# Statistical Geocomputing:
# Spatial Outlier Detection in Precision Agriculture

by

Peter Chu Su

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Environmental Studies
in
Geography

Waterloo, Ontario, Canada, 2011

# AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

# ABSTRACT

The collection of crop yield data has become much easier with the introduction of technologies such as the Global Positioning System (GPS), ground-based yield sensors, and Geographic Information Systems (GIS). This explosive growth and widespread use of spatial data has challenged the ability to derive useful spatial knowledge. In addition, outlier detection as one important pre-processing step remains a challenge because the technique and the definition of spatial neighbourhood remain non-trivial, and the quantitative assessments of false positives, false negatives, and the concept of region outlier remain unexplored. The overall aim of this study is to evaluate different spatial outlier detection techniques in terms of their accuracy and computational efficiency, and examine the performance of these outlier removal techniques in a site-specific management context.

In a simulation study, unconditional sequential Gaussian simulation is performed to generate crop yield as the response variable along with two explanatory variables. Point and region spatial outliers are added to the simulated datasets by randomly selecting observations and adding or subtracting a Gaussian error term. With simulated data which contains known spatial outliers in advance, the assessment of spatial outlier techniques can be conducted as a binary classification exercise, treating each spatial outlier detection technique as a classifier. Algorithm performance is evaluated with the area and partial area under the ROC curve up to different true positive and false positive rates. Outlier effects in on-farm research are assessed in terms of the influence of each spatial outlier technique on coefficient estimates from a spatial regression model that accounts for autocorrelation.

Results indicate that for point outliers, spatial outlier techniques that account for spatial autocorrelation tend to be better than standard spatial outlier techniques in terms of higher sensitivity, lower false positive detection rate, and consistency in performance. They are also more resistant to changes in the neighbourhood definition. In terms of region outliers, standard techniques tend to be better than spatial autocorrelation techniques in all performance aspects because they are less affected by masking and swamping effects. In particular, one spatial autocorrelation technique, *Averaged Difference*, is superior to all other techniques in terms of both point and region outlier scenario because of its ability to incorporate spatial autocorrelation while at the same time, revealing the variation between nearest neighbours.

In terms of decision-making, all algorithms led to slightly different coefficient estimates, and therefore, may result in distinct decisions for site-specific management.

The results outlined here will allow an improved removal of crop yield data points that are potentially problematic. What has been determined here is the recommendation of using *Averaged Difference* algorithm for cleaning spatial outliers in yield dataset. Identifying the optimal nearest neighbour parameter for the neighbourhood aggregation function is still non-trivial. The recommendation is to specify a large number of nearest neighbours, large enough to capture the region size. Lastly, the unbiased coefficient estimates obtained with *Average Difference* suggest it is the better method for pre-processing spatial outliers in crop yield data, which underlines its suitability for detecting spatial outlier in the context of on-farm research.

## ACKNOWLEDGMENTS

It is my pleasure to thank the people who made this thesis possible.

I want to give special thanks to my supervisor, Dr. Alexander Brenning, for his incredible input and help throughout the development of this thesis. Your teachings, good suggestions, and guidance have not only helped me to overcome many obstacles but also inspire me to love what I do as an academic and professional. This work would not exist without you.

Dr. Steven Roberts, Dr. Jonathan Li, and Dr. Yulia Gel deserve special thanks as my thesis committee members and advisors. In particular, I want to highlight Dr. Steven Roberts for getting me involved with academic activity, and for introducing me to the mathematics behind statistical techniques. Dr. Jonathan Li deserves special gratitude for introducing me to remote sensing concepts in class, in which I apply now as a professional. And Dr. Yulia Gel, my deepest appreciation for taking interest in my work.

I am grateful to Susie Castela and Lynch Finch for their administrative assistance and Scott MacFarlane for his technical support and for involving me to participate.

My gratitude extends to the Ontario Ministry of Natural Resources, particularly, Ian Smyth, Paul Sampson, and Gergin Naomouv from IMA, and Steve Leney and Kent Todd from WRIP for their attitude, encouragement, teachings, and benevolence.

I am indebted to UW alumni Myung Kyun Kim and Yan Chen, and my colleagues Alex Parisien and Andrei Balulescu for their gifts of support, unity, and friendship.

It is my family I thank last for everything else.

*To my father, Yam Hing;*

*my mother, Quiac Yuan;*

*my brother, Javier;*

*and my sister, Yuri.*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1:

# INTRODUCTION

The collection of yield data has become much easier with the introduction of technologies such as Global Positioning System (GPS), ground-based yield sensors, and Geographic Information System (GIS). Combine harvesters mounted with a yield sensor and a GPS receiver allows the collection of instantaneous crop yield data as the combine is harvesting the agricultural field. The result of this and other leading technological approaches has led to a new paradigm of agriculture, known as precision agriculture.

Precision agriculture is naturally information-intensive as it requires substantial layers of data in order to provide the necessary information for sound decision-making. The explosive growth and widespread use of spatial data in precision agriculture has challenged the ability to derive useful spatial knowledge, emphasizing the need for better data pre-processing. Particularly, spatial yield datasets obtained by combine harvesters mounted with ground-based yield monitoring sensors and GPS are affected by various random and systematic errors that occur because of natural topographic conditions, management-induced practices, and measurement error (Stafford et al., 1996). These errors need to be appropriately removed from the raw crop yield dataset in order to derive better spatial information.

After the collection of crop yield data, expert filtering software programs are used to remove yield errors. Expert filtering is a system that includes knowledge about the field, combine, crop, GPS, and other characteristics, which assesses the raw data and removes points that experts would not consider reasonable (Blackmore & Marshall, 1996). Expert filtering focuses on removing known systematic errors, which are well defined and described in the literature (Rands, 1995; Blackmore & Marshall, 1996; Nolan et al., 1996; Beck et al., 1999; Arslan and Colvin, 2002; Kleinjan et al., 2002; Sudduth & Drummond, 2007). On the other hand, stochastic errors from mostly unknown sources, commonly referred as yield surges or spatial outliers, are diminished according to the discretion of the analyst. These errors can be correctly removed, completely ignored, or incorrectly removed. In this work, crop yield point measurements that are substantially different than the neighbouring point measurements for the same agricultural field are considered to be spatial outliers.

The precision agriculture community utilizes local neighbourhood statistics to deal with these random errors, which involves the calculation of local statistics and determining outliers based on a moving window. Spatial outlier detection has also received a lot of attention from the data mining community. Data mining, particularly spatial data mining is the process of discovering interesting, previously unknown, and potentially useful patterns from large spatial datasets (Shekhar et al., 2005). In data mining, different spatial outlier detection algorithms have been elaborated and implemented to large spatial datasets such as traffic and census datasets. While both research communities implement similar techniques, the effectiveness of their techniques and the choice of parameters remain non-trivial. In addition, while most of the attention has been given to filtering data for yield mapping purposes, there has been little or no consideration regarding filtering data for the analysis of spatial yield data and possible consequences for decision-making based on statistical analyses in on-farm research. Although recognized as an important yet difficult process, to my knowledge, the study of spatial outlier effects in statistical modelling for site-specific management, particularly in modelling crop yield response functions, has been non-existent in agricultural studies.

## 1.1. Research Goals

The overall goal of this study is to assess the effects of outlying observations in yield datasets and their elimination strategies for site-specific crop management. More specifically, the objectives of this study are as follows:

1. To identify and provide an understanding of the importance of precision agriculture practices for site-specific management;
2. To examine the errors that are present throughout the collection phase of crop yield data;
3. To devise a framework for simulating crop yield data for testing purposes;
4. To examine existing spatial outlier detection techniques that are widely utilized for filtering erroneous crop yield data, and assess their performance via quantitative methods;
5. To examine the effects of outliers and their detection techniques for statistical modelling in a site-specific management context.

## 1.2. Motivation for Research

Spatial and non-spatial outliers and their detection techniques remain a popular research topic in the literature. Outlier detection has been studied as early as 1620 with the work of Sir Francis Bacon (Hadi et al., 2009), while spatial outlier detection gained popularity during the early 21st century. Currently, many spatial outlier detection techniques are available but there is no knowledge about which spatial outlier detection technique is better. Ver Hoef & Cressie (2001) state a problem in statistics is the misuse of statistical techniques: to use lesser statistical methods when more powerful methods are available. And this has been the case in the context of precision agriculture and data mining studies. Sudduth & Drummond (2007) state that there is no standard method for cleaning yield surges, although many different techniques have been suggested to address the specific error in applications of precision agriculture. Several global statistical techniques have been proposed and widely applied in the context of cleaning crop yield datasets (Shekhar et al., 2003; Lu et al., 2003). Global statistical tests remove extreme observations without considering their spatial arrangement, so they cannot detect spatial outliers but global ones. Local neighbourhood statistics became the standard approach to dealing with local errors and have also been widely utilized (Kleinjan et al., 2002; Simbahan et al., 2004; Ping & Dobermann, 2005). More elaborate spatial outlier algorithms have been proposed by the data mining community (Shekhar et al., 2003; Lu et al., 2003; Kou et al., 2006). However, these techniques have yet to be utilized for applications in precision agriculture.

Regardless, outlier detection in spatial data remains a challenge for various reasons. First, the choice of algorithm is non-trivial. Numerous spatial outlier detection techniques have been proposed to supersede previous techniques, but it is unclear whether the new algorithms are better. There is a lack of systematic comparisons of multiple algorithms as many authors have not attempted to compare new algorithms to earlier ones. In addition, the comparisons of spatial outlier detection techniques have been performed by qualitative methods. The current approach at assessing spatial outlier detection techniques is by ranking the top spatial outliers identified by each technique for a particular spatial dataset (Lu et al., 2003; Kou et al., 2006; Kou et al., 2007). However, ranking each detected outlier does not quantitatively measure the performance of each technique, especially when true spatial outliers are unknown.

Second, the choice of a spatial neighbourhood used to calculate the outlierness of an observation is also non-trivial. In all proposed local neighbourhood statistics, the shape of the neighbourhood is distinct. Thylen et al. (2000) and Bachmaier & Auerhammer (2004) utilize Euclidean metrics that result in a circular neighbourhood. However, the neighbourhood of Simbahan et al. (2004) and Ping & Dobermann (2005) resembles a cross band, “+”, with three succeeding and three preceding observations on each direction. Noack et al. (2003) neighbourhood is similar to a letter “H”, where the vertical lines correspond to the neighbouring harvest tramline. And the neighbourhood proposed by Bachmaier (2010) resembles a butterfly. In all cases, the number of neighbouring observations is left to the analyst's discretion. Innately, the definition of spatial neighbourhood affects the effectiveness of detecting true spatial outliers.

Lastly, false positives (swamping effects) and false negatives (masking effects) are not properly explored or treated, which implies most of the study on spatial outliers has been focused on detecting single point outliers. Spatial point outliers are single spatial outliers whose nearest neighbours are all non-outliers. Point outliers can create situations of false positive when an observation is wrongly identified as an outlier because it is surrounded by at least one true outlier. However, region outliers, which are spatial outliers that are clustered together, cause instances of not only false positives but also false negatives. In this particular situation, a true outlier is misclassified as a non-outlier because of the presence of true outliers in its surroundings that make it appear to be a normal observation. This case of region outlier remains largely unexplored in spatial data.

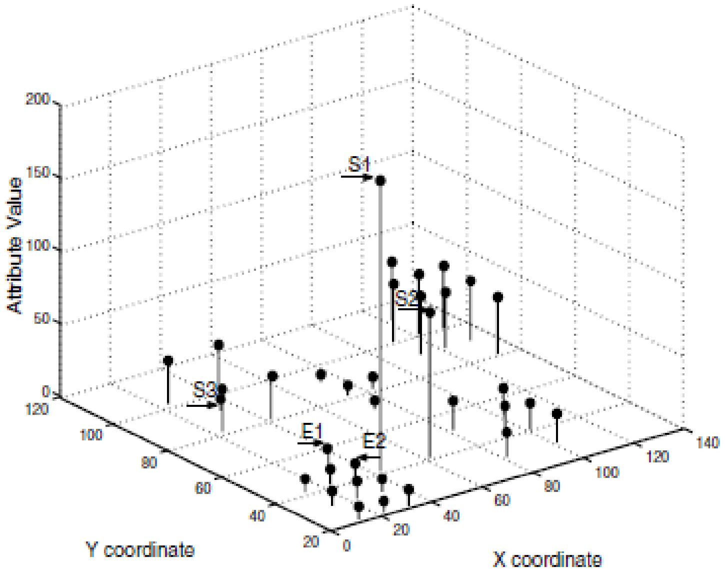


**Figure 1.1: Masking and Swamping Effects**

Neighbours of S3 are masked; their outlier score will be inflated because of the presence of outlier S3. Outlier E1 and E2 are swamped; their outlier score will be diminished because of the presence of the other outlier.

Source: Lu et al. (2003).

Many authors interested in the analysis of yield data for site-specific management, particularly for crop yield economic analyses such as Long (1998), Lambert et al. (2003) and Anselin et al. (2004) perform statistical analysis without outlier pre-processing steps. Others such as Anselin et al. (2004a), Lee et al. (2005), and Vrindts et al. (2005) overlook outliers by only removing extreme values without any focus on removing on local instabilities. Griffin et al. (2008) analyzed seven field-scale on-farm experiments conducted by farmers and concluded that yield data quality affects farm management decisions as five experiments would have led to different farm management recommendations depending upon whether the yield data were adjusted. However, case studies were used by Griffitn et al. (2008) because of insufficient farm management information available as relatively few farmers quantitatively analyze yield data. As such, a need exists to quantitatively determine whether removing outliers quantitatively affects the decision-making in a site-specific management context. As Hadi et al. (2009) notes, outlier detection is much similar to the 'chicken' and 'egg' problem. In order to obtain reliable model estimates, outliers need to be known in the data. But to know outliers in the data, model estimates should not be affected by outliers.

## 1.3. Structure of Thesis

The rest of the thesis consists of the following five chapters:

Chapter 2 provides an introduction to precision agriculture, followed by an overview of its technological components.

Chapter 3 presents a review of previous studies on outlier and spatial outlier detection, and reviews outlier-generating mechanism that are present in yield data.

Chapter 4 introduces the proposed methodology to evaluate spatial outlier detection techniques. This will include a detailed description of the outlier techniques under evaluation.

Chapter 5 presents statistical results for the performance assessment of outlier techniques, followed by a discussion of the findings.

Chapter 6 provides a summary and conclusions of the findings, a discussion about the suggested practices for cleaning yield datasets, and recommendations for future work.

Appendix A provides the results of the Shapiro-Wilk test statistic as a requirement for subsequent statistical tests that appear in Chapter 5.

Appendix B, similar to Appendix A, provides the results of the Brown-Forsythe test statistic as a requisite for subsequent statistical test that appear in Chapter 5.

Appendix C lists all acronyms used throughout this thesis.

Appendix D lists all spatial outlier detection techniques that appear in this thesis.

In this work, spatial outlier detection techniques are referred as spatial outlier detection algorithms. Throughout the text, each algorithm is distinguished by being capitalized and italics (see Appendix D). Similarly, written summary statistics such as the *mean* and the *median* will be in italics.

# CHAPTER 2:

# OVERVIEW OF PRECISION AGRICULTURE

## 2.1. Precision Agriculture: An Introduction

In the literature and industry, the term "precision agriculture" (PA) has been associated with many terms: "precision farming" (PF), "site-specific crop management" (SSCM), "site-specific management" (SSM), and "precision crop management" (PCM). All of these terms attempt to address the same revolutionary agricultural phenomenon that started approximately 30 years ago. In this work, the term "precision agriculture" will be referring to this agricultural phenomenon. The United States National Research Council (1997) defines precision agriculture as a management strategy that incorporates information technology for making decisions associated with agricultural crop production.

In precision agriculture, information technology and technological advances such as the Global Positioning System (GPS) and Variable Rate Technology (VRT) are used in order to geolocate the required information for spatial decision-making and apply the decision about the kind, location, and amount of agricultural input needed to match with the actual crop needs. In this regard, precision agriculture is not only the management and decision-making of labour, equipment, finance, production, but also of information.

The ultimate goal of precision agriculture is to improve economic returns and reduce environmental impact (Fountas, 2004). Economic return and environmental protection is not obtained by maximizing crop yield, but by managing and distributing agricultural inputs efficiently. Through precision agriculture, the farmer administers the exact amount of inputs needed at the exact location on the farm so the use of fertilizer and pesticides is reduced. The economic return is due to the fact that savings in inputs will offset the reduction of crop yield in the long term.

Environmental protection is enhanced by the use of optimal amounts of agricultural inputs, fertilizers and pesticides. The environmental impacts of input application in precision agriculture have not been extensively studied (Pedersen, et al., 2004). Nevertheless, the expected environmental benefits of precision agriculture include reduction of soil erosion, soil compaction, nitrate and pesticide leaching, and energy consumption, as well as focusing on habitat conservation and species protection (Werner, et al., 1999). For instance, Whitley et al. (2000) demonstrates the use of VRT in reducing ground water contamination while Schumacher et al. (2000) examines topographic data used with precision agriculture technologies for reducing erosion. However, the bulk of the literature on precision agriculture focuses on the quantity and quality of crop production, increased labour production, minimization of expenditure in resources, and production profitability (Yakushev et al., 2008).

Not all farms are suitable for precision agriculture because the economics of agriculture is affected by several internal and external factors. The most crucial internal factor for precision agriculture is the degree of spatial variation in the farm. A farm needs to be exhibit spatial variability, and this spatial variation of crop yield is the result of many complex anthropogenic, biological, edaphic, topographic, and climatic factors and their interaction among each other (Corwin & Lesch, 2010). The greater the spatial variation in farming data, the greater the potential of economic return for precision agriculture compared to conventional agriculture practice.

Conventional agricultural practice is based on managing the farm upon a hypothetical average condition, which may not exist anywhere in the farm. It involves constant input application, which is highly inefficient because some locations obtain inadequate inputs; others obtain excessive. In precision agriculture, every location in a farm that exhibits spatial variation can be evaluated according to its site specific characteristics and assigned an optimal input application rate unique to that location, so all locations in the farm obtain optimal inputs. However, if spatial variation is absent or insignificant, precision agriculture is ineffective and therefore not required.

External factors that affect precision agriculture include, but are not limited to, the different crops and their response at different nutrient levels, cost of fertilizer, pesticides and other inputs, market value of crops, and cost of spatial data, equipment, and labour (Lowenberg-DeBoer & Swinton, 1997). A determinant for optimal application rate is the cost of inputs and the crop market value (Havlin & Heiniger, 2009). Other major external factors are the cost of spatial data and VRT equipment.

Decision-making plays an important role for precision agriculture. Farming decisions can be classified into strategic, tactical, and operational decisions (Bouma, 1997). Strategic decisions deal with the overall management of the farm, focusing on issues with long-term consequences, usually 10 years or more. Tactical decisions deal with specific issues of farming, usually spanning 2 to 5 years. Operational decisions are made on a day-to-day basis. These day-to-day decisions include planting, fertilizer and pesticide application, harvesting, yield monitoring, and crop protection measures such as weed detection. Operational decisions are the focus of precision agriculture.

Both traditional agriculture and precision agriculture incorporate management and decision-making in day-to-day activities. The key distinction between both is the quality of information. Traditional agriculture relies on the farmer's mental information approach, whereby the information and knowledge is obtained by years of observation, experimentation, trial and error, and implementation (Davis, et al. 1998). This kind of information is subjective because it is derived from experience and belief. The information is not exact and prone to errors. With years of practice, the farmer will most likely know that spatial variability exists in the farm, and may administer inputs accordingly. However, the farmer does not know the exact magnitude of the variation, or the exact amount of inputs needed to achieve optimal results.

In precision agriculture, information is measured numerically. Advances in information and telecommunication technologies have enabled farmers to collect vast amounts of precise site-specific data with relative ease, and have provided powerful analytical tools for better farm management. This reduces uncertainty in the operational decision-making process (National Research Council, 1997; Blackmore, 2000b). Ultimately, farmers who incorporate better information to their practices are more likely to earn higher economic returns than farmers who do not.

Worldwide adoption of precision agriculture is mostly based on the level of general economic development, the level of government supporting agriculture, and the nature of the production unit (McBratney, et al., 2005a). The bulk of literature about precision agriculture mostly originates in developed countries with strong government support in agriculture such as USA, Japan, and the European Union (McBratney, et al., 2005a). Literature originating from Canada is limited, as the development of precision agriculture practices has been largely driven by technology innovation, private sector crop consultants, and equipment providers (Haak, 2010). Haak (2010) sent a survey to 14,000 Canadian farmers in 2006 and revealed that 23% of them use GPS equipment or products such as digital maps, with greater overall use reported in the Prairie Provinces and Ontario due to larger cropland areas. Out of this 23%, 78% use GPS as a tracking or guidance system on machinery to improve field operations; 50% for VRT input application; 32% for collecting spatial information for soil and crop management; and 4% for water management.

Nonetheless, the precision agriculture community in Canada is strengthening as academia and government has conducted research since the 1990s and provided funding incentives to producers since 2005 (Haak, 2010). As such, there is a strong presence of companies in Canada that provide a variety of precision agriculture services: Geonics Ltd is a worldwide company that provides electromagnetic instrumentation for non-invasive soil sampling; Prairie Precision Network provides differential GPS correction to Western Canada; DynAgra provides VRT service for fertilizer, herbicides, pesticides, fungicides, and insecticides; and companies such as SIGA, Landwise Inc, and Agri-Trend offer multiple services. In addition, indices initially suggested by Swinton & Lowenberg-DeBoer (2001) suggest that among all countries, Canada ranks first for overall suitability for precision agriculture based on a simple index about the number of hectares of cropland per worker. On average, environmental variation increases with area, therefore greater spatial potential for precision agriculture exists.

Given the identified potential for precision agriculture in Canada, the number of farming projects has been steadily increasing over the years. Haak (2010) reports approximately 9,000 precision agriculture projects funded by the National Farm Stewardship Program (NFSP) in Alberta, Saskatchewan, and Manitoba, totalling an amount of $34.5 million in funding.

## 2.2. Components of Precision Agriculture

### *2.2.1. Global Positioning System*

Perhaps the most important component for precision agriculture, the GPS system allows users to automatically determine their location anywhere on Earth, in real time, and while in motion. The GPS consists of a constellation of 24 satellites, a ground station, and a GPS receiver. Launched by the United States Department of Defence (U.S. DOD), the satellites orbit the Earth while broadcasting almanac information of two radio signals with different frequencies (Pfost et al., 1998).

The ground stations continuously adjust the almanac information for each satellite in order to reflect the actual orbit path. Due to gravitational forces from the Sun, moon and Earth, satellites are constantly pulled towards the Earth, causing minor orbital variations and substantial errors while determining the location of the receiver. And a GPS receiver, analogous to an AM/FM radio, receives the satellite signals and translates the almanac information to determine the position of the receiver.

Precision agriculture started during the late 1980s with the introduction of GPS technology into the agriculture sector. One of the first ideas was to mount a GPS receiver and a yield monitor onto a combine harvester (Searcy et al. 1989). While the combine is harvesting the farm field, a yield monitoring system is automatically recording yield at every one or two seconds, and the GPS receiver is obtaining positional information. The result of this combination is geo-referenced yield data. This data collection arrangement has provided empirical evidence of how farming data was spatially autocorrelated.

To obtain better GPS accuracy, signal correction is required. Differential correction, or DGPS, is a technique that adjusts the GPS signals to improve positional accuracy. Corrected GPS signals can achieve 1 to 3 metres accuracy, depending on calibration. Differential correction requires a static and a roaming GPS receiver. The static GPS receiver is placed on a location of known co-ordinates so the actual distance, the true range, between the static receiver and satellites is known and correct at all times. The pseudo-range, the distance between the static GPS receiver and satellites calculated by the static receiver, is a signal that contains the true distance and all the accumulated errors from the atmospheric condition. The difference between the true range and the pseudo-range is the differential correction (Figure 2.1).

The application of GPS technology has been fundamental to the development of precision agriculture. GPS receivers without differential correction can be used for crop scouting, which is an on-site assessment of crops made by farmers or other professionals. Crop scouting is usually required on a mixed-farm system, where a variety of species are grown on different fields.

DGPS is more valuable for precision agriculture as it is utilized for yield mapping, yield monitoring, and soil sampling, which are essential procedures for the characterization of spatial variability of the farm. Real-Time Kinematic GPS (RTK-GPS) is an emergent GPS technology that can be utilized for variable-rate fertilizer application down to centimetre-level accuracy. However, RTK-GPS is still relatively expensive, and requires expensive mapping software and highly accurate soil maps, yield maps, and treatment maps among other deliverables (Stafford, 2000).

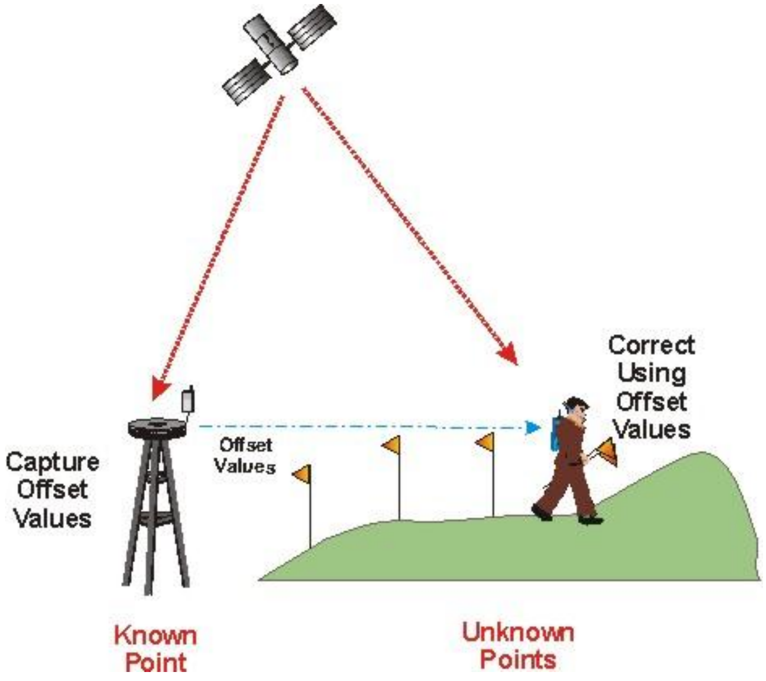


**Figure 2.1: The concept of Differential GPS correction**

Source: http://www.wirelessdictionary.com/Wireless-Dictionary-Real-Time-Kinematic-RTK-Definition.html

## *2.2.2. Yield Mapping*

Yield mapping is considered the initial stage of implementing precision agriculture (Blackmore, 1998). Yield mapping is the process of collecting geo-referenced crop yield data while the crop is being harvested (National Research Council, 1997). Yield mapping was first introduced by

Massey Ferguson in 1982 when a yield meter was mounted onto a combine harvester to obtain continuous yield measurements for the first time, although GPS technology was not available in 1982 (Oliver, 2010).

For yield mapping to work, crop yield per unit area must be determined at exact locations. Indirect methods of measuring crop yield include measurement of the combine engine speed or the torque of the tank filling auger, while direct methods include volumetric flow and mass flow measurements via proximal sensors (Stafford et al. 1996). Mass flow sensors are preferred due to the variation in bulk density and moisture content of volumetric flow sensors (Stafford et al. 1996). Mass flow sensors measure the crop mass as it enters the combine header. For mass flow sensors, grain yield can be calculated as:

$$Y = K \frac{M}{W \times S}$$

$Y$ is the instantaneous yield (volume per unit area), $M$ is the mass flow entering the combine (mass per unit area), $W$is the combine header width (the cutting width), $S$ is the travel velocity of the combine (distance per unit time), and $K$ is a conversion coefficient (Griffin, 2010). The suitability of DGPS over GPS is implied by the fact that the positional accuracy of yield measurement are required to be better than the width of the combine header. The combine header mixes grain across its width, which limits the spatial resolution of yield data up to the width of the header (Blackmore, 1998). Combine headers are approximately seven to eleven metres in width, which satisfies the DGPS accuracy requirements.

The result of yield monitoring is a yield map, which is a document that represents the spatial pattern of crop yield, and all the variables and side effects that were present during the plantation period (Blackmore, 2003). Yield maps are most commonly used for monitoring crop moisture and soil fertility, conducting on-farm experiments, and tile drainage management (Griffin, 2009).

Because some of the variables affecting crops change over time such as weather and nutrient levels, yield maps are only applicable for the survey year and should not be utilized for future years. This concept of variability is addressed in Blackmore & Larscheid (1997). They argue that besides spatial variation, temporal variability and predictive variability are important aspects for precision agriculture. Predictive variability refers to the difference between the farmer's prediction and the actual outcome (Blackmore & Larscheid, 1997). Temporal variability is identified when variables change over time, for example, crop yield has shown change over time.

Since yield mapping has become a less cumbersome process due to GPS and yield sensor technology, it is highly recommended that yield mapping is conducted during each plantation year in order to determine whether the observed yield is accredited to management practice or to environmental conditions (Blackmore, 1998).

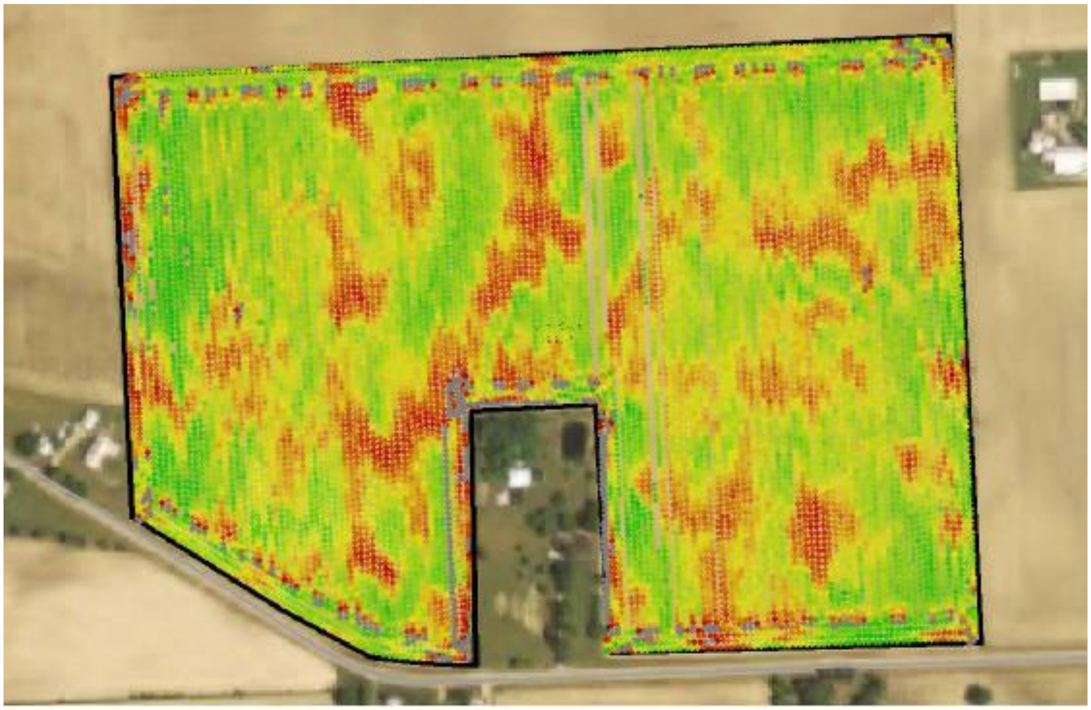


**Figure 2.2: Yield map overlaid on top of an aerial photograph**

Source: http://www.cropstarconsulting.com/id30.html

### *2.2.3. Soil Sampling*

After yield maps have been produced, there will be evidence whether the farm has enough spatial variability to implement site-specific nutrient and fertilizer management. If enough spatial variation exists, then further soil sampling is required for characterizing soil properties. Since soils are the medium for crop growth, characterizing and understanding the spatial variation of soil properties will enable the farmer to manipulate crop growth to meet their economic and environmental goals (McBratney & Pringle, 1998). The ultimate purpose of soil sampling for precision agriculture is to provide enough quality information in order to define management zones (MZ) for the application of inputs. Soil sampling allows farmer to determine the location and magnitude of fertilizer, lime, among other application input (Brase, 2006).

