Spatial and Non-Spatial Statistical Tools for Data Analysis: A Comparative Study Using Malaria Incidence Data in Kassala State

Montasir Ahmed Osman Mohamed
Department of Mathematics, Faculty of Education, University of Kassala, Kassala, Sudan
Corresponding author email: a.montasir@yahoo.com

Abstract

Several advanced statistical analysis tools are recently improved depending on the geographical information system. So, this paper aims at utilizing of spatial statistical analysis to analyze the distribution of malaria incidence in localities of Kassala State and to compare the results with the results of Chi-squared test for goodness of fit. Two software; ArcGis and GeoDa, are used to create and provide the map with annual malaria incidence for each locality. The statistical techniques which are offered by the two software are used to test the malaria distribution randomness among localities. Moran's I test is used in addition to local and global measures to investigate in the distribution of malaria incidence randomness as well as spatial autocorrelation existence. The results reveal that the distribution of malaria incidence in localities is random and spatial autocorrelation does not exist. On the contrary to the goodness of fit test, the study reveal that the distribution of malaria is random and spatial autocorrelation does not exist. The reason of the difference between the two methods is that the spatial statistical analysis depends on the geographical of data as an independent variable, unlike the latter.

Keywords: Spatial Statistical Analysis; Spatial Autocorrelation; Randomness.

الأدوات الإحصائية المكانية وغير المكانية لتحليل البيانات: دراسة مقارنة باستخدام بيانات الإصابة بالملاريا في ولاية كسلا

منتصر أحمد عثمان محمد
قسم الرياضيات، كلية التربية، جامعة كسلا ، كسلا، السودان
البريد الإلكتروني للباحث المرسل:a.montasir@yahoo.com

مستخلص الدراسة

هنالك عدد من الأساليب الإحصائية التي طورت حديثاً اعتماداً على نظم المعلومات الجغرافية. و لذلك هدفت هذه الدراسة إلى استخدام التحليل الإحصائي المكاني لتحليل توزيع الإصابة بمرض الملاريا عبر محليات ولاية كسلا المختلفة و مقارنة النتائج مع نتائج التحليل باستخدام اختبار مربع كاي لجودة التوفيق. و قد تم استخدام برنامجي ArcGis و GeoDa لرسم و تزويد الخريطة بإصابات الملاريا السنوية لكل محلية. و استخدمت التقنيات الإحصائية التي يوفرها البرنامجان لاختبار عشوائية توزيع الملاريا عبر محليات الولاية المختلفة. كما استخدِم اختبار Moran's I لاختبار عشوائية التوزيع إضافة إلى المقاييس المحلية (Local) و العالمية (Global) لاختبار وجود ارتباط ذاتي مكاني للإصابة بالملاريا في المحليات. و قد بينت النتائج أن انتشار الملاريا في محليات الولاية عشوائياً و لا يوجد به ارتباط ذاتي مكاني. و هذه النتائج تخالف ما أظهرته نتائج اختبار جودة التوفيق حيث أظهر وجود علاقة معنوية بين الإصابة بالملايا و المحليات. و يرجع سبب الاختلاف في النتائج بين الطريقتين إلى أن التحليل الإحصائي المكاني يعتمد الموقع الجغرافي للبيانات كمتغير مستقل على عكس الطريقة الأخرى.

كلمات مفتاحية: التحليل الإحصائي المكاني، الارتباط الذاتي المكاني، العشوائية.

1. Introduction:

Several advanced statistical analysis tools are improved depending on geographical information system (GIS). These tools essentially deal with geographical distribution of a phenomenon. The most common of these tools is the analysis of spatial distribution. Its goal is to discover the relationship between that what happens in a region and what happens in neighboring regions. The measurement of this relationship is spatial autocorrelation. This is done according to its value sign classified to positive or negative spatial autocorrelation. The positive spatial autocorrelation occurs when the high value is surrounded by high values or low value is surrounded by low values. While the negative one occurs when the high value is surrounded by low values or low value is surrounded by high values (Ord and Getis, 2001) (Anselin and Bera,1998),(Ord and Getis 1995) (Getis and Ord, 1992), and (Geary, 1954).

Malaria is the most common disease in Africa and considered the major cause of illness and death in these countries. The highest widespread of malaria among the world eventuates in tropical and subtropical regions. The world health organization (WHO) reported that, there are 97 malaria endemic countries and 3.3 million are at risk of malaria, estimated 584 thousand deaths. 90% of all deaths occur in Africa (WHO, 2014). In Kassala State there are4-6% of citizens infected by malaria each year. The children under 5 years are mostly affected by malaria incidence and death (National Program to Control Malaria, 2013).

