Ali Mohammed BABA, PhD Candidate
Email: ambabastats@gmail.com
Institute for Mathematical Research
Universiti Putra Malaysia, Selangor, Malaysia
Professor Habshah MIDI, PhD
Email: habshahmidi@gmail.com
Institute for Mathematical Research and Faculty of Science
Universiti Putra Malaysia, Selangor, Malaysia
Nur Haizum Abd RAHMAN, PhD
Email: nurhaizum@upm.edu.my
Institute for Mathematical Research and Faculty of Science
Universiti Putra Malaysia, Selangor, Malaysia

A SPATIAL OUTLIER DETECTION METHOD FOR BIG DATA BASED ON ADJACENCY WEIGHTED RESIDUALS AND ITS APPLICATION TO COVID-19 DATA

Abstract. *Identification of spatial outlier is essential in revealing hidden useful knowledge in different fields of statistical applications for big data. The Score Statistic (SC_i) has been used as a diagnostic tool for the identification of spatial outliers in big data. Nonetheless, the SC_i method suffers from masking and swamping effects. In order to reduce the swamping effect, we propose two methods denoted as t_{MDR} and t_{EW}. The t_{MDR} and t_{EW} methods adopt location adjacency to construct spatial weights, namely metric distance reciprocal (MDR) weight and exponential weight (EW), respectively, to detect outliers in spatial autoregressive-regressive model (SAR), spatial autoregressive error model (SEM) and general spatial autoregressive-regressive model (GSM). Difference between spatial residuals are calibrated to incorporate adjacency effect into spatial outlier residual. The results of the simulation study and real example show that the performances of the three methods are equally good for SAR model. The t_{MDR} and t_{EW} are comparable and both outperform the SC_i for SEM and GSM models with less swamping effects and less computational running times.*

Keywords*: Adjacency, Calibration, Masking, Spatial autoregressive, Spatial outlier, Swamping.*

JEL Classification : C01, C21, C55

1. Introduction

Outlier is an observation that deviates so much from other observations so as to arouse suspicion that it was generated by a different mechanism (Hawkins, 1980). Detecting outlier results in revealing hidden features that might not have been

DOI: 10.24818/18423264/55.3.21.06

considered. The effect of outlier, if not detected and treated, is disastrous that it renders classical statistical methods of estimators erroneous, thereby resulting in invalid conclusions. In contrast to classical outliers, spatial outliers are location related outliers. They are rather local than global in the sense of spatial position. Kou and Lu (2008) stated that spatial outlier breaks the spatial autocorrelation and continuity in the spatial location. Aggarwal (2013) defined spatial outliers as objects which have behavioral attribute values that are distinct from those of their surrounding spatial neighbors. Haining (2003) noted that attribute values can be extreme depending on their position on the map. Such attributes are termed as spatial outliers because their attribute values are extreme relative to the set of values in their neighborhood on the map. i.e., it is possible for an attribute to be a spatial outlier without necessarily being extreme in the distributional sense. Kou and Lu (2008) noted that: "detecting spatial outliers can help in locating extreme meteorological events such as tornadoes and hurricanes, identify aberrant genes or tumor cells, discover highway traffic congestion points, pinpoint military targets in satellite images, determine possible locations of oil reservoirs and detect water pollution incidents".

There are handful methods of classical outlier detection in the literature. However, spatial outliers are handled in a different way from the classical outliers due to factors such as spatial autocorrelation and spatial dependence. Techniques for detecting spatial outliers include both graphical and quantitative tests. The graphical methods entail visualization of the spatial data in which violating certain criteria prompt observations as outliers. Most common graphical techniques are the Scatterplot (Anselin, 1994) and the Moran Scatterplot (Anselin, 1995) (see Kou and Lu, 2008) for extensive discussion of graphical method of spatial outlier detection). Anselin (1995) suggested a quantitative test for outlier detection based on the global measure of autocorrelation and the Local Indicator of Spatial Autocorrelation (LISA). He assumed that the null hypothesis that there is no spatial autocorrelation and the sum of the LISA is proportional to the global spatial autocorrelation. Based on this, a threshold value is established to decide on declaring a spatial location as outlier. However, this method is prone to masking and swamping effect due to the influence of neighbors with high/low attribute values (Kou and Lu (2008)). Shekhar et al. (2002) proposed the S(x) spatial statistic for quantitative test, where Statistic S(x) is the difference of the attribute value of each data object x and the average attribute value of x's neighbors. Lu et al. (2003) identified masking and swamping as weakness of Shekhar et al. (2002) method, where inlying observations are declared as outliers while outliers are suppressed by the aggregate neighborhood function. They proposed spatial outlier detection techniques, namely the Iterative Z, Iterative R, and Median Z algorithms. The Iterative-Z and the iterative-R detect spatial outliers through multiple

