UNIVERSITI PUTRA MALAYSIA

ROBUST SPATIAL DIAGNOSTIC METHOD AND PARAMETER ESTIMATION FOR SPATIAL BIG DATA REGRESSION MODEL

MOHAMMED BABA ALI

IPM 2022 6

ROBUST SPATIAL DIAGNOSTIC METHOD AND PARAMETER ESTIMATION FOR SPATIAL BIG DATA REGRESSION MODEL

By

MOHAMMED BABA ALI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Doctor of Philosophy

July 2022

DEDICATION

I dedicate this thesis to my late mother: Hajja khadija, my dad: Alhaji Mohammed Baba Gana Goni and my late daughter Khadija.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

ROBUST SPATIAL DIAGNOSTIC METHOD AND PARAMETER ESTIMATION FOR SPATIAL BIG DATA REGRESSION MODEL

By

MOHAMMED BABA ALI

July 2022

Chair: Professor Habshah Midi, PhD
Institute: Mathematical Research

The existing spatial data compression method, namely the Adaptive Spatial Compression Clustering (ASDC) is a very potent method of compressing big data. However, the presence of global outliers in the spatial data affects the formation of spatial dispersion function which subsequently affects the outcome of the spectral clustering; this, in effect, affects spatial contiguity. Hence, a new robust spatial compression technique, which we call Outlier Resistant Adaptive Spatial Clustering (ORASDC) is proposed. Simulation results of synthetic spatial fields and real data application reveal that the proposed method is worthwhile in treating the effect of outliers with over 99% region of similarity retained and over 90% of data similarity maintained. Further research may be carried out to improving the processing speed of the ORASDC and to determining the optimum number of clusters that correspond to a specific data size.

The score statistics (Sc_i) is formulated to identify spatial outliers in big data. Nonetheless, the method not only suffers from masking and swamping effects, but also takes long computational running time. To rectify this problem, a new diag-

nostic measure that adopts location adjacency to construct spatial weights, metric distance reciprocal (MDR) and exponential weight (EW), are developed. Difference between spatial residuals are calibrated to incorporate adjacency effect into spatial outlier residual. Results of simulations in large sample sizes have shown remarkable performance of the proposed methods where both diagnostics measures successfully detect spatial outliers with minimum swamping effect. Applications of our methods to real data have also shown good performance.

This thesis also concerned on the establishment of diagnostic measures for the identification of spatial influential observations (IOs), which are outliers in the x and y directions of spatial regression models. Some of the classical techniques of identification of IOs have been adapted to spatial models. Nonetheless, those adapted methods fail to correctly identify the IOs and show high swamping and masking effects. Thus, we propose a new measure of spatial studentized prediction residuals that incorporate spatial information on the dependent variable and residual. To the best of our knowledge, no research is done on the classification of spatial observations into regular observations, vertical outliers, good and bad leverage points. Hence, the $ISR_s - P_{osi}$ and $ESR_s - P_{osi}$ plots are established to close the gap in the literature. The results signify that the $ESR_s - P_{osi}$ plot, followed by the $ISR_s - P_{osi}$ plot were very successful in classifying observations into the correct groups. The numerical examples and simulation study have shown that the proposed methods possess almost 100% accurate detection and 0% swamping, against their competitors that have lower detection rates and higher swamping rates.

Outliers in spatial applications usually keep vital information about the model; a situation that calls for method that is effective in accommodating the spatial outliers in a special way. Variance Shift Outlier Model (VSOM) in the classical regression is promising in keeping such observations in the model by downweighting their effect in the model. To date, no research has been done to obtain spatial representation of VSOM. To fill the gap in the literature, we formulated the VSOM in the spatial regression model which we call Spatial Variance Shift Outlier Model (SVSOM) using the Residual Maximum Likelihood (REML). Weights based on the detected outliers are used to accommodate the spatial outliers via revised model with the help of the SVSOM. The results of simulation study and real data set indicate that our proposed method has significant improvement in parameter estimation and outlier accommodation.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

KAEDAH TEGUH BERDIAGNOSTIK RUANG DAN PENGANGGARAN PARAMETER BAGI MODEL REGRESI DATA RAYA RUANG

Oleh

MOHAMMED BABA ALI

Julai 2022

Pengerusi: Professor Habshah Midi, PhD
Institut : Penyelidikan Matematik

Kaedah pemampatan data ruang sedia ada, iaitu Adaptive Spatial Compression Clustering (ASDC) merupakan kaedah yang sangat sesuai untuk memampatkan data raya. Walau bagaimanapun, kehadiran titik terpencil global dalam data ruang menjejaskan pembentukan fungsi penyebaran ruang yang seterusnya menjejaskan hasil pengelompokan spektrum; ini, sebenarnya, menjejaskan keterkaitan ruang. Seterusnya, teknik pemampatan ruang baharu yang teguh, yang kami panggil sebagai Outlier Resistant Adaptive Spatial Clustering (ORASDC) telah dicadangkan. Hasil simulasi medan ruang sintetik dan aplikasi data sebenar mendedahkan bahawa kaedah yang dicadangkan adalah berkesan dalam merawat kesan titik terpencil dengan lebih 99% kawasan persamaan dikekalkan dan lebih 90% persamaan data dikekalkan. Kajian lanjut boleh dijalankan untuk meningkatkan kelajuan pemprosesan ORASDC dan untuk menentukan bilangan optimum kelompok yang sepadan dengan saiz data tertentu.

Statistik skor ((Sc_i) dirumuskan untuk mengenal pasti titik terpencil ruang dalam data raya. Walau bagaimanapun, kaedah ini bukan sahaja mengalami kesan topeng dan paya, tetapi juga mengambil masa pengiraan yang lama. Untuk mengatasi masalah ini, langkah diagnostik baharu yang menggunakan lokasi bersebelahan untuk membina pemberat ruang, iaitu timbal balik jarak metrik (MDR) dan berat eksponen (EW), dibangunkan. Perbezaan antara ralat ruang ditentukur untuk memasukkan kesan bersebelahan ke dalam ralat titik terpencil ruang. Keputusan simulasi bagi saiz sampel yang besar telah menunjukkan prestasi yang luar biasa bagi kaedah yang dicadangkan dimana kedua-dua ukuran diagnostik sangat berjaya dalam mengesan titik terpencil ruang dengan kesan paya yang minimum. Aplikasi kaedah kami kepada data sebenar juga menunjukkan prestasi yang baik.