Traditional soil sampling techniques, grid soil sampling and directed soil sampling, are relatively expensive and intensive. When soil sampling is finished, samples need to be taken to laboratory for analysis. Laboratory analysis is required for functional characterization, which is the process of describing the samples in terms of their water regime and nutrient dynamics, as opposed to taxonomic characterization (van Alphen & Stoorvogel, 2000). When functional characterization is done, crop nutrient needs are derived for each soil sample.

The problem is that soil maps produced by field surveys are often not suitable for site-specific management although they are exploited in precision agriculture in practice. In the past, much of the information used in agriculture was coarse information based on field averages which is only adequate for uniform application and field-level management (Kerry et al., 2010). This soil information is not at the appropriate level of detail, and therefore does not have importance for explaining the variation of crop yield nor is useful for reaching the desired economic and environmental goals. This is not a surprise given that soil mapping was not intended for precision agriculture in the first place (van Alphen & Stoorvogel, 2000).

Ground-based proximal sensors can address the soil information needs of precision agriculture. These devices allows for non-invasive sampling as they can measure soil, plant, and crop information from within 2 metres distance from the soil surface (Corwin & Lesch, 2010). Proximal sensors fall into six main categories: electrical and electromagnetic, optical and radiometric, mechanical, acoustic, pneumatic, and electrochemical (Adamchuk et al., 2004). Electrical and electromagnetic sensors are perhaps the most utilized type of proximal sensors in precision agriculture. They include capacitance, electromagnetic induction (EMI), electrical resistivity (ER), and time domain reflectometry (TDR) sensors. Out of these, EMI and ER are the most common whithin-field level devices for soil mapping (Corwin & Lesch, 2005). EMI and ER measure the apparent soil electrical conductivity ($EC_a$).

Soil conductivity measurements are suitable for soil mapping because soil conductivity is highly correlated with soil properties (Pedersen, 2003; Corwin & Lesch, 2005; Kühn et al., 2009). The most cited EMI commercial device in the precision agriculture literature is the EM-38 conductivity meter. The advantage of the EM-38 is that it can be mounted onto a vehicle along with a GPS receiver for automatic and dense sampling (Figure 2.3). The sampling density can be approximately one sample every three metres or even less.

In addition, remote sensing imagery is increasingly being used as a non-invasive approach at soil sampling, particularly hyperspectral imagery (Personal Communication, Brenning, 2011). Before 1970s, aerial photographs were used for large-scale soil mapping, and subsequently, multispectral satellite imagery, such as Landsat TM, SPOT and AVHRR among others, provided the ability to map soils at a small scale, which is only applicable for regional soil mapping requirements (Manchanda et al., 2002). High resolution multispectral imagery such as IKONOS and QuickBird provide the resolution needs required by precision agriculture (Begiebing, et al., 2005). And satellite imagery such as Compact High Resolution Imaging Spectrometer (CHRIS) and Airbourne Visible Imaging Spectrometer (AVIS) provide hyperspectral information, as its narrower bands provide much more detailed information on crop and soil information (Begiebing, et al., 2005).

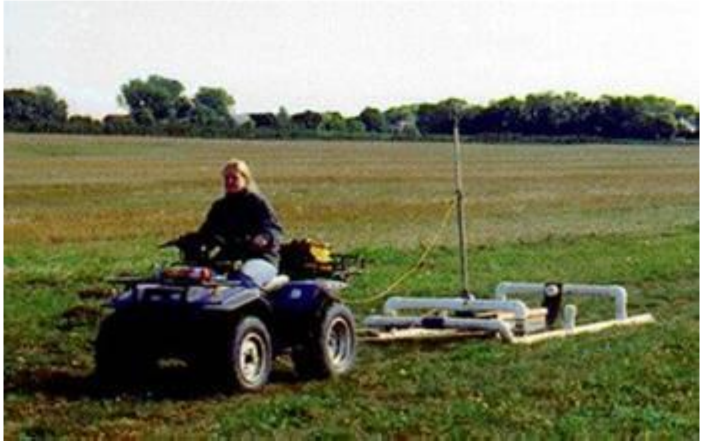


**Figure 2.3: Soil conductivity measurements with the EM-38 on an all-terrain vehicle**

Source: http://www4.agr.gc.ca/AAFC-AAC/display-afficher.do?id=1185562262407&lang=eng

### *2.2.4. Digital Soil Mapping*

Soil mapping is conducted after soil samples have been collected. Traditional soil mapping involved the grouping of continuous pedons together to form polygons representing an area with the same soil type (Rossiter & Hengl, 2002). This exercise requires a thorough knowledge of the soil-landscape model: the relationship between soil and landscape characteristics such as landform, vegetation, geology, and geomorphology (Dobos & Hengl, 2009). Such subjective requirements and the need for accuracy and uncertainty modelling leads to criticism of traditional soil mapping as being too qualitative in nature, especially for precision agriculture (McBratney et al., 2000).

With the emergence of computational statistics, GIS, GPS technology, and remote sensing data, various quantitative methods have been established and subsequently categorized in the emerging field of pedometrics (McBratney et al., 2000). Similarly, the availability secondary variables have aided soil surveyors to estimate soil variables based on these ancillary data (Hengl et al., 2007). This emergent extension of soil prediction has been known as digital soil mapping. Digital soil mapping (DSM) is defined as the creation and population of a geographically referenced soil database generated at a given resolution by using field laboratory observation methods coupled with environmental data through quantitative methods (McBratney & Lagacherie, 2004).

McBratney et al. (2003) proposed a generic framework for soil prediction known as the *SCORPAN* model:

$$S_a = f(s, c, o, r, p, a, n)$$

$$S_{cl} = f(s, c, o, r, p, a, n)$$

The estimated soil attribute value ($S_a$) and estimated soil class ($S_{cl}$) are a function of soil property ($s$), climate ($c$), organisms ($o$), relief ($r$), parent material ($p$), age ($a$), and position ($n$). Soil property ($s$) is usually referred as soil information from a previous soil map or prior expert knowledge. Note that position ($n$) and age ($a$) are implicitly stated in the equation.

Based on this definition and the *SCORPAN* model, DSM has three components: field observations ($S_a$ and $S_{cl}$), environmental variables ($s, c, o, r, p, a, n$), and quantitative methods ($f$). Field observations are obtained by soil sampling (reviewed in Section 2.2.3). In terms of environmental variables, the sources of data are becoming more available and accessible. Remote and proximal active and passive sensors along with pre-existing soil maps or expert knowledge give detailed information about soil properties. Particularly, climate information includes temperature, precipitation, and evapotranspiration, which are derived from remote sensing imagery or gauge measurements (McBratney et al., 2003). Information about organism can be obtained by vegetation, land-cover and land-use, and biomass and crop yield maps (McBratney et al., 2003). These maps are usually derived from remote sensing imagery or from ground measurements as in the case of yield maps.

Variables regarding relief are now mainly derived from digital elevation models (DEMs) (McBratney et al., 2003). These include primary terrain attributes such as slope, aspect, and curvature, while secondary attributes include topographic wetness index and incoming solar radiation among others. And parent material information can be mainly obtained from digitized geological maps (McBratney et al., 2003).

Various quantitative methods have been used to model the relationship between soil and environmental variables. These methods include linear models such as generalized linear models and generalized additive models, non-linear models such as decision tree classification and regression, support vector machines and artificial neural networks, fuzzy systems and expert-knowledge based systems, and geostatistical techniques such as ordinary kriging and co-kriging, among others (McBratney et al., 2003).

Digital soil mapping is a fairly new approach at solving conventional problems by incorporating quantitative techniques. The importance in precision agriculture is emphasized by the fact that efficient, cost-effective, consistent, and reliable techniques are used for the production of soil maps. A comprehensive review of digital soil mapping techniques can be found in McBratney et al. (2003). Their main message is that no singular quantitative technique is best for precision agriculture; all have substantial predictive power and inherent problems. It is the context that determines which particular method is selected.

### *2.2.5. Management Zones*

Management zones are defined as farm areas that exhibit relatively little variation in crop growth conditions (Bouma et al., 1999). The areas in each management zone are treated homogeneously, so application of inputs and decision-making in general, are unique to each zone. The main purpose of defining management zones is to limit the infinite variability of growth conditions throughout the field to a limited set for efficient management. Without this generalization, an extreme amount of zones would encourage the farmer to spend unnecessary time managing inputs, which may not earn him a higher net economic return relative to the committed time and effort.

A significant component of agricultural research has been directed towards delineating management zones. Generally, three factors affect the delineation of management zones: the quality of information, the procedures to process information, and the selection of the optimal number of zones (Fridgen et al., 2004). Many information sources have been postulated. The first information gathering approach is based on the farmers' mental information approach. This approach is subjective since it is based on experience through trial and error.

The second approach is by way of yield mapping successive years. This approach will allow farmers to identify areas where high and low yield occurs, zones where yield growth is most stable, and high grossing zones (Blackmore, 2000a). However, yield mapping alone may not successfully define management zones in terms of site-specific management because the dominance of a factor or a set of factors may change from season to season (Diker et al., 2004). Soil mapping is another information source for defining management zones because it integrates a host of soil physical and chemical properties. However, a large number of samples are required to define statistically significant management zones, which is labour intensive and expensive (Franzen et al., 2002). $EC_a$ measurements are an option for soil mapping. They are fast, relatively inexpensive, and have been used for delineating management zones (McBratney et al., 2005; Kühn et al., 2009).

A third information source is remote sensing imagery, which has been used for agriculture since 1929 (Seelan et al., 2003). It can provide information for the entire farmland without conducting sampling, and is perhaps the easiest and least expensive approach at obtaining spatially intensive farmland information over large areas. Remote sensing for precision agriculture is based on crop spectral reflectance, which can indicate the status of the crop (Seelan et al., 2003). Remote sensing imagery such as aerial photography and high resolution multispectral satellite imagery such as IKONOS and QuickBird are most appropriate for this type of precision agriculture application (Begiebing, et al., 2005).

However, drawbacks of remote sensing imagery, not only for management zone delineation but also for soil mapping includes high cost, dependence on weather and seasonal conditions, and represent static information. Nonetheless, remote sensing remains a viable technological advancement for precision agriculture. Moran et al. (1997) identify eight applications of remote sensing imagery in precision agriculture; in addition to, management zone delineation and soil mapping, they include: crop yield prediction, mapping seasonal variation, production of Digital Elevation Models (DEMs), pest and damage control, recognizing time-critical crop management applications, and mapping spatially-distributed information on climate and meteorological conditions.

The usage of more than one source of information for delineating management zone is highly desirable and practiced. The combination of farmer's experience, soil information and aerial photographs (Fleming et al., 2004), $eC_a$ maps and soil mapping (McBratney et al., 2005), $eC_a$ maps with topographical information (Kühn et al., 2009), satellite imagery and soil properties (Moran et al., 1997) are some examples of multi-information usage for defining management zones.

To process the information, classification schemes are utilized. Unsupervised classification is most applicable to management zone delineation because the analyst does not have a priori knowledge regarding the labels of the outcome management classes. In particular, fuzzy *k*-means clustering has been utilized to delineate management zones (Odeh et al., 1992; Fridgen et al., 2000; Song et al., 2009; Zhang et al., 2010). One particular advantage of the fuzzy *k*-means clustering approach is the ability to optimize the number of classes by deriving two measures: the fuzziness performance index (FPI), and the normalized classification entropy (NCE) Odeh et al. (1992).

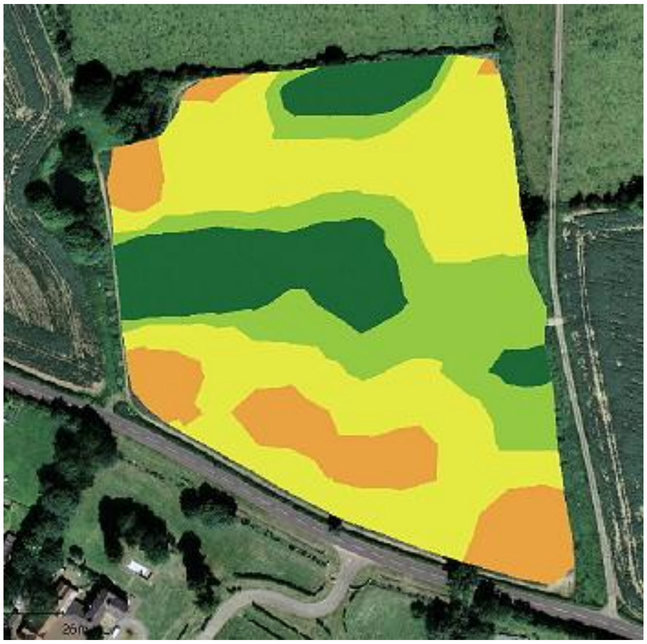


**Figure 2.4: Managament Zones overlaid on top of Google Maps**

http://www.wnif.co.uk/articles/385/1/New-Holland-Precision-farming-systems-for-any-tractor-brand/Page1.html

## *2.2.6. Variable Rate Technology*

Input application will be uniform for areas within the management zones, but vary between management zones. This is all possible with variable rate technology (VRT). VRT, arguably one of the most critical components in precision agriculture, allows agricultural inputs such as fertilizers (nitrogen, potassium, and phosphorous), seeding, pesticides and herbicides, liming, and tillage, to be applied on-the-go throughout the field at appropriate rates according to the pre-set application map (Virin et al., 2008). The application map is loaded onto a computer mounted on a tractor with GPS, fertilizer spreader, speed sensor, and an actuator (Virin et al., 2008). As the tractor is moving, the computer locates the position of the tractor in relation to the application map, and the actuator directs the spreader controller to change the amount or kind of inputs (Virin et al., 2008). Lesser amounts of inputs are applied to areas where they are not needed in excess for optimal crop growth, and saved for areas in the field that need greater amounts.

The two most common VRT fertilizer spraying systems are the centrifugal spreader and the pneumatic boom spreader (Pedersen, 2003). The resulting spread pattern of the centrifugal spreader is about 24 to 36 metres, with a spatial distribution similar to an inverted boomerang with considerable overlap (Pedersen, 2003). The pneumatic spreader uses various nozzles, four to eight on each side, attached to a boom, which are controlled via air flow. The length of the spread is about 18 to 24 metres, and the spread area obtains high uniformity (Scottish Natural Heritage, 2009).

The benefit of VRT is the proper distribution of inputs, which has the potential to reduce environmental impacts, and improve economic returns and crop quality (Pedersen, 2003). For most crops, nitrogen (N) is the most important nutrient, and the right amount at the right place, right time can improve crop yield dramatically. However, inappropriate N application can result in leaching, denitrification, volatilization, and immobilization (Hatfield, 2000). In the case of N-leaching, nitrogen is washed away by excess water, either caused by rainfall or excess irrigation. This runoff can enter nearby biological systems such as lakes or wetlands and can cause eutrophication.

Phosphorous (P) and potassium (K) are more stable nutrients than nitrogen because they are easily held by soil particles (Pedersen, 2003). The precision of their application to the field is not as critical as nitrogen. However, the current practice is the use of pre-mixed NPK fertilizers that also contain essential macro- and micronutrients. Pre-mixed fertilizers allow farmers to efficiently handle and distribute inputs. However, to lessen the environmental impacts of N, P, and K, their ideal application should be separate.

From an economic standpoint, VRT has demonstrated to be mostly profitable. Lambert & Lowenberg-DeBoer (2000) reviewed 108 economic studies of different VRT implementation (VRT N, VRT PK, VRT NPK, VRT pH, VRT seeding, etc) and 69% of them reported positive net returns, 12% indicated negative results, and the remaining 19% indicated mixed results. VRT negative net returns can be associated with insufficient or inappropriate quality of information (Bullock et al., 2002).

Pederson (2003) sent a survey to farmers from Denmark, United Kingdom, and United States about their experiences with precision agriculture technologies. VRT of fertilizers was the most cited practice that would increase profits, either by VRT of phosphorous and potassium, or VRT of phosphorous alone. However, a major drawback about the economics of VRT is the inability to quantify all the benefits and costs in a comprehensive manner. However, many of the 108 studies reviewed in Lambert & Lowenberg-DeBoer (2000) did not consider costs such as information and data collection, labour and time, training, technology, and environmental impact.

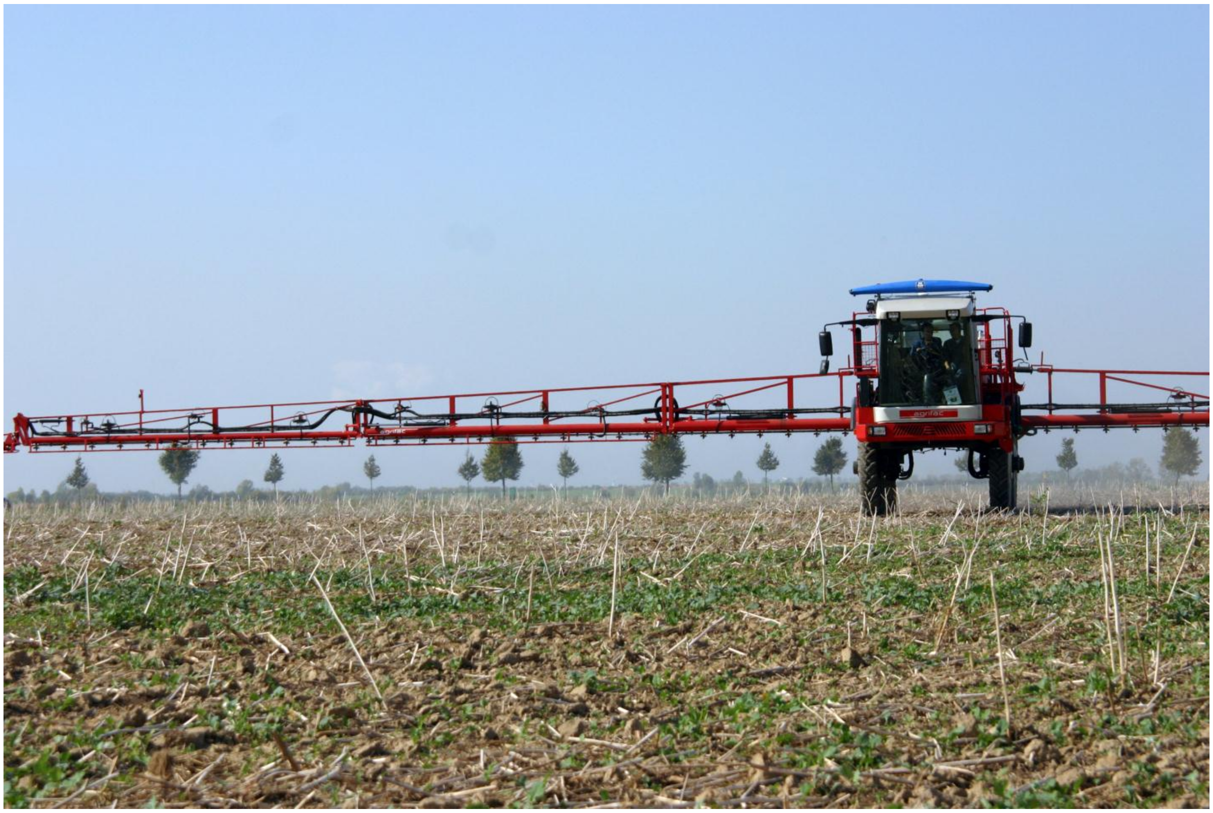


**Figure 2.5: Example of VRT for Pest management**

Source: http://www.agricon.de/en/company/downloads/photos-of-n-sensor/

## 2.3. Chapter Summary

Precision agriculture is a management strategy that incorporates information technology for making decisions associated with agricultural crop production. Precision agriculture deals not only with the management of labour, equipment, finance, production, but also of information. Overall, precision agriculture requires a relatively large field area with enough spatial variability within the fields, a good farm management system already in place, and relatively low market cost of inputs, information, equipment, labour, and specialized skills. In addition, several components such as GPS, VRT, yield sensors, soil sampling and mapping, and management zones, have to be established for a successful agricultural regime.

# CHAPTER 3:

# OUTLIER DETECTION

## 3.1. Outlier Detection: An Introduction

Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Similarly, Barnett and Lewis (1994) state that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Consensually, outliers are a minority of observations that are different from the majority of the observations in a dataset. The majority, referred as the in-lying observations, therefore, consists of at least 50% of the observations of the total dataset that share the same common characteristics, while the remaining outlying observations are different from this common characteristic. Spatial outliers may share the same common characteristic with the remaining data; they are just different in comparison to the characteristics of their spatial neighbours.

Hawkins (1980) identifies two mechanisms by which outliers are generated. The first mechanism is a long-tailed distribution. Depending on the shape of the distribution, observations that arise from the tails of the distribution are considered to be erroneous observations. Barnett and Lewis (1994) refer to the tailed observations as extreme observations, and declaring them as outlier would depend on how they appear in relation to the distribution model. Note that an outlier is always an extreme or relatively extreme observation in the sample, but an extreme observation may not always be an outlier but a form of natural variation in the dataset. The second outlier-generating mechanism is that the data comes from at least two distributions. The in-lying observations come from one distribution while the outliers come from a second distribution. In this mechanism, observations from the second distribution are said to be contaminants of the first distribution (Hawkins, 1980; Barnett & Lewis, 1994).

There are generally three types of outliers: point, contextual, and collective outliers (Chandola, et al., 2009). Point outliers are data instances that are inconsistent with respect to the rest of the dataset. Consider for instance, crop yield measurements with a calibrated mass flow sensor. Let the dataset be repeated yield measurements of the same bulk of yield. A point outlier, or point outliers would be the instance or instances in which the mass flow sensor improperly measured the bulk.

Contextual or conditional outliers are data instances that appear inconsistent to the rest of the data in a specific context, but not otherwise (Song et al., 2007). These outliers are defined by two sets of attributes: contextual and behavioural attribute. The former is used to determine the context in which outliers are assessed, and the latter is any attribute that is tested for outlierness. Defining the context is of particular importance. An observation may be an outlier in a given context, a normal observation given a different context. For example, consider a new house in an established neighbourhood. This house can be considered a contextual outlier in terms of age as its behavioural attribute, but not necessarily in terms of size or in terms of city-wide distribution of the ages of residential houses.

A collective, region, or cluster outlier is a group of observations that are clustered together which have low variance among them but are inconsistent to the rest of the dataset. Unlike point outliers, collective outliers can only occur in sequential datasets, for instance, time-series and spatiotemporal datasets (Chandola, et al., 2009). On the other hand, point or collective outliers can be contextual outliers by defining the contextual attribute in which outliers are assessed.

Point outlier is the simplest type of outlier and is the focus for the majority of research in outlier detection community (Chandola, et al., 2009). Many outlier detection techniques have been proposed as early as in the 19th century, such as Peirce's criterion, Chauvenet's criterion, Grubbs' test, and in the mid-20th century techniques such as Tukey's box plot and Hampel's test (Barnett & Lewis, 1994). This collection of outlier techniques are referred as discordancy tests or distribution-based techniques.

The general idea of discordancy tests is to fit the data set to a known distribution, and develop a test based on the distribution properties, and observations which deviate from the model assumptions are identified as outliers. Discordancy tests rely on the assumption that the data distribution is known, that observations are identically and independently distributed (i.i.d.), that the distribution parameters are known, and that the number of expected outliers are known beforehand (Barnett & Lewis, 1994).

Discordancy tests are unsuitable when model assumptions are not met. Particularly, assumptions are violated for data mining datasets. These datasets are usually of unknown distribution which are high-dimensional and with very large number of observations. Several collections of non-parametric data mining techniques have been proposed, including distance-based, density-based, clustering-based, and depth-based techniques (Preparata & Shamos, 1988; Knorr & Ng, 1997; Breunig et al., 2000; Acuna & Rodriguez, 2004).

Outliers in spatial data are point and collective outliers that occur in a spatial framework. In other words, spatial outliers are a form of contextual outliers, whereby the contextual attribute would be the spatial attributes, for example geographic co-ordinates or spatial relationship such as distance or adjacency. The behavioural attribute would often be a non-spatial attribute, for example, tons per hectare of agricultural yield. Previous data mining techniques are not able to detect spatial outliers with their current definition, as they would identify global extreme observations as spatial outliers (Shekhar et al., 2003).

## 3.2. Spatial Outlier Detection

The identification of outliers took on a new direction with Shekhar et al. (2003) introducing the notion of "spatial outlier", or S-outlier. Previous research in outlier detection focused on the identification of "global outliers" relative to an entire sample. The outlier definition provided by Hawkins (1980) and Barnett and Lewis (1994) are appropriate only for global outliers. Spatial outliers on the other hand, are contextual outliers formally defined as spatially referenced observations whose non-spatial attribute values are significantly different from those of other spatially referenced observations in its spatial neighbourhood (Shekhar et al., 2003). Spatial outliers represent local instability because the outlier observations are extreme relative to its neighbours, even though they may not be markedly different from the entire population (See Figure 3.1).

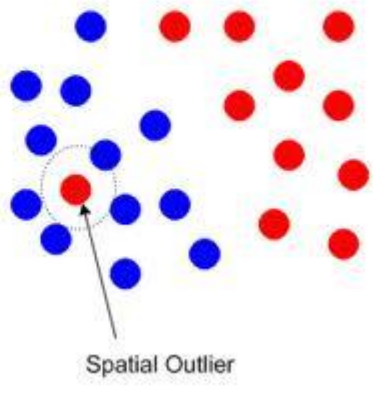


**Figure 3.1: Example of a discrete spatial outlier**

Source: http://fanaee.wordpress.com/2011/05/10/spatial-data-mining/

Shekhar et al. (2003) proposes a unified definition of “spatial outlier”, stating that various statistical techniques for outlier detection in a spatial context can be expressed within this general framework. They include two sets of S-outlier tests: graphical and quantitative. Graphical S-outlier methods are based on the visualization of spatial data to identify spatial outliers. They include the *Variogram Cloud* and the *Moran Scatterplot*. Quantitative methods are based on statistical test to distinguish between spatial outliers from the remainder of the dataset. They include the *Scatterplot*, also known as linear regression, and *Spatial Statistic Z*. All algorithms are introduced formally in Chapter 4.

Lu et al. (2003) identifies a major drawback in Shekhar et al. (2003) general framework of spatial outlier detection techniques: swamping and masking effects are not considered or suppressed when defining the aggregate neighbourhood function. Depending on the spatial relationship of outliers, true outliers can be ignored while in-lying observations can be incorrectly flagged as outliers. The former is referred as masking effect, or false negative classification, while the latter is a known as swamping effect, or false positive classification. Lu et al. (2003) propose three S-outlier algorithms to minimize swamping and masking effects: *Iterative Z, Iterative R, and Median Z* algorithm.

Lu et al. (2003) compare *Iterative Z, Iterative R, Median Z, Spatial Z, Scatterplot*, *and Moran Scatterplot* with a synthetic dataset. Their result *shows Iterative Z, Iterative R,* and *Median Z* successfully identify the top three outliers in the dataset, while *Scatterplot, Moran Scatterplot*, and *Spatial Z* incorrectly flagged in-lying observations as outliers due to masking and swamping. However, the synthetic dataset had a small population size of 36 observations, with a total of three spatial outliers and two global outliers. Additionally, the detection exercise was performed without replication, which does not provide a measure of reliability.

Also, Lu et al. (2003) compare the algorithms on an experimental dataset based on various non-spatial attributes of the U.S Cities compiled by the U.S. Census Bureau. They rank the top 10 spatial outliers detected by each algorithm. The results show the outlierness rank of each City is different for each algorithm, noting that eight spatial outliers are detected by their proposed algorithms but in different order. Chen et al. (2009) update *Median Z* by proposing the use of the median and median absolute deviation instead of the mean and standard deviation for the normalization of differences. They compare *Spatial Z* with *Median Z* on a West Nile virus dataset to identify the top seven counties with West Nile cases. Their results indicate that the top-ranked spatial outliers are different for each algorithm.

Spatial autocorrelation is formally introduced to spatial outlier detection techniques by Kou et al. (2006). Tobler's first law of Geography notes that observations which are closer to each other are most similar than observations farther apart, as "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Previous algorithms ignore the idea that neighbours closer to true spatial outliers have more impact in the calculation of the spatial outlier test statistic. Kou et al. (2006) propose two algorithms: *Weighted Z* and *Averaged Difference*, or *AvgDiff*. *Weighted Z* is simply *Spatial Z* with the neighbourhood aggregated function being calculated by how close the neighbours are to the observation. *AvgDiff* is based on the average absolute difference between an observation and each of its neighbours. Kou et al. (2006) compare *Spatial Z*, *Weighted Z*, and *AvgDiff* using real dataset on Counties infected by West Nile virus. They select the top 30 spatial outliers, which accounts for 1% of all the 3,109 counties. The results show the top-ranked spatial outliers are different for each algorithm.

Chawla and Sun (2006) explore the characteristics of spatial autocorrelation and heteroscedasticity with their measure of spatial outliers: *SLOM* or Spatial Local Outlier Measure. Spatial autocorrelation is accounted for by $\bar{d}(o)$, which is a measure similar to the *Spatial Z* Algorithm $f_{diff}\ (o)$. $\bar{d}$ and $f_{diff}$ represents the distance (Euclidean) between the non-spatial component of object $o$ and its nearest neighbours. The only difference is that $\bar{d}(o)$ factors out the effect of a neighbour *p,* which has the maximum difference between observation *o* compared to all of *o's* neighbours. The benefit of using $\bar{d}$ instead of $f_{diff}$ is that if *o* is indeed an outlier, then $\bar{d}$ will amplify the effect of *o* in its neighbourhood; however, if *p* is not an outlier but a neighbour of *o*, then $\bar{d}(p)$ will be suppressed. $\bar{d}(o)$ behaves much like a trimmed mean.