iterations, and the median based method is a non-iterative algorithm but uses the median instead of the mean to represent the average of a set of neighbors in the Z-value method. However, the authors used a small data on the method and it's not measured for reliability. In a similar way, Liu et al. (2010) emphasized on the problem of masking/swamping. They developed two random walk-based approaches, namely the RW-BP (Random Walk on Bipartite Graph) and RW-EC (Random Walk on Exhaustive Combination) based on the spatial and/or non-spatial attributes of the spatial objects. Two weighted graphs, a BP (Bipartite graph) and an EC (Exhaustive Combination) are used to compute the scores between the spatial objects and outlyingness are ranked, and the top k objects are declared outliers. However, no specific cutoff points were established to identify which observations are the outliers. Singh and Latiha (2018) adopted local quotient (LQ) to compare non-spatial attribute value of a spatial point with its corresponding values in its spatial neighborhood via the ratio similar to location quotients. They subsequently used the algorithms in Lu et al. (2003) to accomplish the detection of outliers. Hadi and Imon (2018) proposed a measure of spatial distance that is calculated for all n observations instead of n - 1. This measure uses the observations in the left and right of a spatial observation to compute its value. Despite the numerous mentioned advantages such as identification of cluster of contiguous outliers, it lacks the capacity to trace multi-neighbor contiguity, which gives room to both masking and swamping effect. Dai et al. (2016) derived a statistic for the mean shift outlier model and variance weight model. Mean shift model is described by extra parameters in the functional or stochastic model. Though, the score statistic used in mean shift model (Dai et al (2016)) has good performance in outlier detection in the spatial autoregressive-regressive model (SAR), spatial autoregressive error model (SEM) and general spatial autoregression model (GSM), it has high swamping effect when the coefficient of spatially lagged parameter, ρ, and that of autoregressive error parameter, λ, are high. Gaspard et al. (2019) noted that residual spatial autocorrelation (which refers to the difference between the observed and the predicted values in the model) indicates the amount of spatial autocorrelation in the variance that are not explained by explanatory variables. They reiterated the fact that failure to appropriately address residual spatial autocorrelation will lead to problems such as underestimating standard error, biased parameter estimates and model misspecification. Another weakness is poor performance in time in the face of large data. These weaknesses motivated us to develop a new method which we call the Adjacency Weighted Spatial Outlier Residual.

The rest of the paper is organized as follows. Section 2 presents the Score Statistic (SC_i) for the identification of spatial outlier. The proposed Adjacency Weighted Spatial Outlier Residual is presented in Section 3. Section 4 discusses the simulation study to evaluate and compare the performance of the proposed methods and the Score Statistic. The application of the proposed methods to a Covid-19 data is presented in Section 5. The concluding remarks are given in Section 6.

2. The Score Statistic (SC_i) for the identification of spatial outlier.

The general spatial autoregression model (GSM) has been employed in various fields of statistical applications such as economics, geography, ecology, public health etc. (Anselin (1988), Lesage (1999)).

The general spatial autogressive model is given by:

$$y = \rho W_1 y + X\beta + \xi,\ \xi = \lambda W_2\ \xi + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n}), \quad (1)$$

where y is an $n \times 1$ vector of dependent variable. X is an $n \times k$ matrix of explanatory variables. W_1 and W_2 are known $n \times n$ spatial weight matrices. $\boldsymbol{I_n}$ is an $n \times n$ identity matrix. ξ is the spatially correlated error terms, ε is the random residual terms. The parameter ρ is a coefficient on the spatially lagged dependent variable Wy and λ is a coefficient on the spatially correlated errors. The estimation of the parameters ρ, λ, σ^2 and β are extensively discussed by Anselin (1988) and Lesage (1999).

The general spatial autoregressive model (1) can be re-written as

$$Ay = X\beta + \xi, \quad (2)$$

which implies that

$$y = A^{-1}(X\beta + \xi), \text{ where,}$$

$$\xi = B^{-1}\varepsilon,\ \ A = \boldsymbol{I_n} - \rho W_1,\ \ B = \boldsymbol{I_n} - \lambda W_2,\ \xi \sim N(\mathbf{0}, \sigma^2 V^{-1}),\ \text{and}\ \ V = B^T B.$$

Different special spatial models can be obtained by imposing different restrictions on Equation 1.

When X = **0** and W_2 = **0**, Equation 1 results in the first order spatial autoregressive model (FAR) given as

$$y = \rho W_1 y + \varepsilon,\ \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n}). \quad (3)$$

The model (3) expresses y as a linear combination of its neighbors.

The mixed regressive-autoregressive model consists of the first order spatial autoregressive and the explanatory variables in X. This is expressed as

$$y = \rho W_1 y + X\beta + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n}). \quad (4)$$

This model is referred as spatial autoregressive model (SAR). It is also termed as mixed regressive - spatial autoregressive model because it combines both the standard regression model with a spatially lagged dependent variable.

We notice when $W_1 = \mathbf{0}$, model (1) becomes the regression model with spatial autocorrelated residuals. This is given by

$$y = X\beta + \xi, \xi = \lambda W_2 \xi + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n}) \quad (5)$$

The model (5) is termed as the spatial autoregressive error model (SEM).

Shi and Chen (2009) adapted the studentized residual in the multiple linear regression to the detection of spatial outlier settings. The adaptive residual is denoted as t_i and given as

$$t_i = \frac{\hat{\varepsilon}^*}{\hat{\sigma}\sqrt{1 - \frac{\hat{p}_{ii}}{\hat{v}_{ii}}}}, \quad (6)$$

where, $\hat{\varepsilon}^* = \frac{\hat{r}_i}{\sqrt{\hat{v}_{ii}}}$, $\hat{r}_i$ is the i^{th} term of the matrix $\hat{V}\hat{\varepsilon}$, $\hat{p}_{ii}$ is the i^{th} diagonal element of the matrix $\hat{P} = \hat{V}X(X^T\hat{V}X)^{-1}\hat{V}$ and $\hat{v}_{ii}$ is the i^{th} diagonal element of the matrix $\hat{V}$.

