as discussed in [6] or in [20]. For instance, understanding the areal spread of a rare disease is crucial to quarantine and protection. In this section, we focus on values in our output from disk 1, across all 5 time points (as an illustrative example), that are considered "extreme" and we model the data to a Generalized Pareto distribution (GPD). GPD's are useful since they can help us describe and understand the distribution function of a variable above a certain threshold. For more details about estimation of GPD, see [4].

The simulation output has some very large values (outliers) that may actually be from the way we defined our area and Moran's index as suggested in [15]. Points generated extremely close ($d << 1$) will result in extremely large Moran values. It is of interest to understand such behavior. Thus, we shift our focus to the first disk at time 1. Using the interquartile range

(IQR) as a measure of spread, since it is resistant to outliers, we can see from Table 3 below that $IQR = Q_3 - Q_1 = 5.17$ hence values $3 * (IQR)$ above Q_3 can be considered extreme value distributions.

Table 3 Quantiles for Disc 1 at time 1.

min	Q_1	median	Q_3	max
0	0	2.158214	5.169383	163.457831

Also notice the curve shape in Figure 4 of the QQ plot increasing from left to right indicates the distribution is right skewed. This is further evidence pointing to extreme value distributions.

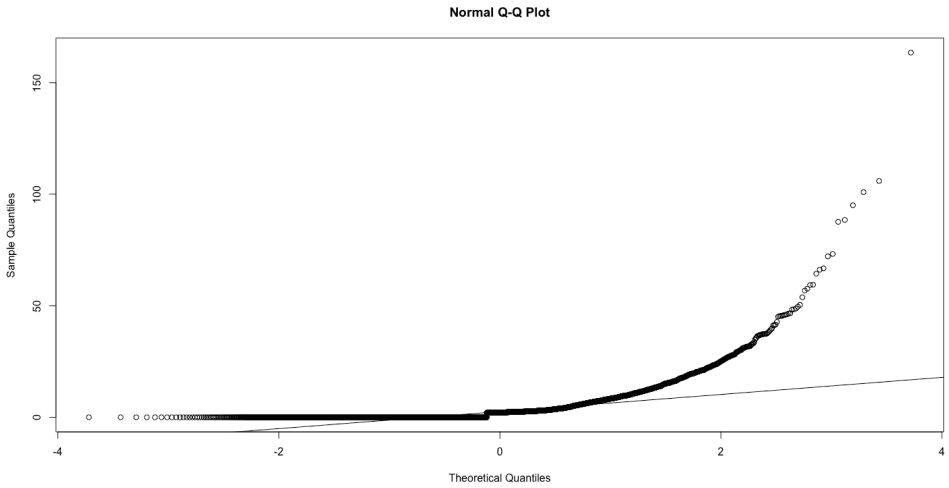

Figure 4 Normal QQ Plot of Moran values from simulation of Disk 1 at $t = 1$.

The generalized Pareto Distribution (GPD) is the classical asymptotically motivated model for excesses above a high threshold. If our data points (Moran statistics for disk 1) are independent and identically distributed above a threshold u, then the limiting distribution will be a GPD. In applications, the GPD is used as a tail approximation to the population distribution from which a sample excesses $x - u$ above some threshold u are observed.

$$G(x|u,\sigma_u,\xi) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-\frac{1}{\xi}}, & \xi \neq 0, \quad x \geq 0 \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma_u}\right)\right]_+ & \xi = 0, \quad x \geq 0 \end{cases} \tag{3}$$

The GPD is parameterized by the shape and scale parameters ξ and σ_u. In particular, the GPD as expressed in Equation (3) is expressed as exceedances $x > u$ where σ_u, u, ξ describe scale, location, and shape parameters, respectively as in [18]. In this representation the mean is equivalent to the threshold.

Parameters of the model are estimated again with Monte Carlo techniques and compared (under maximum likelihood estimation, method of moments, and probability weighted moments estimation) using the data from the 10,000 runs of the algorithm. We then fit the data to a GPD for the first disk for all time periods. For the first time period, we can see that the mean residual life plot (Figure 5) is linear almost everywhere but in particular, becomes slightly erratic above 50. This plot suggests that threshold $u = 12$ is an appropriate choice as sample size of $n_u = 512$ above excess provides a good balance between bias and variance of parameter estimation for the GPD.