Tesis ini juga berkaitan dengan pembinaan ukuran diagnostik untuk mengenal pasti pemerhatian berpengaruh ruang (IOs), yang merupakan titik terpencil dalam arah x dan y bagi model regresi ruang. Beberapa teknik klasik pengecaman IOs telah disesuaikan dengan model ruang. Walau bagaimanapun, kaedah yang disesuaikan gagal mengenal pasti IOs dengan betul dan menunjukkan kesan paya dan topeng yang tinggi. Oleh itu, kami mencadangkan satu ukuran baharu iaitu ralat ramalan pelajar ruang yang menggabungkan maklumat ruang pada pembolehubah bersandar dan ralat. Sepanjang pengetahuan kami, tiada penyelidikan dilakukan untuk mengklasifikasi pemerhatian ruang ke dalam pemerhatian biasa, titik terpencil menegak, titik tuasan yang baik dan buruk. Seterusnya, plot $ISR_s - P_{osi}$ dan plot $ESR_s - P_{osi}$ dibangunkan untuk menutup jurang dalam kesusasteraan. Hasilnya menunjukkan bahawa plot $ESR_s - P_{osi}$, diikuti oleh plot $ISR_s - P_{osi}$ sangat berjaya dalam mengklasifikasikan pemerhatian ke dalam kumpulan yang betul. Contoh numerasi dan kajian simulasi telah menunjukkan hampir 100% pengesanan tepat dan 0% paya, berbanding pesaing mereka yang mempunyai kadar pengesanan yang lebih rendah dan kadar paya yang lebih tinggi

Titik terpencil dalam aplikasi ruang biasanya menyimpan maklumat penting mengenai model; keadaan yang memerlukan kaedah yang berkesan dalam menampung titik terpencil ruang dengan cara yang tersendiri. Variance Shift Outlier Model (VSOM) dalam regresi klasik menjanjikan dalam mengekalkan pemerhatian tersebut dalam model tersebut dengan menurunkan kesan beratnya. Sehingga kini, tiada penyelidikan telah dilakukan untuk mendapatkan perwakilan ruang VSOM. Untuk mengisi jurang dalam kesusasteraan, kami merumuskan VSOM dalam model regresi ruang yang kami panggil Spatial Variance Shift Outlier Model (SVSOM) menggunakan kaedah Reja Kebolehjadian Maksimum (REML). Berat berdasarkan titik terpencil yang dikesan digunakan untuk menampung titik terpencil ruang melalui model yang disemak dengan bantuan SVSOM. Hasil kajian simulasi dan set data sebenar menunjukkan bahawa kaedah yang kami cadangkan mempunyai peningkatan yang bererti dalam anggaran parameter dan penampungan titik terpecil.

ACKNOWLEDGEMENTS

Gratitude and praises are due to Almighty Allah for bringing me this far in my academic career and scholarly pursuits. I also send salutations upon His last messenger, prophet Muhammad (SAW).

Taking on an academic journey like this does not happen on its own or in isolation. Throughout my journey through the School of Graduate Studies, I have been guided by wonderful people to realize my dream. I must acknowledge there support and encouragement without which this journey would not have been a success.

To my supervisor, Prof. Habshah Midi, I wish to express my immeasurable appreciation. You have provided me with immeasurable and indepth knowledge and guidance which I always feel is a singular privilege. I will be eternally thankful for your excellent mentoring, constructive criticism, impartial ideas, and perceptive supervision, which have guided me through the various stages of this research. Your guidance has provided me with a wealth of knowledge and research skills. I would like to acknowledge the advice, support and guidance of my other committee members, late Assoc. Prof. Mohd Bakri Adam and Dr. Nur Haizum Bint Abd Rahman. Your suggestions provided me with remarkable insights, objective contributions to make this study a success. I am grateful for your thoughtful comments and recommendations.

I would also like to acknowledge the Institute and staff of Institute for Mathematical Research (INSPEM) for their tremendous support during my study for their administrative role and logistic support. They really had played significant role through my academic pursuit. Thank you very much.

My special prayers goes to my mother, late Hajja Khadija for her prayers and guidance during the pursuit of the scholarship. May almighty Allah grant her highest position in Jannatul Firdaus. To my father, Alhaji Babagana Goni, I am thankful for your support, guidance and parental advice. May Allah reward you with Jannatul Firdaus. Special thanks goes to my wife Hajja Amne for being with me throughout this struggle, and also to my beloved children Mohammed Kabir, late Khadija and Zahrah, I really appreciate your support and patience especially during the trying times of the COVID19 pandemic. May almighty Allah bless you all abundantly. I will not forget the prayers, support and encouragement I received from my brothers and sisters while undertaking my studies, thanks a lot. Thank you to friends, well wishers at home and here in Malaysia.

Finally, I wish to acknowledge and show my gratitude to my university, Abubakar Tafawa Balewa University Bauchi, and TetFund, without which this study would not have been possible, for choosing me for this programme.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Phylosophy. The members of the Supervisory Committee were as follows:

Habshah binti Midi, PhD
Professor
Faculty of Science and Institute For Mathematical Research
Universiti Putra Malaysia
(Chairperson)

Nur Haizum binti Abd Rahman, PhD
Senior Lecturer
Faculty of Science and Institute For Mathematical Research
Universiti Putra Malaysia
(Member)

Mohd Bakri Adam, PhD
Associate Professor
Faculty of Science and Institute For Mathematical Research
Universiti Putra Malaysia
(Member)

ZALILAH MOH'D SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 8 September 2022

Declaration by Members of Supervisory Committee

This is to confirm that:
- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of Chairman of Supervisory Committee: Professor Dr. Habshah binti Midi

Signature: ______________________________
Name of Member of Supervisory Committee: Dr. Nur Haizum binti Abd Rahman

Signature: ______________________________
Name of Member of Supervisory Committee: Associate Professsor Dr. Moh'd Bakri Adam

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

AD	Accuarte Detection Rate
AIC	Akaike Information Criteria
ASDC	Adaptive Spatial Dispersion Clustering
AWSOR	Adjacency Weighted Spatial Outlier Residual
BACON	Blocked Adaptive Computationally Efficient Outlier Nominators
BAU	Basic Areal Units
BIC	Bayesian Information Criteria
BLP	Bad Leverage Point
BP	Bipartite Graph
CD	Cook's Distance
CD_s	Spatial Cook's Distance
COVID19	Corona Virus 2019
DGR	Data Generating Process
DRGP	Diagnostic Robust Generalized Potential
DS	Data Similarity
ESR	External Studentized Residual
ESR_s	Spatial External Studentized Residual
EW	Exponential Weight
FAR	First Order Spatial Autoregressive Model
GIS	Geographic Information System
GLP	Good Leverage Point
GSM	General Spatial Autoregressive Model
GW	Geograhically Weighted
ID	Identification Number
IDE	Integrated Development Environment
IID	Independently Identically Distributed
IO	Influential Observation
ISR	Internal Studentized Residual
ISR_s	Spatial Internal Studentized Residual
Knn	K Nearest Neighbours