Dai et al. (2016) proposed a score statistic for mean shift outlier model and variance weight model, where they introduced $d_i\gamma$ in the GSM model. The d_i and γ are the outlier indicator and its modelling parameter, respectively. The mean shift model is expressed as

$$y = \rho W_1 y + X\beta + d_i\gamma + \xi, \xi = \lambda W_2 \xi + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n}). \quad (7)$$

They derived a score statistic, SC_i, which has asymptotic chi squared distribution. $SC_i \sim \chi^2_{(1)}$.

$$SC_i = \frac{t_i^2}{1 - \frac{\hat{b}_i^2}{\hat{\kappa}\hat{v}_{ii}(1 - h_{ii})}} \quad (8)$$

where, t_i is as defined in Equation 6.

$\widehat{b_i}$ is the i^{th} element of vector $\hat{b} = \hat{Q}\hat{\eta}, with\ \hat{\eta} = W_1\hat{A}^{-1}X\hat{\beta}$,

$\hat{Q} = \hat{V} - \hat{P}, \hat{A} = I_n - \hat{\rho}W_1, \hat{B} = I_n - \hat{\lambda}W_2$ and $\hat{\beta} = \left(X^T\hat{V}X\right)^{-1}X^T\hat{V}\hat{A}y$.

$\hat{\kappa} = \hat{\sigma}^2\hat{c}_{11} + \hat{\eta}^T\hat{Q}^T\hat{\eta} - \frac{\hat{\sigma}^2(n\hat{c}_{12}^2 - 4\hat{c}_1\hat{c}_2\hat{c}_{12} + 2\hat{c}_1^2\hat{c}_{22}}{n\hat{c}_{22} - 2\hat{c}_2^2}$, $c_i = trace(\hat{C}_i)$, $\hat{C}_{ij} = \hat{C}_i^T\hat{C}_j + \hat{C}_i\hat{C}_j$, $\hat{C}_1 = \hat{B}W_1\hat{A}^{-1}\hat{B}^{-1}, \hat{C}_2 = W_2\hat{B}^{-1}$, $h_{ii} = \frac{\hat{p}_{ii}}{\hat{v}_{ii}}$.

The SC_i depends on t_i in the numerator and $\frac{\hat{b}_i^2}{\hat{\kappa}\hat{v}_{ii}(1-h_{ii})}$ in the denominator. The adaptive studentized residual regression in Equation 6 can be simplified as follows:

$$t_i = \frac{\hat{\varepsilon}^*}{\hat{\sigma}\sqrt{1-\frac{\hat{p}_{ii}}{\hat{v}_{ii}}}},$$

$$= \frac{\frac{\hat{r}_i}{\sqrt{\hat{v}_{ii}}}}{\hat{\sigma}\sqrt{1-\frac{\hat{p}_{ii}}{\hat{v}_{ii}}}}, \quad \hat{\varepsilon}^* = \frac{\hat{r}_i}{\sqrt{\hat{v}_{ii}}}$$

$$= \frac{\hat{r}_i}{\hat{\sigma}\sqrt{\hat{v}_{ii} - \hat{p}_{ii}}},$$

where, t_i is the i^{th} term of the of the matrix $\widehat{V}\hat{\varepsilon}$. Now,

$$\widehat{V}\hat{\varepsilon} = \hat{V}(\hat{A}y - X\hat{\beta}), \hat{\varepsilon} = \hat{A}y - X\hat{\beta}$$

$$= \hat{V}\left(\hat{A}y - X\left(X^T\hat{V}X\right)^{-1}X^T\hat{V}\hat{A}y\right), \hat{\beta} = \left(X^T\hat{V}X\right)^{-1}X^T\hat{V}\hat{A}y$$

$$= (\hat{V} - \hat{V}X\left(X^T\hat{V}X\right)^{-1}X^T\hat{V})\hat{A}y$$

$$= (\hat{V} - \hat{P})\,\hat{A}y,\ P = \hat{V}X\left(X^T\hat{V}X\right)^{-1}X^T\hat{V}.$$

Therefore, $$t_i = \frac{(\hat{v}_i - \hat{p}_i)\hat{a}_i y}{\hat{\sigma}\sqrt{\hat{v}_{ii} - \hat{p}_{ii}}}$$

where, $\hat{p}_i$ and $\hat{v}_i$ are the i^{th} rows of matrix $\hat{P}$ and $\hat{V}$ respectively, $\hat{p}_{ii}$ and $\hat{v}_{ii}$ are as defined in Equation 6, and $\hat{a}_i$ is the i^{th} row of matrix $\hat{A}$.