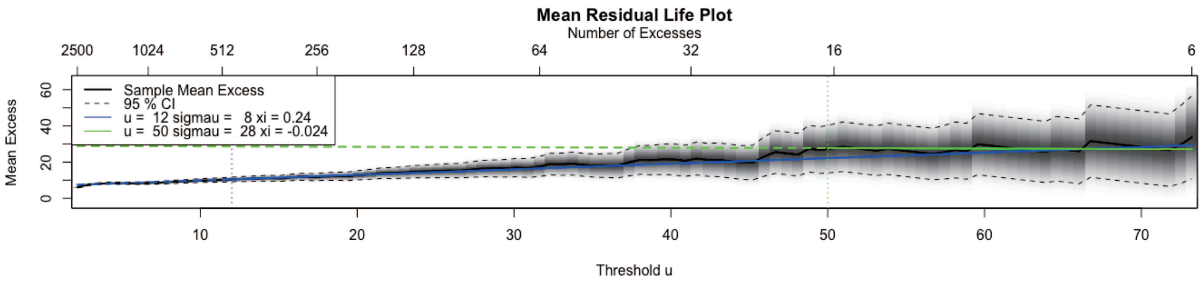

Figure 5

Table 4 Estimation comparison

Threshold u	No. Excess n_u	Shape ξ MLE (SE)	Shape ξ MoM (SE)	Shape ξ PWM (SE)	Scale σ_u MLE (SE)	Scale σ_u MoM (SE)	Scale σ_u PWM (SE)
5	1272	0.209 (0.032)	0.228 (0.101)	0.202 (0.034)	6.543 (0.276)	6.410 (0.817)	6.629 (0.291)
10	611	0.240 (0.049)	0.236 (0.181)	0.240 (0.051)	7.433 (0.468)	7.489 (1.743)	7.441 (0.477)
12	468	0.236 (0.055)	0.233 (0.189)	0.234 (0.058)	8.001 (0.570)	8.047 (1.935)	8.032 (0.587)
15	339	0.288 (0.071)	0.248 (0.687)	0.287 (0.072)	7.865 (0.690)	7.865 (7.525)	007.834 (0.687)
20	193	0.321 (0.099)	0.246 (0.599)	0.317 (0.101)	8.733 (1.052)	9.565 (7.537)	8.665 (1.024)
25	114	0.276 (0.121)	0.220 (0.279)	0.280 (0.124)	10.964 (1.658)	11.720 (4.067)	10.815 (1.630)
30	74	0.288 (0.159)	0.208 (0.286)	0.296 (0.157)	12.181 (2.362)	13.361 (4.698)	11.874 (2.238)

Using R along with the `POT` package, we fit a GPD using the threshold value $u = 12$. We then select threshold values below and above 12 to observe any trend and find one that fits best.

Table 4 shows the parameter estimates using maximum likelihood estimation (MLE), method of moments (MoM), and probability weighted moments (PWM). While the MLE has consistently low standard errors, the PWM may be a better choice for higher threshold values. The MoM on the other hand, has larger standard errors than the other two methods.

Moreover, in extreme value theory, there are three main domains of attraction: Gumbel, Fréchet, and Weibull. The distributions in the Gumbel domain have the exponential, including normal and Gamma as the limiting distribution of their tails. The Fréchet domain contains distributions with infinite yet heavier tails. And lastly, the Weibull domain contains distributions with lighter tails than the exponential distribution. Thus we want to test for the domain of attraction which is determined by the shape parameter ξ, as in [3]. That is, the test of hypothesis $H_0 : \xi = 0$ vs $H_a : \xi > 0$ is equivalent to testing a Gumbel versus a Weibull domain of attraction. Using a 95% confidence interval for all estimated values of ξ under the maximum likelihood estimation, Table 5 indicates that ξ is likely to be positive. So we reject $H_0 : \xi = 0$ for supporting evidence that our data follows a Weibull domain of attraction.

Table 5 95% CI for shape parameter ξ

Threshold u	MLE	CI
5	0.209	(0.147, 0.271)
10	0.240	(0.144, 0.336)
12	0.236	(0.129, 0..343)
15	0.288	(0.150, 0.426)
20	0.321	(0.128, 0.514)
25	0.276	(0.040, 0.512)
30	0.288	(−0.022, 0.598)

5. Conclusion

In this study, we show that for a given general area on some time scale, the global Moran statistic does not adequately summarize or capture the spatial data. Instead we looked into dividing the area into smaller subareas adding time, and obtained spatio-temporal local Moran's values. This process of calibration or refinement allows us to better capture the spatio-temporal information. Because we added a time factor, Moran's values and assumption underlying process generating them were explored. Extreme values of the Moran's values were tested for randomness, and fitted under extreme value distribution. Because some of the Moran's values are large, a Generalized Pareto distribution that models excessiveness above an appropriate threshold was considered and validation performed.

