Prediction and Anomaly Detection Techniques for Spatial Data

Xutong Liu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Chang-Tien Lu, Chair
Ing-Ray Chen
Naren Ramakrishnan
Jason Xuan
Qi Li

May 7, 2013
Northern Virginia Center, Virginia

Spatial, Multivariate, Robust Inference, Anomaly Detection

Abstract

With increasing public sensitivity and concern on environmental issues, huge amounts of spatial data have been collected from location based social network applications to scientific data. This has encouraged formation of large spatial datasets and generated considerable interests for identifying novel and meaningful patterns. Allowing correlated observations weakens the usual statistical assumption of independent observations, and complicates the spatial analysis. This research focuses on the construction of efficient and effective approaches for three main mining tasks, including spatial outlier detection, robust inference for spatial dataset, and spatial prediction for large multivariate non-Gaussian data.

spatial outlier analysis, which aims at detecting abnormal objects in spatial contexts, can help extract important knowledge in many applications. There exist the well-known masking and swamping problems in most approaches, which can't still satisfy certain requirements aroused recently. This research focuses on development of spatial outlier detection techniques for three aspects, including spatial numerical outlier detection, spatial categorical outlier detection and identification of the number of spatial numerical outliers.

First, this report introduces Random Walk based approaches to identify spatial numerical outliers. The Bipartite and an Exhaustive Combination weighted graphs are modeled based on spatial and/or non-spatial attributes, and then Random walk techniques are performed on the graphs to compute the relevance among objects. The objects with lower relevance are recognized as outliers. Second, an entropy-based method is proposed to estimate the optimum number of outliers. According to the entropy theory, we expect that, by incrementally removing outliers, the entropy value will decrease sharply, and reach a stable state when all the outliers have been removed. Finally, this research designs several Pair Correlation Function based methods to detect spatial categorical outliers for both single and multiple attribute data. Within them, Pair Correlation Ratio(PCR) is defined and estimated for each pair of categorical combinations based on their co-occurrence frequency at different spatial distances. The observations with the lower PCRs are diagnosed as potential SCOs.

Spatial kriging is a widely used predictive model whose predictive accuracy could be significantly compromised if the observations are contaminated by outliers. Also, due to spatial heterogeneity, observations are often different types. The prediction of multivariate spatial processes plays an important role when there are cross-spatial dependencies between multiple responses. In addition, given the large volume of spatial data, it is computationally challenging. These raise three research topics: 1).robust prediction for spatial data sets; 2).prediction of multivariate spatial observations; and 3). efficient processing for large data sets.

First, increasing the robustness of spatial kriging model can be systematically addressed by integrating heavy tailed distributions. However, it is analytically intractable inference. Here, we presents a novel Robust and reduced Rank spatial kriging Model (R^3-SKM), which is resilient to the influences of outliers and allows for fast spatial inference. Second, this research introduces a flexible hierarchical Bayesian framework that permits the simultaneous modeling of mixed type variable. Specifically, the mixed-type attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. Finally, the knot-based techniques is utilized to model the predictive process as a reduced rank spatial process, which projects the process realizations of the spatial model to a lower dimensional subspace. This projection significantly reduces the computational cost.

Acknowledgments

First and foremost I am deeply grateful to my advisor, Dr. Chang-Tien Lu, for all his support on both my research and other matters of life during my PhD study. It has been a great fortune for me to work under Dr. Lu's supervision. His systematic training not only focused on methodologies in doing high-standard research but also covered other fundamental skills, including technical writing and giving professional talks that are all essential for being an excellent researcher. Dr. Lu's guidance, patience, and dedication during my PhD study are greatly appreciated.

I would also like to thank my committee member, Dr. Ing-Ray Chen, Dr. Naren Ramakrishnan, Dr. Jason Xuan, and Dr. Qi Li for serving my thesis committee and for their valuable advice and comments.

I owe special thanks to my research group for giving me an inspiring and pleasant work environment. We spent countless hours together discussion research and other fun parts about life. I am grateful to all of them: Feng Chen, Jing Dai, Haili Dong, Bingsheng Wang, Yen-Cheng Lu, Ray Dos Santos, Arnold Boedijardjo, Manu Shukla and Chad Steel. My appreciation also goes to the fellow students at NVC center: Sirui Liu, Zuojin Wang, Yang Chen and Yating Wang. They are all inseparable from my happy years at Virginia Tech.

I would like to thank my parents, my sisters and my brother for their all time love and support. My most special thanks go to my husband, Changshu Jian. It is his love, dedication, and endless support that made me reach this far. Finally, I want to thank my son, Eli, who brought the sunshine into my busiest life during the last year before my final defense.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged formation of large spatial datasets in many fields and has generated considerable interest in statistical modeling for such data. Modeling large spatial data sets have received much attention in the multiple research areas. Illustrative applications include climate prediction [76, 153], environmental monitoring [86], molecular dynamical pattern mining [172], and infectious disease outbreak prediction [103].

Meanwhile, with the ever-increasing volume of spatial data, identifying hidden but potentially interesting patterns of anomalies has attracted considerable attentions, particularly from the areas of data mining experts and geographers. For example, environmental scientists may want to identify abnormally behaving water monitoring sensors; a customs agent may want to discover anomalies among cargo shipments with RFID tags, to identify potentially deviant shipments even before they cross the border; city officials may want to identify threats or malfunctions based on numerous sensors placed around a metropolitan area in subways and tunnels, etc.

1.1 Features of spatial data and spatial prediction

A primary feature driving many methods of spatial analysis is described by Tobler's "First Law of Geography": "Everything is related to everything else, but near things are more related than far things" [37]. Statistically, Tobler's "law" refers to positive spatial autocorrelation in which pairs of observations taken nearby are more alike than those taken farther apart. Allowing correlated observations weakerns the usual statistical assumption of independent observations and complicates analysis in several ways. If we assume

independent observations, any observed spatial patterns can be modeled as a spatial trend in the expected values of the observations. If we allow correlation, observed spatial patterns may be due to a trend in expected values, correlation among observations with the same expectation, or some combination of the two.

In georeferenced studies, data are always collected at particular locations whether these be in a forest, at a particular street address, in a laboratory, or in a particular position on a gene expression array. In many cases, the location may provide additional insight into circumstances associated with the data item, in short "where" we collect a measurement may inform on "what" or "how much" we measure. The field of spatial statistics involves statistical methods utilizing location and distance in inference.

Spatial statistics in the collection of statistical methods in which spatial locations play an explicit role in the analysis of data. Most often, spatial statistics are used to detect, characterize, and make inferences about spatial patterns, primarily in ecology and geography. Spatial statistical methods may be classified by the inferential questions of interest. These questions are motivated by application and often fall into categories based on the type of data available. Cressie [39] provides three useful categories of spatial data that also serve to categorize both inferential questions and inferential questions and inferential approaches. Here present these in order of data complexity.

First, consider spatial point process data consisting of a set of observed locations in a defined study area. We consider the locations themselves as the realization of some random process and seek inference regarding the properties of this process. Examples include the locations of trees in a forest, neurons in the brain, and galaxies in the universe. Questions of interest include:

- Are observations equally likely at all locations? If not, where are observations more or less likely?
- Are there (spatially-referenced) covariates that drive the probability of observations occurring at particular locations?
- Does the presence of an observation at a particular location either encourage or inhibit further observations nearby?
- If observations do impact the probability of nearby observations, what is the range of influence of observations?

Much of the literature on spatial point processes involves modeling of stochastic processes, and comparison of competing models describing observed spatial patterns. Statistical questions address estimation of model parameters and assessment of fit of various models.

Next, suppose we have geostatistical data consisting of a set of measurements taken a fixed set of locations, e.g., ozone levels measured at each of a set of monitoring stations. In this case, the locations are set by

design and not random. An inferential question of interest is prediction of the same outcome at locations where no measurement was taken. Examples of such prediction appear each day in weather maps of current temperatures interpolated from a set of official monitoring stations The literature on spatial prediction builds from a fairly simple concept: spatial correlation suggests that one should given more weight to observations near the prediction location than to those far away. Spatial prediction theory explores how to optimally set these weights based on estimates of the underlying spatial autocorrelation structure.

Finally, we may observe data from a set of regions partitioning the study area. Such data are referred to as lattice data by Cressie [39] and regional data by Waller and Gotway [44]. Lattices may be regularly or irregularly spaced, for example pixels in an image or counties within a state, respectively, so we use the "regional" to avoid confusion with literature that assumes the term "lattice" implies a regular lattice. regional data generally involve summary measures for each region, e.g. number of residents in an enumeration district, average income for residents of the region, or number of items delivered within a postal delivery zone. Inferential questions often involve accurate estimation of summaries from regions with small sample sizes ("small area estimation"), or regression or generalized linear modeling linking outcomes and covariates measured on the same set of regions. In the first case, statistical methods involve how best to "borrow strength" from other regions in order to improve estimates within each region. In the second, methods involve accurate estimation of model parameters with adjustment for spatial correlation between nearby regions.

Since the inferential questions vary with data type, the models for each data type are designed separately normally.

1.2 Tasks of Spatial Anomaly Detection

Spatial anomaly analysis, which aims at detecting abnormal objects in spatial context, has been informally defined as detection of observations in the data set which appear to be inconsistent with the neighborhoods, or which deviate so much from neighborhoods so as to arouse suspicions that they were generated by a different mechanism. The abnormal behaviors represent locations that are significantly different from their neighborhood even though they may not be significantly different from the entire population[144]. Identification of spatial outliers can lead to the discovery of unexpected, interesting, and implicit knowledge, such as local instability. It has a number of practical applications such as : 1) detection of an anomalous sensor, environment[145] and traffic [144] ones; 2) detection of a location with unusual number of disease cases, like West Nile virus[29] or influenza[159]; 3) detection of anomalous spatial weather patterns, like hurricanes[145] and tornadoes[175]; 4) detection of crime hotspot[58, 159]. The importance of spatial anomaly detection is due to the fact that outliers in data translate to significant information in a wide va-

riety of domains. For example, an inconsistent traffic measurement could mean the occurrence of unusual situation. An anomalous hotspot in the meteorological image may indicate a severe weather events, like a hurricane at the gulf. Similarly, irregular temperature measurements and /or precipitation measurements could indicate the effects of **E1 Nino** and **La Nina**. The hotspot identification of Asiatic Cholera in London help make decisions how to stop the epidemic[146].

Spatial objects are associated with both spatial and non-spatial attributes. Existing spatial outlier detection approaches first define a neighborhood, and then perform the outlier detection in it. The outlier detection is performed by identifying the difference (using some distance metric) of an object with other objects in the neighborhood, by considering its non-spatial attribute(s). If this difference is unusual, the object is labeled as an outlier. The process of identification of neighborhood for performing spatial outlier detection must consider spatial auto-correlation, that is, spatial objects are under the influence of nearby spatial objects such that behavior is auto-correlated. Therefore, spatial outliers are the observations in spatial database that do not conform to a well-defined notion of normal behavior. Figure 1.1 illustrates anomalies in a spatial data set. The data has around five normal regions, since most observations in these five areas have common local non-spatial attributes. Points that are significantly far away from the regions, e.g., points s_1, s_2 and s_3 are anomalous.

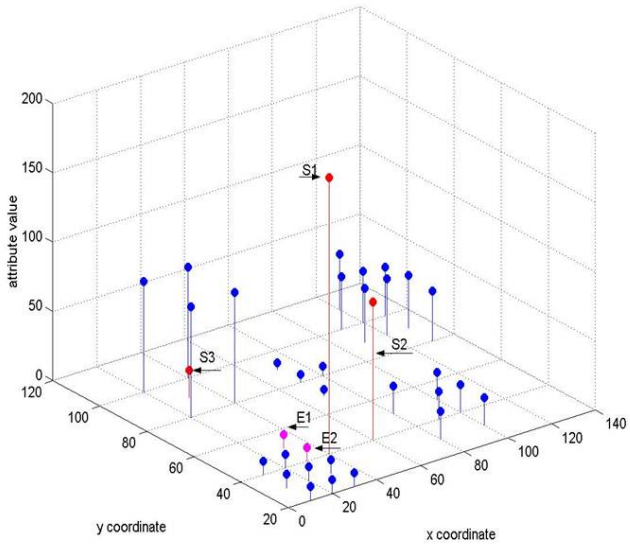


Figure 1.1: Example of spatial outliers[32]: X and Y coordinates denote the spatial locations and Z coordinate represents the value of non-spatial attribute

An important aspect of spatial outlier techniques is the nature of the desired spatial outlier. Spatial outliers can be classified into the following two categories:

- **Point outliers**. If an individual data instance can be considered as anomalous with respect to its spatial neighbors, then the instance is termed as a point outlier. This is the simplest type of anomaly and is the focus of majority of research on anomaly detection. For example, in Figure 1.1, points s_1 and s_2 are point anomalies since they are different from their corresponding neighbors.
- **Region outliers**. If a collection of spatial-closed data instances are anomalous with respect to their common spatial neighbors. The individual data instances in a region outlier may not be anomalies by themselves, but their occurrence together as a collection is anomalous. An typical example is the crime hotspot, which denotes a collection of anomalies because the same outlying value exists.

Another important aspect for spatial outlier detection techniques is the way in which the spatial outliers are output. Generally, the output generated by spatial outlier techniques is one of the following two types:

- **Outlierness**. Outlierness-based approaches assign an outlier score to each observation depending on the local differences between itself and its spatial neighbors. Thus the output of such type of techniques is a ranked list of anomalies. The analyst may choose to either analyze top few anomalies or use a cut-off threshold to select the anomalies.
- **Labels**. The approaches belonging to such categories assign a label (normal or anomalous) to each test instance. It can be controlled indirectly through parameter choices within each method.

Spatial outliers might be induced in the data set for a variety of reasons, such as the occurrence of unusual events, e.g., disease breakout, traffic accident and severe weather, but all of the reasons have a common characteristic that they are interesting to the analysis. The "interestingness" of real outliers is a key feature of spatial outlier detection.

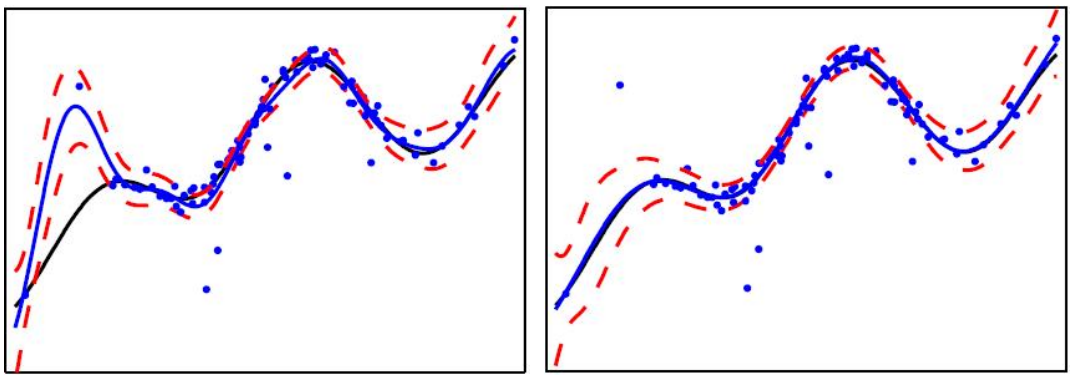
As introduced by Chandola and Kumar[26], spatial outlier detection is distinct from noise removal[154] , which deals with unwanted noise in the data set. Noise can be defined as a phenomenon in data which is not of interest to the analyst, and noise removal is driven by the need to remove the unwanted objects before any analysis is performed on the data.

1.3 Research Issues

A commonly used observation model in the Gaussian process (GP) is the Normal distribution. This is convenient since the inference is analytically tractable up to the covariance function parameters. However, a known limitation with the Gaussian observation model is its non-robustness, and replacing the normal distribution with a heavy-tailed one, such as the Student-t distribution, can be useful in problems with

outlying observations. In both the prior and the likelihood are Gaussian, the posterior is Gaussian with mean between the prior mean and the observations. In conflict this compromise is not supported by either of the information sources. Thus, outlying observations may significantly reduce the accuracy of the inference. For example, a single corrupted observation may pull the posterior expectation of the unknown function value considerably far from the level described by the other observations. (See Figure (1.2)) A robust, or outlier-prone, observation model would, however, weight down the outlying observations the more, the further away they are from the other observations and prior mean.

Spatial prediction is the process of estimating the values of a target quantity at a target quantity at unvisited locations, based on the observed measures at sampled ones. Due to spatial heterogeneity, observations are often of different types, such as continuous, ordinal, and binary, each of which conveys important information. For example, in economics studies, the living area (continuous variable), the age of the dwelling (ordinal variable), and an indicator that shows if the dwelling is located in a certain county (binary variable), are usually measured when characterizing the sale prices of houses. This raises three research challenges: 1).Modeling corss-spatial dependencies between Gaussian and non-Gaussian variables; 2). Prediction of multivariate spatial observations; and 3). Efficient processing for large data sets.



(a) Gaussian observation model (b) Student-t observation model

Figure 1.2: An example of regression with outliers by Neal [112]. On the left Gaussian and on the right the Student-t observation model. The real function is plotted with black line.

At an abstract level, a spatial outlier is defined as a spatial observation that does not conform to expected normal behavior based on its spatial neighborhood[168]. A straightforward spatial outlier detection approach, therefore, is to define areas representing normal behaviors and declare any observations which does not belong to this normal areas as an spatial outlier. However, the following issues make this apparently simple idea very challenge.

- Defining a normal area which encompasses very possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous

observation which lies close to the boundary can actually be normal, and vice-versa.

- In some spatial domains, normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.
- The occurrence of outliers which deviate strongly from the others may incur the failure of measurement process of normal behavior. Given the large volume of spatial data, it is computationally challenging to apply traditional prediction methods in either an allowable memory space limit or an acceptable time limit, even in data mining fields.
- The exact notion of an anomaly is different for different application domains. For example, in the numerical domain, an obvious deviation from its neighbors might be an anomaly, while similar deviation in categorical domain might be considered as normal. Thus, applying a technique developed in numerical domain to another is not straightforward.
- Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue.
- Often, the number of spatial outlier is unknown, and most of existing approach need to pre-define value before the identification work.
- Most of existing research works has been usually to model one variable at a time, with the anomaly detection using the data from the same type of sampled locations. However, most spatial data sets consists of different types of attributes, continuous and discrete.

Due to the above challenges, the problem of spatial outlier detection, in its most general form, is not easy to solve. The objective of this report is to investigate and develop approaches to addressing parts of the above issues. The major research issues are stated as follows:

Identification of Spatial Numerical Outliers. Normally, a Spatial Numerical Outlier (SNO) is defined as a spatial observation which is significantly different with those of its spatial neighbors. Although numerous traditional outlier detection algorithms have been proposed during the past decades, most of them can't be satisfactory with the spatial context since traditional outlier is determined by global differences which do not consider spatial relationship when identifying anomaly patterns. As the geographic rule of thumb, "Nearby things are more related than distant things" requires more considerations on spatial auto-correlation in spatial analysis. Recently, some SNOD approaches have been proposed. However, most of them have three issues: **masking problem** in which the normal objects may be misclassified as outliers; **swamping problem** in which some true outliers may be missed, and **incorrect ranking list** which is estimated due to the incorrect outlierness computation. Accurately identifying relevance spatial objects is one of the fundamental building blocks to resolve these three issues. Among several approaches to the problem of computing the relevance

scores, Random Walk(RW) based algorithms have been proven very effective. RW based techniques have been widely used for varieties of data mining tasks, including clustering and outlier detection. In this research, we investigate the benefits of RW techniques on spatial outlier detection and design two novel SNOD methods.

Identification of Spatial Categorical Outliers. Actually, in real world, the non-spatial attributes of spatial data are usually category-typed, where attributes have no intrinsic order information. A typical example is Rock whose values include *Igneous*, *Sedimentary*, and *Metamorphic*, etc. The special properties make the task of anomaly detection in spatial categorical domain seem more complicated than that of numerical data. Currently, there is a lack of Spatial Categorical Outlier Detection (SCOD) approaches. When encountering categorical data set, some introduce Spatial Numerical Outlier Detection(SNOD) methods by directly mapping the categorical attributes to continuous ones. However, there are several critical issues: 1) **misapplication**. Statistically, the definition of Spatial Categorical Outlier(SCO) is different with that of Spatial Numerical Outlier(SNO). Although both of them take focuses on the identification of abnormal behaviors, SCO is determined by the co-occurrence infrequency with regard to its neighbors, while SNO takes focuses on the differences; 2) **mapping function**; The mapping process is absolutely not straightforward, especial for nominal attributes; 3) **swamping and masking problems**; Without estimating outlierness correctly, some true outliers may be missed and normal ones misclassified as outliers. Among several approaches to capturing the co-occurrence frequency, Pair Correlation Function(PCF) measurement have been proven very effective. The benefits of PCF techniques on SCOD are investigated and the related algorithms are designed to identify SCOs with single or multiple attributes.

Identification of Number of Spatial Numerical Outliers Currently, most of spatial outlier detection algorithms suffer from the limitation that the number of outliers are specified by a human user. This parameter varies from one data set to another. In spatial analysis, the best estimate of the number of spatial outliers determines effect on the outlier detection results. This parameter is typically either l, the number of outliers to return, or some other parameter that indirectly controls the number of outliers to return, such as an error threshold. Setting these parameters requires either detailed pre-existing knowledge of the data, or time-consuming trail. The latter case still requires that the user has sufficient domain knowledge to show what is the appropriate value. It is often impractical to expect a human with sufficient domain knowledge to be available to select the number of spatial outliers. Till now, there is no convincingly acceptable solution to the best number of spatial outliers. In this report, we tackle this problem by designing an entropy based algorithm that can efficiently determine a reasonable number of spatial outliers to return from any SNOD algorithm.

Robust Prediction and Outlier Detection for Spatial Data sets. In the Gaussian process regression, the observation model is commonly assumed to be Gaussian, which is convenient in computational perspective. A commonly used observation model is the Normal distribution. This is convenient since the inference is

analytically tractable up to the covariance function parameters. However, a known limitation with the Gaussian observation model is non-robustness. That is, the predictive accuracy of the model can be significantly compromised if the observations are contaminated by outliers. A robust observation model, such as the Student-t distribution, reduces the influence of outlying observations and improve the spatial prediction and anomaly detections. Outlier The problem, however, is the analytically intractable inference. This report discusses the properties of a Gaussian process linear model with the Student-t likelihood and utilize the Laplace approximation for approximate inference.

Spatial Prediction of Large Multivariate Non-Gaussian Data Spatial prediction is the process of estimating the values of a target quantity at unvisited locations, based on the observed measures at sampled ones. Due to the spatial heterogeneity, observations are often of different types, such as continuous, ordinal, and binary, each of which conveys important information. As most georeferenced data sets are multivariate and concern variables of different types, spatial mapping methods must be able to deal with such data. This raises two difficulties: predicting multivariate discrete random fields and modeling the dependence between continuous and discrete spatial processes. The report presents a new hierarchical Bayesian approach that permits simultaneous modeling the dependent Gaussian, count, and ordinal spatial fields.

1.4 Contributions

The major proposed research contributions can be stated as follows:

Identification of Spatial Numerical Outliers

1. **Model of two different weighted graphs based on spatial and/or non-spatial attributes.** The benefits of Random Walk(RW) techniques are investigated on identifying spatial numerical outliers. Two weighted graphs, a BiPartite(BP) graph and an Exhaustive Combination(EC) graph, are modeled based on the spatial and/or non-spatial attributes of the spatial objects. BP is a bipartite graph in which two independent sets of vertices correspond to spatial objects and clusters generated from their non-spatial attributes. EC consists of all the spatial objects and the edges among them, and each edge value is computed by the spatial and non-spatial attributes.

2. **Design of two RW based SOD algorithms.**Within these two frameworks, RW-BP(Random Walk based on BiPartite) and RW-EC(Random Walk based on Exhaustive Combination) are designed to accurately identify the local differences of spatial objects by operating the RW techniques on the weighted graphs. And, the top k objects with higher difference scores are identified as SNOs.

3. **Extensive experiments to validate the effectiveness and efficiency.** RW-BP and RW-EC methods were applied to hundreds of synthetic data sets and one real data set in which the experiment results

demonstrated their effectiveness.

Identification of Spatial Categorical Outliers

1. **Definition of Spatial Categorical Outlier.** A Spatial Categorical Outlier(SCO) is defined as a spatial observation which occurs infrequently with regards to its spatial neighbors.
2. **Design of Pair Correlation Function based Spatial Categorical Outlier Detection approach**. The capabilities of PCF(Pair Correlation Function) techniques is investigated to estimate the relevance among categories with regards to different spatial distances. PCF based approach is designed to identify SCOs in single attribute domain.
3. **Design of kNN(k Nearest Neighbor)based Spatial Categorical Outlier Detection approaches**. *k*NN based schemes are the approximations of PCF based method. Two nearest neighbor based estimators are proposed to approximate the Pair Correlation Relevance(PCR) values in single and multiple attribute domain, respectively. They allow for more efficient SCOD when memory and processor resources are issues.
4. **Extensive experiments to validate effectiveness.** PCF series of algorithms were applied in synthetic and real data sets to demonstrate their effectiveness and/or efficiencies.

Identification of Number of Spatial Numerical Outliers

1. **Definition of Spatial Local Contrast Entropy.** Spatial Local Contrast Entropy(SLCE) is the measure of the Spatial Local differences in the whole data set. It is motivated by the fact that the spatial local differences of outliers are higher than those of other normal data points. And, an outlier point significantly contributes to the SLCE since its spatial local difference is higher.
2. **Design of SLCE approach (Entropy based approach) to identifying the number of spatial numerical outliers.**. The fundamental idea of SLCE algorithm determines that when spatial outliers are incrementally removed from the data set, the SLCE value will be continuously decreased until it reaches a stable state when all the outliers have been removed.
3. **Extensive experiments to validate effectiveness.** SLCE based approach was applied in real data sets by integrating with existing popular SNOD approach to demonstrate the effectiveness itself.

Robust Prediction and Outlier Detection for Spatial Data sets.

1. **Formulation of the R^3-SKM model.** A Robust and Reduced Rank Spatial Kriging Model is proposed in which the measurement error is modeled by a heavy tailed distribution, and a Bayesian hierarchical framework is integrated to support priors on model parameters.

2. **Design of an approximate algorithm for robust parameter estimation.** The posterior distribution of latent variables conditional on parameters and observations is estimated via Gaussian approximation. Furthermore, the posterior distribution of parameters conditional on observations is estimated via Laplace approximation. It has time complexity of $O(n)$.

3. **Development of robust inference algorithms.** R^3-SP (Robust and Reduced Rank Spatial Prediction) and R^3-SOD (Robust and Reduced Rank Spatial Outlier Detection) algorithms are proposed to perform robust spatial prediction and spatial outlier detection. Their time complexities are analyzed, which scale linearly.

4. **Comprehensive experiments to validate the robustness and efficiency of the proposed techniques**. The R^3-SKM was evaluated by the extensive experiments on simulated and real data sets. The results demonstrated that the three algorithms based on R^3-SKM outperformed existing representative techniques, when the data were contaminated by outliers.

Spatial Prediction of Large Multivariate Non-Gaussian Data

1. **Design of a spatial multivariate non-Gaussian hierarchical framework.** The spatial model is based on a hierarchical framework and is specifically designed to take account of mixed type random variables.

2. **Model of a multivariate reduced-rank predictive process.** This is the first work that applies both knot-based and Laplace Approximation techniques to multivariate non-Gaussian data sets. The knot-based technique is utilized to model the predictive process as a reduced-rank spatial process, which projects the process realizations of the spatial model to a lower dimensional subspace. This projection significantly reduces the computational cost.

3. **Design of an efficient spatial prediction algorithm.** By integrating the Laplace approximation, our approach efficiently makes approximations to the posterior marginal of latent variables for the predictive process, and performs accurate spatial prediction.

4. **Performance analysis and experiment evaluation**. Theoretical analysis and extensive experiments on both simulations and real data sets have been conducted to demonstrate the performance of the proposed hierarchical mixed model. The data sets and the implementation of our model, as well as seven state-of-the-art comparison approaches.

1.5 Organization of This Dissertation

The remainder of this Ph.D. dissertation is organized as follows. Chapter 2 describes the fundamental concepts used in this dissertation for robust reference and anomaly detection in spatial dataset. Chapter 3 presents the random walk based framework for spatial numerical outlier detection. Chapter 4 proposes an entropy based estimation of the number of spatial numerical outliers in the data set. Chapter 5 describes the proposed Pair Correlation Function based techniques for spatial categorical outlier detection. Chapter 6 discusses an robust and reduced rank spatial kriging model for executing accurate parameter estimation and spatial prediction efficiently. Chapter 7 designs a spatial multivariate non-Gaussian hierarchical framework to efficiently make approximations to the posterior marginal of latent variables for the predictive process, and performs accurate spatial prediction. Chapter 8 concludes the research achievement of this dissertation, together with current publications and discussing future directions.

Chapter 2

Theoretical Foundations

In this chapter we describe the fundamental concepts of spatial data mining, including random walk, pair correlation function, kriging model, heavy tail distribution and integrated nest laplace approximation, etc. In , we discuss their relation with our proposed anomaly detection and prediction schemes.

2.1 Random Walk with Restarts

A random walk is a finite Markove chain that is time-reversible and allow weighted edges. Pan et al. [126] defined "random walk with restart" as follows: *considering a random observation that starts from the node A. The observation iteratively transmits to its neightborhood with the probability that is proportional to their edge weights. Aslo at each step, it has some probability c to return to the node A. The relevance score of node B with respect to A is defined as the steady-state probability* $r_{A,B}$ *that the particle will finally stop at node B.*

To compute the steady-state probability, let A be the query observation. An random walk can be operated from node a, and further the steady state probability vector $\vec{r_A} = (\vec{r_A}(1), ..., \vec{r_A}(N))$ is computed which records the related scores of the rest of points with regard to point A. Let $\vec{e_A}$ record the start state of the query point. It is a column vector with all its N elements zero, except for the entry that corresponds to itself which is set as 1.

The computateion of vector $\vec{r_A}$ utilizes the matrix multiplication. Let W be the adjacent matrix which stores the weight values of any two observations. Then we make a column-normalized operation on it.

$$\vec{r_A} = cW\vec{r_A} + (1-c)\vec{e_A} \tag{2.1}$$

where c records the probability of restarting the random walk from A. Equation (2.1) describes the computation of the Steady-state vector, where $\vec{r_A}$ is determined by

$$\vec{r_A} = (1-c)(I-cW)^{-1}\vec{e_A} \tag{2.2}$$

where I is the $N \times N$ identify matrix. The relevance score defined by random walk with restarts can better extract the relationships than pair-wise metrics and other traditional traph distances. It first captures the global structure of the graph, and further it can capture the multi-facet relatiohship between two observations.

Random walk with restart has been receiving increasing interest from both the application and the theoretical point of view, since it provides a good estimation of relevance among objects in a weighted graph. Defining the differences between two observations is one of the fundamental building block in spatial anomaly detection. Random walk techniques can be utilized in the anomaly detection in spatial domain.

2.2 Pair Correlation Function

In statistical mechanics, the pair correlation function in a system of particles describes how density varies as a unction of distance from a reference particle. It is related to the probability of finding an paricle in a shell dr at the distance r of another particle chosen as a reference point.

By dividing the space volum into shells dr, it is possible to compute the number of particles $dn(r)$ at a distance between r and $r + dr$ from a given point. And pair correlation function(PCF), $g(r)$, is defined as the observed probability of finding an object at a given distance, *r*, from a fixed reference particle [133]. The mathematical definition of $g(r)$ is

$$g(r) = \frac{dn(r)/N}{dv(r)/V} = \frac{dn(r)}{dv(r)} \cdot \frac{V}{N} = \frac{dn(r)}{4\pi r^2 dr} \cdot \frac{V}{N} \tag{2.3}$$

Where N and V denote the number of units and the volume of the entire system, respectively; $dn(r)$ and $dv(r)$ represent those in the shell-region; r is the distance from reference unit to the shell of interest.

The volume of the shell is give by

$$V = \frac{4}{3}\pi(r+\delta r)^3 - \frac{4}{3}\pi r^3 \approx 4\pi r^2 \delta r \tag{2.4}$$

For short distances, pair correlation function is related to how the particles are packed together. As shown in Fig. (2.1), consider hard spheres. The particles can't overlap, so the closest distance two centers can be is equal to the diameter of the particles. However, few particles can be touching one particle, then a few more can form a layer around them, which have higher pair correlation probabilities. Further away, these layers

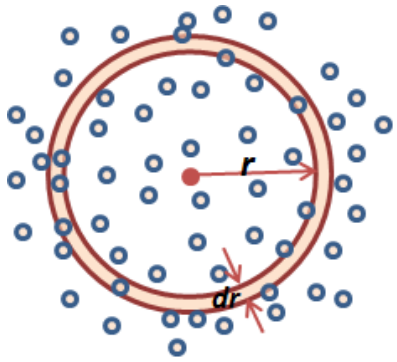


Figure 2.1: PCF using a spherical shell of thickness dr

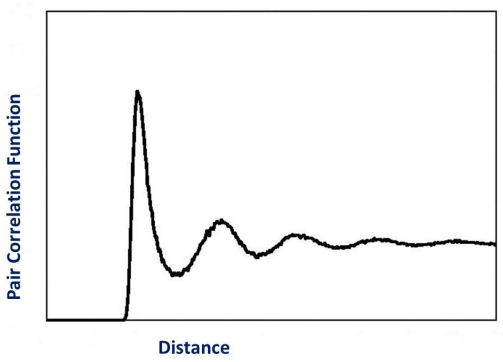


Figure 2.2: Pair Correlation Function $g(r)$ vs r

get more diffuse. For larger distances ,the probability of finding two particles is essentially constant.

Figure (2.2) shows pair correlation function calculated for a simple simulation in Figure (2.1) of two-dimension space. The function is calculated based on all pairs of particle. Note that $g(r)$ continuously decreases just as the distances between particles increases. The pair correlation function is almost uniform for this distribution of particles. Because the particles are randomly located and not tight packed, $g(r)$ is zero for $r < radius$ because the particles have radius and they are not allowed to overlap. The pair correlation function shows some structure for distances close to the reference particle. The higher density forces them to take on some short-range structure as they try to fit in the square without overlapping.

2.3 Spatial Prediction and Kriging

Spatial prediction is the predictive process that incorporates spatial dependence. It has multiple applications, including petroleum exporatio, mining, and water pollution analysis, etc. Spatial prediction is the process of estimating the values of a target quantity at unvisited locations. Development of generic and robust spatial interpolation techniques has been of interest for quite some time []. As geographic information system (GIS) and modelling techniques are becoming powerful tools in natural resource management and biological conservation, Kriging and its variants are widely recognised as primary spatial prediction techniques from the 1970s. And, Kriging is a generic name for a family of generalized least-squares regression algorithms.

2.3.1 Problem Formulation of Spatial Prediction

The general formulation of the spatial interpolation problem can be defined as follows: Given the N values of a studied phenomenon z_i, $j = 1, ..., N$ measured at discrete locations s_i within a certain region of a d-dimensional spaces, find a function $Z(s)$ which passes through the given points, that means, fulfils the condition

$$Z(s_i) = z_i, i = 1, ..., N \tag{2.5}$$

Finding appropriate interpolation methods for GIS applications poses several challenges. The modelled fields are usually very complex, data are spatially heterogeneous. In addition, datasets can be very large, originating from various sources with different accuracies. Reliable interpolation tools, suitable for GIS applications, should therefore satisfy several important demands: accuracy and predictive power, robustness and flexibility in describing various types of phenomena and applicability to large datasets.

In recent years, GIS capablities for spatial interpolation have improved by integration of advanced methods with GIS. Typical examples are conditions based on geostatistical concepts (Kriging). The principles of geostatistics and interpolation by Kriging are described in a large body of literature (e.g. [23, 39, 43, 78, 82, 119]). It is based on a concept of random functions: the surface or volume is assumed to be one realisation of a random function with a certain spatial covariance [82, 106].

2.3.2 Semivariance and Variogram

The empirical variogram provides a description about how the data are correlated with distance. Semivariance and Variogram are two preliminary concepts for the kriging estimator. The concepts semivariance and variogram are often used interchangeably. Be definition, $\gamma(h)$ is the semivariance and the variogram is $2\gamma(h)$ Semivariance (γ) of Z between two dataobjects is defined as:

$$\gamma(s_i, s_j) = \gamma(h) = \frac{1}{2} var[Z(s_i), Z[s_j]] \tag{2.6}$$

where h is the distance between observation s_i and s_j and $\gamma(h)$ is the semivariogram (commonly referred to as variogram)[]. Fig. 2.1 displays several important features by ploting $\gamma(s_i, s_j)$ againt h. The first is the "nugget", a positive value of $\gamma(s_i, s_j)$ when h close to 0. It is the residual reflecting the variance of sampling errors and the spatial variance at shorter distance than the miminum sample spacing. The "range" is the distance at which objects are not longer autocorrelated. The "range" is a value of distance at which the "sill" is reached. Observations discreted by a distance larger than the range are recognized as independent because the estimated semivariance of differences will be invariant with sample discretion distance. Hartkamp et al. [66] pointed that most of the variability is non-spatial when the ratio of sill to nugget is 0. The range provides

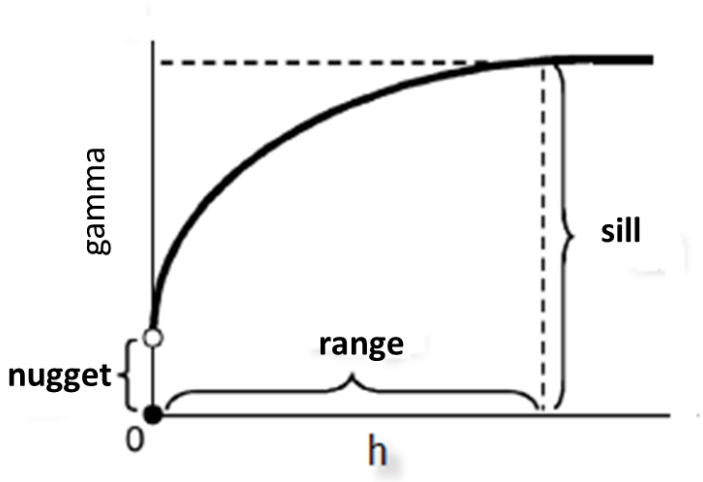


Figure 2.3: A generic variogram showing the *sill*, and *range* parameters along with a *nugget* effect

the information about the size of a search window used in the spatial prediction methods [23].

We can use the following way to estimate the semivariance

$$\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}(z(s_i) - z(s_i) + h)^2 \tag{2.7}$$

where n is the number of pair objects whose spatial distance is h.

For the sake of kriging, the empirical semivariance is replaced with an acceptable semivariogram mode. Belows are the general shapes and the equations of the mathematical medels used to describe the semivariance.

Spherical Semivariogram Model

$$\gamma(h) = \begin{cases} c_0[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}(\frac{h}{a_0})^3], & \text{for } h \leq a_0 \\ a_0, & \text{for } h > a_0, \end{cases} \tag{2.8}$$

Gaussian Semivariogram Model

$$\gamma(h) = c_0[1 - exp(-\frac{h^2}{a_0^2})], \tag{2.9}$$

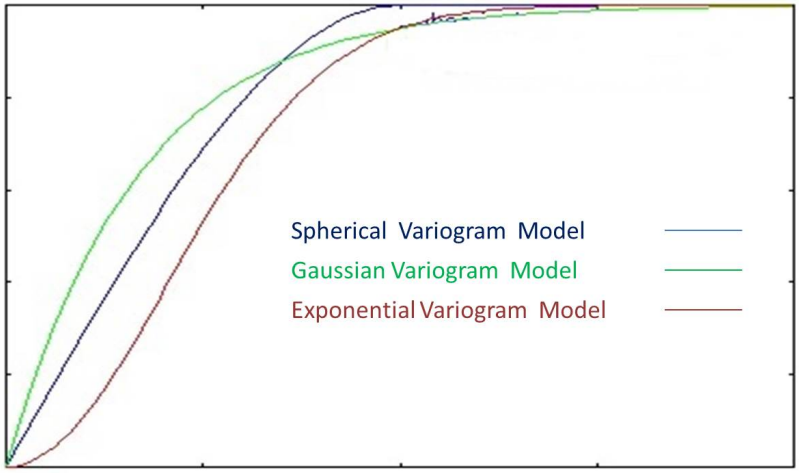


Figure 2.4: Example of three commonly used variogram models

The Exponential Semivariogram Model

$$\gamma(h) = c_0[1 - exp(-\frac{h}{a_0})], \tag{2.10}$$

where a_0 represents the range, h the lag distance, and c_0 the sill. The sperical model actualy reaches the specified sill value, c_0, at the specified range a_0. The exponential and Gaussian approach the sill asymptotically, with a_0 representing the practical range, the distance at which the semivariance reaches 95% of the sill value. The Gaussian model, with its parabolic behavior at the orgin, represents very smoothy varying properties. The spherical and exponential models exhibit linear behavior the origin, appropriate for representing properties with a higher level of short-range variability.

2.3.3 Simple Kriging

"All kriging estimators are but variants of the basic linear regression estimator

$$\hat{Z}(s_0) - \mu(s_0) = \sum_{i=1}^{n} \lambda_i [z(s_i) - \mu(s_i)] \tag{2.11}$$

"[59] where $\mu(s_0)$ is the expected value (meand) of $\hat{Z}(s_0)$ and λ_i is the driging weight assigned to the data s_i for estimation location s_0 and the same data will receive different weight for different estimation location. $\hat{Z}(s)$ is treated as a random field with a trend component, $\mu(s)$, and a residual component, $R(s) =$

$Z(s) - \mu(s)$. λ_i is large when $|s_i - s_0|$ is small. Kriging estimates residual at s as weighted sum of residuals at surrounding data points. Kriging weights are derived from semivariogram, which should characterize residual component.

Assume $\mu(s)$ and $C(s_i, s_j)$ known and take $R(s_i) = Z(s_i) - \mu(s_i)$. Best linear predictor is obtained (mean squared prediction error minimized) by chossing

$$\lambda(s) = \Sigma^{-1} c(s) \tag{2.12}$$

where $\lambda(s)$ is the vector of kriging weights $\lambda(s_i)$ and $c(s)$ is the vector of covariances $C(s, s_i)$.

The minimized mean squared prediction error, or the kriging variance,

$$\mathbb{E}[\hat{Z}(s) - Z(s)]^2 = C(0) - c(s^T \sigma^- 1 c(s)). \tag{2.13}$$

2.4 Robustness with Heavy Tailed Distributions

The heavy tailed distributions are distributions whose tails follow a power-law with low exponent, in contrast to traditional distributions (e.g., Gaussian, Exponential, Poisson) whose tails decline exponentially (or faster). To define heavy tails more precisely, let X be a random variable with cumulative distribution function $Fx) = P[X \leq x]$ and its complement $\bar{F}(x) = 1 - F(x) = P[X > x]$. We say here that a distribution $F(x)$ is heavy tailed if

$$\bar{F}(x) \sim cx^{-\alpha} 0 < \alpha < 2 \tag{2.14}$$

where c is a positive constant. In particular, when sampling random variables that follow heavy tailed distributions, the prbability of very large observations occurring is non-negligible. Therefore, heavy tailed distribution have infinite variance, reflecting the extremely high variability that they capture.

A commonly used observation model in kriging is the Gaussian distribution. It is convenient since the inference is analytically tractable up to the covariance function parameters. If both the prior and the likelihood are Gaussian, the posterior is Gaussian with mean between the prior mean and the observations. Thus, outlying observations may significantly reduce the accuracy of the inference. For example, a single corrupted observation may pull te posterior expectation of the unknown function value considerably far from the level described by the other observations. A robust, or outlier-pron, data mldel would, however, weight down the outlying objects the more, the further away they are from the other observation and prior mean.

Heavy-tailed distributions are often used to enhance the robustness of regression and classification methods

to outliers. The idea of robust regression is not new. A robust data model can reduce the influence of outlying behavior and improve the prediction. Student-t model with linear regression was studied already by West [162] and O'Hagan [56], and Neal [113] introduced it for GP regression. Consider a kriging problem, where the data comprise observations $z_i = x(s_i)+\eta_i(s_i)+\epsilon_i$ at input location $S = \{s_i\}_{i=1}^n$, where the observation errors $\eta_1, ..., \eta_n$ are zero-mean exchangeable random variables. The object of inference is the latent function $(f(s_i) = (x(s_i) + \eta_i(s_i))$, which is given a Gaussian process prior. THis implies that any finite subset of labent variables, $f = \{x(s_i) + \eta_i(s_i)\}_{i=1}^n$, has a mulivariate Gaussian distribution. In particular, at thge observed input location S the latent variables have a distribution

$$\pi(f|S) = \mathcal{N}(f|\mu, \Sigma) \tag{2.15}$$

where Σ is the covariance matrix and μ the mean function.

A formal definition of robustnes is given, for example, in terms of an outier-prone observation model. The observation is defined to be outlier-prone of order n, if $\pi(f|z_1, ..., z_{n+1}) \to \pi(f|z_1, ..., z_n)$ as $y_{n+1} \to \infty$. That is, the effect of a single conflicting observation to the posterior becomes asympotically negligible as the data approaches infinity. This contrasts heavily with the Gaussian observation model where each observation influences the posterior no matter how far it is from the others. The zero-mean Student-t distribution

$$\pi(z_i|f_i, \sigma, \nu) = \frac{\Gamma((\mu+1)/2)}{\Gamma(\mu/2)\sqrt{\mu\pi}\sigma}(1 + \frac{(z_i - f_i)^2}{\mu\sigma^2})^{-\frac{\mu+1}{2}} \tag{2.16}$$

where μ is the degress ofthe freedom and σ the scale parameter, is outlier prone of order 1, and it can reject up to m outliers if there are at leaset $2m$ objects. The challenge with the Student-t model is the inference, which is analytically intractable. Therefore, we discuss the properties of a Gaussian process regression model with the heavy tailed distribution and utilize the Laplace approximation for approximate inference.

2.5 Maximum a Posteriori (MAP) Estimation

Let $x = \{x_1, ..., x_n\}$ and $y = \{y_1, ..., y_n\}$ be realization of discrete-time random processes $\mathcal{X}$ and $\mathcal{Y}$, respectively. Their probabilities are $p(x)$ and $p(y)$. Let us assume that we are given an observed sequence y, and we know the probabilities of $p(y|x)$ and $p(x)$. We do not know which specific sequence x geneared the observation y. MAP offers a method for estimating the generating sequence $\hat{x}(y)$ on the ground of y: $\hat{x}$ is found by maximizing over x the posterior probability $p(x|y) = p(y|x)p(x)/p(y)$. Since $p(y)$ is independent with x, we can equally minimize

$$-ln[\pi(y|x)\pi(x)] \equiv H(y, x) \tag{2.17}$$

Sometimes we have a priori information about predictive process whose parameters we want to estimate. Such information can come from the correct scientific knowledge or from previous empirical evidence. Such prior information can be encoded in terms of a PDF on the parameter to be estimated. The associated probabilities $\pi(\theta)$ are called the prior probabilities. We refer to the inference based on such priors as Bayesian inference. Bayes' theorem shows the way for incorporating prior information in the estimation process. That is, choose the model that maximizes the probability of the model given the data, $\pi(\theta|y) = \pi(y|\theta)\pi(\theta)/\pi(\theta)$. The term on the left hand side is called the posterior. On the right hand, the numerator is the product of the likelihood term and prior term. The likelihood is the probability that this parameter would have produced this dataset.It is high when the parameter is a good fit to the data, and it is low when it would have generate different data. The denominatro serves as a normalization term so that the posterior integrates to unity. Thus, Bayesian inference procudes the maximum a posterior (MAP) estimate

$$argmax_\theta \pi(\theta|y) = argmax_\theta \pi(y|\theta)\pi(\theta) \tag{2.18}$$

Given the MAP formuation, three key issues remain to be addressed: the choice of the prior distribution, the specification ofthe parameters for the prior densities, and the evaluation of the MAP.

Chapter 3

Spatial Numerical Outlier Detection: Random Walk Based Approaches

A Spatial Numerical Outlier(SNO) is a spatially referenced object whose non-spatial attributes are very different from those of its spatial neighbors. Spatial Numerical Outlier Detection(SNOD) has been an important part of spatial data mining and attracted attention in the past decades. Numerous SNOD approaches have been proposed. However, in these techniques, there exist the problems of masking and swamping. That is, some spatial outliers can escape the identification, and normal objects can be erroneously identified as outliers. In this paper, two Random walk based approaches, RW-BP (Random Walk on Bipartite Graph) and RW-EC (Random Walk on Exhaustive Combination), are proposed to detect spatial outliers. First, two different weighed graphs, a BP (Bipartite graph) and an EC (Exhaustive Combination), are modeled based on the spatial and/or non-spatial attributes of the spatial objects. Then, random walk techniques are utilized on the graphs to compute the relevance scores between the spatial objects. Using the analysis results, the outlier scores are computed for each object and the top l objects are recognized as outliers. Experiments conducted on the synthetic and real datasets demonstrated the effectiveness of the proposed approaches.

The chapter is organized as follows. Section 3.1 gives the background and motivation. Section 3.2 reviews the related work on numerical outlier detection methods and data mining techniques with RW. Section 3.3 provides the detailed techniques for RW-BP. Section 3.4 studies the detailed techniques for RW-EC. Section 3.5 evaluates the performance of the proposed approaches on synthetic and real datasets. Section 3.6 concludes our works.

3.1 Background and Motivation

SNOs have been informally defined as observations in a dataset which appear to be inconsistent with their spatial neighbors, or which deviate so much from them so as to arouse suspicions that they were generated by a different mechanism. SNOD is one of the fundamental tasks in data mining, widely used in the discovery of unexpected knowledge and has a number of practical applications in areas such as, meteorological data, traffic control, satellite image analysis and hotspot identification. In contrast to Traditional Numerical Outliers (TNO), SNOs are local anomalies that are extreme compared to their neighbors, but do not necessarily deviate from the remainder of the whole dataset. Informally, spatial outliers can be called "local outliers," because they focus on local differences, while traditional outliers can be called "global outliers," since they are based on global comparison.

During the past decades, numerous TNOD algorithms have been proposed[5, 21, 81, 87]. TNOs are determined by global differences which can't be satisfactory with the spatial context. First, spatial objects have more complex structures. Second, traditional approaches do not consider spatial relationship when identifying anomaly patterns. The special properties of spatial objects require more considerations on spatial autocorrelation in spatial analysis. The outlier score of a spatial object can be evaluated by comparing the non-spatial attribute values of this observation with those of its k-Nearest Neighbors(kNN). Recently,some SOD approaches have been proposed[2, 7, 62, 90, 100, 101, 115, 143, 144, 150]. However, most of them have three issues: 1) **masking problems**: the normal objects may be misclassified as outlier; 2) **swamping problem**: some true outliers may be missed; and 3) **ranking lists**: without correct outlier scores, the outlier list may not be identified correctly. Identifying the relevance score between two spatial objects is one of the fundamental building blocks to resolve these three issues. Among several approaches to the problem of computing the relevance scores, Random Walk (RW) based algorithms have been proven very effective.

RW based techniques have been widely used for varieties of data mining tasks, including clustering [61, 64] and outlier detection[79, 111]. In this paper, we investigate the benefits of RW techniques on spatial outlier detection and then propose two novel SNOD methods, RW-BP (Random Walk on Bipartite graph) and RW-EC (Random Walk on Exhaustive Combination). Both these two approches consider using the concept of RW to compute the similarities or differences among objects. First, two different weighted graphs, a BP (Bipartite graph) and an EC (Exhaustive Combination), are constructed based on the spatial and/or non-spatial attributes. Within the frameworks, RW techniques are utilized to compute outlierness (the differences between spatial objects and their spatial neighbors) for each spatial object, and the top l objects with higher scores are identified as the spatial outliers. The main contributions of the paper are as follows:

1. **Model of two different weighted graphs based on spatial and/or non-spatial attributes.** BP is a bipartite graph in which two independent sets of vertices correspond spatial objects and clusters generated from non-spatial attributes. EC consists of all the spatial objects and the edges among

them, and each edge value is computed by the spatial and non-spatial attributes.

2. **Design of two RW based SNOD algorithms.** By operating the RW techniques on the weighted graphs, RW-BP and RW-EC algorithms are designed to accurately identify SNOs.

3. **Extensive experiments to validate the effectiveness and efficiency.** RW-BP and RW-EC methods were applied to hundreds of synthetic datasets (random generated) and one real dataset (US Housing dataset). The experiment results demonstrated their effectiveness.

3.2 Related Works

In this section, we briefly review related works, which can be categorized into three classes: 1) TNOD (Traditional Numerical Outlier Detection) methods; 2) SNOD (Spatial Numerical Outlier Detection) methods; 3) RW (Random Walk) related methods.

TNOD methods. A TNO in a dataset can be thought of as an observation, which is very different from, or inconsistent with the other observations. Many ways have been proposed to detect and subsequently, predict such anomalies or outliers. These various methods can be divided into statistical/distribution-based methods, distance based methods, density based methods, cluster-based methods.

The Distribution based or the Statistical based Outlier detection[120] attempts to fit the data to a standard distribution, like normal distribution, and then identifies the outliers with respect to the model using Discordancy Tests. However, for this, the distribution of data needs to be known. In some cases the data could follow multiple distributions and in many cases the distribution of the data may not be known.

Distance based outlier detection [87–89] proposes a unified notion of mining outliers which extends on the statistical based outlier detection. It generalizes the notion of outliers provided by many of the discordancy tests. They define outliers as those objects that do not have enough neighbors, here, neighbors are defined based on distances from a given object. However, this approach has been mainly tested with Euclidean distance, which does not scale well for high dimensional data. Moreover, it is not very efficient for large data sets and is sensitive to input parameters. An extension of the Unified approach for mining outliers, which is a modification for large datasets, especially for reducing passes over the data, has been proposed in [6, 88]. However, it does not address the issue of high dimensional outlier discovery.

Density-based algorithms define outliers based on the local densities. OPTICS-OF[176] is a technique to identify local outliers. It is an extension of OPTICS[42] which is a density based clustering algorithms. It talks about outliers relative to their neighborhood. They assign each outlier a degree with which the object

is an outlier in the local neighborhood. LOF[21] is an extension of OPTICS-OF, which proposes the setting of bounds to the local outlier factor and measures the effect of changes in the parameter MinPts. Both these approaches do not address a scenario where a small cluster itself could be outlying, as compared to the other clusters. Secondly, they do not address high dimensional outlier discovery. [3] is one of the few approaches which address this issue. It considers projects of the data and not the entire dataset to see the sparseness of the data. This could lead to a possible loss of information. Moreover, it is very time consuming to find all possible combinations of dimensions and projects.

TNOD method treats spatial and non-spatial attributes equally. However, these two types of attributes should be considered separately in the spatial context. Therefore, TNOD may not be applicable to accurately process spatial data.

SNOD methods. During the past decades, a number of algorithms have been proposed to identify outliers in the spatial databases[140, 141]. There are three basic categories, namely, visualization, statistic and graph-based approaches.

Visualization-based approaches utilize visualization techniques to highlight outlying objects. Representative algorithms include scatterplot [62] and Moran scatterplot [7]. A scatterplot shows attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. Nodes far away from the regression line are flagged as potential spatial outliers. A Moran scatter plot normalizes attribute values against the neighborhood average of value.

Statistic-based approaches execute statistical tests to measure the local inconsistencies. Typical methods include Z-value [143], Median-based and iterative-Z [100] approaches. z-value performs statistical tests to discover local inconsistencies. z-value is the normalized difference between a spatial object and the average of its spatial neighbors. The absolute z-value can determine the outlierness of an object where higher z-value indicates higher likelihood that an object is a spatial outlier.

Graph-based approaches [90, 144] detect spatial outliers by designing a function to compute the difference between an observation and its neighboring points. Kou et al. [90] developed an approach based on k-nearest neighbor relationship in spatial domain, which assigns the differences of non-spatial attribute as edge weights, and continuously cuts high-weight edges to identify isolated pooints or regions that are much dissimilar to their neighboring objects.

Other works study the special property of spatial data. Kou et al. [57] developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weights

when comparing non-spatial attributes. Zhao et al. proposed a wavelet-based approach to detecting region outliers [175]. Cheng et al. [35] presented a multi-scale approach to detect spatial-temporal outliers [101]. Adam et al. proposed an algorithm which considers both spatial relationship among neighbors [2]. A local outlier measure [150] was proposed by Sun and Chawla to capture the local behavior of data in their spatial neighborhood. Liu and Jezek proposed a method for detecting outliers in an irregularly-distributed spatial data set. The outlierness of an object o is measured by both the spatial interpolation residual and surface gradient of its neighborhod. A number of algorithms have been developed to detect spatial anomalies in meteorological images, transportation systems, and contagious disease data. An outlier detection method to identify anomalies in transportation network is given by Shekhar et al. [144]. Their methods are based on road network connectivity and temporal neighborhoods based on time series.

RW related techniques. Random walk technique is one of the important building blocks in many applications, including pagerank, keyword extraction, and content-based image retrieval. In these methods, a graph is constructed to represent the data. And a random walk is performed along all the paths on the graph to evaluate the relevance scores of each object. PageRank method [124] is based on the model where a random walker traverses the hyperlinks of a Web graph. Keywords and sentence extraction [108] studies a TextRank model to vote and recommend the important vetices. Recently, random walk method has been explored in data mining research. Hagen et al. [61] proposed a random walk-based method to perform circuit clustering in the netlist graph. Harel and Koren [64] proposed to decompose the data into arbitrarily shaped clusters of different sizes and densities. Moonesinghe et al. [111] introduced an algorithm, called Outrank, to detect outliers by random walk models. Sun et al. [149] constructed a bipartite graph based on random walks with restart to address two issues: neighborhood formation and anomaly detection. Janeja et al. proposed a random walk based Free-Form spatial scan statistic (FS3)[79] to construct a weighted Delaunay Nearest Neighbor Graph (WDNN) to capture spatial autocorrelation and heterogeneity. These applications of random walk methods showed that it can provide an accurate relevance scores between two nodes in a weighted graph.

3.3 Random Walk on Bipartite Graph(RW-BP)

Intuitively, an SNO is an observation that is exceptionally different from its neighbors. One of the most fundamental issues is how to accurately compute the relevance scores among the observations. In this section, RW-BP method is designed to compute such scores by operating RW techniques on a weighted bipartite graph. The main steps of RW-BP are described as follows.