LISA	Local Indicator of Spatial Autocorrelation
LQ	Local Quotient
MAD	Mean Absolute Deviation
Med	Median
MDR	Metric Distance Reciprocal
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MSOM	Mean Shift Outlier Model
OLS	Ordinary Least Squares
ORASDC	Outlier Resistant Spatial Dispersion Clustering
P_{osi}	Spatial Potential
REML	Residual/Restricted Maximum Likelihood
RS	Regiona; Similarity
RSDP	Robust Spatial Diagnostic Plot
RW-BP	Random Walk on Biparte Autocorrelation
RW-EC	Random Walk on Exhaustive Combination
SAR	Spatial Autoregressive Model
SC_i	Score Statistic
SEM	Spatial Autoregressive Error Model
SVSVOM	Spatial Variance Shift Outlier Model
VIOM	Variance Inflation Outlier Model
VSOM	Variance Shift Outlier Model

CHAPTER 1

INTRODUCTION

1.1 Background of the study

Recent data explosion has awaken researchers to the responsibility of developing solutions to the embarrassment of not being able to explore the valuable information in massive data due to lack of appropriate handling tools. The massive data sets that result from expansion in internet activities and computerization of human lives pose a great challenge to traditional methods of data collection, storage, processing, analysis, presentation and interpretation. Hence, a demand for researchers to thrive for robust techniques that can properly address this trend of obstacles. Large volume, vast variety and high velocity are the main features of big data that pose the challenge of analysis. Large volume problem remains of great importance to researchers because it made the computational cost of most statistical methods in practice too expensive (Zhang et al. (2018); Torrecilla and Romo (2018); Jayasankar et al. (2021)).

Building powerful computing facilities is offered by computer engineering as a solution to big data problem. Notable examples of such solutions are the supercomputers and cloud computing. However, the exponential growth of big data in volume still poses a challenge to the computational capacity for the so said high performance computers.

In statistical applications with fixed computational capacity, analytical and computational methods, *computational capacity constrained statistical methods* adapt these constraints to overcome the problem of big data. *Divide-and-conquer*, for example, divides large data sets into smaller pieces and conduct statistical analysis on each of the smaller manageable pieces. Final results for the full data set is determined by combining results of the smaller pieces of the data set. One great advantage of this method is the significant reduction in time of computation on a distributed computing environment (Härdle et al. (2018)). A major problem with the *divide-and-conquer* method is how to come up with scheme to combine the smaller pieces of results to form the final estimate that satisfy good statistical properties. One of the common assumptions about the distribution of observations in statistics is independent and identically distributed (IID). The IID random samples are used to build models and estimate model parameters. However, there are situations in which data that are close together, either in time or space, are correlated and as such the notion of independence is violated. Time series models, also known as temporal models, are based on identically distributed observations that are time dependent, usually at equal time interval. These data have a unidirectional flow of time that allows the construction of the temporal models (Cressie (1993); Darmofal (2006); Pole et al. (2018)).

In the same vein, Lansley et al. (2019) have pointed out that the problem of massive data have been a long standing problem in the field of spatial/spatio-temporal and have not been properly addressed.

In contrast to the temporal data that is unidirectional in time, spatial data have contextual attribute that is multidirectional in space associated with behavioural attribute. Behavioural attribute is the measurement of interest taken on object, while contextual attribute refers to the location at which the behavioural attribute is measured. The contextual attributes are expressed in terms of coordinates, or using granularity of regions in space, for example, county, zip code and so on (Aggarwal (2015)). Geographic information systems (GIS) support geocoding or address matching which allows address to be converted to coordinates (LeSage and Pace (2009)).

Spatial dependence connotes a scenario where values observed at one location depend on the values of adjacent observations at nearby locations. This adjacent locations can be regions that share borders with each other. Spatial dependence is the degree of spatial autocorrelation between independently measured values observed in geographical spaces (Kitchin and Thrift (2009)).

Spatial autocorrelation is a systematic pattern in attribute values that are recorded in locations on a map (Haining (2001)). Attribute values in one location that are associated with values at neighbouring locations indicate presence of autocorrelation. Positive autocorrelation indicates similar values clustered together. Negative autocorrelation indicates low attribute value in the neighbourhood of high attribute values and vice-versa. One of the measures employed in measuring the spatial autocorrelation is the Moran's I (Anselin (1995)) which was proposed by (Moran (1950)). Moran's I is used to test the hypothesis whether there is no spatial autocorrelation against its opposite.

Spatial data operate according to the first law of geography (Tobler (1970)). The law states that :*"every thing is related to everything else but nearby objects are more related than distance things."* Common fields of real applications (Aggarwal (2015); Zhang et al. (2022)) include meteorological data, traffic data, Earth science data, disease outbreak data, medical diagnostic data, demographic data, among others. Cressie (1993) pointed out that disciplines that work with data that are collected from different spatial locations need established models that indicate when there is dependence between measurements at different locations.

Spatial regression models are designed in such a way that they incorporate spatial relationship within the model (Haines and Thiart (2022)). This is desired to account for spatial relationship in order to generate meaningful inferences about a process under study, which would have otherwise been neglected by classical regression (Anselin (1988)). Researchers believe that the independent variables, X, do not always explain the dependent variable entirely, and perhaps the nearby observations do usually have effect in explaining their nearby observations. Some of the effects of not taking into cognizance the spatial effect include violation of regression assumptions such as independence of residuals (due to autocorrelation in residual). This, in effect, results in biased estimates of coefficients which inflate variance. The outcome of such effects is incorrect inference, which results in misleading conclusions (Anselin (1988); LeSage (1999); Haining and Haining (2003)).

The effects of outliers and influential observations have been subject of discussion for centuries among researchers in various fields of applications, due to their influence on model building and parameter estimation. Ben-Gal (2005) noted that outliers are aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. Influential observation, individually or together with other observation have large impact on the calculated values of estimates that in the case for most of the other observation (Belsley et al. (1980)).

Spatial outlier has peculiar characteristics; that is dependent on its nearby observations. They have extreme values relative to set of observations in their neighbourhood on the map (Haining and Haining (2003)).

1.2 Some Important Definitions

1.2.1 Big Data

Big Data, according to Chen and Zhang (2014), is a collection of very huge data sets with a great diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms. In details, Savitz (2012) defined Big Data as 'high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization'. Though it comes with a lot of opportunities, there are challenges such as data capture, storage, searching, sharing, analysis, and visualisation, which are all demanding tasks in big data.

1.2.2 Characteristics of Big Bata

The four Vs, (volume, velocity, variety and veracity) are widely used to describe the characteristics of Big Data (Chen and Zhang (2014); De Mauro et al. (2015); Davalos (2017); Härdle et al. (2018); Habeeb et al. (2019)), even though most researchers give emphasis to the first three of the aforementioned characteristics.