In the SLOM algorithm, heteroscedasticity is accounted for by a parameter that captures the net variation within a neighbourhood. The idea is that outliers should be more prominent in neighbourhoods with little variation than in neighbourhoods with more variation. Chawla and Sun (2006) compare SLOM against *Spatial Z* with a synthetic dataset. Their synthetic dataset consists of a 10 x 10 data matrix. The non-spatial attribute values were simulated with a Gaussian generator. The locations of some values were changed in order to create a cluster of similar values and a spatial outlier in the centre of the cluster. They also use SLOM on a real dataset compiled by the U.S Census Bureau to detect the top five counties with the highest proportion of people identified as a minority group. The main criticism that can be applied to their simulation study, besides the use of a small dataset and lack of replication, is that no indication was provided whether spatial autocorrelation was included with the Gaussian simulation.

Kou et al. (2007) propose a graph-based approach to detecting spatial outliers. Their motivation is threefold: (1) to minimize masking and swamping, (2) to evaluate region outliers instead of singular outliers and (3) to avoid normalization across the entire dataset. Masking and swamping can lead to erroneous identification of outliers. A similar concept tied to masking and swamping is region outliers. If a region outlier is present, S-outlier algorithms will mask outlying observations and swamp in-lying ones (Lu et al., 2003). Further, the normalization across the entire dataset may be inappropriate for datasets consisting of a number of spatial clusters, with spatially correlated observations in the cluster, while observations in other clusters have no direct correlation (Kou et al., 2007).

Their approach involves a graph where each observation is connected to its *k* nearest neighbours, creating a network. The magnitude of the connection is the absolute difference between observations, so outliers will tend to have larger connections. The algorithm starts by clearing the largest connections until an observation or a region is disconnected from the entire network. This is repeated until *m* spatial outliers are identified. Kou et al. (2007) compare *Spatial Z*, *Scatterplot*, *Moran Scatterplot*, and their graph-based algorithm, based on rental information for each U.S. city. Their objective is to identify the top 10 outliers. Their results indicate that the top spatial outliers detected for each City is different for each algorithm.

## 3.3. Crop Yield Errors and Outliers

While the previous section reviewed the research on spatial outlier detection by the data mining community, this section addresses the possible sources of error and outliers in precision agriculture datasets. In precision agriculture, most of the research has been conducted on errors in yield data due to its importance for site-specific crop management. Yield datasets often contain several errors that arise from a combination of known and unknown sources. These sources of errors can be classified into natural, management, and measurement error (Stafford et al., 1996). Natural sources of error include climate, topography, and soil-landscape features, and site characteristics. For example, poor weather condition affecting the crop growth during a single farming season. These sources are uncontrolled factors that cannot be changed by the farmer, and therefore, cannot be removed from the dataset. The farmer can only identify the factors that were present during the growing season.

Management sources of error are random events that usually occur in small areas due to management decisions, for example, poor crop establishment, inadequate fertilizer or herbicide application, among others, or due to stochastic events such as equipment handling errors (Stafford et al., 1996).

Measurement error is the third source of error and is of particular interest. Measurement error in yield data has been the most studied of the three. This error is further classified into: sensor, positional, and operational errors (Arslan and Colvin, 2002).

Sensor errors are related to the yield sensing mechanism, such as the actual accuracy of the sensor, the sensor response, improper calibration, and grain flow delay (Blackmore & Marshall, 1996; Arslan and Colvin, 2002). Unless multiple harvesters are used in one field (personal communication, Brenning, 2011), these errors mostly affect the entire dataset, and therefore, they are not corrected but acknowledged. However, for yield mapping and analysis of yield data, grain flow delay has to be corrected. Grain flow delay is the time it takes for the crop to move from the cutter bar to the grain tank where the yield flow sensor is located (Blackmore and Marshall, 1996). This delay offsets the position of the observations by a time delay of about 10 to 14 seconds, depending on the combine model, speed, incline, and load (Nolan et al., 1996; Sudduth & Drummond, 2007).

Positional error is the error introduced by the GPS receiver due to calibration, atmospheric condition, measurement noise, signal loss, or any other similar limitation (Rands, 1995). The result is that yield points are incorrectly located, which includes points outside of the field boundary or points that are too far apart (Rands, 1995; Beck et al., 1999). Positional errors are in practice resolved by removal (Rands, 1995). Points outside the field boundary are easy to identify; they are deleted if they do not fall within the field boundary. Points too far apart are identified with a maximum distance threshold, which is derived with knowledge of the combine maximum speed, the time interval between points and the GPS resolution (Rands, 1995).

Operational errors are error introduced to the value, not the location, of the measurements by certain operations during the measurement activity. According to Murphy et al. (1994), Rands (1995), Blackmore and Marshall (1996), Kleinjan et al. (2002), Beck et al. (1999), and Sudduth and Drummond (2007), the sources of operational errors include the following:

1. start-pass & end-pass delay
2. combine header up
3. break-in operations
4. unknown crop width entering the header
5. changes in combine speed

Start-pass and end-pass delays are errors that are always present when measuring yield. Start-pass delay is the error introduced when the combine enters all the tramlines. As the combine starts harvesting at the beginning of each tramline, grain flow storage is not full and takes time to fill up, so yield is underestimated at the beginning of each tract (Figure 3.2). Similarly, as the combine finishes a tramline, the cutting mechanism stops, but the header has not been raised yet. This is commonly referred as end-pass delay, which overestimates yield. These two errors are easy to identify because they are at the beginning and end of each tramline. Start-pass and end-pass delay are estimated to be less than 40 seconds (Thylen and Murphy, 1996; Nolan et al., 1996).

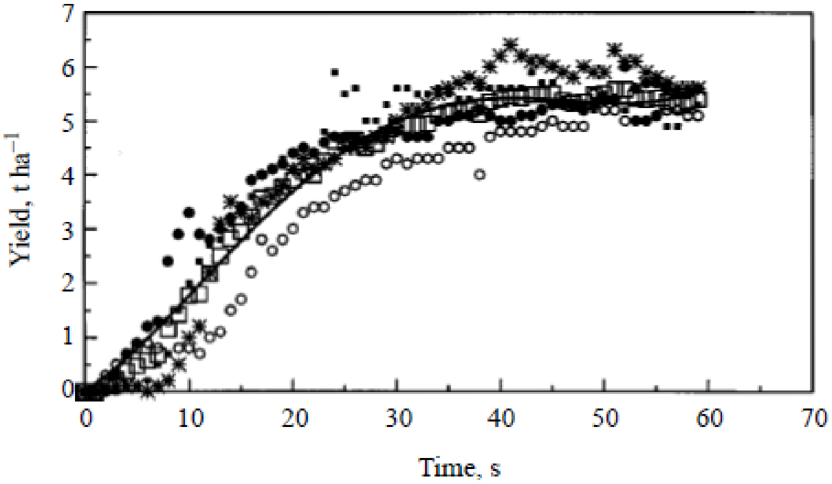


**Figure 3.2: Example of Start-pass delay for yield data logged for the first 60 s of four harvester runs**

Source: Thylen & Murphy (1996).

Another error that is easy to identify is when the combine header is up while the yield sensor is active. Since no crop is entering the combine, yield measurements when the header is up is always zero or small values.

Break-in operations, also known as overlaps, are errors that occur when the combine travels to previously harvested areas with the combine header down and with an active yield sensor. Break-in operations occur when the combine was not able to completely harvest the area due to acute angle turns, narrow lands, or obstacles on the way such as electric posts (Beck et al. 1999). The combine has to return to these areas and harvest the missing crops. The problem is the underestimation of yield in the first and subsequent passes (Figure 3.3).

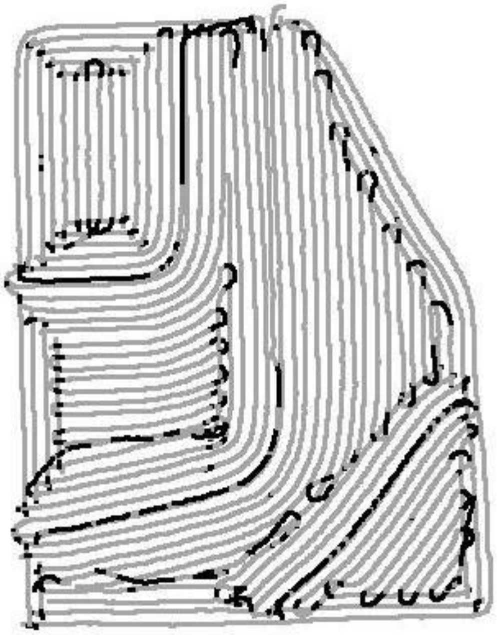
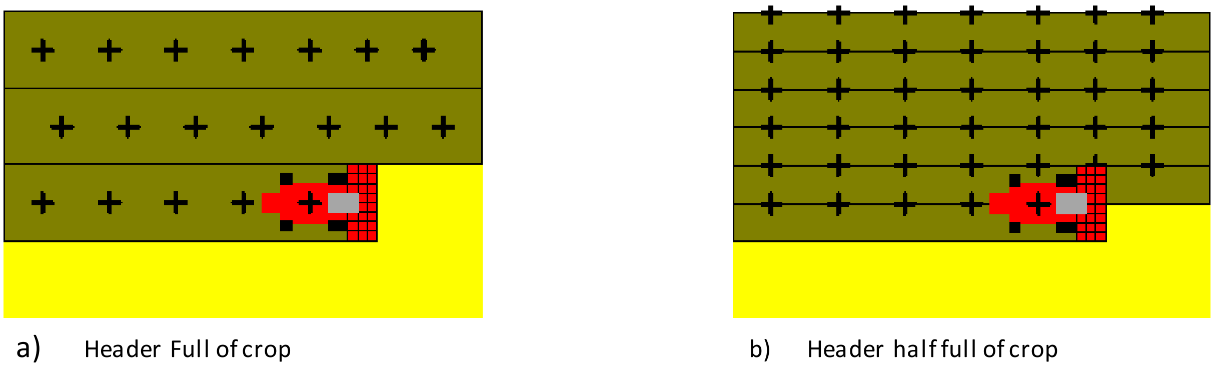


**Figure 3.3: Example of break-in operations (highlighted) in a sorghum field**

Source: Beck et al. (1999).

An error similar to break-in operations is the error of not knowing the crop width as it enters the combine. In the equation to determine yield (p. 11), $Y$, involves the parameter $W$, the width of the combine header. The equation assumes that the crop width is constant and equal to the combine header at all locations. Problems with yield measurement arise because the combine header does not always have a full width of crop entering it. This has been acknowledged as a major problem in yield data collection (Stafford et al., 1996). Yield is underestimated proportionally to the width of the harvested crop. For example, if the combine harvest the entire field with half of the cutter width, then twice as many points will be recorded than if combining with a full cutter width. The problem is that each point will be underestimated by 50% (Figure 3.4), thus, any summary statistic or yield mapping via interpolation will be significantly underestimated.


a) Header Full of crop

b) Header half full of crop

**Figure 3.4: Example of Unknown crop width**

Source: Blackmore & Marshall (1996)

Changes in the combine speed cause erroneous measurements. Again, in the equation of yield (p. 15), if the speed is too slow approaching 0, then the area being harvested will approach 0. So, the grain mass divided by 0 will result in infinite yield, which is incorrect. Similarly, sudden changes in speed introduce errors to the observations. High acceleration or deceleration occurs because of rough changes in the topography or during sharp turns.

## 3.4. Outlier Detection in Yield Datasets

All sources of errors affect yield measurements by creating unrealistic measurements of yield moisture, grain flow, speed of combine, and/or position. Yield minimum and maximum are unrealistic compared to the crop yield's biological potential. Similarly, yield surges, which are the abrupt change of yield values, are widely present in erroneous datasets. Manual filtering by an expert is the common approach at treating erroneous yield observations. The expert starts with identifying and where possible, correcting or removing points affected by primary errors that are known in advance such as combine header up, start- and end-pass delay, grain flow delay, and positional errors. Combine header up errors are dealt by removing yield measurements equal to zero. Start-pass and end-pass delay correction removes the first and last twelve observations, which is about 40 seconds, from each tramline. Grain flow delay assigns a positional shift of approximately 14 seconds to all observations, and positional errors are dealt with deleting the points outside the field and points that are separated by more than a distance threshold.

Secondary filtering attempts to remove errors caused by combine operations, yield sensing, and uncertain values due to localized and extreme variation (Ping & Dobermann, 2005). These errors are removed by using several global statistical tests. Lee et al. (2005) and Vrindts et al. (2005) utilize the frequency distribution of the observations to delete erroneous extreme values. Anselin et al. (2004a) create an outlier percentile map that displays six categories for classification of ranked observations. Outliers are found in the lowest, 0-1, and highest, 99-100, percentile and are labelled as outliers. Robinson and Metternicht (2005) declare yield surges as observations that are outside the lower (upper) quartile – (+) 1.5 times the interquartile range. Similarly, Sudduth and Drummond (2007) identify yield surges as observations that are outside a standard deviation interval.

Local neighbourhood statistics have also been widely utilized and are standard practice. Thylen et al. (2000) identify yield surges as any measured value that falls outside the mean yield of 10 nearest neighbours plus or minus a threshold of acceptance. Kleinjan et al. (2002) advise local outliers as exceeding ± 3 standard deviations within a user-specified moving block. Similarly, Beck et al. (1999) uses the average mean of a moving window composed of 25 nearest neighbours. If the observation falls outside the ± 3 standard deviations, then it is declared a local outlier. Simbahan et al. (2004) and Ping & Dobermann (2005) utilize local inverse distance weighting to detect local outliers (see Table 3.1).

Examples of expert filters are in Rands (1995), Kleinjan et al. (1998), Beck et al. (1999), Simbahan et al. (2004), Ping and Dobermann (2005), and Sudduth and Drummond (2007). And Table 3.1 provides a summary of the secondary filters applied in crop yield data by the precision agriculture literature.

**Table 3.1: Summary of Secondary Filtering**

| Global Methods | Outlier threshold | Neighbourhood | Examples |
|---|---|---|---|
| Histogram (Grubb's Test) | ± 2 or 3 standard deviations;<br>1st & 99th percentile | N.A. | Lee et al. (2003); Anselin et al. (2004a); Vrindts et al. (2005) |
| Boxplot (Tukey's Test) | ± 1.5 interquartile range | N.A. | Robinson & Metternicht (2005) |
| Local Methods | Outlier threshold | Neighbourhood | Examples |
| Beck et al. (1999) | ± 3 standard deviations | 25 nearest neighbours | N.A. |
| Thylen & Algebo (2000) | ± 2 standard deviations | 10 nearest neighbours | N.A. |
| Kleinjan et al. (2002) | ± 3 standard deviations | 30ft by 30ft neighbourhood | N.A. |
| Noack et al. (2003) | undefined | Adjacent tracks resembling an "H" | N.A. |
| Simbahan et al. (2004);<br>Ping & Dobermann (2005) | ± 2 or 3 standard deviations | 3 neighbours in the North, South, East, West direction, resembling a "+" | N.A. |

## 3.5. Chapter Summary

Outliers are observations that deviate so much from other observations as to arouse suspicion that they were generated by a different mechanism (Hawkins, 1980). Spatial outliers, on the other hand, are spatially referenced observations whose non-spatial attribute values are significantly different from those of other spatially referenced observations in their spatial neighbourhood. They are generated under two mechanisms: local extreme observations and contamination from another distribution.

Sources of yield error include natural, management, and measurement. Measurement error is further divided into sensor, positional, and operational, with much of the research emphasis on operational sources of error. Several statistical spatial outlier techniques have been proposed by the data mining community, although the standard approach at removing errors have been via filtering algorithms, either globally or locally, proposed by the precision agriculture community. While the precision agriculture community has not set out to verify detected outliers, the data mining community has investigated them by ranking the top outliers in real datasets or conducting experiments with synthetic datasets composed of small population and lack of replication.

# CHAPTER 4:

# METHODOLOGY

## 4.1. Introduction

The proposed framework for determining the effects of outliers and the effectiveness of spatial outlier detection algorithms is unique among the previous studies reviewed in Chapter 3. The proposed approach is to utilize a simulated spatial dataset with known characteristics and errors known in advance. Simulation is the approach that is often used in statistical literature to assess novel methods as it allows generating datasets with known and controllable properties with an arbitrary replication (Personal Communication, Brenning, 2011). Unlike the approaches reviewed in Chapter 3, a real dataset should not be used to determine whether an algorithm performs better than another because spatial outliers are really not known. In real datasets, spatial observations whose non-spatial attributes significantly deviates from their spatial neighbours can be either real spatial outliers, i.e. observations in a spatial framework that were indeed produced by a differing mechanism, or simply due to the inherent (natural) variability of the spatial data. Algorithms for spatial outlier detection or expert knowledge cannot distinguish between such data properties.

In addition, knowing exactly the characteristics of spatial datasets also allows the effects of the spatial outliers to be determined with great precision because no coefficients have to be estimated from the data. In real experiments, the treatment effects are superimposed onto the natural variability of the data, causing parameters to be unknown (Ver Hoef & Cressie, 2001). Furthermore, because each spatial outlier is known in the dataset, the assessment of spatial outlier algorithms can be conducted as a binary classification problem composed of an outlier and a non-outlier class. Instead of ranking the top outliers of each spatial outlier in the dataset as conducted by the previous studies and making comparisons between algorithms, effective performance measures available for classification problems can be utilized. Lastly, replication has to be emphasized in order to obtain reliable results, as the reliability of results must be inferred from multiple datasets that inherit the same data collection procedures, processes, and environmental variables.

Thus, the idea is to use a geostatistical simulation technique to generate a dataset with known characteristics (refer to section 4.2.1). After the simulation, contaminated datasets are created by randomly adding errors to the simulated the dataset (refer to section 4.2.2). Ten spatial outlier techniques that have been widely used either in data mining or in precision agriculture literature will be compared and assessed with respect to how well they detect the errors in these contaminated datasets (refer to section 4.3). And to determine the effects of spatial outlier algorithms in statistical modeling, each algorithm will be used as a pre-processing step prior to estimating crop yield response function.

All statistical analyses: unconditional simulation, detection of spatial outliers, performance assessment of each algorithm, and modelling crop yield response function are performed with the R statistical language (R Core Development Team, 2010). R is a free language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, and is easily extended via the addition of free packages available on the internet through the Comprehensive R Archive Network (CRAN).

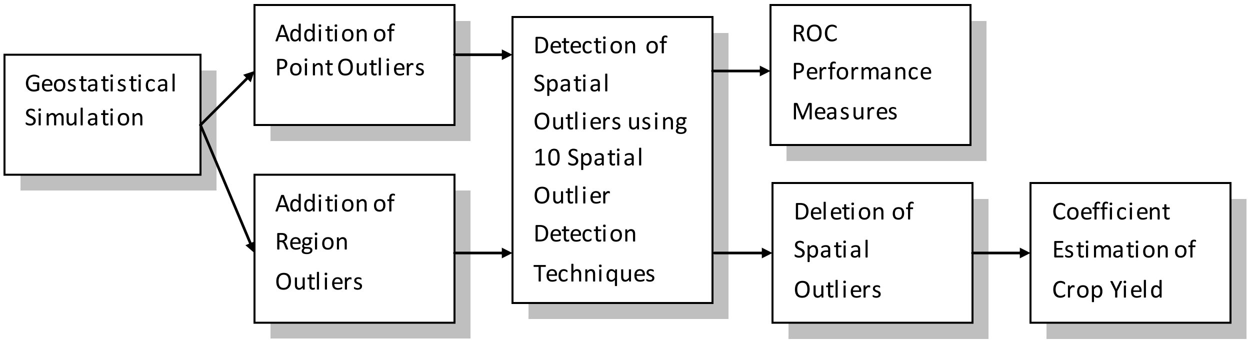

Geostatistical Simulation
Addition of Point Outliers
Addition of Region Outliers
Detection of Spatial Outliers using 10 Spatial Outlier Detection Techniques
ROC Performance Measures
Deletion of Spatial Outliers
Coefficient Estimation of Crop Yield

**Figure 4.1: Workflow of Methodology**

## 4.2. Spatial Data Generator

### *4.2.1. Simulation of On-Farm Experiments*

To obtain spatial data that conforms to the objectives and the specifications mentioned above, a stochastic simulation approach must be employed. Stochastic simulation is the process of selecting random numbers from a predefined probability distribution (Webster & Oliver, 2007). Geostatistical simulation, a particular form of stochastic simulation, is a popular set of techniques that can be used to reproduce spatial variation and uncertainty that is present in precision agriculture datasets.

This simulation design is based on Brenning et al., (2008). Yield point measurements are simulated for a hypothetical on-farm experiment with three treatments on a Gaussian random field $Z(x) = m(x) + e(x)$ on a rectangular 40 ha field (400 m by 1,000 m). An on-farm experiment is a scientifically valid research method to test species varieties, products or equipment performance under specific conditions. The setting of an on-farm experiment is the random application of a treatment in a field to obtain statistical evidence on the treatment effect (Top Crop Manager, 2007). A 40 ha field is relatively common in Southern Ontario farms. The sampling density consists of 50 strips along the length of the farm with 400 data points for each strip, a total of 20,000 raw points. This sampling density is consistent with farming machinery. For each strip, combine harvesters can obtain one sample every one metre and a half or less. For simplicity, the sampling interval is increased to one sample every one metre. The separation distance between strips is usually about 17 m. For simplicity, this distance is increased to 20 m.

The Gaussian random field consists of spatially correlated residual random field $e(x)$ and a deterministic trend modelled as:

$$m(x) = \ a + d_1 f_1(x) + d_2 f_2(x) + \ g_1 t_1(x) + g_2 t_2(x)$$

$f_1$ and $f_2$ represent spatially varying environmental variables, $t_1$ and $t_2$ are 0 and 1 indicator variables indentifying the farmer's treatment over the farm. $a$ is the average crop yield of the farmer's standard treatment approach, and $g_1$ and $g_2$ represent two innovative site-specific management practices. When both $t_1$ and $t_2$ are equal to 0, the farmer's standard treatment was applied, in this case, uniform application of agricultural inputs. These three treatments are applied to 12 alternating blocks, each block containing four strips per treatment, with the 12th block containing two additional strips for a total of 50 strips. $f_1$, $f_2$, and $e$ are simulated unconditionally with sequential Gaussian simulation with mean value of 0 and a spherical semivariogram model. $f_1$ and $f_2$ have a sill of 1, nugget 0, while $e$ has a partial sill of 70 and a nugget value of 3.5 bushels per acre, which represents a 5% relative nugget effect. All three variables have an autocorrelation range of 150 metres.

In this simulation model, $a$ is set equal to 76 bushels per acre, which is consistent with the production of winter wheat (*Triticum aestivum)* in Southern Ontario for the 2009 season (Ministry of Agriculture Food & Rural Affairs, 2009). The effect of environmental variable $f_1$ is set to increase crop yield by 6 units, while innovative practice 1 was is set to increase crop yield by 3 units. Environmental variable $f_2$ and innovative practice 2 are set to have no effect on crop yield. Therefore, the yield model equates to:

$$Z(x) = 76 + 6f_1(x) + 0f_2(x) + \ 3t_1(x) + 0t_2(x) + \ e(x) \qquad \text{(Equation 1)}$$

An unconditional sequential Gaussian simulation is utilized to generate the spatial data. Sequential simulation is widely used and computationally feasible method for simulating continuous variables (Gebbers & Bruin, 2010). Each value is simulated following a random path according to its conditional cumulative distribution function (ccdf), which is determined at each location (Webster & Oliver, 2007). Unconditional simulation is utilized because no initial sample data is available to be conditioned upon. Thus, all distributional characteristics of the simulated data are known, and no initial assumptions have to be made, which allows the testing of any statistical or computational techniques; in this case, to evaluate whether one technique is better than other techniques in a wide variety of situations.

The unconditional sequential simulation algorithm for point simulation is as follows (Gebbers & de Bruins, 2010):

1. Specify the coordinates of the points at which simulation is to be conducted
2. Prescribe the experimental semivariogram
3. Determine the random path in which the points will be simulated
4. Simulate values at each point:
    a. At each un-sampled location, simple kriging with the model semivariogram are used to estimate the sample mean and variance. The estimate will be based on the previously simulated data within a specified search radius or consisting of *n* neighbouring observations.
    b. Use the estimated kriging mean and kriging variance to model the Gaussian cumulative distribution function at the location to be simulated
    c. Draw a random value from the distribution function and insert the value to the point
    d. Proceed to the next un-sampled point of the random path, and repeat from *a* to *c* until all points have been visited

Unconditional sequential Gaussian simulations are conducted with R statistical language (R Core Development Team, 2010). Package *gstat* (Pebesma, 2004) is an R package that provides basic functionality for univariate and multivariate geostatistcal analysis. *gstat* uses sequential simulation algorithm as its default geostatistical simulation platform because it is versatile, efficient, and suitable for very large datasets (Pebesma, 2004).

In order to generate fast and correct simulations, one technicality had to be modified; that is, the a moving window radius for local kriging, or $r_{max}$. $r_{max}$ finds the number of *n* neighbouring observations used for the kriging mean and kriging variance estimate at each single un-sampled point. By default, *gstat* uses all observations. However, this setting significantly slows down the simulation process given that *n* = 20,000. Pebesma (2004) recommends setting the $r_{max}$ value no smaller than the range of autocorrelation, in this case, 150 m. To be more conservative, $r_{max}$ value is set to 400 m.

The simulation is then replicated 20 times in order to obtain a measure of uncertainty. However, because stochastic simulation requires the generation of a large set of random numbers, random number produced by modern computer algorithms are pseudo-random numbers because true random numbers are very difficult to obtain (Gebbers & Bruin, 2010). Pseudo-random number generators (PRNG) are algorithms that generate deterministic series of numbers that are sufficiently similar to random numbers following a uniform distribution. Pseudo-random numbers depend on an initial number, a “seed”, and using the same seed will reproduce the same sequence of numbers. Thus, simulation can be repeated if the seed is known. In this work, each simulation is given a unique seed number multiplied by a constant in order to mimic a truly random set of simulations and achieve reproducible results. Figure 4.2 summarizes the yield simulation procedure.

```
Initialize Yield.Simulation Script
  Create.Grid(x = 50, x.separation = 20, y = 400, y.separation = 1)
  N = Number.of.Simulations
  FOR (i in N){
      Set.Seed(i * Constant)
      Simulate.Env.Variables(formula = trend ~ 1, location = ~ x + y, sill = 1, model = Spherical, range = 150, nugget = 0, beta = 0, n.simulations = 2, nmax = 400)
      f1 = Simulation1; f2 = Simulation2
      Create.Block.in.XCoord(t0, t1, t2, Blocks = 12, Alternating)
      Simulate.Yield.Variable(formula = trend ~ 1 + f1 + f2 + t1 + t2, location = ~ x + y, sill = 70, model = Spherical, range = 150, nugget = 0.05, coefficients = (76, 3, 0, 3, 0), n.simulations = 1, nmax = 400)
  } END FOR
  End Script
```

**Figure 4.2: Yield Simulation Procedure in R-pseudo code**

### *4.2.2. Addition of Spatial Outliers*

Since the simulated fields do not have any errors and no outliers have been added so far, any spatial outliers identified in these simulated fields with any spatial outlier algorithm would be erroneous and could be attributed to the “natural” variability among simulations. Spatial outliers are added once simulated yield measurements are generated. The idea is to randomly select a percentage of the population, add or subtract a substantial error term to the yield value, and label them as spatial or global outliers. Global outliers could be generated by adding and substracting a large error term to a large simulated value. Two scenarios of spatial outliers are used.

The first scenario is the addition of individual point spatial outliers, which are random points in the field that are contaminated. A small percentage of the simulated observations are randomly selected, and these points are further divided randomly into two groups of equal size. In one group, an error term is added to the yield measurement while in the other, the error is subtracted from it. If any of these resulting contaminated yield measurements are greater than the maximum of all original yield values or smaller than the minimum of all original values, then they are labelled as global outliers, otherwise, they are referred as spatial outliers. The outlier term is simulated from a Gaussian distribution with a mean value of two times the nugget (7.0 bu/acre), and with a standard deviation of 1 bu/acre (see Figure 4.3).

```
Initialize Single.Outlier Script
  M = SOutlier.Amount
  Pool.SOutliers = Random.Sample(Population, Size = M)
  Positive.SOutliers = Random.Sample(Pool.SOutliers, Size = M/2)
  Negative.SOutliers = Difference.Between(Pool.SOutliers, Positive.SOutliers)
  Contaminated.Yield[Positive.SOutliers] = Original.Yield + Gaussian(mean = 2*Nugget, st.dev = 1)
  Contaminated.Yield[Negative.SOutliers] = Original.Yield - Gaussian(mean = 2*Nugget, st.dev = 1)
  IF (Contaminated.Yield > Original.Yield OR Contaminate.Yield < Original.Yield ){ THEN "Global.Outlier"
          ELSE "Spatial.Outlier"
  } End IF
  End Script
```

**Figure 4.3: Procedure for adding Point Outliers**

The second scenario involves the addition of region outliers, which are groups of contaminated observations clustered together at random locations. For the number of simulated observations N, given a set cluster size *G* for the number of spatial outliers in a region, random points are selected from the $N - G + 1$ uncontaminated observations. For each random point, the point observation and the next $G - 1$ observations are set as region outliers. This is accomplished by generating a Gaussian error term for each of the observations in a region, and adding or subtracting the error term, as proposed in the single outlier scenario. All observations in a region will have either a positive or negative error, and similarly, the labelling of global and spatial outliers is conducted. The result is a spatial dataset of agricultural yield measurements with a known number of spatial outliers that are clustered together randomly in the dataset (see Figure 4.4).

```
Initialize Region.Outlier Script
M = SOutlier.Amount
G = Region.Size
Pool.Seeds = Random.Sample((Population – G + 1), Size = M/G)
FOR (i in Pool.Seeds){
    R.Outliers = Population[Pool.Seeds[i]:Pool.Seeds[i + Region.Size]
    IF (Random.Number(from 0 to 1) > 0.5){ THEN Contaminated.Yield [R.Outliers] = Original.Yield +
            Gaussian(mean = 2*Nugget, st.dev = 1)
            ELSE Contaminated.Yield [R.Outliers] = Original.Yield - Gaussian(mean = 2*Nugget,
            st.dev = 1)
    } End IF
} End FOR
IF (Contaminated.Yield > Original.Yield OR Contaminate.Yield < Original.Yield ){ THEN “Global.Outlier”
    ELSE “Spatial.Outlier”
} End IF
End Script
```

**Figure 4.4: Procedure for adding Region Outliers**

## 4.3. Detection of Spatial Outliers

Once the simulation is conducted and spatial outliers are added to the dataset, each algorithm will be used for spatial outlier detection. The following section provides a description of all spatial outlier algorithms used for detection. Given the diverse notation encountered in the literature, a need exists to provide a unified notation to describe all spatial outlier algorithms. This section fulfills this need by providing a unified notation to describe spatial outlier algorithms. The key publications from which these algorithm are drawn include works by Han & Kamber (2001), Shekhar et al. (2003), Lu et al. (2003), Simbahan et al. (2004), Ping & Dobermann (2005), Chawla & Sun (2006), Kou et al. (2006), and Chen et al. (2008). In terms of naming convention, algorithm names will be simplified in the text as follows (see Table 4.1): *Spatial Statistic Z* will be referred as *Spatial*; *Median Statistic Z* as *Median*; *Local Area Mean* as *Local; Scatter Plot* as *Scatter*; *Spatial Local Outlier Measure* as *SLOM*; *Weighted Z* as *Weighted; Inverse Distance Weighting Interpolation* as *IDWP; Kriging Interpolation* as *Kriging or Krige; Averaged Difference* as *AvgDiff;* and *Spatial Outlier Test* as *SOTest*. Appendix D provides a full list of naming conventions of all the spatial outlier algorithms reviewed in this work.