The analysis of the simulated Moran values suggests that methods of interpreting measures of spatial correlation will incorporate different assumptions. Many measures of such correlations will deal with critical ways the data is generated, their Geographic Information System, the

interpretation of their values and their relationships. Meanwhile, our interpretation has shown that relative simplistic methods can lead to misspecification of statistical properties and biased decisions.

Acknowledgment

We gratefully acknowledge the critical reviews by the editor and two anonymous reviewers. Their constructive comments and suggestions certainly improved the quality and presentation of the paper.

References

[1] Anselin, L. (1995). Local indicators of spatial association - LISA, *Geographical Analysis*, **27**(2), 93-115.

[2] Baddeley, A. (2007). Spatial Point Processes and their Applications. In Lecture Notes in Mathematics: Stochastic Geometry, Vol 1892, Springer, Berlin, Heidelberg.

[3] Beisel, C., Rokyta, D., Wichman, H., and Joyce, P. (2007). Testing the extreme value domain of attraction for distributions of beneficial fitness effects, *Genetics*, **176**, 2441-2449.

[4] Castillo, J. and Daoudi, J. (2009). Estimation of the generalized Pareto distribution, *Statistics and Probability Letters*, **79**, 684-688.

[5] Cliff, A. D. and Ord, J. K. (1981). Spatial Processes: Models and Applications, Pion Ltd, London.

[6] De Jong, P., Sprenger, C., and van Veen, F. (1984). On extreme values of Moran's I and Geary's c, *Geographical Analysis*, **16**, 17-24.

[7] Deng, X., Xu, Y., Han, L., Yang, M., Yang, L., Song, S., Li, G., and Wang, Y. (2016). Spatial-temporal evolution of the distribution pattern of river systems in the plain river network region of the Taihu Basin, China, *Quaternary International*, **392**(21), 178-186.

[8] Gittleman, J. L. and Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects, *Systematic Zoology*, **39**, pp. 227-241.

[9] Griffith, D. (2005). Spatial Autocorrelation, in: Encyclopedia of Social Measurement, Vol. 3, edited by K. Kempf-Leonard, 581-590, Elsevier, Amsterdam.

[10] Griffith, D. (2009). Methods: Spatial Autocorrelation, in: International Encyclopedia of Human Geography, Edited by R. Kitchin and N. Thrift, 396-402, Elsevier, New York.

[11] Kallenberg, O. (2002). Foundations of Modern Probability, 2nd edition, Springer, New York.

[12] Lee, J. and Li, S. (2017). Extending Morans index for measuring spatiotemporal clustering of geographic events, *Geographical Analysis*, **49**, 36-57.

[13] Lorio, J., Diawara, N., and Waller, L. (2018). Density estimation of spatio-temporal point

patterns using Moran's statistics, *International Journal of Statistics and Probability*, **7**(2), 80-90.

[14] Martin, R. L. and Oeppen, J. E. (1975). The identification of regional forecasting models using space: time correlation functions, *Transactions of the Institute of British Geographers*, **66**, 95-118.

[15] Maruyama, Y. (2015). An alternative to Moran's I for spatial autocorrelation, Retrieved from http://arxiv.org/abs/1501.06260.

[16] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena, *Biometrika*, **37**(1/2), 17-23.

[17] Paradis, E. (2016). Moran's Autocorrelation Coefficient in Comparative Methods. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. APE Package vignette.

[18] Scarrott, C. and MacDonald A. (2012). A review of extreme value threshold estimation and uncertain quantification, *Revstat-Statistical Journal*, **10**(1), 33-60.

[19] Taeger, D. and Kuhnt, S. (2014). Statistical Hypothesis Testing with SAS and R, Wiley, Chichester, UK.

[20] Tiefelsdorf, M. and Boots, B. (1997). A note on the extremities of local Moran's I_is and their impact on global Moran's I, *Geographical Analysis*, **29**(3), 248-257.

[21] Vaillant, J., Puggioni, G., Waller, L. A., and Daugrois, J. (2011). A spatio-temporal analysis of the spread of sugarcane yellow leaf virus, *Journal of Time Series Analysis*, **32**, 392-406.

[22] Waller, L. A. and Gotway, C. A. (2004). Spatial Data, in Applied Spatial Statistics for Public Health Data, John Wiley & Sons, Inc., Hoboken, NJ, USA.

[23] Wang, Y. F. and He, H. L. (2007). Spatial Data Analysis Method, Science Press, Beijing.