1. **Bipartite graph construction.** The vertex sets in the bipartite graph correspond to the spatial objects

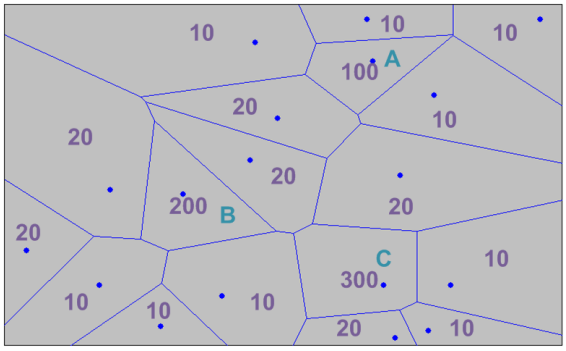


Figure 3.1: Voronoi-based neighborhood formulation

and the clusters generated from the non-spatial attributes of the objects in the spatial database.

2. **Similarity computation between spatial objects.** Random walk is performed on the bipartite graph to compute the similarities of the non-spatial attributes between any pair of the spatial objects.

3. **Neighborhood formulation and outlierness computation.** The spatial neighbor sets for each object can be formed using the Voronoi diagram or *k*NN method. And the outlierness for each object is computed as the differences between itself and its neighborhood.

4. **Outlier identification.** Finally, the outliernesses are ranked in an ascending order and the top *l* objects are identified as spatial outliers.

3.3.1 Modeling Weighted Bipartite Graph

In RW-BP method, the weighted bipartite framework is denoted as $G = \langle P \cup C, E \rangle$, where P is the set of spatial objects, C is the set of clusters generated from the non-spatial attributes of the spatial objects, and E is the set of weighted edges between the spatial objects and the clusters. P and C are two independent sets such that E only exists between them. Constructing such a weighted bipartite graph consists of three fundamental steps. First, non-spatial attributes of the spatial objects are clustered using clustering method. Second, the bipartite graph is constructed in which the left vertex set consists of the spatial objects and the right one consists of the cluster sets. Finally, the edge value is computed based on the non-spatial attributes of the spatial objects and the centroid values of the clusters.

Considering the sample spatial dataset with *18* spatial objects in Figure 3.1, the *K* value (i.e., the number of clusters) equals to *5*. Therefore, the cluster set is *C1(10), C2(20), C3(100), C4(200), C5(300)*. There are *18* spatial objects and *5* separate clusters, and its bipartite graph can be constructed as shown in Figure

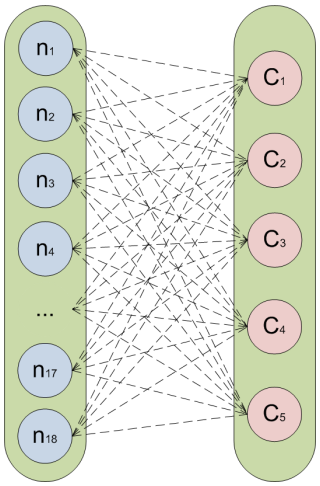


Figure 3.2: Bipartite Framework. The two partitions correspond to spatial objects and clusters

3.2. In particular, the cluster sets are calculated using K-means method in the non-spatial attribute space. The main disadvantage when using K-mean lies in the fact that the optimum *K* value must be pre-specified. To address this issue, a practical approach proposed by Ray et al. [132] is used to experiment with different values of *K* to identify the values that better suit the data set. To generate more accurate results, the non-spatial attributes can be clustered *h* times. And the *K* value at each time is slightly different with that of at the other time, i.e., $K_1, K_2, \cdots, K_h$. The final cluster set in the right part is the union of cluster sets generated individually, i.e., $C = \langle C_1 \cup C_2 \cup \cdots \cup C_h \rangle$. Therefore, in the right part of the bipartite graph, there are $m(= (K_1 + K_2 + \cdots + K_h))$ clusters. For each spatial object p_i in the left part, there will be m edges that connect it with all clusters. In RW-BP method, the edge value in the bipartite graph is defined as the similarities between the spatial object and the cluster, which is shown as follows.

$$E\langle P_i, C_j \rangle = \frac{1}{e^{|Atr(P_i) - Ctr(C_j)|^{\alpha}}}, 0 < \alpha \leq 2 \tag{3.1}$$

where, $Atr(P_i)$ is the non-spatial attribute of the spatial object and $Ctr(C_j)$ is the centroid value of the corresponding cluster. α helps compute more accurate edge value, and is decided by range distribution of the non-spatial attribute values of the whole data set. Normally, when the data values are in a smaller range, α has a larger value, and vice versa.

3.3.2 Similarity Computation Between Spatial Objects

To compute the similarities between spatial objects, RW techniques can be directly applied to the weighted bipartite graph. A random walk means that it starts from node i, and iteratively transmits to its neighborhood with certain probability. At each step, it has the probability c to return to the original node. Random walk

with restarts can be defined as Equation(3.2)[127]:

$$\vec{S_p} = (1-c)W_p\vec{S_p} + c\vec{e_p} \tag{3.2}$$

Where W_p is the NAM (Normalized Adjacency Matrix) of point p. $\vec{e_p}$ is an $(n+m)-$by-1 starting vector. $\vec{S_p}$ is the steady-state probability vector which can describe the **similarity scores** between point p and the other points in the data set. c is known as the damping factor and is normally predefined as 0.1. Based on Equation (3.2), $\vec{S_p}$ can be computed as follows.

$$\vec{S_p} = (1-c)(I - cW_p)^{-1}\vec{e_p} \tag{3.3}$$

Obviously, NAM is a critical factor to compute more accurate solution about vector $\vec{S_p}$. In the following, the procedures of NAM generation are studied step by step.

Normalized Adjacent Matrix (NAM) Construction

The information illustrated by the BP can be stored in a $n-$by$-m$ matrix M, where each entry, $M(i,j)$, is the weight of the edge $\langle i,j\rangle$. The bipartite graph in Figure 3.2 can be represented as follows (α= 1/2).

$$M_{18\times5} = \begin{pmatrix}
1 & e^{-10^{1/2}} & e^{-90^{1/2}} & e^{-190^{1/2}} & e^{-290^{1/2}} \\
1 & e^{-10^{1/2}} & e^{-90^{1/2}} & e^{-190^{1/2}} & e^{-290^{1/2}} \\
 & \cdots & & \cdots & \\
 & \cdots & & \cdots & \\
e^{-10^{1/2}} & 1 & e^{-80^{1/2}} & e^{-180^{1/2}} & e^{-280^{1/2}} \\
e^{-10^{1/2}} & 1 & e^{-80^{1/2}} & e^{-180^{1/2}} & e^{-280^{1/2}} \\
 & \cdots & & \cdots & \\
 & \cdots & & \cdots & \\
e^{-90^{1/2}} & e^{-80^{1/2}} & 1 & e^{-100^{1/2}} & e^{-200^{1/2}} \\
e^{-190^{1/2}} & e^{-180^{1/2}} & e^{-100^{1/2}} & 1 & e^{-100^{1/2}} \\
e^{-290^{1/2}} & e^{-280^{1/2}} & e^{-200^{1/2}} & e^{-100^{1/2}} & 1
\end{pmatrix}$$

As shown in this matrix, the row nodes correspond to the spatial objects and the column ones to the clusters. Intuitively, if two nodes always belong to the same clusters, they have higher similarities. Otherwise, they are very different with each other. Based on the relationship matrix $M_{n\times m}$, we can construct the adjacent

matrix M_p, which is an $(n+m) \times (n+m)$matrix for any spatial object p.

$$M_p = \begin{pmatrix} M_{n\times m}^T & 0_{m\times m} \\ 0_{n\times n} & M_(n \times m) \end{pmatrix} \tag{3.4}$$

	10	**20**	**100**	**200**	**300**
10	1	0.7475	0.0091	1.250e-004	4.658e-006
20	0.7475	1	0.0098	1.305e-004	4.793e-004
100	0.0091	0.0098	1	0.0017	0.0017
200	1.250e-004	1.305e-004	0.0017	1	2.940e-005
300	4.658e-006	4.793e-004	0.0017	2.940e-005	1

Table 3.1: Similarity Computation in RW-BP

Object	**Similarities**	**Rank**
C	4.7251e-005	1
B	1.2772e-004	2
A	0.0091	3
...	...	...
...	0.0134	...
...	...	...
...	1	18

Table 3.2: Outlier Rank in RW-BP

Suppose a walker visits the bipartite graph starting from a random spatial object p_i, the probability of traversing the edge $< p_i, p_j >$ should be in direct proportion to the weight values of all the outgoing edges originating from point p_i. We use the Equation (3.5) to normalize it.

$$W_p(i,j) = M_p(i,j) / \sum_{k=1}^{m+n} M_p(k,i) \tag{3.5}$$

After normalization, the sum of each column in $\vec{W}_p$ is equal to 1.

Similarity Computation

After constructing the NAM, vector $\vec{S}_p$ can be directly computed using Equation (3.3). Before that, we

need to define the vector $\vec{e}_p$. Generally, it is constructed with 1 in the i^{th} row and 0 in the others. Here p is the i^{th} spatial object in $M_{n\times m}$ matrix, then

$$\vec{e_p} = \langle 0_1, \cdots, 1_i, \cdots, 0_n, \cdots, 0_{m+n} \rangle^T \tag{3.6}$$

Here, the subscript character of each entry represents the location of the entry in the vector. For example, 0_1 means that the first entry of the vector is 0. Similarly, 1_i represents that the i^{th} entry of the vector is 1. For the object p_3 in the Figure 3.2, the corresponding starting vector $\vec{e}_3$ can be represented as

$$\vec{e}_3 = \langle 0, 0, 1, 0, \cdots, 0_n, \cdots, 0_{m+n} \rangle^T$$

There, the relevance vector for any specified point p_i can be computed by using the Equation(3.2)or (3.3), that is $\vec{S}_p$. After deriving the relevance vectors of all the points, we can compute the similarities between any pair of spatial objects using Cosine correlation, as shown in Equation(3.7).

$$Sim(p_i, p_j) = \frac{(\vec{S}_{p_i}, \vec{S}_{p_j})}{\sqrt{(\vec{S}_{p_i}, \vec{S}_{p_i})} \cdot \sqrt{(\vec{S}_{p_j}, \vec{S}_{p_j})}} \tag{3.7}$$

3.3.3 Spatial Outlier Identification

Computing the outlier score of any spatial object is to identify the similarity between a specified object and its neighbors. In the example in Figure 3.1, we use Voronoi diagram to determine the spatial neighborhood for each object. Given a set of n points $p_1, p_2, \cdots, p_n$ in the spatial dataset, the Voronoi diagram can be constructed such that each object in the region surrounding the specific object is the closest to that object than any others. For example, for the query point, A, we only need to consider those points whose representative regions border the region of A. Therefore, the neighborhood set of A is $\{n_1(10), n_2(10), n_3(10), n_4(10), n_{10}(20)\}$. Using Equation (3.2) and (3.7), we can identify the similarity between each point and its neighbor. Finally, we can use the geometric mean or arithmetic mean of all the similarity values as the outlier scores for each spatial object. Consider the sample spatial dataset shown in Figure 3.1. Clearly, object A, B, and C are outliers and the rest ones correspond to normal objects. Using RW-BP approach, the non-spatial similarities between each pair of points can be computed. Table 3.1 shows the detailed results.

With the results in Table 3.1, we can determine the relevances(outlierness) between any object and its neighborhood. For example, the outlierness of point C can be computed using the geometric mean value, shown

as follows.

$$\begin{aligned} OutScore(C) &= ((4.658e-006)^3 * (4.793e-004)^3)^{1/6} \\ &= 4.7251e-005 \end{aligned}$$

Repeatedly, we can compute the outlierness values for the other spatial objects. The final outlier scores and the ranking list are described in Table 3.2.

3.3.4 RW-BP Algorithm

Based on the above proposed idea, we generalize the RW-BP algorithm to identify spatial outliers with single attributes in a weighted bipartite graph. The proposed algorithm has *7* input parameters, which are described in Table 3.3.

Parameter	Description
X	A dataset storing the spatial attributes.
Y	A dataset storing the non-spatial attributes.
k	The optimal number of clusters.
r	The pre-defined number of requested outliers.
h	The number of clustering operations on set Y.Generally, $h \leq 10$.
n	The number of spatial objects in the dataset.
c	The damping factor.

Table 3.3: Main Parameters in RW-BP and RW-EC

Algorithm 1 RW-BP SNOD Approach

for $i = 1$ to n **do** {Calculate the neighborhood for each object}
  $Neighbors(x_i) = kNN(X, x_i)$
end for
for $i = 1$ to h **do** {Calculate the cluster sets of nonspatial attribute set Y}
  $C_i = K - mean(Y, k_i)$
end for
$C = \bigcup_{i=1}^{h} C_i$ {Get the overall cluster sets.}
$G = \langle X, C, E(X, C) \rangle$ {Construct Bipartite Graph}
for $i = 1$ to n **do** {Construct the relation matrix of the bipartite graph}
  for $j = 1$ to $|C|$ **do**
    $M(i, j) = 1/e^{|Y_i - Ctr(C_j)|^{\alpha}}$
  end for
end for
$M = \begin{pmatrix} M^T_{n\times m} & 0_{m\times m} \\ 0_{n\times n} & M_(n \times m) \end{pmatrix}$ {Construct adjacent matrix}
$W_{(n+m)\times(n+m)} = ColumnNorm(M_{(n+m)\times(n+m)})$ {Normalize the adjacent matrix}
for $i = 1$ to n **do** {Compute similarity vector for each object}
  $\vec{S_i} = (1 - c)(I - cW_{(n+m)\times(n+m)})^{-1}\vec{e}_i$
end for
for $i = 1$ to n **do** {Compute the relevance scores between specified object and its neighbors}
  for $j = 1$ to k **do**
    $nb = Neighbor(i, j)$ {Get Current Neighbor}
    $Sim(i, nb) = (\vec{S}_{p_i}, \vec{S}_{p_{nb}})/(\sqrt{(\vec{S}_{p_i}, \vec{S}_{p_i})} \cdot \sqrt{(\vec{S}_{p_{nb}}, \vec{S}_{p_{nb}})})$
  end for
end for
for $i = 1$ to n **do** {Compute the Outlierness for each spatial object}
  $OutScores_i = f(Sim_{n\times n}, Neighbors(X, x_i))$
end for
$RankList = RankQueue(OutScores)$ {Rank the objects with the similarities}
$O_r = MaxOutlier(RankList, r)$ {Mark the outliers}

For each data object x_i, the first step is to identify its spatial neighbors, $Neighbors(X, x_i)$. Next, using K-means method, we conduct several clustering on the set of Y. At each loop, we get corresponding cluster set C_i. The overall cluster set is the union of cluster sets, C. We construct a bipartite graph, $G =< X, C, E(X, C) >$, between the spatial datasets and the cluster sets. The edge values between them are computed by the non-spatial attributes and the centroid values of the clusters. With the relationship matrix corresponding to the bipartite graph, we deduce the normalized adjacent matrix which is used in Equation (3.2) to compute the similarity matrix. Cosine similarity equation is also used to compute the final relevance scores between any pair of spatial objects. Finally, outlierness scores $OutScores$ are computed as the differences between the specified objects and their neighbors and the top r objects with the lowest

values are detected as the outliers.

Time Complexity. To form the neighborhood, it will take *O(NlogN)* for Voronoi diagram and *O(logN)*for *kNN*(Space partitioning). When we conduct the clustering on the non-spatial attributes, the time complexity of K-mean method is linear in all relevant parameters: iterations *H*, number of clusters *M*, and number of spatial objects *N*, i.e., *O(IMN)*. Constructing the normalized adjacent matrix has time complexity of *O(NM)*. Calculating the relevance vector for each spatial object costs *O(NlogN)*. Finally, computing the similarity between specified object and its neighbor costs $O(kN^2)$. In summary, assuming *N*>>*M*, *N*>>*k* and *N*>>*I*, the total time complexity of RW-BP approach is $O(N^2)$*(=O(logN)(or O(NlogN))+O(IMN)+O(NM)+O(NlogN)+* $O(kN^2)$*)*.

3.4 Random Walk on Exhaustive Combination (RW-EC)

RW approach is an efficient graph-based technique. It is very powerful to identify the relationship among the points once the graph is well-constructed. In this section, we continue investigating the benefits of RW techniques on spatial outlier detection. Using the spatial and non-spatial attributes of points, we propose another different graph, EC (Exhaustive Combination). The operation of RW techniques on EC constructs another different algorithm, RW-EC (Random Walk on Exhaustive Combination) to identify the spatial outlier. The main steps of RW-EC are described as follows.

1. **Construction of the weighted EC graph.** In EC graph,the vertex set composes of all the spatial objects in the dataset and there is an edge between each pair of spatial objects.

2. **Similarity computation between spatial objects.** Random walk is performed on the EC graph to compute the similarities between any pair of the spatial objects.

3. **Neighborhood formulation and outlierness computation.** Similarly, the spatial neighbor sets are formed by the Voronoi diagram or kNN method. And the outlierness for each object is computed as the similarities between itself and its neighborhood.

4. **Outlier identification.** Finally, the top k objects in the ranked-outlierness list are identified as the spatial outliers.

Actually, RW-EC and RW-BP methods are both the application of RW techniques on the spatial outlier detection based on different weighted graph.They share the same idea. In the following, we introduce the different steps: modeling of the weighted EC graph and construction of normalized adjacent matrix (NAM).

3.4.1 Modeling Weighted EC graph

Given a spatial dataset, the EC graph is constructed with the information of the spatial and non-spatial attributes. For any pair of objects, there is one edge which connects them and the edge value can be computed using Equation(3.8).

$$E\langle P_i, P_j\rangle = \frac{1}{e^{|Atr(P_i)-Atr(P_j)|^{\alpha}}} * \frac{1}{dist(P_i, P_j)}$$
$$0 < \alpha \leq 2 \text{ and } i \neq j \tag{3.8}$$

Normally, $dis(P_i, P_j)$ is decided by the Euclidean Distance. If there is a very large data set, we can consider

	10	**...**	**20**	**...**	**100**	**200**	**300**
10	1	...	0.5207	...	0.7629	0.6200	0.5363
10	0.9076	...	0.5213	...	0.7624	0.6204	0.5367
...	...	...	...	...	...	....	...
20	0.5207	...	0.5213	...	0.7924	0.6477	0.5596
20	0.5131	...	0.5213	...	0.7888	0.6452	0.5574
...	...	..	...	...	...	....	...
100	0.7629	...	0.0.7924	...	1	0.7808	0.6759
200	0.6200	...	0.6477	...	0.7808	1	0.9332
300	0.5363	...	0.5596	...	0.6759	0.9332	1

Table 3.4: Similarities Computation in RW-EC

Object	**Similarities**	**Rank**
C	0.5478	1
B	0.6337	2
A	0.7687	3
...	...	...
...	0.8756	...
...	...	...
...	0.9180	18

Table 3.5: Outlier Rank in RW-EC

only construct partial edges for the sake of efficiency (like $20\% \times |E|$).

3.4.2 Normalized Adjacent Matrix Construction

In RW-EC method, adjacent matrix is an $n-$by$-n$ matrix, where each entry, $M(i,j)$, is the weight of the edge $E\langle p_i, p_j\rangle$. p_i and p_j are two spatial objects. For example, the first row of the adjacent matrix for the EC graph can be represented as follows (α= 1/2).

$$M_{1,1} = \begin{pmatrix} 0 \\ (1/dis(2,1)) \\ ... \\ ... \\ (e^{-(10^{1/2})}/dis(10,1)) \\ ... \\ ... \\ e^{-(90^{1/2})}/dis(16,1)) \\ e^{-(190^{1/2})}/dis(17,1)) \\ e^{-(290^{1/2})}/dis(18,1)) \end{pmatrix}^T$$

In the adjacent matrix in RW-EC, both the row and column nodes correspond to the spatial objects. The NAM is an $n-$by$-n$ matrix and directly constructed by column-normalizing the adjacent matrix.

In the same way, we use Equation (3.2) to compute the similarity vector for each object and then use Equation (3.7) to get the final outlier scores for all spatial objects. Given the same example, the similarities matrix and outlierness vector computed by RW-EC method are given in Table 3.4 and 3.5.

3.4.3 RW-EC Algorithm

RW-EC algorithm is generated in this part and its main input parameters are illustrated in Table 3.3.

In RW-EC algorithm, we compute each edge value during forming the *k*NN neighbors (linear search). And then, the adjacent matrix is constructed based on the edge values. Actually, if the size of the dataset is very large, we can only consider $20 - 50\% \times |E|$ edges. That is, the weights of the first $20 - 50\%$ neighbors are still decided by the spatial and non-spatial attributes, but that of the rest edges is all defined as 0. After normalizing the adjacent matrix, we use the RW techniques to derive the relevance vector for each objects on which the similarities between specified object and its neighbors are computed using the Cosine Similarity. Finally, ranking the outlierness help generate the top *r* outliers.

Algorithm 2 RW-EC SNOD Approach

for $i = 1$ to n **do** {Construct the EC Graph and Calculate the neighborhood}
  for $j = 1$ to n **do**
    $E(i,j) = 1/e^{|Y_i - Y_j|^{\alpha}} * 1/dis(X_i, X_j)$
  end for
  $Neighbors(x_i) = kNN(X, x_i)$
  E
end for
for $i = 1$ to n **do** {Construct the relation matrix of the EC graph}
  for $j = 1$ to n **do**
    $M(i,j) = E(i,j)$
  end for
end for
$W_{n\times n} = ColumnNorm(M_{n\times n})$ {Normalize the adjacent matrix}
for $i = 1$ to n **do** {Compute similarity vector for each object}
  $\vec{S_i} = (1-c)(I - cW_{n\times n})^{-1}\vec{e_i}$
end for
for $i = 1$ to n **do** {Compute the relevance scores between specified object and its neighbors}
  for $j = 1$ to k **do**
    $nb = Neighbor(i,j)$ {Get Current Neighbor}
    $Sim(i, nb) = (\vec{S_i}, \vec{S_{nb}})/(\sqrt{(\vec{S}_{p_i}, \vec{S}_{p_i})} \bullet \sqrt{(\vec{S}_{nb}, \vec{S}_{nb})})$
  end for
end for
for $i = 1$ to n **do** {Compute the Outlierness for each spatial object}
  $OutScores(i) = f(Sim_{n\times k}, Neighbors(X, x_i))$
end for
$RankList = RankQueue(Sim(X))$ {Rank the objects with the similarities}
$O_r = MaxOutlier(RankList, r)$ {Mark the outliers}

Time Complexity. To form the neighborhood, it will take *O(N^2)*for *kNN*(Linear search, which helps construct the EC graph). Constructing the normalized adjacent matrix has the time complexity of *O(N)*. Calculating the relevance vector for each spatial object costs *O(NlogN)*. Finally, computing the similarity between specified object and its neighbor costs $O(kN^2)$. In summary, assuming *N>>k*,the total time complexity of RW-EC approach is $O(N^2)$*(=O(N^2)+O(N)+O(NlogN)+* $O(kN^2)$*)*.

3.5 Experiment Results and Analysis

We conducted an extensive simulation and real datasets to compare the performance among the proposed RW-BP, RW-EC methods, and other related SOD methods proposed in [7, 62, 90, 143, 150].

3.5.1 Simulations

This section studies the extensive simulations to compare the performance between the RW based methods and other related SOD methods. The experimental study followed the standard statistical approach for evaluating the performance of 7 kinds of SOD methods.

Simulation Settings.

Data Set: The simulation data were generated based on a standard statistical model [137] with the decomposition form:

$$Z(s) = \beta + \omega(s) + \epsilon(s) \tag{3.9}$$

where $\beta \sim$ N(0, 1), ω(s) refers to a Gaussian random field with covariogram model *C(h; θ)*, and ϵ(s) refers to measurement error or white noise variation. We considered a popular exponential covariogram model. The exponential model is defined as

$$C(h; b, c) = \begin{cases} b & \text{if } x \geq 0 \\ b(1 - exp(-\frac{h}{c})) & \text{if } 0 < h \leq c \\ 0 & \text{if } h > c \end{cases}$$

where *h* refers to the spatial distance between two sample objects s_i and s_j, the parameter *b* refers to a constant variance for each *Z(s)*, and *c* refers to a valid distance range for nontrivial dependence (or covariance). For the white noise component, we employed the following standard model[39]:

$$\epsilon(s) \sim \begin{cases} N(0, \sigma_0^2) & \text{with probability } 1 - \alpha \\ N(0, \sigma_C^2) & \text{with probability } \alpha \end{cases}$$

There are three related parameters σ_0^2, σ_C^2 and α. ${\sigma_0}^2$ is the variance of a normal white noise, σ_C^2 is the variance of contaminated error that generates outliers, and α is used to control the number of outliers. Note that it is possible that the distribution $N(0, {\sigma_C}^2)$ generates some normal white noises. All true outliers must be only identified based on standard statistical test by calculating the conditional mean and standard deviation for each observation[137]. In the simulations, we tested several representative settings for each parameter, which were summarized in Table 3.6.

Outlier detection methods

Variable	Settings
N	N $\in$ 100, 200. Randomly generate n spatial locations s_i(i $\in$ [1,N])in the range [0,25] $\times$ [0,25]
b,c	b=5; c =5,15,25
β	β_1 N(0,1) and $\beta_i = 0, i = 2, ..., 5$
σ_0, σ_C	${\sigma_0}^2 = 2, 10; {\sigma_C}^2 = 20$
α	$\alpha = 0.05, 0.10, 0.15$
K	$K = 5, 10$

Table 3.6: Combination of Parameter settings

We compared our methods with the state of the art local based SOD methods, including *Z*-test [143], *Scatterplot*[62], *MoranScatterplot*[7], *SLOM*-test[150]and *POD*[90] approach. Our proposed methods are identified as RW-BP and RW-EC approach. The implementations of all existing methods are based on their published algorithm descriptions. **Performance metric:** We tested the performance of all methods for every combination of parameter setting in Table 3.6. For each specific combination, we ran the experiments ten times and then calculated the mean of accuracy for each method. To compare the accuracies of each method, we used the standard ROC curves. For RW-BP approach, the non-spatial attribute set was clustered 6 times (k=6,7,8,9,10,11, respectively). The dumping factor *c* was set as 0.9 and α was set as *2* in both RW-BP and RW-EC.

Detection Accuracy

We compared the outlier detection accuracies of different methods based on different combinations of parameter settings as shown in Table 3.6. Six representative results are displayed in Figure 3.3. Obviously, *RW-EC* and *RW-BP* have very preceding performance increases. RW-based methods achieved 20-30 % improvement over POD and SLOM methods, 40-50 % over Moran-Scatterplot method and 60-70 % over Scatterplot method. Compared with RW-EC, RW-BP is slightly more outperforming.

Meanwhile, *Z*-value test has also very impressive performance on the simulation. *Z*-value is under the null hypothesis stating that the data fits a normal distribution. It computes the mean and standard deviation of the entire dataset to compute the outlierness for each object. As mentioned above, since our simulation data is just generated from standard normalized distribution, there is no doubt that *Z*-value is one of the most appropriate methods for the simulation data. Figure 3.3 depicts that ROC curves derived from RW-based methods have very similar trend with that of *Z*-value method. In a sense, RW based approach can accurately detect the outliers in the dataset with normal distribution although they don't make such hypothesis.

When being utilized into a real dataset with unknown distribution, *Z*-value may not shown such outperforming performance since many datasets do not conform to normal distribution. By contrast, RW based

technique is more practical since it doesn't need to assume any distribution of the data. Its effectiveness has been shown in varieties of real applications [61, 64, 79, 108, 111, 124, 149]. In the following, we will demonstrate their competitive performances by applying them into a real dataset.

3.5.2 Experiments on Real Dataset

In this section, we present the experimental results on the real datasets to further demonstrate the accuracy of the proposed RW-based approaches.

The Real Dataset: The Fair Market Rents data was used for outlying objects identification, which we aimed to find counties whose rental prices were very different from counties in its neighborhood. The Fair Market Rents data was provided by the Policy Development and Research, U.S. Department of Housing and Urban Development (PDR-DHUD). It included the rental prices for apartments of different kinds varying from one-bedroom to four-bedroom apartments in 3000+ counties of the US. The location of each county was determined by the longitude and latitude of its center. The neighboring counties were determined by the *kNN* method.

Parameter Setting in *RW-BP* approach: The dataset was clustered for 6 times ($h = 6$) with six different k values: 8, 10, 12, 14, 16 and 18, respectively. The dumping factor (c) was set to 0.9, a value which was commonly used by other approaches [108, 124, 149], and α was set to *1/2*.

Z-Value	SLOM	ScatPlot	M-ScaPlot	POD	RW-BP	RW-EC
Nantucket(MA)	**Blaine(ID)**	KingGeorge(VA)	**Blaine(ID)**	**Blaine(ID)**	**Blaine(ID)**	**Nantucket(MA)**
Pitkin(CO)	**Teton(WY)**	Plymouth(MA)	**Teton(WY)**	**Teton(WY)**	**Summit(UT)**	**Blaine(ID)**
Summit(UT)	Lubbock(TX)	**Blaine(ID)**	Elbert(CO)	**Summi(UT)**	**Teton(WY)**	**Suffolk(MA)**
Orange(CA)	**Summit(UT)**	Caroline(VA)	Surry(VA)	**Suffolk(MA)**	**Suffolk(MA)**	**Teton(WY)**
Blaine(ID)	Pennington(SD)	Howard(MD)	LaPaz(AZ)	**Coconino(AZ)**	**Fairfield(CT)**	**Fairfield(CT)**
Clarke(VA)	Hughes(SD)	Kern(CA)	Kanabec(MN)	**Fairfield(CT)**	**Coconino(AZ)**	**Summit(UT)**
Suffolk(MA)	**Dane(WI)**	**Teton(WY)**	Dorchester(MD)	**Nantucket(MA)**	**Nantucket(MA)**	Mono(CA)
Frederick(MD)	Boone(MO)	**Summit(UT)**	Sumter(FL)	**Dane(WI)**	**Pitkin(CO)**	St.Mary's(MD)
Coconino(AZ)	Yellowstone(MT)	SanJoaquin(CA)	Blanco(TX)	**Pitkin(CO)**	**Dane(WI)**	**Coconino(AZ)**
Ventura(CA)	Codington(SD)	Worcester(MA)	Sussex(VA)	Eagle(CO)	Rockingham(NH)	SanBenito(CA)

Table 3.7: Top 10 spatial outliers with single attribute detected by seven different approaches

Detection of spatial outliers. We applied seven different algorithms to the Fair Market Rent data, including *Z-Value, Scatterplot, Moran-Scatterplot, POD, SLOM* and *RW-EC, RW-BP* approaches. For all the methods, *k* was set to 10 to compute the neighborhoods. Table 3.7 depicts the top ten outlying counties based on the one-bedroom rent in 2005.

As shown in Table 3.7, POD, RW-BP and RW-EC outperform other approaches. They identify the true outliers (like Blaine(ID), Fairfield(CT), Summit(UT), etc) although the outliers are not ranked in the same order. Compared with these three methods, Z-value tends to miss some true outliers, like ST.Mary's(MD)(as shown in Table 3.8) and FairField(CT), etc.

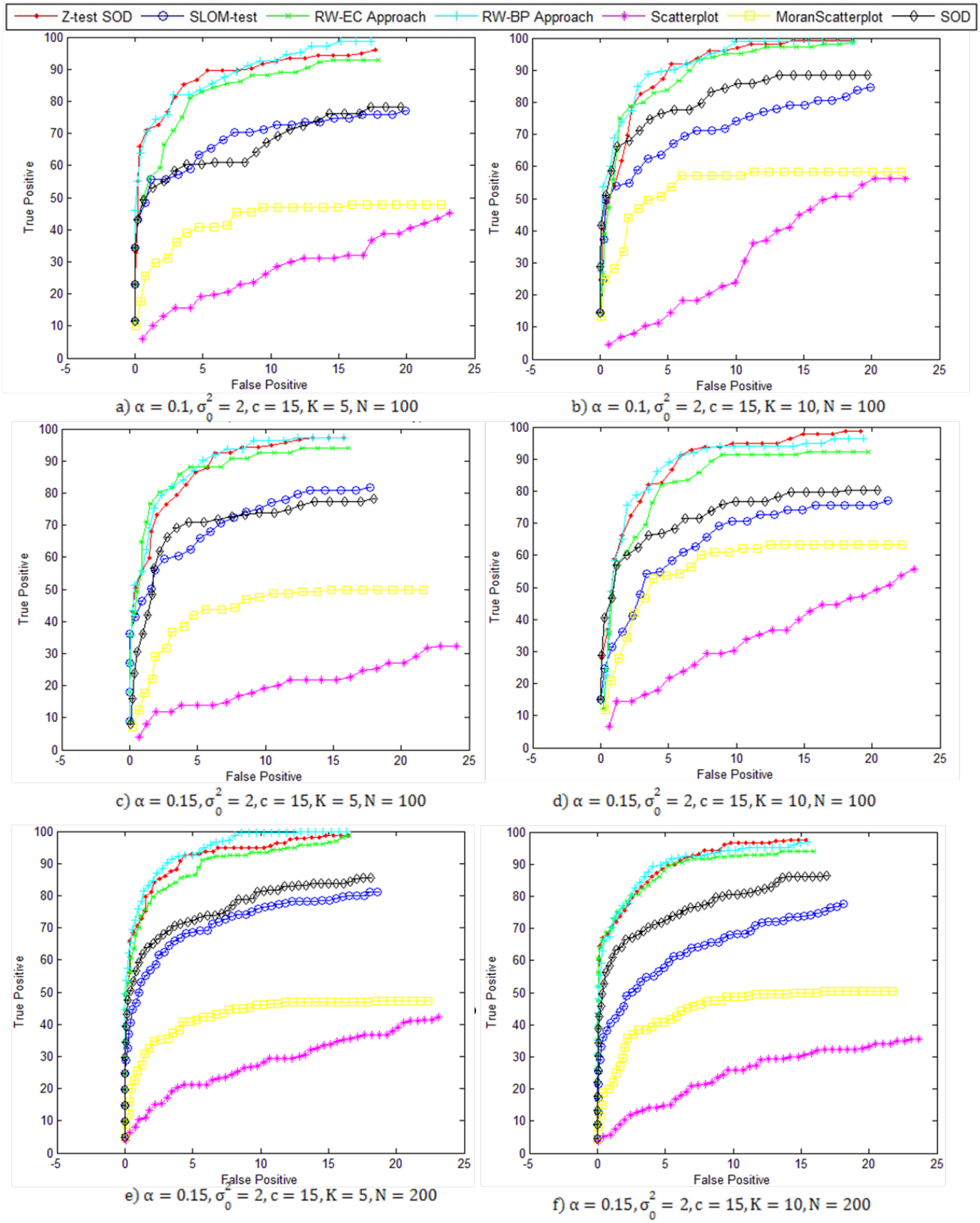


Figure 3.3: Outlier ROC Curve Comparision (the same setting; n =100, b=5, c=5)

CountyName	Rent	Latitude	Longitude
St.Mary's(MD)	702	-76.5976	38.2939
Calvert(MD)	1045	-76.5177	38.5061
Westmoreland(VA)	496	-76.8321	38.1556
Richmond(VA)	496	-76.733	37.9266
Charles(MD)	1045	-76.9723	38.5221
Northumberland(VA)	496	-76.3721	37.8674
Essex(VA)	496	-76.9066	37.9162
King(VA)	611	-77.1525	38.2918
Lancaster(VA)	496	-76.4502	37.6974
Dorchester(MD)	451	-75.9839	38.5466
King(VA)	496	-76.8984	37.7005

Table 3.8: ST.Mary's county

ST. Mary's (MD) is identified as the 8^{th} outlier by RW-BP. Table 3.8 gives the rental prices of the county and its neighbors. As we can see, the rents of some neighbors (such as, *Calvert(1045) and Charles(1045)*) are much higher and the others(*Westmoreland(496),Richmond(196), Northumberland(496), etc*) are much lower. Intuitively, the rent in ST.Mary's is very different with those of its neighbors. However, such outlying behavior cannot be detected by Z-Value, SLOM, scatterplot and Moran scatterplot. This is due to their intrinsic properties when identifying the outlying behavior. For example, Z-Value identifies the outliers by normalizing the difference between a spatial object and **the average of its spatial neighbors**. Moran scatterplot detects the spatial outliers by normalizing the attribute values against **the average values of the corresponding neighborhood**. Averaging the rents of the neighbors neutralizes such significant differences. RW-based approaches address this issue since they accurately compute the similarities among spatial objects on which the outlierness is identified. SanBenito(CA) being identified as the 10^{th} by RW-EC and Rockingham being identified as the 10^{th} by RW-BP are the same case and the information is shown in Table 3.9 and 3.10.

RW based methods can also avoid identifying the false outliers. As can be seen from Table 3.7, 80 % outliers identified by RW-BP and RW-EC are also identified by other approaches. Put differently, what RW based methods identified are true outliers. On the contrary, SLOM, Scatterplot and Moron-Scatteplot not only miss some true outliers, but incorrectly recognize some not very outlying points as *true* outliers. For example, Yellowstone(MT) (Table 3.11) by SLOM approach and Dorchester(MD) (Table 3.12) by Moran-Scatterplot approach.

CountyName	Rent	Latitude	Longitude
SanBenito(CA)	824	-121.2888	36.7458
Monterey(CA)	931	-121.529	36.4507
Santa(CA)	1111	-121.9738	37.0023
Merced(CA)	536	-120.6741	37.2458
Santa(CA)	1107	-121.9128	37.3065
Stanislaus(CA)	645	-120.9588	37.6138
San(CA)	635	-121.2813	37.946
Alameda(CA)	1132	-122.0962	37.7167
Madera(CA)	556	-120.0324	37.0351
San(CA)	1305	-122.3319	37.531
Fresno(CA)	556	-119.9035	36.6384

Table 3.9: SanBenito county

CountyName	Rent	Latitude	Longitude
Rockingham(NH)	750	-71.0776	42.9629
Strafford(NH)	648	-70.9761	43.2583
Essex(MA)	878	-70.9708	42.6355
Hillsborough(NH)	605	-71.5827	42.8956
Middlesex(MA)	884	-71.2756	42.4591
Suffolk(MA)	1120	-71.0735	42.3349
York(ME)	577	-70.6632	43.4458
Merrimack(NH)	624	-71.6373	43.2777
Belknap(NH)	592	-71.4361	43.5152
Norfolk(MA)	914	-71.1544	42.1992
Carroll(NH)	564	-71.1816	43.8226

Table 3.10: Rockingham county

Taking county Dorchester(MD) as an example, most of its neighbors have nearer value. Therefore, it should not be identified as a spatial outlier.It is identified as outlier by Moran-Scatterplot approach mainly because Calvert(MD), one of its neighbors, has higher rent $\{1045\}$ and significantly raises the average rent of the neighborhood. Random walk based method can avoid such problem since it considers not only the relationship with neighborhood when generating the relevance vectors, but the non-spatial attribute distribution of the whole dataset.

Another important issue of existing approaches is the way of identifying the outlierness. They compute the inconsistencies between each object and its neighbors without considering the values of identified object

CountyName	Rent	Latitude	Longitude
Yellowstone(MT)	452	-108.4607	45.8165
Musselshell(MT)	398	-108.3922	46.5546
Carbon(MT)	405	-109.0876	45.3132
Golden(MT)	398	-109.1253	46.3904
Stillwater(MT)	398	-109.3663	45.6301
Big(MT)	398	-107.4838	45.5101
Petroleum(MT)	398	-108.2901	47.0005
Treasure(MT)	398	-107.2915	46.2544
Big(MT)	417	-108.0671	44.5374
Park(MT)	428	-108.999	44.569
Sweet(MT)	398	-109.9178	45.8554

Table 3.11: Yellowstone county

CountyName	Rent	Latitude	Longitude
Dorchester(MD)	451	-75.9839	38.5466
Talbot(MD)	575	-76.1138	38.769
Caroline(MD)	513	-75.8308	38.8752
Wicomico(MD)	576	-75.5945	38.3773
Somerset(MD)	460	-75.7688	38.1057
Queen(MD)	750	-76.0995	39.0478
Calvert(MD)	1045	-76.5177	38.5061
Sussex(DE)	548	-75.3423	38.6514
St.Mary's(MD)	702	-76.5976	38.2939
Kent(DE)	598	-75.5603	39.0927
Kent(MD)	558	-76.0537	39.2605

Table 3.12: Dorchester county

and its neighbors, which may lead to an inaccurate ranking list. The typical method is POD approach which first constructs a graph by assigning the non-spatial attribute differences as edge weights, and then continuously cuts high-weight edges to identify isolated points. Figure 3.4 depicts such issue by comparing two county Eagle(CO) and St.Mary's(MD). Actually, St.Mary's(MD) is ranked as 17^{th} by POD. If we evaluate their outliernesses only by considering the direct differences between the detected object and its neighbors as POD method does, county Eagle will have a little more higher value than county St.Mary's since $[Diff(Eagle) = Avg(117 + 366 + 262 + 199 + 295 + 250) = 256] > [Diff(St.Mary's) = Avg(343 + 206 + 206 + 206 + 206 + 191) = 226]$. However, intuitively, county St.Mary's is more outlying since most non-spatial attributes of itself and its neighbors are not high ([400, 750]). By contrast, those of Eagle is higher ([600, 950]). The difference around 200 makes St.Mary's more outlying and it should

be ranked higher than county Eagle. This issue may also result in identifying false outliers sometimes. In

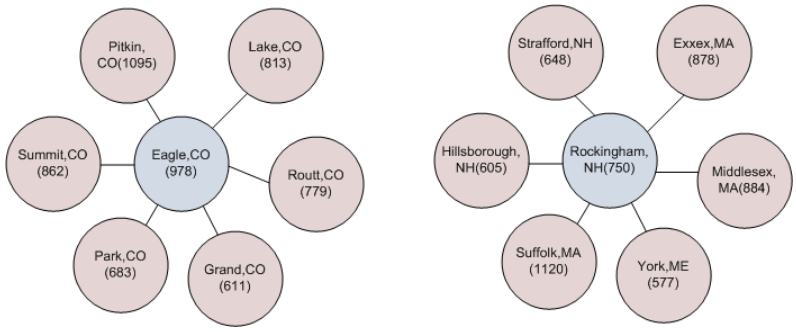


Figure 3.4: Example of two spatial objects

this regard, RW based approaches do better than other methods, including POD. This is because they utilize the Cosine similarity to identifying the outlierness, which means it takes the relationship between these two points and all other points into consideration. Actually, RW-BP does even better than RW-EC since RW-BP also integrates the relationship between any specified object and the clusters into the construction of adjacent matrix before deriving the relevance vector. Although POD performs as well as RW-BP and RW-EC in the real data, the worse results influenced by such issue have been demonstrated by the simulations.

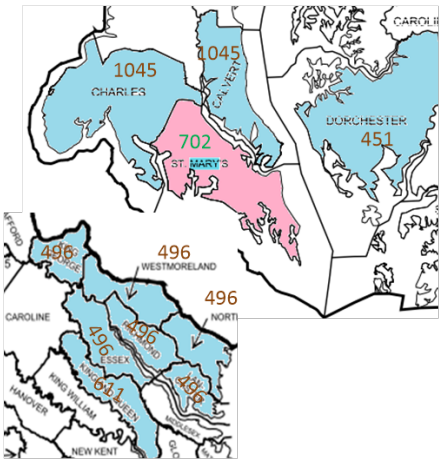


Figure 3.5: Case 1: Masking Problem incurred by SLOM

3.6 Conclusion

In this chapter, we propose two spatial outlier detection approaches based on RW techniques: RW-BP and RW-EC approaches. In these methods, two kinds of weighted graphs, a Bipartite graph and an Exhaustive

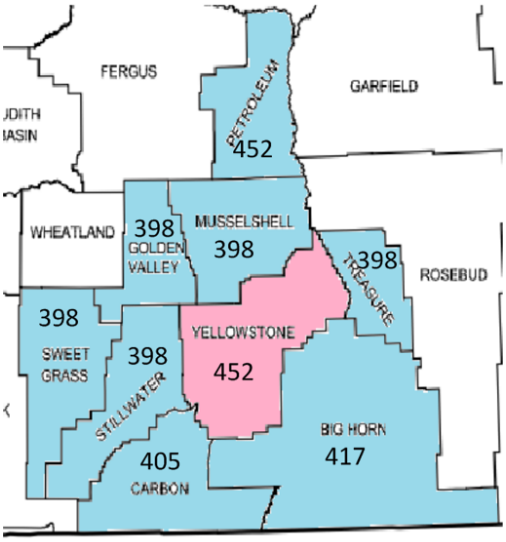


Figure 3.6: Case 2: Swamping Problem solved by RW-SNOD approach

combination, are constructed based on the spatial and/or non-spatial attributes of the spatial objects in the dataset. Secondly, RW techniques are utilized on the graphs to compute the outlierness for each point. The top l objects with higher outlierness are recognized as outliers. The proposed algorithms have three major advantages compared with the existing SOD methods: capable of avoiding the masking and swamping problems and detecting identifying more correct ranking lists. The experiments conducted on the synthetic and real datasets demonstrated the RW based methods significantly outperformed other approaches.

Future efforts could focus on extending this works on spatial categorical dataset and identifying the number of anomalies by integrating the entropy theory.

Chapter 4

An Entropy-Based Method for Assessing the Number of Spatial Numerical Outlier

As introduced by the above two chapters, detecting spatial outliers is an important topic in the field of spatial data mining. A major limitation associated with the existing outlier detection algorithms is that they generally require a pre-specified number of spatial outliers. Estimating an appropriate number of outliers for a spatial data set is one of the critical issues for outlier analysis. This chapter proposes an entropy-based method to address this problem in spatial numerical domain. Based on the relationship between outliers and the overall entropy, that is, the data set with more outiers has a higher entropy value than that with less outliers, we expect that, by incrementally removing outliers, the entropy value will decrease sharply, and reach a stable state when all the outliers have been removed.

This chapter is organized as follows. Section 4.1 surveys the related work. Section 4.2 presents fundamental concepts, including the entropy and the characteristics of SNOs. Section 4.3 proposes an approach to assessing the optimal number of SNOs. Section 4.4 presents experimental evaluation on a real dataset by applying our method to the POD approach [91], and discusses the empirical results. Finally, section 4.5 provides some concluding remarks.

4.1 Backgrounds and related works

Recently, numerous SNOD approaches have been proposed. A major limitation associated with them is that they all assume a pre-specified number of SNOs, typically 5% of the entire dataset. However, there is a well-known problem of masking and swamping effects for SNOD. Masking occurs when true outliers

are not accurately identified; swamping happens when some normal objects are erroneously flagged as outliers [63]. This detrimental effects can be alleviated by determining an appropriate number of outliers. In this regard, estimating an optimal number of SNOs has become one of the most essential issues in outlier analysis. Recently, several works have been focused on assessing an appropriate number of clusters and outliers. Celeux *et al.* proposed entropy-based criterion, Normalized Entropy Criterion (NEC), to estimate the number of clusters associated to a mixture model [25]. This entropy criterion is derived from the relation between the mixture model and cluster analysis. lu *et al.* introduced a new evolutionary algorithm for identifying the optimal number of clusters [102]. They defined an entropy-based fitness function to measure how ell the mixture model fits the data sample. Barbara *et al.* studied the connection between clusters and the entropy, and concluded that the clusters with similar points have lower entropy values than those with dissimilar ones [12]. They designed a clustering algorithm, named as COOLCAT, which groups points within clusters by trying to minimize the expected entropy value of clusters. This algorithm can also be utilized to identify the optimal number of clusters. Nikhil *et al.* proposed a new concept of probabilistic entropy based on the exponential behavior of information gain [125]. Beghdadi *et al.* presented a nonlinear-noise filtering method based on the entropy definition [13].

4.2 Preliminary Concept

This section introduces the theoretical concepts of the proposed approach, including entropy, spatial local contrast, spatial local contrast probability, and spatial local contrast entropy.

Entropy: Entropy is the measure of information and uncertainty of a random variable[13, 139]. If X is a discrete random variable, $S(X)$ is the set of possible distinct value that X can take, and $p(X)$ is the probability function X, the entropy $E(x)$ can be defined as follows:

$$E(X) = -\sum_{X \in S(X)} p(X) \log(p(X)) \tag{4.1}$$

Spatial Local Contrast: The spatial local contrast can be viewed as the differences between an object and its surrounding neighbors. Generally, it is directly related to its spatial outlierness. Given a point object i, the center of an area A_i(A_i includes the point object and its surrounding k neighbors), its spatial outlierness value can be represented by O_i. A spatial location contrast D_i is the function of the outlierness O_i:

$$D_i = f(O_i) \tag{4.2}$$

Spatial Local Contrast Probability: Based on the concept of "Spatial Local Contrast", we define "Spatial Local Contrast probability" as follows:

$$P_i = \frac{D_i}{\sum_{i=1}^{n} D_i} = \frac{f(O_i)}{\sum_{i=1}^{n} f(O_i)} \tag{4.3}$$

In Equation 4.3, n is the number of objects in the spatial dataset. A zero spatial local area, i.e., a homogeneous region, corresponds to a zero probability. Based on the definition of spatial local contrast probability, spatial local contrast entropy can be computed.
Spatial Local Contrast Entropy(SLCE): Spatial local contrast entropy can be formalized as:

$$H = -\sum_{i=1}^{n} (\frac{f(O_i)}{\sum_{i=1}^{n} f(O_i)} \log(\frac{f(O_i)}{\sum_{i=1}^{n} f(O_i)})) \tag{4.4}$$

Motivated by the fact that the spatial local contrast probabilities of outliers are higher than those of other data points, we infer that an outlier point significantly contributes to the spatial local contrast entropy because its spatial local contrast probability is high. The fundamental concept of the proposed technique is that when spatial outliers are incrementally removed from the data set, the spatial local contrast entropy value will be continuously decreased until it reaches a stable state when all the outliers have been removed.

4.3 Proposed Approach

Previous analysis shows that the spatial local contrast entropy will be zero when a dataset is homogeneous. In contrast, the spatial local contrast entropy will be very high if there are a number of outliers in a data set. The outlying objects substantially contribute to the spatial local entropy because the spatial local contrast probabilities of outliers are high. Therefore, if we plot a figure in which the x-coordinate denotes the number of removed outliers and the y-coordinate denotes the spatial local contrast entropy value, we expect a curve which decreases quickly, and then reaches a point where the spatial local contrast entropy value is relatively stabilized. This point is an appropriate estimate for the optimal number of spatial outliers.

4.3.1 The Sliding Window

Removing a point from the data sets will affect the spatial local contrast entropy. This is because the size of the spatial data set has been changed. For example, if the original size is n, it will become $(n-1)$ after removing one point. In this regard, the comparison of entropy values between the data sets of different size is not legitimate.

To address this issue, we introduce the concept of sliding window, which sets a fixed value for dataset as m. That is, after removing one outlier, we compute the spatial entropy of the m point objects. Figure 4.1 shows the concept of applying the sliding window concept to our approach.

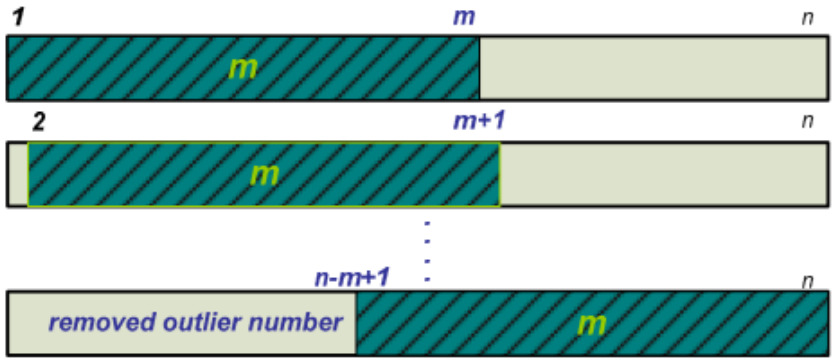


Figure 4.1: A sliding window

First, sort the data based on their outlierness values by a descending order. Suppose there are n data objects in the entire data set. We set a sliding window with m point objects, which specifies the computation of the spatial local contrast entropy.

Second, compute the spatial local contrast entropy from the 1^{st} point to the m^{th} point. The spatial local Local Contrast Entropy(SLCE) is computed as follows:

$$H_0 = -\sum_{i=1}^{m} P_i \log P_i \tag{4.5}$$

After removing the first outlier whose outlierness value is the highest, we reorganize the strucuture of the remaining dataset to reflect the neighborhood change due to the removal of the outlier. We then recompute the spatial local contrasts and their corresponding spatial local contrast probabilities. The sliding window is shifted by one object, that is, the current window is from the 2^{nd} point to the $(m+1)^th$ point. The corresponding SLCE is calculated as follows:

$$H_1 = -\sum_{i=2}^{m+1} P_i \log P_i \tag{4.6}$$

After removing the second outlier, the window is shifted again and the spatial entropy is computed from the 3^{rd} point to the $(m+2)^{th}$ point:

$$H_2 = -\sum_{i=3}^{m+2} P_i \log P_i \tag{4.7}$$

Continue this procedure till the final spatial local entropy is aggregated from the $(n-m+1)^{th}$ point to the n^{th} point:

$$H_{n-m} = -\sum_{i=n-m+1}^{n} P_i \log P_i \tag{4.8}$$

4.3.2 Algorithm description

This subsection introduces the proposed algorithm, named as SLCE, to compute the Spatial Local Contrast Entropy, and discusses the major characteristics of our approach. The input and output of SLCE algorithm are descried in Table 4.1. Algorithm 3 describes this approach as the following 4 steps.

Input	
X	A dataset storing spatial and non-spatial attributes.
$points$	A dataset storing outlier values.
k	The number of neighbors.
m	the size of sliding windows.
kNN	the set of the neighborhood relationship.
$O(x_i)$	An outlierness function
Output	
H	The sum of spatial entropy.

Table 4.1: Input and Output in SLCE

Algorithm 3 Spatial Local Contrast Entropy (SLCE)

$n = size(points)$; {**Step 1: Sort spatial points in a dataset**}
$Y = sort(points)$
for $i = 1$ to $n - m$ **do**
  $outlierID = outliers(i, 1)$;{Retrieve the top outlier based on outlierness values}
  $ruleoutOutlier(outlierID)$; {Remove the outlier from kNN}
  $kNN = adjustNeighbors(kNN)$;{**Step 2: Recompute the k nearest neighbor set**}
  for $j = 1$ to $n - i$ **do** {**Step 3: Calculate the spatial local contrast**}
    $D(x_j) = f(O(x_j))$;
  end for
  for $j = 1$ to m **do** {Calculate the spatial local contrast probability}
    $P(x_j) = \frac{D(x_j)}{sum(D)}$;
  end for
  while $getNode(outlier, j, m)$ **do** {**Step 4: Calculate the SLCE of the sliding window**}
    $H(j) = H(j) + (-1)\log_2(P(x_j))$;{get the $j^{th} outlier from the dataset$}
  end while
end for
$Output(H)$;

The outlierness function $O(x_i)$ is computed based on different approaches to identifying spatial outliers. Given a spatial data set $X = x_1, x_2, ..., x_n$, an outlierness function $O(x_i)$, two positive integer number k (the number of neighbors) and m(the size of sliding window), the major steps in this algorithm are described as follows:

Step 1:Sort spatial points based on outlierness values

Compute a data set $Y = y_1, y_2, ..., y_n$ by decendingly sorting spatial points based on their corresponding outlierness values.

Step 2: Recompute the nearest neighbor set

Remove the top outlier from the current list, and for each spatial point x_i, recompute the k nearest neighbor set $NN_k(x_i)$.

Step 3: Compute local contrast probability

Let $D(x_i)$ and $P(x_i)$ denote the spatial local contrast and spatial local contrast probability of a point x_i, respectively. They are both the functions of $O(x_i)$. That is, $D(x_i) = f(O(x_i))$, and $P(x_i) = \frac{O(x_i)}{\sum f(O(x_i))}$ for $i = 1, 2, ..., m$.

Step 4: Compute the spatial entropy

Using the $getNode$ function, retrieve all the outliers in the sliding window. Based on the $D(x_i)$ and $P(x_i)$, calculate the spatial local contrast entropy: $SLCE = -\sum(\frac{f(O(x_i))}{\sum f(O(x_i))}\log_2(\frac{f(O(x_i))}{\sum f(O(x_i))}))$.

Step 5: Repeat step 2 to step 4 for $(n-m)$ iterations.

From the original data set, we compute the spatial local contrast entropy H_0, from the 1^st point to the m^{th} point. After removing the first outlier, we reorganize the data set, and compute the $NN_k(x_i)$ and $O(x_i)$ for those points with neighborhood updates. We can then calculate the next spatial entropy H_1 from the 2^{nd} point to the $(m+1)^{th}$ point. We continue this procedure until the last spatial entropy H_{n-m} is computed from $(n-m+1)^{th}$ point to the n^{th} point. Based on the set of SLCE values, we plot a curve in which the number of removed outliers is denoted by the x-axis and the corresponding entropy value is denoted by the y-axis. From this curve, we assess the cut-off point to estimate an optimal number of spatial outliers.

4.4 Experiment Results and Analysis

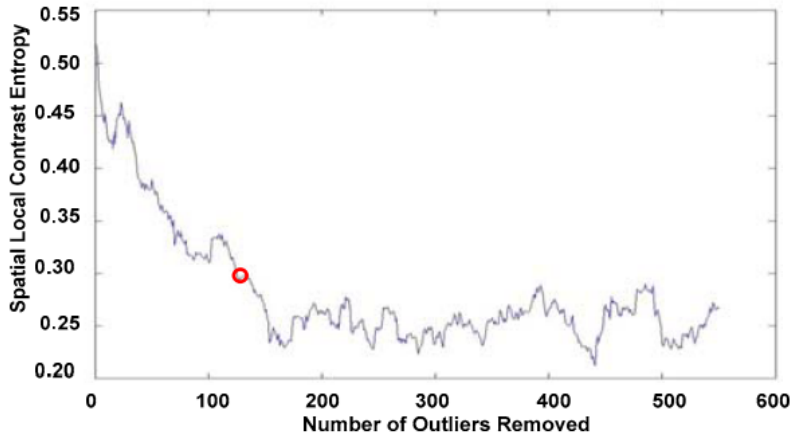


Figure 4.2: Single attribute, SLCE curve (k=8, m=50)

In this section, we present experimental results on a real data set, Fair Market Rents data, provided by the PDR-DHUD(Policy Development and Research, U.S. Department of Housing and Urban Development). The data set includes the rental prices for efficiencies, one-bedroom apartments, two-bedroom apartments, three-bedroom apartments, and four-bedroom apartments in 3000+ counties of the U.S.. In the experiment,

the outlierness function is defined by the POD method[91], which is a graph-based method to identify the spatial outliers. Note that our method can also be applied to other spatial outlier detection methods. The

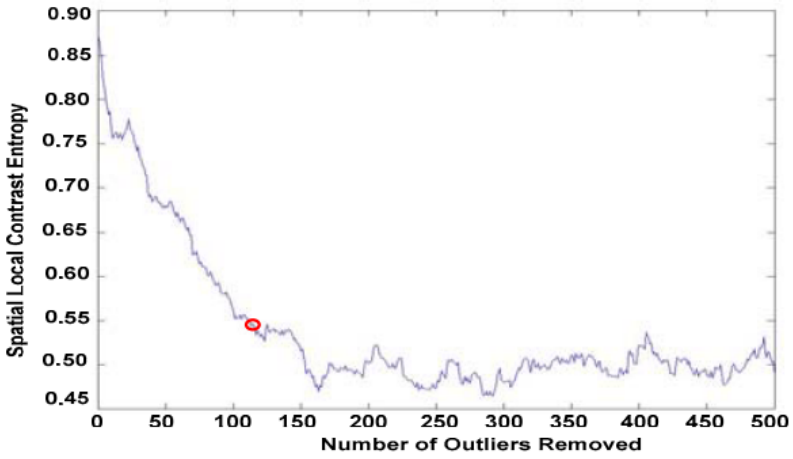


Figure 4.3: Single attribute, SLCE curve(k=10, m=100)

POD method first constructs a graph based on the k nearest neighbor relationship in the spatial domain, assigns the non-spatial attributed differences as edge weighs, and then continuous cuts high weight edges to identify isolated points or regions that are dissimilar to their neighboring objects as spatial outliers. The experiment is composed of two components, including single attribute(the rental prices for one-bedroom apartments) and multiple attributes (the rental prices for one-bedroom apartments, the rental prices for two-bedroom apartments, the rental prices for three-bedroom apartments and the rental prices for four-bedroom apartments).