Notation:

The following notation conventions are used in the sequel:

$S$: an ordered set representing the entire dataset; all locations in the spatial domain. For example, $S = \langle x_1, \dots, x_n \rangle$. Ordered set, i.e. $\langle x_1, x_2, x_3 \rangle \neq \langle x_3, x_2, x_1 \rangle$

$x$: two-dimensional scalar; location of an observation in the spatial domain, $S$.

$k$: integer; number of nearest neighbours

$rk(x,S) = rank\big(distance\,(x,S)\big)$: an ordered set of size $S$ relative to the distance away from $x$, i.e. $rk(x,S) = \langle x, w, y, v, \dots, b, a \rangle$

$N(x) = \langle y \in S: 1 < rk(y, S) < k + 1\rangle$: ordered set of size $k$, representing locations that are neighbours of $x$, excluding $x$

$f(x)$: scalar; attribute value at location $x$

$f_{aggr}(x)$: scalar or vector; an algorithm-specific aggregate function that summarizes the attribute at locations $N(x)$

$f_{diff}(x)$: scalar or vector; a comparison function between $f(x)$ and $f_{aggr}(x)$

$z(x)$: scalar; spatial outlier score for location $x$

$\mu$ = mean of a vector

$sd$ = standard deviation of vector

$MAD$ = median absolute deviation of a vector

Nine of the most popular statistical spatial outlier algorithms are used on the spatial data created with the above process to detect the spatial outliers that were introduced. These include five algorithms that do not account for spatial autocorrelation, and four algorithms that do account for spatial autocorrelation. A proposed novel spatial outlier algorithm, noted as Spatial Outlier Test (SOTest), is formulated as an exploratory exercise using the principles of the spatial algorithms reviewed in here. These 10 algorithms include:

**Table 4.1: Spatial Outlier Detection Algorithms**

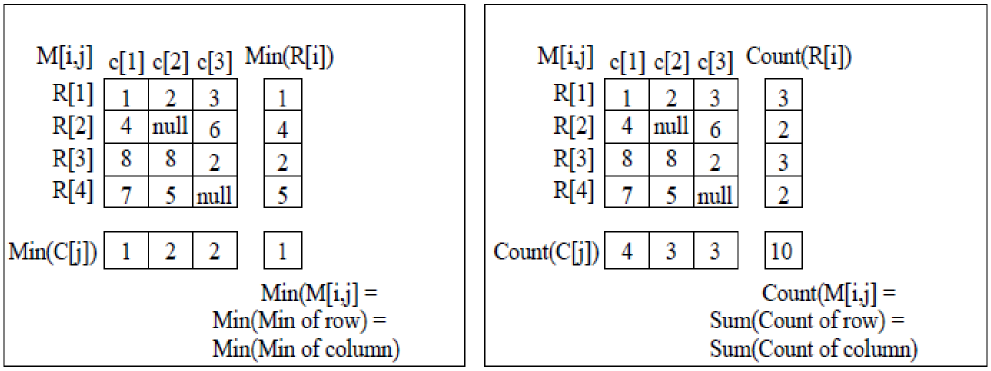
| Without Spatial Autocorrelation | With Spatial Autocorrelation |
|---|---|
| Spatial Statistic Z | Inverse Distance Weighting to the Power (IDWP) |
| Median Statistic Z | Kriging Interpolation |
| Scatter Plot | Weighted Z |
| Local Area Mean | Averaged Difference (AvgDiff) |
| Spatial Local Outlier Measure (SLOM) | Spatial Outlier Test (SOTest) |

All spatial algorithms are based on similar principles: compare the attribute value at each location against an aggregrate function that summarizes the neighbourhood attribute values. This comparison is then normalized across the entire dataset, and observations with the highest outlier score are considered more likely to be spatial outliers than observations with low score. In this work, the attribute refers to crop yield, but these algorithms for spatial outliers are general in nature and apply, in principle, to other numerical spatial variables.

For this spatial data configuration whereby observations are point measurements, the neighbourhood $N(x)$ can be defined as either consisting of the $k$ nearest neighbours (*k-NN*) of $x$ (excluding $x$ itself) according to the Euclidean distance in the two-dimensional spatial domain, or via a search *radius,* i.e. as consisting of all points within a Euclidean distance from $x$ equal to $r$. *k-NN* is used to define $N(x)$ and the subsequent measures for all algorithms because it is the most common among the two in this context (Shekhar et al., 2003; Lu et al., 2003; Kou et al., 2006). *k-NN* always guarantees the same number of neighbours for each $x$, unlike search radius. *k-NN* is useful especially for spatial data that is evenly spaced, while search radius is more suitable for unevenly-spaced spatial data because it can filter out observations that are so far away that they may not be considered actual neighbours.

After defining $N(x)$, an aggregate function, $f_{aggr}(x)$, is computed to summarize the attribute values of $N(x)$. Such function can be classified as distributive, algebraic, or hollistic (Han and

Kamber, 2001). Distributive aggregate functions are functions that can be applied to each partition of the dataset that would be identical as applying the same function to all the data without partition. *count, max, min* are some examples of distributive aggregate functions. For example Figure 4.4 shows that the *min* and *count* of the entire dataset is the same irrespective of whether the dataset was partitioned based on columns or rows.



| M[i,j] | c[1] | c[2] | c[3] | Min(R[i]) |
|---|---|---|---|---|
| R[1] | 1 | 2 | 3 | 1 |
| R[2] | 4 | null | 6 | 4 |
| R[3] | 8 | 8 | 2 | 2 |
| R[4] | 7 | 5 | null | 5 |
| Min(C[j]) | 1 | 2 | 2 | 1 |

Min(M[i,j] =
Min(Min of row) =
Min(Min of column)

| M[i,j] | c[1] | c[2] | c[3] | Count(R[i]) |
|---|---|---|---|---|
| R[1] | 1 | 2 | 3 | 3 |
| R[2] | 4 | null | 6 | 2 |
| R[3] | 8 | 8 | 2 | 3 |
| R[4] | 7 | 5 | null | 2 |
| Count(C[j]) | 4 | 3 | 3 | 10 |

Count(M[i,j] =
Sum(Count of row) =
Sum(Count of column)

**Figure 4.5: Example of Distributive Agreggate Function: Minimum & Count**

Source: Shekhar et al. (2001).

Following the classification and notation of Han & Kamber (2001), algebraic aggregate functions are functions that can be computed using a constant number of distributive aggregate functions for each data partition. *average, standard deviation, variance* are examples of algebraic aggregate function. In the case of average aggregate function, it can be computed with two distributive functions: *sum divided by count*. Holistic aggregate functions on the other hand, are functions that cannot be computed with a constant number of distributive aggregate functions. *median, rank, mode* are some examples of holistic aggregate functions. After deriving $f_{aggr}(x)$, $f_{diff}(x)$ is computed by comparing $f_{aggr}(x)$ to $f(x)$. Such comparison is usually by way of computing their difference, but can also be computed as a ratio, among other measures (Lu et al., 2003). In this work, the arithmetic difference between $f_{diff}(x)$ and $f_{aggr}(x)$ will be used throughout the detection of spatial outliers. Finally, $f_{diff}(x)$ is normalized by finding the centre and spread of $f_{diff}$.

A brief technical description of the spatial algorithms is as follows. This description is based on the references indicated above in Table 4.1.

### 4.3.1. Spatial Statistic Z

$$f_{aggr}(x) = \mu\left(f\left(N(x)\right)\right)$$

$$f_{diff}(x) = f(x) - f_{aggr}(x)$$

$$z(x) = \frac{|\, f_{diff}(x) - \mu\left(f_{diff}(S)\right) |}{sd(f_{diff}(S))}$$

For *Spatial Statistic Z* (Shekhar et al., 2003), $f_{aggr}(x)$ is calculated by first ranking the neighbouring observations of $x$ based on Euclidean distance, and then selecting the $k$ observations that are ranked the highest, excluding $x$. $f_{aggr}(x)$ calculates the mean attribute value of neighbours of $x$, written as $\mu\left(f\left(N(x)\right)\right)$. $f_{diff}(x)$ subtract the attribute value of $x$ with the mean attribute value of its neighbour. This is repeated for all observations in the spatial domain, and the outlier score for $x$ is found by standardizing $f_{diff}(x)$ across the entire dataset, $S$. Note that $f_{diff}(S) = \langle f_{diff}(x_1), f_{diff}(x_2), f_{diff}(x_3), \ldots, f_{diff}(x_n)\rangle$

Most of the computation time is allocated to the calculation of $f_{aggr}$. The $rank$ operation is similar to $sort$, which has on average a quadratic time complexity (Knuth, 1998). Given that $f_{aggr}$ is a computation within a $for$ loop that runs over all observations, the time complexity of $f_{aggr}$ is increased to at least cubic runtime. $f_{diff}$ applies one basic operation outside the $for$ loop. Similarly, $\mu\left(f_{diff}(S)\right)$, $sd\left(f_{diff}(S)\right)$, and $z$ are basic operations that are computed outside the $for$ loop across the entire dataset in a single scan. Unless the sample size is a very small number, each of their time complexities can be considered as constant, without much influence to the overall algorithm runtime.

### *4.3.2. Median Statistic Z*

$$f_{aggr}(x) = median\left(f\big(N(x)\big)\right)$$

$$f_{diff}(x) = f(x) - f_{aggr}(x)$$

$$z(x) = \frac{|\, f_{diff}(x) - median\big(f_{diff}(S)\big)\,|}{MAD\big(f_{diff}(S)\big)}$$

*Median* (Chen et al., 2008) is identical to *Spatial*, except that the *mean* is replaced by the *median*, and the *standard deviation* is replaced by the *median absolute deviation*. The *median absolute deviation* is calculated as follows:

$$MAD = median\,(|X_i - median(X)|)$$

Where $X$ is the vector of values and $X_i$ is the value for the i$^{th}$ observation in the vector.

### *4.3.3. Local Area Mean*

$$f_{aggr}(x) = \mu\left(f\big(N(x)\big)\right)$$

$$f_{aggrsd}(x) = sd\left(f\big(N(x)\big)\right)$$

$$f_{diff}(x) = f(x) - f_{aggr}(x)$$

$$z(x) = \frac{|\, f_{diff}(x)\,|}{f_{aggrsd}(x)}$$

As appearing in Kleinjan et al. (2002), *Local Area Mean's* neighbourhood aggregate function is composed of two functions: $f_{aggr}$ and $f_{aggrsd}$. $f_{aggr}$ is calculated identical to *Spatial*, thus, their complexity time is the same. However, unlike *Spatial* or *Median*, *Local* does not standardize globally, but uses a local standardization method for each neighbourhood. This local standardization is based on $f_{aggrsd}$. $f_{aggrsd}(x)$ computes the standard deviation of the attribute value of neighbours of $x$. This means that each observation is standardized differently.

### *4.3.4. Scatter Plot*

$$f_{aggr}(x) = \mu\left(f\left(N(x)\right)\right)$$

$$f_{diff}(x) = f_{aggr}(x) - (mf(x) + b)$$

$$z(x) = \frac{| f_{diff}(x) - \mu\left(f_{diff}(S)\right) |}{sd\left(f_{diff}(S)\right)}$$

Unlike *Spatial, Median,* and *Local, Scatter* is a graphical spatial outlier technique (Shekhar et al., 2003). If plotted, $f(x)$ is on the X-axis, and $f_{aggr}(x)$ on the Y-axis. Then, a least-square regression line $f_{aggr}(x) = mf(x) + b + \varepsilon(x)$ is fitted, and observations with the largest residuals, $\varepsilon$, are considered as probable spatial outliers. Thus, $f_{diff}(x) = \varepsilon(x) = f_{aggr}(x) - (mf(x) + b)$, where $m$ is the estimated slope of the line and $b$ is the estimated intercept. $m$ and $b$ can be found by using the following formula:

$$m = \frac{\sum_{i=1}^{n}\left(f(x_i) - \mu\left(f(x_i)\right)\right) \times \sum_{i=1}^{n}\left(f_{aggr}(x_i) - \mu\left(f_{aggr}(x_i)\right)\right)}{\sum_{i=1}^{n}\left(f(x_i) - \mu\left(f(x_i)\right)\right)^2}$$

$$b = \mu\left(f_{aggr}(S)\right) - m\mu\left(f(S)\right)$$

$m$, $b$, $f_{diff}$, and $z$ are based on basic operations across the entire dataset without the use of a $for$ loop; thus, their time complexity is almost quadratic.

### 4.3.5. Spatial Local Outlier Measure

$$f_{norm}(x) = \frac{\left(f(x) - \min f(S)\right)}{\left(\max f(S) - \min f(S)\right)}$$

$$f_{aggr}(x) = f(N(x))$$

$$f_{diff}(x) = f_{norm}(x) - f_{aggr}(x)$$

$d(x) = \frac{\Sigma|f_{diff}(x)| - max|f_{diff}(x)|}{k - 1}$, where the sum is over the set $f_{diff}(x)$, which has $k$ elements

```
$If\ d(x) > \mu(d(N(x))$
    $\beta(x) = Count(d(x) > d(N(x)))$
$Else\ if\ d(x) < \mu(d\left(N(x)\right)$
    $\beta(x) = Count(d(x) < d(N(x)))$
$End\ If$
```

$$\beta^{*}(x) = \frac{|\beta(x)|}{(k+1) - 2} \times \frac{1}{1 + \mu\left(d\left(N(x)\right)\right)}$$

$$z(x) = d(x)\beta^{*}(x)$$

Unlike other spatial outlier algorithms, *SLOM* does not standardize the outlier score but it standardizes the attribute (Chawla and Sun, 2006). *SLOM* starts with the normalization of $f(x)$ such that $f(x)$ is between 0 and 1 (i.e. $f_{norm}(x)$). This requires searching for $\max f(S)$ and $\min f(S)$ which are basic operations. $f_{aggr}(x)$ is then computed similar to *Median*; thus, the same time complexity. However, unlike *Median* where $f_{aggr}(x)$ is aggregated during its computation, SLOM has a “dynamic” aggregate and comparison function, because it performs the aggregation and the comparison altogether during the computation of $d(x)$. The aggregation function is technically $\sum\left|f_{diff}(x)\right| - max\left|f_{diff}(x)\right|$, and since $f_{diff}(x) = f_{norm}(x) - f_{aggr}(x)$, $d(x)$ can be re-written as:

$$d(x) = \frac{\sum\left|f_{norm}(x) - f_{aggr}(x)\right| - max\left|f_{norm}(x) - f_{aggr}(x)\right|}{k - 1}$$

$f_{aggr}(x)$ only searches for the neighbours of $f(x)$. $f_{aggr}(x)$ and $f_{diff}(x)$ are not aggregated; they are vectors, not scalars. Because $d(x)$ is computed in a $for$ loop, its time complexity is at least quadratic. The algorithm then calculates $\beta(x)$ which is the net number of times the values around $x$ are bigger or smaller than its neighbours. The idea is that if a neighbourhood has low deviation, a spatial outlier within the neighbourhood would be easier to detect rather than a neighbourhood with high deviation. This concept resembles the *Local Area Mean* algorithm. The neighbourhood deviation, or oscillation, is captured by $\beta(x)$. After computing $\beta(x)$, $\beta(x)$ is divided by $(k + 1) - 2$ for a boundary correction and to standardize $\beta$ so its maximum value is 1. $\beta(x)$ is further divided by $1 + \mu\Big(d\big(N(x)\big)\Big)$ which allows to penalize situations where large values of $d\big(N(x)\big)$ exists around $d(x)$.

### *4.3.6. Weighted Z*

$f_{aggr}(x) = \sum\Big(f\big(N(x)\big) \times weight\big(N(x)\big)\Big)$, where × is used to denote the element-wise vector product

$weight(y) = \frac{distance\,(x,y)^{-1}}{\sum distance\big(x,N(x)\big)^{-1}}$, for $y \in N(x)$

$$\sum weight(N(x)) = 1$$

$$f_{diff}(x) = f(x) - f_{aggr}(x)$$

$$z(x) \ = \ \frac{|\, f_{diff}(x) - \mu\big(f_{diff}(S)\big) \,|}{sd(f_{diff}(S))}$$

Kou et al. (2006) introduces the *Weighted Z* algorithm. $f_{aggr}(x)$ calculates the weighted average of the non-spatial attribute of the neighbours of $x$. This is performed by first determining $N(x)$ with the $rank$ function. The weight represent the impact of each neighbour in relation to $x$. If $y$ is the nearest neighbour of $x$, then $y$ has more impact in $f_{aggr}(x)$ calculation. The weight value for neighbour $y$ is between 0 and 1, and the sum of weights for all $x$'s neighbours is 1. Thus, the weight of $y$ is calculated by inverting the Euclidean distance of $(x, y)$ and dividing it with the sum of all the inverse distances between $x$'s neighbours.

### 4.3.7. *Inverse Distance Weighted to a Power (IDWP)*

$$f_{aggr}(x) = \sum \Big(f\big(N(x)\big) \times \ weight\big(N(x)\big)\Big)$$

$$weight(y) = \frac{distance\ (x,y)^{-p}}{\sum distance\big(x,N(x)\big)^{-p}} \ , p = 2, \text{for } y \in N(x)$$

$$\sum weight(N(x)) \ = \ 1$$

$$f_{diff}(x) \ = \ f(x) - f_{aggr}(x)$$

$$z(x) \ = \ \frac{|\, f_{diff}(x) - \mu\big(f_{diff}(S)\big) \,|}{sd(f_{diff}(S))}$$

*Inverse Distance Weighted to a Power* is very similar to *Weighted*. The only difference is that an exponent is applied to the inverse distances. Thus, closer observations will have more impact in the calculation of $f_{aggr}$ than in *Weighted*. Similarly, observations farther from $x$ will have less influence in the calculation of $f_{aggr}(x)$ than in *Weighted*. In this case, the power function, $p$, is set to equal to 2, which is the same exponent used by Simbahan et al. (2004) and Ping & Dobermann (2005).

### *4.3.8. Kriging Interpolation*

$f_{aggr}(x) = \sum \left(f\big(N(x)\big) \times \ weight\big(N(x)\big)\right)$

$weight\big(N(x)\big) = \Big(A\big(N(x)\big)\Big)^{-1} B(x)$, where $A\big(N(x)\big)$ and $B(x)$ are a matrix and vector of semivariances as defined later in this section

$f_{diff}(x) \ = \ f(x) - f_{aggr}(x)$

$z(x) \ = \ \dfrac{\mid f_{diff}(x) - \mu\big(f_{diff}(S)\big) \mid}{sd(f_{diff}(S))}$

*Kriging* is an interpolation technique that estimates the value at a location based on linear weighted combination of the neighbouring locations. Thus, the idea is to interpolate each point in the dataset and compare the interpolated value against the true value to test for outlierness. *Kriging* starts by calculating the experimental semivariogram. For observations, $f(x_i), i = 1, \dots, k$ at locations $x_1, \dots \ , x_k$ the empirical semi-variogram is defined as (Cressie, 1993):

$$y^*(h) = \frac{1}{2\left|\dot{N}(h)\right|} \sum_{(i,j)\,\epsilon \dot{N}(h)} \left(f(x_i) - \ f(x_j)\right)^2$$

where $h$ is the lag distance between $x_i$ and $x_j$ such that $\left|x_i - \ x_j\right| = h$, and $\left|\dot{N}(h)\right|$ is the number of pairs in the set. Since a spherical semivariogram is used in the sequential simulation, a spherical model is used here as well, which has the form:

$$y^{sph}\left(h;\ \sigma^2_{nugg};\ \sigma^2;r\right) = \begin{cases} \sigma^2_{nugg} & if\ h = 0 \\ \sigma^2\left(\dfrac{3h}{2r} - \dfrac{h^3}{2r^2}\right) & if\ h < r \\ \sigma^2 & otherwise \end{cases}$$

where $\sigma^2_{nugg} \geq 0$ is the nugget semivariance, $\sigma^2$ is the sill, and $r$ is the range of autocorrelation. Semivariogram functions available in *gstat* are utilized to calculate the empirical and model semivariogram. Both semivariograms are computed only once, globally for all observations. As the default, *gstat* uses iteratively reweighted least-squares (WLS) (Cressie, 1985) to estimate model semivariogram parameters but estimation by generalized least-squares (GLS) and restricted maximum likelihood (REML) are also available (Pebesma, 2004). $A\big(N(x)\big)$ is a $k+1$ by $k+1$ matrix of model semivariances between neighbours of $x$. This matrix characterizes the spatial autocorrelation of $N(x)$. Prior to computing this matrix, $h$ has to be determined, which are pairwise Euclidean distances among points. Once $h$ is determined, $y^{sph}\left(h;\ \sigma^2_{nugg};\ \sigma^2;r\right)$ is utilized to calculate matrix $A(N(x))$.

$B(x)$ is a $k+1$ vector that characterizes the spatial autocorrelation between $x$ and its neighbours, $N(x)$, which is computed with $y^{sph}\left(h;\ \sigma^2_{nugg};\ \sigma^2;r\right)$. $weight$ is then calculated by multiplying $B$ with the inverse of $A$. The $(k+1)^{th}$ row in $weight$ is then removed in order for $\sum weight = 1$. This is the effect of the Langrange multiplier.

Kriging interpolation is perhaps the most complex among all algorithms used. $h$ is computed for $A(N(x))$ in a $for$ loop with Euclidean distance. Given that $h$ is computed for each observation, the computation requires a nested $for$ loop; the inner loop is $k$ times, and the outer loop which is $N$ times. $A(N(x))$ and $B(x)$ are simple calculations, but are calculated $N$ times. So, their complexity is at least $N$ times each. $A$ has to be inverted, and matrix inversion is at most cubic runtime (Strassen, 1969). Matrix multiplication is at most a $nmp$ runtime for a $n \times m$ and $m \times p$ matrix (Strassen, 1969). And $f\big(N(x)\big)$ is computed identical as in *Weighted's* $f\big(N(x)\big)$, thus the same time complexity.

### *4.3.9. Averaged Difference (AvgDiff)*

$$f_{aggr}(x) = f(N(x))$$

$$f_{diff}(x) = |f(x) - f_{aggr}(x)|$$

$$z(x) = \sum f_{diff}(x) \times weight(N(x))$$

$$weight(y) = \frac{distance\,(x,y)^{-1}}{\sum distance\big(x,N(x)\big)^{-1}}\text{, for } y \in N(x)$$

$$\sum weight(N(x)) = 1$$

As appearing in Kou et al. (2006), *AvgDiff* has a dynamic aggregate function because neighbourhood aggregation does not occur at the initial stages of the algorithm, which is similar to *SLOM*. $f_{aggr}(x)$ only searches for the neighbours of $f(x)$. $f_{diff}(x)$ computes the absolute difference between $f(x)$ and $f_{aggr}(x)$, and the actual aggregation occur during the computation of $z$. $weight$ is calculated same as *Weighted* with the same time complexity. But unlike *Weighted*, *AvgDiff* does not standardize outlier scores.

### *4.3.10. Spatial Outlier Test*

$$f_{aggr}(x) = f\big(N(x)\big)$$

$$f_{diff}(x) = slope\big(x, N(x)\big) = \frac{f(x) - f_{aggr}(x)}{Distance\big(x, N(x)\big)}$$

$$z(x) = \sum \left|f_{diff}(x)\right| - \left|max\left(f_{diff}(x)\right)\right| - \left|min\left(f_{diff}(x)\right)\right|$$

The idea of *SOTest* is to compare the slope, rise over run, between $x$ and its neighbours, $N(x)$, over their respective distances. If the sum of all of these slopes is a large value, then $x$ is likely a spatial outlier. However, given that there may be more than one (positive or negative) spatial outlier in $N(x)$, the maximum slope and the minimum slope are taken away from the sum, i.e. a trimmed sum is used. The maximum slope is taken in order to suppress spatial outliers that are above the neighbourhood average, while the minimum slope is taken away to suppress spatial outliers that are below the neighbourhood average. If there are no spatial outliers in $N(x)$, then removing both the maximum and minimum would not make a significant change to the computation of the outlier score.

## 4.4. Assessment of Spatial Outlier Techniques

### *4.4.1. Introduction*

The previous section on detecting spatial outliers involves using each spatial outlier algorithm to assign an outlierness score to each observation in the simulated dataset. This section evaluates whether the outlier scores derived by each algorithm is correct. Previous performance assessments of spatial outlier techniques involved the ranking and the comparison of the top $m$ spatial outlier detected by each algorithm (Lu et al., 2003; Chandola et al., 2006; Kou et al., 2006; Chen et al., 2008). This is problematic because experiments were conducted on real datasets whereby the spatial outliers are really not known in advance. Thus, the comparison between algorithms by ranking top outliers does not necessarily determine the algorithm correctness because there is no point of reference of what a spatial outlier really is. On the other hand, the simulated spatial data contains known spatial outliers, which is the point of reference needed to make such comparisons. Given that there are only two class labels, outliers and non-outliers, the assessment of algorithms can be conducted as a binary classification problem, treating each spatial outlier algorithm as a classifier. Thus, the question is, 'how accurate are the prediction and classification of each algorithm?'

The answer is to use performance measures available in classification problems such as accuracy, misclassification error, sensitivity, specificity, among others. Two very popular analytical tools that encompass such performance measures are a confusion matrix and the receiver operating characteristics (ROC) curve. Both the ROC curve and the confusion matrix are techniques to visualize, organize, and select classifiers based on their performance (Fawcett, 2006). However, the ROC curve is utilized to assess the algorithm performance because it summarizes multiple confusion matrices at different decision thresholds.

### *4.4.2. ROC Curve*

The ROC curve is constructed by plotting the sensitivity (true-positive rate or $TPR$) and 1 – specificity (false-positive rate or $FPR$) of the classifier against each other as a function of $c$, a threshold criterion (Hanczar et al., 2010). In this case, the threshold criterion is the spatial outlier score, $z$. Informally, the ROC curve is equal to the collection of multiple confusion matrices with differing thresholds for class selection. This means that the information of one confusion matrix represents a single point, $(TPR(c), FPR(c))$ in the ROC curve. Thus, the ROC curve can be used to summarize all the confusion matrices that could have been produced with differing thresholds.

A simple method to compare classifiers is to reduce the information contained in the ROC curve down to a single convenient scalar value that represents the classifier performance. Various indices have been used to summarize the ROC curve. The most popular one is the AUC, the area under the ROC curve, noted as the most recommendable index of detectability (McClish, 1989). The AUC is a scalar that summarizes across all thresholds, reflecting the overall quality of the classifier. The AUC of a classifier is equivalent to the probability that the classifier will score a randomly chosen positive sample higher than a randomly negative sample (Fawcett, 2006).

The AUC is a portion of the area of the unit square, so AUC ∈ $[0,1]$, $AUC(0,1) = \int_0^1 ROC(t)dt$. AUC = 1, corresponds to the perfect classifier that will correctly detect all spatial outliers without any false positives. AUC = 0.5 corresponds to an uninformative classifier that is not better than a classifier that randomly guesses whether observations are spatial outliers or not. Technically, AUC = 0.5 usually corresponds to the diagonal of the ROC curve, $y = x$, although it can sometimes meander around the diagonal. When a straight diagonal line is depicted, TPR will always equal to FPR. For example, the classifier may correctly detect 80% of the spatial outliers, but will also incorrectly detect 80% of non-outliers. If there are 100 outliers and 100 non-outliers, the classifier will label 160 observations as outliers; 80 of them correctly, 80 of them incorrectly. AUC = 0 corresponds to a classifier assigning all observations to the wrong class. In this situation, all spatial outliers would be classified as non-outliers and all non-outliers as spatial outliers.

The AUC may be a misleading measure for classifier performance. Total area is, in some sense, not the ideal measure of the classifier performance. AUC is a single global measure that summarizes over the region of the ROC curve in which one would rarely operate (Dodd & Pepe, 2003). In practical situations, researchers may only be interested in a few situations rather than all of them. For instance in medical studies, population screening may result in large monetary costs of follow-up examinations if FPR is high; thus, the focus would be on the ROC areas corresponding to low FPR. Similarly, in diagnostic testing, high TPR is emphasized in order to not miss-detecting subjects with disease; hence the area with high TPR is of particular interest (Dodd & Pepe, 2003).

Similarly, when comparing ROC curves, the curves may be identical for some range, but one curve may be superior to the other in other ranges. This can imply that a high-AUC classifier can perform worse than a low-AUC classifier for a particular range of the curve. This subtlety is not captured with the AUC. One naive approach is to compare ROC curves at individual points on the curve (McClish, 1989). The novel approach would be to compare the partial area under the

ROC curve, $PAUC(t_0, t_1) = \int_{t_0}^{t_1} ROC(t)dt$, for a fixed range of FPR or TPR values (see Figure 4.5).

The 5% FPR and 80% TPR are chosen as the performance thresholds to compare the algorithms. This is because outlier detection algorithms with a high TPR and a low FPR are highly desirable. For FPR, $t_0 = 0, t_1 = 0.05$, and for TPR, $t_0 = 0.8, t_1 = 1.0$. These two conditions are evaluated for each algorithm. Note that for ROC curves, the FPR is on the x-axis and TPR on the y-axis. Thus theoretically, finding the PAUC with respect to TPR will involve integrating on the y-axis.

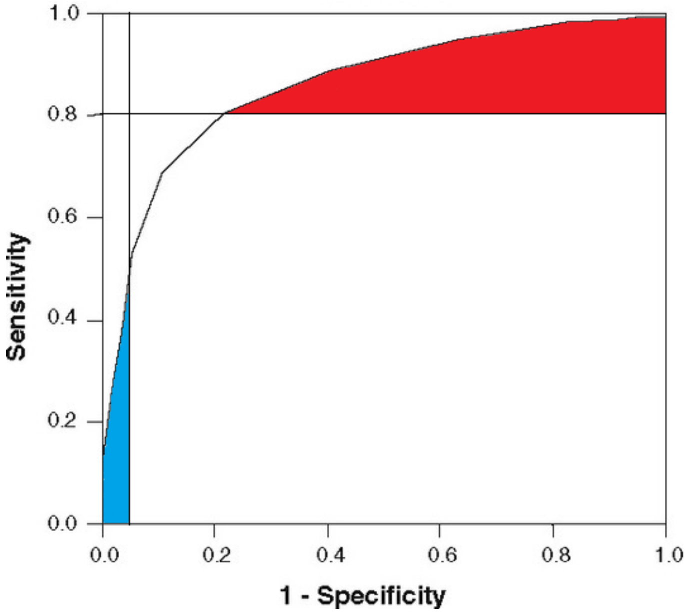


**Figure 4.6: Selected partial area under ROC curve at 5% FPR (blue) and from 80% TPR (red)**

The R package *ROCR* provides the tools to construct ROC curves along with performance measures such as the AUC, and PAUC up to a fixed FPR can be computed by passing an optional parameter, *fpr.stop=0.05*. However, *ROCR* is not capable of restricting a fixed TPR to calculate PAUC. It can only restrict the ROC region of interest to FPR. The solution has been to transform the ROC curve to a specificity-ROC curve, which is a 270$^{0}$ rotation of the original curve having the TPR on the x-axis (Dodd & Pepe, 2003). However, *ROCR* does not recognize this reconfiguration. The alternative solution to find the PAUC from 80% TPR, based on simple considerations, is as follows:

1. compute the AUC,
2. find the FPR in which TPR = 0.8,
3. compute the PAUC at this FPR, and
4. Subtract the AUC in (1) minus the PAUC in (3) and minus 0.8 multiplied by (1 – FPR).