The amount of the data sets that need to be evaluated and processed, which are today frequently larger than terabytes and petabytes, is referred to as data volume. The sheer volume of data necessitates processing solutions that are separate from standard storage and processing capabilities. To put it in other words, the data sets in Big Data are too enormous to be processed by a standard laptop or desktop CPU.

The rate at which data is generated is referred to as velocity. Because high-velocity data is created at such a rapid rate, it necessitates the requirement for use of separate (distributed) processing procedures.

Big Data is made even bigger by its diversity. Big Data can come from a variety of places and can be classified into one of three categories: structured, semi-structured, or unstructured data. The diversity of data kinds frequently necessitates specialised processing capabilities and algorithms.

The quality of the data being studied is referred to as veracity. Many records in high veracity data are beneficial to evaluate and add meaningfully to the overall results. Low veracity data, on the other hand, comprises a large amount of information that have no apparent value.

1.3 Spatial dependence

Spatial dependence, according to LeSage and Pace (2009); Basile et al. (2014), typically reflects a situation where values observed at one location or region depend on the values of neighboring observations. Usually measured through spatial autocorrelation, spatial dependence is a data property that occurs when there is a spatial pattern in the attribute values, as opposed to a random pattern which implies no spatial autocorrelation. Consider an illustration by LeSage and Pace (2009) on spatial dependence: suppose two locations i and j are neighbours, then

$$y_i = \beta_{0i} y_j + X_i + \varepsilon_i \quad and$$
$$y_j = \beta_{0j} y_i + X_j + \varepsilon_j.$$

indicate that y_i depends on y_j and vice-versa.

1.4 Spatial weight

Spatial weight imposes structure that ignores the interactions that are between observations that are not neighbors. This structure constraints the number neighbours so that the spatial weight matrix is a sparse matrix. Sparse matrix is a matrix whose most of its entries are zeros. The spatial interaction between locations are measured using spatial interaction coefficient and the spatial interaction matrix. Smaller spatial weight yields large coefficient and vice versa.

1.4.1 Spatial weight matrix

Let W be an $n \times n$ matrix with entries w_{ij} such that

$$w_{ij} = \begin{cases} \omega & if \quad i \quad and \quad j \quad are \quad neighbours \\ 0 & if \quad i \quad and \quad j \quad are \quad not \quad neighbours \end{cases} \tag{1.4.1}$$

where $0 \leq \omega \leq 1$, and $w_{ii} = 0$ (i.e. locations are not self neighbours). Spatial weights are measured using geography based spatial weight that are classified as binary contiguity and distance based weight.

1.4.1.1 Binary contiguity

In the binary contiguity weight, the $(ij)^{th}$ entry in the weight matrix is 1 if i and j share a boarder, otherwise the entry is zero.
There are basically three kinds of binary contiguity matrices:

1. The Rook contiguity matrix: this contiguity recognize neighbours as boarders that share common edge. Figure 1.1 demonstrates the example of Rook contiguity, where location E share border with locations B, D, F and H neighbours.

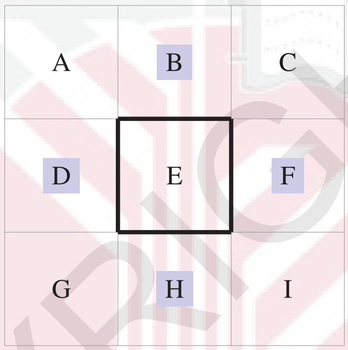

Figure 1.1: An example of the Rook contiguity to location E indicated by bold lines

2. The Bishop contiguity: The Bishop contiguity are the boarders that share common vertices. In the example of the Bishop contiguity illustrated in Figure 1.2, locations A, C, G and I are neighbours to location E.

3. The Queen contiguity:
The queen contiguity recognize locations that share both edges and vertices as spatial neighbours. It combines both rook's and bishop's contiguity together. In the example of the Bishop contiguity illustrated in Figure 1.3, locations A, B, C, D, F, G, H and I are recognized as neighbours to location E.

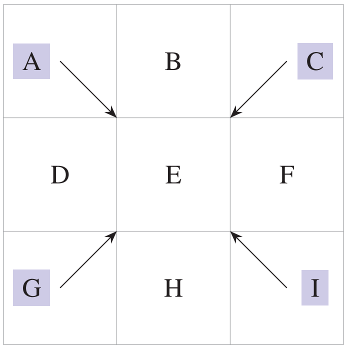

Figure 1.2: An example of the Bishop method showing contiguity to location E

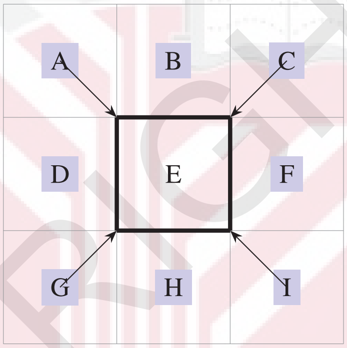

Figure 1.3: An example of the Queen method showing contiguity to location E

1.4.1.2 Distance based Weight

This based weight is between points. These points are usually between the centroid of polygons. The distance based measure can be any distance metric such as the Manhattan distance, Euclidean distance, Great circle, e.t.c. (Anselin (1988)). However, distance decay measures with respect to the metric distance are employed. This include the inverse distance with negative exponential. The distance based weight is defined as

$$w_{ij} = \begin{cases} \kappa & if \quad d_{ij} < d \\ 0 & \qquad 0 \end{cases} \tag{1.4.2}$$

where, κ is the similarity indexed. The problem with the distance based weights is that there are locations that have no neighbours (called isolates or islands). In practice, such locations are removed from the data before analysis (Anselin and Rey (2014)).

1.4.1.3 K-Nearest Neighbours Weights (knn)

This weight considers the k nearest neighbours for the computation of weights. However, a major drawback of this is the decay will be stiffer for dense distribution than in a sparse distribution locations. Another problem is that of equidistance locations (whose number is greater than k, the number of nearest neighbours). The decision of which nearest neighbours to consider becomes problem.

1.4.2 Measures of spatial dependence

Measures of spatial dependence are used to detect if there exist any spatial pattern in a spatial data set. Measures of spatial autocorrelation are usually obtained from matrix cross-product. This is typically referred to as the general cross-product statistic as defined in Huber and Ronchetti (1981) and Upton and Fingleton (1985). The commonly used measures of spatial autocorrelation are the Geary's C (Geary (1954)), the G statistics (Getis (1992)), the Moran's I (Anselin (1995)) and the GLISA (Bao and Henry (1996)), and they all have some common features. These features as noted by Bao (1999) are:

1. they first assume that the data are spatially randomly distributed.
2. the spatial pattern of the location, spatial structure of the locations and form of spatial dependence are obtained from the data.