The effect of the coefficients of spatial autoregressive and that of spatial autocorrelation error term are discernible from the simplified t_i. Both numerator and denominator of SC_i (Equation 8) depend on $\hat{\sigma}$; the numerator through t_i and the denominator through $\hat{\kappa}$. Thus, the SC_i score statistic (Dai et al, 2016) relies on non-robust variance that is affected by extreme values. In the same vein, high coefficient of spatially lagged dependent variable and residual term coupled with high contamination result in high variance and vice-versa. In the context of spatial statistics, directly adopting robust variance measures such as the median absolute

deviation (Huber and Rochettic, 1981) does not capture the true variance due to spatial dependence. We adopt a robust measure that takes into consideration the spatial dependence as described in Section 3.2.

3. The Proposed Adjacency Weighted Spatial Outlier Residual

The main aim of this work is to produce spatial outlier residual that incorporate the contiguity of spatial locations based on relative distance among the locations and difference of residuals of the corresponding spatial locations in large data. It also reasonably addresses the dominant masking and swamping effect in spatial outlier detection techniques. Prior to the development of the proposed method, let us first discuss two proposed weight functions to be incorporated in the proposed method.

3.1 Proposed weights

Two weights functions, namely the metric distance reciprocal weight (MDR) and the exponential weight (EW), are considered in this work to measure the relative spatial positions of attributes. The distance decay weights are discussed by Anselin and Rey (2014) as inverse distance weights. The weights are chosen to improve on controlling the problem of masking and swamping. We employ the geographically weighted (GW) concept of calibration to measure spatial adjacency of a location relative to another. Harris et al. (2014) noted that the GW can be extended to any statistical method. Most popular applications of the GW methods are Brunsdon et al. (1996) and Fotheringham et al. (2002). Skov-Petersen (2001) used a reciprocal like function and exponential function as distance decay to measure environmental indicators. Similarly, Von Luxburg (2004) pointed out that if d is a dissimilarity measure, such as distance, then $\exp\left(-\frac{d}{t}\right)$ is a similarity function for some parameter t. In the same way is $\frac{1}{1-d}$ for a scaled d. Also, Geurs and Van (2004) used the exponential function as a cost function in estimating the accessibility of opportunities at different locations with reference to a specific location, in which more distance opportunities provide diminishing influences and vice-versa.

3.1.1 Metric distance reciprocal weight (MDR)

In the metric distance reciprocal weight (MDR), two places that are far away would have a small reciprocal value, and vice versa. Hence, multiplying the attribute residual difference with the reciprocal of the metric distance contributes a proportionate effect of the difference based on distance. We define the metric

distance, $\|S_i - S_j\|$, as the Euclidean distance between the spatial locations S_i and S_j. The metric distance is re-scaled such that

$$m_{ij} = \frac{\|s_i - s_j\|}{max(\|s_i - s_j\|)}, \quad i,j = 1,2,\cdots,n, \tag{9}$$

Where, $max(\|S_i - S_j\|)$ is the largest distance between any two locations on the spatial data. Hence the metric distance reciprocal weight is defined as

$$mdr_{ij} = \frac{1}{m_{ij}+1}.$$

In matrix form, define a matrix such that the entries are the reciprocal of $mdr_{ij}, i,j = 1,2,\cdots,n$.

The metric distance reciprocal weight is given by,

$$\boldsymbol{W} = \{mdr_{ij}, i,j = 1,2,\ldots,n\} \tag{10}$$

3.1.2 Exponential weight (EW)

The exponential weight (EW) is defined as the exponent of the negative metric distance between pair of points. The exponential of negative metric distance decays with distance, where metric distance is as defined in Equation 9. The exponential weight is given by $exp(-m_{ij})$, where $i,j = 1,2,\ldots,n$. In matrix form, the exponential weights are given by Equation 11, with elements $exp(-m_{ij})$.

$$W = \{exp(-m_{ij}), i,j = 1,2,\ldots,n\} \tag{11}$$

3.2 The Proposed Adjacency Weighted Spatial Outlier Residual

Let W be the weight matrix defined in either of the Equation 10 and 11, where W is scaled so that each row elements sum up to unity. Let D be a matrix such that each element, d_{ij}, is the difference between the i^{th} residual term and the j^{th}residual term, i.e. $d_{ij} = \varepsilon_i - \varepsilon_j$, $i,j = 1,2,\cdots,n$. Let w_i be the i^{th} row of the weight matrix, W, and d_i be the i^{th} row of the residual difference matrix, D. Express w_i and d_i as column vectors:

$$w_i = (w_{i1}, w_{i2}, \ldots, w_{in})^T, d_i = (d_{i1}, d_{i2}, \ldots, d_{in})^T.$$

Then ε_i is projected such that

$$\varepsilon_i \rightarrow \varepsilon_{awsor} = w_i^T d_i \tag{12}$$

It follows from the normal distribution that if$\varepsilon_i \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$, and $d_{ij} = \varepsilon_i - \varepsilon_j$, then $d_{ij} \sim N\left(\mu_{\varepsilon_i} - \mu_{\varepsilon_j}, \sigma_{d_{ij}}^2\right)$, where $\sigma_{d_{ij}}^2 = \frac{\sigma_{\varepsilon_i}^2}{n} + \frac{\sigma_{\varepsilon_j}^2}{n} - 2cov(\varepsilon_i, \varepsilon_j)$. It then follows that

$$\varepsilon_{awsor} = w_i^T d_i \sim N(w_i^T \mu_{d_i}, w_i^T cov(d_i, d_j) w_i), \tag{13}$$

where, $w_i^T = [w_{i1}, w_{i2}, \cdots, w_{in}]$, $cov(d_i, d_j)$ is the covariance of difference between the i^{th} and the j^{th} residual difference. If $i = j$, then $cov(d_i, d_i) = var(d_i) = \sigma_{d_i}^2$.