Accurate statistical estimates are helpfulness to anti-malaria organizations in term of control and spread of malaria incidence. So, advanced statistical techniques are used in this paper to figure out the patterns of malaria distribution in Kassala State. The capability of spatial statistical methods, which enable capturing of spatial pattern among data are motivated this research.

2. Spatial autocorrelation and its measures:

Spatial autocorrelation investigates in term of what happens in one region related or not related to what happens in neighboring regions and its measuring processes depends on geographical data. The required data should be point or polygons. The measures which are used to scale spatial autocorrelation are classified into two types. The first type is global measures of spatial autocorrelation which focus on the relationship among the whole data. The second type is local measures of spatial autocorrelation which focus on the relationship between each sub-region and neighboring regions.

2.1 Global measures of spatial autocorrelation:

Global measures of spatial autocorrelation are measures applied to display in a single value of the pattern of distribution for single variable (there is spatial autocorrelation or there is no spatial autocorrelation). The distribution pattern investigates in randomness of data spread among the whole region. The investigation has to figure out one of two cases, either there is clustering in data distribution or there is no clustering. Several measures are used (in the calculation process) to calculate these measures. Three of them are discussed in the following paragraphs.

2.1.1 Moran's I

The most common measure of spatial autocorrelation is Monran's I. It uses the points or polygons of the regions to (without writing to) as well as variable values to compute the value of spatial autocorrelation (Moran, 1950). Its formula is:

Where: $$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(x_i - \bar{x})^2} \quad \dots \quad (1)$$

N is the number of observations (points or polygons)

$\bar{x}$ is the mean of the variable values

x_i is the variable value at particular location

x_j is the variable value at another location

W_{ij} is a weight indexing location of i relative to j

n is number of neighboring regions.

2.1.2 General G-statistics

General G-statistics is a common used measure in term of spatial autocorrelation according to this formula (Getis and Ord, 1992):

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \quad \ldots (2)$$

Where:

x_i is the variable value at particular location?

x_j is the variable value at another location?

d is neighborhood distance

W_{ij} is a weight matrix has only 1 or 0. 1 if j is within d distance of i and 0 if it is beyond that distance

n is number of neighboring regions

2.2 Local measures of spatial autocorrelation:

Local measures are used to calculate the spatial autocorrelation between each sub-region and its neighboring which shared the borders. There are local versions of the above three measures as follow:

2.2.1 Local Moran's I

The local version of Monran's I as follows (Anselin, 1995):

$$I_i = z_i \sum_{j=1}^{n} W_{ij} z_j \quad \ldots (3)$$

Where:

$z_i = \frac{x_i - \bar{x}}{SD_x}$

W_{ij} is a weight indexing location of i relative to j

n is number of neighboring regions

2.2.2 Local general G-statistics:

Local general G-statistics is a common used measure in term of spatial autocorrelation according to this formula (Ord and Getis, 2001):

$$G = \frac{\sum_{j=1}^{n} W_{ij} x_j}{\sum_{i=1}^{n} x_j} \qquad \dots \quad (4)$$

Where:

x_j is the variable value at another location?

W_{ij} is a weight matrix has only 1 or 0. 1 if j is within d distance of i and 0 if it is beyond that distance

n is number of neighboring regions

3. The study area and data:

Area of study is Kassala State which located in eastern Sudan. Kassala State includes 11 localities, which are all covered in the study. These localities are: Hamashkoraib, Northern Aldalta, New Halfa, Atbara River, Kassala, Talkook, Girba Wad Alhelaiw, Aroma, Kassala Rural and Western Kassala. The following map shows Kassala localities. Limitation of the study is stranded in Kassala State borders.

Figure 1: Kassala State map

Figure 1 shows the 11 Kassala localities. Hamashkoraib is in the northern state,
Wad Alhelaiw in the southern, Talkook in the eastern and Atbara River in western.

The study units are citizens infected by malaria in Kassala localities in 2013 according to National Program to Control Malaria (2013). The data is analyzed and visualized by AtcGis and Geoda software.

4. Results

The results which include the global and local tests for spatial autocorrelation as well as goodness of fit test are displayed in this paragraph. Moran's I is used in global and local versions to investigate in spatial autocorrelation among the whole data and clustering.