1.4.2.1 Moran's I

The Moran's I, originally proposed by Moran (1950) is a measure of spatial autocorrelation that lies between -1 and +1. The Moran's I is defined by Equation (1.4.3)

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{S^2 \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}}, \qquad (1.4.3)$$

where, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2$. x_i is the observed value at location i, $\bar{x}_i$ is the average of the observed values at the neighbouhood of location i, W_{ij} is the measure of spatial weight which takes the value 1 if location i and j share common border and 0 otherwise.

The mean and variance of the Moran's I are given by $E(I) = -\frac{1}{n-1}$ and $Var(I) = \left(\frac{1}{S_0^2(n^2-1)}(n^2S_1 - nS_2^2 + 3S_0^2) - E(I)^2\right)$, respectively. $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$, $S_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(W_{ij}+W_{ji})}{2}$ (this simplifies to $S_1 = 2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$ if the weight matrix, W_{ij} is symmetric). $S_2 = \sum_{j=1}^{n}(W_{\bullet j} + W_{i\bullet})$ (this simplifies to $S_2 = 4\sum_{j=1}^{n} W_{i\bullet}$). $W_{\bullet j}$ and $W_{j\bullet}$ are the i^{th} column and the j^{th} row of weight matrix W_{ij}.

1.4.2.2 Geary's C

The Geary's C, also known as Geary's contiguity ratio or Geary's ratio, is defined as in Equation (1.4.4),

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}(x_i - x_j)^2}{2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1.4.4}$$

where C_{ij} is a proximity measure of values between location i and location j. Such measure of proximity are the Euclidean distance, Manhattan distance spherical distance, e.t.c.

The Geary's C has values between 0 and some value greater than 1. Values that are significantly lower than 1 indicate increasing positive spatial autocorrelation, while values that are significantly greater than 1 indicate increasing negative spatial autocorrelation.

1.4.2.3 Variogram

The concept of autocorrelation is quantified in geostatistics using a function called a variogram. Usually defined using semivariogram, the variogram is a fundamental piece of geostatistics from which one can get the model form that applies to natural mineral resources, the kriging weights, and the resulting standard errors of kriging estimation. Given two points S_i and S_j in space, the semivariogram is defined as

$$\begin{aligned}\gamma(S_i, S_j) &= \frac{1}{2} var(Z(S_i) - Z(S_j)) \\ &= \frac{1}{2} E\left[(Z(S_i) - Z(S_j))\right],\end{aligned}$$

where $Z(.)$ is the observed value at location (.). Thus, the variogram is given by $2\gamma(S_i, S_j)$.

1.5 Statement of Problems

In modeling spatial big data, researchers employ a variety of techniques, including spectral density to establish inversion for the large matrix's covariance, tapering of the covariance matrix to reduce computational burden, dimension reduction to reduce computational burden, sparsity of the precision matrix with markov fields, and approximation of the covariance function with reduced covariance (Besag and Kooperberg (1995); Furrer et al. (2006); Kaufman et al. (2008); Banerjee et al. (2008); Finley et al. (2009); Lindgren et al. (2011); Sang and Huang (2012); Eidsvik et al. (2014); Gramacy and Apley (2015); Datta et al. (2016)). Adaptive spatial dispersion clustering (ASDC) (Marchetti et al. (2018)) is another data compression technique that provides a compressed representation of the data using features that capture the basic information (spatial dependence) of the spatial field under consideration, and has demonstrated remarkable performance in spatial data compression applications where other methods failed (Fouedjio (2020); Asokan et al. (2020)). In the ASDC, data points outside the region of interest are compressed in such a way that geographic locations associated with them are allocated to spatial clusters using spectral clustering; where mean spatial observations for each cluster represents the whole data points in the cluster. However, the fact that the spatial dispersion function depends on the spatial variability of the observed spatial value, Z, implies that global outliers in observed values influence the outcome of the spatial dispersion function. The effect of such outliers, which has not been addressed by the ASDC, would have an impact on the accuracy of the outcome of the compression. The shortcomings of the ASDC has motivated us to construct Outlier Resistant Adaptive Spatial Dispersion Clustering (ORASDC). We expect the ORASDC to counter the effect of outliers in the development of the weights that would subsequently be used for the data compression.

Researchers have pointed out that most of the methods propose to detect spatial outliers are mostly prone to the problem of masking and swamping due to the aggregate of neighbourhood function (Shekhar et al. (2002); Lu et al. (2003); Liu et al. (2010); Singh and Lalitha (2018); Hadi and Imon (2018)). Masking occurs when an outlying observations are incorrectly declared as inliers. Swamping on the other hand, occurs when clean observations are incorrectly classified as outliers (Hadi and Simonoff (1993)). Other methods are only effective for small data size with no reliability measures (Lu et al. (2003)). Some adopt measures that do not capture multi-neighbour contiguity (Hadi and Imon (2018); Imon and Hadi (2020)). Non-robustness of most measures pose suspicion to the performance of some methods. Residual spatial autocorrelation reflects the amount of spatial autocorrelation in the variance that is not explained by explanatory variables, according to Gaspard et al. (2019), which failure to incorporate properly might result in issues including underestimating standard error, biased parameter estimations, and model mispecification. Another flaw is slow performance in the face of large amount of data (Dai et al. (2016)). These flaws prompted us to develop a new method of identification of spatial outlier that appropriately capture spatial contiguity in spatial big data, which we call the Adjacency Weighted Spatial Outlier.

Representation of internally spatial studentized residuals requires a spatial statistic that contains the spatial neighbourhood information in both the dependent variable and the residuals. Most works in the literature on spatial field focused mainly on the statistic that contains residual spatial autocorrelation (Martin (1992); Christensen et al. (1992); Haining (1994); Shi and Chen (2009)). Not only does including spatial neighbouhood information on the both the error term and the dependent variable is expected to help in detecting spatial outliers, but also to improve the performance of model fitting in the spatial statistics. This inspired us to construct an internally studentized spatial residual which can be used to construct externally studentized spatial residual, detect observations with large spatial residuals and subsequently be used for robust spatial model fitting.

Addressing the problem of outliers in the vertical direction does not suffice in fishing the effect of influential observations in model fitting. Large studies in the literature have indicated the effect of leverage in the classical regression (Hoaglin and Welsch (1978); Belsley et al. (1980); Huber and Ronchetti (1981); Cook and Weisberg (1982); Chatterjee and Hadi (1988); Rousseeuw and Van Zomeren (1990); Hadi (1992); Martin (1992); Christensen et al. (1992); Imon (2002); Habshah et al. (2009); Midi and Mohammed (2015); Bagheri and Midi (2015)). Thus, spatial regression model would not be an exception, and so require proper definition of leverage in spatial regression, which would help in coping with the effect of spatial leverage. Spatial leverage in model with spatial autocorrelation in the error term has been expressed by Martin (1992); Christensen et al. (1992); Haining (1994); Shi and Chen (2009). Dai et al. (2016) detected the spatial outliers without given due consideration to the effect of the leverage in the derived statistics. A measure that contains spatial information in both the dependent variable and the residual term and spatial leverage definition are imperative to appropriately classify influential observations. This motivated us to develop a spatial outlier detection technique in spatial regression that adopts the classification of spatial observations into the categories: regular observations, good leverage points, bad leverage points (outlier in the x direction) and vertical outliers.