### *4.4.3. Sensitivity Analysis*

An additional increased uncertainty exists regarding algorithm ROC performance under differing parameters, particularly when different numbers of nearest neighbours can be used to compute the neighbourhood aggregation function, $f_{aggr}$. Uncertainty arises because there is no consensus regarding how many nearest neighbours to use, and subsequently, no knowledge about how the ROC performance of an algorithm is influenced by the number of nearest neighbours. As such, determining the uncertainty of algorithm performance can be accomplished by way of a sensitivity analysis. The basic idea of sensitivity analysis is to change a single parameter while holding all remaining parameters constant. This would determine the influence of the single parameter in relation to the remaining parameters, which would allow the identification of algorithms that are unstable under a user-specified number of nearest neighbours. Thus, the investigation of neighbourhood definition in spatial outlier detection is conducted by applying spatial outlier algorithms to the simulated dataset at different user-specified number of nearest neighbours and utilizes the ROC measures to determine algorithm performance at each defined neighbourhood. The number of nearest neighbours under investigation will range from four nearest neighbours to 20.

A need also exists to determine the robustness of each algorithm given that algorithm performance may be influenced by the structure of the dataset. For example, algorithms may perform differently in local areas with large variation (large nugget effect) than in areas with little local variation. Certainly, spatial outliers located in areas with large variation would be harder to detect than spatial outliers that occur in areas with little variation. Similarly, algorithm performance may be influenced by the overall dataset itself. An approach to determine algorithm robustness is by way of exploring the variation of ROC performance measurements obtained from all 20 simulations. Particularly, the standard deviation would convey whether an algorithm is capable of obtaining consistent performance under different data structures. Thus, the approach is to calculate the standard deviation of the ROC curve measures for each of the investigated number of nearest neighbours used to compute $f_{aggr}$. 20 replications are performed for each NN setting.

### *4.4.4. Neighbourhood Sensitivity*

A statistical approach at determining the algorithms' neighbourhood stability, i.e. to determine whether changing the NN parameter alters algorithm performance, is to perform a one-way analysis of variance (ANOVA) test. ANOVA provides a test to determine whether or not multiple means or proportions are statistical different (De Veaux et al., 2005). ANOVA relies on the F-statistic, which is the ratio between the treatment mean square ($MS_T$), the variation between groups, and the error mean square ($MS_E$), the variation within groups (De Veaux et al., 2005). Three assumptions must be satisfied: independence, equal variance, and normality. Independence is checked for between and within groups. Between-group independence may be questionable. ROC performances, AUC, PAUC TPR, and PAUC FPR, in the groups are generated from the same algorithm but with different NN parameter; they are nonetheless derived from the same dataset. Within-group independence is met as each group contains 20 datasets that were simulated independently.

Equal variance can be checked by observing the spread of the boxplots, particularly the spread of the IQR. A more objective approach is to use a statistical equal variance test such as the Brown-Forsythe homogeneity of variance test (Appendix B). Normality can also be checked by visual inspection of the boxplot, normal quantile plot, or histogram, or via a normality test such as the Shapiro-Wilk test (Appendix A). Both Brown-Forsythe test and Shapiro-Wilk test are preferred because they provide a test statistic. The null hypothesis for the Brown-Forsythe test is that the population variances are equal while the null hypothesis from the Shapiro-Wilk test is that the sample comes from a normally-distributed population.

The Kruskal-Wallis test, a non-parametric extension to ANOVA test, can be utilized to assess neighbourhood impact on AUC score without making the assumption of a normally-distributed population. Because the Kruskal-Wallis test first ranks all observations for all groups together, the test is analogous to testing population medians instead of means (Hollander & Wolfe, 1973).

### *4.4.5. Algorithm Performance Similarity*

This section evaluates whether algorithms are statistically similar or different. The parametric statistical approach to determining whether algorithms are different is by way of performing a two-sample t-test or a paired t-test. The idea is to perform the test on the ROC performance measures: AUC, PAUC at TPR and PAUC at FPR. The two sample t-test is commonly used to determine if two independent means or proportions are statistically different. However, independence assumption is broken because the data, in this case the ROC performance measures, come from the same simulated datasets but generated from different algorithms. In such situation, a paired t-test is more appropriate.

A paired t-test requires two assumptions to be met. First, the data has to be paired. Pairing is met, as mentioned. ROC performance measurements are derived from the same set of simulations for each algorithm. This is analogous to having the same subjects try different treatments. Secondly, the data must follow a normal distribution. Normality is checked with the Shapiro-Wilk test (Appendix A). Again, the normally-distributed population assumption is broken for SLOM. Thus, a non-parametric approach is utilized. The paired Wilcoxon test, also known as the Mann-Whitney test, is the non-parametric version of the paired t-test which does not require a normal distribution (Hollander & Wolfe, 1973).

## 4.5. Evaluating Spatial Outlier Effects in Site-Specific Management

The most common approach at determining the effects of spatial outliers have been the *ex ante* and *ex post* analysis of yield for a particular statistical analysis; that is, comparing the raw data against the pre-processed data. For example, such analyses include the *ex ante* and *ex post* estimation of the summary statistics of crop yield and its semivariogram parameters as well as yield mapping (Thylen & Murphy, 1996; Beck et al., 1999; Kleinjan et al., 2002; Simbahan et al., 2004; Ping & Dobermann, 2005; Sudduth & Drummond, 2007). With the exception of yield mapping, there is little of or no value for site-specific management regarding the information conveyed in summary statistics and semivariogram parameters because the true parameters are unknown in real situations. In addition, it is difficult to observe differences in yield maps obtained in *ex ante* and *ex post* yield mapping, as the maps will appear almost identical, depending on the spatial resolution and level of outlier contamination.

Several studies have been conducted to investigate crop yield response functions (Long, 1998; Bullock & Lowenberg-DeBoer, 2002; Lambert et al., 2003; Anselin et al., 2004; Liu et al., 2006; Brenning et al., 2008). A few of them have compared the effectiveness of different spatial regression models regarding coefficient estimation (Lambert et al., 2003; Anselin et al., 2004; Brenning et al., 2008). But none have addressed the potential effects of spatial outliers in their spatial analysis. The proposed approach takes a similar path at comparing coefficient estimates derived from a spatial regression model. Here, different spatial outlier detection algorithms will be applied and each resulting dataset will be then used to estimate the crop yield coefficients. In this case, spatial outlier algorithms are being compared in terms of how effectively coefficients are estimated after outlier detection algorithms are applied.

Several types of spatial regression models are popular for estimating crop yield response functions. Classical ordinary least-squares (OLS) regression has been shown to underestimate field heterogeneity and has led to biased or misleading inferences about crop response function because crop yield data is almost always spatially correlated (Bullock & Lowenberg-DeBoer, 2002). As a response, regression models that account for spatial correlation have been proposed. Particularly, four spatial regression models are commonly used, which include classical nearest neighbour, polynomial trend, spatial autoregressive model (SAR), and a geostatistical approach (Lambert et al., 2003; Brenning et al., 2008). SAR and geostatistical approach remain the most popular techniques among agronomists.

In this work, the geostatistical approach is chosen for coefficient estimation. This is because although SAR and geostatisical are similar in obtaining similar parameter estimates, the latter approach is about 30% more efficient in terms of computation time (Brenning et al., 2008). In addition, SAR can fail because of numerical singularities that cannot be avoided by sub-sampling (Brenning et al., 2008). And both SAR and geostatistical approach have shown to be more precise than the classical nearest neighbour and polynomial trend approaches in terms of coefficient estimation (Lambert et al., 2003).

The geostatistical approach in Cressie (1993) serves a backbone for spatial regression, but several geostatisticians have elaborated upon the approach. To estimate coefficients, the approach that appears in Goovaerts (1997) is going to be utilized, which is as follows:

1. Determine a linear model of the variables. In this case, Ordinary Least-Squares: $y = X\beta + \varepsilon$, which $y$ is the regressed value, $X$ is the regressors, $\beta$ is the vector of

coefficients, in this case, coefficients for the environmental variables and treatment variables ($d_1$, $d_2$, $g_1$, $g_2$), and $\varepsilon$ is the error term at all locations.

2. Derive OLS residuals in the form $\varepsilon = y - X\beta$
3. Compute the empirical semivariogram for the OLS residuals, obtain the model semivariogram, and model $C$, the covariance matrix of the OLS residuals, with the nugget, sill, and range of the computed model semivariogram, $c_0, c_1$, and $r$ respectiely.
4. Use Generalized Least-Squares (GLS) to determine coefficients. Cressie (1993) provides the estimation of the coefficients, which are solved by $\hat{\beta} = (X^T C^{-1} X)^{-1} X^T C^{-1} y$

However, a disadvantage of using spatial regression models is that they are computationally intensive. Neither GLS nor SAR can be computed for a spatial dataset consisting of 20,000 point because of insufficient random-access memory (RAM), even for a computer equipped with 4 gigabytes of RAM. The common approach is to spatially aggregate the data to a point density that is consistent with the scale at which agricultural machinery operates (Brenning et al., 2008). Aggregation of spatial point data consists of summarizing points by computing the mean centre of a local neighbourhood within a user-specified distance and then taking the local neighbourhood median attribute value. For instance, if three spatial points were to be aggregated, the centre location would obtain the median attribute value. In this case, a three-metre nearest neighbour distance is utilized, which would derive a point density of approximately 5,000 points.

The R package *nlme* provides functions to fit linear and non-linear mixed effects models, in this case, generalized least squares (GLS) linear regression. In GLS, the errors are allowed to be correlated and/or have unequal variances. The covariance matrix constructed with the spatial correlation structure given by the spherical model, or any user-specified semivariogram model, is also derived with functions in the *nlme* package. Calculation of the empirical semivariogram and semivariogram modelling is implemented in the R package *gstat*.

## 4.6. Chapter Summary

Unconditional sequential Gaussian simulation is performed to generate crop yield data along with two explanatory variables. Point and region spatial outliers are added separately to the simulated datasets by randomly picking observations and adding or subtracting a Gaussian error term to the observed value. Given that each spatial outlier is known in advance, the assessment of spatial outlier techniques can be conducted as a binary classification problem, treating each spatial algorithm as a classifier. Performance assessment is evaluated with the area and partial area under the ROC curve at 80% true positive and 5% false positive rates. Two additional analyses involves determining whether changing the number of nearest neighbours affect the algorithm performance, and determining which algorithms are most similar in terms of ROC performance. Further investigation of the spatial outlier effects is conducted by coefficient estimation with a geostastical approach, which involves incorporating semivariogram parameters into the covariance matrix of a generalized least-square regression to fit a model into the dataset that has been spatially aggregated.

# CHAPTER 5:

# RESULTS AND DISCUSSION

## 5.1. Geostatistical Simulation

The results of the simulations are summarized in Table 5.1. The minimum yield value of 46.04 bu/acre and the maximum of 107.55 bu/acre correspond to a three and a half standard deviations away from the mean. This spread is reasonable given the large sample size. The fourth column indicates the simulation with the addition of 2,000 point spatial outliers, which amounts to 5% of the total number of observations. And the fifth column depicts the addition of region outliers of size 5 for the same 5% contamination. A few differences can be inferred between both.

First, the addition of point and region spatial outlier has generated global outliers since the minimum has decreased about 2.5 bu/acre and the maximum has increased 2 bu/acre. As such, the standard deviation has been inflated, but other summary measures remain almost unchanged. The reason global outliers appeared on the simulation is because the spatial outlier generator selects random observations as spatial outliers. Thus, observations which are relatively extreme could be selected and superimposing the outlier term on them could result in the generation of global outliers.

**Table 5.1: Summary of Simulations**

| *Summary Statistics* | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Initial Parameters | $D_o$ | | $D_p$ | | $D_r$ | |
| Yield | | Mean | Between-replication std. error | Mean | Between-replication std. error | Mean | Between-replication std. error |
| Minimum | | 46.04 | 0.80 | 43.37 | 0.81 | 43.87 | 0.78 |
| 1st Quartile | | 70.92 | 0.30 | 70.85 | 0.30 | 70.74 | 0.31 |
| Median | | 76.80 | 0.27 | 76.82 | 0.27 | 76.81 | 0.27 |
| Mean | 76 | 76.81 | 0.29 | 76.83 | 0.29 | 76.81 | 0.28 |
| 3rd Quartile | | 82.71 | 0.31 | 82.82 | 0.32 | 82.90 | 0.32 |
| Maximum | | 107.55 | 0.65 | 109.48 | 0.54 | 110.53 | 0.64 |
| Standard Deviation | | 8.77 | 0.14 | 8.92 | 0.13 | 9.06 | 0.13 |
| *Semivariogram Parameters of Yield* | | | | | | | |
| Nugget | 3.5 | 3.65 | 0.52 | 6.07 | 0.56 | 8.25 | 0.57 |
| Sill | 70 | 73.67 | 3.17 | 73.77 | 3.23 | 74.19 | 3.27 |
| Range | 150 | 145.62 | 3.27 | 145.47 | 3.24 | 145.23 | 3.36 |
| Nugget-to-Sill ratio | 5.0% | 4.9% | | 8.2% | | 11.1% | |

$D_o$ − original simulation; $D_p$ −simulation after adding point outliers; $D_r$ – simulation after adding region outliers.

Units is bushels per acre.

The most significant difference between the original simulations and the simulation with spatial outliers is the inflation of the estimated nugget effect. For point outliers, the nugget almost doubled from 3.65 bu/acre to 6.07 bu/acre, making the nugget-to-sill ratio increase to 8.2%, and for region outliers, the nugget almost tripled with an 11% nugget-to-sill ratio. This can be attributed to the simulated spatial outliers. The nugget describes the short range micro-scale variability that is present because of measurement error or in this case, inherent variability. Spatial outliers produce local instability by introducing observations that are markedly different from their nearest neighbours. This implies that micro-scale variation is increased as nearest neighbours are on average more different when spatial outliers are present. This is further exasperated by region outliers given that a cluster is now more dissimilar to nearest observations.

**Table 5.2: Average Pearson correlation between simulated variables**

| | $D_p$ | $D_o$ | $d_1$ |
|---|---|---|---|
| $D_o$ | 0.98 | | |
| $d_1$ | 0.31 | 0.32 | |
| $d_2$ | 0.02 | 0.03 | -0.01 |

$D_o$ – original simulation; $D_p$ – simulation after adding point outliers; $d_1$ – environmental variable 1; $d_2$ – environmental variable 2

The average Pearson correlation between the simulated variables for the 20 simulations is shown in Table 5.2. This table suggests that the spatial outliers have the slightest impact on the correlation between simulated variables as the correlation difference is 0.01 between coefficients of $D_o$ and $D_p$ . This may be credited to the magnitude and quantity of spatial outliers. The correlation structure between variables may have been affected significantly if more spatial outliers were introduced with a higher error value.

## 5.2. Point Outlier Algorithm Performance

### *5.2.1. Area under ROC curve*

Figure 5.1 shows the area under ROC curve for each algorithm under different number of nearest neighbours used to compute $f_{aggr}$. All algorithms that do not account for spatial autocorrelation, *Spatial, Median, Local, Scatter,* and *SLOM* falter against the number of nearest neighbours (NN). As the number of neighbours increases, the AUC decreases rapidly, even for *Median*. *Median* has the highest AUC between 4 and 8 NN but decreases rapidly once NN reaches to 20. The AUC pattern for *Local* differs from all other algorithm. This is perhaps because *Local Area Mean* restricts each observation to the statistics computed from the moving window, unlike all other outlier algorithms.

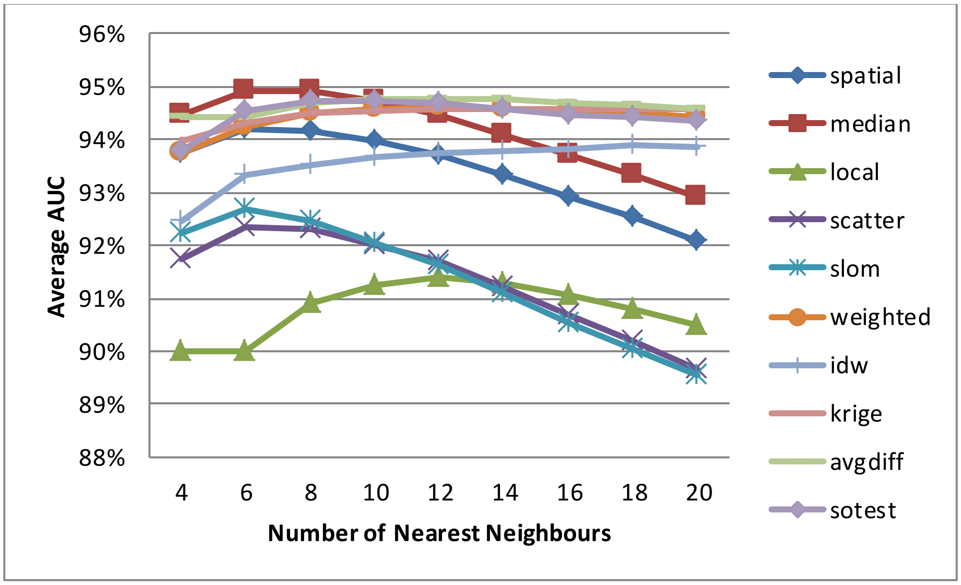


**Figure 5.1: AUC Sensitivity analysis over 20 simulated datasets**

Algorithms that account for spatial autocorrelation are less influenced by the change of NN. This may be related to the fact that spatial autocorrelation algorithms assign a different weight to each neighbour during the computation of $f_{aggr}$. So, observations in the neighbourhood which are weakly autocorrelated will therefore provide a minor contribution to the spatial neighbourhood. AUC for *Weighted, Kriging, AvgDiff,* and *SOTest* decrease slightly as NN increases. However, AUC increases for *Inverse Distance Weighting* as NN increases.

Spatial autocorrelation based algorithms obtain the lowest standard deviation as well as two standard techniques, *Spatial* and *Median* (Figure 5.2). In particular, *Median* obtains the lowest variation across most of the NN settings, which suggests that *Median* is an algorithm that performs consistently on different datasets across all NN definition. *SOTest, AvgDiff, Weighted,* and *Kriging* are subsequent algorithms that have lowest standard deviation across the NN settings.

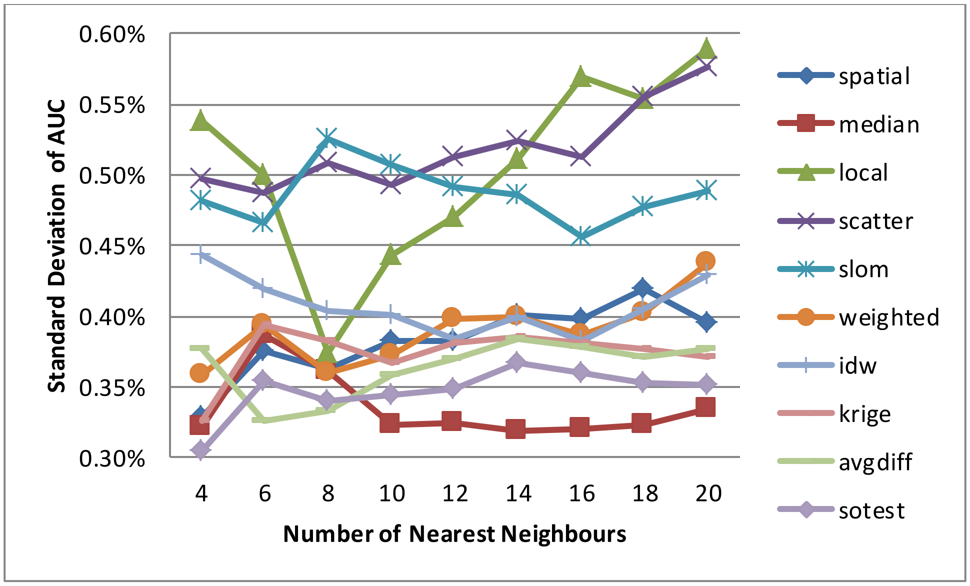


**Figure 5.2: Standard deviation of AUC over 20 simulated datasets**

### *5.2.2. Partial area under ROC curve from 80% true positive rate*

Given that in this case TPR is restricted at 0.8, the maximum area that can be obtained is 0.2 or 20%. As evidenced, spatial autocorrelation algorithms obtain higher PAUC than algorithms that do not account for spatial autcorrelation. *Median* performs well with small number of nearest neighbours (Figure 5.3). The biggest contrast is the poor PAUC performance of *SLOM*. *SLOM* obtains less than a 10% PAUC for all tested nearest neighbours. This implies that *SLOM* obtains a very high rate of false positives when obtaining a true positive rate of 80%.

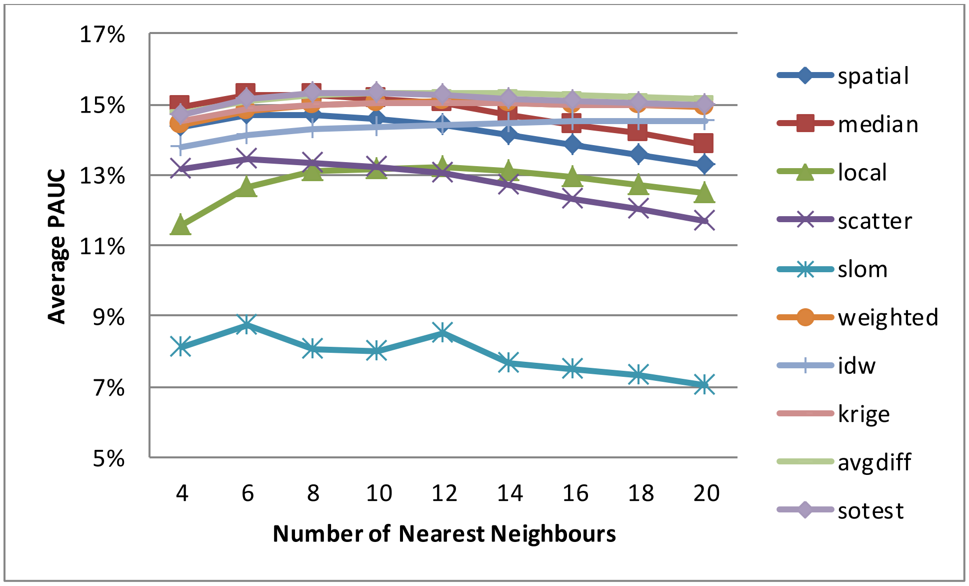


**Figure 5.3: PAUC from 80% TPR sensitivity analysis over 20 simulated datasets**

*SLOM* obtains significantly the highest error across all NN settings, which is about four times higher than all other algorithms (Figure 5.4). Given that *SLOM* obtains the lowest PAUC and the highest variation for all NN settings suggests that SLOM does not adapt too well to different datasets and has difficulties detecting different spatial outliers.

Given that the variation of *SLOM* ROC performance is extreme, Figure 5.5 depicts the standard deviation for all algorithms with the exception of SLOM. There is no discernable pattern as most algorithms fluctuate with the change of NN. Nevertheless as depicted, the two most stable algorithms with the smallest error are *AvgDiff* and *SOTest*, in which the error slightly increases as the NN increases.

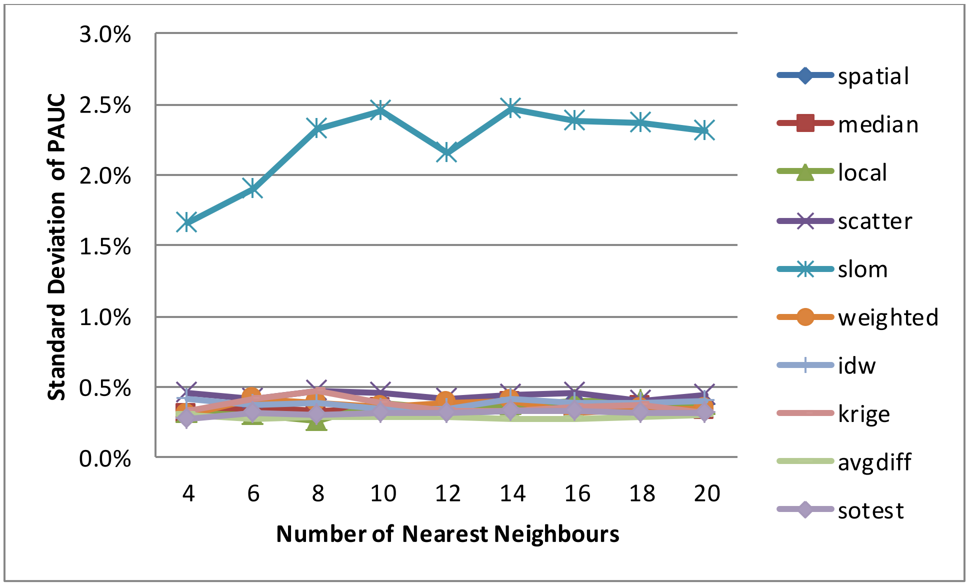


**Figure 5.4: Standard deviation of PAUC at 80% TPR over 20 simulated datasets**

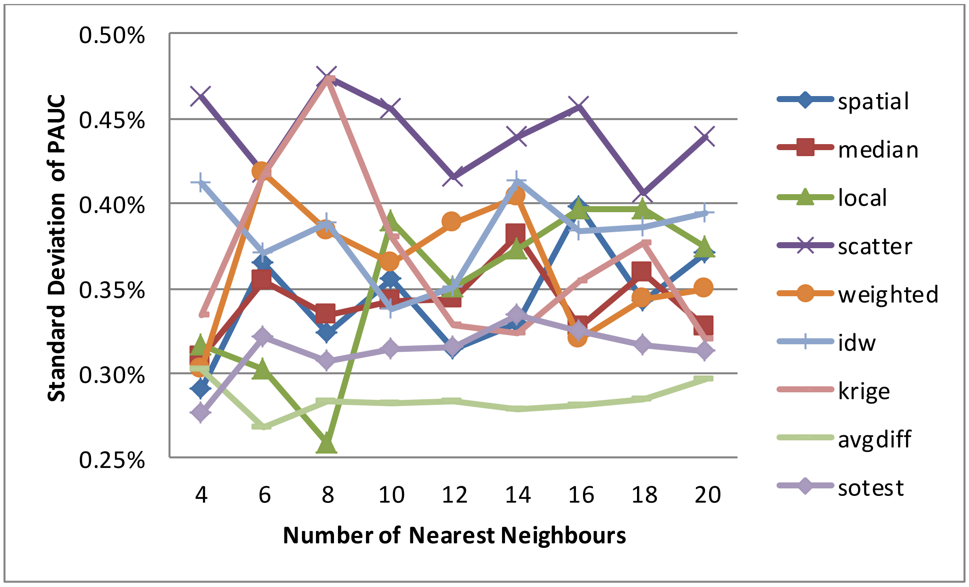


**Figure 5.5: Standard deviation of PAUC at 80% TPR without SLOM over 20 simulated datasets**

### *5.2.3. Partial area under ROC curve at 5% false positive rate*

Figure 5.6 provides the partial area under ROC curve at 5% false positive rate for each algorithm under different number of nearest neighbours. The maximum area that can be obtained is 5% given that the FPR is restricted at 0.05. In this case, obtaining a large PAUC by implies that an algorithm obtains a relatively high true positive rate given a false positive of 5%. Figure 5.8 shows spatial autocorrelation algorithms obtain the highest PAUC, much similar to the AUC in Figure 5.1. The revealing information is that *Loca*l is the algorithm with poorest performance, especially when NN is small.

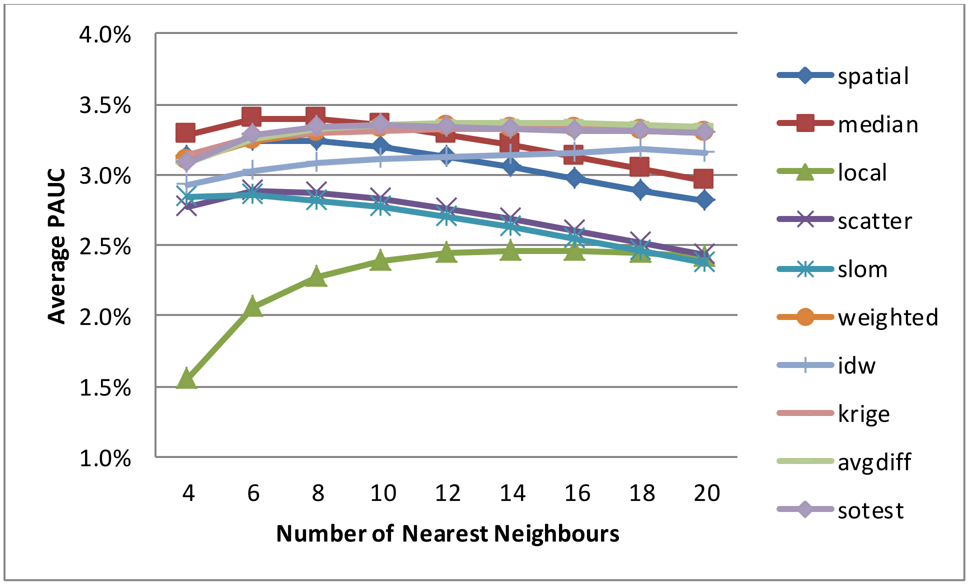


**Figure 5.6: PAUC at 5% FPR sensitivity analysis over 20 simulated datasets**

Figure 5.7 provides the standard deviation for the PAUC at 5% FPR. *SOTest* and *IDWPP* obtain the lowest variation while *SLOM* and *Scatter* obtain the highest variation in most of the NN settings. However, the total variation range is very small compared to the range in PAUC TPR and AUC. Range for AUC variation is about 0.35% and for PAUC TPR is approximately 2.5% (0.25% without SLOM), while PAUC FPR is about 0.05%. A clear trend can be evidenced in all the figures depicting standard deviation. *Local* obtains a sudden change in error when NN equals 8.

In Figure 5.2, error drops the lowest when NN is 8 and then increases almost linearly. Similarly, Figure 5.5 shows the error dropping when NN is 8 and then a sudden increase and fluctuation. Finally, Figure 5.7 depicts the error suddenly drop and then remaining constant after NN equals 8. This trend may be evidence of *Local* algorithm over-fitting the data.