4.4.1 Experiment on spatial dataset with single and multiple attributes

In the experiment in spatial dataset with single attribute, we set m, the size of the sliding window, as 50, 100 and 150, respectively, and k, the number of the nearest neighbors, as 8 and 10. Figures 4.2-4.4 shows the generated curves by our algorithm for the single attribute dataset. As can be seen, the values of spatial entropy decreases in the beginning and reaches a stable state at a certain point, which shows an estimate of the optimal number of outliers is around 120. When considering the multiple attributes of the spatial objects, the value of m is set as 50, 100 and 150, respectively, and k is equal to 10. Figures 4.5-4.7 were generated by the SLCE algorithm for multiple attributes. We can observe a similar precipitating trend in which an estimated outlier number is around 90.

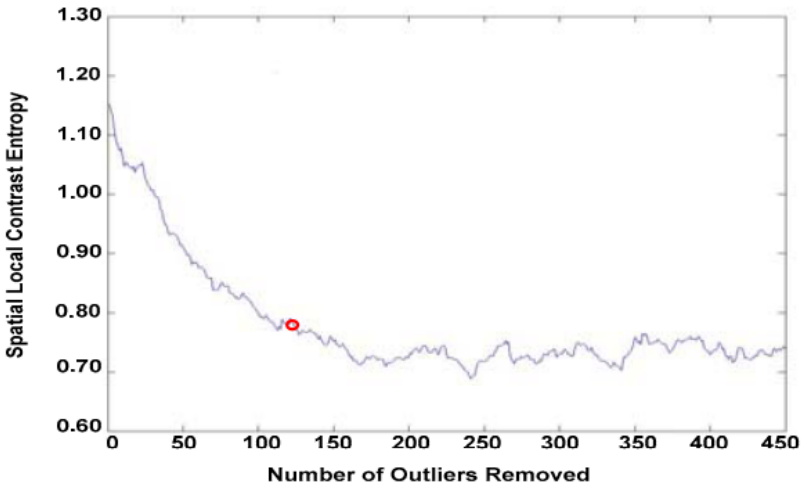


Figure 4.4: Single attribute, SLCE curve(k=10, m=150)

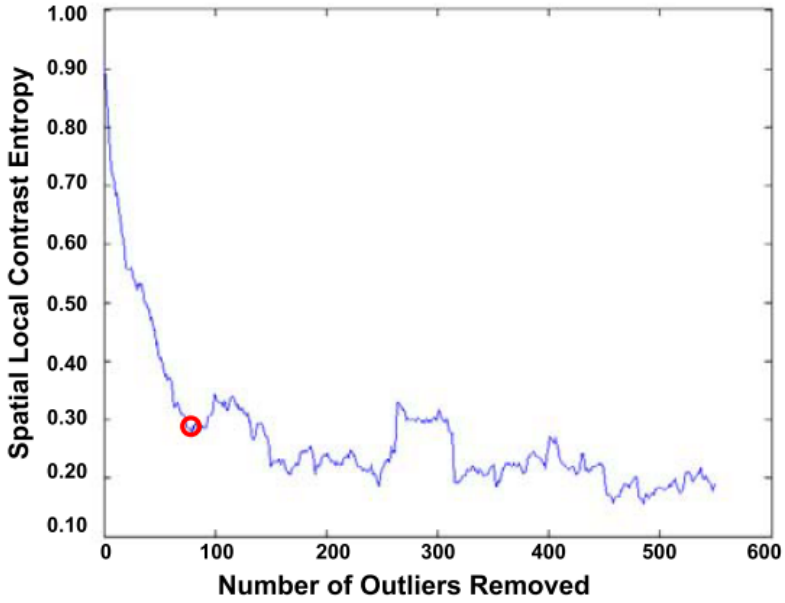


Figure 4.5: Multiple attributes, SLCE curve (k=10, m=50)

4.4.2 Analysis of Experiment Results

As shown in Figures 4.2-4.7, for both cases of single attribute and multiple attributes, the dominant patters as exhibited in the experiment is consistent with the theoretical analysis discussed in Section 4.4. That is, there does exist an "reflexion" point which corroborates the relationship between the spatial local contrast entropy and the number of outliers removed. Because the object with the highest outlierness value will

contribute most to the spatial local contrast entropy, we can greedily reduce the overall entropy value by incrementally removing the outlying object, whose local inconsistence value is the highest. When all outliers

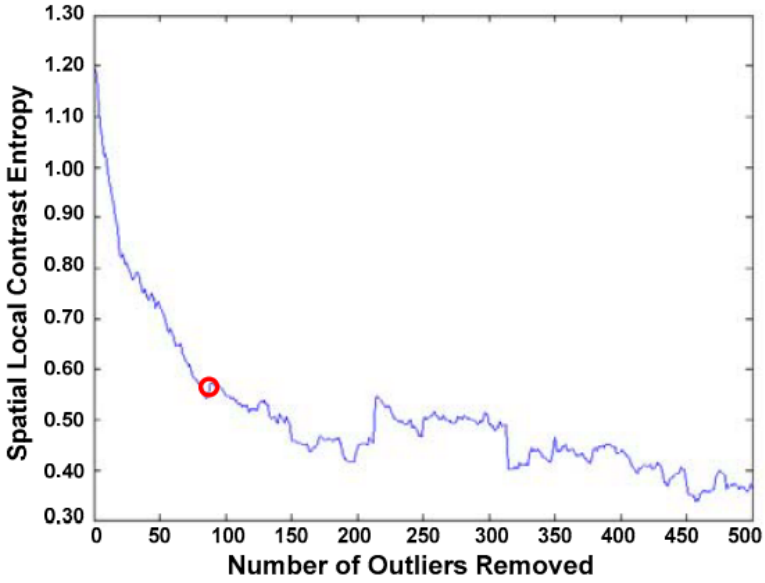


Figure 4.6: Multiple attributes, SLCE curve (k=10, m=100)

have been removed, if we continue to remove normal objects, the spatial local contrast entropy will not change significantly. It is reasonable to assume that the outlierness values of normal objects are randomly distributed. Then the removal of normal objects based on the ranks of their corresponding outlierness values will not drastically disturb the background distribution. Therefore, the background of spatial local contrast entropy will keep persistent. Take Figure 4.2 as an example, we can observe that the estimate of an optimal number of outliers is around 120. When the number of candidate outliers removed is greater than 120, the spatial local contrast entropy indicates a constant mean value around 0.25. The trembling of the curve is due to a normal variance.

Ideally, the SLCE curve will monotonously decrease with a continuously decreasing slope until the spatial local contrast entropy becomes stable when all the outliers have been removed. However, the outlierness value of each object, which is approximated based on the POD method, may not be identical to the true outlierness value. That is, during each incremental process, the object removed is not certainly to the most "significant" outliers. Therefore, the slope will not monotonously decrease in strict relation to the number of outliers removed. Considering the case of false positives, in which normal objects are misidentified as outliers, potentially the ratio of outliers to normal objects will be erroneously increased, and hence will the spatial local contrast entropy. As a result, the curve will not have the distinct characteristic of decreasing monotonicity. It is reasonable that the SLCE curve may have some small jumps before the removal of all

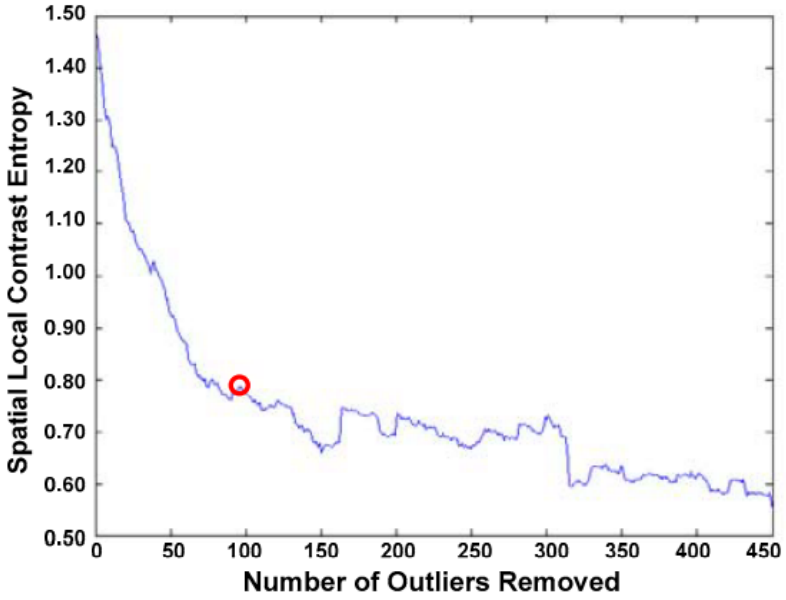


Figure 4.7: Multiple attributes, SLCE curve (k=10, m=150)

outliers. As shown in Figure 4.4, when the number of outliers removed is around 22, the spatial local contrast entropy stops decreasing and incurs a small jump, but it quickly resumes the decreasing trend when the number if larger than 25. This pattern reveals that, it is necessary to evaluate the curve entirely, in order to estimate a close-to-optimal number of outliers.

4.5 Conclusion

Detecting spatial outliers is an important topic in the field of spatial data mining. The goal of identifying such anomalies is to discover hidden but potentially useful knowledge. A major limitation of existing outlier detection algorithms is that they generally require a pre-determined number of spatial outliers, which is not suitable in real applications due to the well-known masking and swamping effects. In this paper, we present an entropy-based approach to assess an optimal number of spatial outliers in a spatial data set. Specifically, we define the formula of spatial local contrast entropy, analyze the theoretical characteristics and foundataion, and propose an effective algorithm. Experiments were conducted on the cases of single attribute and multiple attributes. The empirical results validate that our proposed approach can appropriately reveal an “inflexion” point for the outlier number assessment.

Chapter 5

On Detecting Spatial Categorical Outliers

Spatial outlier detection is an important research problem that has received much attentions in recent years. Most existing approaches are designed for numerical attributes, but are not applicable to categorical ones (e.g., binary, ordinal, and nominal) that are popular in many applications. The main challenges are the modeling of spatial categorical dependency as well as the computational efficiency. This paper presents the first outlier detection framework for spatial categorical data. Specifically, a new metric, named as Pair Correlation Ratio (PCR), is measured for each pair of category sets based on their co-occurrence frequencies at specific spatial distance ranges. The relevances among spatial objects are then calculated using PCR values with regards to their spatial distances. The outlierness for each object is defined as the inverse of the average relevance between the object and its spatial neighbors. The objects with the highest outlier scores are returned as spatial categorical outliers. Several algorithms are further designed for single-attribute and multi-attribute spatial categorical datasets. Extensive experimental evaluations on both simulated and real datasets demonstrated the effectiveness and efficiency of our proposed approaches, compared with the existing ones.

The chapter is organized as follows. Section 5.1 gives the background and motivation. Section 5.2 provides some critical definitions used in SCOD, and introduces a general SCOD framework. Section 5.3 presents two SCOD approaches to identifying SCOs with single attribute, named PCF-SCOD and *k*NN-SCOD. *k*NN-SCOD work is extended to detect the SCOs with multiple attributes in Section 5.4. Experimental evaluations on both simulated and real life datasets are presented in Section 5.5. The paper concludes with a summary of the research presented in Section 5.6.

5.1 Background and Motivation

With the ever-increasing volume of spatial categorical data, identifying hidden but potentially interesting patterns of anomalies has attracted considerable attentions, particularly from the areas of data mining experts and geographers. Spatial Categorical Outlier (SCO) analysis, which aims at detecting abnormal objects in spatial context, becomes one of the most important spatial data mining branches. The identification of SCOs can help extract important knowledge in many applications, including geological data, meteorological data, satellite image analysis, and hotspot identification.

During the past decades, numerous Traditional Categorical Outlier Detection (TCOD) algorithms [31, 40, 92] have appeared in some literatures. TCOD approaches can be categorized into four groups: rule based, probability distribution based, entropy based and similarity based. Rule based approaches [4, 30, 70, 71, 121, 167] mine rules from the dataset, and observations which are significantly uncommon are recognized as anomalies. Typical algorithms include LERAD[30], WSARE[167], FP-Outlier[71], etc. Distribution based approaches[22, 40, 109, 128] model the normal data as a specific probability density distribution. Each object that significantly deviates the normal distribution is identified as an outlier. Representative models include Bayesian network and dependency trees, etc. Entropy based methods[68, 69] define TCOD as an optimization problem. That is, identifying l objects such that after removing them, the expected entropy of the rest of dataset is minimized. Similarity based approaches combine some typical TNOD approaches[21, 130] with certain well-designed dissimilarity measures together to identify TCOs. Meanwhile, some research works focus on more efficiently identifying categorical outliers, including AVF[92] and MapReduce AVF[93]. When encountering Spatial Categorical Outlier Detection(SCOD), TCOD approaches sometimes can't be satisfactory with the spatial context. First, spatial objects have complex structures (e.g., points, lines, polygons and locations, etc.). Second, traditional approaches do not consider spatial dependencies when identifying anomaly patterns. As the geographic rule of thumb, "Nearby things are more related than distant things [155]" requires more considerations on spatial autocorrelation in spatial analysis. Third, TCOD methods treat spatial and non-spatial attributes equally, which should be considered separately for spatial anomaly identification.

Recently, a number of algorithms [2, 7, 32, 60, 150] have been proposed to identify outliers in spatial databases [140, 142]. There are three basic classes, namely, visualization based, statistic based, and graph based. Visualization based approaches utilize visualization techniques to highlight outlying objects. Representative algorithms include scatterplot [62] and Moran scatterplot [7]. Statistic based approaches apply statistical tests to measure the local inconsistencies. Typical methods include Z [143], median-based Z[100], iterative-Z [100], and GLS-SOD [33] approaches. Graph based ones [90, 99, 144] detect spatial outliers by designing a function to compute the difference between specific observation and its neighboring points. Other works identified outliers by studying the property of specific spatial data. Zhao et al. proposed a

wavelet-based method to detect region outliers [175]. Lu et al. presented a multi-scale approach to detecting spatial temporal outliers [101]. Adam et al. introduced an approach that considers both the spatial and semantic relationship among neighbors [2]. A local outlier measure [150] was proposed by Sun and Chawla to capture the local behaviors of data in their spatial neighborhood. However, most of the aforementioned techniques have only concentrated on continuous real-valued data attributes. There is no mechanism for processing spatial categorical data with no implicit ordering.

Actually, in real world, the non-spatial attributes of spatial data are usually category-typed, where attributes have no intrinsic order information. A typical example is Rock whose values include Igneous, Sedimentary, and Metamorphic. The special property makes anomaly detection in spatial categorical domain more complicated than that in numerical one. Currently, there is a lack of Spatial Categorical Outlier Detection (SCOD) approaches. When encountering categorical dataset, some introduce Spatial Numerical Outlier Detection (SNOD) methods by directly mapping the categorical attributes to continuous ones. However, there are several critical issues: 1)**mis-utilization:** statistically, the definition of an SCO is different with that of a Spatial Numerical Outlier (SNO). Although both of them focus on the identification of spatial abnormal behaviors, SCOD is determined by the co-occurrence infrequency, while SNOD focuses on the numerical differences; 2)**complicated function:** the mapping process is not straightforward, especially for nominal attributes; 3)**swamping and masking problems:** without estimating outlying degrees accurately, some true outliers may be missed and normal ones misclassified as outliers.

Pair Correlation Function (PCF) has been proven very effective [77] to capture how observations are packed together, which could be utilized to estimate the relevance among spatial categorical objects. It is a probability measure to find a unit at a distance of d away from a reference unit. PCF techniques have been widely used to analyze the behavioral characteristics of the individual objects in a variety of natural systems, like electrostatic, magnetic and biological application. This paper investigates the benefits of PCF techniques on SCOD, and design algorithms for spatial datasets with single and multiple categorical attributes. First, PCF techniques are utilized to measure the Pair Correlation Ratio (PCR) between any pair of category sets as a function of spatial distances. And, the discrete relevance between a reference object and its neighbors is fitted by the PCR function. Then, the outlying degree for each object is computed as the inverse of average PCR between the object and its neighbors. Finally, the top l objects with higher outlierness are identified as SCOs. The key contributions of this paper include:

- **Formalization of the SCOD problem**. This is the first work that specifically focuses on Spatial Categorical Outlier Detection(SCOD). The SCOD problem is differentiated from the SNOD one: an SCO is identified as a spatial observation which occurs infrequently with regard to its spatial neighbors.
- **Design of two SCOD algorithms for single attribute dataset.** We first present a PCF-SCOD (Pair

Correlation Function based Spatial Categorical Outlier Detection) algorithm to identify SCOs by investigating the capability of PCF techniques of calculating the Pair Correlation Ratios (PCRs) for each pair of categories at specific distances. Further, considering the computational cost of PCF-SCOD, a *k*NN-SCOD(*k* Nearest Neighbor based Spatial Categorical Outlier Detection) approach is proposed to approximate the outlier scores. It allows for more efficient SCOD when memory and processor resources are issues.

- **Design of one SCOD algorithm for multi-attribute dataset.** The *k*NN-SCOD work is extended to the SCOD issue in multi-attribute domain. By mapping the *k*NN relationship from the raw dataset into a well-defined pair object dataset, the PCR of possible pair category sets is computed to capture the relevance among objects which are spatial neighbors with each other.
- **Comprehensive experiments to validate the effectiveness and efficiencies of the proposed techniques.** The proposed approaches were evaluated by the extensive experiments on simulated and real datasets. The results demonstrated that PCF series of algorithms outperformed 14 existing techniques for both single and multiple attribute dataset.

5.2 Preliminary Concept

This section introduces PCF techniques, summarizes some key notations used, and defines the SCOD problem. The deficiencies of existing methods are also examined.

5.2.1 Pair Correlation Function

In mathematical mechanics, PCF, $g(r)$, is defined as the observed probability of finding an object at a given distance, *r*, from a fixed reference particle [133]. The mathematical definition of $g(r)$ is

$$g(r) = \frac{dn(r)/N}{dv(r)/V} = \frac{dn(r)}{dv(r)} \cdot \frac{V}{N} = \frac{dn(r)}{4\pi r^2 dr} \cdot \frac{V}{N} \tag{5.1}$$

Where N and V denote the number of units and the volume of the entire system, respectively; $dn(r)$ and $dv(r)$ represent those in the shell-region; r is the distance from reference unit to the shell of interest. Fig. 5.1 depicts the 2D-projection of a typical example which describes the PCF computation in Eq.(5.1). In this paper, the relevances among spatial objects are determined by the frequency of co-occurrence of a pair of categories at specific distances. PCF is capable of estimating how observations are packed together, which could be utilized to capture the relationship among spatial categorical objects.

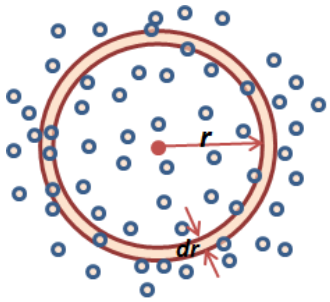


Figure 5.1: PCF using a spherical shell of thickness dr

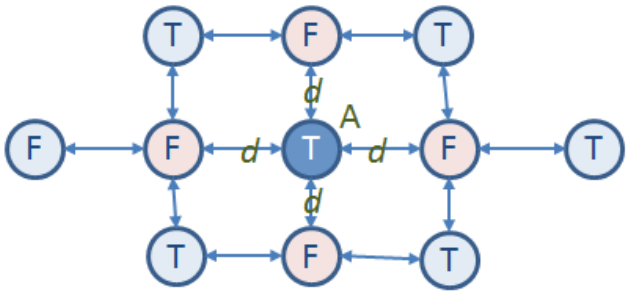


Figure 5.2: An example of differentiating an SNO and an SCO

5.2.2 Preliminary Definition

To formalize the SCOD framework, we need to understand some basic definitions.

Definition 1 (Spatial Categorical Dataset) *Let s denote a spatial location on a domain S of the d dimensional Euclidean space R^d. Let $A_1, \cdots, A_m$ be a set of categorical attributes and $\mathcal{C}_1, \cdots, \mathcal{C}_m$ non-empty sets over these attributes where $\mathcal{C}_i \bigcap \mathcal{C}_j = \phi$ for $i \neq j$.*

A set $\mathcal{D} \subseteq S \times \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$ is called a spatial categorical dataset over the domains, S, $\mathcal{C}_1, \cdots, \mathcal{C}_m$. Each record $r_i \in \mathcal{D}$ $(i \in 1, \cdots, n)$ can be denoted as a vector $(r.s, r.A_1, \cdots, r.A_m)\prime$, where $r.A_i \in \mathcal{C}_i$. The number of categorical attributes, m, is also referred as the dimensionality of the spatial dataset.

Categorical attributes can be classified into two types: ordinal and nominal ones. The key characteristic of nominal attributes is that different values which a data takes in an attribute domain are absolutely not inherently ordered, like the different colors. The issue of distance or dissimilarity for nominal data is not as straightforward as for ordinal or numerical one. Thus it is difficult to directly compare two nominal values. This paper is focused on such type of dataset as consists solely of nominal attributes. However, our approach could be directly applied into spatial categorical dataset with ordinal attributes.

Informally, the anomalous behavior in spatial domain can be truly captured by the local difference, which is determined by the irrelevance between a specific object and its spatial neighbors. In the paper, k-Nearest Neighbor (kNN) is utilized to construct the neighborhood relationship.

Definition 2 (Spatial Neighborhood) *Given a dataset $\mathcal{D}$ with n points and parameter k, for $r_i \in \mathcal{D}$, its spatial neighborhood is constructed by the top k points according to its spatial Euclidean Distance vector with the rest of observations in the dataset, such that $\forall j \in 1, ..., n$, $j \neq i$, $r_j \in kNN(r_i) : d^E(r_i, r_j) \leq d_k^E(r_i)$, where $d_k^E(r_i)$ represents the distance between r_i and its k^{th} spatial neighbor.*

In numerical domain, an SNO is defined as the one whose non-spatial attributes are significantly different

with those of its neighbors. Such definition is not applicable in categorical domain. For example, as shown in Fig. 5.2, based on the idea of SNOD approaches, object *A* will be recognized as an outlier since it has the categorical attribute, *T*, which is very different with its neighbors', *Fs*. However, the contrary is the case in categorical domain. This is because the pair of attributes, $< T, F >$ or $< F, T >$ occurs normally at the spatial distance, *d*. Object *A* should be treated as a normal observation. In this sense, the definition of spatial outliers in categorical domain is totally different with that of SNOs.

Definition 3 (SCO) *Let r_i be an observation in $\mathcal{D}$, and $r_{i\_1}, \cdots, r_{i\_k}$ be its spatial neighbors. Its outlierness, for $k \geq 1$,is defined as*

$$OutScore(r_i) = -\frac{\sum_{j=1}^{k} PCR(r_i.A, r_{i\_j}.A, d^E(r_i, r_{i\_j}))}{k} \tag{5.2}$$

r_i will be considered as an outlier if $OutScore(r_i) > \theta$. Here, θ is a user-defined threshold. $PCR(r_i.A, r_{i\_j}.A, d^E(r_i, r_{i\_j}))$ denotes the co-occurrence frequency of the pair category sets, $< r_i.A, r_{i\_j}.A >$, of objects, r_i and r_{i_j}, at the specified distance, $d^E(r_i, r_{i\_j})$.

In one word, an SCO is an observation which has lower co-occurrence frequency with its spatial neighbors. PCF is capable of estimating such frequencies as how objects are packed together, which could be utilized to calculate PCR, further $OutScore(r_i)$. Section 5.3 and 5.4 will discuss the PCR computation in spatial categorical dataset with single and multiple attributes, respectively.

Based on the above definitions, the SCOD problem can be modeled as follows. **Given:**

- $\mathcal{D}$ is a set of spatial objects r_1, ..., r_n with single or multiple categorical attributes.
- ***k*** is an integer denoting the number of adjacent data objects which form the spatial neighborhood.
- ***l*** is the number of outliers to be identified, generally, $l \ll n$.

Objective:

- Design a mapping function $f : \mathcal{D} \times \mathcal{D} \longrightarrow R^+$, which estimates PCR for each pair of objects as a function of spatial distances.
- Estimate the *OutScore* for each observation, and identify a set of O_1, ..., $O_l \in \mathcal{D}$ with higher values as SCOs.

5.3 Spatial Categorical Outlier Detection in Single Attribute Dataset

Intuitively, given a spatial dataset, a normal observation is the one that behaves normally with regard to its spatial neighborhood. In single categorical domain, this corresponds to the higher frequency of co-occurrence of a pair of categories at a specified distance. The categorical outlier has rarely occurring category attribute with regard to the ones of its neighborhood. This section presents two SCOD approaches to detecting SCOs with single attribute, namely PCF-SCOD (Pair Correlation Function based Spatial Categorical Outlier Detection) and kNN-SCOD-S (k Nearest Neighbor based Spatial Categorical Outlier Detection in Single attribute dataset).

5.3.1 Pair Correlation Function based SCOD

In this section, we investigate the benefits of PCF techniques to capture the rare behaviors of SCOs. The main components in PCF-SCOD are described as follows.

- **PCR(Pair Correlation Ratio) estimation.** PCR is defined to characterize the co-occurrence frequency of each pair of categories at different specified distances. With the set of discrete points in a 2-D space, determined by PCR values against spatial distances, we can statistically learn a continuous PCR function which can easily estimate the PCRs among spatial objects.
- **Neighborhood formulation and outlierness computation.** The spatial neighborhood for each object can be formed using kNN. And the outlier degree for each object is computed by the mean of PCRs between itself and its spatial neighbors.
- **Outlier identification.** Finally, the outer scores are ranked in an descending order and the top l objects are identified as SCOs.

For the above components, the first one is extremely critical since it determines the estimation quality of the relevances among observations. Section 5.1.1 introduces PCR computation in particular, and then PCF-SCOD algorithm is described step by step in Section 5.1.2.

Pair Correlation Ratio Computation

For each random variable r, $r.A$ is a multilevel categorical variable taking values in $\mathcal{C} = A^1, ..., A^L$. We denote Eq.(5.3) as the frequency of observing category A^l in the dataset,

$$Freq(A^l) = P[A(r_i.A) = A^l] = \frac{n^{A^l}}{n} \tag{5.3}$$

where n^{A^l} represents the number of objects whose non-spatial attribute are A^ls, and n the number of objects in the whole dataset. Fig. 5.3(a) depicts a small spatial categorical dataset which consists of 64 objects, of which 37 ones take category "+", and 27 ones "-". With Eq.(5.3), we get $Freq(+) = 37/64$ and $Freq(-) = 27/64$.

Let $SPF(< A^l, A^{l'} >, d^E(r_i.X, r_j.X))$ denote the Spatial Pair Frequency associated to two objects, r_i and r_j, where $r_i.A = A^l$ and $r_j.A = A^{l'}$. Then PCR can be defined as follows:

Definition 4 (Pair Correlation Ratio-PCR) *Considering a spatial pair correlation process in which there are two observations, r_i, r_j in $\mathcal{D}$, each of them is tagged with one category, A^l and $A^{l'}$, respectively. The PCR of r_i,r_j is defined as the normalized spatial pair frequency of the pair of categories, $< A^l, A^{l'} >$, happen to occur at r_i and r_j.*

The mathematical definition of PCR is

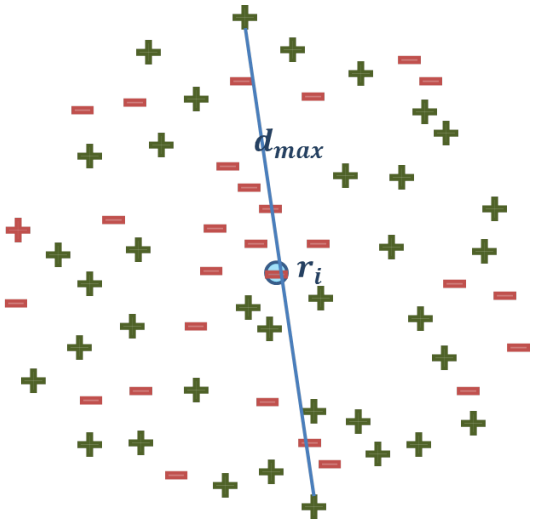
$$PCR(r_i, r_j) = \frac{SPF(< A^l, A^{l'} >, d^E(r_i.S, r_j.S))}{Freq(A^l) \cdot Freq(A^{l'})} \tag{5.4}$$

As shown in Eq.(5.4), the PCR value between two spatial objects, is only determined by the co-occurrence frequency of categories they take and their spatial Euclidean Distance, not their specific spatial locations. In the following, SPF computation is discussed step by step by utilizing the example shown in Fig. 5.3.

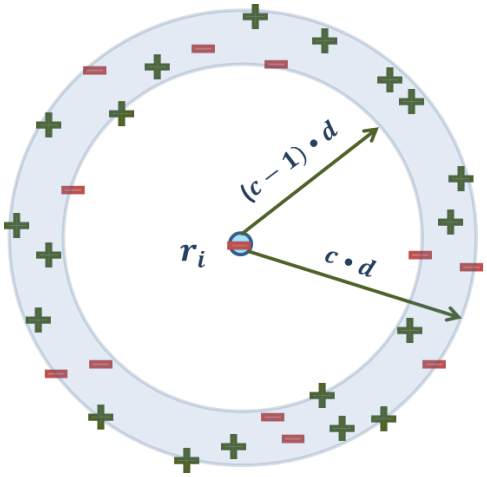
- **Distance division.** Compute the Euclidean distance for each pair of spatial objects, identify the maximal and minimal ones, ${d^E}_{Max}$ (as shown in Fig. 5.3(a)) and ${d^E}_{Min}$ (set as 0), and divide the distance into b small bins whose sizes are computed as:

$$d = \frac{{d^E}_{Max}}{b} \tag{5.5}$$

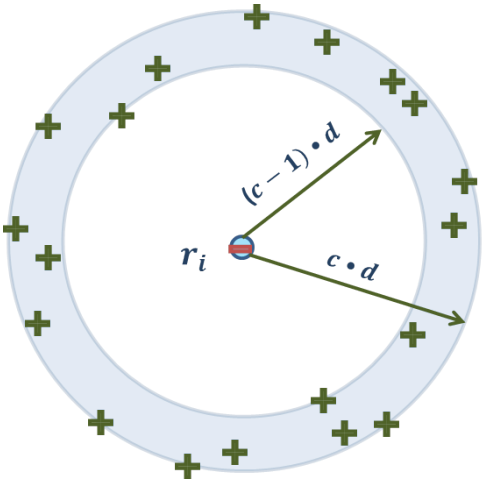
  As we know, it is common for spatial objects to be autocorrelated at shorter distances. It is not necessary to take the pair correlation at longer distances into considerations. Simply, ${d^E}_{Max}$ can be



(a) An example dataset



(b) $\mathcal{D}^c$ identification w.r.t. r_i



(c) $\mathcal{D}^c_{-+}$ identification w.r.t. r_i

Figure 5.3: An example of identifying B-PD and B-PC-PD.

approximated by

$$d^E{}_{Max} = \frac{1}{2} max\{|max(Proj_x(r_i.S)) - min(Proj_x(r_i.S))|, \\ |max(Proj_y(r_i.S)) - \min(Proj_y(r_i.S))|\}, i, j = 1, ..., n \quad (5.6)$$

where $Proj_x(\cdot)$ and $Proj_y(\cdot)$ represent the projection operations of *S* location on *X*, *Y* coordinates, respectively. It is reasonable for such approximation since SCOD focuses on the local relevance estimation.

- **Identification of Bin based Pair Dataset (B-PD).** Based on the spatial distance, map each pair of objects into their corresponding distance bin.

$$\mathcal{D}^c = \{< r_i, r_j >, (c-1) \cdot d \leq d^E(r_i.S, r_j.S) < c \cdot d, c \in [1, b]\} \quad (5.7)$$

For example, for the reference object r_i shown in Fig. 5.3, based on its spatial distances from others, we can identify 30 objects, $\{r_j\}_{j=1}^{30}$, which make $< r_i, r_j > \in \mathcal{D}^c$ since they satisfy the condition: $(c-1) \cdot d \leq d^E(r_i.S, r_j.S) < c \cdot d$. As depicted in Fig. 5.3(b), the 30 objects are contained in the blue shaded circular ring.

- **Identification of Bin and Pair Category based Pair Dataset (B-PC-PD).** By scanning the identified B-PD, we map each pair of objects into its pair category based subset so that their categorical attributes are just the specified pair of categories. That is, we can construct $\mathcal{D}^c_{A^l A^{l'}}$ as follows:

$$\mathcal{D}^c_{A^l A^{l'}} = \{\langle r_i, r_j \rangle, [(r_i.A == A^l) \,\&\, \&(r_j.A == A^{l'})] \| \\ [(r_i.A == A^{l'}) \,\&\, \&(r_j.A == A^l)], < r_i, r_j > \in \mathcal{D}^c, c \in [1, b]\} \quad (5.8)$$

In Fig. 5.3(b), in the identified objects $\{r_j\}_{j=1}^{30}$ with regard to reference object r_i in $\mathcal{D}^c$, we can map 11 pairs of $\{< r_i, r_j >\}_{j=1}^{11}$ into $\mathcal{D}^c_{--}$, and the other 19 pairs into $\mathcal{D}^c_{-+}$ based on their corresponding categorical attributes. Fig. 5.3(c) depicts the pair objects in $\mathcal{D}^c_{-,+}$ with regard to the reference object r_i.

- **Spatial Pair Frequency computation.** Therefore, the SPF of the pair of categories in the c^{th} bin can be computed by

$$SPF(\langle A^l, A^{l'} \rangle, [(c-1) \cdot d, c \cdot d]) = \frac{|\mathcal{D}^c_{A^l A^{l'}}|}{|\mathcal{D}^c|} \quad (5.9)$$

where $|\mathcal{D}^c_{A^l A^{l'}}|$ and $|\mathcal{D}^c|$ represent the number of pair objects in $\mathcal{D}^c_{A^l A^{l'}}$ and $\mathcal{D}^c$, respectively. Overall, for each pair of $\langle A^l, A^{l'} \rangle$, we can estimate b pair frequency values corresponding with b bins. Based on the b discrete points in a 2-D space, we can statistically learn a pair frequency function $SPF(\langle A^l, A^{l'} \rangle, d^E)$ by polynomial and curve fitting, subjecting to the following constraints:

Table 5.1: Main parameters used in this paper

Parameters	Description
S	A dataset storing the spatial attributes
A	A dataset storing the non-spatial categorical attributes
b	The number of bins to divide the distance values
m	The number of categorical attributes
n	The number of spatial objects in the dataset
k	The number of spatial neighbors
l	The number of SCOs

1. $SPF(\langle A^l, A^{l'}\rangle, d^E) = SPF(\langle A^{l'}, A^l\rangle, d^E)$

   Proof. By definition, it is easy to prove this constraint.

2. $SPF(\langle A^l, A^{l'}\rangle, 0) = \begin{cases} Freq(A^l) & A^l = A^{l'} \\ 0 & A^l \neq A^{l'} \end{cases}$

   Proof. If $A^l = A^{l'}$, $SPF(\langle A^l, A^{l'}\rangle, 0)$=$\frac{|\mathcal{D}^0_{A^l A^{l'}}|}{|\mathcal{D}^0|}$=$\frac{n^{A^l}}{n}$=$Freq(A^l)$, and if $A^l \neq A^{l'}$, $PF(\langle A^l, A^{l'}\rangle, 0)$=$\frac{|\mathcal{D}^0_{A^l A^{l'}}|}{|\mathcal{D}^0|}$=$\frac{0}{n}$

3. $\sum_{l'=1}^{L} SPF(\langle A^l, A^{l'}\rangle, d^E) = Freq(A^l)$.

   Proof. We can identify $D^c_{A^l}$ as $D^c_{A^l} = \{\langle r_i, r_j\rangle, (r_i.A == A^l)\&((c-1)\cdot d \leq d^E(r_i.S, r_j.S) < c\cdot d), < r_i, r_j > \in \mathcal{D}, c \in [1, b]\}$.

   There is a deduction as follows:
   $\sum_{l'=1}^{L} SPF(\langle A^l, A^{l'}\rangle, d^E) = \frac{\sum_{l'=1}^{L} |\mathcal{D}^c_{A^l A^{l'}}|}{|\mathcal{D}^c|} = \frac{n^{A^l}}{n} = Freq(A^l)$

PCF-SCOD Algorithm

The proposed PCF-SCOD algorithm for single attribute domain has 6 input parameters, *S*, *A*, *b*, *n*, *k* and *l*, which are described in Table 5.1. Algorithm 1 describes this approach as the following 4 critical steps.

Step-1(*line:1-3*) **Formalization of spatial neighborhood**. First, we construct distance matrix, $Dismat$, in which the i^{th} row records the spatial distances between r_i and the rest of objects in the dataset. With it, the spatial neighborhood matrix, $Neighbor$, can be identified for each spatial object.

Step-2(*line:4-17*) **Computation of SPFs among spatial objects**.

a. (*line:4-5*) `Distance division`. With the stored values in $DisMat$, identify its maximum and minimum values (0). Then, the size of unit bin, d, can be computed using Eq.(5.5).

b. (*line:6*) `Computation of category frequency`. We construct the frequency array, $Freq_A$, which records the occurrence frequencies for all the observing category, and the pair category array,

Algorithm 4 PCF-SCOD-S Approach

1: **for** $i = 1$ to n **do** {Calculate the neighborhood and distance matrix}
2: $[Neighbor(i,:), DistMat(i,:)] = kNN(S, r_i.S, k)$
3: **end for**
4: $d^E_{Max} = max(DistMat)$;{Identify the maximum spatial distance}
5: $d = \frac{d^E_{Max}}{b}$;{Calculate the size of unit bin}
{Computation of category frequency, pair category array and its sizes.}
6: $[Freq_A, PC\_Arr, N_p] = CateFreq(A)$;
7: **for** $c = 1$ to b **do** {Identify B-PD}
8: $\mathcal{D}^c = B\_PD\_Iden(DistMa, Cond^{d,c})$;
9: **for** $p = 1$ to N_p **do** {Identify B-PC-PD}
10: $A^l A^{l'} = PC\_Arr(p)$;
11: $\mathcal{D}^c_{A^l A^{l'}} = B\_PC\_PD\_Iden(\mathcal{D}^c, Cond^{A^l A^{l'}})$;
12: $SPF(A^l, A^{l'}, c \cdot d) = \frac{|\mathcal{D}^c_{A^l A^{l'}}|}{|\mathcal{D}^c|}$ {Calculate its corresponding spatial pair frequency.}
13: **end for**
14: **end for**
15: **for** $c = 1$ to N_p **do** {Model continuous PCR function}
16: $SPF(A^l, A^{l'}, d^E) = FitModel(\{SPF(A^l, A^{l'}, [(c-1) \cdot d, c \cdot d])\}_{c=1}^{b})$;
17: **end for**
18: **for** $i = 1$ to n **do** {Calculate PCR matrix between spatial object and its neighbors}
19: **for** $j = 1$ to k **do**
20: $f = Neighbor(i, j)$;
21: $PCRMat(i, j) = \frac{SPF(r_i.A, r_f.A, DistMat(i,f))}{|Freq_{r_i.A}| \cdot |Freq_{r_f.A}|}$;
22: **end for**
23: **end for**
24: $RelevanceArr = mean(PCRMat)$; {Compute relevances for spatial objects}
25: $RankList = Rank(RelevanceMat, ascend)$; {Rank objects with ascending relevance values}
26: $O_l = Outlier(RankList, 1:l)$ {Mark the outliers}

PC_Arr, which stores all the possible pairs of categories (N_p represents the number of possible pairs of categories) in the dataset.

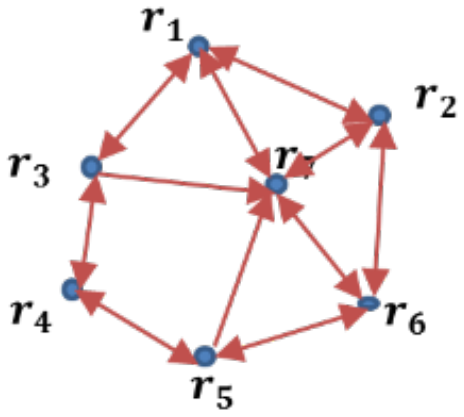
c. (*line:7-14*) `SPF computation.` This step includes three important procedures: bin based pair set identification, bin and pair category based pair set identification and the discrete SPF computations. At step 8, we use function, B_PD_Iden, extract all the pair objects for $\mathcal{D}^c$, which satisfy certain distance conditions, $Cond^{d,c}$ (in Eq.(5.7)). At step 11, function $B\_PC\_PD\_Iden$, is used to construct $\mathcal{D}^c_{A^l A^{l'}}$ by scanning the pair objects in $\mathcal{D}^c$, which satisfy category attribute condition, $Cond^{A^l A^{l'}}$ (in Eq.(5.8)). With the above two pair sets, the b discrete SPF values for each pair of categories can be computed at step 12.

c. (*line:15-17*) `Learn of continuous SPF function.` With the above discrete SPF values against

b different distance range, we statistically learn a continuous SPF function for each pair of categories using polynomial and curve fitting.

Step-3(*line:18-24*) **Construction of PCR matrix**. Utilizing Eq.(5.4), with SPF function, the relevance scores (PCR) can be simply calculated for any pair of spatial objects. Furthermore, PCR matrix, $PCRMat$, is constructed as the mean of the PCR values between the reference observation and its neighbors.
Step-4(*line:25-26*) **Outlier identification**. Finally, the objects are sorted with ascending PCR values, and the l objects with lower relevance scores are recognized as outliers.

Computational Complexity. To form the distance and neighborhood matrices will take $O(n^2)$. It takes $O(n)$ to construct the category frequency array and pair category array. Identifying $\mathcal{D}^c$ and $\mathcal{D}^c_{A^l A^{l'}}$ takes around $O(b \cdot N_p \cdot |\mathcal{D}^c| \cdot n^2)$. Finally, computing the PCR matrix costs $O(k \cdot n)$. In summary, assuming $n \gg k$, $n \gg b$, $n \gg N_p$ and $n \gg |\mathcal{D}^c|$. The total computational complexity of PCF-SCOD approach is $O(n^2)$=($O(n^2)$+$O(n)$+$O(b \cdot N_p \cdot |\mathcal{D}^c| \cdot n^2)$)+$O(k \cdot n)$).



ID	Attr.1	Attr.2	3NN		
r_1	T	$\{T, P\}$	r_2	r_3	r_7
r_2	F	$\{F, Q\}$	r_1	r_6	r_7
r_3	F	$\{F, P\}$	r_1	r_4	r_7
r_4	F	$\{F, Q\}$	r_3	r_5	r_7
r_5	T	$\{T, P\}$	r_4	r_6	r_7
r_6	F	$\{F, Q\}$	r_2	r_5	r_7
r_7	F	$\{F, P\}$	r_1	r_2	r_6

Figure 5.4: A sample of spatial categorical dataset. (Attr.1 means the observed attributes in single attribute domain, which is used in Section 5.2; Attr.2 means the observed attributes in multiple attribute domain, which is used in Section 5.4.)

5.4 Spatial Categorical Outlier Detection in Multiple Attribute Dataset

The work of *k*NN based PCR approximation can be easily extended to solve the SCOD issue in multi-attribute domain. That is, given a spatial dataset, an outlying observation is the one whose non-spatial attribute set occurs infrequently with regards to those of its spatial neighborhood. And, the calculation of PCR is computed by the frequency of co-occurrence of a pair of category sets at a specific spatial distance.

Since PCF based approach will be a time-consuming process, we only introduce how to compute PCR values in *k*NN-SCOD-M (*k* Nearest Neighbor based Spatial Categorical Outlier Detection in Multiple attribute dataset).

5.4.1 PCR Computation in Multi-Attribute Dataset

Similarly, *k*NN-SCOD-M approach first extracts the *k*NN relationship from the raw dataset $\mathcal{D}$ and maps it into F^k with $2m$ dimensions. In F^k, each data frame contains the attribute set, $\mathbb{A} = \{< A_1, .., A_m >, < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i$, $i \in [1, m]$. First, we need to learn two important concepts about attribute subset in $\mathcal{D}$ and F^k.

Table 5.2: Observations for PAS $\{< A_1, A_2 >, < A_1, A_2 >\}$ in F^3

Pair objects	$\{< A_1, A_2 >, < A_1, A_2 >\}$
$< r_1, r_3 >$	$< \{T, P\}, \{F, P\} >$
$< r_1, r_2 >$	$< \{T, P\}, \{F, Q\} >$
$< r_1, r_7 >$	$< \{T, P\}, \{F, P\} >$
$< r_2, r_6 >$	$< \{F, Q\}, \{F, Q\} >$
$< r_2, r_7 >$	$< \{F, Q\}, \{F, P\} >$
$< r_3, r_4 >$	$< \{F, P\}, \{F, Q\} >$
$< r_3, r_7 >$	$< \{F, P\}, \{F, P\} >$
$< r_4, r_5 >$	$< \{F, Q\}, \{T, P\} >$
$< r_4, r_7 >$	$< \{F, Q\}, \{F, P\} >$
$< r_5, r_6 >$	$< \{T, P\}, \{F, Q\} >$
$< r_5, r_7 >$	$< \{T, P\}, \{F, P\} >$
$< r_6, r_7 >$	$< \{F, Q\}, \{F, P\} >$

Table 5.3: PCR Computation

Pair Categories	Freq.	Prob.	PCR
$<$ $\{T, P\}$,$\{F, Q\} >$	3	0.25	2.00
$<$ $\{T, P\}$,$\{F, P\} >$	3	0.25	2.00
$<$ $\{F, Q\}$,$\{F, Q\} >$	1	0.083	0.99
$<$ $\{F, Q\}$,$\{F, P\} >$	4	0.34	4.04
$<$ $\{F, P\}$,$\{F, P\} >$	1	0.083	0.99
$Freq(\{F, P\})$=0.29;$Freq(\{F, Q\})$=0.29 $Freq(\{T, P\})$=0.43			

Definition 6(Attribute Subset-AS). *Given a dataset $\mathcal{D}$ with m categorical attributes, $A = \{A_1, .., A_m\}$, $A_i \subseteq \mathcal{C}_i$, its AS is defined as follows:*

$$\mathcal{AS} = \{A^*, \{A^* = \{A_x, ..., A_y\}\}, 1 \leq x \leq y \leq m, A_i \subseteq D_i, i \in [x, y]\} \tag{5.10}$$

Considering $Attr.2$ of the sample spatial dataset as shown in Fig. 5.4, there are two category attributes: $A_1 = \{T, F\}$ and $A_2 = \{P, Q\}$. By definition, we can easily generate all of its ASs:
$\mathcal{AS}$: { $\{A_1\}, \{A_2\}, \{A_1, A_2\}$ }.
Fig. 5.4 also describes the *3*NN relationship among objects. By utilizing Definition 5, we can map it into a

dataset set F^3 with 4-dimension attributes, as shown in Table 5.4.

Definition 7(Pair Attribute Subset-PAS). *Given a dataset F^k with $2m$ categorical attributes, $\mathbb{A} = \{< A_1, .., A_m >, < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i$, its PAS is defined as follows:*

$$\begin{aligned} \mathcal{PAS} &= \{< A^*, A^{*'} >, \{A^* = \{A_x, ..., A_y\}\} \ \& \ \&\{A^{*'} = \{A_x, ..., A_y\}\} \\ &\quad \& \ \& \ \{|A^*| == |A^{*'}|\}, 1 \leq x \leq y \leq m, A_i \subseteq \mathcal{C}_i, i \in [x, y]\} \end{aligned} \tag{5.11}$$

Apparently, A^* and $A^{*'}$ always exist in pairs in PAS. They are originated from the same attribute domain in dataset $\mathcal{D}$. For pair attribute set $\mathbb{A}$, there are $(2^m - 1)$ pair subsets. Similarly, we can enumerate all the PASs for the sample dataset in Fig. 5.4 as follows.
$\mathcal{PAS}$: $\{ \{< A_1 >, < A_1 >\}, \{< A_2 >, < A_2 >\}, \{< A_1, A_2 >, < A_1, A_2 >\}$.

In this paper, we call AS and PAS observations as **AS and PAS combinations** which are generalized by the following concepts .

Definition 8(AS Combination). *Given a dataset $\mathcal{D}$ with m categorical attributes, $A = \{A_1, .., A_m\}$, $A_i \subseteq \mathcal{C}_i$. Suppose for each attribute A_i, it takes value in $\mathcal{L}_i$=$\{A_i^1, ..., A_i^{L_i}\}$. Therefore, an AS Combination (ASC) is one of the possible category examples of existing attributes in the AS. The ASC of A^* (defined in Definition 7) is,*

$$\mathcal{ASC}_{A^*} = \{\mathcal{A}, \mathcal{A} = \{A_x^{l_x}, ..., A_y^{l_y}\}, A_i^{l_i} \subseteq \mathcal{L}_i, i \in [x, y]\} \tag{5.12}$$

As we can see, each observation may take one of values $\{T, F\}$ for A_1, and $\{P, Q\}$ for A_2. $\{A_1\}$ is one of its ASs, then $ASC_{A_1} = \{\{T\}, \{F\}\}$. Similarly, $ASC_{\{A_1, A_2\}} = \{\{T, P\}, \{T, Q\}, \{F, P\}, \{F, Q\}\}$.

Definition 9(PAS Combination). *Given a dataset F^k with $2m$ category attributes, $\mathbb{A} = \{< A_1, .., A_m > , < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i$, $i \in [1, m]$. Similarly, it takes value in $\mathcal{L}_i$=$\{A_i^1, ..., A_i^{L_i}\}$. Therefore, an PAS Combination (ASC) is one of the possible pair category examples of existing attributes in the PAS. That is*,

$$\begin{aligned} \mathcal{PASC}_{<A^*, A^{*'}>} &= \{< \mathcal{A}, \mathcal{A}' >, \{\mathcal{A} = \{A_x^{l_x}, ..., A_y^{l_y}\}\} \ \& \ \{\mathcal{A}' = \{A_x^{l_{x'}}, ..., A_y^{l_{y'}}\}\} \\ &\quad \&\{|\mathcal{A}| == |\mathcal{A}'|\}, 1 \leq x \leq y \leq m, A_i^{l_i}, A_i^{l_{i'}} \subseteq \mathcal{L}_i, i \in [x, y]\} \end{aligned} \tag{5.13}$$

For example, there are three possible PASCs: $\{< T >, < T >\}$,$\{< T >, < F >\}$ and $\{< F >, < F >\}$ for PAS $\{< A_1 >, < A_1 >\}$ in F^3. And, all the observed PASCs for $\{< A_1, A_2 >, < A_1, A_2 >\}$ in Fig. 5.4 are enumerated in the second column in Table 5.4.

We can compute the PCR value for each PASC of a specified PAS $< A, A' >$ by utilizing the following 5

steps.

- **Dataset identification for specific ASC.** For a specific ASC $\mathcal{A} = \{A_x^{l_x}, ..., A_y^{l_y}\}$, we scan the dataset $\mathcal{D}$ and identify $\mathcal{D}_\mathcal{A}$ as follows.

$$\mathcal{D}_\mathcal{A} = \{r_i, (r_i.A_x = A_x^{l_x}) \,\&\, \cdots \,\&\, (r_i.A_y = A_y^{l_y}), r_i \in \mathcal{D}, A_j^{l_j} \subseteq \mathcal{L}_j, j \in [x, y]\} \tag{5.14}$$

- **Frequency computation of ASC.** Calculate the frequency for $\mathcal{D}_\mathcal{A}$ using Eq.(4.18).

$$Freq(\mathcal{D}_\mathcal{A}) = \frac{|\mathcal{D}_\mathcal{A}|}{|\mathcal{D}|} \tag{5.15}$$

  For ASC $\{T\}$ in Fig. 5.4, $D_T = \{r_1, r_5\}$, and its frequency is $2/7 = 0.29$.

- **Dataset identification for specific PASC**. Suppose $< \mathcal{A}, \mathcal{A}' >$ =< $\{A_x^{l_x}, ..., A_y^{l_y}\}$, $\{A_x^{l_{x'}}, ..., A_y^{l_{y'}}\} >$ is one of the PASCs of $< A^*, A^{*'} >$($\{A^* = \{A_x, ..., A_y\}\}$, $\{A^{*'} = \{A_x, ..., A_y\}\}$). Scan the whole set F^k and identify all observations which satisfy the following conditions.

$$\begin{aligned}\mathcal{F}^k_{\{\mathcal{A},\mathcal{A}'\}} \;=\; & \{f_i, [(f_i.A^*.A_x = A_x^{l_x}) \,\&\, ... \,\&\, (f_i.A^*.A_y = A_y^{l_y})\} \,\&\\ & \{(f_i.A^{*'}.A_x = A_x^{l_{x'}}) \,\&\, ... \,\&\, (f_i.A^{*'}.A_y = A_y^{l_{y'}})]|| \\ & [(f_i.A^{*'}.A_x = A_x^{l_x}) \,\&\, ... \,\&\, (f_i.A^{*'}.A_y = A_y^{l_y})\} \,\&\\ & \{(f_i.A^*.A_x = A_x^{l_{x'}}) \,\&\, ... \,\&\, (f_i.A^*.A_{y'}y = A_y^{l_{y'}})], f_i \in F^k\}\end{aligned} \tag{5.16}$$

  In $F^k_{\{\mathcal{A},\mathcal{A}'\}}$, it stores all pairs of objects which have the same PASC $< \mathcal{A}, \mathcal{A}' >$.

- **Frequency computation of PASC**. Calculate the frequency for each PASC $< \mathcal{A}, \mathcal{A}' >$.

$$Freq(< \mathcal{A}, \mathcal{A}' >) = \frac{|F^k_{\{\mathcal{A},\mathcal{A}'\}}|}{|F^k|} \tag{5.17}$$

  For example, for PASC $\{< T, P >, < F, P >\}$, $\mathcal{F}^3_{\{<T,P>,<F,P>\}}$= $\{< r_1, r_3 >, < r_1, r_7 >, < r_5, r_7 >\}$. Its frequency is equal to $3/12 = 0.25$.

- **PCR computation**. Finally, PCR for a PASC can be calculated as Eq.(5.21).

$$PCR_k(< \mathcal{A}, \mathcal{A}' >) = \frac{Freq(< \mathcal{A}, \mathcal{A}' >)}{Freq(\mathcal{D}_\mathcal{A}) \cdot Freq(\mathcal{D}_{\mathcal{A}'}))} = \frac{|F^k_{\{\mathcal{A},\mathcal{A}'\}}|/|F^k|}{(|\mathcal{D}_\mathcal{A}| \cdot |\mathcal{D}_{\mathcal{A}'}|)/|\mathcal{D}|^2} \tag{5.18}$$

Therefore, we can compute PCR on $\{< T, P >, < F, P >\}$ among object r_1 and r_3 as $PCR(< r_1, r_3 >) = 0.25/(0.29 * 0.43) = 2.00$. Table 5.5 shows all the PCR computation of all possible PASC for Fig. 5.4

on the PAS $\{< A_1, A_2 >, < A_1, A_2 >\}$. In the same way, we can compute the PCR values of other PASCs with regards to differnt PAS.

By scanning the dataset, we can determine, for each pair of spatial objects, there are at most $2^m - 1$ PCR scores which correspond to $2^m - 1$ PASC. After that, we choose the smallest one as their final relevance score. We identify relevances among objects in this way because, sometimes, an outlier only exists in the subspace of multiple attributes. Exhaustively estimating outlier scores in different PASC will help identify SCOs more effectively. With the smallest PCR vector, we can construct a PCR matrix (n-by-k). Further, the outlierness value can be computed for each object using the mean of neighborhood relevances.

5.4.2 Algorithm of *k*NN-SCOD-M

In this part, we generalize the *k*NN-SCOD-M approach which is utilized to detect multi-attribute outliers. There are *6* input parameters, *S*, *A*, *n*, *k*, *m* and *l*, which are described in Table 5.1. As shown in Algorithm 3, identifying SCOs with multiple attributes includes the following *5* steps.
Step-1(*line:1-4*)(**Construction of spatial neighborhood and mapping process of *k*NN relationships**). For each data observation r_i, identify its k spatial neighbors and map *k*NN relationship into F^k.
Step-2(*line:5-18*). **PCR computation**.

a. (*line:5*) `Identification of AS and PAS`. **Algorithm 4** describes this function in detail. Intuitively, there are $\binom{m}{i}$ ASs which consist of i attributes, and $\binom{m}{i}$ PASs with size as $2 * i$. Therefore, in each loop, $\binom{m}{i}$ ASs and PASs are identified, respectively.

b. (*line:6-7*) `Frequency computation of` $\mathcal{D}_{ASC}$ `and` F^k_{PASC}. We first operate the identification of possible ASCs and PASCs for each AS^* and PAS^* Meanwhile, their corresponding subset, $\mathcal{D}_{\mathcal{ASC}^*}$ and $\mathcal{F}^k_{\mathcal{PASC}^*}$ are identified by scanning $\mathcal{D}$ and $\mathcal{F}^k$. As shown in **Algorithm 5**, for each object in $\mathcal{D}$, any subset of its attributes can be mapped as one of the possible ASCs(Step 11 and 14). In the same way, for each pair observations in F^k, any subset of their attributes originated from the same domains can be recognized as one of the possible PASCs(Step 16). After that, we map each object into its corresponding ASC subset (Step 12, 15), and each pair of the object and its current spatial neighbor into the corresponding PASC subset. Finally, the frequencies of all the $\mathcal{D}_{ASC}$ and F^k_{PASC} are computed by using Eq.(5.18) and (5.20).

c. (*line:8-18*) `PCR computation for specific PASC`. Compute the PCR values between reference object and its *k*NN neighbors for each possible PASC.