Building robust statistical models requires detection and removal of the effects of outliers and influential observations in most statistical applications through down weighing techniques (Huber and Ronchetti (1981); Cook and Weisberg (1982); Beckman and Cook (1983); Midi and Mohammed (2015); Alguraibawi et al. (2015); Insolia et al. (2021)). These methods usually assume shift in the mean of the outlier observations; hence, Mean Shift Outlier Model (MSOM). While the influential observations are assigned weights that results in eliminating them in the MSOM (Insolia et al. (2021)), models that attach value to the influential observations for revealing important features insist on retaining such observations in the model in a fashion that construct special weights according to their relevance (Beckman and Cook (1983); Insolia et al. (2021)). This adopts models that assume shift or inflation in variance of the outliers, called Variance Inflation Outlier Model (VIOM). The detected outliers or influential observations in spatial applications require spatial weights, that contains the spatial information of the observations, which can be used to appropriately accommodate the detected observations with inflated variance or variance shift in the spatial regression model. Dai et al. (2016) used the ML, which is deficient of loss in degrees of freedom, in estimation. They accommodated the spatial outliers as a group in a fashion similar to classical regression instead of way that will capture the spatial contiguity of the outliers as a block. These shortcomings motivated us to develop spatial accommodation method that accommodate the spatial outliers and improve the spatial estimation performance of the parameters in the spatial regression model.

1.6 Research Questions

1. Does the effect of the global outliers affect the ASDC data compression method?
2. Can incorporating the spatial contiguity of the residuals improve the outlier detection performance in the spatial Big data?
3. Does developing a test statistic that incorporate the neighbourhood information help in detecting large spatial studentized residuals?
4. Can adopting the classification of spatial observations into the categories: regular observations, good leverage points, bad leverage points (outlier in the x direction) and vertical outliers improve in detecting spatial influential observations?
5. Does the spatial variance shift outlier model shows improvement in performance due to incorporating spatial neighbourhood information and subsequent

development of spatial weight results in a better fitting, spatial outlier detection and accommodation?

1.7 Research objectives

The study aims at developing a robust spatial regression model for big data using robust Adaptive Spatial Dispersion Clustering. The specific objectives are:

1. To construct Adaptive Spatial Dispersion Clustering (ASDC) that is resistant to the effect of global outliers.

2. To develop a new diagnostic measure for the identification of spatial outliers in a large spatial data set .

3. To formulate a new diagnostic measure for identification of influential observations in spatial data using a statistic that capture neighbouhood information of observations.

4. To develop a new spatial diagnostic plot to classify observations into four categories: regular observation, good leverage points, bad leverage points and vertical outlier.

5. To establish spatial weights that will determine inflation in variance and accommodate the detected spatial outliers in the spatial regression model.

1.8 Scope and limitations of the study

Robust spatial regression modeling in spatial big data as a relatively new field in statistics has not received adequate attention. In particular, research on spatial regression model that has autoregession in both the dependent variable and the residual terms on big spatial data has not been addressed in the literature to the best of our knowledge.

The importance of robust spatial regression modeling in big data is apt due to its wide range of applications and the ever-increasing data size as a result of the availability of new data collection devices. As a relatively new subject, there is no literature on techniques for robust data compression or models for robustly detecting and accommodating influential geographical observations. The algorithms created contain codes that address the issue of influential spatial observations.

Relatively larger compression sizes are considered because as sample size increases, the confidence in estimates is expected to increase, and uncertainty decreases thereby producing greater precision (Biau et al. (2008)). Positive spatial autocorrelation are used in both dependent variables and the residuals due to its importance in revealing significant features of spatial dependence in applications. Moreover, the autocorrelations are considered as low and high in simulations studies.

1.9 Outline of the thesis

The contents of this thesis are divided into seven chapters in accordance with the study objectives. The thesis chapters are organized in such a way that the goals are clear and organized in a logical order.

Chapter Two: This chapter starts by reviewing measures of spatial dependence which include spatial weights, measures of spatial autocorrelation, the general spatial autocorrelation with its variations, maximum likelihood estimations and information criteria are discussed.

Chapter Three: Using robust adaptive spatial dispersion clustering, this chapter primarily addresses the problem of data compression. The effects of local spatial dispersion on outlier detection are examined. Adaptive spatial dispersion clustering, spectral clustering, and spatial dispersion function are also studied. Simulation results findings and demonstration of the ORASDC on the California housing data are presented.

Chapter Four: The fourth chapter uses weighted adjacency residuals to detect spatial outliers. The proposed adjacency weighted spatial outlier residuals are discussed and compared to the score statistic based on the general spatial model. The weights used, the metric distant reciprocal (MDR) and the exponential weight (EW), are described in detail on how to obtain the t_{awsor}. In comparison to the score statistics, simulation studies on the first order spatial autoregressive model (FAR), mixed regressive-spatial autoregressive model (SAR), spatial autoregressive error model (SEM), and general spatial model (GSM) are presented. Examples of real data applications are also presented. Finally, illustration of the AWSOR on the compressed California housing data using ORASDC are presented.

Chapter Five: In chapter five, a robust spatial diagnostic plot is proposed. Some diagnostic measures in the classic linear regression model are reviewed in relation to the spatial regression model. We represented the leverage values of hat matrix of linear regression to GSM model and extended the internally studentized residual and externally studentized residual of linear regression to GSM model. We also extended the Cook's distance and the overall potential influence of linear regression to GSM model and developed a method of identification of influential observations of GSM model by proposing a procedure of classification of observations into regular observations, vertical outliers, good and bad leverage points. Simulation studies are used to evaluate the performances of the proposed methods and finally applied the proposed methods on gasoline price data for retail sites in Sheffield, UK, Covid-19 data at Georgia, USA, and the Life expectancy data in USA counties. The results of application on the ORASDC compressed California housing data are also presented.

Chapter Six: In this chapter, variance shift outlier model are presented in the classical regression and its equivalence, called spatial variance outlier model (SVSOM), in the general spatial model is obtained using procedure based on the restricted maximum likelihood estimation. Spatial weight based on the inflated variance are obtained to accommodate spatial outliers. Simulation studies are performed to classify and accommodate spatial outliers; real application to Georgia counties COVID-19 data are also presented.