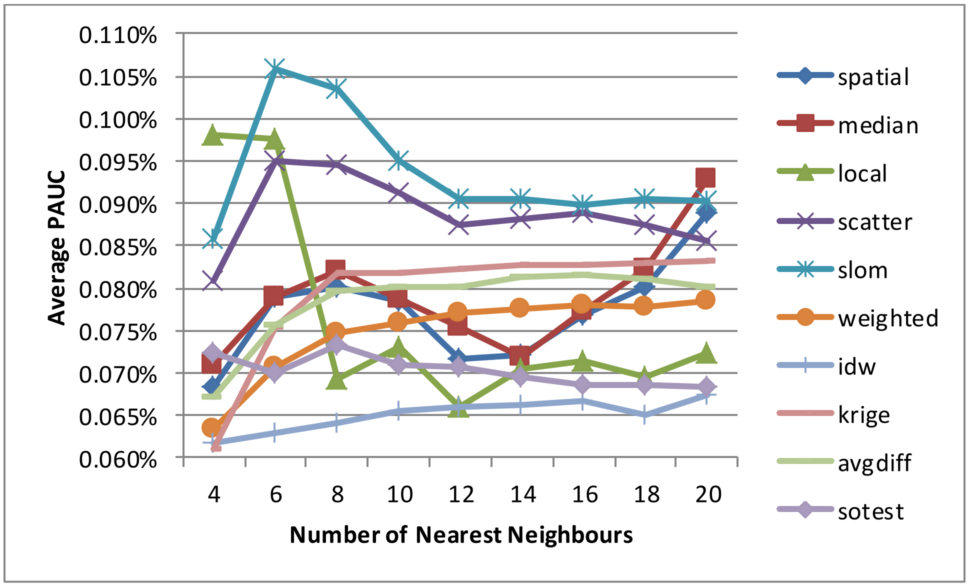


**Figure 5.7: Standard deviation of PAUC at 5% FPR over 20 simulated datasets**

## 5.3. Region Outlier Algorithm Performance

### *5.3.1. Introduction*

Section 5.2 presented the ROC performance results of the spatial outlier algorithms for single point outlier situations. Such situation where a single outlier is present is implicative that swamping but rarely masking effect exists when a single spatial outlier is present in the computation of the neighbourhood aggregation function. In this case, the spatial outlier in the spatial neighbourhood will inflate the neighbourhood statistic, making the observation more dissimilar to the neighbourhood. When an algorithm fails to detect a spatial outlier, then it is due to the confusion exhibited in the inherent natural variability of the spatial data.

However, this section explores the situations of region outliers where multiple spatial outliers are clustered together, which is implicative that more than one spatial outlier are present in the computation of the neighbourhood aggregate function. In addition to confusion with natural inherent variability and swamping effects, masking effects affecting true spatial outliers is present.

Again, a sensitivity analysis is a viable approach at determining the influence of parameter on algorithm performance. This time two parameters exist: the number of nearest neighbours utilized in the computation of the spatial neighbourhood function and the number of spatial outliers clustered in the region. The following section investigates the sensibility of these two parameters for the AUC, PAUC 80% TPR, and PAUC at 5% FPR.

### *5.3.2. Area under ROC curve*

As evidenced, all algorithms are weakened by the size of the region outlier (Figure 5.8). The algorithm performances drop approximately linear, and this is not of much surprise. The larger the region outlier size, the more instances of masking occurs. And this is particularly critical for *Local. Local* drops to an AUC close to 50% when the region size equals to 5. At this point, *Local* is no longer an informative algorithm as it obtains the same number of true positives as false positives for any detection threshold.

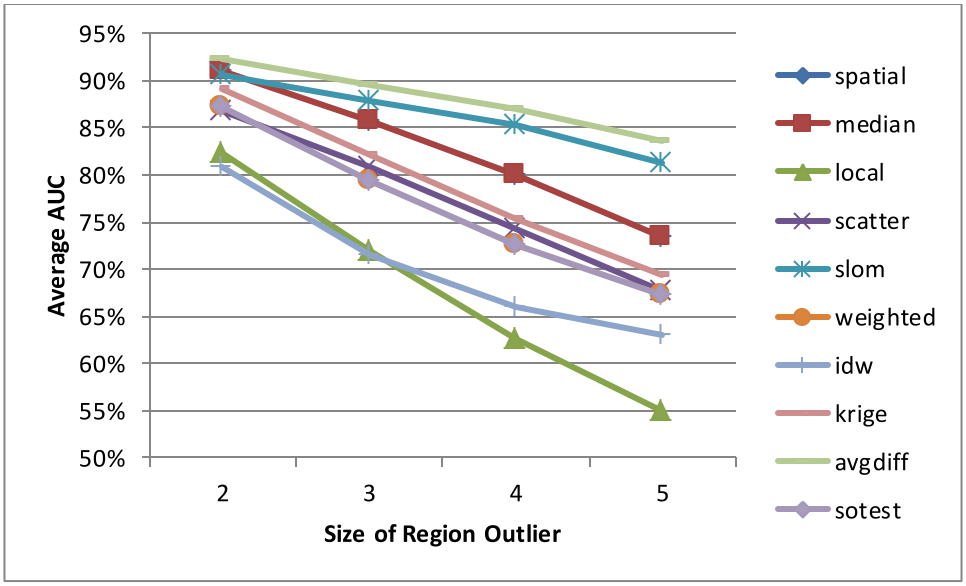


**Figure 5.8: AUC sensitivity at 8 NN over 20 simulated datasets**

Figure 5.9 shows the AUC performance against the sensitivity of NN used to calculate the neighbourhood aggregate function. The clear best algorithm is *AvgDiff* because it obtains the highest AUC in all NN settings followed by *Spatial*, *Median*, and surprisingly, *SLOM,* although *SLOM* deteriorates with increasing NN. Spatial autocorrelation techniques perform variably. Very poor performance is evidenced in *IDWP,* while *Kriging, Weighted,* and *SOTest* are inferior to *Spatial* and *Median.* This time, spatial outliers are clustered together. So, spatial autocorrelation algorithms assign more weight to neighbours which are spatial outliers. In turn, the aggregate function and subsequent statistics are contaminated by multiple spatial outliers.

Note that, unlike point outlier situations where all algorithms, especially algorithms that do not account for spatial autocorrelation, are weakened by increasing NN, here the opposite is true for region outlier. In this situation, AUC performance increases as NN increases, which would ultimately reach a plateau where AUC performance can no longer increase.

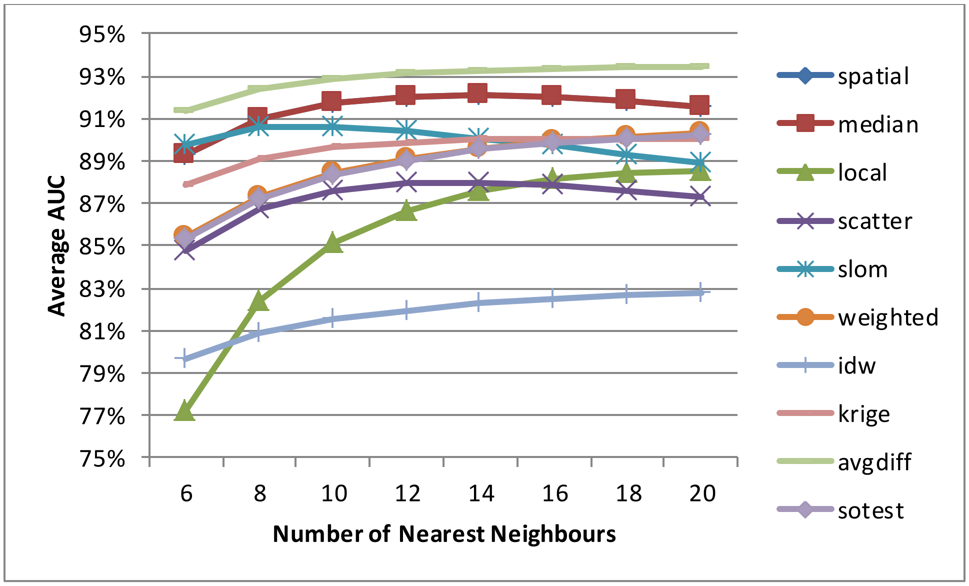


**Figure 5.9: AUC sensitivity at region outlier size 2 over 20 simulated datasets**

Figure 5.10 now presents the AUC performance given region outlier of size 5. Not much difference exists between Figure 5.9 and Figure 5.10. For example, AUC performance for *Spatial* and *Median* remain identical along with *Scatter* and *SOTest.* This evidence suggests that the size of the region size has the same influence on the performance of all algorithms. This time however, the AUC performance increase more sharply as NN increases, compared to the smooth increase as depicted in Figure 5.9. The reason for this sharp increase is due to the size of the region outlier. As the region outlier size increases, the number of nearest neighbours required to properly describe the spatial property of the dataset also increases.

Another difference between Figure 5.9 is that all spatial autocorrelation algorithms, except for *AvgDiff,* perform worse than standard algorithm. In particular, Figure 5.9 depicted *Kriging, SOTest,* and *Weighted* obtaining higher AUC performance than *Scatter*. At region outlier size of 5, *Scatter* outperforms *Kriging, SOTest,* and *Weighted.*

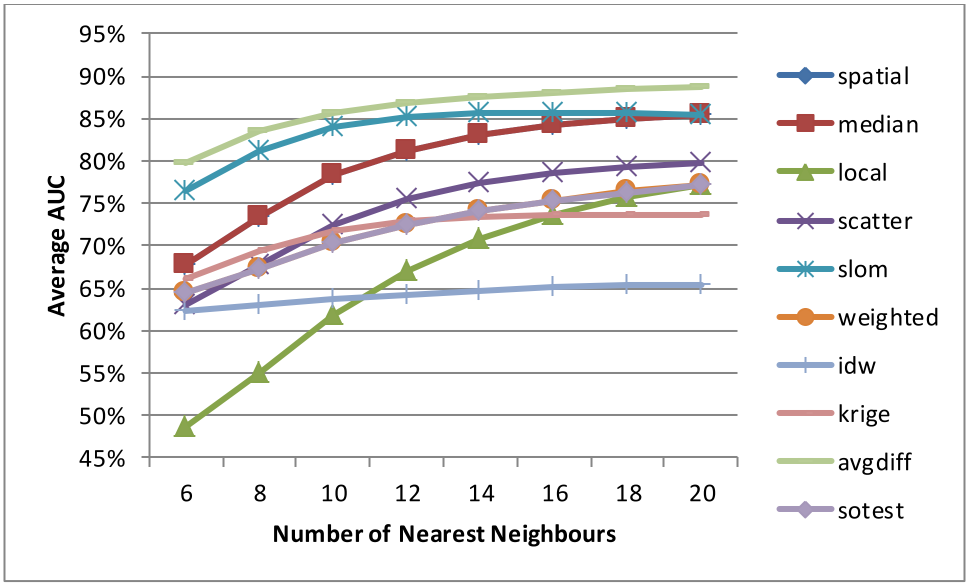


**Figure 5.10: AUC sensitivity at region outlier size 5 over 20 simulated datasets**

Figure 5.11 depicts that all algorithms, except *Kriging*, obtain low standard deviation across all NN settings, suggesting that these algorithm obtains consistent performance for different datasets. *Kriging* approximately obtains more than double the variation of the rest algorithms. *Kriging* requires the computation of the semivariogram parameters: nugget, sill, and range. The nugget, as evidenced in section 5.1, is clearly affected by spatial outliers. Given the pair-wise comparisons in order to compute the semivariogram, region outliers of size 2 are mostly influential in contaminating the semivariogram parameters. Because the nugget is computed globally, it would not fit well locally on areas in which region outliers of size 2 occur, suggesting that a local computation of semivariogram parameters would be preferred.

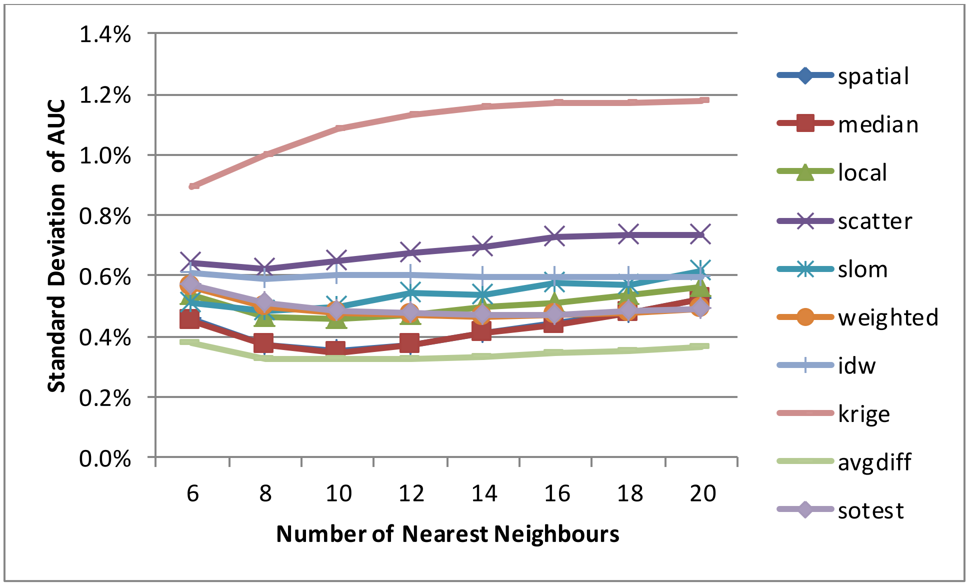


**Figure 5.11: Standard deviation of AUC at region outlier size 2 over 20 simulated datasets**

Now in Figure 5.12, *Kriging* obtains about three times variation as all other algorithms. Since more outliers are clustered in a region, the nugget is more contaminated and harder to be correctly fitted by the model semivariogram. This contamination is accredited to the fact that small variation exists within the outliers in a region, thus, the interpolated values would not correctly match the true values. Consequently, *Kriging* performs less consistently with larger size of region outliers. Further revealing information from Figure 5.12 is that *Scatter* has higher standard deviation value, which are about double the variation obtained in Figure 5.10.

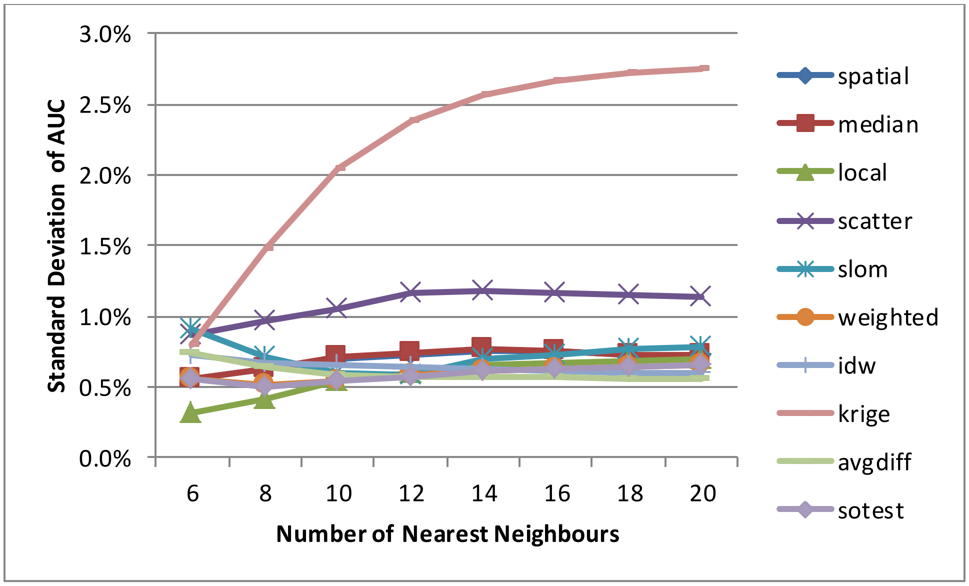


**Figure 5.12: Standard deviation of AUC at region outlier size 5 over 20 simulated datasets**

### *5.3.3. Partial area under ROC curve from 80% true positive rate*

The performance of all algorithms is weakened by the region outlier size (Figure 5.13). All algorithms except *AvgDiff*, *IDWP*, and *SLOM* succumb in a linear fashion. In particular, *SLOM* obtains lowest PAUC when detecting region outliers of size 3 and 4, but PAUC strangely increases when region outliers equal 5. On the contrary, the decay of *AvgDiff* PAUC performance occurs smoothly. At region size of 5, the difference in performance between *AvgDiff* and all other algorithms is evident.

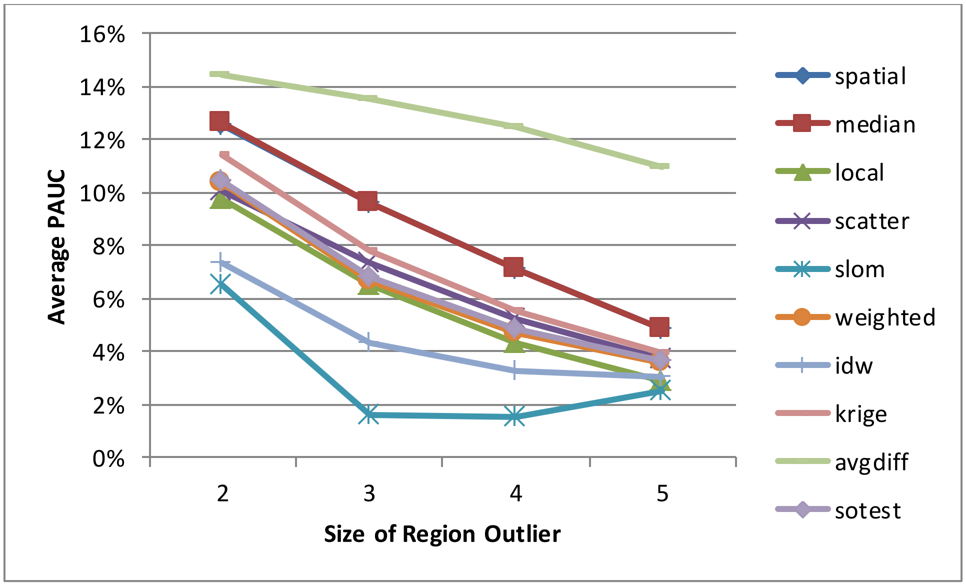


**Figure 5.13: PAUC 80% TPR at 8 NN over 20 simulated datasets**

For region outlier size 2, Figure 5.14 resembles the AUC performance of the same region outlier size (see Figure 5.9). The only difference is the performance of *SLOM.* Unlike AUC performance whereby *SLOM* obtained the fourth highest performance, here *SLOM* obtains the worst performance among all algorithms, which is the same trend evidenced for the single outlier situation. This suggests that for single and region outliers, *SLOM* perform relatively well on all decision thresholds with the exception of decision thresholds that achieve high sensitivity.

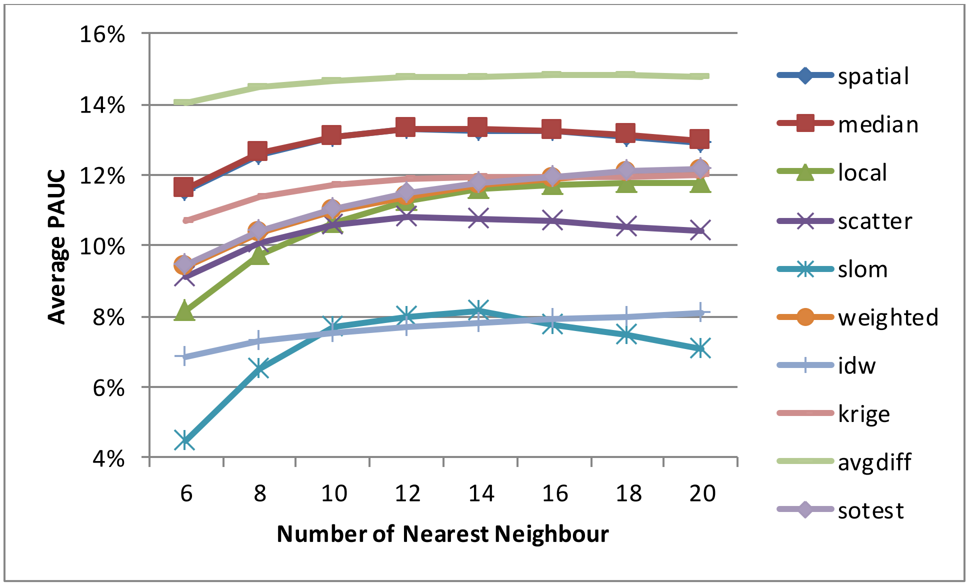


**Figure 5.14: PAUC 80% TPR at region outlier size 2 over 20 simulated datasets**

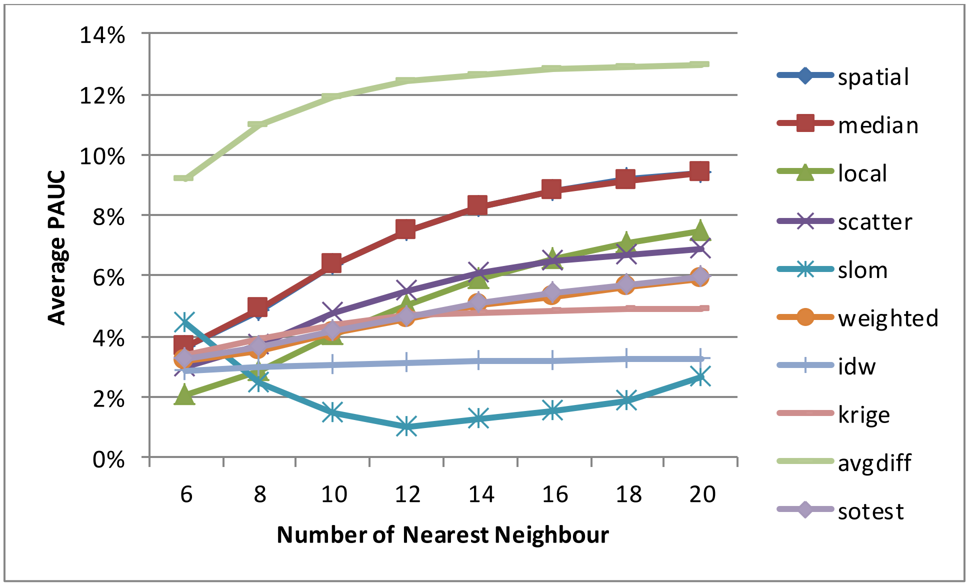


**Figure 5.15: PAUC 80% TPR at region outlier size 5 over 20 simulated datasets**

Although all algorithms have decreased performance, *AvgDiff* is depicted substantially superior to all other algorithms (Figure 5.15). For instance, *AvgDiff* performs about 5% better than

*Spatial* and *Median*. This is surprisingly unexpected given that the performance gap between these algorithms is approximately less than 2% for detecting region outliers of size 2. As a result, *AvgDiff* is able to obtain the lowest FPR when obtaining 80% TPR for all region outliers, as compared to other algorithms, and the performance gap increases with increasing region outlier size.

Figure 5.16 provides the variation of the PAUC performances from 80% TPR for detecting region outliers of size 2. Similar to the single outlier situation, *SLOM* obtains the highest variation across all NN settings, approximately five times the variation evidenced in all other algorithms. This further proves *SLOM* is a very inconsistent algorithm at obtaining high sensitivity performance not only for situations of single outlier but also of region outliers. Other than *SLOM*, all algorithms obtain similar variation.

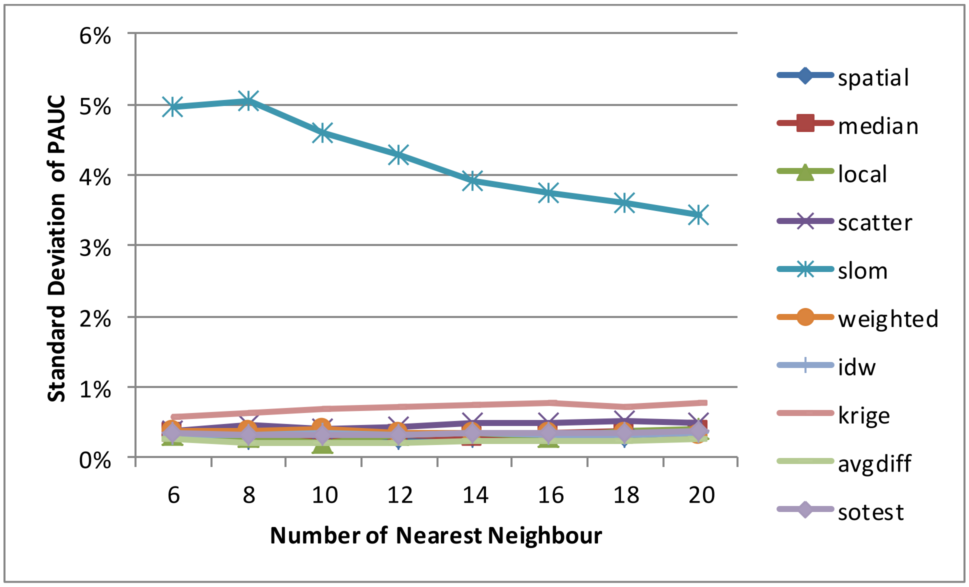


**Figure 5.16: Standard deviation of PAUC 80% TPR at region outlier size 2 over 20 simulated datasets**

Figure 5.17 depicts the standard deviation for the algorithms at PAUC 80% TPR for region outliers of size 5. Again, *SLOM* obtain considerable higher variation than other algorithms. However, the standard deviation is variable across NN settings, with the lowest value at 12 NN.

Notice that in Figure 5.17, PAUC for *SLOM* is lowest at NN equals 12, which may suggest that *SLOM* over-fits the data at this neighbourhood configuration as it obtains lowest performance and lowest variation at the same time.

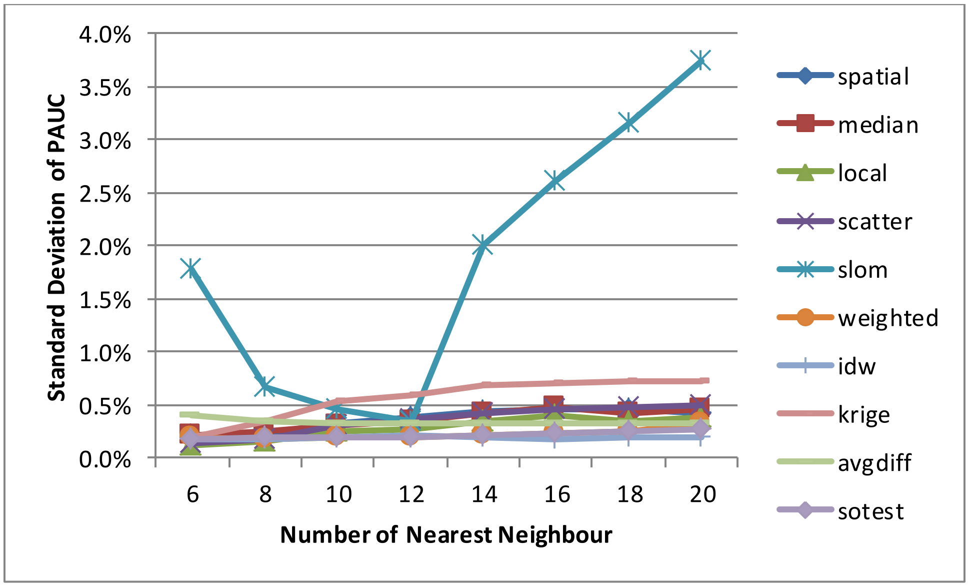


**Figure 5.17: Standard deviation of PAUC 80% TPR at region outlier size 5 over 20 simulated datasets**

## *5.3.4. Partial area under ROC curve at 5% false positive rate*

Figure 5.18 shows the PAUC at 5% FPR performance at 8 NN. Similar to its AUC and PAUC TPR counterpart, all algorithms drop performance as the region outlier size increases. Note that, PAUC of *Local* drops to almost zero when detecting group outliers of size 5. This suggests that *Local* detects only 95% of inliers without detecting any spatial outliers. And similar to previous evidence, *Spatial* and *Median,* and *Weighted* and *SOTest* obtain identical PAUC trend. The difference here however, is that *SLOM* obtains much better PAUC at FPR performance than PAUC at TPR. *SLOM* is third only to *Spatial* and *Median.*

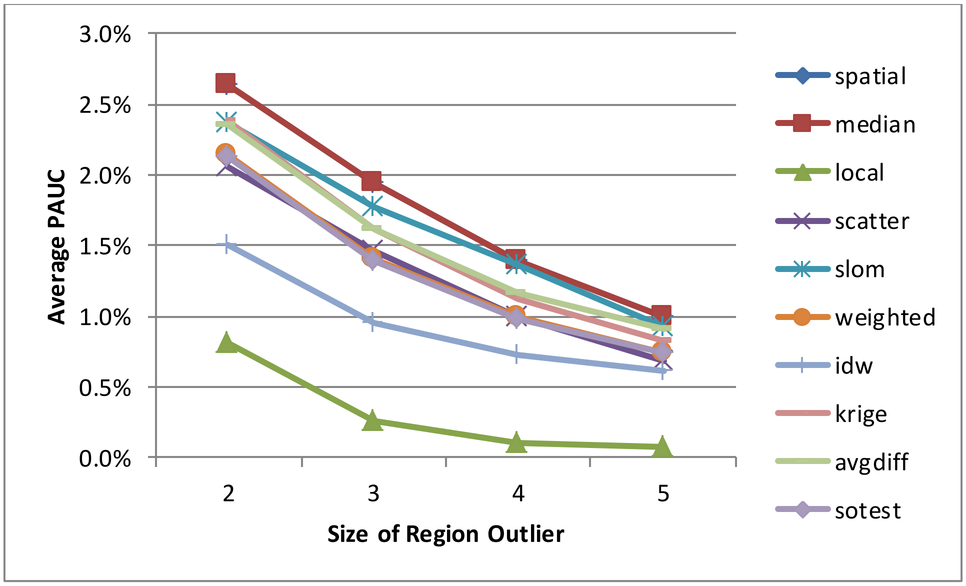


**Figure 5.18: PAUC 5% FPR sensitivity at 8 NN over 20 simulated datasets**

The results in Figure 5.19 depict the same similarities as the AUC and PAUC TPR performance. The only difference is that *AvgDiff* obtains the third highest PAUC FPR behind *Spatial* and *Median,* whereas *AvgDiff* obtained the highest overall AUC and PAUC TPR performance. This suggests that *AvgDiff* performs best on all decision thresholds with the exception of decision thresholds that achieve 5% false positive rate or less. Notice that *Local* and *IDWP* are significantly inferior to all other algorithms. Additionally, *SLOM* and *Kriging* are very similar.

Not of much surprise, the performances of all algorithms in Figure 5.20 resemble the PAUC FPR performances at region outlier of size 2. The main distinction is the *SLOM* and *Kriging* are no longer similar. *Kriging* stabilizes by obtaining a PAUC of 1.0%, while *SLOM* continues to increase.