Step-3(*line:19-28*). **Computation of Relevances among objects**. Then, use the mean of *k* PCRs as the relevance value in each PAS subspace. And, the smallest one of the $2^m - 1$ PCR values is recognized as its final relevance score.

Algorithm 5 kNN-SCOD-M Approach

1: **for** $i = 1$ to n **do** {Construct the neighborhood matrix.}
2: $[Neighbor(i,:)] = kNN(S, r_i.S, k)$;
3: **end for**
4: $F^k = MapFunction(Neighbor, A)$; {Map the kNN relationship into dataset F^k.}
5: $[PAS, AS] = ASIdentify(A)$; {Identify all possible PASs and ASs.}
{Identify all possible PASCs and ASCs for each PAS and AS, and then extract their corresponding subsets.}
6: $[PASC, ASC, \mathcal{D}_{ASC}, \mathcal{F}^k_{PASC}] = ASCIdentify(A, PAS, AS)$;
{Compute the frequency vectors for each $\mathcal{D}_{ASC}, \mathcal{F}^k_{PASC}$.}
7: $[Freq_{asc}, Freq_{pasc}] = FreqCompute(\mathcal{D}_{ASC}, F^k_{PASC})$;
8: **for** $i = 1$ to n **do** {Calculate PCR matrix between spatial object and its neighbors.}
9: $A_i = r_i.A$;
10: **for** $j = 1$ to k **do**
11: $r_j = Neighbor(i, j)$;
12: $A_j = r_i.A$;
13: **for** $a = 1$ to $2^m - 1$ **do**
14: $curPAS = PAS(a); curAS = AS(a)$; {With the information from r_i and r_j, identify its corresponding PASC and ASCs.}
15: $[curPASC, curASC_1, curASC_2]$=$PASCIdenfity(curPAS, curAS, r_i, r_j)$; {Compute the PCR values for neighbor objects $< r_i, r_j >$ in current PAS space.}
16: $PCRMat_a(i,k) = \frac{Freq_{pasc}(curPASC)}{Freq_{asc}(curASC_1) \cdot Freq_{asc}(curASC_2)}$;
17: **end for**
18: **end for**
19: **for** $a = 1$ to $2^m - 1$ **do** {Identify the smallest one of PCRs in all the PAS space as its final PCR value.}
20: **if** a==1 **then**
21: $PCR(i) = mean(PCRMat_a(i,:))$;
22: $tempValue = PCR(i)$;
23: **else**
24: $PCR(i) = min\{mean(PCRMat_a(i,:)), tempValue\}$;
25: $tempValue = PCR(i)$;
26: **end if**
27: **end for**
28: **end for**
29: $RelevanceMat = PCR$; {Compute relevances for spatial objects.}
30: $RankList = Rank(RelevanceMat, ascend)$; {Rank objects with ascending PCR values.}
31: $O_l = Outlier(RankList, 1:l)$ {Mark the outliers.}

Step-4(*line:30-32*). **Outlier detection**. Finally, the objects are sorted with ascending relevance values, and the top l objects with lower relevance scores are recognized as outliers.

Algorithm 6 PAS and AS Identification
$[PAS, AS] = ASIdentify(A)$

$Label = 0; m = |A|$;
for $i = 1$ to m **do** {Identify AS,PAS whose size are i, $2 * i$, respectively.}
  $AS((i + Label) : (\binom{m}{i} + Label)), :) = \{AS^i, (AS^i \subseteq A)\ \&\ (|AS^i| = i)\}$;
  $PAS((i + Label) : (\binom{m}{i} + Label)), :) = \{< AS^i, AS^i >, (AS^i \subseteq A)\ \&\ (|AS^i| = i)\}$;
  $Label = \binom{m}{i} + Label$;
end for

Algorithm 7 Identification of $PASC$, ASC, $\mathcal{D}_{ASC}$ and $\mathcal{F}_{PASC}$.
$[PASC, ASC, \mathcal{D}_{ASC}, \mathcal{F}_{PASC}] = ASCIdentify(A, PAS, AS, \mathcal{D})$

$PASC = \{\}; ASC = \{\}; \mathcal{D}_{ASC} = \{\}; \mathcal{D}_{PASC} = \{\}$;
for $i = 1$ to n **do**
  A_j=$r_i.A$;
  for $j = 1$ to k **do**
    $r_j = Neighbor(i, j)$;
    A_j=$r_j.A$;
    for $a = 1$ to $2^m - 1$ **do**
      $\mathcal{AS}$=$AS(a)$;{Identify current AS .}
      $\mathcal{PAS}$=$PAS(a)$;{Identify current PAS.}
      {Generate one of the ASCs based on the attributes of object r_i.}
      ASC_i^a=$A_i\{\mathcal{AS}\}$;
      $ASC_{\mathcal{AS}} = Add\_ASC(ASC_{\mathcal{AS}}, ASC_i^a)$;
      {Add object r_i into its corresponding ASC subset.}
      $\mathcal{D}_{ASC_{\mathcal{AS}}} = AddObjInASCSet(D_{ASC_{\mathcal{AS}}}, r_i)$;
      {Generate one of the ASCs based on the attributes of object r_j.}
      ASC_j^a=$A_j\{\mathcal{AS}\}$;
      $ASC_{\mathcal{AS}} = Add\_ASC(ASC_{\mathcal{AS}}, ASC_j^a)$;
      {Add object r_j into its corresponding ASC subset.}
      $\mathcal{D}_{ASC_{\mathcal{AS}}} = AddObjInASCSet(D_{ASC_{\mathcal{AS}}}, r_j)$;
      {Generate one of the PASCs based on the pair attribute sets of $< r_i, r_j >$.}
      $PASC_{\mathcal{PAS}} = Add\_PASC(PASC_{\mathcal{PAS}}, < ASC_i^a, ASC_j^a >)$;
      {Add $< r_i, r_j >$ into its corresponding PASC subset.}
      $\mathcal{F}^k_{PASC_{\mathcal{PAS}}} = AddObjInPASCSet(F^k_{PASC_{\mathcal{PAS}}}, < r_i, r_j >)$;
    end for
  end for
end for

Computational Complexity. To form the neighborhood, it will take $O(n \log n)$ for *k*NN (Space partitioning) construction and mapping process. As shown in Algorithm 4, it takes around $O(2^m - 1)$ to identify all possible PASs and ASs. Algorithm 5 demonstrates that it takes $O(n * k * (2^m - 1))$ to detect all possible

Table 5.4: Three Simulation Dataset

Dataset	Size	Dimension	The number of observing categories in each dimension
Syn_1	4000	1	A_1: 3
Syn_2	4000	3	A_1: 3 ; A_2: 3 ; A_3: 3 ;
Syn_3	4000	3	A_1: 3 ; A_2: 3 ; A_3: 3 ;
Syn_4	1000	4	A_1: 3 ; A_2: 4 ; A_3: 5 ; A_3: 6 ;
Syn_5	4000	2	A_1: 5 ; A_2: 8 ;

PASCs, ASCs and their corresponding subsets. Finally, computing the final PCR value for each observation takes $O(n * k * (2^m - 1))$. In summary, assuming $n \gg k$ and $n \gg m$, the total computational complexity of *k*NN-SCOD-M approach is $O(n * (2^m - 1))$(= $O(n \log n)$ + $O(2^m - 1)$ + $O(n * k * (2^m - 1))$ + $O(n * k * (2^m - 1))$).

5.5 Experiment Results and Analysis

We conducted extensive experiments on both simulated and real datasets to compare the performances of PCF-SCOD and *k*NN-SCOD, with other popular outlier detection approaches[18, 21, 40, 100, 130].

5.5.1 Experiment Settings

This subsection introduces simulation and real life datasets, the outlier detection methods, and performance metrics.

Simulation and real dataset

For experiments in the single attribute domain, we applied all the approaches into one simulated and three real datasets. For those in the multiple attribute domain, since there was no public baseline dataset, we evaluated them on simulation datasets.

Simulation Dataset

The simulation categorical datasets were generated by discretization from some numerical simulation datasets. Denote a numerical dataset S as $\{Z(s_1),...,Z(s_n)\}$, $Z(s_i) \in R^m$ $(i = 1 \cdots n)$, where *m* is the number of non-spatial attributes. The simulation datasets $S_1, ..., S_m$ were generated by a Gaussian random field model

defined as follows:

$$
\begin{aligned}
&[Z(s_1)^T, ..., Z(s_n)^T]^T \sim N\left(0, \begin{bmatrix} \sum_{11}(\theta_{11}) & \cdots & \sum_{1n}(\theta_{1n}) \\ \vdots & \ddots & \vdots \\ \sum_{n1}(\theta_{n1}) & \cdots & \sum_{nn}(\theta_{nn}) \end{bmatrix}\right) \\
&[\theta_{ij}]_1 \sim Uniform(1.17, 1.85), [\theta_{ij}]_2 \sim Uniform(2.00, 3.24), i = j \\
&[\theta_{ij}]_1 \sim Uniform(1.00, 1.44), [\theta_{ij}]_2 \sim Uniform(2.30, 2.80), i \neq j \\
&s_1, \cdots, s_n \sim Uniform(0, 5).
\end{aligned} \tag{5.19}
$$

where $Z(s_i) = [z_1(s_i), \cdots, z_m(s_i)]^T$, $\sum_{ij}(\theta_{ij}) = Var(Z(s_i), Z(s_j)), \theta \in R^2$. $[\sum_{ij}(\theta_{ij})]_{km}$ is defined by an exponential model as

$$
[\sum_{ij}(\theta_{ij})]_{km} = [\theta_{ij}]_1 \cdot e^{\frac{\|s_i - s_j\|}{|\theta_{ij}|_2}} \tag{5.20}
$$

$[\theta_{ij}]_1$ and $|\theta_{ij}|_2$ are named as sill and range parameters, respectively. The above simulation model parameters were determined based on the distribution of a benchmark data set *97data.dat* available in GSLIB software[147], which has two attributes. It was fitted by the Gaussian random field model which has a quadratic trend (mean), and the cross covariance functions were approximated by exponential models with sill and range parameters (1.85, 2.00) for the first attribute, (1.17, 3.24) for the second attribute, and (1.22, 2.55) for the cross covariance between the two attributes. Note that, in our simulation model, we did not consider any trend, and the data distribution was determined purely by the cross covariance functions, which potentially increases the complexity of the distribution. We did not fix the sill and range parameters, but instead sampled the parameters from uniform distributions around the estimated sill and range parameters for *97data.dat*. Our model was able to flexibly generate spatial simulation data sets with multiple attributes. After the numerical data sets were generated, we applied a discretization process to convert the numerical data into categorical data. To illustrate the discretization strategy, suppose we need to convert a numerical attribute data $\{Z_1(s_1), \cdots, Z_1(s_n)\}$ into 3 categories data, we sorted the values and then separated them into three groups such that the orders among the three groups were preserved. This means, the objects in group 2 is always larger than those in group 1. The similar situation is for group 2 and 3. Since we focused on nominal data, we assigned each data a label with a unique category so that the data attributes could not be ordered.

For the five generated simulation data sets, shown in Table 6, Syn_1 was utilized in the experiments for the single attribute domain, while Syn_2, Syn_3, Syn_4 and Syn_5 were used in those for the multiple attribute one.

Real dataset

We also executed the SCOD approaches on three real datasets with single attribute to further demonstrate their effectiveness. The three datasets include $Jura$, $Soil_1$ and $Soil_2$. $Jura$ data is a well-known categorical dataset from Pieere Goovarerts book[59]. In the original dataset, five rock types are available. Following Bel et al. [14], Portlandian is grouped with Quanternary into category 4, because its frequency of occurrence is very low (1.2%), which makes those observations taking Portlandian general outliers. $Soil_1$ and $Soil_2$ dataset were both extracted from Harmonized World Soil Database[75]. Table 7 describes their detailed information. Fig. 5.5 provides the data distribution of parts of raw soil datasets which were utilized in our experiment. $Soil_1$ data seems to distribute uniformly, but that of $Soil2$ is more complicated, which needs higher identification qualities for SCOD approaches.

Table 5.5: Three Real Datasets

Dataset	Size	Dimension	The observing categories
$Jura$	359	1	A^1:Argovian; A^2:Kimmeridgian; A^3:Sequanian; A^4:Quaternary;
$Soil_1$	1000	1	A^1:LP-Leptosol; A^2:CL-alcisol; A^3:RK-utcrops; A^4:DS-Sand Dunes;
$Soil_2$	3000	1	A^1:LV-Luvisol; A^2:LP-Leptosol; A^3:PT-Plinthosol; A^4:VR-Vertisol; A^5:NT-Nitisol; A^6:LX-Lixisol; A^7:FL-Fluvisol;

Figure 5.5: Data distribution of three real-life datasets.(Left:$Jura$;Middle:$Soil_1$;Right:$Soil_2$)

Outlier Detection Approach

We compared the performances of our proposed methods, denoted as PCF-SCOD and *k*NN-SCOD, to other existing methods introduced in this subsection.

Univariate Detection Methods

TCOD. Considering the local correlation property of spatial data, we chose NN (Nearest Neighbor) based techniques for SCOD tasks. Typical approaches include *k*NN [130] and LOF[21] methodologies. To compute the similarities among nominal data, we used Lin and OF measurements which showed high performances[18, 31]. We combined the NN based techniques with categorical similarity measurements together to get overall 4 different comparable "TCOD" approaches: LOF-Lin, LOF-OF, *k*NN-Lin and *k*NN-OF.

It is noted that *k*NN-SCOD is not related to *k*NN-Lin and *k*NN-OF, although they have the same prefix "*k*NN". As we introduced above, *k*NN-Lin and *k*NN-OF were generated from one of the most popular traditional numerical outlier detection approaches: *k*NN [130], while *k*NN-SCOD is proposed in this paper by introducing a novel *k*NN mapping function process as an effective and efficient approximation of the general PCR computation.

SCOD. Z-test is one of the most typical methods to identify SNOs. When operating it on categorical data, we integrated Z-test with Lin and OF measurements. As a result, there were 2 comparable "SCOD" approaches: Z-OF and Z-Lin. Also, we directly applied Z-test into the categorical datasets by assuming that the nominal categories can be ordered, denoted by Z-SNOD.

Multivariate Detection Methods

TCOD. Several advanced TCOD methods have been proposed for multivariate categorical data, including Bayes Net Method, Marginal Method, LERAD, Conditional Test, Conditional Test-Combining Evidence, and Conditional Test-Partitioning. Experiments had shown that Conditional Test and its two variants outperformed all other methods[40]. Therefore, we focused on the comparison of our method with the three best methods, denoted as Conditional Test, Conditional Test-Combining Evidence, and Conditional Test-Partitioning.

SCOD. For the competing methods in SNOD group, we chose Multivariate Z-SNOD, which is an extension version of single attribute Z-test, by considering Mahalanobis distance and MCD (Minimum Covariance Determinant) techniques. In addition, we integrated the preceding method with multiple categorical similarity measurements, Lin and OF, named as Multivariate Z-Lin, and Multivariate Z-OF.

Performance metrics

We generated synthetic outliers in both simulation and real datasets, which enable us to analyze the effectiveness of outlier detection approaches in a controllable way. We assumed the raw dataset as a ground truth, and contaminated around α percent of the data records as outliers. In our paper, for each dataset, including both simulation and real life ones, we randomly selected 2%, 3% and 5% of the data to be anomalies by modifying them from its original category to anyone of others. For each contamination rate (*2*, *3* and

5), the synthetic outliers were generated 10 times and the mean and standard deviation of accuracies were calculated for each method.

To compare the accuracies among all methods, we used the common evaluation measures: *precision* (detection rate), i.e., *the fraction of examples labeled as outliers that are true outliers*, and *recall* (detection precision), i.e., *the fraction of true outliers that are correctly identified*. The precision is plotted against recall, and the curves that are higher and farther to the right denote better performances since it corresponds to a higher precision for a given recall. Each point corresponds precision and recall when a specified number of outlier is predefined, from 1 to n (the number of objects in the whole dataset). As another measure of accuracy, *average precision* was computed to approximate the area under the precision-recall curve.

All the experiments were conducted in a PC with Intel (R) Core (TM) Duo CPU, CPU 2.80 GHz, and 2.00 GB memory. The development tool was MATLAB 2008.

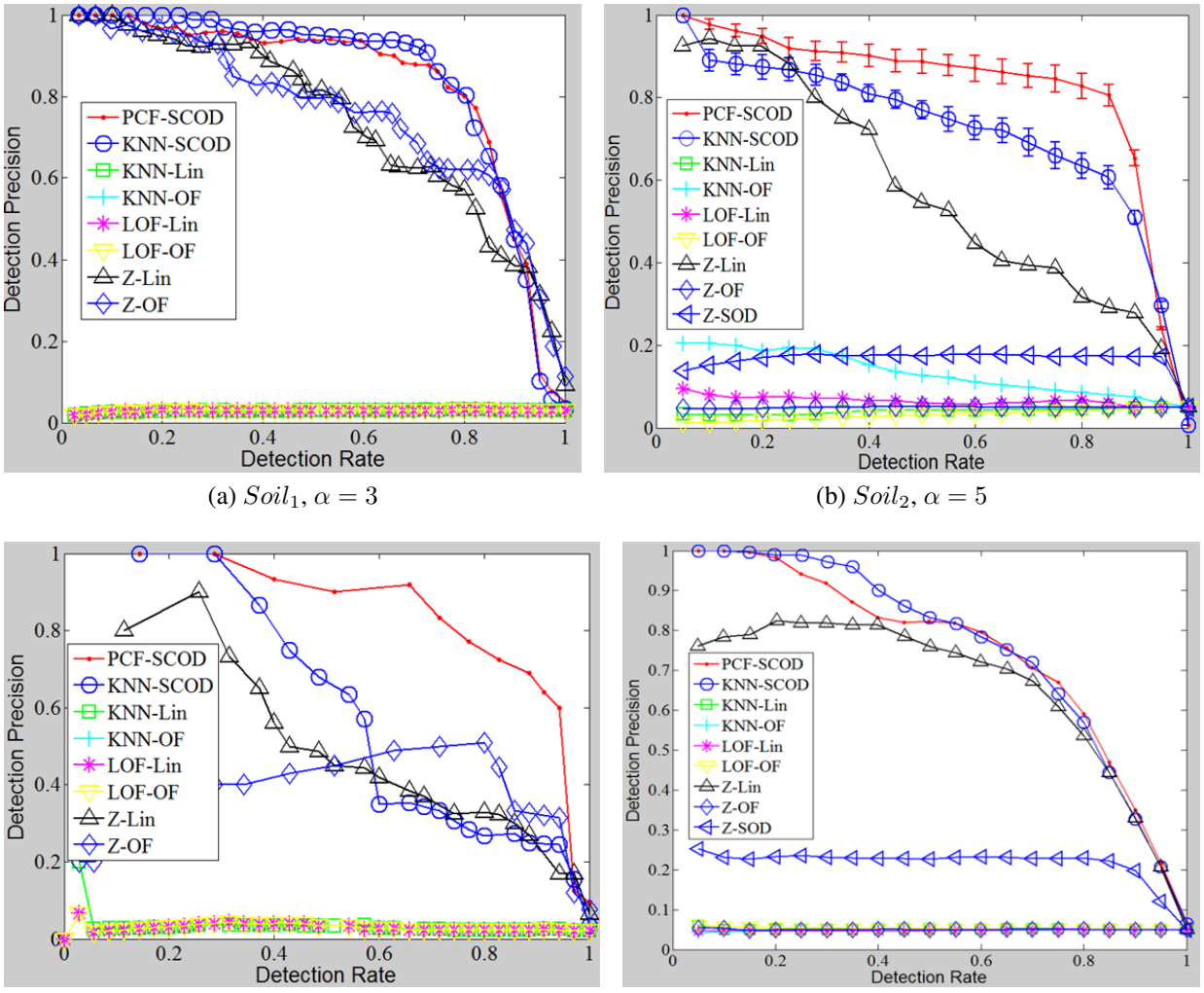
5.5.2 Experiment Results and Analysis

This section presents experimental evaluations for the above approaches on simulation and real datasets. We compared the SCOD accuracies among different methods based on different parameter combinations.

Results on single attribute datasets

Detection Accuracy

Fig. 5.6 depicts the comparison of our methods against the other 7 existing approaches on the single attribute datasets. The contamination rate α was set as *3*, *5*, *2* and *5* in $Soil_1$, $Soil_2$, $Jura$ and Syn_1, respectively. Each point in the curves corresponds to the average performance over *10* randomly generated datasets for each algorithm. We observed that both PCF-SCOD and *k*NN-SCOD methods achieved $20-40\%$ improvement over Z-OF and Z-Lin, and $60-70\%$ over LOF-Lin, LOF-OF, *k*NN-Lin, *k*NN-OF and Z-SNOD(Z-SOD). From these results, we found that *k*NN and LOF can't handle the categorical outlier detection in spatial context. After integrating Z-test with OF and Lin similarity measurements together, the outlier identification quality was increased. Z-Lin was always better than Z-OF. As introduced in [18], OF



(a) $Soil_1$, $\alpha = 3$ (b) $Soil_2$, $\alpha = 5$

(c) $Jura$, $\alpha = 2$ (d) Syn_1, $\alpha = 5$

Figure 5.6: Comparison of algorithm performances for the spatial dataset with single attribute

and Lin compute similarities for categorical attributes in different ways:

$$Sim_{OF}(X,Y) = \begin{cases} 1 & \text{if } X = Y; \\ \frac{1}{1+\log(N/f_k(X_k))\times\log(N/f_k(Y_k))} & \text{otherwise.} \end{cases}$$
$$\omega_k = 1/d \tag{5.21}$$

$$Sim_{Lin}(X,Y) = \begin{cases} 2\log p_k(X_k) & \text{if } X = Y; \\ 2\log(p_k(X_k) + p_k(Y_k)) & \text{otherwise.} \end{cases}$$
$$\omega_k = \frac{1}{\Sigma_{i=1}^{d} \log p_i(X_i) + \log p_i(Y_i)} \tag{5.22}$$

Table 5.6: Average precision (normalized area under precision-recall curve) for spatial categorical datasets with single attribute, comparing PCF-SCOD, *k*NN-SCOD-S and other 7 approaches.

Approach	$Soil_1$	$Soil_2$	$Jura$	Syn_1
PCF-SCOD	0.7805	0.7822	0.7481	0.7646
*k*NN-SCOD-S	0.7811	0.7763	0.6521	0.7148
*k*NN-Lin	0.0279	0.0389	0.0276	0.0489
*k*NN-OF	0.0284	0.1261	0.0279	0.0454
LOF-Lin	0.0284	0.0621	0.0279	0.0455
LOF-OF	0.0284	0.0300	0.0279	0.0502
Z-Lin	0.6781	0.5407	0.4362	0.6298
Z-OF	0.6966	0.0473	0.3668	0.0477
Z-SNOD	0.1845	0.1603	0.0807	0.2072

where $f_k(X_k)$ denotes the number of times attribute to take the value X in the k^{th} dimension, $p_k(X_k)$ the sample probability to take the value X_k in the dataset, and ω_k is the weight value of the k^{th} dimension. When identifying the relevance score among objects with the same categorical attribute, OF always assigns a constant value 1 to it, while Lin computes the value based on the occurrence probability of the category. When two objects have different categorical attributes, OF assigns lower relevance to the objects with higher frequencies, while Lin assigns higher values. Consequently, Lin could better capture the spatial relationship than OF by more accurately computing spatial relevance among objects. If the category of one of the pair objects occurs frequently in the dataset, it means the higher probability to co-occur with another category in the whole dataset. That is why the methods integrating with Lin always achieved better performance than those with OF. However, when identifying the spatial categorical outliers, Lin and OF measurements are based on the category frequencies that are determined by the whole data distribution, not the co-occurrence frequencies which take spatial dependency into considerations. That was why the performances of Z-Lin and Z-OF were worse than those of PCF-SCOD and *k*NN-SCOD.

Compared with *k*NN-SCOD, PCF-SCOD had better performance. Especially, when applied to the datasets with more complicated distribution, like $Jura$ and $Soil_2$ (as shown in Fig. 5.5), PCF-SCOD can accurately capture the relationships among objects by considering PCRs among pair of categories at different spatial distances. Furthermore, it can get the highest precision of identifying spatial categorical outliers. Also, with the increasing data size, *k*NN-SCOD got better approximations of PCF-SCOD, like in $Soil_1$(1000), $Soil_2$(3000) and Syn_1(4000), since with the more objects in the dataset, *k*NN-SCOD could capture sufficient mapping information from the raw dataset, which helps accurately approximate PCRs for pair objects. Finally, we found that the identification qualify of PCF-SCOD was not affected by different contamination rates. For each contamination value, PCF-SCOD always achieves higher accuracy with stable process

abilities, as shown by its small standard deviations of detection precisions in Fig. 5.6(b).

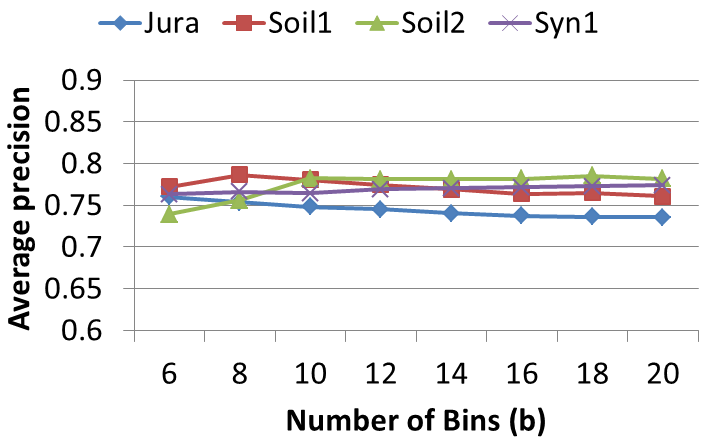


Figure 5.7: Average precisions of PCF-SCOD by varying b value

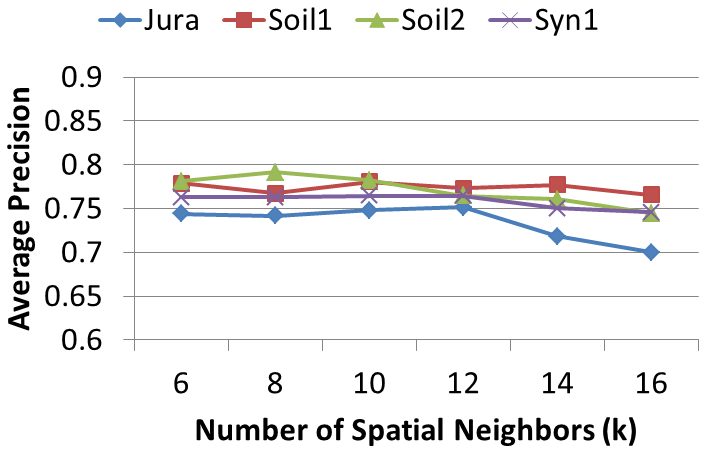
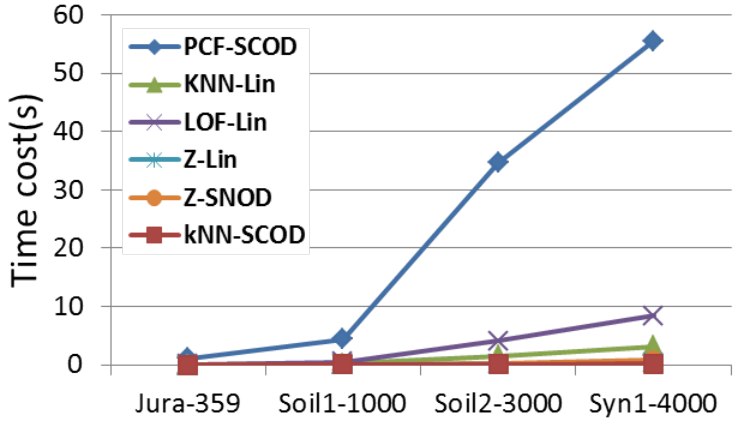


Figure 5.8: Average precisions of PCF-SCOD by varying k value

The average precision values are also given in Table 8 for all the SCOD approaches in single attribute domain. Note that for most datasets, PCF-SCOD and kNN-SCOD achieved higher accuracy than other approaches. We notice that the performance of the methods also depends on the detection tasks. For example, kNN-SCOD has comparable or better performance than PCF-SCOD in $Soil_1$ and Syn_1. In $Soil_1$, only 3% of data are contaminated which makes the outlying behavior more obvious based on the information derived from the normal objects. There are 3 categories in Syn_1, and only 6 possible pair attributes. It is sufficient for 4000 observations to extract the normal pair attributes which co-occur frequently by analyzing those behaviors. On the contrary, kNN-SCOD can't work as well as PCF-SCOD in $Jura$. There are only 359 observations which take four different categories, which means there are overall 10 pair attributes. The

neighborhood information in these 359 observations can't provide substantial information to help derive the normal behaviors. On the other hand, by observing Fig. 5.5, we notice that the distribution of the whole dataset is not uniform. It seems that there are various kinds of pair attributes co-occurring in near distances, such as, blue-blue, blue-orange, red-red, red-orange, red-while, and white-orange, etc. In this sense, only considering the neighborhood based information is not sufficient to mine the normal patterns. Although some pair attributes often co-occur within a near region, but they maybe not within a little distant one. PCF-SCOD can extract the co-occurrence frequencies of pair attributes at different distances. That is why PCF-SCOD performs well in $Jura_1$.



Data Size	359	1000	3000	4000
PCN-SCOD	1.14	4.41	34.6	55.5
*k*NN-SCOD	0.01	0.02	0.11	0.18
*k*NN-Lin	0.01	0.14	1.48	3.14
LOF-Lin	0.04	0.44	4.11	8.39
Z-Lin	0.001	0.02	0.12	0.23
Z-SOD	0.001	0.02	0.16	0.75

Figure 5.9: Runtime in seconds for datasets with varying size

Impacts of neighborhood sizes

We also evaluated the anomaly detection performances of proposed approaches by varying the sizes of spatial neighborhood. Fig. 5.8 shows various k values from 6 to 16, respectively. The curves depict the effects of varying the number of spatial neighbors on the average precisions of PCF-SCOD on 4 different sizes of datasets. In general, its anomaly detection performance seems stable as the neighborhood size increases. In $Soil_1$, $Soil_2$ and Syn_1, the optimal k values are around 8 to 14. But for $Jura$, which is a small-size data set, higher k value leads to a worse performance. This is because, higher neighborhood size makes distant objects involve in evaluating the specified observation behaviors, which violates the spatial

correlation theory. This same situation occurred in the results generated by *k*NN-SCOD method, as shown in Fig. 5.10. Since *k*NN-SCOD work is based on the neighborhood information, which makes it more sensitive to the *k* value. For the dataset with larger data size, 8-16 neighborhood size is appropriate to collect the co-occurrence information of pair attributes. This is proved by the curves of $Soil_1$, $Soil_2$ and Syn_1. For small dataset, both lower and higher k values result in worse identification performances.

Impact of bin sizes

The effects of bin sizes were examined on the performances of PCF-SCOD method. Fig. 5.7 shows its performances keep impressive by varying b values from 6 to 20. Apparently, SCOD identification quality achieves stable after the points at 10. PCF-SCOD computes the pair attribute frequencies at different bins. If *b* is smaller, e.g., $b < 8$, the pair observations in more distant region are mixed with those in nearer region that we are interested. This results in the incorrect computation of the co-occurrence frequency of pair attributes, which further leads to the worse identification performances. On the contrary, higher *b* value helps compute the pair frequency more accurately. However, too much bins will cost a lot, and normally, it is sufficient to set *b* as 10, which is demonstrated by the curves in Fig. 5.8.

Computational Cost Analysis

Finally, we showcase the speed and respective scalability of the algorithms. Fig. 5.9 contains the runtime performances of algorithms in the datasets with varying number of data observations. As observed, the methods based on Lin have similar runtime with those based on OF. Therefore, we only show the Lin based approaches. As shown in Fig. 5.9, PCF-SCOD finished execution at around *1.14* seconds for *Jura* data, , while *k*NN-SCOD had a running time of *0.01* seconds. And, in Syn_2 dataset, PCF-SCOD is at around *55.5* seconds, while *k*NN-SCOD is at only *0.18* seconds. By analyzing the results in Fig. 5.9, *k*NN-SCOD approximated the accuracy of PCF-SCOD very well, while it outperformed PCF-SCOD for larger-size datasets. For other compared approaches, although they finished running more quickly than PCF-SCOD, they had lower identification accuracies.

Results on multiple attribute datasets

Detection Accuracy

Fig. 5.8 shows the performances when contamination rate was 5% in simulation datasets. Obviously, *k*NN-SCOD-M still had very preceding performance increases. The curves demonstrate that *k*NN-SCOD performs the best, followed by the Multivariate Z-Lin and Multivariate Z-SNOD. The series of Conditional Test perform very poorly in comparison. The worst one is Multivariate Z-OF, since OF measurement can't handle well the similarities among nominal data with multiple attributes. Meanwhile, the curves of all methods are also depicted with the standard variances of precision values of the *10* randomly generated

Table 5.7: Average Precision for spatial categorical datasets with multiple attribute datasets, comparing *k*NN-SCOD-M and other six approaches.

Approach	Syn_2	Syn_3	Syn_4	Syn_5
*k*NN-SCOD-M	0.7282	0.7862	0.8003	0.7886
Conditional Test	0.0526	0.2084	0.3895	0.2509
Conditional Test-Combining Evidence	0.0526	0.2084	0.4084	0.2509
Conditional Test-Partitioning	0.0526	0.2082	0.3948	0.2468
Multivariate Z-SNOD	0.3984	0.5536	0.3596	0.4257
Multivariate Z-Lin	0.5950	0.6413	0.5498	0.4068
Multivariate Z-OF	0.0442	0.0512	0.0476	0.0512

datasets. The smaller standard variance of *k*NN-SCOD indicates that it has more stable performance to detect spatial multivariate categorical outliers.

Similarly, TCOD approaches didn't get competitive results when being applied into the spatial context, as shown by the ROC curves generated by Conditional Test, Combining Evidence, and Conditional Test-Partitioning. The performance of Multivariate Z-Lin is still much better than that of Multivariate Z-OF in the multiple dimension domain. Besides the different ways of similarity computation for a specific attribute domain, OF and Lin applied different weight values when calculating the final similarities by combining the different values from different attribute domain. OF assigns the same weight to different attributes, while Lin computes the corresponding weigh based on the category distribution in it. As shown in Eq.(5.25), Lin measure gives higher weight to the same categories with frequent values, and lower weight to different categories with infrequent values. Such a way could better reflect the case that if two objects having same category co-occur with higher frequency, they will have higher relevances. Whereas, if they co-occur frequently with different categories, their relevance score will be lower since they are assigned an lower weight. However, such measurement to capture the relevance among pair objects with different categories is significant inconsistent with the concept of SCOs. That is why the performance of Multivariate-Lin is much worse than that of kNN-SCOD. It is worthy of note that Z-SNOD approach performs well compared with TCOD methods since it takes the characteristic of spatial auto-correlation into considerations when identifying SCOs, although it treats the categorical attributes as numerical ones.

Similarly, TCOD approaches didn't achieve competitive results when applied to the spatial context, as shown by the PR(Precision-Recall) curves generated by Conditional Test, Combining Evidence, and Conditional Test-Partitioning. The performance of Multivariate Z-Lin was still much better than that of Multivariate Z-OF in the multiple dimension domain. Besides the different ways of similarity computation for a

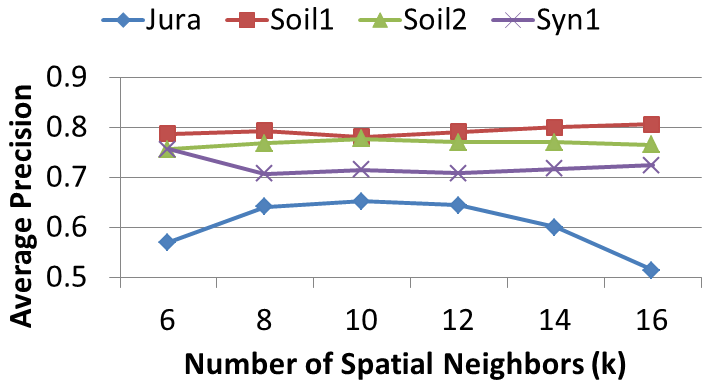


Figure 5.10: Average precisions of kNN-SCOD-S by varying k value

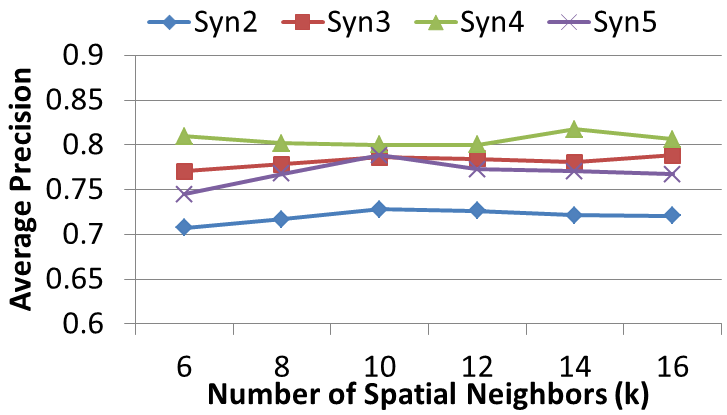
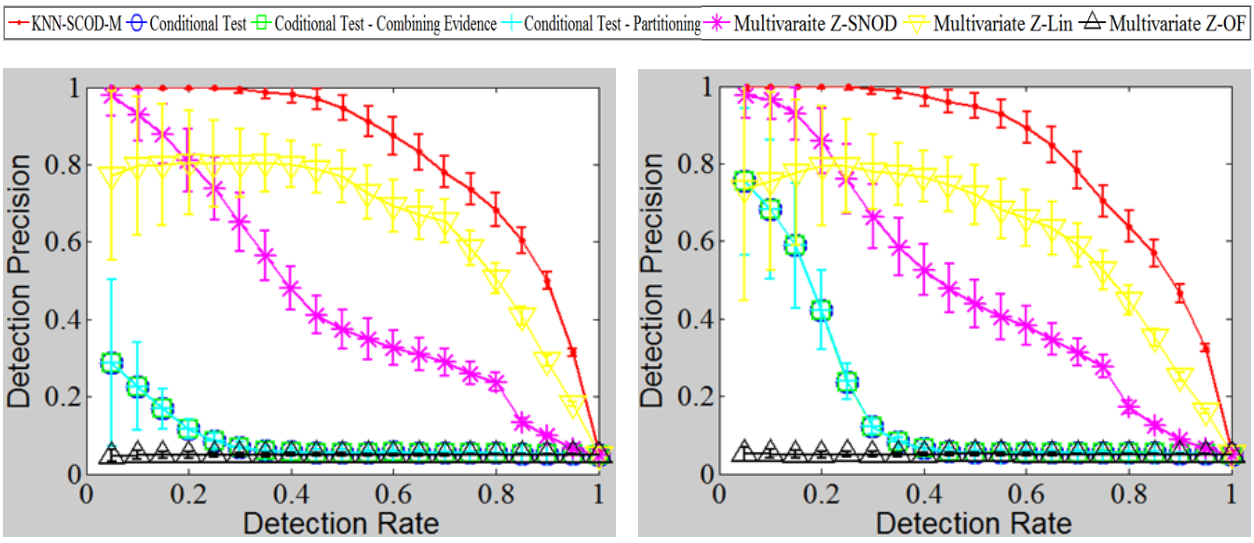


Figure 5.11: Average precisions of kNN-SCOD-M by varying k value

specific attribute domain, OF and Lin applied different weight values when calculating the final similarities by combining the different values from different attribute domains. OF assigns the same weight to different attributes, while Lin computes the corresponding weigh based on the category distribution in it. As shown in Eq.(25), Lin measure gives higher weight to the same categories with frequent values, and lower weight to different categories with infrequent values. Such a way could better reflect the case that if two objects having the same category co-occur with a higher frequency, they will have higher relevance. Whereas, if they co-occur frequently with different categories, their relevance score will be lower since they are assigned a lower weight. However, such measurement to capture the relevance between pair objects with different categories is significant inconsistent with the concept of SCOs. That is why the performance of Multivariate-Lin is much worse than that of kNN-SCOD. It is worthy to note that Z-SNOD approach performs well compared with TCOD methods since it takes the characteristic of spatial auto-correlation into considerations when identifying SCOs, although it treats the categorical attributes as numerical ones.



(a) Syn_2, $\alpha = 5$ (b) Syn_3, $\alpha = 5$

Figure 5.12: Comparison of algorithm performances for the spatial data with multiple attributes

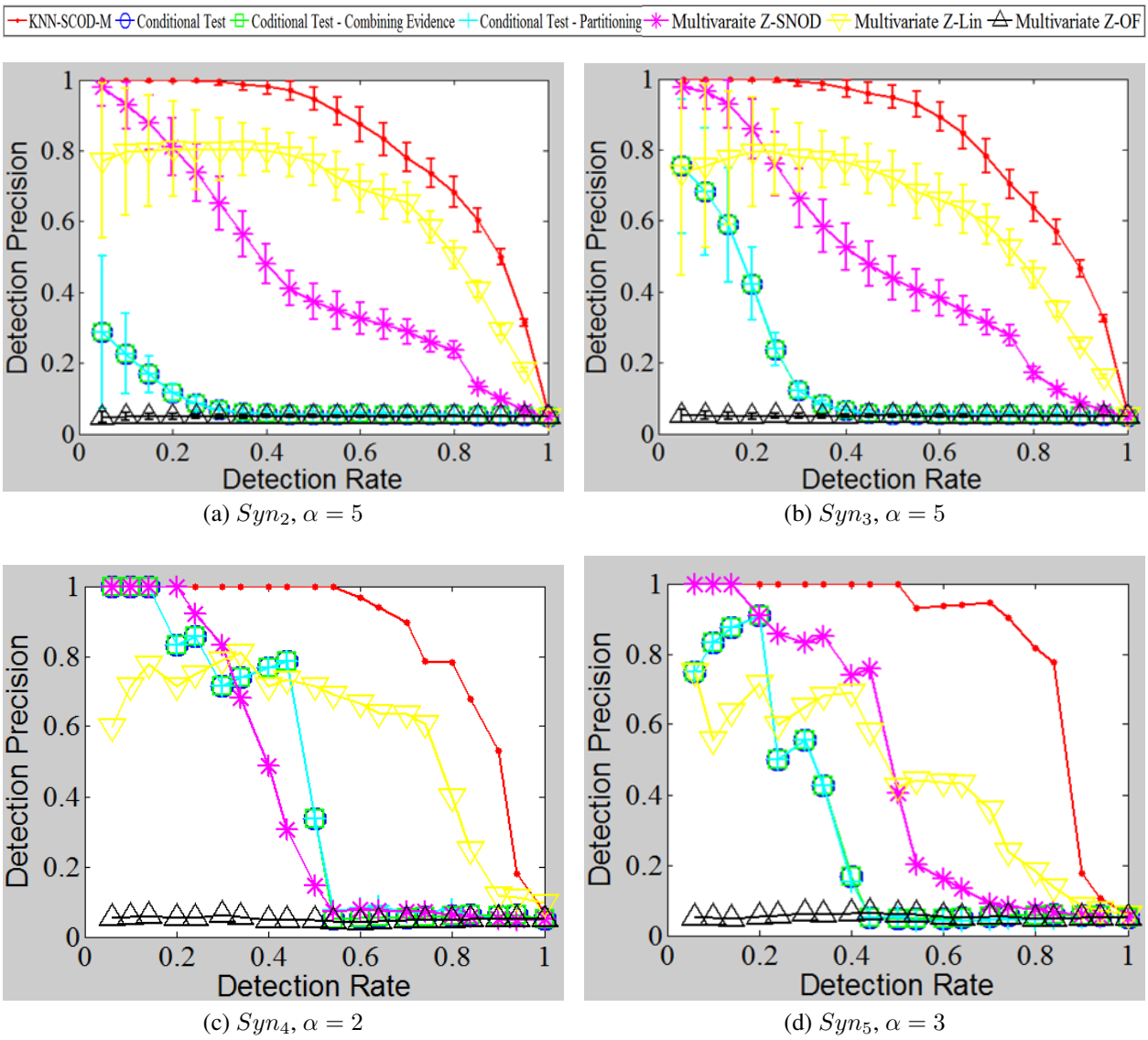
The average precision values are given in Table 9 for all the SCOD approaches in the multi-attribute domain. Note that for most datasets,*k*NN-SCOD achieved much higher accuracy, from 0.7282 to 0.8003, than other approaches, from 0.0442 for Multivariate Z-OF, to 0.6413 for Multivariate Z-Lin, and 0.5536 for Multivariate Z-SNOD. Similarly, the performance of the methods also depends on the detection tasks. In Syn_4 and Syn_5, the contamination rate are 2 and 3, respectively, which means there exist less outliers in the whole data sets. The lower contaminated data alleviated the side-effects of outlying behaviors on the identification quality of *k*NN-SCOD-M approach. This is also demonstrated by most of the higher average precisions generated by other methods, like the Conditional Test series. For Multivariate Z series, they all performs poorer in Syn_4 compared in other datasets, since there are 4 dimensions in it. It is apt to lose information when computing the spatial relevance by integrating with Mahalanobis distance if there exist more attributes in datasets.

Impacts of neighborhood sizes

In the same way, we showcase the effects of neighborhood size on the performance of *k*NN-SCOD-M on multiple attribute domain. Fig. 5.11 depicts its identification quality is not sensitive to different sizes, from 6 to 16, of spatial neighborhood. The sizes of these four data sets are 1000 and 4000. It is sufficient to set the *k* as around 8 to perform the computation of spatial relevance scores.

Analysis and discussion

Based on the above experimental evaluations, PCF techniques have shown to be very effective in modeling the relevances among spatial category objects in both single and multiple attribute datasets. As a result,



(a) Syn_2, $\alpha = 5$

(b) Syn_3, $\alpha = 5$

(c) Syn_4, $\alpha = 2$

(d) Syn_5, $\alpha = 3$

Figure 5.13: Comparison of algorithm performances for the spatial data with multiple attributes

PCF-SCOD and kNN-SCOD demonstrated superior identification qualities over the competing techniques in both simulated and real datasets. The evaluations verify two observations: 1) first, SCOs are identified in a different way with that of SNOs. Two objects taking different attributes are not necessarily irrelevant with each other. Sometimes, their frequent co-occurrence exactly illustrates their higher spatial correlation. This can be demonstrated by comparisons of Z-SNOD against PCF series of methods; 2) when identifying SCOs, the existing TCOD and SCOD approaches can't avoid the well-known swamping and masking problems. TCOD approaches treat spatial and non-spatial attributes equally and don't take the spatial dependency and spatial auto-correlation into considerations, which are the specific properties of spatial data. Z-OF and Z-

Lin outperformed TCOD methods since they differentiate spatial and non-spatial attributes. However, they performed worse than PCF-SCOD and kNN-SCOD since their dissimilarity computation is based on the global frequencies, not local frequencies.

Notice that there might be white noise in the original data set. Considering that data noise is usually uniformly distributed over the space, our defined pair correlation ratio is able to capture the spatial correlation as a small but nontrivial ratio value between noise observations and normal observations as long as there exist some correlation patterns between them based on their spatial distances. The pair correlation ratio will increase if the signal-noise ratio decreases. In the situation with high signal-noise ratio, noise will not be identified as outliers. In the situation with low signal-noise ratio, some noise observations may be identified as outliers, but should not be highly ranked outliers. In the extreme case where the signal-noise ratio is very high, then all the noise observations will be returned as top ranked outliers, since they are rare observations and should be regarded as outliers.

This paper assumes that the spatial locations are uniformly distributed. The case of sparsely distributed data may refer to two scenarios. The first scenario refers to the situation where the data set has a very low signal-to-noise ratio. In this case, noise will be handled well as explained above. The second scenario refers to the situation where some categorical types are rare. This still depends on the sample size of these rare categorical types is still sufficient to calculate the stable pair correlation ratios. If it is not sufficient, we may need to remove them in the pre-processing step, since we are not able to calculate stable statistics for them. However, note that sparse distribution is not common in spatial categorical data. The datasets that we collected are all not sparsely distributed.

5.6 Conclusion

This paper investigates the benefits of PCF technique on the SCOD, and designs three algorithms which can identify SCOs with single and multiple attributes. General idea in PCF-SCOD is that, first, for each pair of categories, we compute its Pair Correlation Ratios (PCR) as a function of distance. Then, the outlier scores are computed by the mean of estimated PCR values between each object and its spatial neighbors. Finally, the top l objects with higher infrequent behaviors are recognized as SCOs. Furthermore, we propose two kNN based estimators which only utilize kNN neighborhood information to estimate the co-occurrence frequency of pair objects in single and multiple attribute domains, respectively. The proposed approaches have several advantages: 1)they can identify SCOs with both single and multiple categorical attributes; 2) they can process not only ordinal, but nominal categorical datasets; 3)compared with existing approaches, they can better avoid swamping and masking issues. The experiments conducted on the synthetic and real datasets demonstrated PCF series of approaches significantly outperformed other existing popular approaches.

Chapter 6

Robust Prediction and Outlier Detection for Spatial Datasets

Spatial kriging is a widely used predictive model for spatial datasets. In spatial kriging model, the observations are assumed to be Gaussian for computational convenience. However, its predictive accuracy could be significantly compromised if the observations are contaminated by outliers. This deficiency can be systematically addressed by increasing the robustness of spatial kriging model using heavy tailed distributions, such as the Huber, Laplace, and Student's t distributions. This paper presents a novel Robust and Reduced Rank Spatial Kriging Model (R^3-SKM), which is resilient to the influences of outliers and allows for fast spatial inference. Furthermore, three effective and efficient algorithms are proposed based on R^3-SKM framework that can perform robust parameter estimation, spatial prediction, and spatial outlier detection with a linear-order time complexity. Extensive experiments on both simulated and real data sets demonstrated the robustness and efficiency of our proposed techniques.

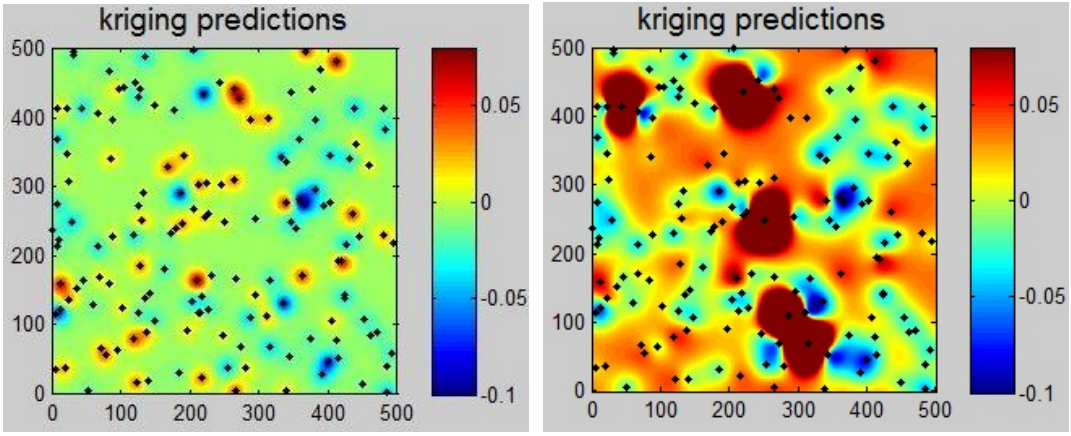
This chapter is organized as follows. Section 6.1 gives the background and motivation. Section 6.2 reviews the theoretical preliminaries. Section 6.3 presents the R^3-SKM framework. A general approach based on R^3-SKM is proposed to perform robust parameter estimation in Section 6.4, and two inference algorithms are discussed in Section 6.5. Experiments on both simulated and real datasets are presented in Section 6.6. The paper concludes with a summary of the research in Section 6.7.

6.1 Background and Motivation

With the increasing public sensitivity and concern on environmental issues, as well as the development of remote sensing technologies, huge amounts of spatial data have been collected from location based social network applications to scientific data, and the volume keeps increasing at fast pace over recent decades. Especially, the rapid advances in Geographical Information System (GIS) and Global Positioning System (GPS) enable accurate geo-coding of locations where scientific data are collected. As one of the major research issues, the prediction of spatial data has attracted significant considerations. Illustrative applications include climate prediction, environmental monitoring, molecular dynamical pattern mining, and infectious disease outbreak prediction in fields such as environmental monitoring, biology, epidemiology, geography, and economics. Spatial prediction is the process of estimating the values of a target quantity at unobserved locations. When applied to a whole study area, it is also referred to as spatial interpolation or mapping. Given the large volume of spatial data, it is computationally challenging to apply traditional prediction methods in either an allowable memory space limit or an acceptable time limit, even in supercomputing environments. Efficient prediction for large spatial data has therefore become one of the emerging challenges in data mining fields. Most existing spatial prediction methods have the time complexity of $O(n^3)$. Recently, a number of approximate methods have been proposed to tackle the "Big N" problem using different techniques, such as kernel convolutions [73], low rank basis functions or splines [95], moving averages, likelihood approximation [123], and Markov random field [39]. Recent advance by Banerjee et al. [11] proposed a reduce rank spatial kriging approach that projects the spatial process onto a subspace generated by realizations of the original process at a specific set of locations named as knots. All these methods assume that the observations follow a multivariate Gaussian distribution.

However, a well-known limitation with the above Gaussian observation model is non-robustness. In many regression problems, observations may include outliers which deviate strongly from the other members of the sample. Such outliers may occur, for example, because of failures in the measurement process or absence of certain relevant explanatory variables in the model. The estimations of the mean and variance-covariance matrices are sensitive to outliers due to the well-known masking and swamping effects [105]. In addition to the impacts on parameters estimation, outliers also significantly reduce the accuracy of spatial predictions. For example, a single corrupted observation will deviate the posterior expectation of predictions at unobserved locations far away from the level described by the other observations. As demonstrated by Figure 6.1, the kriging prediction result is heavily distorted by the existence of 5 outliers (in the dark red areas in Figure 6.1(b)). In such cases, a robust observation is required. This limitation can be properly addressed by considering robust statistics techniques.

Robust inference has been studied extensively. De Finetti [41] described how Bayesian inference on the mean of a random sample, assuming a suitable observation model, naturally leads to giving less weight to



(a) Original Data (b) Contaminated Data

Figure 6.1: Impacts of spatial outliers on prediction

outlying observations. However, in constrast to simple rejection of outliers, the posterior depends on all data but in the limit, as the separation between the outliers and the rest of the data increases, the effect of outliers becomes negligible. More theoretical results on this kind of outliers rejection were presented by Dawid [41] who gave sufficient conditions on the observation model $p(y|\theta)$ and the prior distribution $p(\theta)$ of an unknown location parameter θ, which ensure that the posterior expectation of a given function $m(\theta)$ tends to the prior as $y \to \infty$. He also stated that the Student-t distribution combined with a normal prior has the property.

A more formal definition of robustness was give by O'Hagan [117] in terms of an outlier-prone observation model. The observation model is defined to be outlier-prone of order n. That is, the effect of a single conflicting observation to the posterior becomes asymptotically negligible as the observation approaches infinity. O'Hagan showed that the Student-t distribution is outlier prone of order 1, and that it can reject up to m outliers if there are at least 2m observations altogether. this contrasts heavily with the commonly used Gaussian observation model in which each observation influences the posterior no matter how far it is from the others.

Currently, a number of robust methods have been proposed for different learning problems, including multivariate regression, Kalman filtering and smoothing, clustering, and independent component analysis. (e.g., [1, 47, 67, 80, 83, 105, 116, 151]). The majority of these methods can be summarized by using a probabilistic framework [105] in which the measurement error is modeled by a heavy tailed distribution, such as the Huber, Laplace, Student's t, and Cauchy distributions, instead of the traditional Gaussian distribution. The prediction problem can then be reformulated as a Maximum-A-Posterior (MAP) prediction problem conditional on observations. However, employing heavy tailed distributions makes the prediction process analytically intractable. Although stochastic simulation methods have been applied to estimate an approximate

posterior distribution, for example, via MCMC or particle filtering [47, 67, 83] , these versatile methods are very computationally intensive. Jylanki et al. [84] presented an efficient expectation propagation algorithm for robust Gaussian process regression based on the Student's t distribution, while Svensn and Bishop [151] proposed a variational inference approach to robust Student's t mixture clustering. Gandhi and Mili [53] proposed a robust Kalman filter based on the Huber distribution and the iterative reweighted least squares (IRLS) method. An efficient Kalman smoother was presented by Aravkin et al. [8] based on the Laplace distribution and the convex composite extension of the Gauss-Newton method.

This paper aims to address the robust prediction problem for large spatial dataset. It considers the same probabilistic framework as that was used in existing robust methods. Specifically, a Robust and Reduced-Rank Spatial Kriging Model (R^3-SKM) is formulated, and then efficient algorithms are proposed by utilizing Laplace approximation to perform parameter estimation, robust spatial prediction, and spatial outlier detection.

To the best of our knowledge, this is the first statistical approach that can perform robust spatial prediction in linear time. The main contributions can be summarized as follows:

- **Formulation of the R^3-SKM model.** A Robust and Reduced Rank Spatial Kriging Model is proposed in which the measurement error is modeled by a heavy tailed distribution, and a Bayesian hierarchical framework is integrated to support priors on model parameters.
- **Design of an approximate algorithm for robust parameter estimation.** The posterior distribution of latent variables conditional on parameters and observations is estimated via Gaussian approximation. Furthermore, the posterior distribution of parameters conditional on observations is estimated via Laplace approximation. It has time complexity of $O(n)$.
- **Development of robust inference algorithms.** R^3-SP (Robust and Reduced Rank Spatial Prediction) and R^3-SOD (Robust and Reduced Rank Spatial Outlier Detection) algorithms are proposed to perform robust spatial prediction and spatial outlier detection. Their time complexities are analyzed, which scale linearly.
- **Comprehensive experiments to validate the robustness and efficiency of the proposed techniques**. The R^3-SKM was evaluated by the extensive experiments on simulated and real datasets. The results demonstrated that the three algorithms based on R^3-SKM outperformed existing representative techniques, when the data were contaminated by outliers.

6.2 Preliminary Concept

In this section, we review the approximate inference methods considered in this paper. First we give a short description of Spatial Kriging Model (SKM). Then we review knot based reduced-rank techniques, as well as the Laplace Approximation.