Chapter Seven: In this chapter, the summary, conclusion and recommendations of the thesis are presented. Recommendations for future researches are also presented in the chapter.

REFERENCES

Aggarwal, C. C. (2013). Spatial outlier detection. In *Outlier Analysis*, pages 345–368. Springer.

Aggarwal, C. C. (2015). Outlier analysis. In *Data mining*, pages 237–263. Springer.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Alguraibawi, M., Midi, H., and Imon, A. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, 2015.

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):85–93.

Anselin, L. (1988). *Spatial econometrics: methods and models.* Kluwer Publishing Academics.

Anselin, L. (1990). Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, 20(2):141–163.

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. *New tools for spatial analysis*, 17:45–54.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical analysis*, 27(2):93–115.

Anselin, L. and Rey, S. J. (2014). *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL.* GeoDa Press LLC.

Asokan, A., Anitha, J., Ciobanu, M., Gabor, A., Naaji, A., and Hemanth, D. J. (2020). Image processing techniques for analysis of satellite images for historical maps classification—an overview. *Applied Sciences*, 10(12):4207.

Bagheri, A. and Midi, H. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions*, pages 51–70.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Bao, S. (1999). An overview of spatial statistics. *University of Michigan, USA, China Data Center*.

Bao, S. and Henry, M. (1996). Heterogeneity issues in local measurements of spatial association. *Geographical Systems*, 3:1–14.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245.

Beckman, R. J. and Cook, R. D. (1983). Outlier.......... s. *Technometrics*, 25(2):119–149.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Ben-Gal, I. (2005). Outlier detection: Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers, red. o. maimon, l. rokach.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Biau, D. J., Kernéis, S., and Porcher, R. (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clinical orthopaedics and related research*, 466(9):2282–2288.

Billor, N., Hadi, A. S., and Velleman, P. F. (2000). Bacon: blocked adaptive computationally efficient outlier nominators. *Computational statistics & data analysis*, 34(3):279–298.

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298.

Cerioli, A. and Riani, M. (2002). Robust methods for the analysis of spatially autocorrelated data. *Statistical Methods and Applications*, 11(3):335–358.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons.

Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, 275:314–347.

Chen, D., Lu, C.-T., Kou, Y., and Chen, F. (2008). On detecting spatial outliers. *Geoinformatica*, 12(4):455–475.

Christensen, R., Johnson, W., and Pearson, L. M. (1992). Prediction diagnostics for spatial linear models. *Biometrika*, 79(3):583–591.

Cook, R. D. (1977). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Cook, R. D., Holschuh, N., and Weisberg, S. (1982). A note on an alternative outlier model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):370–376.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, 12(2):115–125.

Dai, X., Jin, L., Shi, A., and Shi, L. (2016). Outlier detection and accommodation in general spatial models. *Statistical Methods & Applications*, 25(3):453–475.

Darmofal, D. (2006). Spatial econometrics and political science. *Society for Political Methodology Working Paper Archive: http://polmeth. wustl. edu/workingpapers. php*.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Davalos, S. (2017). Big data has a big role in biostatistics with big challenges and big expectations. *Biostatistics and Biometrics Open Access Journal*, 1(3):63–64.

De Mauro, A., Greco, M., and Grimaldi, M. (2015). What is big data? a consensual definition and a review of key research topics. In *AIP conference proceedings*, volume 1644, pages 97–104. American Institute of Physics.

Dowd, P. (1984). The variogram and kriging: robust and resistant estimators. In *Geostatistics for natural resources characterization*, pages 91–106. Springer.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.

Finley, A. O., Banerjee, S., and McRoberts, R. E. (2009). Hierarchical spatial models for predicting tree species assemblages across large domains. *The annals of applied statistics*, 3(3):1052.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fouedjio, F. (2016a). Discovering spatially contiguous clusters in multivariate geostatistical data through spectral clustering. In *International Conference on Advanced Data Mining and Applications*, pages 547–557. Springer.

Fouedjio, F. (2016b). A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, 18:333–351.

Fouedjio, F. (2020). Clustering of multivariate geostatistical data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(5):e1510.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gaspard, G., Kim, D., and Chun, Y. (2019). Residual spatial autocorrelation in macroecological and biogeographical modeling: a review. *Journal of Ecology and Environment*, 43(1):19.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146.

Genton, M. G. (1998). Highly robust variogram estimation. *Mathematical Geology*, 30(2):213–221.

Getis, D. (1992). Spatial autocorrelation and spatial association by the use of distance statistics. *Geographical Analysis*, 24:93–116.

Geurs, K. T. and Van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport geography*, 12(2):127–140.

Gramacy, R. B. and Apley, D. W. (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Gumedze, F. N. (2019). Use of likelihood ratio tests to detect outliers under the variance shift outlier model. *Journal of Applied Statistics*, 46(4):598–620.

Gumedze, F. N. and Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Medical Research Methodology*, 11(1):1–9.

Gumedze, F. N., Welham, S. J., Gogel, B. J., and Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model. *Computational Statistics & Data Analysis*, 54(9):2128–2144.

Habeeb, R. A. A., Nasaruddin, F., Gani, A., Hashem, I. A. T., Ahmed, E., and Imran, M. (2019). Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, 45:289–307.

Habshah, M., Norazan, M., and Rahmatullah Imon, A. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36(5):507–520.

Hadi, A. S. (1990). Two graphical displays for the detection of potentially influential subsets in regression. *Journal of Applied Statistics*, 17(3):313–327.

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis*, 14(1):1–27.

Hadi, A. S. and Imon, A. R. (2018). Identification of multiple outliers in spatial data. *International Journal of Statistical Sciences*, 16:87–96.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American statistical association*, 88(424):1264–1272.

Haines, L. M. and Thiart, C. (2022). The impact of spatial statistics in africa. *Spatial Statistics*, 50:100580.

Haining, R. (1994). Diagnostics for regression modeling in spatial econometrics. *Journal of Regional Science*, 34(3):325–341.

Haining, R. (2001). Spatial autocorrelation. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 14763–14768. Pergamon, Oxford.

Haining, R. P. and Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

Härdle, W., Lu, H. H.-S., and Shen, X. (2018). *Handbook of big data analytics*. Springer.

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke, A. (2014). Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences*, 46(1):1–31.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Hawkins, D. M. and Cressie, N. (1984). Robust kriging—a proposal. *Journal of the International Association for Mathematical Geology*, 16(1):3–18.

Hawkins, D. M. and Zamba, K. (2005). A change-point model for a shift in variance. *Journal of Quality Technology*, 37(1):21–31.

Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, 95(449):197–211.