Let $\hat{\mu}_{d_i} = w_i^T \hat{d}_\iota$ and $\hat{\sigma}_{d_i} = \sqrt{w_i^T cov(\hat{d}_i, \hat{d}_i) w_i}$.

The proposed adjacency weighted spatial outlier residual, ε_{awsor}, incorporates the spatial features of the residual term, $\hat{\varepsilon}_\iota$ into $\hat{d}_\iota$ using the spatial weight w_i, in line with the first law of geography; "everything is related to every other thing but closer things are more related than distant things" (Tobler, 1970). We use weights defined in Equation 10 and 11 to construct ε_{awsor} in Equation 13, which is subsequently used to construct t_{awsor}. We refer to t_{awsor} based on metric distance reciprocal as t_{MDR} and based on exponential weight as t_{EW}.

The t_i in Equation 6 is distributed as standard normal distribution. However, in the ε_{awsor}, both the $\hat{\mu}_{d_i}$ and $\hat{\sigma}_{d_i}$are weighted. We replace ε^* in Equation 6 with a robust residual measure that weights the residual difference of a location relative to spatial distance of reference location, ε_{awsor}, and $\hat{\sigma}$ with adaptive spatial scale measure, $\hat{\sigma}_{d_i}$. The t_{awsor} statistic is given by

$$t_{awsor} = \frac{\hat{\varepsilon}_{awsor}}{\hat{\sigma}_{d_i}\sqrt{1 - h_{ii}}}, \tag{14}$$

where, $h_{ii} = \frac{\hat{p}_{ii}}{\hat{v}_{ii}}$. $\hat{p}_{ii}$ and $\hat{v}_{ii}$ are as defined in Equation 6.

Since both $\hat{\mu}_{d_i}$ and $\hat{\sigma}_{d_i}$ are weighted, the statistic t_{awsor} follows the student t - disribution, $t_{awsor} \sim t_{(2n-2)-k-1}$, where n is the number of observations and k is the number of regressors in the model.

A spatial location S_i is declared outlier if its corresponding t_{awsor} is such that $t_{awsor} > t_{(2n-2)-k-1}$.

4. Simulation Study

In this section, we discuss a simulation study to assess the performance of our proposed methods, namely the t_{awsor} based on metric distance reciprocal and based on exponential weight, denoted as t_{MDR} and t_{EW}, respectively. The two proposed methods were compared to the existing score statistic of Dai et al. (2016), denoted as SC_i, The simulation formulation follows Dai et al. (2016) with modifications to capture low and high coefficients of spatial autoregression and larger sample sizes. The three models, specifically the SAR, SEM and GSM are considered where , $x_i \sim N(0,1)$, $\varepsilon \sim N(0,0.01)$, $\beta_0 = \mathbf{0}$ and $\beta_1 = \mathbf{1}$.

As per Lesage (1999) and Dai et al. (2016), the weight matrix for all the models is the same, i.e $W_1 = W_2 = W$. W is an $n \times n$ spatial contiguity matrix with entries unity where locations are neighbours, and 0 otherwise. The queen's contiguity matrix is obtained from the Pysal library of Rey and Anselin (2010). The row of W is scaled to sum equals to 1.

Regular square grids of sizes 400, 900, 2500 and 10000 of spatial fields and 4% and 10% contamination levels were considered. In each contamination level, say δ, expressed in percentage, the total contamination of size $p = n\delta$ is obtained, where n is the sample size. p random integers are generated and modulus, mod(n), of each random integer is taken to avoid integers greater than the sample size, n. The new set of generated integers are sorted in ascending order. Contamination are taken with uniform distribution within the minimum and maximum values of generated y in the SAR, SEM and GSM models.

In each of the model, divide the sorted random integers into two groups, with the first group containing the 1^{st} to $(p/2)^{th}$ terms and the second group containing the $(p/2+1)^{th}$ to the p^{th} terms. Uniform random variables of size $p/2$, from 0 to the maximum value of y, are assigned to the first group, and the other half are assigned the uniform random variables from the minimum value of y to 0. As such, all the values in the first half group are positive while that of the second half are all negative. In this way, low values contamination are embedded among high spatial data values and vice-versa. However, there is a possibility that some of the contamination would be similar to the attribute of its neighborhood. The simulation is repeated 10000 times for sample of sizes 400 and 900; 1000 times for sample of size 2500 and 100 times for sample of size 10000.

The coefficients of spatially lagged dependent variable, ρ, and that of the coefficient on the spatially correlated errors, λ, that are used for simulation of the four models are as follow:

In the mixed regressive-spatial autoregressive model (SAR) (model (4), $\rho = 0.5,\ 0.9$.

In spatial autoregressive error model (SEM) (model (5)), $\lambda = 0.5,\ 0.7$.

In the general spatial general model (GSM) (model (1)), $\rho = 0.3,\ \lambda = 0.5$, and $\rho = 0.9,\ \lambda = 0.7$. Python programming language, version 3.7.6 with IDE Spyder 4.1.3 is used for the analysis of the simulated and real dataset.