6.2.1 Spatial Kriging Model

Let us define a numerical random field $Y(s)$ on a domain $D \subseteq \mathcal{R}^2$, and $Y = (Y(s_1), \cdots, Y(s_n))'$ be the $n \times 1$ vector of observed responses, each of which is along with a $p \times 1$ vector of spatially referenced predictors $x(s)$. The associated spatial kriging model can be represented as

$$Y(s) = x^T(s)\beta + \eta(s) + \epsilon(s) \tag{6.1}$$

where $\epsilon(s)$ is a spatial white noise process with mean zero, $var(\epsilon(s)) = \tau^2 > 0$, and τ^2 is a parameter to be estimated. The white noise assumption implies that $R_{i,j}(\phi) = cov(\epsilon(s_i), \epsilon(s_j)) = 0$, unless $i = j$. $x(s)$ refers to a vector of known predictors, and the coefficients β are unknown. $x^T(s)\beta$ is a vector of deterministic (spatial mean) or trend functions, which models large scale variations, and the spatial random process $\eta(s)$ captures the small scale variations. The hidden process $\eta(s)$ captures spatial association. It is assumed to follow a Gaussian process with zero mean and the covariance function $\sigma^2 C(s, s'; \phi)$, where σ^2 refers to the variance, and $C(\cdot; \phi)$ the correlation function of the process controlled by the parameter ϕ. Function C controls the smoothness and scale among latent variables $\eta(s_i)$, and can be selected freely as long as the resulting covariance matrix is symmetric and positive semi-definite.

6.2.2 Reduced Rank Methodology

The spatial inference (e.g., spatial prediction, outlier detection) based on the SKM model involves the inversion of the n by n correlation matrix, which has the time complexity of $O(n^3)$. This makes the SKM model prohibitively expensive for large n. The knot-based model proposed by Banerjee et al. [11] considers a fixed set of "knots" $S^* = (s_1^*, \cdots, s_{n^*}^*)$ with $n^* \ll n$. The Gaussian process $\eta^*(s)$ yields an n^*-vector of realizations over the knots, that is, $\eta^* = (\eta(s_1^*), \cdots, \eta(s_{n^*}^*))$, which follows a $GP\{0, C^*(s_i^*, s_j^*; \theta)\}$. Spatial estimation at a generic site s is operated through

$$\tilde{\eta}(s) = E\{\eta(s)|\eta^*\} = c^T(s; \theta)C^{*-1}(\theta)\eta^* \tag{6.2}$$

Table 6.1: Description of Major Symbols

Sym.	Description
S	$S = \{s_i\}_{i=1}^n$, a set of n training locations.
S^*	$S^* = \{s_i^*\}_{i=1}^m$, a set of m knot locations.
Y	A given set of observations with numerical attributes which follow Gaussian distribution. $Y = \{Y(s_i)\}_{i=1}^n$
X	A set of explain variables. $\{X(s_i)\}_{i=1}^n$ is a $p \times 1$ vector of covariates or explain variables at location s_i.
η	Spatial random effects of the observations, which provide local adjustments to the means. $\eta = \{\eta(s_i)\}_{i=1}^n$
η^*	Spatial random effects of the knots. $\eta^* = \{\eta^*(s_i)\}_{i=1}^m$
$\tilde{\eta}$	The predicted values of η by η^*.$\tilde{\eta} = \{\tilde{\eta}(s_i)\}_{i=1}^n$
$\tilde{\epsilon}$	$\{\tilde{\epsilon}(s_i)\}_{i=1}^n$ is the nugget measurement error.
v^*	$v^* = (\eta^{*'}, \beta')'$, it is a $(m+p) \times 1$ vector comprising the realizations of the spatial predictive process and the regression parameters.
ϕ	The decay and smoothness parameter.
$\rho(\cdot;\phi)$	The function for computing the correlations between Y, between Z, or between Y and Z.
H	$H = [F(\phi)X]$. $F(\phi)$ is a transformation matrix which describes that $\tilde{\eta}$ is defined as a spatially varying linear transformation of η^*.
Θ	The set of sample locations of θ, based on the mode and Hessian at it of $\hat{\pi}(\theta\|Y,Z)$. $\Theta = \{\theta\}_{k=1}^K$.
Δ	The set of weight values of sample θ, which are computed by their corresponding posterior distributions. $\Delta = \{\Delta\}_{k=1}^K$

where $c(s;\theta) = [C(s, s_j^*;\theta)_{j=1}^{n^*}]$. The reduced rank SKM model can be formalized as

$$Y(s) = x^T(s)\beta + \tilde{\eta}(s) + \epsilon(s) \tag{6.3}$$

It is important to select a reasonable number of knots as well as their spatial locations. This is related to the problem of spatial design. There are two popular knots selection strategies. One is to draw a uniform grid to cover the study region and each grid is considered as a knot. Another is to place knots such that each covers a local domain and the regions with dense data have more knots. In practice, it is feasible to validate models by using different number of knots and different choices of knots to obtain a reliable and robust configuration.

6.2.3 Laplace Approximation (LA)

The Laplace approximation for the conditional posterior of the latent function is constructed from the second order Taylor expansion of $\log p(f|\mathcal{D}, \theta, \sigma^2, v)$ around the mode $\hat{f}$, which gives a Gaussian approximation to

the conditional posterior

$$p(f|\mathcal{D},\theta,\sigma^2,v) \approx q(f|\mathcal{D},\theta,\sigma^2,v) = N(f|\hat{f},\Sigma_{LA}) \tag{6.4}$$

where $\hat{f} = argmax_f p(f|\mathcal{D},\theta,\sigma^2,v)$ [131]. Σ_{LA}^{-1} is the Hessian of the negative log conditional posterior at the mode, that is,

$$\Sigma_{LA}^{-1} = -\nabla\nabla \log p(f|\mathcal{D},\theta,\sigma^2,v)|_{f=\hat{f}} = K^{-1} + W, \tag{6.5}$$

where W is a diagonal matrix with entries $W_{ii} = \nabla_{fi}\nabla_{fi} \log p(y|f_i,\sigma^2,v)|_{f_i=\hat{f}_i}$.

The inference in the hyperparameters is done by approximating the conditional marginal likelihood $p(f|\mathcal{D},\theta,\sigma^2,v)$ with Laplace's method and searching for the approximate maximum a posterior estimate for the hyperparameters

$$\{\hat{\theta},\hat{\sigma}^2,\hat{v}\} = \arg\max_{\theta,\sigma^2,v}[\log q(\theta,\sigma^2,v|\mathcal{D})] = \arg\max_{\theta,\sigma^2,v}[\log q(y|X,\theta,\sigma^2,v) + \log p(\theta,\sigma^2,v)] \tag{6.6}$$

where $p(\theta,\sigma^2,v)$ is the prior of the hyperparameters. The gradients of the approximate log marginal likelihood can be solved analytically, which enables the MAP estimation of the hyperparameters with gradient based optimization methods. Following Williams and Barber, the approximation schema [165] is called the Laplace method, but essentially the same approach is named Gaussian approximation by Rue et al. [136] in their Integrated nested Laplace approximation (INLA) software package for Gaussian Markov random field models [156].

6.3 Robust and Reduced Rank Spatial Kriging Model

The Robust and Reduced-Rank Spatial Kriging Model (R^3-SKM) integrates robust, reduced-rank, and Bayesian hierarchical techniques together.

The proposed R^3-SKM is defined as

$$Y = X\beta + \tilde{\eta} + \tilde{\epsilon} \tag{6.7}$$

in which most of the variables are defined in section II, except the measurement error $\tilde{\epsilon}$ now follows a heavy tailed distribution with the probability density function $f(\tilde{\epsilon};\mu,\varrho^2) = \frac{1}{\varrho}h((\tilde{\epsilon}-\mu)/\varrho)$, where μ refers to the mean, and ϱ the dispersion parameter. Examples of the h function include: 1) Laplace distribution: $h(x) = \frac{1}{2}e^{-|x|}$; 2) Student's *t* distribution: $h(x) = c(x+\nu)^{(p+\nu)/2}$, where *c* is a normalization constant, the

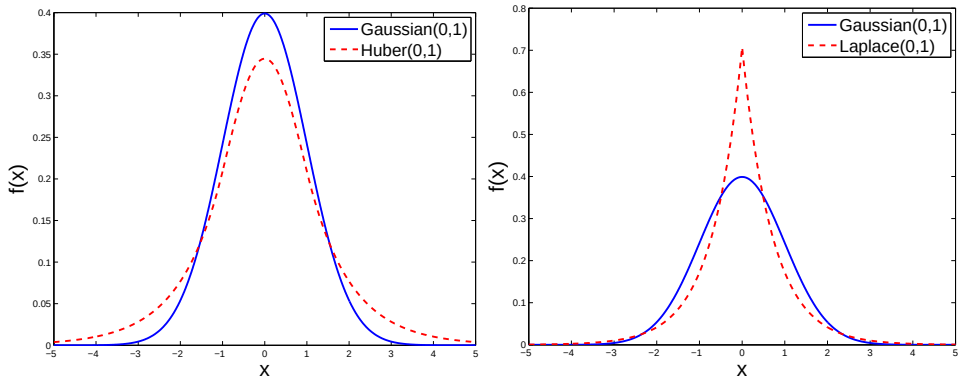


Figure 6.2: pdfs of Heavy Tailed Distributions

case $v = 1$ is the Cauchy density, and the limiting case $v \to \infty$ yields the normal distribution; and 3) Huber distribution: $h(x) = ce^{-\varphi(x;\varrho)}$,

$$\varphi(x;\kappa) = \begin{cases} \kappa|x| - \dfrac{1}{2}\kappa^2, & \text{for } |x| > \kappa \\ \dfrac{1}{2}x^2, & \text{for } |x| \leq \kappa, \end{cases} \tag{6.8}$$

where c is a normalization constant that ensures $\int \frac{c}{\varrho} e^{-\varphi(x;\kappa)} = 1$, and ϱ is a range parameter of the distribution. The probability density functions (pdf) of the Huber, Laplace and Gaussian distribution are compared in Figure 6.2. The robustness of the R^3-SKM is realized by the latent variation component $\tilde{\epsilon}_i, i = \{1, \cdots, n\}$, which follows a heavy tailed distribution. For example, the parameter ν in Student's t, or κ in Huber distribution controls the degree of the robustness. When the value of ν increases, the robustness of the Student's t will decrease.

R^3-SKM can be formalized in the framework of Bayesian hierarchical model with three layers, including the observation layer, the latent robust Gaussian process layer, and the parameter layer. The observation layer contains the observations $Y = \{Y(s_1), \cdots, Y(s_n)\}$. It is assumed that each $Y(s_i)$ follows a Gaussian distribution. Each random variable $Y(s_i)$ is related to the latent Gaussian effects in the second layer, $v^* = (\eta^{*\prime}, \beta')'$, which is the $(m + p) \times 1$ vector. Specifically, β is assigned a multivariate Gaussian prior, i.e., $\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$. The third level of the hierarchical model consists of the related parameters with the latent variables. In the R^3-SKM model, the parameters include $\theta = (\sigma^2, \phi, \nu, \varrho^2)$. That is, σ^2 and ϕ for modeling η^*, μ and ϱ^2 for modeling $\tilde{\epsilon}$. σ^2 has an inverse gamma prior distribution: $\sigma^2 \sim IG(\alpha_\sigma, \gamma_\sigma)$, where α_σ and γ_σ are sufficiently small informative prior distribution. The correlation parameter ϕ is usually assigned an informative prior decided based on the underlying spatial domain, i.e., $\phi \sim \mathcal{U}(a_\phi, b_\phi)$, a uniform distribution over a finite range. In Student's t distribution, $\nu \sim \mathcal{U}(a_\nu, b_\nu)$ and $\varrho^2 \sim IG(\alpha_\varrho, \gamma_\varrho)$. Taking the Student's t

as an example, the graphic representation of the R^3-SKM is depicted in Figure 6.3.

6.4 Robust Parameter Estimation

This section presents a novel approach, R^3-PE (Robust and Reduced-Rank Parameter Estimation), to execute the robust parameter estimation by integrating Laplace approximation [136]. It consists of two critical steps: 1) Gaussian approximation of the posterior distribution of latent variables conditional on parameters and observations; 2) Laplace approximation of the posterior distribution of corresponding parameters conditional on observations. Student's t distribution is selected to model the *pdf* of $\tilde{\epsilon}$.

6.4.1 Gaussian Approximation of Posterior Distribution of v^*

First, we need to compute $\pi(v^*|Y,\theta)$, where $v^* = (\eta^{*'}, \beta')'$, which consists of the spatial predictive process and the regression parameters. Its mean and covariance matrix are computed as follows.

$$\mu_{v^*} = (0_m, \mu_\beta)', \Sigma_{v^*} = \begin{bmatrix} \sigma^2 C^*(\phi) & 0_{m\times p} \\ 0_{p\times m} & \Sigma_\beta \end{bmatrix} \tag{6.9}$$

we have prior $v^* \sim N(\mu_{v^*}, \Sigma_{v^*})$.

Case 1: Integrated with Student-t distribution

With the information depicted by the graphical model in Figure 6.3, we determine that the full conditional distribution of $\pi(Y|v^*,\theta)$ follows the heavy tailed distribution, which can be approximated as a Gaussian distribution of v^* by Taylor expansion. For example, if $\tilde{\epsilon}$ accords to the Student's t distribution, then $\pi(y_i|\nu, T_iHv^*, \varrho^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\varrho^2}}(1+\frac{1}{\nu}\frac{(y_i-T_iHv^*)^2}{\varrho^2})^{-\frac{\nu+1}{2}}$, where $T_iHv^* = x_i\beta + \tilde{\eta}_i$ and T_i is the i^{th} row of unit matrix I_n. H=$[F(\phi)X]$, and $F(\phi) = \mathcal{C}(\phi)'C^{*-1}(\phi)$, where $\mathcal{C}(\phi)'$ is an $n \times m$ matrix whose i^{th} row is the $1 \times m$ vector in which the j^{th} element is given by $C(s_i, s_j; \phi)$. We Taylor expand it to second order by

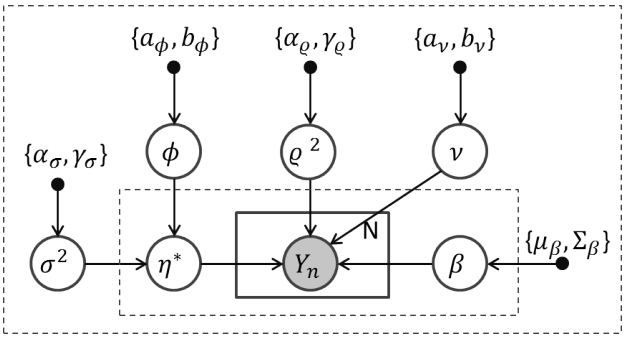


Figure 6.3: Graphic Model Representation of R^3-SKM

expressing the result in a quadratic form of v^*,

$$
\begin{aligned}
&\log\left(\hat{\pi}(Y|v^*,\theta)\right) = -\frac{1}{2}v^{*'}Q_Y v^* + v^{*'}b_Y + const \\
&Q_Y = \sum_{i=1}^{n} Q_{y_i}, b_Y = \sum_{i=1}^{n} b_{y_i} \\
&Q_{y_i} = m_1 m_2 \left(\frac{m_2 \nabla D(\hat{v}^*)\{\nabla D(\hat{v}^*)\}'}{(1+m_2 D(\hat{v}^*))^2} - \frac{\nabla^2 D(\hat{v}^*)}{1+m_2 D(\hat{v}^*)}\right) \\
&b_{y_i} = \frac{m_1 m_2 \nabla D(\hat{v}^*)}{1+m_2 D(\hat{v}^*)} + Q_{y_i}\hat{v}^* \\
&m_1 = -\frac{\nu+1}{2}, m_2 = \frac{1}{\nu\varrho^2} \\
&D(\hat{v}^*) = (y_i - T_i H\hat{v}^*)'(y_i - T_i H\hat{v}^*) \\
&\nabla D(\hat{v}^*) = -2H'T_i'y_i + 2H'T_i'T_i H\hat{v}^* \\
&\nabla^2 D(\hat{v}^*) = 2H'T_i'T_i H
\end{aligned}
\tag{6.10}
$$

With the above Gaussian approximation, $\pi(v^*|Y,\theta)$ is analytically available and numerical routines can be applied. For the R^3-SKM, the full conditional for v^* is

$$
\begin{aligned}
&\pi(v^*|Y,\theta) \propto \hat{\pi}(Y|v^*,\theta)\pi(v^*|\theta) \qquad (6.11)\\
&\propto exp[(-\frac{1}{2}v^{*'}Q_y v^* + v^{*'}b_y) + (-\frac{1}{2}v^{*'}\Sigma_{v^*}^{-1}v^* + v^{*'}\Sigma_{v^*}^{-1}\mu_{v^*})] \\
&\propto exp[-\frac{1}{2}v^{*'}Qv^* + v^{*'}b] \\
&\propto exp[-\frac{1}{2}v^{*'}(Q_y + \Sigma_{v^*}^{-1})v^* + v^{*'}(b_y + \Sigma_{v^*}^{-1}\mu_{v^*})] \\
&\propto N(Y|v^*,\theta)N(\eta^*|\theta)\pi(\tilde{\epsilon}|\theta)nonumber \qquad (6.12)\\
& \qquad (6.13)
\end{aligned}
$$

where the full conditional precision matrix $Q = Q_Y + \Sigma_{v^*}^{-1}$, and the canonical parameter $b = b_Y + \Sigma_{v^*}^{-1}\mu_{v^*}$. Thus, the full conditional is $\pi(v^*|Y,\theta) \sim N(Q^{-1}b, Q^{-1})$. We can compute the required inverse and determinant of the size $(m+p)\times(m+p)$ matrix Q by utilizing the structure of Σ_{v^*} and H. The main cost of matrix inversion is thus $O(m^3)$, since the number of knots is m, assuming $m \gg p$.

Case 2: Integrated with Laplace distribution

If $\hat{\epsilon}$ accords to the Laplace distribution, then $\pi(y_i|B, T_iH\nu^*) = \frac{1}{2B}exp(-\frac{|y_i - T_iH\nu^*|}{B})$. Here, $B \geq 0$, where $T_iH\nu^* = x_i\beta + \tilde{\eta}_i$, and T_i is the i^{th} row of unit matrix I_n. $H = [F(\phi)X]$, and $F(\phi) = \mathcal{C}(\phi)'C^{*-1}$, where $\mathcal{C}(\phi)'$ is an $n\times m$ matrix whose i^{th} row is the $1\times m$ vector in which the j^{th} element is given by $C(s_i, s_j; \phi)$. We Taylor expand it to second order by expressing the result in a quadratic form of ν^*,

$$\log\left(\hat{\pi}(Y|v^*,\theta)\right) = \begin{cases} \dfrac{T_iH\nu^*}{B} + const, & \text{for } y_i > T_iH\nu^* \\ -\dfrac{T_iH\nu^*}{B} + const, & \text{for } y_i \leq T_iH\nu^*, \end{cases} \tag{6.14}$$

For the R^3-SKM, corresponding to Eq. (6.11), the full conditional precision matrix $Q = \Sigma_{\nu^*}^{-1}$, and the canonical parameter

$$b = \begin{cases} \dfrac{H'T_i'}{B} + \Sigma_{\nu^*}^{-1}\mu_{\nu^*}, & \text{for } y_i > T_iH\nu^* \\ -\dfrac{H'T_i'}{B} + \Sigma_{\nu^*}^{-1}\mu_{\nu^*}, & \text{for } y_i \leq T_iH\nu^*, \end{cases} \tag{6.15}$$

Thus, the full conditional is $\pi(v^*|Y,\theta) \sim N(Q^{-1}b, Q^{-1})$. We can compute the required inverse and determinant of the size $(m+p)\times(m+p)$ matrix Q by utilizing the structure of Σ_{v^*} and H. The main cost of matrix inversion is thus $O(m^3)$, since the number of knots is m, assuming $m \gg p$.

Case 3: Integrated with Huber distribution

Here we explores the special structure of the R^3-SKM model based on the Huber distribution. The Huber distribution is used to model the measurement error: the random variable $\tilde{\epsilon}(s_i) \sim Huber(0, \sigma, \kappa)$. The pdf of the Huber distribution is defined as $\pi(\tilde{\epsilon}; \mu, \sigma, \kappa)$, $h(y_i|T_iH\nu^*, \sigma, \kappa) = cexp(-\varphi(\frac{y_i - T_iH\nu^*}{\sigma}; \kappa))$, and

$$\varphi(\frac{y_i - T_iH\nu^*}{\sigma}; \kappa) = \begin{cases} \kappa(y_i - T_iH\nu^*) - \dfrac{1}{2}\kappa^2, & \text{for } (y_i - T_iH\nu^*) > \kappa \\ \dfrac{1}{2}(y_i - T_iH\nu^*)^2, & \text{for } -\kappa \leq (y_i - T_iH\nu^*) \leq \kappa \\ -\kappa(y_i - T_iH\nu^*) - \dfrac{1}{2}\kappa^2, & \text{for } (y_i - T_iH\nu^*) < -\kappa, \end{cases} \tag{6.16}$$

We Taylor expand it to second order by expressing the result in a quadratic form of v^*,

$$\log\left(\hat{\pi}(Y|v^*,\theta)\right) = \begin{cases} \nu^{*\prime}\kappa H' T_i' + const, & \text{for } (y_i - T_i H\nu^*) > \kappa \\ -\frac{1}{2}\nu^{*\prime} H' T_i' T_i H\nu^* + \nu^{*\prime}(y_i H' T_i'), & \text{for } -\kappa \leq (y_i - T_i H\nu^*) \leq \kappa \\ -\nu^{*\prime}\kappa H' T_i' + const, & \text{for } (y_i - T_i H\nu^*) < -\kappa, \end{cases} \tag{6.17}$$

That is, we can express it in the following quadratic form of ν^*,

$$\log\left(\hat{\pi}(Y|v^*,\theta)\right) = -\frac{1}{2} v^{*'} Q_Y v^* + v^{*'} b_Y + const \tag{6.18}$$

where

$$Q_Y = \begin{cases} 0, & \text{for } (y_i - T_i H\nu^*) > \kappa \\ H' T_i' T_i H, & \text{for } -\kappa \leq (y_i - T_i H\nu^*) \leq \kappa \\ 0, & \text{for } (y_i - T_i H\nu^*) < -\kappa, \end{cases} \tag{6.19}$$

and

$$b_Y = \begin{cases} \kappa H' T_i', & \text{for } (y_i - T_i H\nu^*) > \kappa \\ y_i H' T_i', & \text{for } -\kappa \leq (y_i - T_i H\nu^*) \leq \kappa \\ -\kappa H' T_i', & \text{for } (y_i - T_i H\nu^*) < -\kappa, \end{cases} \tag{6.20}$$

The value of κ is chosen in order to ensure a given asymptotic variance - hence a given asymptotic efficiency - at the normal distribution.

6.4.2 Laplace Approximation of Posterior Distribution of θ

Different from $\pi(v^*|Y,\theta)$, the posterior $\pi(\theta|Y)$ is usually skewed and the approximation as a Gaussian distribution is inappropriate. The posterior $\pi(\theta|Y)$ plays an important role in the inference of marginal posterior of latent variables that are of interests. Take v^* as an example, the interest is to estimate the marginal posterior $\pi(v^*|Y)$, which has

$$\pi(v^*|Y) = \int \pi(v^*|Y,\theta)\pi(\theta|Y)d\theta \tag{6.21}$$

Algorithm 8 Exploring posterior distribution of $\pi(\theta|Y)$

Input: S, S^*, Y, X

Output: Θ, Δ

1: Choose an initial value $\theta = (\sigma^2, \phi, \nu, \varrho)$;
2: **repeat**
3: Construct μ_{v^*}, Σ_{v^*} with θ (See Equation(6.9));
4: Calculate the transformation matrix H;
5: Gaussian approximation of $\pi(Y|v^*, \theta)$ as the form of v^*.
6: Apply IRLS to identify the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y, \theta)$.
7: Compute the gradient and Hessian of $\hat{\pi}(\theta^*|Y)$ and apply one Newton's step to update θ.
8: **until** Convergence
9: Explore the contour of $\tilde{\pi}(\theta|Y)$ based on its mode and Hessian at the mode, obtain K sample locations, $\Theta = \{\theta_1, ..., \theta_K\}$.
10: Compute and normalize $\{\hat{\pi}(\theta_1|Y), \cdots, \hat{\pi}(\theta_K|Y)\}$ to obtain $\Delta = \{\Delta_1, \cdots, \Delta_K\}$ as $\Delta_k = \frac{\hat{\pi}_k(\theta_k|Y)}{\sum_{k=1}^{K} \hat{\pi}_k(\theta_k|Y)}$.

It is possible to obtain a sample set of $\{\theta_1, \cdots, \theta_K\}$ from the input space of θ that represents an approximate discrete form of the posterior $\pi(\theta|Y)$. We can estimate the approximate $\hat{\pi}(v^*|Y)$ by

$$\hat{\pi}(v^*|Y) = \sum_{k=1}^{K} \pi(v^*|Y, \theta_k)\pi(\theta_k|Y)\Delta_k \tag{6.22}$$

where Δ_k is the weight of the sample θ_k that can be measured by its normalized probability density. The critical step is to efficiently identify a suitable sample set $\{\theta_1, \cdots, \theta_K\}$, as well as its corresponding weight set $\{\Delta_1, \cdots, \Delta_K\}$. The posterior $\pi(\theta|Y)$ can be reformalized as

$$\pi(\theta|Y) \propto \frac{\pi(Y|v^*, \theta)\pi(v^*|\theta)\pi(\theta)}{\pi(v^*|Y, \theta)} \tag{6.23}$$

Laplace Approximation (LA) can be applied to approximate the denominator $\pi(v^*|Y, \theta)$ as a Gaussian distribution, and then set the vector of variables, v^*, to the mode. The LA method uses similar ideas for Bayesian spatial inference:

$$\hat{\pi}(\theta|Y) \propto \left.\frac{\pi(Y|v^*, \theta)\pi(v^*|\theta)\pi(\theta)}{\hat{\pi}(v^*|Y, \theta)}\right|_{v^*=\hat{v}^*} \tag{6.24}$$

where $\hat{\pi}(v^*|Y, \theta)$ is a Gaussian approximation as shown in Equation (6.11). We can get the mode $\hat{v}^*$ and the curvature at the mode of this full conditional expression. The preceding Gaussian approximation can be efficiently conducted by using the popular Iterated Re-weighted Least Squares (IRLS) algorithm. The above detailed procedures can be summarized as Algorithm 1.

Algorithm 1 iterates l_1 times, from Step 2 to Step 8 until convergence. Among these steps, Step 6 has the

highest time cost, because the solution is analytically intractable and numerical optimization techniques are applied. An efficient IRLS algorithm is proposed to conduct this process. Step 6 is first reformulated as the following optimization problem

$$\underset{v^*}{\operatorname{argmax}}\hat{\pi}(v^*|Y,\theta) = \underset{v^*}{\operatorname{argmin}} - \ln \pi(Y|v^*,\theta) - \ln \pi(v^*|\theta) \tag{6.25}$$

Expanding the density functions $\pi(Y|v^*,\theta)$ and $\pi(v^*|\theta)$, we have that $\underset{v^*}{\operatorname{argmin}}\{\frac{1}{2}v^{*\prime}\{\sum_{i=1}^{n} Q_i\}v^* - v^{*\prime}\{\sum_{i=1}^{n} b_i\} + \frac{1}{2}\left(v^* - \mu_{v^*}\right)'\Sigma_{v^*}^{-1}(v^* - \mu_{v^*})\}$. The gradient and Hessian matrix of the above objective function can be obtained as

$$\begin{aligned} \bigtriangledown\hat{\pi}(v^*|Y,\theta) &= (\sum_{i=1}^{n} Q_i + \Sigma_{v^*}^{-1})v^* - (\sum_{i=1}^{n} b_i + \Sigma_{v^*}^{-1}\mu_{v^*}) \\ \bigtriangledown^2\hat{\pi}(v^*|Y,\theta) &= \sum_{i=1}^{n} Q_i + \Sigma_{v^*}^{-1} \end{aligned} \tag{6.26}$$

The IRLS algorithm for Step 6 is described as follows:

1. Select an initial $\hat{v}^*$
2. Until convergence

   Update $\hat{v}^* = \hat{v}^* - \left(\bigtriangledown^2\hat{\pi}(\hat{v}^*|Y,\theta)\right)^{-1} \bigtriangledown \hat{\pi}(\hat{v}^*|Y,\theta)$**.**
3. Output $\hat{v}^*$ as the mode of $\hat{\pi}(v^*|Y,\theta)$.

Computational Complexity. In Algorithm 1, suppose that it needs l_2 iterations to find the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y,\theta)$, the time cost of Step 6 is $O(l_2 * (n * m^2 + m^3))$. The Step 5, Gaussian approximation of $\pi(Y|v^*,\theta)$, takes $O(n*m)$. Overall, Steps 2-8, which generate the converged gradient and Hessian of $\pi(\theta|v^*)$ take $O(l_1 * l_2 * (n * m^2 + m^3) + l_1 * n * m)$. Finally, sampling the θ set and computing their corresponding weighted values take $O(K)$. In summary, assuming $n \gg K$, $n \gg m$, $n \gg l_1$ and $n \gg l_2$, the total computational complexity of robust parameter estimation based on R^3-SKM is $O(n)$.

6.5 Robust Spatial Inference

This section formalizes the Robust and Reduced Rank Spatial Prediction (R^3-SP) , and Robust and Reduced Rank Spatial Outlier Detection (R^3-SOD) based on the R^3-SKM.

Algorithm 9 Robust Reduced Rank Spatial Prediction

Input: $S, S^*, S^0, Y, X, X^0, \Theta, \Delta$
Output: Y^0

1: **for** $k = 1$**to**K **do**
2: Construct μ_{v^*}, Σ_{v^*} with θ_k and S^* (See Equation (6.9)).
3: Calculate the transformation matrix H with θ_k, S^*, S, X.
4: Gaussian approximation of the likelihood of Y.
5: Calculate the mode, Hessian at the mode of $\hat{\pi}(v^*|Y, \theta_k)$, and its Gaussian approximation (See Equation (6.11)).
6: Predict Y_k^0 for new locations S^0. (See Equation (7.54))
7: **end for**
8: Calculate the final Y^0 values as $Y^0 = \sum_{k=1}^{K} Y_k^0 \times \Delta_k$

6.5.1 Robust Spatial Prediction

Given a set of unsampled locations $\{s_1^0, \cdots, s_{N_{te}}^0\}$, we are interested in predicting the Y values at these locations, denoted as $Y^0 = [Y(s_1^0), \cdots, Y(s_{N_{te}}^0)]$. The first step is to estimate the posterior distributions of the corresponding latent variables $\pi(v_0|Y)$, where $v^0 = [v(s_1^0), \cdots, v(s_{N_{te}}^0)]'$. Then, the posterior distributions of Y^0 can be obtained as

$$\pi(Y^0|Y) = \int \pi(Y^0|v^0)\pi(v^0|Y)dv^0 \tag{6.27}$$

Given the approximated $\hat{\pi}(v^*|Y, \theta)$ and $\tilde{\pi}(\theta|Y)$ as obtained in Sections IV.A and IV.B, the posterior distribution $\pi(v^0|Y)$ can be estimated by

$$\begin{aligned}
\pi(v^0|Y) &= \int\int \pi(v^0|v^*, Y, \theta)\pi(v^*|Y, \theta)\pi(\theta|Y)dv^*d\theta \\
&= \int \left\{ \int \pi(v^0|v^*, \theta)\pi(v^*|Y, \theta)dv^* \right\} \pi(\theta|Y)d\theta \\
&\approx \sum_k \left\{ \int \pi(v^0|v^*, \theta_k)\hat{\pi}(v^*|Y, \theta_k)dv^* \right\} \hat{\pi}(\theta_k|Y)\Delta_k \\
&\approx \sum_k \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})\hat{\pi}(\theta_k|Y)\Delta_k
\end{aligned} \tag{6.28}$$

where

$$\begin{aligned}
&\Sigma^0 = Cov(v^0), \Sigma^* = Cov(v^*), \Sigma^{0*} = Cov(v^0, v^*) \\
&\tilde{\mu} = \Sigma^{0*}\Sigma^{*-1}Q^{-1}b \\
&\tilde{\Sigma} = \Sigma^0 - \Sigma^{0*}\Sigma^{*-1}\Sigma^{0*\prime} + \Sigma^{0*}\Sigma^{*-1}Q^{-1}\Sigma^{*-1}\Sigma^{0*\prime}
\end{aligned}$$

Based on the above theoretical analysis, the main procedures of R^3-SP are described by Algorithm 2. In the R^3-SP algorithm, we first derive the K samples of θ and their weight values, Δ, by utilizing the R^3-SKM framework, and then use each generated sample, θ_k, to construct the corresponding mean and covariance matrix of latent variables, v^*. Next, the transformation matrix $H = [F(\phi)X]$ is computed, in which $F(\phi)$ describes the spatially varying linear transformation of $\tilde{\eta}$ on η^*. Furthermore, the likelihood of Y are approximated as the result of a quadratic form of v^*. Next, the mode of $\hat{\pi}(v^*|Y, \theta_k)$ are calculated to predict the new observations Y_k^0 at sample θ_k. Finally, the predicted Y is calculated as $Y^0 = \sum_{k=1}^{K} Y_k^0 \times \Delta_k$.

Computational complexity. Similarly, for the R^3-SP algorithm, Steps 4 and 6 dominate most time costs, because they are naturally analytical intractable. With the numerical optimization discussed in Section IV, it takes $O(n * m)$ to operate a Gaussian approximation of Y at each sample θ_k. And computing the mode and Hessian of $\hat{\pi}(v^*|Y, \theta_k)$ costs $O(l_2 * (m^3 + n * m^2))$. Repeating Steps 2-6 at K sample θs takes $O(K * (n * m + l_2 * (m^3 + n * m^2)))$. In summary, the total computational complexity of the R^3-SP algorithm is $O(n)$, assuming $n \gg K$, $n \gg l_2$, $n \gg p$ and $n \gg m$.

6.5.2 Robust Spatial Outlier Detection

Statistically, spatial outlier can be interpreted as observations that have abnormally low correlations with their spatial neighbors, considering normal deviations caused by measurement error (white noise). For the regular SKM framework, when a data set contains outliers, the additional variation due to those outliers will be captured by distorting spatial dependence. The white noise component is unable to handle large deviations due to the light tailed feature of the Gaussian distribution. In comparison, the proposed R^3-SKM uses heavy tailed distribution to model the measurement error. When outliers appear, our model directly captures the additional large variation due to outliers as the measurement error, which will control the resulting accurate parameter estimation and spatial outlier detection.

Therefore, the R^3-SKM can also be utilized to identify spatial outliers as objects with higher predicted $\tilde{\epsilon}$ values(measurement error). First, we apply the R^3-SKM to accurately estimate the latent variables and parameters for the contaminated spatial dataset. Second, the estimated values are utilized to operate a spatial prediction for each observed location. Finally, the differences between observed and predicted values are computed to measure their outlying degrees. The objects which have higher measurement errors are labeled as spatial outliers.

The main procedures of spatial outlier detection are described in Algorithm 3. In the R^3-SOD Algorithm, we use the K samples of θ to predict the corresponding $\{Y_i\}_k^p (k = 1, \cdots, K)$, and the predicted $\{Y_i\}^p$ is finalized by the sum over values derived from different θ_k with weight Δ_k. If the predicted $\{Y_i\}^p$ has a large deviation compared with its original value, and this deviation is higher than the cut-off value, $c \cdot \varrho (c = 3)$,

Algorithm 10 Robust Reduced Rank Spatial Outlier Detection (R^3-SOD)

Input: $S, S^*, Y, X, \Theta, \Delta$
Output: Y^0

1: Repeat Steps 1-7 in Algorithm 2 to predict $\{Y_i\}_k^p$ for each locations s_i.($k = 1, \cdots, K$, and $i = 1, \cdots, n$).
2: **for** $i = 1$**to**n **do**
3: Calculate the final $\{Y_i\}^p$ values as $\{Y_i\}^p = \sum_{k=1}^{K} \{Y_i\}_k^p \times \Delta_k$.
4: Calculate the abstract difference Diff$_i$= $|\{Y_i\}^p - Y_i|$.
5: **end for**
6: Rank the objects by sorting Diff with an descending order.
7: Label the top ones that have Diff$\geq c \cdot \varrho$ as spatial outliers.

then the corresponding objects will be identified as spatial outliers.

Computational complexity. As analyzed in Algorithm 2, predicting $\{\{Y_i\}_k^p\}_{i=1}^n$ takes around $O(K(l_2 n + m^3))$. Finalizing the predicted $\{\{Y_i\}^p\}_{i=1}^n$ costs $O(n)$. In summary, R^3-SOD algorithm takes $O(n)$, assuming $n \gg K$, $n \gg l_2$, $n \gg p$ and $n \gg m$.

6.6 Experiment

This section evaluates the robustness and efficiency of the proposed R^3-SKM model based on an analysis of simulated and real data sets. Student's t distribution was selected to model the probability density function of $\tilde{\epsilon}$. All experiments were conducted on a PC with Intel(R) Core(TM) I5-2400, CPU 3.1 Ghz, and 8.00 GB memory.

6.6.1 Experiment Setting

Dataset Description

Simulation Dataset. The simulations were generated based on the following statistical model:

$$Y(s) \quad \sim \quad \mathcal{N}(x^T\beta + \eta(s), \tau^2) \tag{6.29}$$

where $\eta(s)$ is from a latent spatial Gaussian process with the variogram model $Var(\eta(s_i), \eta(s_j)) = \sigma^2 C(h|\phi)$, and $h = |s_i - s_j|$. $C(h|\phi)$ refers to the spatial correlation, where ϕ is the range parameter that controls its degree. The popular exponential function was used to model $C(h|\phi)$. The parameter settings used in our experiments are shown in Table 6.2. We also evaluated different combinations of parameters, and observed similar patterns.

Table 6.2: Parameter settings in the simulations

Variable	Setting Description
$[N_{tr}, N_{te}]$	Training and testing points were randomly generated at N_{tr} spatial locations $\{s_i\}_{i=1}^{N_{tr}}$ and N_{te} spatial locations $\{s_i\}_{i=1}^{N_{te}}$, respectively, in the range [0,50]×[0,50] units. $N_{tr} = 300, 500$, $N_{te} = 30, 100$
β	The regression coefficient $\beta = [0.5, 1.5]'$.
σ	$\sigma^2 = 4$ in all simulations
ϕ	$\phi = 25$.
τ	The nugget variance, τ^2, was set to 0.1.
α	The contamination rate was set to 25, 20, 15, 10, 5.
γ	The shift rate was set to 3, 3.5, 4, 4.5, 5, 5.5, 6.
$C(h\|\phi)$	An exponential spatial correlation function $C(h\|\phi) = \sigma^2 exp(-\frac{h}{\phi})$ was used in all simulations.

Real Dataset. We validated our approach on five real datasets, namely, *Lake*, *MLST*, *BEF*, *HR*, and *House*. *Lake* was originally published by Varin et al. [157] and was used to model trout abundance in Norwegian lakes as a function of lake acidity. The explained attributes used include *Intercept*, *X coordinate*, *Y coordinate*, *Product of X* and *Y coordinates*, *X coordinate squared* and *Y coordinated squared*. *MLST* [46] came from multiple listings containing structural descriptors of houses, their sale prices, and their addresses for Baltimore, Maryland, in 1978. Dubin estimated a spatial autocorrelation model which calculated the portion of the price by multiplying the vector of attributes by the estimated coefficients. The explained attributes used contain *X coordinate*, *Y coordinate*, *Product of X and Y coordinates*, *X coordinate squared*, and *Y coordinate squared*. *BEF* [49] is a forest inventory dataset from the U.S. Department of Agriculture Forest Service, Barlett, NH. The explained variables include levels of nitrogen oxides, particulate concentrations, average number of rooms, proportion of structured built before 1940, black population proportion, lower status population proportion, crime rate, proportion of area zoned with large lots, proportion of nonretail business area, property tax rate, pupil-teacher ratio, location contiguous to the Charles River, weighted distances to the employment centers, and an index of accessibility. Variables include species, specific basal area, total tree biomass, inventory plot coordinates, and slope, etc. Finley and Banerjee made the detailed analysis of the non-spatial logistic regression to the data. *HR* is a Boston Housing dataset from 1978 which discusses issues related to the demand for clean air. [65] made statistical analysis on it. The explained variables are *the slope*, *elevation of the object*, *the tasseled cap brightness*, *greenness*, and *wetness components* from summer 2002. *House* contains information collected for a range of variables for all the block groups in California from the 1990 Census. The spatial regression model of the data was analyzed by Pace and Barry [122]. *BEF*, Both *House* and *HR* data are included in the spBayes R package [148]. Specifically, it contains median house value, median income, housing median age and total room, etc. The analyzed explained variables include *Median Income*, *Median Income*^{2}, *Median Income*^{3}, *ln(Median Age)*, *ln(TotalRooms/Population)*,

ln(Bedrooms/Population), *ln(Population/Households)* and *ln(Households)*. Table 6.3 summarizes the main information types of each of these datasets used in our experiments.

Table 6.3: Settings in 5 real datasets

Dataset	Data Size	N_{tr}	N_{te}	Y	Y-SD
BEF	437	337	100	BE basal area	0.17
Lake	371	271	100	Trout abundance	0.007
MLST	211	150	61	House price	0.17
HR	506	406	100	House price	0.10
House	20,640	1000	200	House price	0.25

Spatial Inference Method

Spatial Estimation and Prediction Methods. There are currently two popular methods used for parameter estimation and spatial prediction. The Spatial Kriging Model(SKM) predicts unobserved values as a linear combination of the known values of observed locations, while Linear Regression[15] models data using linear predictor functions, and estimates unknown model parameters from the data. It can be used to fit a predictive model to an observed dataset of Y and X value. However, Linear Regression does not take the spatial dependency into considerations, since it is focused on the general modeling. We identify them as *SRM* and *Regression* in our experiment.

Outlier Detection Methods. We compared R^3-SOD with eight existing representative SOD approaches: Z-test [143], Median Z-test, Iterative Z-test, trimmed Z-test [100], Scatterplot [7], MoranScatterplot [62], SLOM [9] and POD [90]. The implementations of the above methods were all based on their published algorithm descriptions.

Performance Metric

For both the synthetic and real dataset, there might be some outlying observations in raw dataset which are unknown for us. To demonstrate the effectiveness of proposed approach, we need to generate some synthetic outliers. We assumed the raw dataset as a ground truth, and contaminated around α percent of the data records as outliers. In our paper, *Data Contamination.* For each dataset, including both the simulations and real datasets, we randomly selected $\alpha\%$(contamination rate) of the data to be anomalies by shifting them from their original values with γ(shift rate) times standard deviation of Y. For each α, the synthetic outliers were generated 10 times, and the mean values of the results from the parameter estimation, spatial prediction and spatial outlier detection were calculated for each approach.

Parameter Estimation. Parameter estimation was executed only in simulations since the true values of parameters were known. We compared the estimation results from SKM, Regression and R^3-SKM, with the

Table 6.4: Comparison of parameter estimation results on simulations

Data	Para.	True Value	R^3-SKM	Regression	SKM
Sim_300_30_36_0.25_7	β	[0.50, 1.50]	[0.49, 1.61]	[0.44, 2.31]	[0.06, 1.91]
	ϕ	25.00	24.73	–	6
	σ^2	2.00	1.11	–	1.85
Sim_500_100_49_0.05_2.5	β	[0.50, 1.50]	[0.44, 1.54]	[0.62, 1.84]	[0.78, 1.23]
	ϕ	25.00	25.92	–	5.78
	σ^2	2.00	1.98	–	1.82
Sim_700_200_81_0.05_5	β	[0.50, 1.50]	[0.48, 1.60]	[0.39, 1.82]	[0.29, 1.76]
	ϕ	25.00	26.41	–	5.66
	σ^2	2.00	1.80	–	1.77
Sim_1000_400_100_0.05_5	β	[0.50, 1.50]	[0.49, 1.46]	[0.63, 1.31]	[0.61, 1.38]
	ϕ	25.00	27.39	–	19077.44
	σ^2	2.00	1.73	–	48.33

true values to validate their effectiveness.

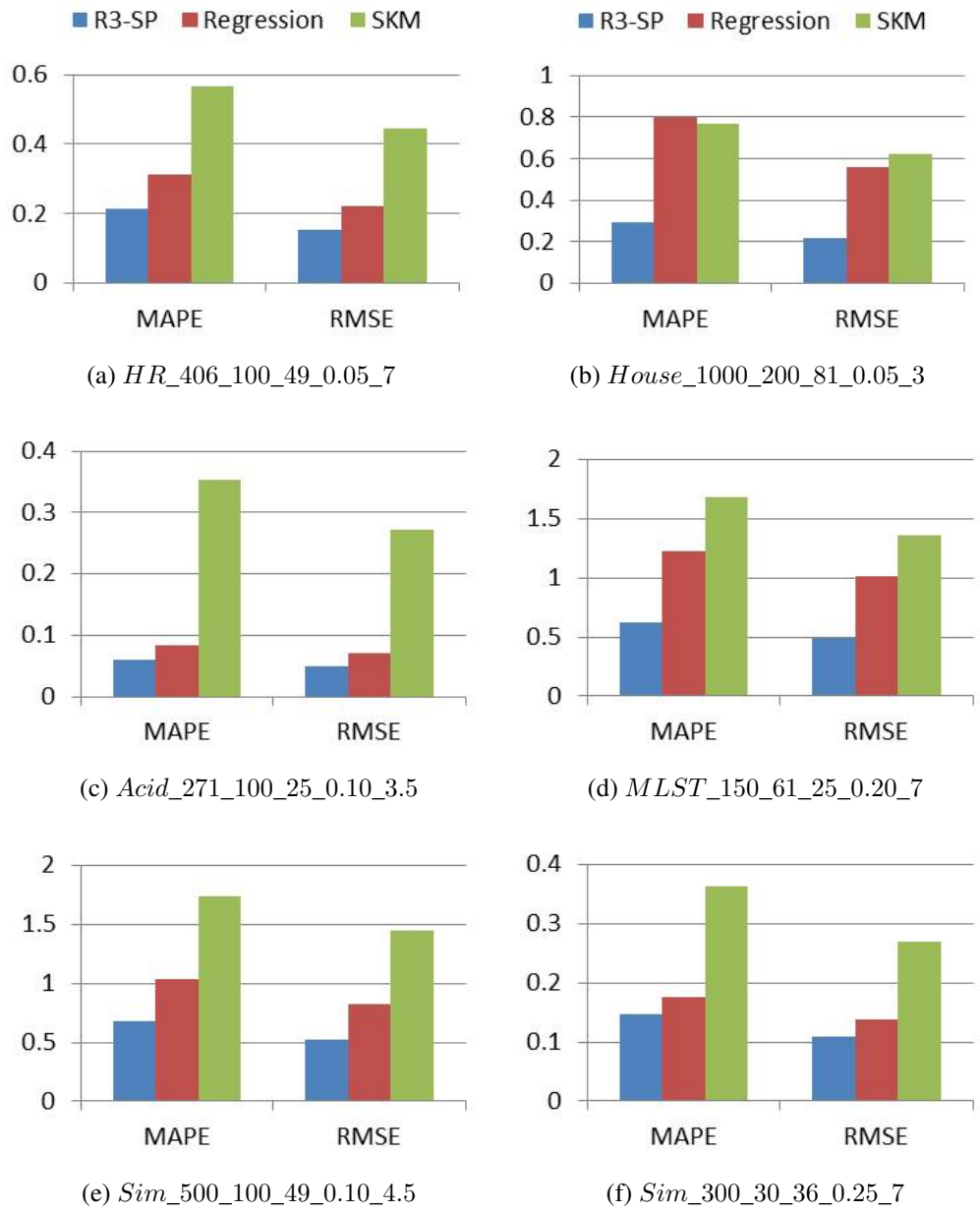
Spatial Prediction. SKM, Regression and R^3-SKM were also applied in both simulation and real datasets to obtain the predicted $\tilde{Y}$. The Mean Absolute Percentage Error ($MAPE = \frac{\sum_{i=1}^{N_{te}} |Y_i - \tilde{Y}_i|}{N_{te}}$) and Root Mean Square Error ($RMSE = \{\frac{\sum_{i=1}^{N_{te}} (Y_i - \tilde{Y}_i)^2}{N_{te}}\}^{1/2}$) between Y and $\tilde{Y}$ were calculated to evaluate the prediction performance.

Spatial Outlier Detection. Nine different outlier detection approaches were applied to both simulation and real datasets. To compare the accuracies among them, we used the following common evaluation measures: detection rate (precision) and detection precision (recall). The precision was plotted against recall and the curves that are higher and farther to the right denote better performance.

6.6.2 Experiment analysis and discussion

Robustness on Parameter Estimation. Table 6.4 shows the parameter estimation results on four simulations with training data sizes of 300, 500, 700 and 1000. The data name depicts the parameter combination information. For example, "Sim_500_100_49_0.05_2.5" indicates that it was generated by **simulation** data, and there are **500** training data, **100** testing data, **49** knots, and **5**% of the training data were contaminated as outliers by shifting the original Y to $(Y + \mathbf{2.5} * std(Y))$.

Comparing the estimated parameters with the true values, R^3-SKM was able to more accurately estimate most of the parameters. For "Sim_500_100_16_0.05_2.5", only 5% the data are distorted with a relatively small shift rate(2.5), which means the contaminated data had a similar distribution to that of the original.



(a) HR_406_100_49_0.05_7

(b) $House$_1000_200_81_0.05_3

(c) $Acid$_271_100_25_0.10_3.5

(d) $MLST$_150_61_25_0.20_7

(e) Sim_500_100_49_0.10_4.5

(f) Sim_300_30_36_0.25_7

Figure 6.4: Comparison of prediction performances on simulation and real datasets

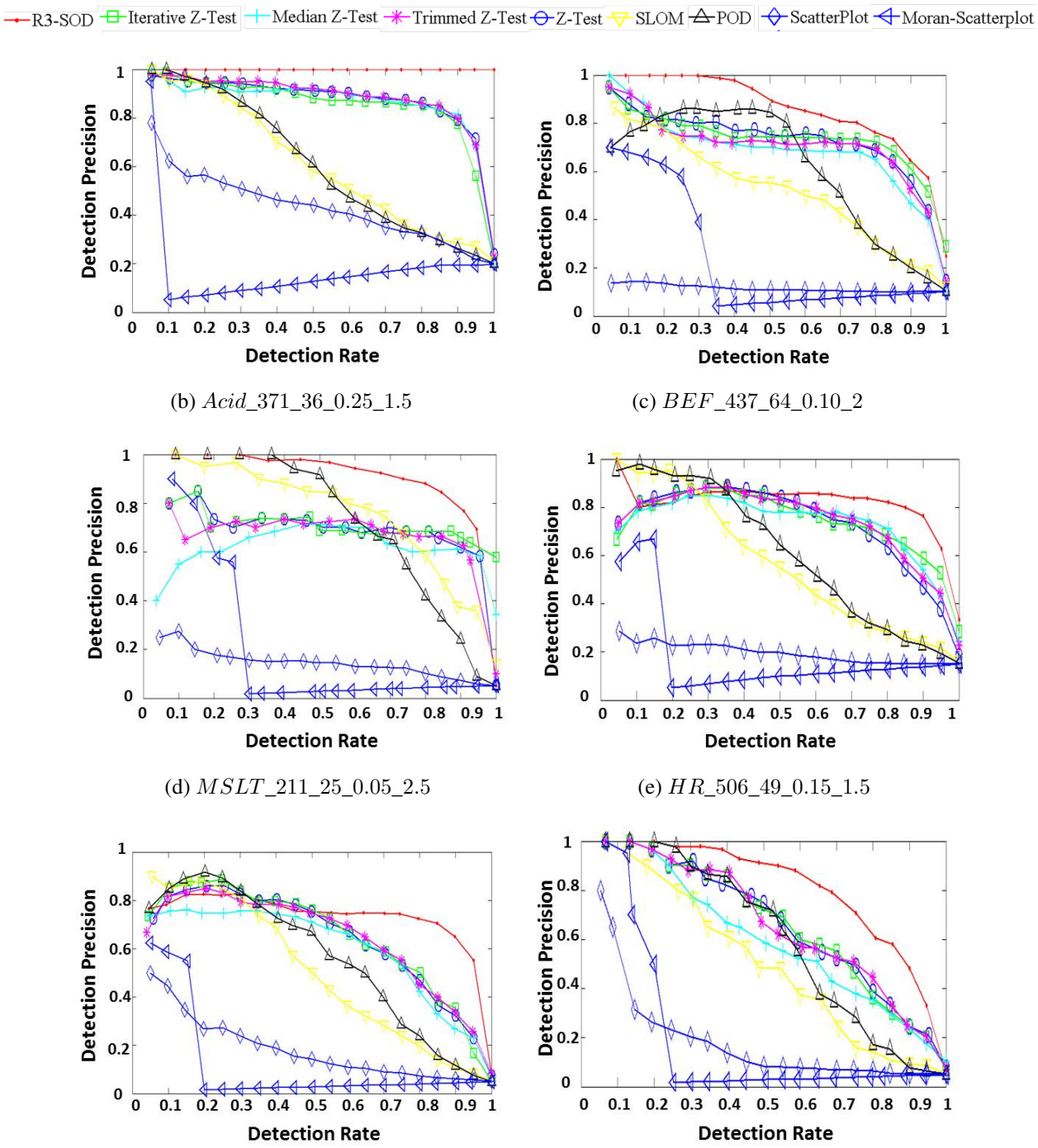
Even so, SKM and Regress performed much worse than R^3-SKM on this data. As depicted by Table 6.4, when estimating β value, the estimation errors for R^3-SKM were 12% and 2.7% for β_1 and β_2, respectively. However, SKM had values of 56% and 18%, and Regression 24% and 22.6% for the same data. When estimating σ and ϕ, the estimation errors for R^3-SKM were 4% and 1%, while for SKM, they were 76.8% and 9%. This considerable difference in the estimation errors implies that SKM was highly influenced by

the existence of outliers. There is no result for σ and ϕ for the Regression model, since this approach does not take the spatial dependency into consideration. And, as a consequence, it always incorrectly estimated β with higher errors, which absorbed the spatial variations into the spatial mean($X^T\beta$). Compared with SKM and Regression, R^3-SKM can be resilient to the influence of outliers, even in datasets in which more data were heavily contaminated, such as, "Sim_300_30_36_0.25_7" in which 25% of data were skewed with a higher shift rate(7). Not surprisingly, some experimental results indicate that if the data is severely contaminated, it is more difficult to accurately estimate parameters. This is demonstrated by the lower performances of SKM and Regression in Sim_* **_5** and Sim_* **_7**. Still, R^3-SKM was able to achieve very impressive estimation results since the integration of the heavy tailed distribution helped alleviate the impact of outliers.

Robustness on Spatial Prediction. To better analyze prediction performances, we utilized Moran's I-statistic to capture the spatial dependency of Y observations. This made it possible to learn more about the degree of spatial auto-correlation for each dataset. The spatial dependency for Y in simulations was computed to be 0.70, which means that the simulations have a higher spatial dependency and this must be accurately captured during the estimation and prediction processes. The last column in Table 6.3 shows the calculated spatial dependencies for real datasets. Most are lower values, which implies that the non-spatial attribute(X) contributes a great deal to the prediction the outcome variables.

Figure 6.4 compares the performances of different prediction models for simulated and real datasets. The calculated RMSEs and MAPEs demonstrate that R^3-SP outperforms both Regression and SKM. In particular, SKM did not perform as good as the Regression model in spite of the fact it takes into account the spatial auto-correlation in its spatial predictions. This is because the SKM model is considerably more complicated. It consists of a vector representing the spatial mean($x^T\beta$), the spatial random process(η) and measurement error ϵ. Regression is composed of $x^T\beta$ and ϵ. The hidden process η captures spatial association which is assumed to be a multivariate Gaussian process. When there are outliers in the dataset, SKM treats their outlying behaviors as natural spatial variations in the dataset, which therefore affects the computation of η, and further degrades the prediction quality. Meanwhile, the greater the outlying degrees of outliers, the worse its prediction performance: for the cases of HR_406_100_49_0.05_7, $MLST$_150_61_25_0.20_7 and Sim_300_30_36_0.25_7, where the shift rate are all 7. Interestingly, R^3-SP generated very similar prediction result to that for the Regression model for $Acid$_271_100_25_0.10_3.5. This is because Acid data has a very low spatial dependency, 0.007. So, spatial mean($x^T\beta$) dominates the prediction results. But for datasets with high spatial dependencies, like Sim_300_30_36_0.25_7 and $House$_1000_200_25_0.05_3, R^3-SP has preceding performance increases by benefiting from the integration of the heavy tailed distribution.

Accuracy of Spatial Outlier Detection. The outlier detection accuracies of different methods were compared based on different combinations of parameter settings. Figure 6.5 shows six representative results from



(b) $Acid\_371\_36\_0.25\_1.5$ (c) $BEF\_437\_64\_0.10\_2$

(d) $MSLT\_211\_25\_0.05\_2.5$ (e) $HR\_506\_49\_0.15\_1.5$

(f) $House\_1000\_81\_0.05\_1.5$ (g) $Sim\_500\_64\_0.10\_1.5$

Figure 6.5: Comparison of SOD performances on simulation and real datasets

simulated and real datasets. Clearly, R^3-SOD has impressive identification performances, achieving 10-15%

improvement over Z, Median-Z, Iterative-Z and Trimmed-Z, 20-30% over POD and SLOM, 40-50% over Moran-Scatterplot, and 60-70% over Scatterplot. Z series of approaches identify outliers by normalizing the difference between a spatial object and the average of its spatial neighbors. However, this difference value is easily influenced by the presence of one or more outliers in its neighborhood, which leads to worse outlier detection qualities. This is especially true for higher numbers of outliers in the neighborhood, as demonstrated by "BEF_437_64_0.10_2" and "Sim_500_64_0.10_1.5". POD method constructs a graph based on k nearest neighbors, assigns the non-spatial attribute differences as edge weights, and then continuously cuts high weight edges to identify isolated points as outliers. Its performance degrades significantly with increasing outlier sizes. Such as in "$MSLT$_211_25_0.05_2.5", its performance was better than others since only 5% of the data were contaminated. The MoranScatterplot and Scatterplot approaches detect outliers by normalizing the attribute values against the average values for the corresponding neighborhood, which greatly neutralizes the significant differences caused by outliers and results in poor performances. It is worth mentioning that, if the outlying degrees of outliers are much higher(such as the shift rate is set to 4 and 5), all the SOD approaches can get good identification results. However, if the outlying behavior is less differentiated, they did not accurately capture spatial outliers at all except R^3-SOD. In contrast, R^3-SOD can be easily resilient to the outliers with different outlying degrees. When identifying outliers, it does not rely on neighborhood differences, which are susceptible to the neighborhood size and presence of outliers. Rather, it statistically analyzes the data model by integrating a heavy tailed distribution to minimize the effects of outliers. Its competing identification results are demonstrated by Figure 6.5.