Hiekkalinna, T., Göring, H. H., and Terwilliger, J. D. (2012). On the validity of the likelihood ratio test and consistency of resulting parameter estimates in joint linkage and linkage disequilibrium analysis under improperly specified parametric models. *Annals of human genetics*, 76(1):63–73.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Huber, P. J. and Ronchetti, E. (1981). Robust statistics john wiley & sons. *New York*, 1(1).

Imon, A. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, 3:207–218.

Imon, A. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied statistics*, 32(9):929–946.

Imon, A. R. and Hadi, A. S. (2020). Identification of multiple unusual observations in spatial regression. *STATISTICS APPLICATIONS*.

Insolia, L., Chiaromonte, F., and Riani, M. (2021). A robust estimation approach for mean-shift and variance-inflation outliers. *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, page 17.

Jayasankar, U., Thirumal, V., and Ponnurangam, D. (2021). A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. *Journal of King Saud University-Computer and Information Sciences*, 33(2):119–140.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Kitchin, R. and Thrift, N. (2009). *International encyclopedia of human geography*. Elsevier.

Kou, Y. and Lu, C.-T. (2008). Outlier detection, spatial. *Encyclopedia of GIS*, pages 1539–1546.

Lansley, G., de Smith, M., Goodchild, M., and Longley, P. (2019). Big data and geospatial analysis. *arXiv preprint arXiv:1902.06672*.

Lark, R. M. (2002). Modelling complex soil properties as contaminated regionalized variables. *Geoderma*, 106(3-4):173–190.

Lax, D. A. (1975). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80(391):736–741.

Lehmann, R., Lösler, M., and Neitzel, F. (2020). Mean shift versus variance inflation approach for outlier detection—a comparative study. *Mathematics*, 8(6):991.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Taylor & Francis Group, LLC.

LeSage, J. P. (1999). The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio*, 28(11).

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Liu, X., Lu, C.-T., and Chen, F. (2010). Spatial outlier detection: Random walk based approaches. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 370–379. ACM.

Lord, D., Qin, X., and Geedipally, S. R. (2021). Chapter 2 - fundamentals and data collection. In Lord, D., Qin, X., and Geedipally, S. R., editors, *Highway Safety Analytics and Modeling*, pages 17–57. Elsevier.

Lu, C.-T., Chen, D., and Kou, Y. (2003). Detecting spatial outliers with multiple attributes. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 122–128. IEEE.

Marchetti, Y., Nguyen, H., Braverman, A., and Cressie, N. (2018). Spatial data compression via adaptive dispersion clustering. *Computational Statistics & Data Analysis*, 117:138–153.

Martin, R. J. (1992). Leverage, influence and residuals in regression models when observations are correlated. *Communications in statistics-theory and methods*, 21(5):1183–1212.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the american Statistical association*, 88(423):968–978.

Meloun, M. and Militkỳ, J. (2011). *Statistical data analysis: A practical guide*. Woodhead Publishing Limited.

Midi, H. and Mohammed, A. (2015). The identification of good and bad high leverage points in multiple linear regression model. *Mathematical Methods and System in Science and Engineering*, pages 147–158.

Midi, H., Sani, M., Ismaeel, S. S., and Arasan, J. (2021). Fast improvised influential distance for the identification of influential observations in multiple linear regression. *Sains Malaysiana*, 50(7):2085–2094.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14:849–856.

Okunlola, O. A., Alobid, M., Olubusoye, O. E., Ayinde, K., Lukman, A. F., and Szűcs, I. (2021). Spatial regression and geostatistics discourse with empirical application to precipitation data in nigeria. *Scientific Reports*, 11(1):1–14.

Pole, A., West, M., and Harrison, J. (2018). *Applied Bayesian forecasting and time series analysis*. Chapman and Hall/CRC.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). Bootstrap sampling distributions. In *Subsampling*, pages 3–38. Springer.

Puterman, M. L. (1988). Leverage and influence in autocorrelated regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 37(1):76–86.

Rashid, A. M., Midi, H., Slwabi, W. D., and Arasan, J. (2021). An efficient estimation and classification methods for high dimensional data using robust iteratively reweighted simpls algorithm based on nu-support vector regression. *IEEE Access*, 9:45955–45967.

Rey, S. J. and Anselin, L. (2010). Pysal: A python library of spatial analytical methods. In *Handbook of applied spatial analysis*, pages 175–193. Springer.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

Savitz, E. (2012). Gartner: Top 10 strategic technology trends for 2013. *URL http://www. forbes. com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years*.

Schall, R. and Dunne, T. T. (1988). A unified approach to outliers in the general linear model. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 157–167.

Shekhar, S., Lu, C.-T., and Zhang, P. (2002). Detecting graph-based spatial outliers. *Intelligent Data Analysis*, 6(5):451–468.

Shi, L. and Chen, G. (2009). Influence measures for general linear models with correlated errors. *The American Statistician*, 63(1):40–42.

Singh, A. K. and Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods*, 47(1):247–257.

Skov-Petersen, H. (2001). Estimation of distance-decay parameters: Gis-based indicators of recreational accessibility. In *ScanGIS*, pages 237–258.

Sun, X.-L., Wu, Y.-J., Zhang, C., and Wang, H.-L. (2019). Performance of median kriging with robust estimators of the variogram in outlier identification and spatial prediction for soil pollution at a field scale. *Science of the Total Environment*, 666:902–914.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):53–55.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46.

Torrecilla, J. L. and Romo, J. (2018). Data learning from big data. *Statistics & Probability Letters*, 136:15–19.

Upton, G. and Fingleton, B. (1985). *Spatial data analysis by example. Volume 1: Point pattern and quantitative data.* John Wiley & Sons Ltd.

Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242.

Von Luxburg, U. (2004). *Statistical learning with similarity and dissimilarity functions*. PhD thesis, Technische Universität Berlin Berlin, Germany.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.

Yee Peng, L., Midi, H., Rana, S., and Fitrianto, A. (2016). Identification of multiple outliers in a generalized linear model with continuous variables. *Mathematical Problems in Engineering*, 2016.

Yildirim, V. and Mert Kantar, Y. (2020). Robust estimation approach for spatial error model. *Journal of Statistical Computation and Simulation*, 90(9):1618–1638.

Zhang, M., Li, Y., Lu, J., and Shi, L. (2019). Outlier detection and accommodation in meta-regression models. *Communications in Statistics-Theory and Methods*, 50(8):1728–1744.

Zhang, X., Xie, R., and Ma, P. (2018). Statistical leveraging methods in big data. In *Handbook of Big Data Analytics*, pages 51–74. Springer.

Zhang, Y., Chang, H. H., Iuliano, A. D., and Reed, C. (2022). Application of bayesian spatial-temporal models for estimating unrecognized covid-19 deaths in the united states. *Spatial statistics*, page 100584.

Zimmerman, D. L. (2010). Likelihood-based methods. *Handbook of spatial statistics*, pages 45–56.