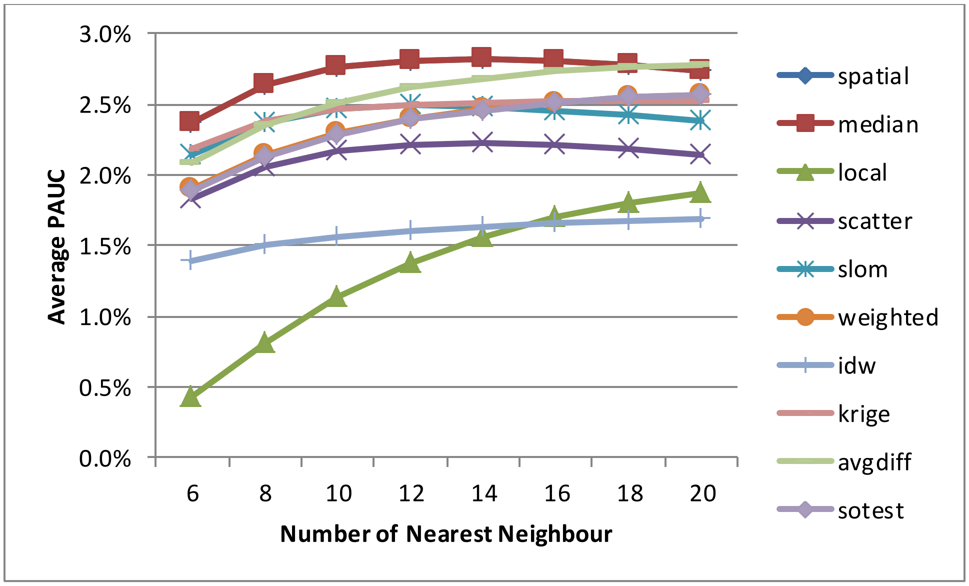


**Figure 5.19: PAUC 5% FPR at region outlier size 2 over 20 simulated datasets**

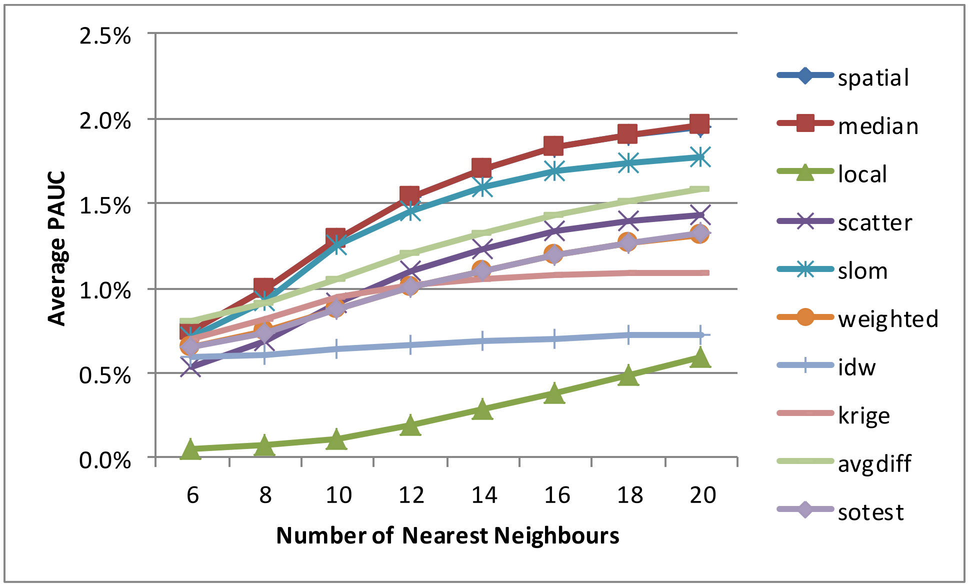


**Figure 5.20: PAUC 5% FPR at region outlier size 5 over 20 simulated datasets**

Figure 5.21 provides the variation of the PAUC performances at 5% FPR for detecting region outliers of size 2. Not surprisinglyg, *Kriging* obtains the highest variability due to the inability to correctly to compute the nugget semivariogram because of the presence of region outliers. However, the range of the standard deviation quite low about less than 0.16%, which is not an indication that the algorithm is significantly inconsistent. Besides *Kriging,* all spatial autocorrelation algorithms obtain lower standard deviation than algorithms that do not account for spatial autocorrelation, except *Local* for the lower range of NN.

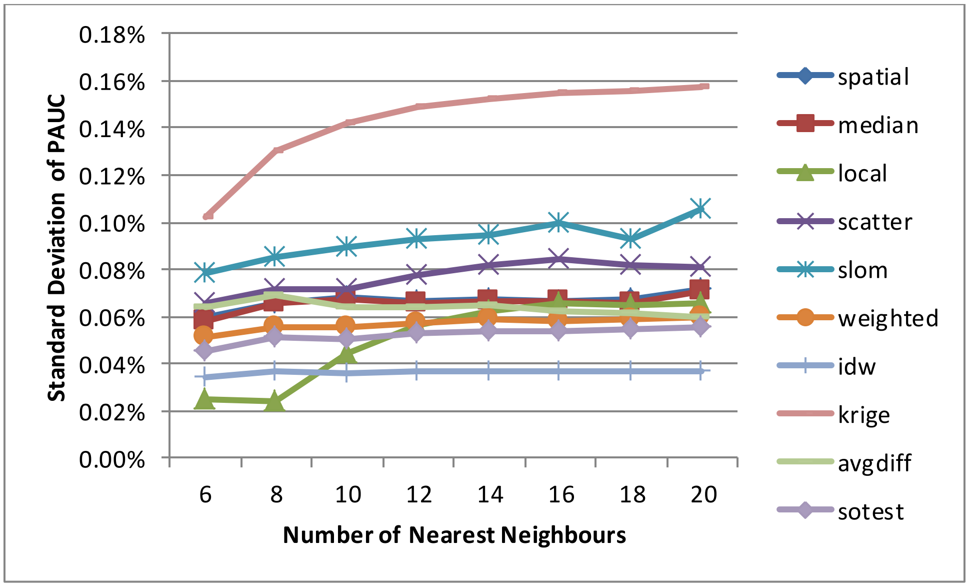


**Figure 5.21: Standard deviation of PAUC 5% FPR at region outlier size 2 over 20 simulated datasets**

Figure 5.22 shows the PAUC 5% FPR performance variation for detecting region outliers of size 5. *Local* obtains lowest standard deviation among all NN settings, while *Scatter* obtains relatively high variation, and *Kriging* obtains the highest variation. However, similar to the single outlier scenario, the range of the variation, about 0.16%, is not substantial to suggest significant performance inconsistency among NN settings.

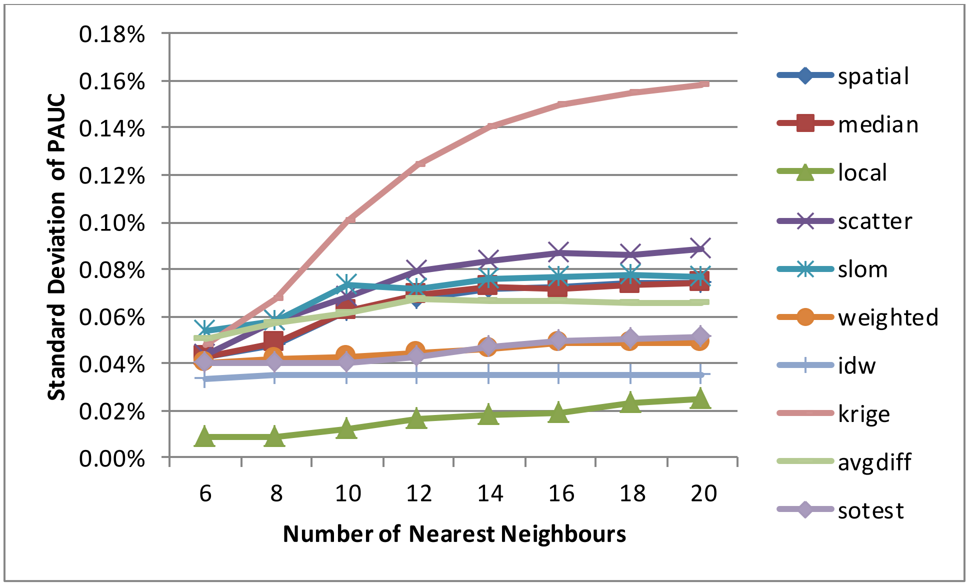


**Figure 5.22: Standard deviation of PAUC 5% FPR at region outlier size 5 over 20 simulated datasets**

## 5.4. Neighbourhood Size Stability

The test results for the Shapiro-Wilk are shown in Appendix A. Most of the p-values are higher than 0.2, which suggest that there is no statistical evidence to reject the null hypothesis that the ROC samples for each NN setting come from a normally-distributed population. However, *SLOM* obtains p-values of less than 0.01 for few NN settings which suggest that normality assumption may be broken only for *SLOM*. To be more conservative, an alternative test, the Kruskal-Wallis rank sum test is implemented as well (Table 5.3).

In terms of the Brown-Forsythe test (Appendix B), there is no evidence for non-equal variances among NN groups (all p-values are not significant at a 5% critical level). Therefore, all statistical assumptions precluding the comparisons of multiple means are met.

**Table 5.3: Point outlier test for neighbourhood stability. Only reported test statistics that are not significant at a 1% significance level**

| | Spatial | Median | Local | Scatter | SLOM | Weighted | IDWP | Kriging | AvgDiff | SOTest |
|---|---|---|---|---|---|---|---|---|---|---|
| | ANOVA Test | | | | | | | | | |
| AUC | | | | | | 0.03 | | 0.32 | 0.01 | |
| TPR | | | | | 0.19 | 0.26 | 0.03 | 0.84 | 0.13 | 0.97 |
| FPR | | | | | | 0.01 | | 0.31 | | |
| | Kruskal-Wallis Rank Sum Test | | | | | | | | | |
| AUC | | | | | | 0.06 | | 0.47 | 0.02 | 0.33 |
| TPR | | | | | 0.16 | 0.47 | 0.03 | 0.93 | 0.11 | 0.97 |
| FPR | | | | | | 0.01 | | 0.26 | | |

For AUC, all spatial outlier detection algorithms that account for spatial autocorrelation except *IDWP* and *SOTest* were not significant at a 1% level, which suggests that means and/or medians of ROC performance from the several NN groups are statistically equal. This implies that varying the NN parameter for the calculation of $f_{aggr}$ does not have a significant effect on the algorithm AUC performance (see Table 5.3).

In regards to PAUC 80% TPR, there is more consistency among performance within different nearest neighbour aggregation. Unlike the neighbourhood stability results for AUC, performance for *IDWP* and *SLOM* are statistically stable. *SLOM* is resistant to changes in the neighbourhood definition at a 10% significant level. However, its TPR performance is significantly low compared to all others, which may imply that SLOM obtains a very high and relatively constant number of false positives at any given user-specified neighbourhood.

For PAUC at 5% FPR, only *Weighted* and *Kriging* are stable across the tested neighbourhood definitions at 1% significance. Overall, there is indication that most spatial autocorrelation algorithms, especially *Kriging*, are capable of maintaining high performance, whether obtaining high true positive and/or low false positive rate, at most given user-specified neighbourhood.

In terms of performance stability with respect to neighbourhood size for situations of region outlier (table not shown), test statistics of IDWP for AUC, PAUC TPR, and PAUC FPR is not significant at a 5% level. All other algorithms were significant. That is, IDWP is the lone algorithm in which altering its nearest neighbour parameter does not influence performance significantly.

## 5.5. Algorithm Performance Similarity

Given that there are ten algorithms, nine NN settings, and three ROC performance measures, which give 1,215 totals number of Wilcoxon tests, the best approach at summarizing all tests is by way of counting the number of times the test statistic is greater than a particular significance level. If the p-value is greater than a significance level, then the null hypothesis is retained, meaning that the difference between population means is expected to be zero, which implies that the algorithm performance is statistically similar. However, given that the Wilcoxon is a two-tailed test, if the test statistics are less than the significant level, conclusions cannot be made regarding which of the two algorithms performs better.

**Table 5.4: Number of non-significant Wilcoxon tests at 1% significance out of 27 tests performed for each combination of algorithm for point outlier**

| | Spatial | Median | Local | Scatter | SLOM | Weighted | IDWP | Krige | AvgDiff |
|---|---|---|---|---|---|---|---|---|---|
| Spatial | - | | | | | | | | |
| Median | | - | | | | | | | |
| Local | | | - | | | | | | |
| Scatter | | | 3 | - | | | | | |
| SLOM | | | 3 | 11 | - | | | | |
| Weighted | 4 | 2 | | | | - | | | |
| IDWP | 3 | 3 | | | | | - | | |
| Krige | 3 | 2 | | | | 21 | | - | |
| AvgDiff | 1 | 5 | | | | 1 | | 4 | - |
| SOTest | 1 | 4 | | | | 9 | | 5 | 11 |

Table 5.4 summarizes the Wilcoxon tests for all nine NN settings and three ROC performance measures a 1% significant levels. Each number in the table represents the number of times the p-value is greater than the significance level, or the number of times the difference of means are within the specified confidence level. For example, Table 5.4 shows that the Wilcoxon tests between Kriging and Weighted 21 times a p-value greater than 0.01 was obtained, or 21 times that the difference of means between Kriging and Weighted is within the 99% confidence interval. The total number of possible counts is 27 given that there are 9 NN settings and 3 ROC performance measures.

As evidence in Table 5.4, there is strong evidence that the performance of *Kriging* and *Weighted* are similar across the NN settings. There is also some evidence that *SOTest* and *AvgDiff*, *SLOM* and *Scatter,* and *SOTest* and *Weighted* are statistically similar. In the case of *Spatial* and *Median*, both obtain a few matches with spatial autocorrelation algorithms, which indicate performance similarities at specific NN settings. This is evidenced in the sensitivity analysis in Section 5.2.2, *Spatial* and *Median* are most similar to spatial autocorrelation algorithms when NN is small However, as NN increases, the performance of *Spatial* and *Median* drop substantially while performance of spatial autocorrelation algorithms remains stable.

The algorithm performance similarity for region outliers is shown in Table 5.5. Now, the total number of possible counts is 48 as there are 3 ROC measures (AUC, PAUC TPR, and PAUC FPR), 8 NN settings, and 2 region outlier settings (region 2 and region 5). *Spatial* and *Median* perform identical, along with *Weighted* and *SOTest*. Other than this, there is moderate similarity between *Krige* and *Weighted*, and *SOTest* and *Krige*. Note that algorithms considering spatial autocorrelation differ statistically to algorithms without considering spatial autocorrelation (Table 5.4). However, in terms of region outliers (Table 5.5), there is a mix of similarities between algorithms considering spatial autocorrelation to algorithms without spatial autocorrelation. *SLOM* and *Local* are similar to *Krige, SOTest*, and *Weighted.* Such trend suggests that algorithms considering spatial autocorrelation are most suitable for detecting point outliers, while spatial autocorrelation algorithms work best for region outliers.

**Table 5.5: Number of non-significant Wilcoxon tests at 1% significance out of 48 tests performed for each combination of algorithm for region outlier**

| | Spatial | Median | Local | Scatter | SLOM | Weighted | IDWP | Krige | AvgDiff |
|---|---|---|---|---|---|---|---|---|---|
| Spatial | - | | | | | | | | |
| Median | 48 | - | | | | | | | |
| Local | | | - | | | | | | |
| Scatter | | | 4 | - | | | | | |
| SLOM | 8 | 8 | 1 | | - | | | | |
| Weighted | | | 3 | 8 | 5 | - | | | |
| IDWP | | | 2 | | 3 | | - | | |
| Krige | 1 | 1 | 4 | 3 | 9 | 16 | | - | |
| AvgDiff | 3 | 3 | | | 3 | | | 3 | - |
| SOTest | | | 3 | 8 | 5 | 48 | | 16 | |

## 5.6. Effects of Spatial Outliers on GLS Regression

Table 5.6 shows the results of the bias in coefficient estimation by incorporating the different methods of outlier removal (recall Equation 1, pg. 47).

Unlike algorithm performance depicted in the ROC analysis that consisted of 20 equiprobable simulations, here 100 simulations are conducted. Because coefficient estimation depends more on the properties of each simulated field, 20 simulations may not be enough to obtain a reliable coefficient estimate. For instance, in Sections 5.2 and 5.3, it has been determined that the performance of most algorithms is consistent at different detection rates; however algorithm performance is less dependent on model coefficients. Thus, a need exist to obtain coefficient estimates from more trials. The approach is to simulate additional 80 on-farm trials with the original simulation parameters. Then, for each of the 100 simulated fields, one thousand random observations are contaminated by introducing an error; 500 random observations are point outliers while the remaining 500 are random region outliers. Each spatial algorithm is applied to the simulations, and the top 5% are removed for each algorithm before spatially aggregating the dataset. Finally, the geostatistical regression approach is applied to estimate coefficients.

An assumption must be made to effectively compare the coefficients. First, because unknown effects have been introduced to the simulations by spatial aggregation, it is assumed that the coefficients in *Clean* are the initial input coefficients. For example, $g_1$is initially set to 3 before spatial aggregation. After spatial aggregation, $g_1$is unknown, and has to be estimated. And the closest estimate to $g_1$ is *Clean* as it does not contain any spatial outliers prior to the spatial regression analysis

Table 5.6 indicates the capabilities of each spatial outlier removal method to effectively estimate coefficients. Methods *Raw, Global*, and *Random* produce unacceptable estimations, each obtaining at least four significantly different coefficient estimates out of the total seven. Particularly, *Global* obtains five unacceptable estimates, which leads to the conclusion that the choice of utilizing global outlier tests to detect spatial outliers will lead to making totally wrong estimates of agricultural fields. *Random* produces better estimates than *Global* although 1,000 in-lying observations were incorrectly removed. This observed contrast is explained by the fact that *Global* eliminates all extreme observations which have a significant effect on the statistics of each simulation, and in the case of *Random*, the removed in-lying observations may not have as significant impact as evidenced in *Global*. For example, *Global's* estimated range of autocorrelation of 137 m is significantly lower than the true value of 150 m.

*Local, Scatter,* and *SLOM* are also incapable algorithms, as each obtain three unacceptable estimates. *Spatial, Median,* and three spatial autocorrelation techniques, *Weighted, IDWP,* and *SOTest* each obtain two unacceptable estimates. Lastly, *Kriging* and *AvgDiff* only produce one unacceptable estimate. Note that almost all techniques are incapable of correctly estimating the farmer's innovative treatment 1 ($g_1$) and the nugget (the nugget-to-sill ratio), which may suggest that the spatial aggregation introduced somewhat substantial unknown effects, and/or the lack of simulated iterations.

Table 5.7 is a supplementary table to Table 5.6. It depicts the empirical Type I errors of the t-tests on coefficients at the 5% significance level. Type I error (coefficient $d_2 = 0$, and $g_2 = 0$) is an important assessment criterion for decision-making. Table 5.6 demonstrate the percentage, or the number of times the Type I error occurred to each technique given a 5% significance level. For instance, in *Global*, the frequency of Type I error for $d_2$ and $g_2$ are 12% and 25% respectively. Given that 100 simulations were performed and tested at 5% significance level, *Global* technique would have led farmers to believe that environmental variable 2 ($d_2$) has a significant effect on crop yield (i.e. $d_2 \neq 0$) in 12 out of 100 simulations, while *Global* depicts innovative treatment 2 ($g_2$), which does not influence crop ($g_2 = 0$), having an effect on yield in 29 of the total 100 simulations.

Evidence in Table 5.7 reinforce the idea that in a decision-making context, *Raw, Global,* and *Random* produce relatively elevated Type I errors, followed by *Scatter, Local,* and *SLOM.* The remaining techniques obtain similar Type I error frequencies to *Clean* without evidence of superiority. However, given that the high frequency of Type I errors suggests again, the lack of simulations and/or the effects of spatial aggregation and/or the choice of spatial model may have caused the Type I error divergence. That is, given that the Type I error is tested at a 5% level; the frequency of errors should converge to 5%, which is not the case as evidenced in Table 5.6.

**Table 5.6: Coefficient Estimation.** *Raw* indicates spatial aggregation and spatial regression model performed without any prior spatial outlier removal. *Clean* indicates all spatial outliers, 1,000 or 5% of the dataset, were correctly removed. *Global* indicates the removal of the 1,000 most extreme observations via Grubbs' Test. *Random* is the incorrect removal of spatial outliers by randomly picking in-lying observations instead of actual true spatial outliers. The numbers indicate the mean value over the 100 simulations, while the ones in parenthesis refer to the standard error. Numbers in bold indicate that they are significantly different from the true coefficient (coefficient in *Clean*) by way of a paired t-test at the 5% critical level.

| | $a$ | $d_1$ | $d_2$ | $g_1$ | $g_2$ | $range$ | $nugget*$ |
|---|---|---|---|---|---|---|---|
| Clean | 76.06(0.50) | 5.77(0.64) | 0.14(0.83) | 2.93(0.06) | -0.02(0.06) | 150.22(3.94) | 0.04(0.00) |
| Raw | **76.16**(0.57) | 5.66(0.74) | **-0.06**(1.00) | **2.88**(0.05) | -0.03(0.05) | 150.62(3.23) | **0.03**(0.00) |
| Global | **76.45**(0.42) | **4.77**(0.55) | 0.02(0.67) | **2.59**(0.04) | -0.03(0.04) | **137.53**(2.81) | **0.07**(0.00) |
| Random | 75.97(0.48) | 5.62(0.62) | **-0.13**(0.82) | **2.88**(0.06) | -0.02(0.06) | **151.03**(3.90) | **0.04**(0.00) |
| Spatial | 76.11(0.50) | 5.82(0.65) | **-0.04**(0.83) | **2.88**(0.06) | -0.01(0.06) | 150.65(3.84) | 0.04(0.00) |
| Median | 76.14(0.50) | 5.71(0.64) | -0.03(0.85) | **2.81**(0.06) | -0.02(0.06) | 150.88(3.95) | **0.03**(0.00) |
| Local | 76.05(0.51) | 5.85(0.66) | **0.11**(0.87) | **2.86**(0.06) | -0.02(0.06) | 150.65(3.95) | **0.03**(0.00) |
| Scatter | 76.02(0.49) | 5.88(0.65) | 0.17(0.84) | **2.82**(0.06) | -0.02(0.06) | **150.90**(3.87) | **0.03**(0.00) |
| SLOM | **76.29**(0.48) | 5.62(0.61) | **-0.35**(0.79) | 2.95(0.06) | -0.02(0.06) | 150.66(3.96) | **0.05**(0.00) |
| Weighted | 76.08(0.51) | 5.77(0.66) | 0.08(0.85) | **2.86**(0.06) | -0.02(0.06) | 150.50(3.99) | **0.03**(0.00) |
| IDWP | 76.09(0.50) | 5.78(0.65) | 0.05(0.85) | **2.86**(0.06) | -0.01(0.06) | 150.77(3.91) | **0.03**(0.00) |
| Krige | 76.09(0.45) | 5.75(0.59) | 0.08(0.77) | 2.95(0.06) | -0.02(0.06) | 150.67(3.89) | **0.04**(0.00) |
| AvgDiff | 76.04(0.49) | 5.84(0.63) | 0.13(0.83) | **2.85**(0.06) | -0.01(0.06) | 150.45(3.89) | 0.04(0.00) |
| SOTest | **76.24**(0.50) | 5.77(0.64) | -0.04(0.84) | **2.87**(0.06) | -0.03(0.05) | 150.84(3.90) | 0.04(0.00) |
| True values | 76.00 | 6.00 | 0.00 | 3.00 | 0.00 | 150.00 | 0.05 |

* – $nugget$ refers to the nugget effect, the nugget-to-sill ratio

### Table 5.7: Frequency of Type I Errors

| | $Percentage\ (\%)$ | |
|---|---|---|
| | $d_2$ | $g_2$ |
| Clean | 12 | 25 |
| Raw | 15 | 26 |
| Global | 16 | 29 |
| Random | 15 | 28 |
| Spatial | 13 | 26 |
| Median | 13 | 26 |
| Local | 15 | 28 |
| Scatter | 15 | 29 |
| SLOM | 16 | 29 |
| Weighted | 13 | 26 |
| IDWP | 13 | 28 |
| Krige | 13 | 26 |
| AvgDiff | 13 | 26 |
| SOTest | 13 | 26 |

The parallel coordinates plot for all outlier removal techniques is shown in Figure 5.23. The parallel coordinates plot is a visualization tool to explore high-dimensional data with multiple variables. The parallel lines/axes represent each dimensional space, in this case, the 14 outlier removal techniques; the colours represent each variable, in this case, the four coefficients. The y-axis depicts the coefficient value, and each line represents a single simulation result. Sharp angles in a line or line-crossing imply that the coefficient estimate is substantially incorrect for that particular parallel line (outlier technique). Similarly, a straight line along all parallel lines indicates that all techniques obtained similar or identical estimates. Figure 5.23 mostly depicts *Global* as having sharp angles and line-crossings which indicates that most of its coefficient estimates are far different that all other techniques. *Local, Scatter, SLOM,* and *IDWP* are depicted to have few sharp angles, but other than these, the plot does not depict substantial sharp angles or line-crossings. This indicates that no clear evidence exists about which outlier techniques are most successful in coefficient estimation and subsequent decision-making.

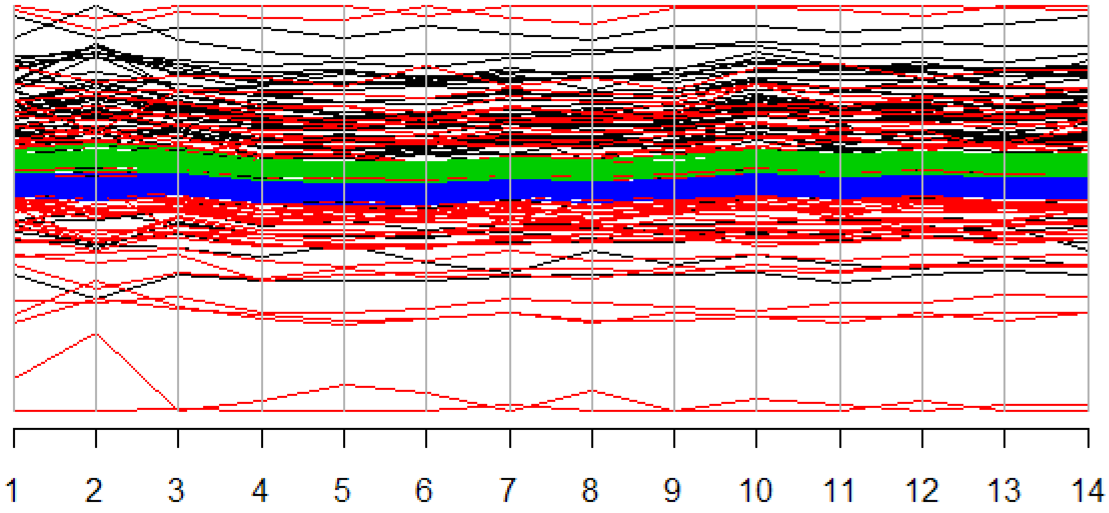


**Figure 5.23: Parallel Coordinates Plot of Coefficients**

1:Raw, 2:Global, 3:Random, 4:Spatial, 5:Median, 6:Local, 7:Scatter, 8:SLOM, 9:Weighted, 10:IDWP, 11:Krige, 12:AvgDiff, 13:SOTest, 14:Clean
Red: $d_1$, Black: $d_2$, Green: $g_1$, Blue: $g_2$

## 5.7. Discussion of Findings

Previous work in spatial outlier detection overlooks the quantitative performance of detection algorithms and lacks the comparison of the numerous detection algorithms proposed by various authors. That is, the studies on spatial outlier detection algorithm are not as comprehensive given that comparisons are made between few algorithms, usually three or four. In this work, the objective is to compare multiple spatial outlier detection algorithms in hopes to determine their performance and the conditions in which these algorithms perform best and worst. However, as comprehensive as this study can be, the weakness remains in the fact that no real-life dataset is utilized to conduct the analysis, which may limit the results to a limited range of outcomes. The main reason at rejecting the usage of real-life datasets is that the assessment of spatial outlier detection algorithm performance will be flawed as all spatial outliers are not known in advance. Identifying all spatial outliers in a real-life dataset is likely impossible since natural variability can introduce confusion. Even if the possibility exists for identifying all spatial outliers, the time requirement for this feat would be substantial. Thus, having a simulated dataset with known spatial outliers seems the most feasible approach at determining spatial outlier detection algorithm performance.

On another note, although the sensibility of ROC performance measures and variation is studied, the sensibility of the error term added to the simulated dataset in relation to the ROC measures is excluded from analysis. Even though not reported, it is found the higher value added (or subtracted) to the original yield values, the better the ROC values (AUC, PAUC TPR, PAUC FPR) for all algorithms. For example, AUC for *Spatial* is on average 92% for point outlier with errors having a mean value of 7 bu/acre. At a mean of 3.5 bu/acre, *Spatial* obtains about 88% AUC. This omission is due to the fact that varying the error term affects equally all algorithms in terms of their ROC performance, which makes sense because no algorithm should have a special association with the value of the error; they should instead have an association with the location of the error in the spatial dataset, as evidenced. For instance, *AvgDiff* is considerably effective in detecting spatial outliers clustered together. Similarly, although not in the analysis, the change of TPR and FPR threshold for PAUC analysis affects equally all spatial outlier algorithms. Keeping these considerations in mind, the following section provides a technical discussion of the results.

Shekhar et al. (2003) introduces *Spatial* and *Scatter*, but do not to provide evidence of algorithm performance. Similarly, Kou et al. (2007) compares *Spatial* and *Scatter* against a *Graph-Based* approach, but do not provide information about the performances of *Spatial* and Scatter. In this work, results suggest *Spatial* is a much better spatial outlier detection algorithm than *Scatter*, as *Spatial* obtains higher overall ROC measures (AUC, PAUC TPR, and PAUC FPR), lower ROC measure variation, lower number of significantly different coefficients, and lower number of Type I errors (Table 5.7). *Scatter* obtains poorer ROC performance most probably because it requires the estimation of slope, $m$, and intercept, $b$, both which are sensitive to outliers, and masking and swamping effects. Thus, it can be generalized that spatial outlier detection algorithm s with more operations, particularly involving operations using the *mean*, will most likely be less efficient in detecting spatial outliers.

Lu et al. (2003) are the first to compare the performance of spatial outlier algorithms as they evaluated *Median* and *Spatial*, and concluded *Median* performed better because it detected the top 10 spatial outliers while *Spatial* miss-detected one outlier. Similarly, Wang et al. (2004) and Chen et al. (2008) confirm *Median* is a more robust spatial detection algorithm than *Spatial* because *Spatial* falsely judged spatial objects as outliers in their study. Both studies conclude *Median* is effective in reducing the risk of falsely identifying regular spatial points as outliers, and this work confirms *Median* a superior spatial outlier detection algorithm than *Spatial*. However, this is true only for point outlier situations. For region outlier situations, with the exception of this work, no study has been conducted to compare *Median* against *Spatial*.