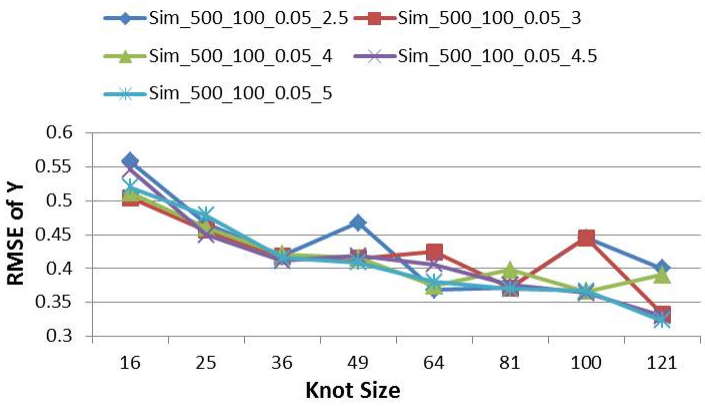


Figure 6.6: Prediction performances by varying knot sizes

Impact of Knot Sizes. We also evaluated the prediction performance by varying the knot sizes. Figure 6 shows various knot sizes from 16 to 121 (representing 3.2% and 24.2%, respectively, of the total number of observations). The curves show the effects of varying the number of knots on prediction accuracy. The simulations with different outlying degrees(2.5, 3, 4, 4.5, 5) have very similar affected trends for the different

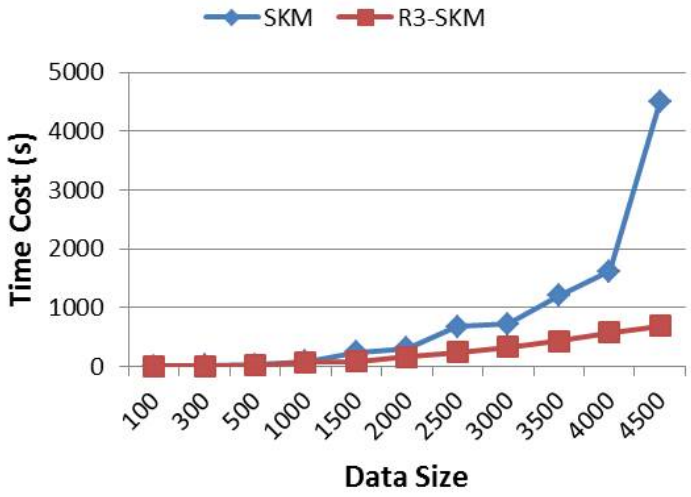


Figure 6.7: Total response time by varying data size

knot sizes. As shown in Figure 6.5, RMSEs decline as the knot size increases till they reach a stable state. In the simulations, which include 500 training and 100 testing points, the optimum prediction performance was achieved when knot size is equal to 64. That is, the knot size can be as high as 10%-15% of the total dataset. We also evaluated the optimal knot size in different sized datasets, and observed similar patterns. The optimal selection of the knot size enables not only more accurate spatial prediction, but also faster inferences.

Computational Cost. Finally, we examined the speed and associated scalability of SKM and R^3-SKM. Figure 6.7 displays the comparison of their runtime in datasets with varying numbers of training points. For all the simulations, the knot sizes were set to 10. Consequently, for R^3-SKM, when the data size is smaller than 1000, the time complexity is dominated by the knot size, $O(m^3) = 1000$, rather than data size, $O(n)$. Above 1000, the time cost increases linearly. In comparison, the time cost of SKM was observed to increase in a nonlinear fashion for larger datasets. In summary, the reduced-rank techniques enables our algorithms to perform efficiently with a linear time complexity.

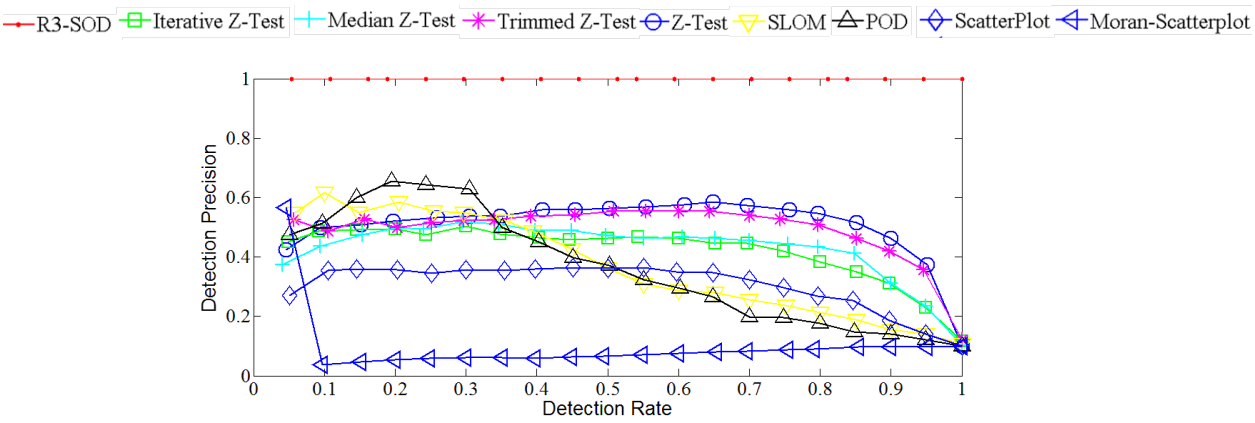
Result Discussion. R^3-SKM has been shown to be very robust for parameter estimation and spatial inference. It has superior performance over existing techniques in both real and simulation datasets. The experimental results verify three observations. First, if there is a good selection of knots that cover most of the domain interests, the predictive process cost will be significantly reduced to a linear order. Second, by being combined with numerical routines, Laplace approximation can provide much faster and more accurate parameter estimation. Third, integrating the heavy detailed distribution into the modeling process clearly minimizes the impact of outliers to a reasonable value, which provides a very good demonstration of the

new method's robustness.

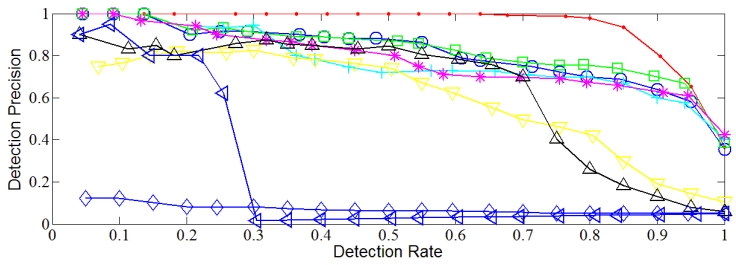
6.7 Conclusion

This paper proposes a Robust and Reduced-Rank Spatial Kriging Model for large spatial datasets, abbreviated as R^3-SKM. This approach integrates a Bayesian hierarchical framework to support priors on model parameters. Meanwhile, the measurement error is modeled by a heavy tailed distribution, which enables it to be resilient to the influences of outliers and allow for fast spatial inferences. Furthermore, three algorithms are proposed to perform robust parameter estimation, spatial prediction and spatial outlier detection, respectively, in linear time. Their robustness and efficiency were demonstrated by extensive experimental evaluations. R^3-SKM provides critical functionality for stochastic processes on large spatial datasets.

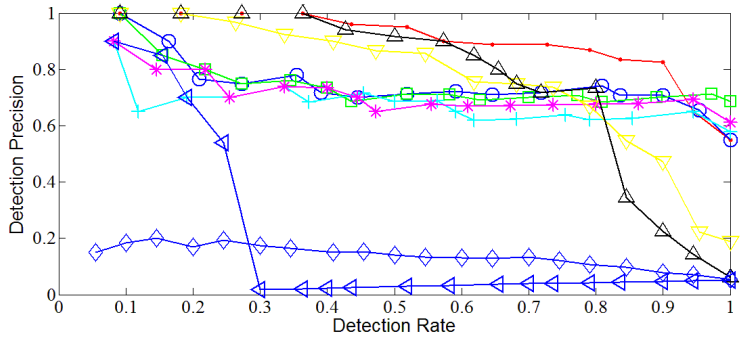
The limitation of the model is that it assumes the spatial data accords to the generalized linear model for the calucated local differences, but no justifications for this crtical assumption have been presented. Acutally, the prediction performances on geostatistic data with both linear and non-linear trend are required to studying in real application.



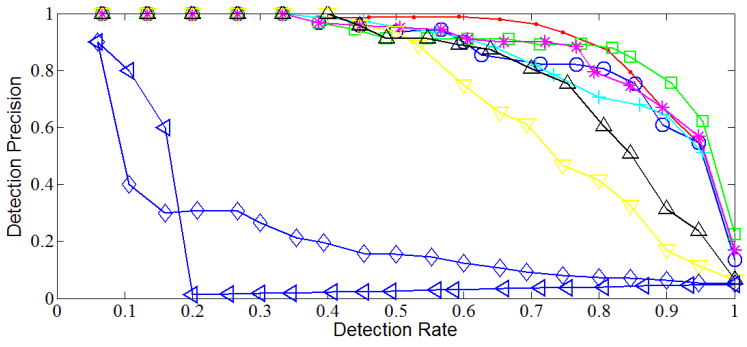
(b) $Acid\_371\_36\_0.25\_1.5$



(c) $BEF\_437\_64\_0.10\_2$

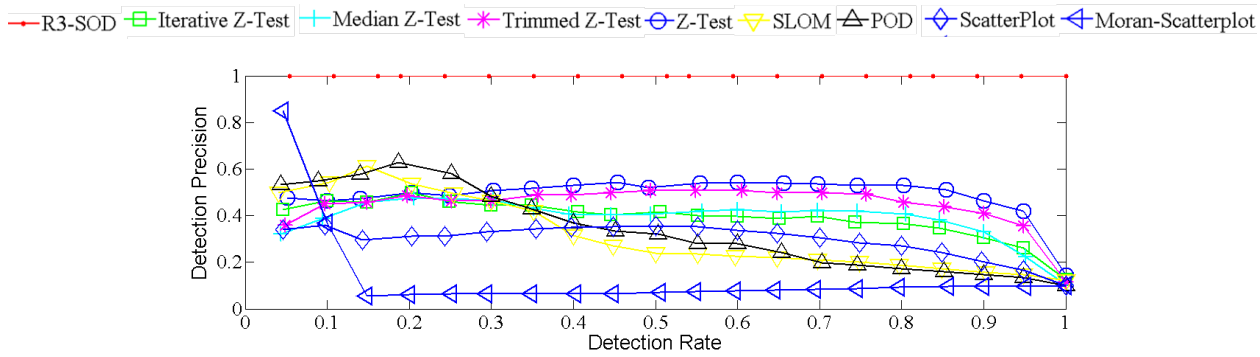


(d) $HR\_506\_49\_0.15\_1.5$

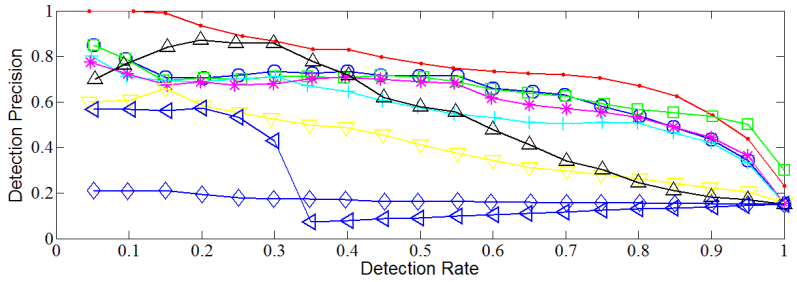


(e) $House\_1000\_81\_0.05\_1.5$

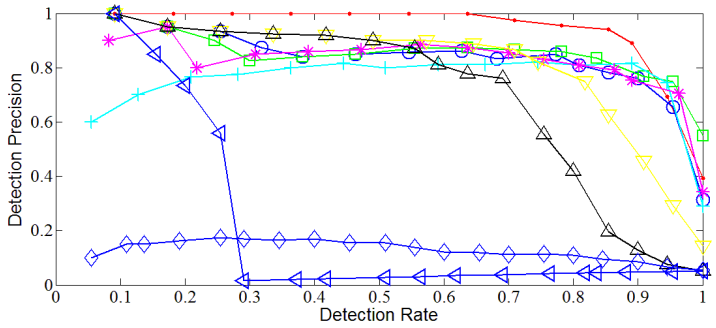
Figure 6.8: Comparison of SOD performances on real datasets: **Evaluation on Laplace Distribution**



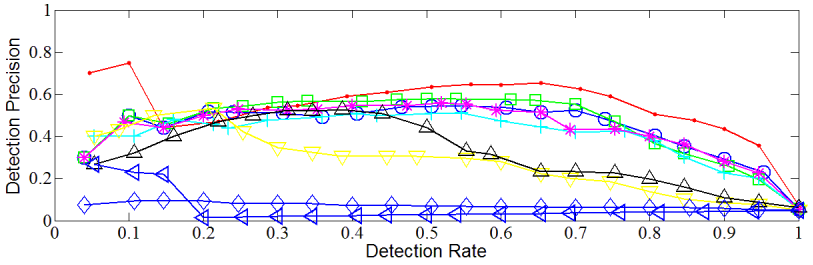
(b) *Acid*_371_36_0.25_1.5



(c) BEF_437_64_0.10_2



(d) HR_506_49_0.15_1.5



(e) $House$_1000_81_0.05_1.5

Figure 6.9: Comparison of SOD performances on real datasets: **Evaluation on Huber Distribution**

Chapter 7

Spatial Prediction of Large Multivariate Non-Gaussian Datasets

With the ever increasing volume of geo-referenced datasets, there is a real need for better statistical estimation and prediction techniques for spatial analysis. Most existing approaches focus on predicting multivariate Gaussian spatial processes, but as the data may consist of non-Gaussian (or mixed type) variables, this incurs two challenges: 1) how to accurately capture the dependencies among different data types, both Gaussian and non-Gaussian; and 2) how to efficiently predict multivariate non-Gaussian spatial processes for large datasets. In this paper, we propose a generic approach for predicting multiple response variables of mixed types. The proposed approach accurately captures cross-spatial dependencies among response variables and reduces the computational burden by projecting the spatial process to a lower dimensional space with knot-based techniques. In addition, efficient approximations are provided that estimate posterior marginals of latent variables for the predictive process. Extensive experimental evaluations based on both simulation and real-life datasets demonstrate the effectiveness and efficiency of this new approach.

The chapter is organized as follows. Section 7.1 gives the background and motivation. Section 7.2 reviews the theoretical background, including the exponential family, knot-based techniques and the INLA framework. Section 7.3 presents the new multivariate hierarchical non-Gaussian model and the reduced rank spatial predictive process. An approximate inference for multivariate spatial predictions is proposed in Section 7.4 and experiments on both simulated and real datasets are presented in Section 7.5. The paper concludes with a summary of the research in Section 7.6.

7.1 Introduction

The increasing public sensitivity and concern on environmental issues have led to as well as the developments of remote sensing technology, huge amounts of spatial data being collected, and this volume keeps increasing at an ever faster pace. As one of today's major research issues, the prediction of multivariate spatial observations has attracted significant attentions, particularly from those working in areas such as biology [107], epidemiology [16], geography [72], and economics [36]. Spatial prediction is the process of estimating the values of a target quantity at unvisited locations, based on the observed measures at sampled ones. When applied to a whole study area, it is also referred to as spatial interpolation or mapping. Observations are often of different types, such as continuous, ordinal, and binary, each of which conveys important information. For example, in economics studies, the living area (continuous variable), the age of the dwelling (ordinal variable), and an indicator that shows if the dwelling is located in a certain county (binary variable), are usually measured when characterizing the sale prices of houses. Most spatial prediction models focused on predicting one variable at a time. However, the spatial prediction of several related variables simultaneously also attracts significant attractions. In Ecology, ecologists are interested in predicting the joint abundance of species at unsampled spatial locations [74]. In the meantime, the interesting joint patterns of species can be identified. In geology, it is often desirable to predict related variables of different types, such as element concentrate, granularity and coloration for pedological data [27], based on data collected at nearby locations. This raises three research challenges: 1) Modeling cross-spatial dependencies between Gaussian and non-Gaussian variables; 2) Prediction of multivariate spatial observations; and 3) Efficient processing for large datasets.

In the univariate case, spatial prediction has been well studied for different data types, including continuous spatial processes [39, 160], discrete spatial processes [118, 161], and Poisson spatial processes [166]. Diggle et al. [45] presented an embedding linear Kriging framework for non-Gaussian attributes, which considers a latent spatial Gaussian process in the framework of Generalized Linear Mixed Model (GLMM). Kammann and Wand [85] extended Diggle et al.'s work and proposed a geoadditive model that integrates both the GLMM and additive models. All these approaches can be described in a hierarchical Bayesian framework, which decomposes a complicated model into a series of simpler conditional levels [164]. Recently, a number of methods have been proposed for processing large univariate spatial datasets of different types, including the fixed rank kriging [38], the knot-based spatial process [11], and the INLA based predictive process [136].

In many cases, geo-referenced data sets are multivariate. The prediction of multivariate spatial processes at unsampled locations plays an important role when there are cross-spatial dependencies between multiple response variables of interests and there is a considerable literature on the modeling and prediction of multivariate spatial processes [39]. Most of the related works focus on multivariate Gaussian processes, the key component of which is the modeling of cross-covariance functions between attributes at different spatial

locations. Commonly used constructive frameworks include the separable forms proposed by Mardia and Goodall [104]; the linear models of coregionalization (LMC) proposed by Wackernagel [160], which was extended by Gelfand and Sirmans [55]; and the moving average model designed by Ver Hoef and Barray [158]. The book by Wackernagel [160] and the review by Gelfand and Banerjee [54] provide a comprehensive survey of different spatial Gaussian multivariate modeling and prediction techniques.

However, only a limited amount of research has been proposed to support non-Gaussian multivariate spatial processes. One potential reason is the challenge of analytical intractable inference. Wibrin et al. [163] explored the Bayesian Maximum Entropy (BME) approach in which both continuous and categorical values are considered by a "cross-covariance" function. Schmidt and Rodriguez proposed Markov Chain Monte Carlo (MCMC) methods for modeling multivariate counts [138], while Chagneau et al. proposed a hierarchical Bayesian model for the modeling of Gaussian, count, and ordinal variables, and designed MCMC methods using the Gibbs sampler with Metropolis-Hastings (M-H) steps [28].

Both a KDD panel discussion [129] and a position paper [173] identified the prediction of multivariate non-Gaussian (or mixed type) data as one of the 10 most important challenges in data mining for the next decade. Most existing predictive algorithms for multivariate non-Gaussian spatial processes are designed within the framework of MCMC, which is a popular way to address problems that are analytically intractable [10, 17, 94, 134, 174]. Although MCMC has an immense flexibility for a variety of hierarchical models, it becomes prohibitively expensive for large spatial datasets [28]. This raises the question: How can we process large non-Gaussian multivariate spatial data as efficiently as Gaussian data? To the best of our knowledge, there is no previous work that addresses this challenging problem. In recent works, Rue et al. proposed an Integrated Nested Laplace Approximation (INLA) algorithm as an alternative to MCMC that achieves significant computation savings while preserving high accuracy [136]. Banerjee et al. proposed a knot-based model for multivariate Gaussian processes, which is also called "subset of regressors" method [11]. Banerjee's method reduces the time complexity from $O(n^3)$ to $O(nm^3)$, where n is the data cardinality and m is the number of knots, with $m \ll n$. This paper aims to address the prediction problem of multivariate non-Gaussian processes.

This paper presents a flexible hierarchical Bayesian framework that permits the simultaneous modeling of mixed type variables for larger datasets. Specifically, the mixed-type attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. Each attribute is mapped to a corresponding latent numerical variable via a specific link function, such as a logit function for binary attributes or a log function for count attributes. The dependency between mixed type attributes is then modeled by the relationship between their latent numerical random variables using a variance-covariance matrix. Computationally, we first utilize knot-based techniques [11] to model the multivariate predictive process as a reduced rank spatial process, and further reduce the dimensionality of the model. We then go on to develop a computational approach for multivariate variables using a Laplace Approximation [136] approach.

Our generic approach can be applied to a variety of spatial prediction applications where mixed-type attributes are involved, including geographical information systems [24], medical imaging [34], urban traffic modeling [96], weather forecasting [76, 153], and disease outbreak detection [103]. This technique can also be extended to other data mining problems, including spatial outlier detection [170, 171], spatial temporal outlier detection [96, 169], spatial-temporal scan [110], and spatial anomaly cluster identification [48, 114]. For example, for mixed-type spatial prediction, our approach can be extended by adding an additional variation component with a student-t distribution to absorb large variations due to outliers. The spatial outliers can then be detected based on the posterior distribution of this new variation component conditional observations.

Our major contributions can be summarized as follows:

- **Design of a spatial multivariate non-Gaussian hierarchical framework.** The spatial model is based on a hierarchical framework and is specifically designed to take account of mixed type random variables.
- **Model of a multivariate reduced-rank predictive process.** This is the first work that applies both knot-based and Laplace Approximation techniques to multivariate non-Gaussian datasets. The knot-based technique is utilized to model the predictive process as a reduced-rank spatial process, which projects the process realizations of the spatial model to a lower dimensional subspace. This projection significantly reduces the computational cost.
- **Design of an efficient spatial prediction algorithm.** By integrating the Laplace approximation, our approach efficiently makes approximations to the posterior marginal of latent variables for the predictive process, and performs accurate spatial prediction.
- **Performance analyses and experiment evaluation**. Theoretical analysis and extensive experiments on both simulations and real datasets have been conducted to demonstrate the performance of the proposed hierarchical mixed model. The datasets and the implementation of our model, as well as seven state-of-the-art comparison approaches, can be downloaded from [98] for evaluation.

7.2 Preliminary Concept

This section introduces the exponential family and the framework for the knot-based spatial process.

7.2.1 The exponential family

Let $Y(\mathbf{s})$ be a response variable at the location $\mathbf{s} \in \mathcal{D} \subset \mathcal{R}^2$. It is assumed that $Y(\mathbf{s})$ follows an exponential family distribution with the probability density

$$f(Y(\mathbf{s})|\theta(\mathbf{s}), \tau) = \exp\left(\frac{Y(\mathbf{s})\theta(\mathbf{s}) - a(\theta(\mathbf{s}))}{d(\tau)} + h(Y(\mathbf{s}), \tau)\right), \tag{7.1}$$

where $\theta(\mathbf{s})$ and τ are model parameters. $\theta(\mathbf{s})$ is related to the mean of the distribution that varies by location, and τ is a dispersion parameter related to the variance of the distribution. The functions $h(y(\mathbf{s}), \tau)$, $a(\theta(\mathbf{s}))$, and $d(\tau)$ are known. $Y(\mathbf{s})$ has mean and variance

$$E(Y(\mathbf{s})) := \mu(\mathbf{s}) = a'(\theta(\mathbf{s})), \tag{7.2}$$

$$Var(Y(\mathbf{s})) := \sigma(\mathbf{s})^2 = a''(\theta(\mathbf{s}))d(\tau), \tag{7.3}$$

where $a'(\theta(\mathbf{s}))$ and $a''(\theta(\mathbf{s}))$ are the first and second derivatives of $a(\theta(\mathbf{s}))$. Many popular distributions belong to this family, including the Gaussian, exponential, Binomial, Poisson, gamma, Inverse Gaussian, Dirichlet, and Chi-Squared Beta distributions.

For example, the Binomial distribution $B(n(\mathbf{s}), \pi(\mathbf{s}))$ has the density

$$p(Y(\mathbf{s})) = \binom{n(\mathbf{s})}{Y(\mathbf{s})} \pi(\mathbf{s})^{Y(\mathbf{s})} (1 - \pi(\mathbf{s}))^{n(\mathbf{s}) - Y(\mathbf{s})}. \tag{7.4}$$

Taking logs, we can rewrite the density function as

$$\log p(Y(\mathbf{s})) = Y(\mathbf{s}) \log(\frac{\pi(\mathbf{s})}{1 - \pi(\mathbf{s})}) + n(\mathbf{s}) \log(1 - \pi(\mathbf{s})) + \log \binom{n(\mathbf{s})}{Y(\mathbf{s})}. \tag{7.5}$$

This shows that $\theta(\mathbf{s}) = \log(\frac{\pi(\mathbf{s})}{1-\pi(\mathbf{s})})$, $a(\theta(\mathbf{s})) = n(\mathbf{s}) \log(1 + \exp\theta(\mathbf{s}))$, and $h(Y(\mathbf{s}), \tau) = \log \binom{n(\mathbf{s})}{Y(\mathbf{s})}$, where the second term in the density function is rewritten as $\log(1 - \pi(\mathbf{s})) = -\log(1 + \exp\theta(\mathbf{s}))$.

7.2.2 Knot-based spatial process model

Estimation and prediction in spatial process models often involve a high computational complexity, which is cubic order with the number of spatial locations. To facilitate the spatial process, Banerjee et al. [11] proposed a knot-based spatial predictive model to reduce the computational cost through lower dimensional process observations.

Let us define a numerical random field $Y(s)$ on a domain $D \subseteq \mathcal{R}^2$, and let $Y = (Y(s_1), \cdots, Y(s_n))'$ be the

$n \times 1$ vector of observed responses, each of which is accompanied by a $p \times 1$ vector of spatially referenced predictors, $x(s)$. The associated spatial regression model can be represented as

$$Y(s) = x^T(s)\beta + \omega(s) + \epsilon(s). \tag{7.6}$$

The spatial process $\omega(s)$ captures spatial correlations and is a Gaussian process with zero mean and a covariance function $C(s, s'; \theta)$. Spatial prediction requires matrix factorizations involving the dense $n \times n$ covariance matrix which may become prohibitively expensive for a large *n*. The $\omega(s)$ are spatial random effects, providing local adjustment(with structured dependence) to the mean, interpreted as capturing the effects of unmeasured or unobserved covariates with spatial pattern. Recently, Banerjee et al.[11]proposed a class of knot-based spatial process models for large spatial datasets. Instead, knot-based models consider a fixed set of "knots" $S^* = (s_1^*, \cdots, s_{n^*}^*)$ with $n^* \ll n$. The Gaussian process $\omega^*(s)$ yields an n^*-vector of realizations over the knots, that is, $\omega^* = (\omega(s_1^*), \cdots, \omega(s_{n^*}^*))'$, which follows a $GP\{0, C(s_i^*, s_j^*; \theta)\}$. Spatial estimation at a generic site s is operated through

$$\tilde{\omega}(s) = E\{\omega(s)|\omega^*\} = c^T(s; \theta)C^{*-1}(\theta)\omega^*, \tag{7.7}$$

where $c(s; \theta) = [C(s, s_j^*; \theta)_{j=1}^{n^*}]$. As shown in Eq. (2), the *predictive process* $\tilde{w}(s)$ is derived from the *parent process* $\omega(s)$. The realizations of $\tilde{\omega}(s)$ are referred to as the predictions that are conditional on a realization of $\omega^*(s)$. Replacing $\omega(s)$ in model (7.6) with $\tilde{\omega}(s)$, we obtain the predictive process model

$$Y(s) = x^T(s)\beta + \tilde{\omega}(s) + \epsilon(s), \tag{7.8}$$

where $\tilde{\omega}(s)$ is defined as a spatially varying linear transformation of ω^*. The dimension reduction is reduced from the original n to n^*, thus the spatial interpolation process involves only $n^* \times n^*$ matrices.

It is important to select an appropriate number of knots as well as their spatial locations. This is related to the problem of spatial design. There are two popular knots selection strategies. One is to draw a uniform grid to cover the study region and each grid is considered as a knot. Another is to place knots such that each covers a local domain and the regions with dense data have more knots. In practice, it is feasible to validate models by using different number of knots and different choices of knots to obtain a reliable and robust configuration.

7.2.3 The INLA approach

The INLA(Integrated Nested Laplace approximation) [136] is a computational approach which is proposed as an alternative of the time consuming MCMC method. The INLA approximation performs Bayesian

Table 7.1: Description of Major Symbols

Symbol	Description
S	$S = \{s_1, \cdots, s_n\}$, a set of n training locations, where $s_i \in \mathbb{R}^2$;
S^*	$S^* = \{s_1^*, \cdots, s_m^*\}$, a set of m knot locations, where $s_i^* \in \mathbb{R}^2$;
Y	A given set of observations with a numerical attribute that follows a Gaussian distribution. $Y = \{Y(s_i)\}_{i=1}^n$;
Z	A given set of observations with a discrete attribute. If a count dataset (Z_c), it follows a Poisson distribution; if a binary dataset (Z_b), it follows a Binomial distribution. $Z = \{Z(s_i)\}_{i=1}^n$;
X	A set of explanation variables. $\{X(s_i)\}_{i=1}^n$ is a $p \times 1$ vector at location s_i. $X = \{X(s_i)\}_{i=1}^n$;
ω, γ	Spatial random effects of the observations, which provide local adjustments to the means, and is interpreted as the effects of unmeasured covariates with spatial patterns. $\omega = \{\omega(s_i)\}_{i=1}^n$,$\gamma = \{\gamma(s_i)\}_{i=1}^n$;
ω^*, γ^*	Spatial random effects of the knots. $\omega^* = \{\omega^*(s_i)\}_{i=1}^m$, $\gamma^* = \{\gamma^*(s_i)\}_{i=1}^m$;
$\tilde{\omega}, \tilde{\gamma}$	The predicted values of ω, γ by ω^*, γ^*. $\tilde{\omega} = \{\tilde{\omega}(s_i)\}_{i=1}^n$,$\tilde{\gamma} = \{\tilde{\gamma}(s_i)\}_{i=1}^n$;
$\tilde{\omega}_\epsilon, \tilde{\gamma}_\epsilon$	The corrected predictive values of $\tilde{\omega}, \tilde{\gamma}$. $\tilde{\omega}_\epsilon = \{\tilde{\omega}_\epsilon(s_i)\}_{i=1}^n$, $\tilde{\gamma}_\epsilon = \{\tilde{\gamma}_\epsilon(s_i)\}_{i=1}^n$;
$\tilde{\epsilon}_y, \tilde{\epsilon}_z$	The estimation bias values for $\tilde{\omega}$,$\tilde{\gamma}$. $\tilde{\epsilon}_y(s) = \{\tilde{\epsilon}_y(s_i)\}_{i=1}^n$, $\tilde{\epsilon}_z(s) = \{\tilde{\epsilon}_z(s_i)\}_{i=1}^n$. And, $\{\tilde{\omega}_\epsilon(s_i) = \tilde{\omega}(s_i) + \tilde{\epsilon}_y(s_i)\}_{i=1}^n$, $\{\tilde{\gamma}_\epsilon(s_i) = \tilde{\gamma}(s_i) + \tilde{\epsilon}_z(s_i)\}_{i=1}^n$;
ϵ	$\{\epsilon(s_i)\}_{i=1}^n$ is the nugget measurement error for $\{Y(s_i)\}_{i=1}^n$. $\epsilon = \{\epsilon(s_i)\}_{i=1}^n$;
v^*	$v^* = ((\omega^{*'}, \gamma^{*'}), (\beta_y', \beta_z'))'$, it is a $(2m + 2p) \times 1$ vector comprising the realizations of the spatial multivariate predictive process and the regression parameters;
ϕ	The decay and smoothness parameter;
$F(\phi)$	A transformation matrix that defines $\{\tilde{\omega}, \tilde{\gamma}\}$ as a spatially varying linear transformation of $\{\omega^*, \gamma^*\}$. See Eq.(7.24,7.27) for $F(\phi)$;
η_z	The expected value of Z which is linear on a transformed scale. $\eta_z = \{\eta_z(s_i)\}_{i=1}^n$. $\eta_z = H_z^* v^*$.
H_y^*	$H_y^* = [F_y(\phi), \begin{bmatrix} X & 0_{n\times p} \end{bmatrix}]$. $F_y(\phi)$ consists of the first n rows of matrix $F(\phi)$;
H_z^*	$H_z^* = [F_z(\phi), \begin{bmatrix} 0_{n\times p} & X \end{bmatrix}]$. $F_z(\phi)$ consists of the last n rows of matrix $F(\phi)$;
Θ	The set of sample locations of θ based on the mode and Hessian at it of $\hat{\pi}(\theta\|Y, Z)$. $\Theta = \{\theta_k\}_{k=1}^K$;
w	The set of weighted values of sample θ, which are computed by their corresponding posterior distributions. $w = \{w_{\theta_k}\}_{k=1}^K$;

inferences in latent Gaussian fields, that is, the models of n observations that are conditinally independent given latent field v and the hyperparameter set, θ. It approximates the marginal posteriors of latent variables.

$$\pi(v_i|Y) = \int \pi(v_i|\theta, Y)\pi(\theta|Y)d\theta. \tag{7.9}$$

This approximation is an efficient combination of Laplace approximations to the full conditionals $\pi(\theta|Y)$ and $\pi(v_i|\theta, Y)$, and finally executes numerical integration routines by integrating out the parameter θ.

The INLA approach consists of three main approximations to obtain the marginal posterior for each latent variable. The first step is to approximate the full posterior $\pi(\theta|Y)$, which is executed using the Laplace approximation

$$\tilde{\pi}(\theta|Y) \propto \frac{\pi(v, \theta, Y)}{\tilde{\pi}_G(v|\theta, Y)}\Big|_{v=v^*(\theta)}. \tag{7.10}$$

As shown above, we need to approximate the full conditional distribution of $\pi(v|Y, \theta)$, which can be achieved by a multivariate Gaussian density $\tilde{\pi}_G(v|Y, \theta)$ [135]. The $v^*(\theta)$ is the mode of the full conditional distribution of v for a given θ and can be estimated using $\tilde{\pi}_G(v|Y, \theta)$. The posterior $\tilde{\pi}(\theta|Y)$ will be used later to integrate out the uncertainty with respect to θ when approximating $\pi(v_i|Y)$.

The second step executes the Laplace approximation of the full conditionals $\pi(v_i|\theta, Y)$ for specified θ values. The density $\pi(v_i|\theta, Y)$ is approximated using Laplace approximation defined by

$$\tilde{\pi}_{LA}(v_i|\theta, Y) \propto \frac{\pi(v, \theta, Y)}{\tilde{\pi}_G(v_{-i}|v_i, \theta, Y)}\Big|_{v_{-i}=v^*(v_i,\theta)}, \tag{7.11}$$

where $\tilde{\pi}_G(v_{-i}|v_i, \theta, Y)$ refers to the Gaussian approximation of $\pi(v_{-i}|v_i, \theta, Y)$ which takes the v_i as a fixed value. $v^*(v_i, \theta)$ is the mode of $\pi(v_{-i}|v_i, \theta, Y)$.

Finally, we can approximate the marginal posterior density of v_i by combining the full posteriors obtained in the previous steps. The approximation expression is shown as follows.

$$\pi(v_i|Y) \approx \sum_k \tilde{\pi}(v_i|\theta_k, Y)\tilde{\pi}(\theta_k|Y)\triangle_k. \tag{7.12}$$

It is a numerical summation on a representative set of θ_k, with the area weight, $\triangle_k$ for $k = 1, \cdots, K$. Note that a good choice of the set of θ_k is crucial to the accuracy of the above numerical integration. Rue et al.[136] suggested to compute the negative Hessian matrix at the modal configuration, and then select the set θ_k of evaluation points by stepping along the main directions away from the model until $log[\hat{\pi}\{\theta|Y\}]$ is negligible. All θ-points satisfying

$$log[\hat{\pi}\{\theta(0)|Y)\}] - log[\hat{\pi}\{\theta_k|Y\}] < \eta_\pi \tag{7.13}$$

With the θ_k values, $\pi(v_i|Y)$ in (7.13) is evaluated and normalized.

7.3 Spatial Multivariate Non-Gaussian Model

This section presents a spatial process model for random variables that consists of one Gaussian and one non-Gaussian variables from the exponential family. The multivariate non-Gaussian Model is designed based on a Bayesian hierarchical framework that allows any number of response variables. The computational challenges in modeling large mixed type spatial datasets are addressed by integrating with the knot-based predictive process. Table 1 summarizes the key notations used in this paper.

7.3.1 Model formulation

The spatial multivariate predictive model is specifically designed to deal with variables of different types. The model can be defined for any number of response variables. Here, we consider two different types: Gaussian and non-Gaussian variables (e.g., Binomial or Poisson).

Let $s_1, \cdots, s_n$ be the n sampled locations, $Y(s_i)$ be a Gaussian variable at location s_i, and $Z(s_i)$ be a non-Gaussian variable, such as a Poisson variable. Let $Y = (Y(s_1), \cdots, Y(s_n))'$ and $Z = (Z(s_1), \cdots, Z(s_n))'$. Geostatistics typically assumes that the Gaussian response variable $Y(s)$ is modeled as a spatial regression model with a $p \times 1$ vector of spatially referenced predictors, $x(s)$, such as

$$Y(s) = x(s)^T \beta_y + \omega(s) + \epsilon(s). \tag{7.14}$$

The residual includes the spatial random effect, $\omega(s)$, and the independent process $\epsilon(s)$, known as the nugget. Usually, $\epsilon(s) \sim N(0, \tau^2)$. The $\omega(s)$ provides a local adjustment to the mean, interpreted as the effect of unmeasured covariates on the spatial pattern.

Let the first stage of Z be the non-Gaussian process. Essentially, we assume that the function of the expected value of $Z(s_i)$ is linear on a transformed scale, such as

$$\eta_z(s) \equiv g(E(Z(s))) = x(s)^T \beta_z + \gamma(s), \tag{7.15}$$

where $g(\cdot)$ is a suitable link function.

The Gaussian variable $Y(s_i)$ and the non-Gaussian variable $Z(s_i)$ depend on the latent variables $\omega(s_i)$ and $\gamma(s_i)$, respectively, which are together responsible for the spatial dependences. Given $\omega(s_i)$ and $\gamma(s_i)$, the variables $Y(s_i)$ and $Z(s_i)$ are conditionally independent. The customary process specification for $(\omega', \gamma')'$

is a mean zero Gaussian process with covariance function, $\Sigma_{(\omega',\gamma')'}$, denoted as $GP(0, \Sigma_{(\omega',\gamma')'})$. The most obvious specification of a valid cross-covariance function for $(\omega', \gamma')'$ is to let ρ be a valid correlation function for a univariate spatial process. Let T be a $d \times d$ (here $d = 2$ refers to the dimension of the dataset) positive definite matrix $T = \begin{pmatrix} \sigma_y^2 & \sigma_{yz}^2 \\ \sigma_{yz}^2 & \sigma_z^2 \end{pmatrix}$, which is interpreted as the covariance matrix associated with $(\omega', \gamma')'$. T follows an Inverse Wishart distribution, denoted as $T \sim IW(\Psi, m)$. And, $\rho(s_i, s_j; \phi)$ attenuates association as s_i and s_j become farther apart. The covariance matrix for $(\omega', \gamma')'$ is easily shown to be

$$\Sigma_{(\omega',\gamma')'} = R(\phi) \otimes T, \tag{7.16}$$

where $R(\phi)_{i,j} = \rho(s_i, s_j; \phi)$ is a correlation function, ϕ includes decay and smoothness parameters, yielding constant process variances, and $\otimes$ denotes the Kronecker product.

The prior distributions of the remaining parameters construct the third level of the hierarchical model. Customarily, the regression parameters β_y and β_z are assigned multivariate Gaussian priors, i.e., $\beta_y \sim N(\mu_{\beta_y}, \Sigma_{\beta_y}), \beta_z \sim N(\mu_{\beta_z}, \Sigma_{\beta_z})$, while the latent variance components σ_y, σ_z, and σ_{yz} are assigned IW as introduced above. The nugget variance τ^2 is assigned an $IG(a_\tau, b_\tau)$ prior (Inverse Gamma). The process correlation parameter ϕ is usually assigned an informative prior (e.g., uniform over a finite range) based on the underlying spatial domain.

With n locations, say $S = (s_1, \cdots, s_n)$, the process realizations are collected into an $2n \times 1$ vector, say $\begin{pmatrix} \omega \\ \gamma \end{pmatrix} = (\omega(s_1), \cdots, \omega(s_n), \gamma(s_1), \cdots, \gamma(s_n))'$, which follows a multivariate normal distribution with mean 0 and dispersion matrix $\Sigma_{(\omega',\gamma')'} = \begin{pmatrix} \sigma_y^2 R(\phi) & \sigma_{yz}^2 R(\phi) \\ \sigma_{yz}^2 R(\phi) & \sigma_z^2 R(\phi) \end{pmatrix}$ with $\rho(s_i, s_j; \phi)$ being the $(i,j)^{th}$ element of $R(\phi)$.

Let Y and Z be two $n \times 1$ vectors of observed responses. The mixed data likelihood can be obtained by combining with hierarchical specifications, as shown in Fig. 7.1, to derive a posterior distribution $\pi(\beta_y, \beta_z, \omega, \gamma, T, \tau^2, \phi | Y, Z)$ that is proportional to

$$\begin{aligned} &\pi(\phi) \times IG(\tau^2 | a_\tau, b_\tau) \times IW(T | \Psi, m) \times N(\beta_y | \mu_{\beta_y}, \Sigma_{\beta_y}) \times N(\beta_z | \mu_{\beta_z}, \Sigma_{\beta_z}) \\ &\times N\left(\begin{pmatrix} \omega \\ \gamma \end{pmatrix} | 0, \Sigma_{(\omega',\gamma')'} \right) \times \prod_{i=1}^{n} N(Y(s_i) | x(s_i)^T \beta_y + \omega(s_i), \tau^2) \\ &\times \prod_{i=1}^{n} \pi(Z(s_i) | x(s_i)^T \beta_z + \gamma(s_i))). \end{aligned} \tag{7.17}$$

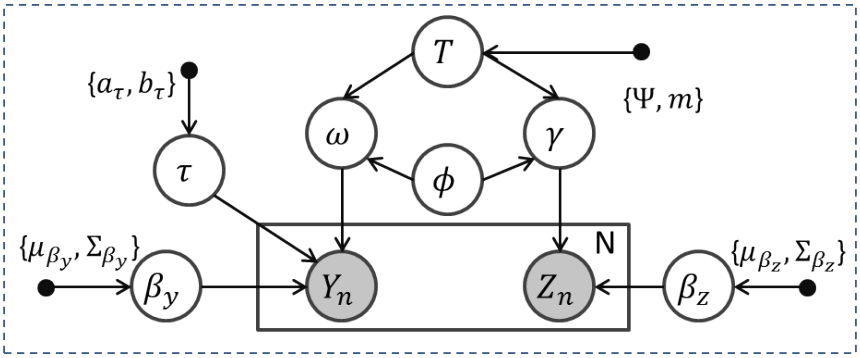


Figure 7.1: Graphical Model Representation

7.3.2 Reduced-rank spatial multivariate non-Gaussian process

For the spatial multivariate non-Gaussian process model, both the estimation and prediction steps require evaluating the $(d * n) \times (d * n)$ covariance matrix among d dependent response variables. Therefore, fitting hierarchical mixed models often involves expensive matrix decompositions whose computational cost is $O((d * n)^3)$, thus making such model not scalable for large spatial data sets. To address this challenge, we take the predictive process models into consideration for multivariate datasets. In this section, the spatial multivariate predictive model, known as the reduced-rank spatial multivariate non-Gaussian process, is designed by projecting the full process into a subspace generated by a specified set of representative locations.

Consider a set of "knots", $S^* = \{s_1^*, \cdots, s_m^*\}$, the vector of corresponding centroids of the m spatial clusters generated by the spatial attributes of the dataset. The latent variables $(\omega^{*'}, \gamma^{*'})'$ follow a mean zero Gaussian distribution with the covariance function, $\Sigma_{(\omega^{*'},\gamma^{*'})'} = R^*(\phi) \otimes T$, denoted as $GP(0, \Sigma_{(\omega^{*'},\gamma^{*'})'})$. $R^*(\phi)$ is the corresponding $m \times m$ covariance matrix, where $R^*(\phi)_{i,j} = \rho(s_i^*, s_j^*; \phi)_{i,j=1,\cdots,m}$. The spatial interpolant at a site s_0 is estimated by

$$\begin{aligned} \begin{pmatrix} \tilde{\omega}(s_0) \\ \tilde{\gamma}(s_0) \end{pmatrix} &= E\left\{ \begin{pmatrix} \omega(s_0) \\ \gamma(s_0) \end{pmatrix} \middle| \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} \right\} \\ &= \Upsilon(s_0)\Sigma^{-1}_{(\omega^{*'},\gamma^{*'})'} \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} = \begin{pmatrix} f_\omega^\omega(s_0) & f_\omega^\gamma(s_0) \\ f_\gamma^\omega(s_0) & f_\gamma^\gamma(s_0) \end{pmatrix} \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix}. \end{aligned} \tag{7.18}$$

Here, $\Upsilon(s_0) = r(s_0; \phi)' \otimes T$, and $r(s_0; \phi)$ is an $m \times 1$ vector whose j^{th} element is given by $\rho(s, s_j^*; \phi)$. The f series represent four $1 \times m$ matrices. This yields a spatial Gaussian process $(\tilde{\omega}', \tilde{\gamma}')' \sim GP(0, \tilde{\rho} \otimes T)$, where $\tilde{\rho}(s_i, s_j; \phi) = \Upsilon(s_i)\Sigma^{-1}_{(\omega^{*'},\gamma^{*'})'}\Upsilon'(s_j)$ and $(\tilde{\omega}', \tilde{\gamma}')'$ is referred to as the *predictive process* derived from the *parent process* $(\omega', \gamma')'$. As shown in Eq. (7.18), $(\tilde{\omega}(s)', \tilde{\gamma}(s)')'$ is a spatially adaptive linear transformation of the realizations of $(\omega(s)', \gamma(s)')'$ over S^* with $\Upsilon(s_0)\Sigma^{-1}_{(\omega^{*'},\gamma^{*'})'}$ comprising the coefficients

of the transformation.

Proof. Assume $x \sim N_x(\mu, \Sigma)$, where

$$x = \begin{pmatrix} \begin{pmatrix} \omega(s_0) \\ \gamma(s_0) \end{pmatrix} \\ \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} \end{pmatrix}, \mu = \begin{pmatrix} 0_{2\times 1} \\ 0_{2m\times 1} \end{pmatrix}, \Sigma = \begin{pmatrix} T & r(s_0;\phi)' \otimes T \\ r(s_0;\phi) \otimes T & R(\phi) \otimes T \end{pmatrix} \tag{7.19}$$

then,

$$\begin{aligned} \begin{pmatrix} \tilde{\omega}(s_0) \\ \tilde{\gamma}(s_0) \end{pmatrix} &= 0_{2\times 1} + \{r(s_0;\phi)' \otimes T\} * \{R(\phi) \otimes T\}^{-1} * \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} \\ &= \Upsilon(s_0)\Sigma^{-1}_{(\omega^{*\prime},\gamma^{*\prime})'} \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} = \begin{pmatrix} f^{\omega}_{\omega}(s_0) & f^{\gamma}_{\omega}(s_0) \\ f^{\omega}_{\gamma}(s_0) & f^{\gamma}_{\gamma}(s_0) \end{pmatrix} \begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} \end{aligned} \tag{7.20}$$

Replacing $\omega(s)$ and $\gamma(s)$ in Eqs. (7.14-7.15) with $\tilde{\omega}$ and $\tilde{\gamma}$, we obtain the reduced-rank predictive model,

$$Y(s) = x(s)^T\beta_y + \tilde{\omega}(s) + \epsilon(s), \tag{7.21}$$

$$\eta_z(s) \equiv g(E(Z(s))) = x(s)^T\beta_z + \tilde{\gamma}(s). \tag{7.22}$$

Using Eqs. (7.21-7.22) as the likelihood, we obtain the predictive process counterpart of Eq. (7.17) as

$$\begin{aligned} &\pi(\phi) \times IG(\tau^2|a_\tau, b_\tau) \times IW(T|\Psi, m) \times N(\beta_y|\mu_{\beta_y}, \Sigma_{\beta_y}) \times N(\beta_z|\mu_{\beta_z}, \Sigma_{\beta_z}) \\ &\times N\left(\begin{pmatrix} \omega^* \\ \gamma^* \end{pmatrix} |0, \Sigma_{(\omega^{*\prime},\gamma^{*\prime})'} \right) \times \prod_{i=1}^{n} N(Y(s_i)|x(s_i)^T\beta_y + \tilde{\omega}(s_i), \tau^2) \\ &\times \prod_{i=1}^{n} \pi(Z(s_i)|x(s_i)^T\beta_z + \tilde{\gamma}(s_i))). \end{aligned} \tag{7.23}$$

The reduced variability in $\tilde{\omega}$ often incurs an overestimation of the nugget variance τ^2. Banerjee et al. [11] detailed these biases. With regard to this issue, Finley et al. [50] proposed replacing $\tilde{\omega}(s)$ and $\tilde{\gamma}(s)$ in Eqs. (7.21-7.22) with $\tilde{\omega}_\epsilon(s) = \tilde{\omega}(s) + \tilde{\epsilon}_y(s)$ and $\tilde{\gamma}_\epsilon(s) = \tilde{\gamma}(s) + \tilde{\epsilon}_z(s)$. Using $\tilde{\omega}_\epsilon(s)$ and $\tilde{\gamma}_\epsilon(s)$ instead of $\tilde{\omega}(s)$

and $\tilde{\gamma}(s)$ as the spatial process yields

$$\pi(\phi) \times IG(\tau^2|a_\tau, b_\tau) \times IW(T|\Psi, m) \times N(\beta_y|\mu_{\beta_y}, \Sigma_{\beta_y}) \times N(\beta_z|\mu_{\beta_z}, \Sigma_{\beta_z}) \times$$
$$N\left(\begin{pmatrix}\omega^* \\ \gamma^*\end{pmatrix}|0, \Sigma_{(\omega^{*\prime},\gamma^{*\prime})^\prime}\right) \times N\left(\begin{pmatrix}\tilde{\omega}_\epsilon \\ \tilde{\gamma}_\epsilon\end{pmatrix}\middle| F(\phi)\begin{pmatrix}\omega^* \\ \gamma^*\end{pmatrix}, \Sigma_{(\epsilon_y^\prime,\epsilon_z^\prime)^\prime}\right) \times$$
$$\prod_{i=1}^{n} N(Y(s_i)|x(s_i)^T\beta_y + \tilde{\omega}_\epsilon(s_i), \tau^2) \times \prod_{i=1}^{n} \pi(Z(s_i)|x(s_i)^T\beta_z + \tilde{\gamma}_\epsilon(s_i))). \tag{7.24}$$

$F(\phi) = (\mathcal{R}(\phi)^\prime \otimes T)\Sigma^{-1}_{(\omega^{*\prime},\gamma^{*\prime})^\prime}$, where $\mathcal{R}(\phi)^\prime$ is an $n \times m$ matrix whose i^{th} row is given by $r(s_i;\phi)^\prime$, and $r(s_i;\phi)$ is an $m \times 1$ vector whose j^{th} element is given by $\rho(s_i, s_j^*;\phi)$, for $i = 1, \cdots, n$, $j = 1, \cdots, m$. And, $\Sigma_{(\epsilon_y^\prime,\epsilon_z^\prime)^\prime}$ is a $2n \times 2n$ matrix which consists of four diagonal matrices ($n \times n$) in which the following 4 specified diagonal elements $\begin{pmatrix}(i,i)^{th} & (i+n,i)^{th} \\ (i,i+n)^{th} & (i+n,i+n)^{th}\end{pmatrix}$ are computed as $T - \Upsilon(s_i)\Sigma^{-1}_{(\omega^{*\prime},\gamma^{*\prime})^\prime}\Upsilon^\prime(s_i)$, where $\Upsilon(s_i) = r(s_i;\phi)^\prime \otimes T$.

Let $v^* = ((\omega^{*\prime}, \gamma^{*\prime}), (\beta_y^\prime, \beta_z^\prime))^\prime$ be a $(2m + 2p) \times 1$ vector comprising the realizations of the spatial multivariate predictive process and the regression parameters. Since Z is related to the discrete variables, we assume there is no estimation bias [136]. The posterior $\pi(v^*, T, \phi, \tau^2|Y, Z)$ is proportional to

$$\pi(\phi) \times IG(\tau^2|a_\tau, b_\tau) \times IW(T|\Psi, m) \times N(v^*|\mu_{v^*}, \Sigma_{v^*}) \times$$
$$\prod_{i=1}^{n} N(Y(s_i)|x(s_i)^T\beta_y + f_\omega^\omega(s)\omega^* + f_\omega^\gamma(s)\gamma^* + \tilde{\epsilon}_y(s), \tau^2) \times$$
$$\prod_{i=1}^{n} \pi(Z(s_i)|x(s_i)^T\beta_z + f_\gamma^\omega(s)\omega^* + f_\gamma^\gamma(s)\gamma^*), \tag{7.25}$$

where $\mu_{v^*} = (0_{1\times 2m}, \mu_{\beta_y}^\prime, \mu_{\beta_z}^\prime)^\prime$ and the $(2m + 2p) \times (2m + 2p)$ covariance matrix

$$\Sigma_{v^*} = \begin{bmatrix}\Sigma_{(\omega^{*\prime},\gamma^{*\prime})^\prime} & 0_{2m\times p} & 0_{2m\times p} \\ 0_{p\times 2m} & \Sigma_{\beta_y} & 0_{p\times p} \\ 0_{p\times 2m} & 0_{p\times p} & \Sigma_{\beta_z}\end{bmatrix}. \tag{7.26}$$

The likelihood of Y for the modified predictive process is

$$N(Y|H_y^* v^*, \tau_y^2 I_n + \tilde{\epsilon}_y I_n), H_y^* = [F_y(\phi), \begin{bmatrix}X & 0_{n\times p}\end{bmatrix}]. \tag{7.27}$$

Here, $F_y(\phi)$ consists of the first *n* rows of matrix $F(\phi)$. The GLM likelihood model of Z can be defined by

$$\eta_z = H_z^* v^*, H_z^* = [F_z(\phi), \begin{bmatrix} 0_{n\times p} & X \end{bmatrix}], \tag{7.28}$$

Similarly, $F_z(\phi)$ consists of the last *n* rows of matrix $F(\phi)$.

7.3.3 Predictive model for two non-Gaussian variable

Replacing $\omega(s), \gamma(s)$ in (7.14-7.15) with $\tilde{\omega}, \tilde{\gamma}$, we obtain the predictive model,

$$\eta_y(s) \equiv g_y(E(Y(s))) = x(s)^T\beta_y + \tilde{\omega}(s) \tag{7.29}$$

$$\eta_z(s) \equiv g_z(E(Z(s))) = x(s)^T\beta_z + \tilde{\gamma}(s) \tag{7.30}$$

Using (7.29-7.30) as the likelihood, we obtain the predictive process counterpart of (7.17) as

$$\begin{aligned} &\pi(\phi) \times IW(T|\Psi, m) \times N(\beta_y|\mu_{\beta_y}, \Sigma_{\beta_y}) \times N(\beta_z|\mu_{\beta_z}, \Sigma_{\beta_z}) \times \\ &N\left(\begin{pmatrix}\omega^* \\ \gamma^*\end{pmatrix}|0, \Sigma_{(\omega^{*\prime}, \gamma^{*\prime})^\prime}\right) \times \prod_{i=1}^{n} \pi_y(Y(s_i)|x(s_i)^T\beta_y + \tilde{\omega}(s_i)) \\ &\times \prod_{i=1}^{n} \pi(Z(s_i)|x(s_i)^T\beta_z + \tilde{\gamma}(s_i)) \end{aligned} \tag{7.31}$$

Since Y and Z are both discrete variables, we assume there does not exist the estimation bias.

Let again $v^* = ((\omega^{*\prime}, \gamma^{*\prime}), (\beta_y^\prime, \beta_z^\prime))^\prime$ be a $(2m+2p)\times 1$ vector. The predictive posterior $\pi(v^*, T, \phi, \tau^2|Y, Z)$ is proportional to

$$\begin{aligned} &\pi(\phi) \times IW(T|\Psi, m) \times N(v^*|\mu_{v^*}, \Sigma_{v^*}) \times \\ &\prod_{i=1}^{n} \pi_y(Y(s_i)|x(s_i)^T\beta_y + f_\omega^\omega(s)\omega^* + f_\omega^\gamma(s)\gamma^*) \times \\ &\prod_{i=1}^{n} \pi_z(Z(s_i)|x(s_i)^T\beta_z + f_\gamma^\omega(s)\omega^* + f_\gamma^\gamma(s)\gamma^*) \end{aligned} \tag{7.32}$$

The GLM likelihood mode of Y and Z can be defined by

$$\begin{aligned} \eta_y = H_y^* v^*, H_y^* = [F(\phi_y), \begin{bmatrix} X & 0_{n\times p} \end{bmatrix}] \\ \eta_z = H_z^* v^*, H_z^* = [F(\phi_z), \begin{bmatrix} 0_{n\times p} & X \end{bmatrix}] \end{aligned} \tag{7.33}$$

$F(\phi) = (\mathcal{R}(\phi)' \otimes T)\Sigma^{-1}_{(\omega^{*\prime},\gamma^{*\prime})'}$, where $\mathcal{R}(\phi)'$ is an $n \times n^*$ matrix whose i^{th} row is given by $r(s_i;\phi)'$, for $i = 1, \cdots, n$. Here, $F(\phi_y)$ consists of the first *n* rows of matrix $F(\phi)$ and $F(\phi_z)$ the last *n* rows.

7.4 Approximate Bayesian Inference

Since the likelihood model of the spatial multivariate observations is non-Gaussian, it makes the predictive process no longer analytical available. To address this issue, we can formalize the multivariate predictive process by applying approximate Bayesian inference methods.