The performances of the three methods are evaluated based on the percentage of correct detection of spatial outliers and percentage of swamping effect. A good method is the one that has the highest percentage of correct detection of outliers and smaller percentage of swamping. Swamping refers to inliers incorrectly declared as outliers. The results of the study are exhibited in Tables (1-3) and Figure 1.

SAR model: Let us first focus our attention to Table 1, for SAR model. The results in Table 1 show that the three methods are equally good in terms of having closed values of percentage of correct detection of spatial outliers and swamping effect irrespective of ρ values, percentage of contaminations and sample size, except at $\rho = 0.5$, 4% contamination, where the SC_i has slightly lower value of swamping effect compared to the t_{MDR} and t_{EW}. It can also be observed that the percentage of correct detection of outliers for the three methods is higher at 4% contamination levels than at 10%, irrespective of ρ values and sample size.

Table 1. Percentage of Correct Detection of outliers and swamping, SAR model

			Correct Detection (%)			Swamping (%)		
ρ	%cont	n	t_{MDR}	t_{EW}	SC_i	t_{MDR}	t_{EW}	SC_i
0.5	4%	400	74.30	74.27	74.91	0.89	0.95	0.77
		900	73.30	73.88	74.02	0.35	0.40	0.22
		2500	74.36	73.54	74.83	0.11	0.16	0.05
		10000	73.24	72.06	73.56	0.00	0.22	0.00
	10%	400	57.63	58.28	58.97	0.05	0.02	0.04
		900	58.15	58.58	59.38	0.02	0.02	0.01
		2500	58.70	58.39	59.69	0.00	0.03	0.00
		10000	59.58	59.23	60.20	0.00	0.00	0.00
0.9	4%	400	72.38	72.68	72.82	13.52	13.76	13.52
		900	72.55	72.66	72.92	8.89	9.29	8.70
		2500	73.28	72.61	73.42	6.34	6.48	6.12
		10000	71.60	73.91	73.91	4.67	4.37	4.37

	10%	400	56.21	56.36	56.58	2.86	3.10	3.04
		900	55.81	56.14	55.38	2.44	2.56	2.45
		2500	56.64	56.50	57.01	1.74	1.89	1.73
		10000	57.00	56.27	57.33	2.48	1.55	2.41

SEM model: We will now discuss the results of SEM model presented in Table 2. Similar to SAR model, the three methods have reasonably close values of percentage of correct detection of outliers. The percentage of correct detection of outliers for the three methods is higher at 4% contamination levels than at 10%, irrespective of λ values and sample size. It is interesting to note that there is no swamping effect for both t_{MDR} and t_{EW} methods. Nonetheless, the SC_i shows swamping problem where lowest swamping is seen at $\lambda = 0.5$ with 10% contamination, followed by at $\lambda = 0.7$ with 10% contamination, at $\lambda = 0.5$ with 4 % contamination and the highest being at $\lambda = 0.7$ with 4% contamination. The results seem to suggest that t_{MDR} and t_{EW} methods are comparable, and they outperform the SC_i.

Table 2. Percentage of Correct Detection of outliers and swamping, SEM model

			Correct Detection (%)			Swamping (%)		
λ	%cont	n	t_{MDR}	t_{EW}	SC_i	t_{MDR}	t_{EW}	SC_i
0.5	4%	400	76.63	75.63	75.42	0.00	0.00	9.54
		900	74.10	75.45	74.20	0.00	0.00	6.71
		2500	75.38	74.97	74.83	0.00	0.00	4.46
		10000	75.26	75.94	74.66	0.00	0.00	4.73
	10%	400	59.85	60.48	58.57	0.00	0.00	1.14
		900	59.88	60.14	58.28	0.00	0.00	0.88
		2500	61.09	60.81	59.07	0.00	0.00	0.55
		10000	60.68	61.43	59.45	0.00	0.00	0.56
0.7	4%	400	76.60	74.19	75.12	0.00	0.00	25.92
		900	74.16	75.83	74.78	0.00	0.00	19.71
		2500	74.86	74.41	74.08	0.00	0.00	17.29
		10000	74.80	74.53	74.53	0.00	0.00	17.76
	10%	400	59.62	59.41	58.82	0.00	0.00	6.03
		900	60.61	60.80	59.38	0.00	0.00	5.18
		2500	61.05	60.97	59.19	0.00	0.00	4.67
		10000	60.91	60.50	59.27	0.00	0.00	4.44

GSM model: Finally, we discuss the results obtained from GSM model exhibited in Table 3. Several interesting points emerge from these results. It can be seen that the percentage of correct detection of outliers are reasonably closed for the three

methods at $\rho = 0.3, \lambda = 0.5$ irrespective of the percentage of contamination and size of samples. However, the performance of SC_i is not encouraging since it has swamping problems. On the other hand, both t_{MDR} and t_{EW} shows no swamping problems. At $\rho = 0.9, \lambda = 0.7$, the SC_i is slightly better than t_{MDR} and t_{EW} in terms of having slightly higher percentage of correct detection of outliers. Nonetheless, its performance still inferior than both methods evidenced by having large percentage of swamping especially at 4% contamination, $\rho = 0.9, \lambda = 0.7$. The t_{MDR} and t_{EW} are comparable with respect to the values of percentage of correct detection of outliers and percentage of swamping.