4.1 Global tests of spatial autocorrelation:

Global Moran's I is used to test spatial autocorrelation among malaria incidence in Kassala State localities. Table 1 shows the results.
Table 1: Results of global Moran's I test

Coefficient	Observed	Expected	St. d	Z	P-value
Moran's I	0.17	-0.10	0.14	1.77	0.08

From table 1 the p-value is greater than 0.05 and that means acceptance of null hypothesis$H_0: \rho = 0$, in other words there is no significant spatial autocorrelation in malaria incidence in localities of Kassala State.

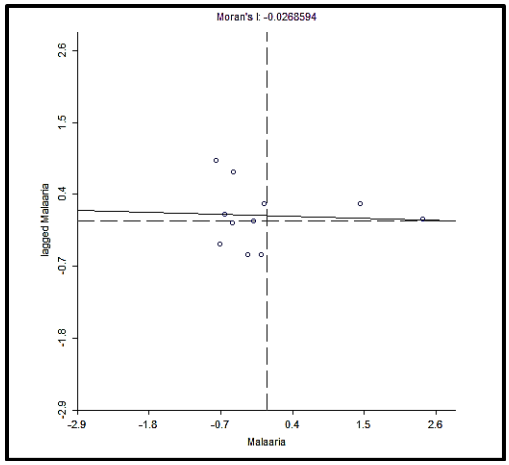

Figure 2: Global Moran's I scatter plot for malaria incidence.

Figure 2 shows that the values of malaria incidence in a locality and its neighboring are widely scattered in different directions and most of them are far from the estimated line. That wide spread means that there is no relationship between incidence in locality and its neighboring.

4.2 Local tests of spatial autocorrelation:

Local version of Moran's I and general G statistics are used to test the clustering of malaria incidence in Kassala localities. Tables 2 and 3 below show the results.

Table 2: Results of local Moran's I test

Locality	Observed	Cluster	P-value
Hamashkoraib	-0.064	0	0.440
Northern Aldalta	0.153	2	0.020
New Halfa	0.042	0	0.090
Atbara River	0.261	0	0.100
Kassala	0.370	0	0.270
Talkook	0.002	0	0.360
Girba	-0.010	0	0.340
Wad Alhelaiw	-.0382	0	0.110
Aroma	0.021	0	0.260
Kassala Rural	0.060	0	0.330
Western Kassala	-0.722	3	0.010

The first column in table 2 is locality name, the second one is observed value for Moran's I test, the third one is the number of localities which are clustered as high or low values compared to specific locality and the last column is the probability value for Moran's I test.

Table 2 shows that, p-values for all localities are greater than 0.05 except northern Aldalta and western Kassala. Northern Aldalta locality has a significance cluster with positive Moran's I, which means locality has similar rate of malaria incidence high or low as same as its neighbors. western Kassala locality has a significance cluster with negative Moran's I, the rate of malaria incidence is different contrast to the locality neighbors.

Table 3: Results of local general G statistics test

Locality	Observed	Cluster	P-value
Hamashkoraib	0.051	0	0.360
Northern Aldalta	0.036	2	0.030
New Halfa	0.041	2	0.050
Atbara River	0.041	0	0.130
Kassala	0.148	0	0.330
Talkook	0.085	0	0.350
Girba	0.114	0	0.350
Wad Alhelaiw	0.148	0	0.090
Aroma	0.075	0	0.230
Kassala Rural	0.148	0	0.250
Western Kassala	0.158	1	0.030

The first column in table 3 is locality name, the second one is observed value for general G test, the third one is the number of localities which are clustered as high or low values compared to neighbor localities and the last column is the probability value.

Local general G statistics shows no clustering in malaria incidence in Kassala localities except northern Aldalta, New Halfa and western Kassala. Regarding to the difference in local Moran's I and local general G statistics formulas, the different classification for the same

data are expected. In this research, the significant spatial autocorrelation is revealed in two localities by local Moran's I contrast to three localities by local general G statistics.

The following graph display clustering of malaria incidence in Kassala localities.