7.4.1 Gaussian approximation to the posterior distribution of υ^*

First, we need to approximate $\pi(\upsilon^*|Y, Z, \theta)$. For the predictive process model, the covariance parameters would be $\theta = (T, \phi, \tau^2)$. The simplest approximation to $\pi(\upsilon^*|Y, Z, \theta)$ is the Gaussian approximation. We have

$$\pi(\upsilon^*|Y, Z, \theta) \propto \pi(Y, Z|\upsilon^*, \theta)\pi(\upsilon^*|\theta) \propto \pi(Y|\upsilon^*, \theta)\pi(Z|\upsilon^*, \theta)\pi(\upsilon^*|\theta), \tag{7.34}$$

where $\pi(Y, Z|\upsilon^*, \theta) = \pi(Y|\upsilon^*, \theta)\pi(Z|\upsilon^*, \theta)$ is derived based on the D-separation rules in the graphic model theory (see Fig. 7.1). As discussed in Section (3.2), $\pi(Y|\upsilon^*, \theta)$ follows a Gaussian distribution, but $\pi(Z|\upsilon^*, \theta)$ does not. We therefore need to conduct a Gaussian approximation on $\pi(Z|\upsilon^*, \theta)$, and then on $\pi(Y, Z|\upsilon^*, \theta)$. Under the Gaussian distribution assumption, $N(Y|H_y^*\upsilon^*, \tilde{\epsilon}_y I_n + \tau^2 I_n)$, and we have the prior $\upsilon^* \sim N(\mu^*, \Sigma^*)$. The full conditional distribution of υ^* conditional to $\{Y, \theta\}$ is thus

$$\begin{aligned} &\pi(\upsilon^*|Y, \theta) \propto N(Y|H_y^*\upsilon^*, U)N(\mu_v^*, \Sigma_v^*) \\ &\propto exp\{[-\frac{1}{2}(Y - H_y^*\upsilon^*)'U^{-1}(Y - H_y^*\upsilon^*)] - \frac{1}{2}(\upsilon^* - \mu_v^*)'\Sigma_v^{*-1}(\upsilon^* - \mu_v^*)\} \\ &\propto exp(-\frac{1}{2}\upsilon^{*'}Q_y\upsilon^* + \upsilon^{*'}b_y), \end{aligned} \tag{7.35}$$

where $U = \tilde{\epsilon}_y I_n + \tau^2 I_n$, the full conditional precision matrix $Q_y = H_y^{*'}U^{-1}H_y^* + \Sigma_v^{*-1}$, and the canonical parameter $b_y = H_y^{*'}U^{-1}Y + \Sigma_v^{*-1}\mu_v^*$.

The likelihood model of Z is non-Gaussian, so we need to expand the likelihood in a quadratic form utilizing the Gaussian approximation. The GLM likelihood of Z is $\prod_i \pi(Z(s_i)|\eta_z(s_i))$, where the GLM parameter $\eta_z = H_z^*\upsilon^* = [F(\phi_z), [0_{n\times p}X]]\upsilon^*$.

Based on the discussion in Section 7.2.1, the distributions in a natural exponential family have the form

$$\pi(Z|\eta_z) = exp\{\eta_z Z - f(\eta_z)\}h(Z). \tag{7.36}$$

For example, for binomial distribution, $Binomial(1,\pi)$, $\eta_z = \log(\frac{\pi}{1-\pi})$, $f(\eta_z) = \log(1+exp(\eta_z))$, and $h(Z) = 1$. In the Poisson case, $Poisson(\lambda)$, $\eta_z = \log(\lambda)$, $f(\eta_z) = exp(\eta_z)$, and $h(Z) = \frac{1}{Z!}$.

By performing a Taylor expansion of $f(\eta_z) = f(H_z^* v^*)$ to the second order, we obtain the quadratic form of v^*,

$$\begin{aligned} \pi(Z|\eta_z) &\propto exp\{-\frac{1}{2}v^{*'}Q_z v^* + v^{*'}b_z\}, \\ Q_z = H_z^{*'}\nabla^2 f(H_z^*\hat{v}^*)H_z^*, b_z &= H_z^{*'}(Z - \nabla f(H_z^*\hat{v}^*) + \nabla^2 f(H_z^*\hat{v}^*)H_z^*\hat{v}^*). \end{aligned} \tag{7.37}$$

Combining Eqs. (7.34), (7.35) and (7.37) gives

$$\pi(v^*|Y,Z,\theta) \propto exp[-\frac{1}{2}v^{*'}(Q_y+Q_z)v^* + v^{*'}(b_y+b_z)]. \tag{7.38}$$

Finally, the full conditional precision matrix $Q = Q_y + Q_z$, and the canonical parameter $b = b_y + b_z$. Thus, the full conditional distribution is $\pi(v^*|Y,Z,\theta) \sim N(Q^{-1}b, Q^{-1})$. We can compute the required inverse and determinant of the size $(2m+2p)\times(2m+2p)$ matrix Q by utilizing the structure of H_z^*, H_y^*, and Σ_v^*. The main cost of the matrix inversion is thus $O(m^3)$, since the number of knots is m and assuming $m \gg p$. The supplementary material provides further details of the Taylor expansion for the Binomial and Poisson distributions.

Taylor expand for Binomial and Poisson distribution

With count data which follows the Poisson distribution,

$$\pi(Z_c(s_i)|\eta_{z-c}(s_i)) \propto exp(Z_c(s_i)\eta_{z-c}(s_i) - exp(\eta_{z-c}(s_i)) \tag{7.39}$$

We operate a Taylor expand $exp(\eta_{z-c}(s_i))$ to second order. By expressing the result in a quadratic form of v^*, we obtain,

$$\begin{aligned} &log(\pi(Z_c|\eta_{z-c})) = -\frac{1}{2}v^{*\prime}Q_{z-c}v^* + v^{*\prime}b_{z-c} + const \\ &Q_{z-c} = H_z^{*\prime}D_{z-c}H_z^*, b_z = H_z^{*\prime}(Z_c - d_{z-c} + D_{z-c}H_z^*\hat{v}^*) \\ &D_{z-c} = diag(exp(H_z^*\hat{v}^*)), d_{z-c} = D_{z-c} * 1_n \end{aligned} \tag{7.40}$$

where 1_n is an $n \times 1$ vector of ones.

Geo-statistic usually assumes the binary data follows the Binomial distribution,

$$\pi(Z_b(s_i)|\eta_{z-b}(s_i)) \propto exp(Z_b(s_i)\eta_{z-b}(s_i) - n(s_i)log(1 + exp(\eta_{z-b}(s_i)))) \tag{7.41}$$

where $n(s_i)$ is the fixed number of trials. Taylor expansion is utilized to express υ^* as the quadratic form [136].

$$\begin{aligned} &log(\pi(Z_b|\eta_{z-b})) = -\frac{1}{2}\upsilon^{*\prime}Q_{z-b}\upsilon^* + \upsilon^{*\prime}b_{z-b} + const \\ &Q_{z-b} = H_z^{*\prime}D_{z-b}H_z^*, b_{z-b} = H_z^{*\prime}(Z_b - d_{z-b} + D_{z-b}H_z^*\hat{\upsilon}^*) \\ &d_{z-b} = \{n \odot exp(H_z^*\upsilon^*)\} \oslash \{1_n + exp(H_z^*\upsilon^*)\} \\ &D_{z-b} = diag(\{n \odot exp(H_z^*\upsilon^*)\} \oslash \{(1_n + exp(H_z^*\upsilon^*))^{\circledast 2}\}) \end{aligned} \tag{7.42}$$

where $n = ((n(s_1), \cdots, n(s_n)))'$.

Approximating $\pi(\upsilon^*|Z_b, Z_c, \theta)$ for Binomial+Gaussian response

With Eqs.(7.40) and (7.42), we can obtain

$$\pi(Z_b, Z_c|\upsilon^*, \theta) \propto exp[-\frac{1}{2}\upsilon^{*'}(Q_{z-b} + Q_{z-c})\upsilon^* + \upsilon^{*'}(b_{z-b} + b_{z-c})] \tag{7.43}$$

Finally, the full conditional precision matrix $Q = Q_{z-b} + Q_{z-c}$, and the canonical parameter $b = b_{z-b} + b_{z-c}$. Thus, the full conditional distribution is $\pi(\upsilon^*|Z_b, Z_c, \theta) \sim N(Q^{-1}b, Q^{-1})$.

7.4.2 Laplace approximation to posterior distribution of θ

Unlike $\pi(\upsilon^*|Y, Z)$, the posterior $\pi(\theta|Y, Z)$ is usually highly skewed and its approximation as a Gaussian distribution is thus inappropriate [136]. The posterior $\pi(\theta|Y, Z)$ plays an important role in the inference of the marginal posterior of latent variables. Taking υ^* as an example, we can estimate the marginal posterior $\pi(\upsilon^*|Y, Z)$, which takes the form of

$$\pi(\upsilon^*|Y, Z) = \int \pi(\upsilon^*|Y, Z, \theta)\pi(\theta|Y, Z)d\theta. \tag{7.44}$$

It is possible to obtain a sample set $\{\theta_1, \cdots, \theta_K\}$ from the input space of θ that represents an approximate discrete form of the posterior $p(\theta|Y, Z)$. We can estimate the approximate $\hat{p}(\upsilon^*|Y, Z)$ by

$$\hat{\pi}(\upsilon^*|Y, Z) = \sum_{k=1}^{K} \pi(\upsilon^*|Y, Z, \theta_k)\pi(\theta_k|Y, Z)w_{\theta_k}, \tag{7.45}$$

where w_{θ_k} is the weight of the sample point θ_k that can be measured by its normalized probability density. The critical step is to efficiently identify a representative sample set $\{\theta_1, \cdots, \theta_K\}$, as well as the corresponding set of weights $\{w_{\theta_1}, \cdots, w_{\theta_K}\}$.

The posterior $\pi(\theta^*|Y, Z)$ can be re-formalized as

$$\pi(\theta|Y, Z) \propto \frac{\pi(Y, Z|\upsilon^*, \theta)\pi(\upsilon^*|\theta)\pi(\theta)}{\pi(\upsilon^*|Y, Z, \theta)}. \tag{7.46}$$

The Laplace Approximation (LA) can be applied to approximate the denominator $\pi(\upsilon^*|Y, Z, \theta)$ as a Gaussian distribution. The LA method uses a similar approach to that for Bayesian spatial inference:

$$\hat{\pi}(\theta|Y, Z) \propto \frac{\pi(Y, Z|\upsilon^*, \theta)\pi(\upsilon^*|\theta)\pi(\theta)}{\hat{\pi}(\upsilon^*|Y, Z, \theta)}\bigg|_{\upsilon^*=\hat{\upsilon}^*}, \tag{7.47}$$

where $\hat{\pi}(\upsilon^*|Y, Z, \theta)$ is a Gaussian approximation of

$$\hat{\pi}(\upsilon^*|Y, Z, \theta) \propto \hat{\pi}(Y, Z|\upsilon^*, \theta)\hat{\pi}(\upsilon^*|\theta). \tag{7.48}$$

Utilizing the above approximation yields the mode $\hat{\upsilon}^*$ and the curvature at the mode of this full conditional expression. In our framework, we apply the generalized linear model (GLM) to capture the distributions of non-Gaussian variables. The preceding Gaussian approximation can be efficiently conducted using the popular Iterated Re-weighted Least Squares (IRLS) algorithm. The detailed procedures for this are summarized in Algorithm 1.

Algorithm 1 iterates l_1 times from Step 2 to Step 8 until convergence. Among these steps, Step 6 has the highest time cost. Because the solution is analytically intractable, numerical optimization techniques need to be applied. An efficient IRLS algorithm is proposed to conduct this process. For the purpose of illustration, suppose observations Z are count data and follow a Poisson distribution (See Eq.(7.39)). Step 6 is first reformulated as the following optimization problem

$$\underset{\upsilon^*}{\operatorname{argmax}}\, \hat{\pi}(\upsilon^*|Y, Z, \theta) = \underset{\upsilon^*}{\operatorname{argmin}}\, -\ln \pi(Y|\upsilon^*, \theta) - \ln \pi(Z|\upsilon^*, \theta) - \ln \pi(\upsilon^*|\theta). \tag{7.49}$$

Algorithm 11 Exploring the posterior distribution of $\pi(\theta|Y,Z)$

Input: S, S^*, S^0, Y, Z, X
Output: Θ, w

1: Choose an initial value $\theta = \{\tau^2, T, \phi\}$;
2: **repeat**
3: Construct μ_{v^*}, Σ_{v^*} with θ (See Eq. (7.26)).
4: Calculate the transformation matrix $F(\phi)$ (See Eq. (7.24)).
5: Calculate the likelihood of Y for Gaussian variables (See Eq.(7.27)) and GLM likelihood of Z for exponential ones (See Eq.(7.28)).
6: Apply IRLS to find the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y,Z,\theta)$, then make a Gaussian approximation by applying Eq.(7.38).
7: Compute the gradient and Hessian of $\hat{\pi}(\theta^*|Y,Z)$ and apply one Newton's step to update θ.
8: **until** Convergence
9: Explore the contour of $\hat{\pi}(\theta|Y,Z)$ based on its mode and Hessian at the mode, obtain K sample locations, $\Theta = \{\theta_1, ..., \theta_K\}$.
10: Compute and normalize $\{\hat{\pi}(\theta_1|Y,Z), \cdots, \hat{\pi}(\theta_K|Y,Z)\}$ to obtain the set of weights $w = \{w_{\theta_1}, \cdots, w_{\theta_K}\}$ as $w_{\theta_k} = \frac{\hat{\pi}_k(\theta_k|Y,Z)}{\sum_{k=1}^K \hat{\pi}_k(\theta_k|Y,Z)}$.

Expanding the density functions $\pi(Y|v^*,\theta)$, $\pi(Z|v^*,\theta)$, and $\pi(v^*|\theta)$, we have

$$\begin{aligned}\underset{v^*}{\operatorname{argmin}} \quad & \frac{1}{2}\left(Y - H_y^* v^*\right)^T U^{-1}\left(Y - H_y^* v^*\right) - \left(Z^T H_z^* v^* - \mathbf{1}^T \exp(H_z^* v^*)\right) \\ & + \frac{1}{2}\left(v^* - \mu_{v^*}\right)^T \Sigma_{v^*}^{-1}(v^* - \mu_{v^*}) \,.\end{aligned} \tag{7.50}$$

The gradient and Hessian matrix of the above objective function can be obtained as

$$\begin{aligned}\bigtriangledown \hat{\pi}(v^*|Y,Z,\theta) &= \left(H_y^{*T} U^{-1} H_y^* + \Sigma_{v^*}^{-1}\right) v^* + H_z^{*T} \exp(H_z^* v^*) \\ &\quad - H_y^{*T} U^{-1} Y - H_z^{*T} Z - \Sigma_{v^*}^{-1} \mu_{v^*}, \end{aligned} \tag{7.51}$$

$$\bigtriangledown^2 \hat{\pi}(v^*|Y,Z,\theta) = H_y^{*T} U^{-1} H_y^* + \Sigma_{v^*}^{-1} + H_z^{*T} diag(\exp(H_z^* v^*)) H_z^*, \tag{7.52}$$

where $diag(\exp(H_z^* v^*))$ reshapes the vector $\exp(H_z^* v^*)$ as a diagonal matrix.

The IRLS algorithm for Step 6 is described as follows:

1. Select an initial $\hat{v}^*$
2. Until convergence
   (a) Update $\hat{v}^* = \hat{v}^* - \left(\bigtriangledown^2 \hat{\pi}(\hat{v}^*|Y,Z,\theta)\right)^{-1} \bigtriangledown \hat{\pi}(\hat{v}^*|Y,Z,\theta)$ by Eqs. (7.51) and (7.52)
3. Output $\hat{v}^*$ as the mode of $\hat{\pi}(v^*|Y,Z,\theta)$.

Computational Complexity. In Algorithm 1, suppose that l_2 iterations are required to find the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y,Z,\theta)$, and the time cost of Step 6 is $O(l_2 * (n * m^2 + m^3))$. For Step 5, the Gaussian approximation of $\hat{\pi}(v^*|Y,Z,\theta)$ takes $O(n*m)$. Overall, Steps 2-8, which generate the converged gradient and Hessian of $\pi(\theta|v^*)$, take $O(l_1 * l_2 * (n * m^2 + m^3) + l_1 * n * m)$. Finally, sampling the θ set and computing their corresponding weighted values take $O(K)$. The overall framework is designed based on Newton's method, whose convergence is generally rapid. The performance on problems in $\mathcal{R}^{10000}$ is thus similar to that on problems in $\mathcal{R}^{10}$, and the required number of Newton's steps (l_1) only increases modestly [19]. Step 6 applies IRLS to capture the mode of $\hat{\pi}(v^*|Y,Z,\theta)$, and in practice five iterations ($l_2 = 5$) are sufficient. In summary, assuming $m \gg K$, $m \gg l_1$ and $m \gg l_2$, the total computational complexity of parameter estimation is $O(n * m^2)$.

7.4.3 Spatial prediction via Laplace Approximation

Given a set of unsampled locations $\{s_1^0, \cdots, s_{N_{te}}^0\}$, we are interested in predicting the Y and Z attribute values at these locations, denoted as $Y^0 = (Y(s_1^0), \cdots, Y(s_{N_{te}}^0))'$ and $Z^0 = (Z(s_1^0), \cdots, Z(s_{N_{te}}^0))'$. The first step is to estimate the posterior distributions of the corresponding latent variables $\pi(\omega^0|Y,Z)$ and $\pi(\gamma^0|Y,Z)$, where $\omega^0 = (\omega(s_1^0), \cdots, \omega(s_{N_{te}}^0))'$ and $\gamma^0 = (\gamma(s_1^0), \cdots, \gamma(s_{N_{te}}^0))'$. Then the posterior distributions of Y^0 and Z^0 can be obtained as

$$\pi(Y^0|Y,Z) = \int \pi(Y^0|\omega^0)\pi(\omega^0|Y,Z)d\omega^0, \tag{7.53}$$

$$\pi(Z^0|Y,Z) = \int \pi(Z^0|\gamma^0)\pi(\gamma^0|Y,Z)d\gamma^0. \tag{7.54}$$

We denote $v^0 = (\omega^{0\prime}, \gamma^{0\prime})'$. Given the approximated $\hat{\pi}(v^*|Y,Z,\theta)$ and $\hat{\pi}(\theta|Y,Z)$ as obtained in Sections 4.1 and 4.2, the posterior distribution $\pi(v^0|Y,Z)$ can be estimated by

$$\begin{aligned}
\pi(v^0|Y,Z) &= \int\int \pi(v^0|v^*,Y,Z,\theta)\pi(v^*|Y,Z,\theta)\pi(\theta|Y,Z)dv^*d\theta \\
&= \int \left(\int \pi(v^0|v^*,\theta)\pi(v^*|Y,Z,\theta)dv^* \right) \pi(\theta|Y,Z)d\theta \\
&\approx \sum_{k=1}^{K} \left(\int \pi(v^0|v^*,\theta_k)\hat{\pi}(v^*|Y,Z,\theta_k)dv^* \right) \times \hat{\pi}(\theta_k|Y,Z)w_{\theta_k} \\
&\approx \sum_{k=1}^{K} \mathcal{N}(\tilde{\mu},\tilde{\Sigma})\hat{\pi}(\theta_k|Y,Z)w_{\theta_k},
\end{aligned} \tag{7.55}$$

where

$$\begin{aligned}
&\Sigma^0 = Cov(v^0), \Sigma^* = Cov(v^*), \Sigma^{0*} = Cov(v^0, v^*),\\
&\tilde{\mu} = \Sigma^{0*}\Sigma^{*-1}Q^{-1}b,\\
&\tilde{\Sigma} = \Sigma^0 - \Sigma^{0*}\Sigma^{*-1}\Sigma^{0*^T} + \Sigma^{0*}\Sigma^{*-1}Q^{-1}\Sigma^{*-1}\Sigma^{0*^T}.
\end{aligned} \tag{7.56}$$

Note that, an alternative approximation strategy is to reformulate $\pi(v^0|Y, Z, \theta)$ as

$$\pi(v^0|Y, Z) = \int \pi(v^0|Y, Z, \theta)\pi(\theta|Y, Z)d\theta. \tag{7.57}$$

Given that

$$\pi(v^0|Y, Z, \theta) \propto \frac{\pi(Y, Z|v^0, v^*)\pi(v^0, v^*|\theta)}{\pi(v^*|v^0, Y, Z, \theta)}, \tag{7.58}$$

we apply Laplace approximation in Section 4.2 to obtain the approximate marginal as

$$\tilde{\pi}(v^0|Y, Z, \theta) \approx \frac{\pi(Y, Z|v^0, v^*)\pi(v^0, v^*|\theta)}{\hat{\pi}(v^*|v^0, Y, Z, \theta)}\bigg|_{v^*=\hat{v}^*}, \tag{7.59}$$

where $\hat{\pi}(v^*|v^0, Y, Z, \theta)$ is a Gaussian approximation of

$$\pi(v^*|v^0, Y, Z, \theta) \propto \pi(Y, Z|v^*, v^0)\pi(v^*, v^0|\theta). \tag{7.60}$$

The posterior distribution $\pi(v^0|Y, Z, \theta)$ can be estimated by

$$\pi(v^0|Y, Z) \approx \sum_{k=1}^{K} \hat{\pi}(v^0|Y, Z, \theta_k)\hat{\pi}(\theta_k|Y, Z)w_{\theta_k}. \tag{7.61}$$

In practice, the above two alternative strategies provide similar accuracies for the predictions of Y^0 and Z^0, but the former strategy has a lower computational cost, so this was chosen as the default implementation. However, studies have shown that the latter strategy works better for the prediction of the regression parameters β_Y and β_Z.

Based on the above theoretical analysis, the main procedures involved in predicting multivariate non-Gaussian variables are described by Algorithm 2.

After obtaining the θ samples $\{\theta_1, \cdots, \theta_K\}$ and their weight values $\{w_{\theta_1}, \cdots, w_{\theta_K}\}$ from Algorithm 1, we first use each $\theta_k (k = 1, \cdots, K)$, to construct the mean and covariance matrix of latent variables, v^*. Next,

Algorithm 12 Spatial Multivariate Non-Gaussian Prediction

Input: $S, S^*, S^0, Y, Z, X, X^0, \Theta, w$
Output: Y^0, Z^0

1: **for** $k = 1$ **to** K **do**
2: Construct μ_{v^*}, Σ_{v^*} with θ_k and S^* (See Eq. (7.26)).
3: Calculate the transformation matrix $F(\phi)$ with θ_k, S^*, S, X (See Eq.(7.24)).
4: Calculate the likelihood of Y for Gaussian variables (See Eq.(7.27)) and the GLM likelihood of Z for exponential ones (See Equation (7.28)).
5: Calculate the mode, the Hessian at the mode of $\hat{\pi}(v^*|Y, Z, \theta_k)$, and its Gaussian approximation (See Equation (7.38)).
6: Predict Y_k^0, Z_k^0 for new locations S^0. (See Eqs.(7.53 and 7.54))
7: **end for**
8: Calculate the final Y^0, Z^0 values as $Y^0 = \sum_{k=1}^{K} Y_k^0 \times w_{\theta_k}$, $Z^0 = \sum_{k=1}^{K} Z_k^0 \times w_{\theta_k}$

the transformation matrix $F(\phi)$ is computed, which describes the spatially varying linear transformation of $(\tilde{\omega}', \tilde{\gamma}')'$ on $(\omega^{*\prime}, \gamma^{*\prime})'$. As shown in Eqs. (7.27-7.28), the likelihood of Gaussian observations and the GLM likelihood of exponential ones are defined by $F(\phi)$, μ_{v^*}, Σ_{v^*}, and X. The mode of $\hat{\pi}(v^*|Y, Z, \theta)$ is then calculated to predict the multivariate observations Y_k^0 and Z_k^0 at sample θ_k. Finally, the predicted Y and Z are calculated as $Y^0 = \sum_{k=1}^{K} Y_k^0 \times w_{\theta_k}$, $Z^0 = \sum_{k=1}^{K} Z_k^0 \times w_{\theta_k}$.

Computational complexity. Step 6 dominates the computational costs because it is analytical intractable. With the numerical optimization discussed in Section 4.1-4.2, it takes $O(n * m)$ to operate a Gaussian approximation of $\hat{\pi}(v^*|Y, Z, \theta_k)$ for each sample θ_k. Computing the mode and Hessian of $\hat{\pi}(v^*|Y, Z, \theta_k)$ costs $O(l_2 * (m^3 + n * m^2))$. Repeating Steps 1-7 for K sample θs therefore takes $O(K * (n * m + l_2 * (m^3 + n * m^2)))$. In summary, the total computational complexity of the Spatial Multivariate Non-Gaussian Prediction algorithm is $O(n * m^2)$, assuming $m \gg K$ and $m \gg l_2$.

7.5 Experimental Result and Analysis

This section evaluates the effectiveness and efficiency of our proposed framework based on experiments on simulations and four real life datasets. Because of the space limit, we focused on three bivariate scenarios: 1) the response variables include one Gaussian and one Binomial; 2) the response variables include one Gaussian and one Poisson; 3) the response variables include one Binomial and one Poisson. The datasets and the implementation of our approach can be downloaded from [98] for evaluations. All the experiments were conducted on a PC with Intel(R) Core(TM) I5-2400, CPU 3.1Ghz, and 8.00 GB memory. The development tool was MATLAB 2011.

7.5.1 Simulation study

Simulation settings

Data set: We used a similar simulation model as that used in [28]. Taking the Gaussian+Binomial(G+B) simulation as an example, we considered $Y(s)$ as a Gaussian random variable at location s, and $Z(s)$ as a binomial random variable. The simulation data were generated based on the following statistical model:

$$\begin{aligned} Y(s) &\sim \mathcal{N}(x^T\beta_y + \omega(s), \tau^2), \\ Z(s) &\sim Binomial(m, g_b(x^T\beta_z + \gamma(s))), \end{aligned} \tag{7.62}$$

where $g_b(\cdot)$ is the link function of the generalized linear model for binomial distribution and $\omega(s)$ and $\gamma(s)$ are defined together as a latent spatial co-kriging Gaussian process [54] with the covariogram model

$$Var([\omega(s);\gamma(s)],[\omega(o);\gamma(o)]) = \hat{\Sigma}_{tn} = \begin{vmatrix} \sigma_y^2 & \sigma_{yz}^2 \\ \sigma_{yz}^2 & \sigma_z^2 \end{vmatrix} C(h|\phi),$$

and $h = |s - o|$. The correlation component $C(h|\phi)$ refers to the spatial correlation between two random variables. We used the popular exponential kernel function to model the correlation, which has the form $C(h|\phi) = -exp\left(\frac{h}{\phi}\right)$, where ϕ is the range parameter that controls the degree of spatial autocorrelations. The model design indicates that the spatial dependency between the variables $Y(s)$ and $Z(s)$ is realized through their corresponding latent variables.

The simulations for count data were generated based on a Poisson distribution

$$Z(s) \sim Poisson(g_p(x^T\beta_z + \gamma(s))), \tag{7.63}$$

where $g_p(\cdot)$ is the link function of the generalized linear model for Poisson distribution.

In Gaussian+Poisson(G+P), $Y(s)$ and $Z(s)$ were considered as Gaussian and Poisson random variables, respectively. In Binomial+Poisson(B+P), $Y(s)$ and $Z(s)$ were Binomial and Poisson random variables, respectively. The parameter settings used in our experiments are shown in Table 2. We also tested different combinations of parameters, and observed similar patterns. Fig. 7.2 depicts density maps of the numerical(Y) and binary(Z) responses from a typical G+B simulation, which shows the complicated distributions involved and illustrates why a higher processing ability is required for the predictive models.

Seven State of the Art Competing Methods: Our literature survey revealed only two methods for predicting multivariate non-Gaussian spatial data. One is the BME method proposed by Wibrin et al., which supports the mixture of one numerical and one categorical variables[163], and the other is MCMC designed

Table 7.2: Parameter settings in simulations

Variable	Setting Description
Data Type	Gaussian(Y)+Binomial(Z), Gaussian(Y)+Poisson(Z), Binomial(Y)+Poisson(Z)
N_{tr}, N_{te}	$N_{tr} = 1000, N_{te} = 400, 500$. Training data were randomly generated at N_{tr} spatial locations $\{s_i\}_{i=1}^{N_{tr}}$ for the range [0,50]×[0,50] units. Testing data were generated at N_{te} spatial locations $\{s_i\}_{i=1}^{N_{te}}$ over the same range.
β_y, β_z	The regression coefficient $\beta_y = [2, 2]', \beta_z = [2, 1]'$ in G+P; $\beta_y = [0.5, 0.5]', \beta_z = [0.1, 0.1]'$ in G+B; $\beta_y = [0.1, 0.1]', \beta_z = [2, 1]'$ in B+G.
$\sigma_y, \sigma_z, \sigma_{yz}$	$\sigma_y^2 = 4, \sigma_z^2 = 3.24, \sigma_{yz}^2 = 2.52$ in all types of simulations.
ϕ	$\phi = 25$ in all types of simulations.
τ	The nugget variance, τ^2, was set to 1 in both G+B and G+P simulations.
Correlation model	An exponential spatial correlation function $C(h, \phi) = \sigma^2 exp(-\frac{h}{\phi})$ was used in all types of simulations.

by Chagneau et al., which is based on the Gibbs sampler with Metropolis-Hastings (M-H) steps [28]. As BME is restricted to bivariate data with one Gaussian and one Categorical variables, and MCMC is flexible for a variety of mixture types. We implemented MCMC using the same framework (Gibbs sampler with M-H steps), and denoted it as Spa-Multi-MCMC.

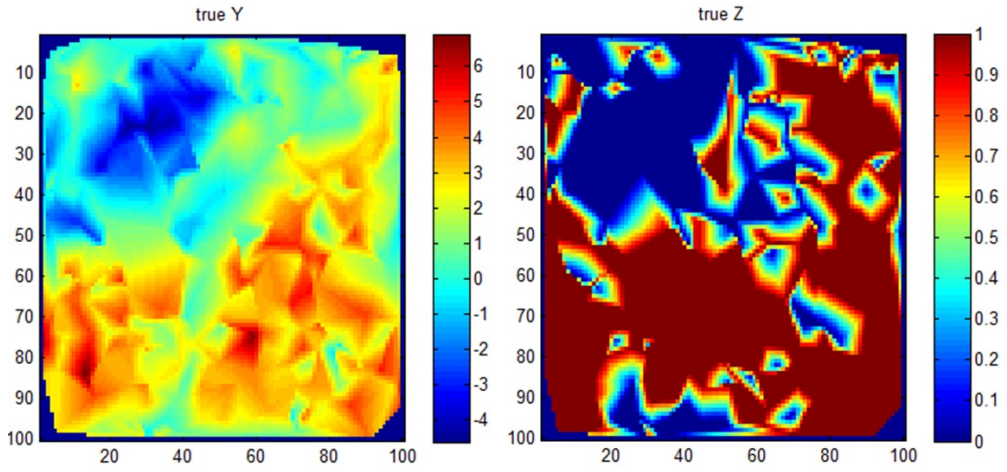


Figure 7.2: Density maps of a typical G+B simulation

There is also an R toolbox function named "MCMCglmm" that supports general multivariate non-Gaussian prediction without considering spatial dependency. We implemented this method using the same MCMC

framework, denoted as Multi-MCMC. In practice, Spa-Multi-MCMC and Multi-MCMC do not scale well to large datasets (e.g., 1,000 points), since they need to sample a large set of latent variables, ω and γ. Therefore, we implemented two approximate versions of them, namely, Spa-Multi-MCMC-K and Multi-MCMC-K. The basic idea is to split the full conditional distribution $\pi(\omega(s_1), \cdots, \omega(s_n))$ into the product of block conditional distributions, where each block includes a set of neighbor points. The blocks are obtained by using K-Means clustering algorithm. Using this strategy, we reduced the dimensionality of latent ω and γ, from n to the block size. The model proposed in this paper is identified as Spa-Multi-INLA.

There are several techniques that can be used for non-spatial predictive modeling, including CART (Classification and Regression Trees) [20], MARS (Multivariate Adaptive Regression Splines) [51], and Treenet (also known as MART, Multiple Additive Regression Trees) [52]. These three methods can all determine the predictors from a large number of variables and estimate their interactions, which helps accurately predict the outcome variables. All are flexible regarding classification and regression modeling of high dimensional data. However, none take spatial dependency into consideration since they are focused on general multivariate modeling. These three techniques have been implemented into a flexible and powerful data mining tool, known as the Salford System [152], which was used to make predictions for Y and Z separately in our experiment. We identify them as CART, MARS, and Treenet.

Performance metric: We ran the experiments with 20 realizations of each parameter combination and then calculated the mean and standard deviation of the approximated value of every parameter combination in the multivariate process model. For each observation, we computed the MAE (Mean Absolute Error) for numerical and count observations, and the accuracy for binary ones based on their corresponding predicted and true values. To validate the new model's effectiveness and efficiency, we compared the results of estimations, predictions, and response times for the Spa-Multi-INLA, MCMC based approaches, CART, MARS, and Treenet. Finally, we utilized Moran's I-statistic to capture the spatial dependency of numerical observations.

Simulation results

Model Parameter Estimates

Tables 3-5 show the estimation results for the model parameters on the data sets of size 1000 for the G+B, G+P, and B+P simulations, respectively. "Spa-Multi-INLA(64)" refers to our approach with a knot size equal to 64. No results are shown for Multi-MCMC and Spa-Multi-MCMC because they became very slow when the data size exceeded 1000. The iterations of Spa-Multi-MCMC-K and Multi-MCMC-K were set to 3000 iterations, and K equal to 170 blocks (clusters). Also, there are no results for CART, MARS and Treenet in these tables. This is because these models do not include these parameters. Instead, these ran on the Salford tool, which is a well-developed and optimized powerful tool and it was not considered reasonable

Table 7.3: Comparisons of the parameter estimation and computational cost in G+B.(Spa-Multi-MCMC and Multi-MCMC are unable to process datasets with data sizes greater than 1000)

Para. / Approach	β_y	β_z	ϕ	σ_y^2	σ_z^2	σ_{yz}^2	τ^2	Time(m)
True Values	$\begin{bmatrix}0.50\\0.50\end{bmatrix}$	$\begin{bmatrix}0.10\\0.10\end{bmatrix}$	25	4	3.24	2.52	1.00	—
Spa-Multi-INLA(64)	$\begin{bmatrix}0.60(0.05)\\0.63(0.05)\end{bmatrix}$	$\begin{bmatrix}0.18(0.07)\\0.15(0.07)\end{bmatrix}$	11.44 (2.64)	5.65 (1.86)	3.20 (1.02)	2.47 (0.12)	1.31 (0.06)	1.3
Spa-Multi-INLA(256)	$\begin{bmatrix}0.60(0.03)\\0.63(0.03)\end{bmatrix}$	$\begin{bmatrix}0.21(0.07)\\0.19(0.07)\end{bmatrix}$	10.64 (2.25)	4.73 (1.21)	3.02 (0.86)	2.40 (0.19)	1.12 (0.06)	1.5
Spa-Multi-MCMC-K	$\begin{bmatrix}0.18(6.89)\\-0.13(6.77)\end{bmatrix}$	$\begin{bmatrix}0.22(0.47)\\-0.09(0.43)\end{bmatrix}$	12 (1.05)	22.48 (24.39)	51.41 (58.19)	4578.56 (1054.58)	0.12 (0.15)	171.93
Multi-MCMC-K	$\begin{bmatrix}0.29(7.11)\\0.06(7.29)\end{bmatrix}$	$\begin{bmatrix}-2.84(0.48)\\0.28(1.11)\end{bmatrix}$	$--$	$--$	$--$	$--$	0.13 (0.16)	234.27

to directly compare their running times with those of the LA(Laplace Approximation) and MCMC based approaches both of which ran on Matlab. However, we did compare the prediction performances of Y and Z among all of these approaches and the results are shown in Fig. 7.3.

By comparing the estimated parameters with the true values, we observed that our method was able to accurately estimate most of the model parameters with only small deviations, compared to the other two MCMC based methods for these three types of simulations. The true range parameter ϕ is 25, but both the LA and MCMC based approaches underestimated the range parameter at around 11. This indicates the difficulty of capturing the degree of spatial autocorrelation based on the spatial distance. In addition, we found that it was more difficult to estimate β_z than β_y in G+B and G+P, and β_y than β_z in B+P. This is reasonable since for binary data, it is usually more difficult to estimate the corresponding β than for count and numerical data. Compared with numerical data, count data is also more difficult to model.

Finally, there was an interesting result for MCMC based methods. In some cases, they performed well, as in the B+P simulation, where both Spa-Multi-MCMC-K and Multi-MCMC-K did approximately estimate the β_y and β_z. However, they sometimes failed to make good estimations, as in the G+P and G+B simulations, where all of the estimated parameters deviated substantially from their true values. The reason for this is that MCMC approaches are sampling-based. If there insufficient iterations are provided, their performances will not be stable, as this depends on the sample selections of the latent variables. Executing MCMC approaches with appropriate iterations provided comparable results but were very time consuming, requiring around 1-2 days for our simulations. Since our focus is on not only the effectiveness, but also the efficiency, we

Table 7.4: Comparisons of the parameter estimation and computational cost in G+P

Para / Approach	β_y	β_z	ϕ	σ_y^2	σ_z^2	σ_{yz}^2	τ^2	Time(m)
True Values	$\begin{bmatrix}2.00\\2.00\end{bmatrix}$	$\begin{bmatrix}2.00\\1.00\end{bmatrix}$	25	4	3.24	2.52	1.00	–
Spa-Multi-INLA(64)	$\begin{bmatrix}2.51(0.22)\\2.01(0.02)\end{bmatrix}$	$\begin{bmatrix}1.38(0.17)\\1.03(0.01)\end{bmatrix}$	13.87 (3.29)	4.90 (1.27)	3.02 (0.77)	3.31 (0.67)	1.28 (0.06)	1.83
Spa-Multi-INLA(256)	$\begin{bmatrix}0.92(0.15)\\1.94(0.01)\end{bmatrix}$	$\begin{bmatrix}0.44(0.12)\\0.94(0.01)\end{bmatrix}$	9.42 (2.29)	3.37 (0.94)	2.08 (0.48)	2.03 (0.35)	1.07 (0.05)	3.20
Spa-Multi-MCMC-K	$\begin{bmatrix}-0.07(7.01)\\0.13(6.94)\end{bmatrix}$	$\begin{bmatrix}0.94(0.05)\\0.89(0.10)\end{bmatrix}$	2.97 (1.75)	36.31 (35.36)	2.82 (2.64)	1.84 (0.48)	0.26 (0.29)	72
Multi-MCMC-K	$\begin{bmatrix}0.39(6.95)\\-0.03(7.12)\end{bmatrix}$	$\begin{bmatrix}1.39(0.05)\\0.99(0.07)\end{bmatrix}$	$--$	$--$	$--$	$--$	0.14 (0.15)	27

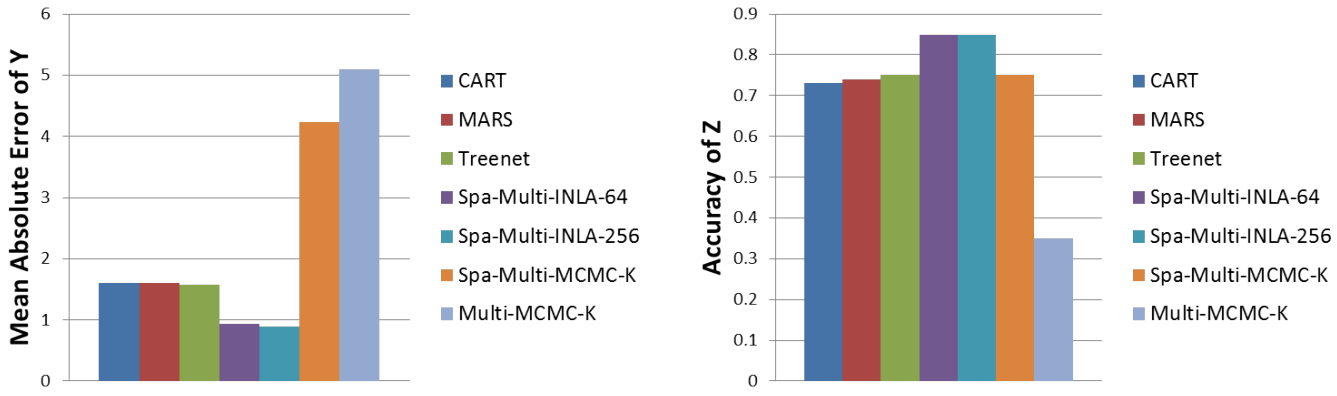
therefore abandoned the attempt to improve the accuracy of the MCMC based approaches.

Prediction Accuracy

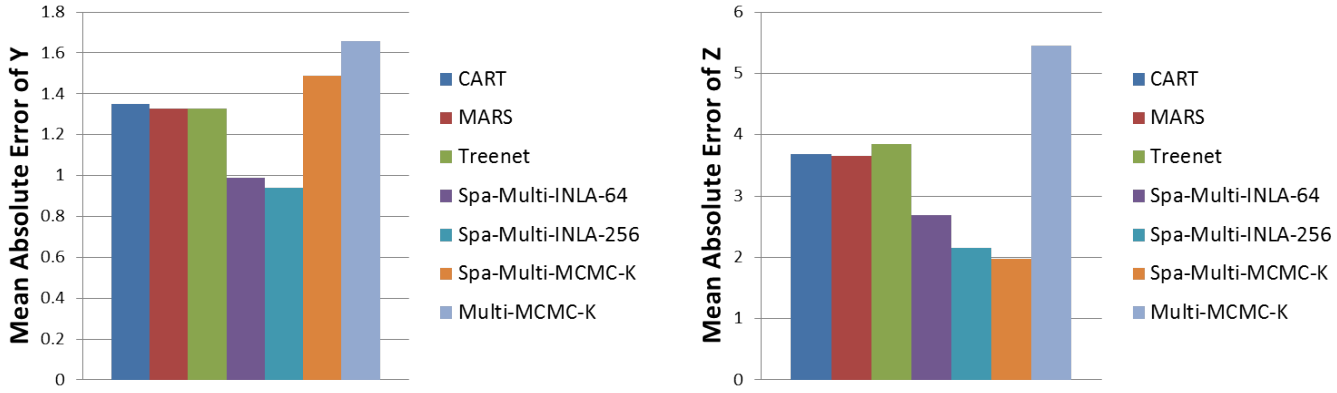
Fig. 7.3 provides the prediction results of different approaches for the G+B, G+P and B+P simulations. Applying Moran's I-statistic, we computed the spatial dependencies for Y numerical attributes in G+B and G+P, as 0.7006 and 0.7222, which indicates that the existing high spatial auto-correlation needs to be considered during the estimation and prediction processes. In fact, when we generated these simulations, to demonstrate the effectiveness and efficiency of our approach, the degree of spatial dependencies among different objective attributes were set with higher values.

As we can see in Fig. 7.3, for the case of G+B, Spa-Multi-INLA(256) has the lowest MAE(0.89) for Y and the highest accuracy (0.85) for Z. By contrast, CART, MARS and Treenet have higher MAEs (1.60, 1.61 and 1.58) and lower accuracies (0.73, 0.74 and 0.75) since they are unable to capture the spatial dependencies. Spa-Multi-MCMC-K has the worse performance (MAE: 4.23, Accuracy: 0.75) at the cost of large computational iterations. Multi-MCMC-K approach failed to accurately execute spatial predictions since it was unable to learn the spatial dependency and operated without sufficient iterations. Spa-Multi-MCMC and Multi-MCMC cannot process mixed type datasets whose data sizes are greater than 1000 because in MCMC based approaches the un-marginalized models used to fit the Binomial+Poisson outcome data required more MCMC iterations.

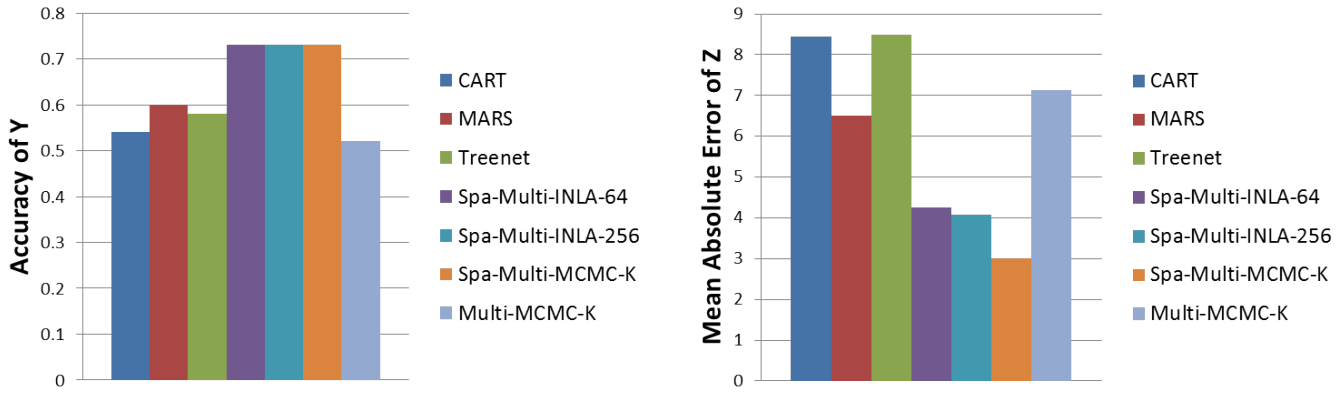
Fig. 7.3(b-c) also provides the prediction comparisons of Spa-Multi-INLA models with 64 and 256 knots



(a) Gaussian+Binomial



(b) Gaussian+Possion



(c) Gaussian+Possion

Figure 7.3: Comparison of the performances for six approaches on simulation datasets

Table 7.5: Comparisons of the parameter estimation and computational cost in B+P

Para / Approach	β_y	β_z	ϕ	σ_y^2	σ_z^2	σ_{yz}^2	Time(m)
True Values	$\begin{bmatrix}0.10\\0.10\end{bmatrix}$	$\begin{bmatrix}2.00\\1.00\end{bmatrix}$	25	4	3.24	2.52	–
Spa-Multi-INLA(64)	$\begin{bmatrix}0.16(0.03)\\0.23(0.03)\end{bmatrix}$	$\begin{bmatrix}2.09(0.01)\\1.04(0.01)\end{bmatrix}$	20.05 (3.58)	2.38 (0.87)	2.89 (0.54)	1.47 (0.03)	0.95
Spa-Multi-INLA(256)	$\begin{bmatrix}0.12(0.03)\\0.35(0.03)\end{bmatrix}$	$\begin{bmatrix}1.97(0.01)\\1.08(0.01)\end{bmatrix}$	20.05 (3.58)	3.56 (1.21)	3.87 (0.88)	2.60 (0.32)	9.17
Spa-Multi-MCMC-K	$\begin{bmatrix}0.17(0.10)\\0.26(0.10)\end{bmatrix}$	$\begin{bmatrix}1.91(0.03)\\1.01(0.03)\end{bmatrix}$	6.37 (0.61)	1.34 (0.06)	1.51 (0.09)	1.04 (0.06)	101
Multi-MCMC-K	$\begin{bmatrix}0.15(0.08)\\0.20(0.08)\end{bmatrix}$	$\begin{bmatrix}3.15(0.02)\\1.36(0.02)\end{bmatrix}$	$--$	$--$	$--$	$--$	95

against other approaches for the G+P and B+P simulations. These exhibit the same estimation patterns as those in G+B. By comparing different knot intensities, we see the predictive process with 64 knots has quite a close performance to that with 256 which indicates that the parameter effects can be accurately estimated with the proper knot selections.

Computational Cost

The last columns in Tables 3-5 show the computing times required to deliver the estimation and prediction results for each simulations. For the MCMC based approaches, the main evaluation cost is matrix inversion at $O((2 * 1000)^3)$. For the Spa-Multi-INLA model, the main cost is $O(2 * n * (2 * m)^2)(m = 64$ or $256)$, which is the cost of building the required inverse and determinant of the size $(2m + 2p) \times (2m + 2p)$ matrix Q as shown in Eq.(30) by assuming $m >> p$. As shown in Tables 3-5, there is a clear reduction in the computational cost when using Spa-Multi-INLA approach. Meanwhile, the predictive process with 64 knots has a similar prediction capability but lower computational burden compared to the 256 knots. Integrating the Laplace approximation into the spatial multivariate predictive model clearly helps achieve sufficiently accurate results in a moderate time.

7.5.2 Real life datasets

We validated our approach using four real datasets, namely, *Lake*, *MLST*, *BEF*, and *House*. All are G+B datasets.

The datasets

Lake was originally published by Varin et al. [157]. It was used to model the trout abundance in Norwegian lakes as a function of lake acidity. The explained attributes include *Intercept*, *X coordinate*, *Y coordinate*, *Product of X* and *Y coordinates*, *X coordinate squared*, and *Y coordinated squared*. *MLST* came from multiple listings containing structural descriptors of houses, their sale prices, and their addresses for Baltimore, Maryland in 1978. Dubin [46] estimated a spatial autocorrelation model that calculated the portion of the price by multiplying the vectors of attributes by their estimated coefficients. The explained attributes used *X coordinate*, *Y coordinate*, *Product of X and Y coordinates*, *X coordinate squared*, and *Y coordinate squared*. *BEF* is a forest inventory dataset from the U.S. Department of Agriculture Forest Service. Variables include species, specific basal area, total tree biomass, inventory plot coordinates, and slope, etc. *BEF* data is included in the spBayes R package [148]. Finley and Banerjee [49] made a detailed analysis of the non-spatial logistic regression to the data. The explained variables are *the slope*, *elevation of the object*, *the tasseled cap brightness*, *greenness*, and *wetness components* from summer 2002. *House* contains information collected for a range of variables for all the block groups in California from the 1990 Census. Specifically, it contains median house value, median income, housing median age and total room, etc. The spatial regression model of *House* was analyzed by Pace and Barry [122]. Explained variables include *Median Income*, *Median Income*^{2}, *Median Income*^{3}, *ln(Median Age)*, *ln(TotalRooms/Population)*, *ln(Bedrooms/Population)*, *ln(Population/Households)* and *ln(Households)*.

Table 6 summarizes the main information for each of these datasets used in our experiment. The spatial dependencies were computed using Moran's I-statistic function, and the results are shown in the last column.

Table 7.6: Settings in the 4 real datasets

Dataset	Size	N_{tr}	N_{te}	Y	Z	Spatial Dependence in Y
BEF	437	337	100	BE basal area	EH basal area	0.1672
Lake	371	271	100	Trout abundance	Lake acid	0.0072
MLST	211	150	61	House price	If located in county	0.1753
House	20,640	$2,000$ $5,000$	200 500	House price	House age	0.2529

Fig. 7.4–7.7 summarizes the comparisons among Spa-Multi-INLA(64), four MCMC based approaches, CART, MARS and Treenet. The data name "Lake.271.100.1" indicates that it is the first realization generated from the original *Lake* data with 271 training data and 100 test data points. By learning their spatial

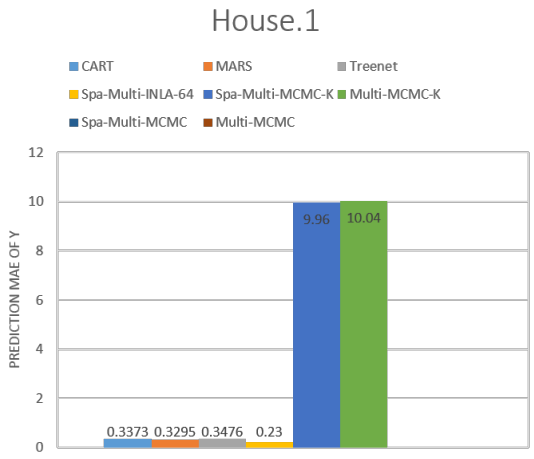
dependencies, we determined that most of the real datasets have lower spatial auto-correlations, which suggests that non-spatial attributes will contribute a lot when predicting the outcome variables.

Experimental results

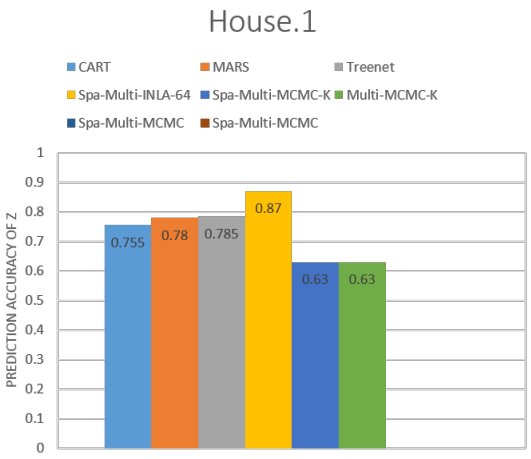
Prediction accuracy

For the predicted Y (numerical observations), the MAEs were computed to demonstrate the prediction performance. Neither Multi-MCMC nor Spa-Multi-MCMC could process the *House* data because of the large data sizes (2000 and 5000 points). Also, the MAE values from Spa-Multi-MCMC-K and Multi-MCMC-K were much higher (around 10) than those (0.21-0.33) of others. To better plot the performance comparisons among Spa-Multi-INLA, CART, MARS and Treenet, we did not draw MCMC based plots for *House* as MCMC based approaches generated such poor results due to the larger datasets (2000 and 5000) which incurred excessive computation times of around 2 days. In our experiments, the iteration values for all the datasets were set to 3000, although this still cost around 1-3.5 hours with Multi-MCMC-K, and 2.5-4.5 hours with Spa-Multi-MCMC-K for *House* datasets. As shown in Fig. 7.4–7.7, Spa-Multi-INLA achieved an average 10 % improvement over CART, MARA and Treenet, 40-50 % over Spa-Multi-MCMC-K and Spa-Multi-MCMC, and 60-70 % over Multi-MCMC-K and Multi-MCMC. For the predicted Z(binary observations), the accuracies were again computed. As shown in Fig. 7.4–7.7, Spa-Multi-INLA achieved average improvements of 10% over CART, MARS and Treenet, 40-50% over Spa-Multi-MCMC-K and Spa-Multi-MCMC, and 60-70% over Multi-MCMC-K and Multi-MCMC. The predicted results from different methods for *House.2000.200.1* are shown as density maps in Fig. 7.9. The visualization indicates that our predictive model was able to accurately capture the distributions of both the Gaussian(Y) and the Binary(Z) variables. One interesting observation is that MCMC based methods tended to under-estimate the Gaussian variable, while over-estimating the Binary variable, and the number of positive values is much greater then the number of true positive values. It is worth noting that CART, MARS and Treenet generated impressive prediction results, and they are much better than MCMC based approaches. This is because the degrees of spatial auto-correlations of the four real datasets are not obvious. The predictions of outcome variables are mainly controlled by the non-spatial predictors, and have less relationship with the spatial distances among the objects. For *Lake*, which has the lowest spatial dependency (0.0072), their performance is close to that of the Laplace approximation based approach. In *MLST* and *BEF*, the spatial dependencies increase a little but are still lower; the performances of CART, MARS, and Treenet are a little worse than that of the Laplace approximation based approach. In *House*, the degree of spatial auto-correlation is more obvious and clearly demonstrates the effectiveness of the Laplace approximation based approach since it takes spatial dependency into consideration during the predictive process.