We further investigated the performances of our proposed methods, t_{MDR} and t_{EW} by computing their execution times and compared with the SC_i. Due to space limitation, we only report the execution time, in seconds, of the three methods as sample size, n, increases for the GSM model (see Figure 1). However, the results for SEM and SAR models are consistent. It is very interesting to observe from Figure 1 that the execution time for t_{MDR} and t_{EW} are indistinguishable and much shorter than that of the SC_i. The execution time of SC_i increases drastically as sample size increases.

Table3. Percentage of Correct Detection of outliers and swamping, GSM model

			Accurate Detection (%)			Swamping (%)		
ρ, λ	%cont	n	t_{MDR}	t_{EW}	SC_i	t_{MDR}	t_{EW}	SC_i
0.3, 0.5	4%	400	75.31	74.59	75.32	0.00	0.00	22.45
		900	73.28	75.21	74.89	0.00	0.00	17.14
		2500	75.31	74.85	75.75	0.00	0.00	13.09
		10000	75.01	74.98	75.62	0.00	0.00	15.00
	10%	400	59.50	59.04	60.10	0.00	0.00	5.16
		900	60.00	59.06	59.98	0.00	0.00	4.16
		2500	60.58	60.49	60.59	0.00	0.00	3.45
		10000	60.28	60.31	60.42	0.00	0.00	4.05
0.9, 0.7	4%	400	73.14	72.98	76.53	12.98	12.29	51.92
		900	73.47	73.18	76.74	8.63	7.55	48.29
		2500	72.65	72.87	76.18	6.41	6.69	47.26
		10000	72.33	72.75	76.25	7.19	6.56	47.93
	10%	400	56.85	56.10	62.43	2.67	2.64	29.83
		900	56.76	55.92	62.04	2.50	2.53	28.25
		2500	57.14	56.89	63.10	2.15	2.36	27.31
		10000	56.72	56.22	61.88	1.80	2.12	27.96

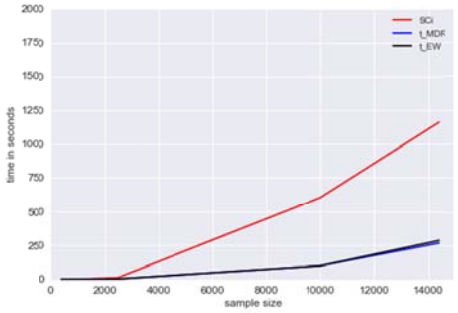

Figure 1. GSM Execution time in seconds.

5. The US counties COVID-19 Data

The available COVID-19 data for 2841 of 3142 counties in the United States are used to illustrate the performances of our t_{MDR}, t_{EW}, and SC_i in real life applications. The data are obtained from the Kaggle website (https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data). The dependent variable is the total number of cases in each of the counties as at 4/12/2020. The independent variables are population density (pd), percentage of female population (female), life expectancy (le), percentage fair/poor health (fph), percentage of adult with obesity (obesity), percentage of population in rural area (rural), percentage of smokers (smokers), percentage of physical inactivity (inactivity), exercise opportunity (eo), percentage of people that are 65 years and above (ab65), percentage Black race (black), percentage Asians (asians), and percentage population that are Hawaiians (hawaiians). The result of Moran's I spatial autocorrelation test on the dependent variable and the residual are significant (p-value <<0.01). Therefore, there is spatial autocorrelation in the data. In other words, there are clusters of cases. The data were then applied to the SAR, SEM and GSM models. The findings show that the data best fitted to the SEM model evident by having the smallest value of AIC (57437) and highest value of R^2 (0.8901). Subsequently, we apply the t_{MDR}, t_{EW}, and SC_i methods to the data fitted with SEM model in order to identify spatial outliers. It is very interesting to note that the t_{MDR} and t_{EW} can identify eighty-six counties as outliers. On the other hand, the SC_i identifies eighty-nine counties as outliers. The extra three counties detected by SC_ias outliers can be treated as swamping effect. It's worth noting that, in the simulation results of SEM model with small coefficient of residual autocorrelation ($\lambda = 0.4$), for sample size of 2500 (closer to the sample size of 2841for this data set), all the three methods able to correctly detect almost the same percentage of spatial outliers with no swamping effects for t_{MDR} and t_{EW}, but SC_i shows certain percentage of swamping. These results are consistent

with the results obtained from the simulation study where for ($\lambda = 0.4$, n=2500), the SC_i suffers from the swamping effect.

Conclusions

In this article, we propose a technique that uses metric distance reciprocal weight (MDR) and exponential weight (EW), to detect spatial outliers in large data sets. The technique calibrates the difference between residual of a spatial location and the residuals of other locations, according to relative distance, to construct spatially weighted residuals. The results of simulation study and real data set signify that the percentage of correct detection of outliers and percentage of swamping for t_{MDR} , t_{EW} and SC_i are fairly closed to each other for SAR model. For SEM model, the t_{MDR} and t_{EW} are comparable and both outperform the SC_i in the detection of spatial outliers without any swamping effects. Moreover, our proposed t_{MDR} and t_{EW} methods also have better performance compared to SC_i for GSM model where the swamping effects for both methods are much lower than the SC_i method. Furthermore, both t_{MDR} and t_{EW} methods are very appealing because their computational running time are much faster than the SC_i method.