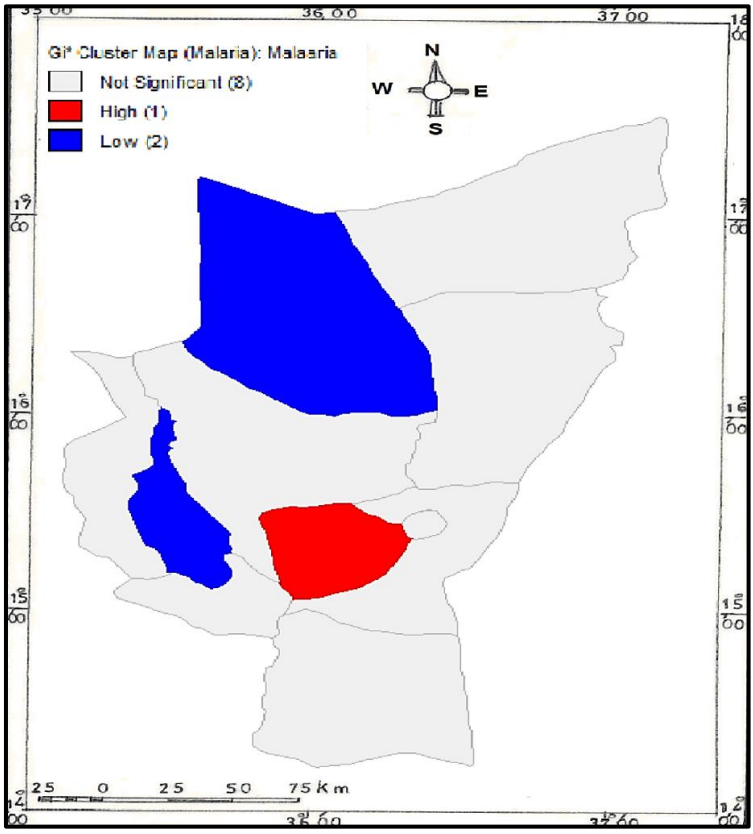

Figure 3: Local general G statistics clustering map of malaria incidence in Kassala State.

Figure 3shows that western Kassala is clustered as surrounded by high values and northern Aldaltaand new Halfa are clustered as surrounded by low values. (more light should be thrown here, in form of a conclusion to the non-statistician, or what do you mean?).

5. Goodness of fit test:

Goodness of fit test represented in Chi-squared as a traditional method is used here to test the randomness of malaria incidence in Kassala localities.

Table 4: Chi-squared test for goodness of fit

χ^2-value	d. f	P-value
4198.60	10	0.00

Table 4 shows that there is a significant relationship between malaria incidence and localities and this is opposite to what figured out in spatial autocorrelation tests (more concentration or elaboration is needed to define the different columns and to explain what is meant by the figures).

5. Discussion:

The results of the study showed that there is no spatial autocorrelation in the malaria incidence among Kassala State localities according to Moran's I test and general G statistics. The local tests revealed that there is one Low-High cluster in case of western Kassala locality and two Low-Low clusters in case of northern Aldalta and new Halfa. The spatial tests result is different from goodness of fit test which pointed out relationship (is it statistically significant or not?) between malaria incidence and localities. The reason of this contradiction refers to use of geographical attributes as explore variable in the spatial statistical analysis case. So, spatial statistical methods are recommended when the data is collected in spatial dimensions.

References:

1. Anselin, L. (1988). Spatial Econometrics: Methods and Models, New York.
2. Anselin, L. (1995). Local Indicators of Spatial Association – LISA, Geographical Analysis 27: 93-105.
3. Anselin, L. and Bera, K. (1998). Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics, Handbook of Applied Economic Statistics. Marcel Dekker, New York.
4. Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping, The Incorporated Statistician, 5 (3): 115–145.
5. Getis, A and Ord, J.K. (1992). The analysis of spatial association by use of distance statistics, Geographic Analtsis, 24(3): 189-206.
6. Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena, Biometrika 37 (1): 17–23.
7. National Program to Control Malaria (2013). Kassala State, Report.
8. Ord, J.K. and Getis, A (1995), Local Spatial Autocorrelation Statistics: Distributional Issues and an Application, Geographical Analysis 29: 93-115.
9. Ord, J.K., and A. Getis (2001). Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation, Journal of Regional Science 41 (3): 411– 432.
10. WHO (2014). World Health Organization, Available http://www.who.int/malaria/media/world_malaria_report_2014/en/

Appendix A
Research data

Locality	Population 2013	No. Maria Incidence 2013
Hamashkoraib	304334	1490
Northern Aldalta	109497	1882
New Halfa	252567	6137
Atbara River	163214	357
Kassala	355882	27104
Talkook	327807	6535
Girba	117947	3036
Wad Alhelaiw	100951	974
Aroma	122501	1086
Kassala Rural	184337	20130
Western Kassala	94626	306