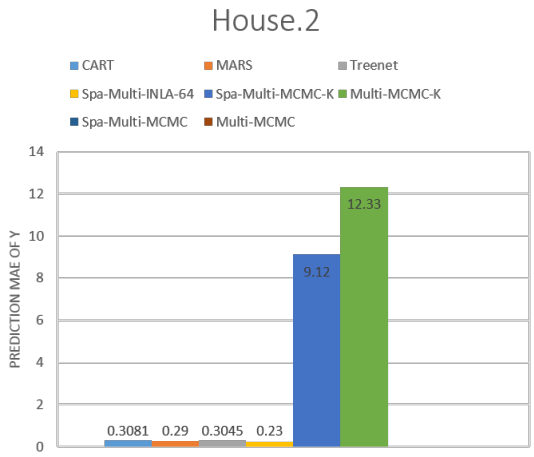
MCMC based approaches have similar estimation patterns to these in simulations and cannot perform well



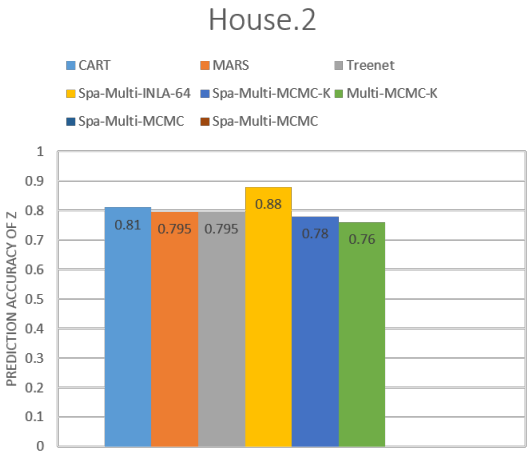
(a) Prediction MAE for Y



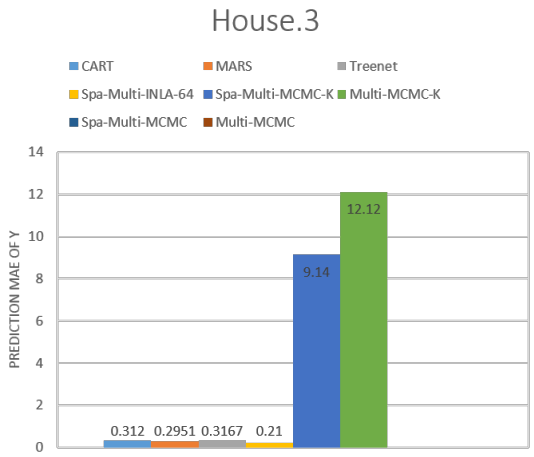
(b) Prediction accuracy for Z



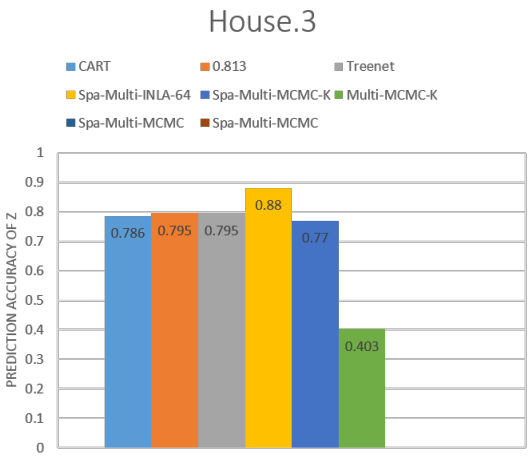
(c) Prediction MAE for Y



(d) Prediction accuracy for Z

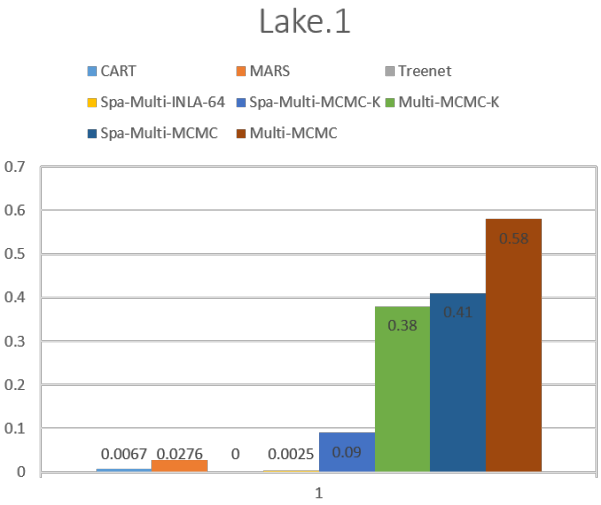


(e) Prediction MAE for Y

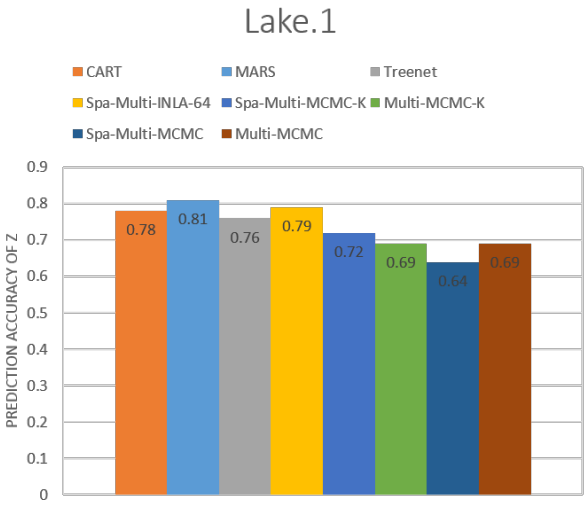


(f) Prediction accuracy for Z

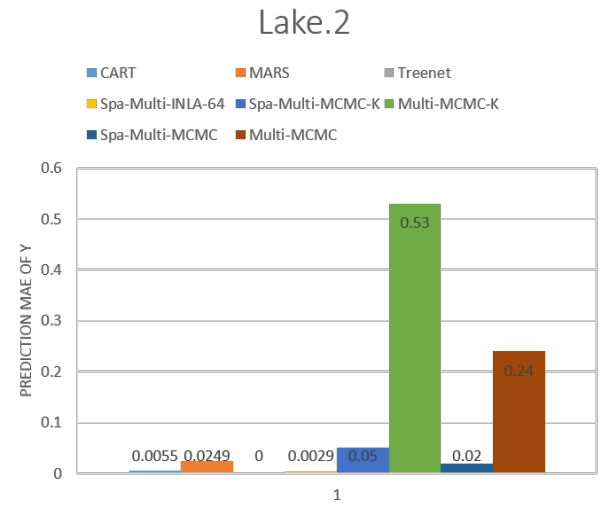
Figure 7.4: Comparison of the performances for eight approaches on House dataset



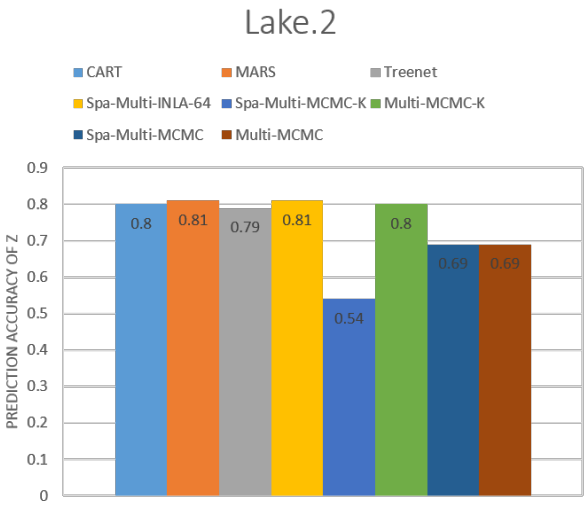
(a) Prediction MAE for Y



(b) Prediction accuracy for Z

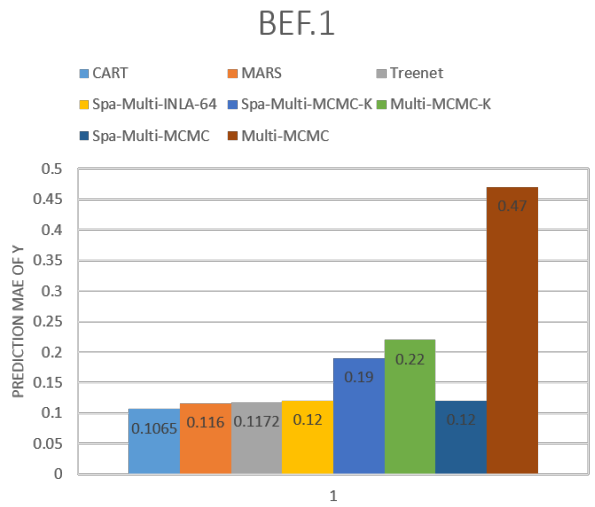


(c) Prediction MAE for Y

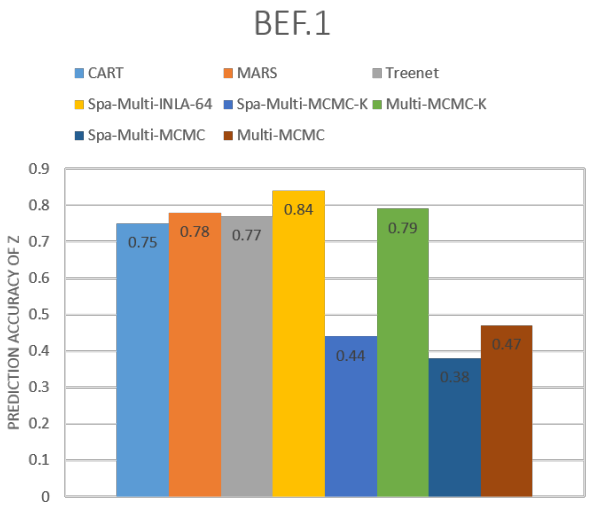


(d) Prediction accuracy for Z

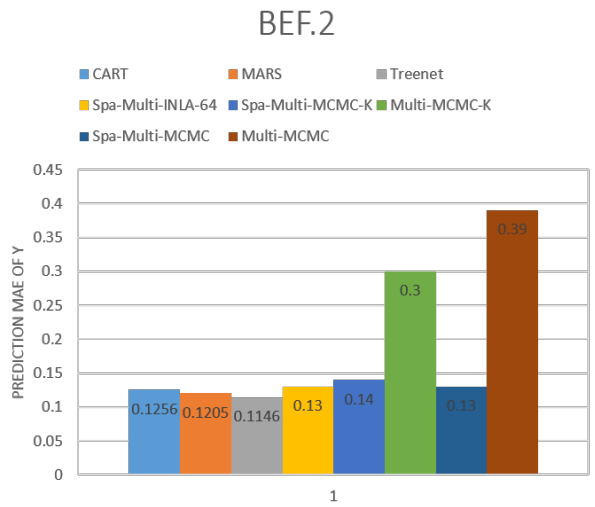
Figure 7.5: Comparison of the performances for eight approaches on Lake dataset



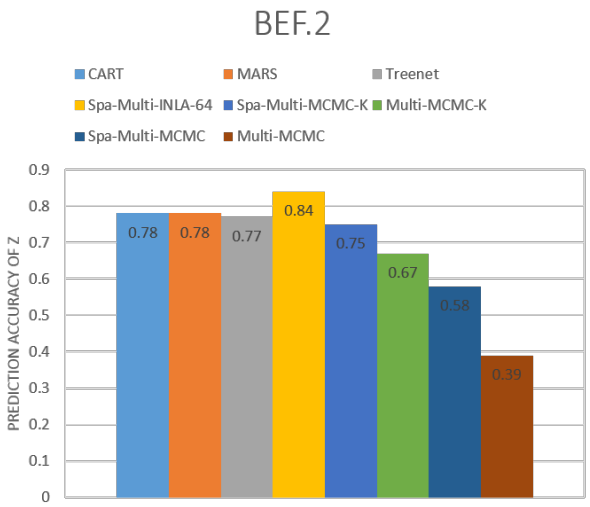
(a) Prediction MAE for Y



(b) Prediction accuracy for Z

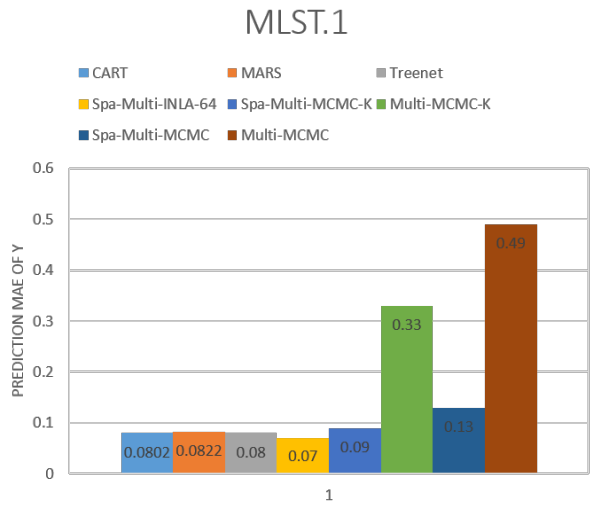


(c) Prediction MAE for Y

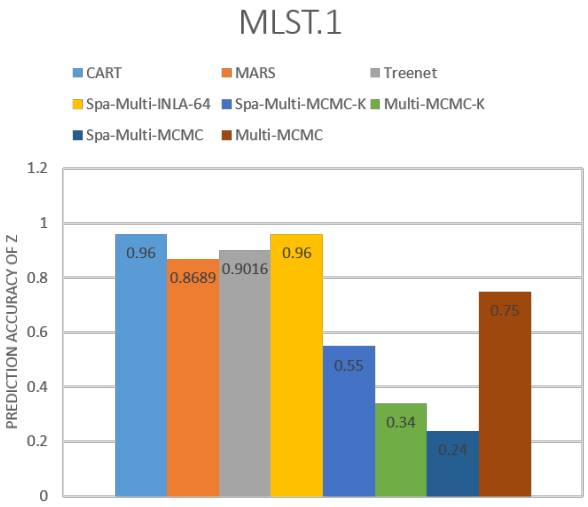


(d) Prediction accuracy for Z

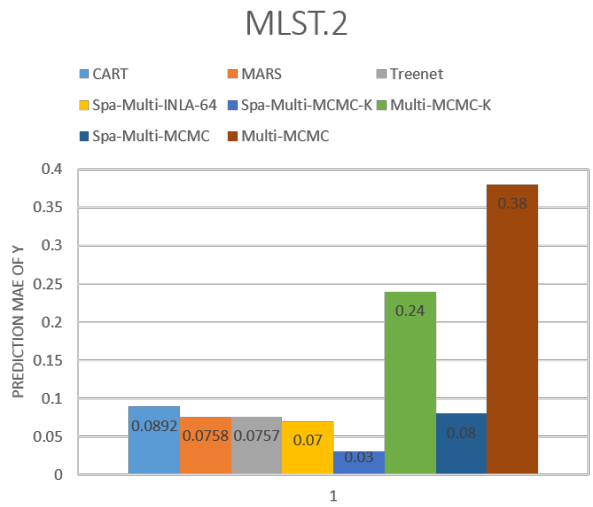
Figure 7.6: Comparison of the performances for eight approaches on BEF dataset



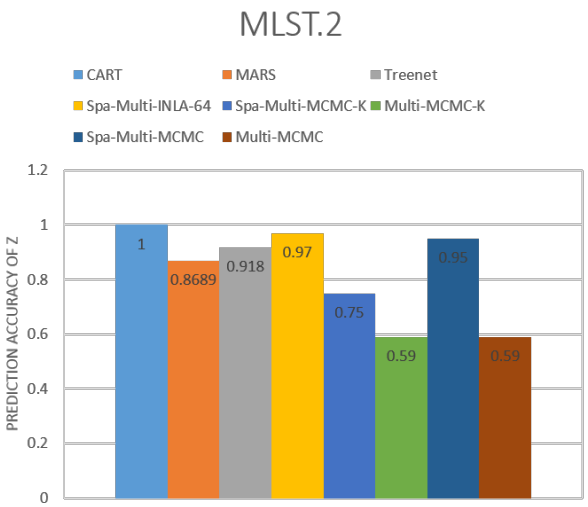
(a) Prediction MAE for Y



(b) Prediction accuracy for Z



(c) Prediction MAE for Y

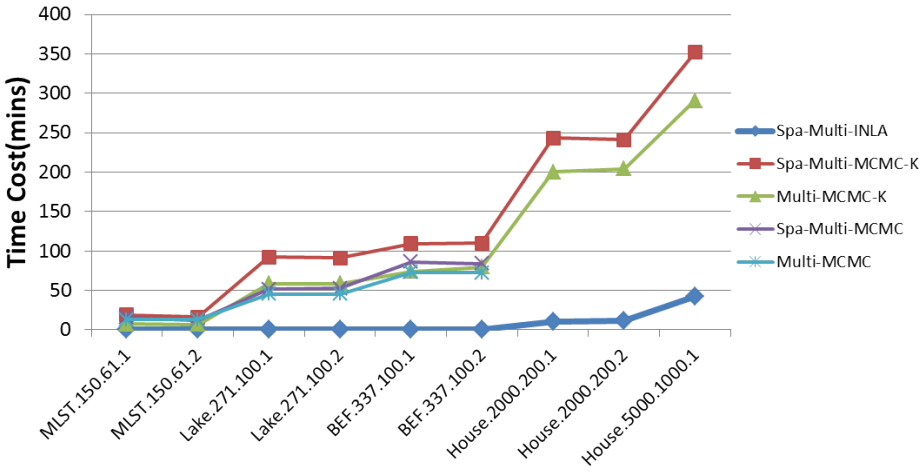


(d) Prediction accuracy for Z

Figure 7.7: Comparison of the performances for eight approaches on MLST dataset

when there are no appropriate iterations provided.

Computational cost



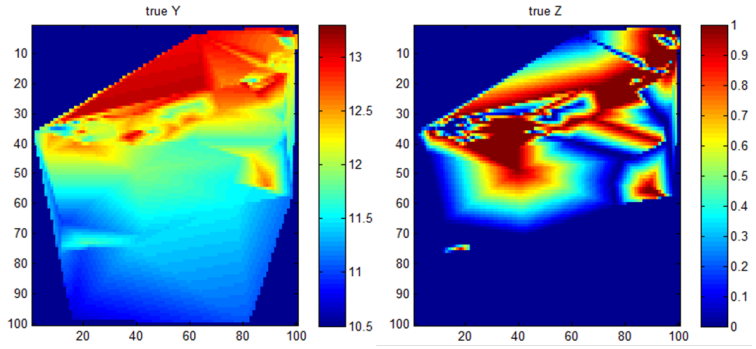
(a)

Figure 7.8: Comparison of the performances for eight approaches on real life datasets

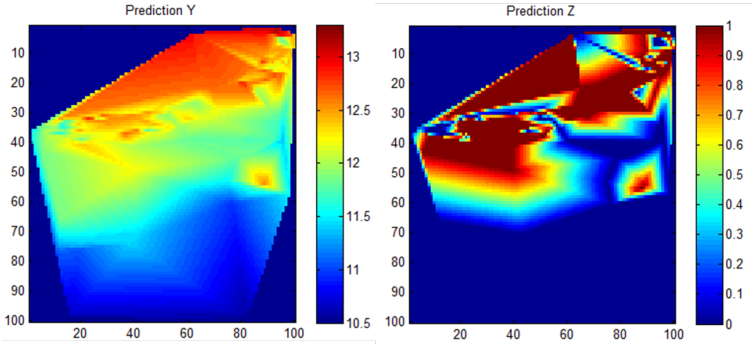
Real datasets showcase the speed and associate scalability achieved by the approaches that we evaluated. Fig. (7.8) compares the runtime performance of these algorithms in the datasets for varying numbers of training and testing points. For example, for *BEF.337.100.2*, Spa-Multi-INLA finished execution in around 0.63 mins, while Spa-Multi-MCMC-K completed in around 15.19 mins, and Multi-MCMC-K, Spa-Multi-MCMC and Multi-MCMC took from 11.43 to 32.64 mins. In particular, for *House.2000.200.2*, our methods finished running in 10 mins, while Spa-Multi-MCMC-K and Multi-MCMC-K took several hours. The other two approaches were not able to execute this dataset since it exceeds 2000 points.

7.5.3 Result analysis

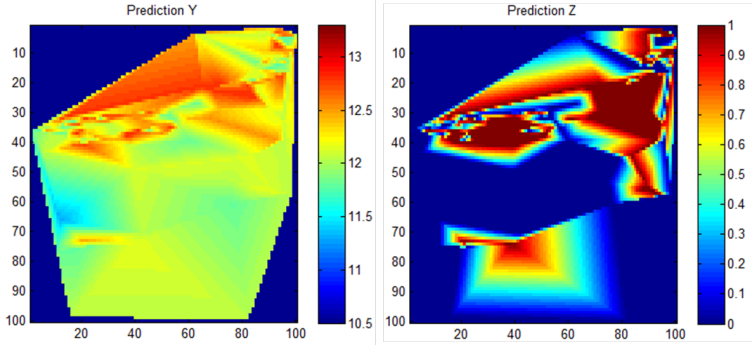
The above experimental results demonstrate that Spa-Multi-INLA is both effective and efficient in estimating the parameters and predicting different types of variables. It has a superior identification quality over existing techniques, achieving around 10-30% improvement over CART, MARs and Treetnet, and 40-50% over MCMC based approaches. The experimental results verified three observations. First, if there is an appropriate selection of knots that covers most of the domain interests, the predictive process cost will be significantly reduced to a linear order. Second, when combined with numerical routines, Laplace approximation techniques can provide much faster and more accurate parameter estimation than MCMC based



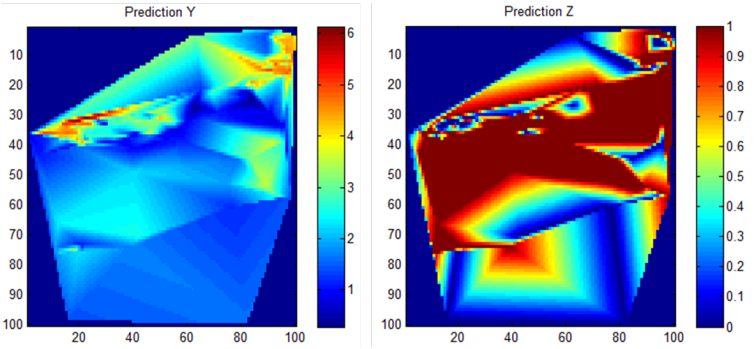
(a) True values



(b) Prediction values by Spa-Multi-INLA



(c) Prediction values by CART



(d) Prediction values by Spa-Multi-MCMC-K

Figure 7.9: Density map comparisons of the predicted values for *House* dataset. Y: numercial response; Z: binary response

algorithms for spatial multivariate non-Gaussian prediction. Third, when processing more sophisticated datasets, such as the simulation data shown in Fig. 7.2, MCMC based approaches need a very high number of iterations to achieve acceptable results at the cost of larger computational cost, and CART MARS and Treenet cannot handle such data with higher spatial dependencies, but our new approach can finish the prediction in moderate times with no loss of accuracy.

7.6 Conclusions

This paper proposes a novel framework for estimating multivariate predictive process models that is designed to take into account mixed type response variables. It integrates multivariate predictive process models with approximate Bayesian inference using INLA. The predictive model consists of a representative selection of knot locations which projects the spatial process to a lower dimensional subspace. The INLA provides more accurate and much faster inference for spatial multivariate predictive models. Experimental results on synthetic and real datasets conclusively demonstrated that our proposed non-Gaussian prediction model achieved a much higher processing capability in terms of prediction accuracy and computation time. The limitation of the model is that it assumes the spatial data accords to the generalized linear model fro the calucated local differences, but no justifications for this crtical assumption have been presented. The prediction performances on geostatistic data with both linear and non-linear trend are required to studying in real application.

Future work will focus on using the designed prediction model to solve multiple spatial data mining issues, including spatial outlier detection [33, 97, 99], spatial temporal outlier detection [96, 169] and spatial clustering [48, 114] for large mixed type dataset.

Chapter 8

Completed Work and Future Directions

This Ph.D. research focuses on the construction of efficient and effective approaches for several spatial data mining tasks, including robust prediction for large spatial data set, spatial prediction for large multivariate non-Gaussian data and spatial anomaly detection for both numerical and categorical data. Empirical results have shown the effectiveness and scalability of the proposed approaches.

8.1 Research Achievement

The proposed research is following four branches, spatial numerical outlier detection, spatial categorical outlier detection, robust inference for large spatial data set and spatial prediction for multivariate non-Gaussian data sets. Specifically, there are several research topics on each branch that are studied in this research. These topics as well as the corresponding tasks are listed as follows.

8.1.1 Spatial Numerical Outlier Detection (Chapter 3 and Chapter 4)

RW based Spatial Outlier Detection on Bipartite Graph

In this work, RW-BP based method has been proposed to compute the relevance by operating RW techniques on a weighted graph. In the bipartite graph, the vertex sets correspond to the spatial objects and the clusters generated from the non-spatial attributes of the objects in the spatial database. Random walk is performed on the bipartite graph to compute the similarities of the non-spatial attributes between any pair of the spatial objects. The spatial neighbor sets for each object can be formed using kNN method. And the outlierness for

each objects is computed as the differences between itself and its neighborhood. The experiments on both synthetic data and real data sets verify the outstanding performances of RW-BP in terms of its effectiveness.

RW based Spatial Outlier Detection on Exhaustive Combination Graph

However, sometimes the distribution of non-spatial attributes does not have the tendency to form clusters. That is, it is too difficult to execute a cluster operation, like uniform distribution. In this case, we can consider to construct another different graph. With the spatial and non-spatial attributes of the observations, we proposed an weighted Exhaustive Combination. In this graph, the vetex set composes of all the spatial objects and there is an edge between each pair of spatial objects. In the same way, random walk techniques is performed on the Exhaustive Combination Graph to compute the relevance vector for each spatial object. And then, using the sames steps as in the RW-BP approach to identify the spatial outliers. The operation of RW-EC constructs another different algorithm:1)the construction of the weighted EC graph; 2)the random is performing on EC graph to compute the relevance vector of any spatial object; 3)The outlierness value is identified as the local differences of the spatial object.

Identification of the Number of Spatial Outliers

A major limitation associated with the existing outlier detection algorithms is that they generally require a pre-specified number of spatial outiers. Estimating an appropriate number of outliers for a spatial data set is one of the critical issues for outlier analysis. This work is interested in designing an entropy-based method to address this problem in spatial numerical domain. Based on the relationship between outliers and the overall entropy, that is, the dataset with more outliers has a higher entropy value than that with less outliers. We expect, that is, the data set with more outliers has a higher entropy value than that with less outliers, we expect that, by incrementally removing outliers, the entropy value will decrease sharply, and reach a stable state when all the outliers have been removed.

8.1.2 Spatial Categorical Outlier Detection (Chapter 5)

Spatial Categorical Outlier Concept

The concept of the Spatial Categorical Outlier is proposed to differentiate SCOD with SNOD due to their intrinsically differences between numerical and categorical data. Actually, in real world, category-typed data has no intrinsic order information, which makes SCOD more complicated than SNOD. In this work, a spatial categorical outlier is defined as a spatially referenced observation which occurs infrequently with

regard to its neighborhood.

Design of PCF Based SCOD Algorithm)

Among several approaches to capturing the co-occurrence frequency, Pair Correlation Function has been proven very effective. In this work, we investigate the benefits of PCF techniques on SCOD and design algorithms in spatial categorical data sets. Firstly, we use the thoughts of PCF to define the Pair Correlation Ratio, which evaluate the relevance between each pair of categories with regards to different specified distances.With the set of discrete points in a 2-D space, determined by distances against PCR values, we can statistically learn a continuous PCR function which can easily estimate the PCRs among spatial objects. Then, the spatial neighbor set for each object can be formed using kNN technique. And the outlier degree for each object is computed by the mean of PCRs between itself and its neighborhood. Finally, the outlier scores are ranked in an ascending order and the top l objects are identified as SCOs.

Design of Approximate PCF Schemes: kNN Based Approach

For PCF-SCOD approach, in a larger size dataset, it is a time-consuming process to estimate the probability of each pair of categories as a function of distances. Since spatial outlier detection only takes focuses on the local differences, we consider only extract the co-occurrence among spatial objects which are spatial neighbors with each other. Based on such idea, kNN based estimator has been proposed, which is an approximate PCF-SCOD which only utilizes the kNN relationship to identify SCOS. Firstly, a kNN mapping function is defined to store the kNN information from the raw dataset into a pair dataset. Then, the co-occurrence of different pairs of categories are extracted based on the pair dataset so that the PCR between any pair of spatial objects which are spatial objects with each other can be computed.Finally, the outlierness for each object is computed by the mean of PCRs between itself with its neighbors.

Extension of kNN Based Approaches to multivariate spatial data

The current PCF series of frameworks was designed for univariate spatial data. This work will further generalize kNN based approaches to multivariate spatial data. That is, given a spatial dataset, an outlying observation is the one whose non-spatial attribute set occurs infrequently with regards to that of its neighborhood. A new concept, Pair Attribute Subset(PAS), was defined to evaluate the PCRs in multiple attribute domain. For each PAS combination, the PCR values among spatial objects were computed and constructed a PCR matrix in which each entry describes the relevance between a reference object with its neighbors. Furthermore, the outlier scores were computed for each pair of objects in every PAS combination. In the

end, the smallest ones is chosen as its final outlierness. We identify outliers in this way because sometimes, an outlier only exists in the subspace of the multiple attributes. Exhaustively estimating outliers scores in different PAS will help identify SCOs more effectively.

8.1.3 Robust Prediction and Outlier Detection for Spatial Datasets (Chapter 6)

Design of Robust and Reduced Rank Spatial Kriging Model (R^3-SKM)

R^3-SKM integrates robust, reduced-rank and Bayesian hierarchical techniques together. It is formalized in the framework of Bayesian hierarchical model with three layers, including the observation layer, the latent robust Gaussian process layer, and the parameter layer. The observation layer contains the observations which are assumed to follow a Gaussian distribution. Each random variable is related to the latent Gaussian effects in the second layer. The third level of the hierarchical model consists of the related parameters with the latent variables.

Design of Robust and Reduced Rank Parameter Estimation

Within R^3-SKM framework, R^3-PE is proposed to execute the robust parameter estimation by integrating Laplace approximation. It consists of two critical steps: 1). Gaussian approximation of the posterior distribution of latent variables conditional on parameters and observations; 2). Laplace approximation of the posterior distribution of corresponding parameters conditional on observations.

Design of Robust and Reduced Rank Spatial Prediction (R^3-SP)

Given a set of unsampled locations, R^3-SP enables to operating accurate spatial prediction at these locations. The first step is to estimate the posterior distribution of the corresponding latent variables conditional the observed samples. Then, with it, we can predict the interested values for the unsampled locations conditional the observed values at the sampled ones. Specifically, we first derive the K samples of parameters and their weight values; and then utilizes each generated sample to construct the corresponding mean and covariance matrix of latent variables. Furthermore, the likelihood of the observations are approximated as the result of a quadratic form of the latent variables. Next, its distribution is analytically intractable inference.

Design of Robust and Reduced Rank Spatial Anomaly Detection (R^3-SOD)

Statistically, spatial outlier can be interpreted as observations that have abnormally low correlations with their spatial neighbors. R^3-SKM can also be utilized to identify spatial outliers as objects with higher predicted values. First, we apply R^3-SKM to accurately estimate the latent variables and parameters for the contaminated spatial dataset. Second, the estimated values are utilized to operate a spatial prediction for each observed location. Finally, the differences between observed and predicted values are computed to measure their outlying degree. The objects which have higher measurement errors are labeled as spatial outliers.

8.1.4 Spatial Prediction for Multivariate Non-Gaussian Datasets (Chapter 7)

Design of A Spatial Multivariate Non-Gaussian Hierarchical Framework

The spatial model is based on a hierarchical framework and is specifically designed to take account of mixed type random variables. It is the first work that applies both knot-based and Laplace approximation techniques to multivariate non-Gaussian datasets. The knot based technique is utilized to model the predictive process as a reduced-rank spatial process, which projects the process realizations of the spatial model to a lower dimensional subspace. This projection significantly reduces the computational cost.

Design of A Prediction Algorithm for G + B Mixture Data

By integrating the Laplace approximation, our approach efficient makes approximations to the posterior marginal of latent variables for the predictive process, and performs accurate spatial prediction. Suppose the output of the observations consists of Gaussian and Binary data, the mixed-type attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. For binary data, the success probability of the observations is defined by a logit link function. The dependency between mixed type attributes is then modeled by the relationship between their latent numerical random variables using a variance-covariance matrix.

Design of A Prediction Algorithm for G + C Mixture Data

Here, one output of the observation is assumed to be a Gaussian variable and the first stage of the other non-Gaussian process be the count response modeled using Poisson regression. Essentially, we assume the

expected value of count responses is linear on a transformed scale mapped by a suitable link function.The Gaussian variable and the Poisson variable depend on latent variables, which are responsible for the spatial dependence together. We will execute experiments on synthetic data sets to validate the effectiveness and efficiency of proposed models.

Design of A Prediction Algorithm for C + B Mixture Data

Here, the first stages of two non-Gaussian processes are the binary and count responses, which are modeled using Binomial and Poisson regression, respectively. Essentially, we assume the expected values of binary and count responses are both linear on the transformed scales mapped by their corresponding suitable link functions. In the same way, the Binomial and Poisson variables depend on latent variables, which are responsible for the spatial dependence together. We will also execute experiments on synthetic data sets to validate the effectiveness and efficiency of proposed models.

8.2 Future Direction

This section discusses important directions of the following topics for future work, including spatio-temporal outlier detection and Spatial anomaly detection for multivariate non-Gaussian Datasets.

8.2.1 Anomaly Detection for Spatial Mixed Type Dataset

As introduced, approaches to anomaly detection include distance-based, local density based, graph based, and statistical based methods. Most of these approaches are designed for single type datasets, whereas most real world the non-spatial attributes of datasets are composed of a mixture of different data types, such as numerical, binary, ordinal, nominal, and count. Direct applications of these approaches to mixed type data results in the loss of significant correlations between attributes, and their extension to mixed-type data is technically challenging. For example, the statistical mode based approach relies on modeling the correlations between different attributes, but there is no uniform correlation measure available for mixed-type attributes. There are two main challenges for mixed-type datasets, namely modeling mutual correlations between mixed-type attributes and capturing large variations due to anomalies.

As an extend of the work in Chapter 7, we consider to design a statistical based approach to address the above challenges. We begin by presenting a new variant of the spatial generalized linear model that can capture the mutual correlations between mixed-type attributes. Specifically, the mixed-type attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. Each attribute is mapped

to a corresponding latent numerical variable via a specific link function, such as logit function for binary attribute and log function for count attribute. Using link functions to model attribute. Using link function to model attributes of different types is one of the most popular strategies for modeling non-numerical data. Based on this strategy, the dependency between mixed type attributes is captured by the relationship between their latent variables using variance-covariance matrix. Then, an "error buffer" component will be incorporated based on heavy tailed distribution to capture the large variaions caused by anomalies. Heavy tailed distributions have been widely used in robust statistics to minimize the effects of outliers in variety of statistical models, e.g., multivariate regression, Kalman filtering, clustering, and independent component analysis. By fitting the data into the model, the error buffer absorbs all of the errors. The detection process then revisits the error buffer and detects those abnormal instances with irregular magnitudes of error.

Based on the above idea, we will design an unsupervised framework suitable for general purpose spatial anomaly detection on mixed-type data. The framework incorporates a novel statistical model for capturing abnormal behaviors based on improving the INLA method to approximate Bayesian inference. We will conduct a series of experiments on synthetic and real-life datasets to verify the idea.

8.2.2 Spatio-Temporal Outlier Detection

We have presented several generic solutions to the problems of numerical and non-numerical anomaly detection. We will extend these works to the spatio-temporal domain. There is a lack of effective and efficient approaches to idnetifying outliers from large scale multivariate mixed type spatio-temporal datsets. For multivariate mixed type spatio-temporal dataset, it raises three research challenges: 1). How to model spatial correlations among mixed type attributes; 2). How to remove the side-effect of the outliers; 3). How to reduce the computational cost.

Currently, there are two popular frameworks which are used to model spatio-temporal data, including spatio-temporal random effects mode (STRE) and the spatio-temporal Kriging (STK) model. First, we will try to design a framework which integrates these two models with a heavy tailed distribution to absorb the large variations cuased by outliers. Second, the knot based technique will be utilized to reduce the computation burden from a high dimension (the number of the oservations) to a lower domain (the number of representative knots). Finally, we can map the mixed-type variables to laten numerical random ones that are multivariate Gaussian. Further, the spatial correlations among mixed type attribute are modeled using a variance-covariance matrix. After integrating the heavy-tailed distribution and mapping non-Gaussian variables into Gaussin domain by the link function, the inference of the designed model will be analytically intractable. For this issues, we consider apply the approximationg inference techniques, like Gaussian approximation and Laplace approximation and expectation propagation.

8.3 Current Publications

In Progress

Xutong Liu, Feng Chen, Chang-Tien Lu. "Spatial Outlier Detection in Multivariate Non-Gaussian Dataset", target to *IEEE International Conference on Data Mining* ***(ICDM)***, 2013.

Feng Chen, **Xutong Liu**, Chang-Tien Lu. "On Local Based Techniques for Spatial Outlier Detection", target to *ACM Transactions on Intelligent Systems and Technology* ***(TIST)***, 2013.

Feng Chen, **Xutong Liu**, Chang-Tien Lu. "A Unified Outlier Detection Framework for Spatial Continuous, Categorical, and Count data", target to *IEEE Transactions on Knowledge and Data Engineering* ***(TKDE)***, 2013.

Journals

Xutong Liu, Feng Chen, Chang-Tien Lu. "On Detecting Spatial Categorical Outliers", *Journal of Geoinformatica*, 2013, Submitted after Major Revision.

Xutong Liu Feng Chen, Yen-Cheng Lu, Chang-Tien Lu, "Spatial Prediction of Large Multivariate Non-Gaussian Data", *IEEE Transaction on Knowledge and Data Engineering* ***(TKDE)***, 2013, submitted.

Chang-Tien Lu, Raimundo F. Dos Santos, **Xutong Liu**, Yufeng Kou. "A Graph-Based Approach to Detect Abnormal Spatial Points and Regions", *International Journal on Artificial Intelligence Tools*, Volume: 20, Issue: 4, pp. 721-751, 2011.

Xutong liu, Huijin Wang, Changshu Jian. "Applying Grid to Evolutionary Computation", *Journal of Computer Application*, Volume: 25, Issue: 11, pp. 2635-2637, 2005.

Conferences

Xutong Liu, Feng Chen, Chang-Tien Lu. "Robust Prediction and Outlier Detection for Spatial Data Sets", *IEEE International Conference on Data Mining (****IEEE ICDM****)*, pp. 370-379, Brussels, Belgium, December 10-13, 2012.

Xutong liu, Feng Chen, Chang-Tien Lu. "Spatial Categorical Outlier Detection: Pair Correlation Function Based Approach", *Proceedings of the 19th ACM SIG SPATIAL International Conference on Advances in Geographic Information Systems (****ACM SIGSPATIAL GIS****)*, pp. 465-468, Chicago, Illinois, November 1-4, 2011.

Xutong liu, Chang-Tien Lu, Feng Chen. "Spatial Outlier Detection: Random Walk based Approaches", *Proceedings of the 18th ACM SIG SPATIAL International Conference on Advances in Geographic Information Systems (**ACM SIGSPATIAL GIS**)*, pp. 370-379, San Jose, California, November 2-5, 2010.

Xutong liu, Changshu Jian, Chang-Tien Lu. "Demo paper: Spatial Temporal Search Engine", *Proceedings of the 18th ACM SIG SPATIAL International Conference on Advances in Geographic Information Systems (**ACM SIGSPATIAL GIS**)*, pp. 528-529, San Jose, California, November 2-5, 2010.

C. Kwan, B. Ayhan, J. Yin, **X. Liu**, P. Ballal, A. Athameneh, A. Ramani, W. Lee, F. Lewis. "Real-Time System Condition Monitoring Using Wireless Sensors", *IEEE Aerospace Confernce*, pp. 1-8, 2009.

Xutong Liu, Chang-Tien Lu, Feng Chen. "An Entropy Based Method for Assessing the Number of Spatial Outliers", *IEEE International Conference on Information Reuse and Integration (**IEEE IRI**)*, pp. 244-249, Las Vegas, Nevada, July 13-15, 2008.

C. Kwan, S. Chun, J. Yin, **X. Liu**, M. Kruger, I. Sityar. "Enhanced Speech in Noisy Multiple Speaker Environment", *International Joint Conference on Neural Networks (IJCNN*), pp. 1640-1643, 2008.

C. Kwan, J. Yin, B. Ayhan, S. Chu, **X. Liu**, K. Puckett, Y. Zhao, K.C. Ho, M. Kruger, I. Sityar. "An Integrated Approach to Robust Speaker Identification and Speech Recognition". *International Joint Conference on Neural Networks (IJCNN)*, pp. 1635-1639, 2008.

C. Kwan, J. Yin, B. Ayhan, S. Chu, **X. Liu**, K. Puckett, Y. Zhao, K.C. Ho, M. Kruger, I. Sityar. "An Integrated Approach to Robust Speaker Identification and Speech Recognition". *International Joint Conference on Neural Networks (IJCNN)*, pp. 1635-1639, 2008.

C. Kwan, J. Yin, B. Ayhan, S. Chu, **X. Liu**, K. Puckett, Y. Zhao, K.C. Ho, M. Kruger, I. Sityar. "Speech Separation Algorithm for Multiple Speaker Environment". *International Joint Conference on Neural Networks (IJCNN)*, pp. 1644-1648, 2008.

Bibliography

[1] J. Abernethy, T. Evgeniou, O. Toubia, and J.-P. Vert. Eliciting consumer preferences using robust adaptive choice questionnaires. *IEEE Trans. on Knowl. and Data Eng.*, 20(2):145–155, 2008.

[2] N. R. Adam, V. P. Janeja, and V. Atluri. Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. In *Proceedings of the 2004 ACM symposium on Applied computing*, SAC '04, pages 576–583, New York, NY, USA, 2004. ACM.

[3] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01, pages 37–46, New York, NY, USA, 2001. ACM.

[4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, 1994.

[5] F. Angiulli and F. Fassetti. Detecting distance-based outliers in streams of data. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 811–820, New York, NY, USA, 2007. ACM.

[6] F. Angiulli and F. Fassetti. Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets. *ACM Trans. Knowl. Discov. Data*, 3:4:1–4:57, March 2009.

[7] L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27(2):93–115, 1995.

[8] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto. An 1 -laplace robust kalman smoother. *IEEE Trans. Automat. Contr.*, 56(12):2898–2911, 2011.

[9] B. Arunasalam, S. Chawla, P. Sun, and R. Munro. Mining complex relationships in the sdss skyserver spatial database. In *COMPSAC Workshops*, pages 142–145, 2004.

[10] S. Bandyopadhyay. Simulated annealing using a reversible jump markov chain monte carlo algorithm for fuzzy clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):479–490, 2005.

[11] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets, 2008.

[12] D. Barbará, Y. Li, and J. Couto. Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 582–589, New York, NY, USA, 2002. ACM.

[13] A. Beghdadi and A. Khellaf. A noise-filtering method using a local information measure. *IEEE Transactions on Image Processing*, 6(6):879–882, 1997.

[14] L. Bel, D. Allard, J. M. Laurent, R. Cheddadi, and A. Bar-Hen. Cart algorithm for spatial data: Application to environmental and ecological data. *Comput. Stat. Data Anal.*, 53(8):3082–3093, June 2009.

[15] r. Björck. *Numerical methods for least squares problems*. SIAM, 1996.

[16] J. V. Z. B.M. Golam Kibria, Li Sun and N. D. Le. Bayesian spatial prediction of random space-time fields with application to mapping pm2.5 exposure. *Journal of the American Statistical Association*, 97:112–124, 2002.

[17] M. Boley and H. Grosskreutz. A randomized approach for approximating the number of frequent sets. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 43–52, 2008.

[18] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM*, pages 243–254, 2008.

[19] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[21] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 93–104, New York, NY, USA, 2000. ACM.

[22] R. Bronstein, J. Das, M. Duro, R. Friedrich, G. Kleyner, M. Mueller, S. Singhal, I. Cohen, G. Kleyner, M. Mueller, S. Singhal, and I. Cohen. Self-aware services: Using bayesian networks for detecting anomalies in internet-based services. In *Northwestern University and Stanford University. Gary (Igor*, pages 623–638. Publishing, 2001.

[23] P. Burrough. *Principles of geographical information systems for land resources assessment*. Monographs on soil and resource surveys. Clarendon Press, 1994.

[24] H. Cao, N. Mamoulis, and D. W. Cheung. Discovery of periodic patterns in spatiotemporal sequences. *IEEE Trans. Knowl. Data Eng.*, 19(4):453–467, 2007.

[25] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

[26] V. Ch, A. Banerjee, V. Kumar, and V. Chandola. Outlier detection: A survey, 2007.

[27] P. Chagneau, F. Mortier, N. Picard, and J.-N. Bacro. Hierarchical bayesian model for gaussian, poisson and ordinal random fields. volume 16 of *Quantitative Geology and Geostatistics*, pages 333–344. Springer Netherlands, 2010.

[28] P. Chagneau, F. Mortier, N. Picard, and J.-N. Bacro. A hierarchical bayesian model for spatial prediction of multivariate non-gaussian random fields. *Biometrics*, 67(1):97–105, 2011.

[29] N. Chaikaew, N. K. Tripathi, and M. Souris. Exploring spatial patterns and hotspots of diarrhea in chiang mai, thailand. *International Journal of Health Geographics*, 8(1):36, 2009.

[30] P. K. Chan, M. V. Mahoney, and M. H. Arshad. A machine learning approach to anomaly detection. 2003.

[31] V. Chandola, S. Boriah, and V. Kuman. Understanding categorical similarity measures for outlier detection. Technical report, University of Minnesota, 2008.

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

[33] F. Chen, C.-T. Lu, and A. P. Boedihardjo. Gls-sod: a generalized local statistical approach for spatial outlier detection. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1069–1078, 2010.

[34] Y. Chen, K. Chen, and M. A. Nascimento. Effective and efficient shape-based pattern detection over streaming time series. *IEEE Trans. on Knowl. and Data Eng.*, 24(2):265–278, Feb. 2012.

[35] T. Cheng and Z. Li. A multiscale approach for spatio-temporal outlier detection. *T. GIS*, 10(2):253–263, 2006.

[36] J. Chica-Olmo. Prediction of housing location price by a multivariate spatial method: Cokriging. *Journal of Real Estate Research*, 29(1):95–114, 2007.

[37] N. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1991.

[38] N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008.

[39] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

[40] K. Das and J. Schneider. Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 220–229, New York, NY, USA, 2007. ACM.

[41] A. P. Dawid. Posterior expectations for large observations. *Biometrika*, 60:664–667, 1973.

[42] A. Delis, C. Faloutsos, and S. Ghandeharizadeh, editors. *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*. ACM Press, 1999.

[43] C. Deutsch and A. Journel. *GSLIB: geostatistical software library and user's guide*. Number v. 1. Oxford University Press, 1992.

[44] P. Diggle. Applied Spatial Statistics for Public Health Data. *Journal of the American Statistical Association*, 100(470):702–703, June 2005.

[45] P. Diggle, R. A. Moyeed, and J. A. Tawn. Model-based geostatistics. *Applied Statistics*, 47:299–350, 1998.

[46] R. A. Dubin. Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22(3):433–452, 1992.

[47] J. Durbin and S. J. Koopman. Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84:669–684, 1997.

[48] M. Ester, H.-p. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining*, 1996(6):226–231.

[49] A. O. Finley and S. Banerjee. Hierarchical modeling for non-gaussian spatial data in r. 2009.

[50] A. O. Finley, H. Sang, S. Banerjee, and A. E. Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884, 2009.

[51] J. H. Friedman. Rejoinder: Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):123–141, 1991.

[52] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[53] M. A. Gandhi and L. Mili. Robust kalman filter based on a generalized maximum-likelihood-type estimator. *Trans. Sig. Proc.*, 58(5):2509–2520, May 2010.

[54] A. E. Gelfand and S. Banerjee. Multivariate spatial process models. *Handbook of Spatial Statistics*, pages 495–515, 2010.

[55] A. E. Gelfand, A. M. Schmidt, S. Banerjee, and C. F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion). *Test*, 13(2):1–50, 2004.

[56] J. Geweke. Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40, 1993.

[57] J. Ghosh, D. Lambert, D. B. Skillicorn, and J. Srivastava, editors. *Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA*. SIAM, 2006.

[58] A. R. Gonzales, R. B. Schofield, S. V. Hart, J. E. Eck, S. Chainey, J. G. Cameron, M. Leitner, R. E. Wilson, and S. V. Hart. Mapping crime: Understanding hot spots, 2005.

[59] P. Goovaerts. *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press, 1997.

[60] G. Grekousis and Y. N. Fotis. A fuzzy index for detecting spatiotemporal outliers. *Geoinformatica*, 16(3):597–619, July 2012.

[61] L. Hagen and A. B. Kahng. A new approach to effective circuit clustering. In *Proceedings of the 1992 IEEE/ACM international conference on Computer-aided design*, ICCAD '92, pages 422–427, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.

[62] R. Haining. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, 1990.

[63] J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational & Graphical Statistics*, 14(4):928–946, December 2005.

[64] D. Harel and Y. Koren. Clustering spatial data using random walks. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 281–286, New York, NY, USA, 2001. ACM.

[65] D. J. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[66] A. Hartkamp, I. Maize, and W. I. Center. *Interpolation Techniques for Climate Variables*. Geographic information systems series. International Maize and Wheat Improvement Center, 1999.

[67] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[68] Z. He, S. Deng, X. Xu, and J. Z. Huang. A fast greedy algorithm for outlier mining. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge and Data Discovery*, pages 567–576, 2006.

[69] Z. He, X. Xu, and S. Deng. An optimization model for outlier detection in categorical data. *CoRR*, abs/cs/0503081, 2005.

[70] Z. He, X. Xu, J. Z. Huang, and S. Deng. A frequent pattern discovery method for outlier detection. In *WAIM*, pages 726–732, 2004.

[71] Z. He, X. Xu, J. Z. Huang, and S. Deng. Fp-outlier: Frequent pattern based outlier detection. *Comput. Sci. Inf. Syst.*, 2(1):103–118, 2005.

[72] T. Hengl. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2):75–93, 2004.

[73] D. Higdon. Space and space-time modeling using process convolutions. *Technique Report*, 2008.

[74] J. Hoef and N. Cressie. Multivariable spatial prediction. volume 25, pages 219–240. Kluwer Academic Publishers-Plenum Publishers, 1993.

[75] http://www.iiasa.ac.at/Research/LUC/External-World-soil database/HTML/.

[76] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Trans. on Knowl. and Data Eng.*, 20(4):433–448, Apr. 2008.

[77] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. Statistical analysis and modelling of spatial point patterns. 2008.

[78] E. Isaaks and R. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, 1989.

[79] V. P. Janeja and V. Atluri. Fs3: A random walk based free-form spatial scan statistic for anomalous window detection. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 661–664, Washington, DC, USA, 2005. IEEE Computer Society.

[80] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16:1370–1386, 2004.

[81] W. Jin, A. K. H. Tung, and J. Han. Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 293–298, New York, NY, USA, 2001. ACM.

[82] A. Journel and C. Huijbregts. *Mining Geostatistics*. BLACKBURN Press, 2003.

[83] B. Jungbacker and S. J. Koopman. Monte carlo estimation for nonlinear non-gaussian state space models. *Biometrika*, 94(4):827–839, 2007.

[84] P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust gaussian process regression with a student-t likelihood. *J. Mach. Learn. Res.*, 12:3227–3257, 2011.

[85] E. E. Kammann and M. P. Wand. Geoadditive models. *Applied Statistician*, 52:1–18, 2003.

[86] M. Katzfuss and N. Cressie. Spatio-temporal smoothing and em estimation for massive remoteâĂŘsensing data sets. *Journal of Time Series Analysis*, 32:430–446, 2011.

[87] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, pages 392–403, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[88] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 211–222, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[89] E. M. Knorr, R. T. Ng, and R. H. Zamar. Robust space transformations for distance-based operations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 126–135, New York, NY, USA, 2001. ACM.

[90] Y. Kou, C.-T. Lu, and R. F. D. Santos. Spatial outlier detection: A graph-based approach. In *ICTAI (1)*, pages 281–288, 2007.

[91] Y. Kou, C.-T. Lu, and R. F. D. Santos. Spatial outlier detection: A graph-based approach. *Tools with Artificial Intelligence, IEEE International Conference on*, 1:281–288, 2007.

[92] A. Koufakou, E. G. Ortiz, M. Georgiopoulos, G. C. Anagnostopoulos, and K. M. Reynolds. A scalable and efficient outlier detection strategy for categorical data. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI '07, pages 210–217, Washington, DC, USA, 2007. IEEE Computer Society.

[93] A. Koufakou, J. Secretan, J. Reeder, K. Cardona, and M. Georgiopoulos. Fast parallel outlier detection for categorical datasets using mapreduce. *IEEE World Congress on Computational Intelligence (WCCI)*, 2008.

[94] G. Li, N. Japkowicz, T. J. Stocki, and R. K. Ungar. Border sampling through coupling markov chain monte carlo. pages 393–402, 2008.

[95] X. Lin, G. Wahba, D. Xiang, F. Gao, and R. K. M. B. Klein. Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *Ann. Statist*, pages 1570–1600, 2000.

[96] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1010–1018, New York, NY, USA, 2011. ACM.

[97] X. Liu, F. Chen, and C.-T. Lu. Spatial categorical outlier detection: pair correlation function based approach. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 465–468, 2011.

[98] X. Liu, F. Chen, Y.-C. Lu, and C.-T. Lu. http://filebox.vt.edu/users/xutongl/spinla/exppackage.zip.

[99] X. Liu, C.-T. Lu, and F. Chen. Spatial outlier detection: random walk based approaches. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 370–379, New York, NY, USA, 2010. ACM.

[100] C.-T. Lu, D. Chen, and Y. Kou. Algorithms for spatial outlier detection. In *ICDM*, pages 597–600, 2003.

[101] C.-T. Lu, D. Chen, and Y. Kou. Detecting spatial outliers with multiple attributes. In *ICTAI*, pages 122–128, 2003.

[102] W. Lu and I. Traore. Determining the optimal number of clusters using a new evolutionary algorithm. In *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '05, pages 712–713, Washington, DC, USA, 2005. IEEE Computer Society.

[103] V. Malbasa and S. Vucetic. Spatially regularized logistic regression for disease mapping on large moving populations. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1352–1360, New York, NY, USA, 2011. ACM.

[104] K. Mardia and C. Goodall. Spatiotemporal analyses of multivariate environmental monitoring datay. *Multivariate environmental statistics: Elsevier, Amsterdam*, pages 347–386, 1993.

[105] R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley and Sons, New York, 2006.

[106] G. Matheron. *The Theory of Regionalized Variables and Its Applications*. Cahiers. École national supérieure des mines, 1971.

[107] A. B. McBratney, I. O. A. Odeh, T. F. A. Bishop, M. S. Dunbar, and T. M. Shatar. An overview of pedometric techniques for use in soil survey. *Geoderma*, 97(3-4):293–327, 2005.

[108] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *EMNLP*, pages 404–411, 2004.

[109] N. Y. Mingming. Probabilistic networks with undirected links for anomaly detection. In *Proceedings of IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, pages 175–179, 2000.

[110] S. H. Mohammadi, V. P. Janeja, and A. Gangopadhyay. Discretized spatio-temporal scan window. *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 1195–1206, 2009.

[111] H. D. K. Moonesignhe and P.-N. Tan. Outlier detection using random walks. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 532–539, Washington, DC, USA, 2006. IEEE Computer Society.

[112] R. M. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical report, 1997.

[113] R. M. Neal. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Jan. 1997.

[114] D. B. Neill and A. W. Moore. Rapid detection of significant spatial clusters. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265, 2004.

[115] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 144–155, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[116] J. Oh and K.-D. Kang. A predictive-reactive method for improving the robustness of real-time data services. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2012.

[117] A. O'Hagan. On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society, Series B*, 41(3):358–367, 1979.

[118] V. D. Oliveira. Bayesian prediction of clipped gaussian random fields. *Computational Statistics and Data Analysis*, 34(3):299–314, 2000.

[119] M. A. Oliver and R. Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Science*, 4(3):313–332, 1990.

[120] K. Ord. Outliers in statistical data : V. barnett and t. lewis, 1994, 3rd edition, (john wiley & sons, chichester), 584 pp., [uk pound]55.00, isbn 0-471-93094-6. *International Journal of Forecasting*, 12(1):175–176, March 1996.

[121] M. E. Otey, A. Ghoting, and S. Parthasarathy. Fast distributed outlier detection in mixed-attribute data sets. *Data Min. Knowl. Discov.*, 12:203–228, May 2006.

[122] K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[123] C. J. Paciorek. Computational techniques for spatial logistic regression with large data sets. *Comput. Stat. Data Anal.*, 51(8):3631–3653, 2007.

[124] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[125] N. Pal and S. Pal. Entropy: a new definition and its applications. In *IEEE Transactions on Systems, Man dna Cybernetics*, pages 1260 – 1270, 1991.

[126] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 653–658, New York, NY, USA, 2004. ACM.

[127] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 653–658, New York, NY, USA, 2004. ACM.

[128] D. Pelleg. *Scalable and practical probability density estimators for scientific anomaly detection*. PhD thesis, Pittsburgh, PA, USA, 2004. AAI3126928.

[129] G. Piatetsky-Shapiro, C. Djeraba, L. Getoor, R. Grossman, R. Feldman, and M. Zaki. What are the grand challenges for data mining? kdd-2006 panel report. *SIGKDD Explor. Newsl.*, 8:70–77, December 2006.

[130] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 427–438, New York, NY, USA, 2000. ACM.

[131] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[132] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, ICAPRDT'99, pages 137–143, New Delhi, India, 2000. Narosa Publishing House.

[133] T. Reed and K. Gubbins. *Applied statistical mechanics: thermodynamic and transport properties of fluids*. Butterworth-Heinemann reprint series in chemical engineering. Butterworth-Heinemann, 1973.

[134] G. Ridgeway and D. Madigan. Bayesian analysis of massive datasets via particle filters. pages 5–13, 2002.

[135] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

[136] H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society*, 71(2):319–392, 2009.

[137] O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman Hall/CRC, 2005.

[138] A. Schmidt and M. Rodriguez. Modelling multivariate counts varying continuously in space. *Bayesian Statistic*, 2011.

[139] C. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, Urbana, 1949.

[140] S. Shekhar and S. Chawla. *Spatial databases - a tour*. Prentice Hall, 2003.

[141] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. T. Lu. Spatial databases: Accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering*, 11:45–55, 1999.

[142] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. T. Lu. Spatial databases: Accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering*, 11:45–55, 1999.

[143] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *KDD*, pages 371–376, 2001.

[144] S. Shekhar, C.-T. Lu, P. Zhang, S. Shekhar, C. T. Lu, and P. Zhang. A unified approach to spatial outliers detection. *GeoInformatica*, 7:139–166, 2003.

[145] A. Shevyrnogov, G. Vysotskaya, and E.Shevyrnogov. Spatial and temporal anomalies of sea surface temperature in global scale (by space-based data). *Advances in Space Research*, 33(7):1179–1183, 2004.

[146] J. Snow. *On the mode of communication of cholera*. John Churchill, London, 1855.

[147] G. Software. http://www.gslib.com/.

[148] spBayes. spbayes: Univariate and multivariate spatial modeling. *http://cran.r-project.org/web/packages/spBayes/*, 2012.

[149] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 418–425, Washington, DC, USA, 2005. IEEE Computer Society.

[150] P. Sun and S. Chawla. On local spatial outliers. In *IEEE International Conference on Data Mining*, pages 209–216, 2004.

[151] M. Svensén and C. M. Bishop. Robust bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

[152] S. System. http://www.salford-systems.com/.

[153] P. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Finding spatio-temporal patterns in earth science data. In *Proc. of KDD Workshop on Temporal Data Mining*, 2001.

[154] H. S. Teng and K. Chen. Adaptive real-time anomaly detection using inductively generated sequential patterns. *Security and Privacy, IEEE Symposium on*, 0:278, 1990.

[155] W. R. Tobler. *Cellular Geography*, pages 379–389. Reidel, Dordrecht, Netherlands, 1979.

[156] J. Vanhatalo, P. Jylänki, and A. Vehtari. Gaussian process regression with student-t likelihood. pages 1910–1918, 2009.

[157] C. Varin, G. Host, and O. Skare. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis*, 49(4):1173–1191, 2005.

[158] J. M. Ver Hoef and R. P. Barry. Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion). *J. Stat. Plan. Inference*, 69(2):275–294, 1998.

[159] C. J. Vilalta. The spatial dynamics and socioeconomic correlates of drug arrests in mexico city. *Journal of Applied Geography*, 30(2):263–270, 2010.

[160] H. Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer-Verlag, 2nd edition, 2003.

[161] R. Webster and M. Oliver. *Statistical methods in soil and land resource survey*. Oxford University Press, 1990.

[162] M. West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B*, 46(3):431–439, 1984.

[163] M. Wibrin, P. Bogaert, and D. Fasbender. Combining categorical and continuous spatial information within the bayesian maximum entropy paradigm. *Stochastic Environmental Research and Risk Assessment*, 20:423–433, 2006.

[164] C. K. Wikle. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394, 2003.

[165] C. K. Williams and D. Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[166] R. L. Wolpert and K. Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267, 1997.

[167] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *Eighteenth national conference on Artificial intelligence*, pages 217–223, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[168] M. Worboys and M. Duckham. *GIS: A Computing Perspective, 2nd Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2004.

[169] E. Wu, W. Liu, and S. Chawla. Spatio-temporal outlier detection in precipitation data. *KDD Workshop on Knowledge Discovery from Sensor Data*, pages 115–133, 2008.

[170] M. Wu, C. Jermaine, S. Ranka, X. Song, and J. Gums. A model-agnostic framework for fast spatial anomaly detection. *TKDD*, 4(4):20, 2010.

[171] M. Wu, X. Song, C. Jermaine, S. Ranka, and J. Gums. A lrt framework for fast spatial anomaly detection. *KDD*, pages 887–896, 2009.

[172] H. Yang, S. Parthasarathy, and S. Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 716–721, New York, NY, USA, 2005. ACM.

[173] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4):597–604, 2006.

[174] H. Zhang and S. Sheng. Learning weighted naive bayes with accurate ranking. *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 567–570, 2004.

[175] J. Zhao, C.-T. Lu, and Y. Kou. Detecting region outliers in meteorological data. In *Proceedings of the 11th ACM international symposium on Advances in geographic information systems*, GIS '03, pages 49–55, New York, NY, USA, 2003. ACM.

[176] J. M. Zytkow and J. Rauch, editors. *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99, Prague, Czech Republic, September 15-18, 1999, Proceedings*, volume 1704 of *Lecture Notes in Computer Science*. Springer, 1999.