REFERENCES

[1] **Aggarwal, C. C. (2013),** ***Spatial Outlier Detection, in Outlier Analysis;*** *Springer,* pp. 345-368;

[2] **Anselin, L. (1994),** ***Exploratory Spatial Data Analysis and Geographic Information Systems, New Tools for Spatial Analysis.*** *Eurostat Luxembourg*, pp. 45-54.

[3] **Anselin, L. (1995),** ***Local Indicators of Spatial Association LISA;*** Geographical analysis, *Wiley Online Library,* pp. 93-115;

[4] **Geurs, K. T. and Van Wee, B. (2004),** ***Accessibility Evaluation of Land-Use and Transport Strategies: Review and Research Directions***; *Journal of Transport geography,* vol. 12, 2, *Elsevier,* pp. 127—140;

[5] **Hadi, A. S. and Imon, A. H. M. R. (2018),** ***Identification of Multiple Outliers in Spatial Data;*** *International Journal of Statistical Sciences.*, vol. 16, pp. 87—96;

[6] **Haining R. P. and Haining R. (2003),** ***Spatial Data Analysis: Theory and Practice;*** *Cambridge University Press;*

[7] **Halas, M., Klapka, P. and Kladivo, P. (2014),** ***Distance-decay Functions for Daily Travel-to-work Flows***; *Journal of Transport Geography*, vol. 35, *Elsevier,* pp. 107—119;

[8] **Harris, P., Brunsdon, C., Charlton, M., Juggins, S. and Clarke, A. (2014),** ***Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods;*** *Mathematical Geosciences*, vol. 46, 2, *Springer,* 1—31;

[9] **Hawkins, D. M. (1980),** ***Identification of Outliers***, *Biometrical Journal,* vol. 29, 2, *Wiley Online Library,* 189—198;

[10] **Kou, Y. and Lu, C. (2008),** ***Outlier Detection*****;** *Spatial, Encyclopedia of GIS, Springer,* 1539—1546;

[11] **Liu, X., Lu, C. and Chen, F. (2010),** ***Spatial Outlier Detection: Random Walk-Based Approaches;*** Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 370—379;
[12] **Lu, C., Chen, D. and Kou,Y. (2003),** ***Detecting Spatial Outliers with Multiple Attributes;*** Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence, IEEE, 122—128;
[13] **Rey, S. J. and Anselin, L. (2010), PySAL:** ***A Python Library of Spatial Analytical Methods;*** Handbook of applied spatial analysis, *Springer,* pp. 175—193;
[14] **Shekhar, S., Lu, C. and Zhang, P. (2002),** ***Detecting Graph-Based Spatial Outliers;*** *Intelligent Data Analysis*, vol. 6, 5, *IOS Press,* 451—468;
[15] **Singh, A. K. and Lalitha, S. (2018),** ***A Novel Spatial Outlier Detection Technique;*** Communications in Statistics-Theory and Methods, vol. 47, 1, *Taylor & Francis,* 247—257;
[16] **Skov-Petersen, H. (2001),** ***Estimation of Distance-decay Parameters: GIS-Based Indicators of Recreational Accessibility***; *ScanGIS,* 237—258;
[17] **Tobler, W. R. (1970),** ***A Computer Movie Simulating Urban Growth in the Detroit Region;*** *Economic geography,* vol. 46, supl.T*aylor & Francis,* 234--240;
[18] **Anselin, L. and Rey, S. J. (2014),** ***Modern Spatial Econometrics in Practice: A Guide to Geoda, Geodaspace and Pysal;*** *GeoDa Press LLC;*
[19] **Von Luxburg, U. (2004),** ***Statistical Learning with Similarity and Dissimilarity Functions;*** Technische University Berlin Berlin, Germany;
[20] **Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2003),** ***Geographically Weighted Regression: The Analysis of Spatially Varying Relationships****; John Wiley & Sons;*
[21] **Shi, L. and Chen, G. (2009),** ***Influence Measures for General Linear Models with Correlated Errors;*** *The American Statistician*, vol 63, 1, *Taylor & Francis,* 40-42;
[22] **Dai, X., Jin, L., Jin, A. and Shi, L. (2016),** ***Outlier Detection and Accommodation in General Spatial Models;*** Statistical Methods & Applications, vol. 25, 3, *Springer,* 453-475;
[23] **Anselin, L. (1988),** ***Spatial Econometrics: Methods and Models****; Kluwer Publishing Academics;*
[24] **LeSage, J. P. (1999),** ***The Theory and Practice of Spatial Econometrics;*** University of Toledo. Toledo, Ohio, vol. 28, 11, Citeseer;
[25] **Brunsdon, C., Fotheringham, A. S. and Charlton, M. E. (1993),** ***Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity.***Geographical analysis, vol. 28, 4, *Wiley Online Library*, pp. 281-298;
[26] **Huber P. J. and Ronchetti, E. (1981),** ***Robust Statistics;*** vol. 1, 1, J*ohn Wiley & sons,* New York;
[27] **Gaspard, G., Kim, G. and Chun, Y. (2019),** ***Residual Spatial Autocorrelation in Macroecological and Biogeographical Modeling: A Review;*** Journal of Ecology and Environment, vol. 43, 1, *Springer.*