For point outlier situations, *Median* is statistically superior to *Spatial* in all ROC aspects. *Median* computes $f_{aggr}$ with the $median$ and standardizes $f_{diff}$ with $median$ and $MAD$. Thus, effects of masking and swamping are more properly suppressed when $f_{aggr}$ is computed with $median$ than the $mean$. However, in terms of region outlier situation, *Median* and *Spatial* obtain identical AUC, PAUC TPR, and PAUC FPR. This suggests that both algorithms obtain identical outlier scores. This may be accredited to the low variance in $f_{aggr}$ due to absent extreme values. A spatial neighbourhood that contains a region outlier will be of low variance because the outliers tend to have similar values. Therefore, $f_{aggr}$ will be identical or very similar when computed either by $mean$ or $median$. Overall, *Spatial* and *Median* are very similar; *Median* is slightly superior, but both obtain the same results in terms of coefficient estimation (Table 5.8), suggesting that the slight superiority in performance is not of much importance for site-specific decision-making.

**Table 5.8: Summary of Results.** Overall Performance is calculated by standardizing all the AUC, PAUC TPR, and PAUC FPR values to percentage up to 100% and a nested average approach is applied in order to avoid weighting the two region outlier situations (region outlier 2 and region outlier 5) in the calculation. Average Variation refers to the average standard deviation of all ROC values that were tested.

| | ROC Curve | | | Coefficient Estimation | |
|---|---|---|---|---|---|
| Algorithm | Overall Performance | Average Variation | Average Performance Stability | Number of Significant Coefficients | Combined Type I Errors |
| *AvgDiff* | 71.8 | 0.26 | 0.01 | 1 | 39 |
| *IDWP* | 58.3 | 0.30 | 0.53 | 2 | 41 |
| *Kriging* | 65.1 | 0.62 | 0.18 | 1 | 39 |
| *Local* | 55.4 | 0.30 | 0.00 | 3 | 43 |
| *Median* | 68.4 | 0.30 | 0.00 | 2 | 39 |
| *Scatter* | 60.3 | 0.42 | 0.00 | 3 | 44 |
| *SLOM* | 55.3 | 1.14 | 0.02 | 3 | 45 |
| *Spatial* | 67.3 | 0.30 | 0.00 | 2 | 39 |
| *SOTest* | 65.5 | 0.28 | 0.14 | 2 | 39 |
| *Weighted* | 64.8 | 0.29 | 0.06 | 2 | 39 |

Chawla and Sun (2006) compare *SLOM* and *Spatial*, and conclude *SLOM* is sharper in detecting spatial outliers. However, their performance lacked quantitative evidence. According to the results in this work, *Spatial* is a better spatial outlier algorithm than *SLOM* in terms of higher overall ROC performance, lower variation in ROC measures, and less number of significantly different coefficients and Type I errors. Similar to *SLOM*, *Local* is a poor spatial outlier algorithm. Both *SLOM* and *Local* are the worst spatial outlier detection algorithms, as they obtain the lowest overall ROC performance, and lowest correct coefficient estimates.

Unlike all other spatial outlier algorithms, *SLOM* and *Local* require the calculation of two values that are influenced by the neighbourhood structure of the dataset. *SLOM* is the product between a difference function and an oscillation parameter while *Local* is a difference function divided by the neighbourhood's standard deviation. There may be instances where the one of those two components may not be able to clearly distinguish between spatial outliers. *SLOM*'s oscillation parameter and *Local* standard deviation may add additional error to the outlier score computation. In the case of *SLOM*, a high oscillation parameter multiplied by a low difference score will produce a similar score to a low oscillation parameter multiplied with a high difference value.

A unique feature of *SLOM* is that it uses a deterministic value to capture the neighbourhood variation, which is essentially based on a count of neighbouring observations which are larger or smaller than that of the observation (whichever returns the more neighbours) divided by the average of the neighbour's difference value. Changes in the number of neighbour count substantially affect the resulting computation of the outlier score (Figure 5.24).

In Figure 5.24(a), the count in SLOM's oscillation parameter is 4 since four observations are higher and four are lower than the value at (2,2). In Figure 5.24(b), three observations (1,1), (2,1), and (3,1) are changed so their value are a bit larger but very similar to (2,2). In this case, there are six neighbours larger and two neighbours smaller than (2,2). Thus, the oscillation parameter is calculated with six (the highest among the two counts), producing a value of 0.49, which is larger than in Figure 5.24(a). However, the local standard deviation for both Figure 5.24(a) and 5.24(b) remain unchanged, suggesting that counting the number of nearest neighbours may not be a good indicator for representing the neighbourhood variation in point outlier situations. On the other hand, in region outliers, if (2,2), (1,1), (2,1), and (3,1) is defined as a region outlier because of similar values, the oscillation parameter at (2,2) would correctly incorporate the region neighbourhood because of their similarities.

Xue et al. (2008) argue *SLOM* is more biased in detecting global outliers. However, it seems *SLOM* is a much more appropriate algorithm for region outlier, although it was proposed to detect point outliers in the first place (Chawla and Sun, 2006). This is shown inSection 5.3 where *SLOM* performs much better in region outlier situations than in point outlier situations (Section 5.2). However, *SLOM* obtains the highest performance variation, perhaps due to its deterministic value representing neighbourhood variation, which suggests that it does not perform consistently on different dataset.

**Figure 5.24(a)**

| | | | |
|---|---|---|---|
| 3 | 0.64 | 0.80 | 0.00 |
| 2 | 0.85 | 0.83 | 1.00 |
| 1 | 0.61 | 0.87 | 0.85 |
| | 1 | 2 | 3 |

SLOM

$$d_{(2,2)} = 0.15$$
$$B_{(2,2)} = 0.43$$
$$Score_{(2,2)} = 0.29$$

Local Area Mean

$$f_{diff(2,2)} = 0.16$$
$$f_{aggrsd(2,2)} = 0.31$$
$$Score_{(2,2)} = 0.51$$

**Figure 5.24(b):**

| | | | |
|---|---|---|---|
| 3 | **0.85** | 0.80 | 0.00 |
| 2 | **0.85** | 0.83 | 1.00 |
| 1 | **0.85** | 0.87 | 0.85 |
| | 1 | 2 | 3 |

SLOM

$$d_{(2,2)} = 0.11$$
$$B_{(2,2)} = 0.49$$
$$Score_{(2,2)} = 0.37$$

Local Area Mean

$$f_{diff(2,2)} = 0.07$$
$$f_{aggrsd(2,2)} = 0.31$$
$$Score_{(2,2)} = 0.21$$

**Figure 5.24: Comparing *SLOM* and *Local***

On the other hand, neighbourhood standard deviation may be responsible for the *Local's* inferior AUC, PAUC TPR, and PAUC FPR performance, especially for region outliers. For detecting region outliers, $f_{diff}$ for true outliers will most likely be a small value because $f_{aggr}$ is masked by several region outliers in the neighbourhood. However, *Local's* $f_{aggrsd}$ will be also small value since small variation occurs in neighbourhoods with region outliers. As a result, the detection of spatial outliers for *Local* will not truly reflect region outliers.

Kou et al. (2006) conclude *Weighted* and *AvgDiff* are better spatial outlier detection algorithms than *Spatial*. However, in this work, this claim is only applicable to point outlier situations. For region outlier detection, *Spatial* is superior to *Weighted* given that *Weighted* assigns more weight to adjacent neighbours, which can be region outliers. *SOTest* and *Weighted* are very similar algorithms when applied to point outliers, but identical when applied to region outliers. In addition, *Weighted* and *IDWP* obtained results that are relatively distinct, as *Weighted* and *IDWP* are statistically different (Table 5.4 & 5.5). This is particularly revealing as *Weighted* and *IDWP* are similar algorithms with the only difference that the distance decay function of *IDWP* is twice the value in *Weighted*. *IDWP* has a quadratic distance decay function, which allocates more importance to closer neighbours for the computation of $f_{aggr}$ than *Weighted's* linear distance decay function.

In this regard, masking problems are exacerbated in *IDWP* when spatial outliers are present or when high natural variability is present. For example, the nearest observations to each spatial outliers would give more importance to the spatial outliers in the calculation of $f_{aggr}$, resulting in erroneous estimated neighbourhood average value. This is evidenced in the region outlier situations where *Weighted* substantially outperforms *IDWP* in all ROC performance measures. Similarly, a noisy local neighbourhood where substantial variability exists would introduce more confusion to outlier scores in *IDWP* than in *Weighted*. This is particularly evidenced in region outlier scenarios where *IDWP* obtains significantly inferior performance than Weighted, which suggest that *IDWP* is most susceptible to masking effects. Therefore, the choice of the power parameter for the distance decay function should be kept to a minimum value.

In point outlier scenario, *Kriging* obtains better ROC performance than *Weighted* and *IDWP*, although not significantly higher than *Weighted*, by being able to model the changes in spatial autocorrelation. Additionally, *Kriging* obtains a higher test statistic for the ANOVA and Kruskall-Wallis, implying higher neighbourhood stability among all other algorithms. *Kriging* is unique in the calculation of distance weights, as it depends on the autocorrelation structure set by the semivariogram. For instance, neighbours which are far away and not autocorrelated obtain negative weights. In contrast, algorithms *Weighted, AvgDiff, IDWP,* and *SOTest* incorporate spatial autocorrelation by assigning positive weights to all neighbours. Thus, neighbours which are not autocorrelated will obtain a minor but positive weight, and will add an error to the computation of $f_{aggr}$.

In addition, unlike all other spatial algorithms, *Kriging* requires additional input parameters from the semivariogram. *Kriging* has to first compute the empirical semivariogram and then model it to obtain the nugget, sill, and range for the computation of *Kriging* weights. There is an additional uncertainty about selecting the correct semivariogram model and semivariogram parameters which can result in reduced performance (Table 5.9).

**Table 5.9: Kriging ROC performance measures at 8 NN**

| | Single Outlier | | | Region Outlier (size 5) | | |
|---|---|---|---|---|---|---|
| Model | AUC (%) | PAUC TPR (%) | PAUC FPR (%) | AUC (%) | PAUC TPR (%) | PAUC FPR (%) |
| Spherical | 94.5 (0.08) | 15.0 (0.02) | 3.3 (0.02) | 69.3 (0.07) | 3.9 (0.08) | 0.7 (0.02) |
| Gaussian | 94.2 (0.08) | 14.7 (0.02) | 3.2 (0.02) | 73.3 (0.14) | 4.9 (0.05) | 0.9 (0.02) |
| Exponential | 93.4 (0.19) | 14.2 (0.04) | 3.1 (0.04) | 65.6 (0.11) | 3.3 (0.08) | 0.7 (0.02) |
| Power | 93.4 (0.10) | 14.2 (0.09) | 3.1 (0.01) | 62.9 (0.15) | 3.0 (0.04) | 0.6 (0.03) |

Note: reported mean and, in parenthesis, standard error for the 20 simulated datasets

Although all ROC measures are very similar for single outliers, each semivariogram model obtains statistically different performance measures given the low standard error value (Table5.9). This is suggestive the choice of NN for *Kriging* is irrelevant as long as the correct semivariogram model is selected. In addition, in region outlier situations, the region outliers introduce error to the computation of the semviariogram parameters, particularly the nugget, which in turn cause *Kriging* to be a very inconsistent algorithm as evidenced in the high standard deviation of AUC, PAUC TPR, and PAUC FPR.

Another major drawback of utilizing *Kriging* is its computational time complexity (Figure 5.27). Each algorithm was run on a DELL laptop equipped with an Intel Core i7 820 QM at 1.6 GHz and

4 GB of RAM. The procedures taken to determine computation time without bias was to run a single algorithm, turn the laptop off when completed, wait a couple of minutes, re-boot the laptop, and run the subsequent algorithm.

*Spatial* and *Scatter* obtain the lowest computation time because they are mostly composed of basic operations. As predicted, *Median* is more complex than *Spatial* as it takes about twice the computation time given that $median$ and $MAD$ are more complex operations than $mean$ and $standard\ deviation$. Also, spatial autocorrelation algorithms are approximately twice the computation time of *Spatial* mainly because of their computation of distance weight for each observation. *Local* and *SLOM* take about three times more than *Spatial* and *Scatter* because they both have to calculate two local statistics: local centre and local spread for each observation's neighbourhood. And the computation time for *Kriging* is about nine times more than all other spatial autocorrelation algorithms because of the combination of computing the empirical and model semivariogram, and for all observations, matrix multiplication and matrix inversion to calculate weights, and of course, calculating the neighbourhood function.

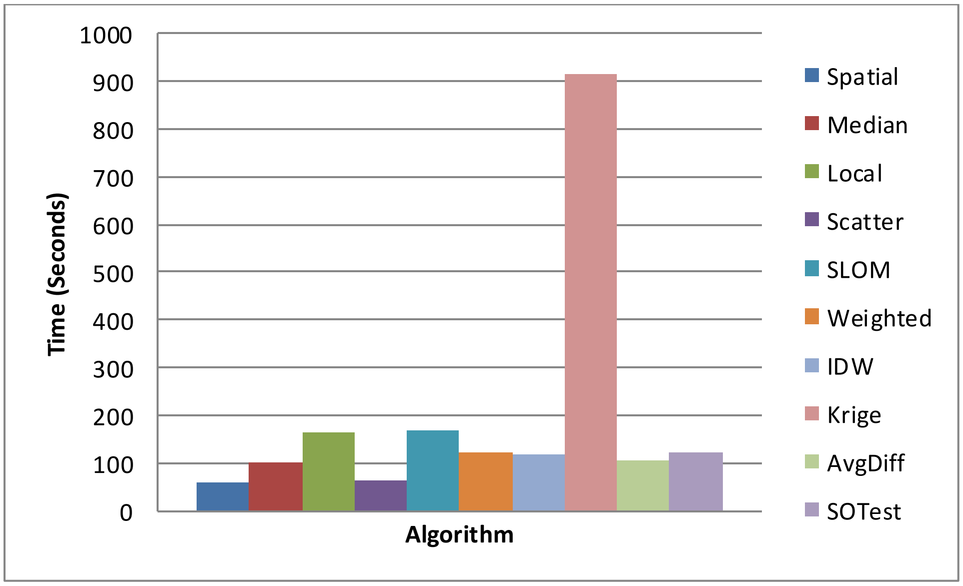


**Figure 5.25: Computing time of spatial outlier algorithms**

Thus, according to the results summarized in Table 5.7, the best algorithm is *AvgDiff*. For single outlier scenario, *AvgDiff* obtains the highest AUC, PAUC TPR, and PAUC FPR at most NN settings and lowest variation for all ROC measures. For region outlier scenario, it obtains the lowest performance decay for AUC and PAUC TPR, highest AUC and PAUC TPR performance at all NN and region outlier size settings. Additionally, *AvgDiff* obtains a relatively fast computation time. Two technical reasons can be formulated on why *AvgDiff* is the better algorithm.

*AvgDiff* compares an observation with each of its neighbours on a one-by-one basis and then averaging the comparisons, whereas all other algorithms start by averaging the neighbourhood value and then making comparisons with the average neighbourhood value. This is advantageous because the averaging of neighbourhood values before comparison may conceal their variance (Kou et al., 2006). For example, if one observation $x$ has a value of 50, with two neighbours of value 0 and 100 that are spaced evenly so distance weight will be 0.5 and 0.5, then Weighted's $f_{aggr}(x)$ will be $(0 \times 0.5) + (100 \times 0.5) = 50$, and $f_{diff}(x)$ will be $50 - 50 = 0$. However, 0 and 100 are quite different from 50. *AvgDiff* retains variance by first calculating the absolute differences, $|0 - 50| = 50$ and $|100 - 50| = 50$, and then calculating the weighted average, $f_{diff}(x) = (50 \times 0.5) + (50 \times 0.5) = 50$ . Weighted's $f_{diff}(x)$ of 0 is quite different from *AvgDiff's* $f_{diff}(x)$ of 50. Thus, the first advantage of *AvgDiff* is its capability of properly adapting to the neighbourhood variance. When the neighbourhood variance is high, which may be accredited to masking and swamping, *AvgDiff* will reveal it, and when variance is low, *AvgDiff* will obtain the same or similar results as in Weighted.

The second advantage evidenced in *AvgDiff* is that unlike all other algorithms, outlier scores are not normalized, which also allows the algorithm perform faster than other spatial autocorrelation algorithms. Since the difference between an observation and its neighbours are absolute, the resulting scores will not follow a normal distribution, thus normalization is not required (Kou et al., 2006). Normalization may add additional confusion to detecting spatial outliers since the distribution of $f_{diff}$ will contain outliers, so estimates of centre and spread will be biased. Although the bias may not be substantial, the confusion that will be introduced to the scores will be substantial given the class disproportion between outlier and non-outlier.

Next to *AvgDiff* in overall ROC performance is *Median* and *Spatial* (Table 5.7). Both are very influenced by the NN used to define the neighbourhood aggregate function. In detecting point outliers, ROC performance rapidly decays with increasing NN; however, in region outliers, ROC performance increases rapidly with increasing NN. Spatial outlier detection algorithms considering spatial autocorrelation, *Kriging*, *SOTest*, and *Weighted*, are the subsequent algorithms. The ROC performances of these three are not influenced by the change of NN; however, they cannot properly deal with region outliers because their neighbourhood aggregate function is computed by incorrectly assigning higher weights to adjacent outliers that are present in the spatial neighbourhood. Finally, *Scatter, IDWP, SLOM,* and *Local* are at the bottom four in overall performance. *Scatter* requires the estimating slope and intercept, both which are affected by spatial outliers. *IDWP* performs poorly especially for detecting region outliers because its power function assigns more weight to adjacent spatial outliers than other spatial autocorrelation algorithms. And, *SLOM* and *Local* have two local statistics that can introduce confusion to detecting spatial outliers.

Differences in ROC performances can be attributed to the estimation of coefficients. The evidenced trend is that algorithms obtaining higher overall ROC performance for instance, *AvgDiff*, *Spatial, Median, Krige*, obtain better coefficient estimates, and lower Type I error. Similarly, lower performance algorithms such as *SLOM* and *Local* obtain poor coefficient estimates and higher Type I errors. As such, the level of correct decisions made based on the coefficients obtained through the GLS regression approach will be influenced by the spatial outlier detection algorithm chosen for pre-processing. Evidence suggests the possibility of classifying spatial outlier detection algorithm into four classes in terms of their decision-making effect: “poor decisions” (*Random, Global*, and *Raw),* ‘moderate decisions’ (*Local, Scatter,* & *SLOM*), “good decisions” (*Spatial, Median, IDWP, Weighted,* and *SOTest*), and “great decisions” (*Kriging* and *AvgDiff*). However, the difference in decision-making (i.e., coefficient estimates and Type I errors) is not substantial between classes perhaps because of the effect of spatial aggregation or because each spatial outlier detection technique is tested with the same coefficient estimation approach. Investigating different coefficient estimation techniques may provide more depth to the assessment of spatial outlier detection techniques in site-specific decision-making. Overall, there are differences in coefficient estimates if data is pre-processed by removing global extremes versus removing spatial outliers. However, little difference exists regarding the choice of spatial outlier technique.

# CHAPTER 6:

# CONCLUSIONS

## 6.1. Summary

This thesis has set out to investigate the random and systematic error-generating mechanisms that occur during the collection of crop yield data, the performance of detection techniques that are utilized to clean spatial yield datasets, and the effects of cleaned datasets on site-specific decision-making. To determine the correctness of spatial outlier techniques, a geostatistical simulation study was conducted to generate crop yield data that contains known spatial errors in advance. Given the known information about yield errors, the assessment of each spatial outlier technique is conducted as a binary classification exercise, treating each spatial technique as a classifier. Classifier performance was evaluated with the area and partial area under the ROC curve from 80% sensitivity and at 5% false positive rate. The value of each spatial outlier technique for statistical inference in GLS models was investigated with the bias in coefficient estimation of a spatial linear model that utilizes semivariogram parameters of OLS residuals as the spatial correlation structure for a generalized least-squares regression.

The results indicate that in situations with point outliers, techniques which account for spatial autocorrelation are far superior to techniques that do not account for spatial autocorrelation in terms of higher sensitivity and lower false positive detection rate at any given decision threshold. Spatial autocorrelation techniques are also more resistant to changes in the definition of spatial neighbourhood, and obtain more consistent performance results across different datasets than algorithms that do not account spatial autocorrelation. In terms of region outlier situations, the latter are superior in all performance aspects because they are less affected by masking and swamping effects.

In terms of algorithms that do not account for spatial autcorrelation, *Median* obtains better and more consistent performance results because it is composed of robust, outlier-resistant operations that suppress masking and swamping effects. *Scatter*, *SLOM*, and *Local* on the other hand, perform poorly because of additional operations which add unnecessary confusion to the outlier scores. In particular, *SLOM* and *Local* require more computational requirements given their additional local neighbourhood operations.

In terms of spatial autocorrelation techniques, *AvgDiff* obtains the best results because of its ability to reveal variance among neighbours and because its outlier scores do not require standardization. On the other hand, *IDWP* performs relatively poorly because masking and swamping have a substantial effect on the inverse distance weight calculation. *Kriging*, *Weighted*, and *SOTest* are closely similar to *AvgDiff* in performance. However, the computation of *Kriging* is significantly far more complex than all other algorithms, and it also requires further user-input semivariogram parameters. Overall, spatial autocorrelation techniques, especially techniques that assign more weight to closest observations such as *IDWP* and *Kriging*, obtain good performance on single outlier scenario but perform poorly in situations where region outliers are present.

In terms of outlier removal for decision-making, all algorithms have led to different coefficient estimates, and therefore, distinct decisions for site-specific management. For instance, an incorrectly estimated coefficient would have led to a Type I error; suggsting that such coefficient significantly influences yield when in fact it does not, or a Type II error; suggesting that the coefficient is not significant when in fact it is. In both situations, farmers may have made investments to improve the wrong explanatory variable.

However, evidence suggests four distinct classes can be elaborated to classify algorithms in terms of their decision-making effect: ‘poor decisions’ (*Random, Global*, and *Raw),* ‘moderate decisions’ (*Local, Scatter,* & *SLOM*), ‘good decisions’ (*Spatial, Median, IDWP, Weighted,* and *SOTest*), and ‘great decisions’ (*Kriging* and *AvgDiff*).

## 6.2. Implications

Erroneous data and associated variability that result from inconsistent data collection practices can corrupt data analysis and produce poor decisions. The results outlined here will allow a producer to remove many of the harvest yield data points that are potentially problematic. Not only the data mining algorithms are applicable for precision agriculture applications, their algorithms far exceed the common techniques used by the precision agriculture community. Three types of spatial algorithms have been utilized by the precision agriculture community for filtering yield datasets: *Local, IDWP*, and *Kriging.* The data mining community have developed the remaining algorithms.

Both communities have overlooked instances of region outliers, and have only focused on single outlier scenarios. For instance, although *SLOM* obtains better performance in region outlier than single outlier situation, it was never proposed to detect the former. Yield surges are errors that occur randomly, unlikely to occur in the same areas on successive years. In this respect, yields surges are not only single outliers, but region outliers, as outliers can randomly be clustered together. In this regard, the precision agriculture techniques will most likely fail against determining true spatial outliers. What has been determined here is the recommendation to use *Averaged Difference* algorithm for cleaning yield surges and all other spatial datasets that exhibits spatial dependence. Determining the optimal nearest neighbour parameter for the neighbourhood aggregate function is still non-trivial. As evidenced in the results, the recommendation is to specify a large number of nearest neighbours, large enough to capture the region size as *AvgDiff* performance does not decrease substantially with a high nearest neighbour value. In addition to superior performance in scenarios of single and region outliers, and fast computational requirement, correctness of the majority of estimated coefficients is obtained with *AvgDiff*, suggesting it is the best method for pre-processing spatial outliers for crop yield data.

## 6.3. Recommendations for Future Research

Although this thesis has investigated computational effectiveness and efficiency of spatial outlier algorithms in precision agriculture yield datasets, there are still several topics that remain unexplored. The following section addresses the selected topics for further investigation.

First, a need exists to investigate the computational efficiency and correctness of iterative and spatial outlier algorithms: *Iterative Z, Iterative R,* and *Graph-Based.* These algorithms were specifically proposed to deal with masking and swamping problems, but their actual effectiveness remains unknown. These algorithms were left out of the analysis because they are extremely difficult to be evaluated with ROC performance measures as they are sequential outlier techniques based on inward procedures. They do not provide outlier scores, but classify the utmost outlier at each step. In other words, unlike the algorithms evaluate in here, they do not require a thresholding value, but a stopping criteria. As such, they are highly computational intensive.

For example, given the same 5% outlier contamination rate in the dataset, *Iterative Z* and *Iterative R* are estimated to take approximately 1,000 longer than *Spatial*, about 17 hours for a single run. The computational time for *Graph-Based* is projected to be more intensive, depending on the complexity of the spatial neighbourhood definition. Similarly, iterative versions, inward or outward procedures, of other algorithms such as *Median, Weighted,* *AvgDiff,* among others, can be postulated and investigated. Thus, the question that may be posed is "whether iterative spatial algorithms are more effective than non-iterative spatial algorithms in dealing with masking and swamping situations? If so, would the discrepancy in performance offset their high computational complexity?"

Similarly, a need exists to explore graphical methods for spatial outlier detection, mainly the *Variogram Cloud* and *Moran Scatterplot.* The *Variogram Cloud* is based on pair-wise comparisons, which would flag a spatial outlier and its spatial neighbour for all point clouds. Post-processing is required to separate and identify between the real spatial outlier and its neighbour. And *Moran Scatterplot* identifies spatial outliers as points that are situated in the upper left and lower right quadrants of the *Moran* graph, which indicates that the spatial association of these observations is dissimilar to their neighbourhood: they are either low values surrounded by high neighbours or vice versa. They key issue with graphical methods is the difficulty to use ROC performance measures because an additional step to summarize the visualization of spatial outliers into a scalable calculation is required.

Second, multivariate spatial outlier algorithms remain unexplored. In many cases, outliers cannot be detected when multiple non-spatial attributes are considered independent. The standard approach has been to detect spatial outliers for a single attribute, independently of other attributes. Expert filters examine observation outlierness based on one attribute at a time; most commonly crop yield, combine velocity, and crop moisture.

For multivariate attributes, the definition of spatial neighbourhood will remain the same, but the neighbourhood aggregate function, the comparison function, and the statistic test will have to be modified. Additionally, a distance function, such as the Mahalanobis distance, has to be defined to convey the multivariate data space. And the correlation structure of the attributes has to be modelled as well.

Another option for multivariate spatial outlier detection would be to create spatial versions of different data mining outlier algorithms. Distance-based, density-based, clustering-based, and depth-based algorithms are non-parametric techniques that are capable of dealing with high dimensional datasets. The problem remains that they are not capable of detecting spatial outliers, but global outliers. The idea would be to utilize the spatial relationship among observations as an additional variable; however, the weight of each variable remains in question.

Another obstacle is the approach to contaminating multiple attributes. The issue is the lack of knowledge about the relationship between multiple attributes, for example, the relationship between combine velocity, crop moisture, and crop yield. Most importantly, there is a lack of knowledge about the relationship between their outliers. In addition, spatial versions of data mining outlier algorithms will require additional input parameters, which translate to additional uncertainty and more complex sensitivity analyses. These new spatial algorithms will also imply more complex computational requirements, perhaps double the time required for the current spatial outlier algorithms. Thus, the question is "whether multivariate spatial outlier methods are be more convenient and more effective than analyzing spatial outliers on an attribute-by-attribute basis?"

Third, all algorithms for spatial outlier detection do not provide a natural critical value for the final classification of outliers. The final output of each algorithm is a list of all observations with the spatial outlier score. The user is required to decide upon a suitable threshold between the outlier and non-outlier space. This can be accomplished by selecting out a specified percentage of the outlier histogram, for example, selecting the top 5% observations with the highest outlier score, as in this case. The option to automatically select spatial outliers would be to implement histogram-based thresholding techniques. However, there is simply no knowledge about how many outliers are present in the dataset. Therefore, the detection of spatial outliers is very sensitive and dependent on the threshold value. Histogram thresholding remains an impending topic in outlier detection.

An innovative alternative to histogram thresholding involves an entropy-based approach at detecting the number of spatial outliers present in a spatial dataset (Liu et al., 2008). In information theory, entropy is a measure of uncertainty in a random variable (Liu et al., 2008). A dataset with more outlying observations has naturally higher entropy value than one with less outlier. The idea is to continually remove top spatial outliers until the entropy value of the dataset stabilizes, which would imply that most if not all spatial outliers have been removed from the dataset. Given the iterative nature however, this entropy-based method is surely computational intensive for large datasets.

Lastly, further research is needed to develop scalable and numerically stable spatial algorithms with reduced computational requirements on large datasets. Currently, a single *for loop* for nearest neighbour search for large datasets require the usage of substantial physical memory. Processing each algorithm requires approximately 3.0 GB of physical memory on a Dell Intel Core i7 laptop with 4 GB of RAM. Furthermore, the geostatistical or the spatial autoregressive approach for estimating model coefficients would fail because of insufficient memory requirements that are currently ameliorated by spatially aggregating the data.

A promising solution is to parallelize task elements to increase performance by reducing the amount of load over many processors. Parallel computing enables the simultaneous use of multiple computer resources to solve a computation task. The task is broken down into independent sub-tasks, and each sub-task is processed simultaneously on different central processor units (CPUs). However, this solution not only requires compatible and correct computer hardware structure but also well-designed software interface that matches user-end requirements. Solving these requirements would be a major undertaking given the diverse hardware and software configuration. For example, R CRAN lists about 57 packages for parallel computing, each with differing level of usability, performance, and acceptance. Consequently, migrating current spatial outlier algorithms to a parallelized version would be a challenging activity.

# REFERENCES

Adamchuk, V.I., Hummel, J.W., Morgan, M.T., & Upadhyaya S.K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture, 44,* pp. 71 - 91.

Acuna, E., & Rodriguez, C.A. (2004). Meta analysis of outlier detection methods in classification. In *proceedings IPSI 2004,* Venice.

Anselin, L., Bongiovanni, R., & Lowenberg-DeBoer, J. (2004). A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics, 86,* pp. 675 - 687.

Anselin, L., Wook Kim, J., & Syari Ibnu. (2004a). Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems, 6,* pp. 197 - 218.

Arslan, S., & Colvin, T.S. (2002). Grain yield mapping: yield sensing, yield reconstruction, and errors. *Precision Agriculture, 3,* pp. 135 - 154.

Bachmaier, M., & Auernhammer, A. (2004). A method for correcting raw yield data by fitting paraboloid cone. *In AgEng 2004: Proceedings of the Agricultural Engineering Conference*, Session 10, Leuven, Belgium.

Bachmaier, M. (2010). Yield mapping based on moving butterfly neighbourhoods and the optimization of their length and width by comparing with yield data from a combine harvester. *EE'10 Proceedings of the 5th IASME/WSEAS international conference on Energy & Environment.* pp. 76 - 82.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley.

Beck, A.D., Roades, J.P., & Searcy, S.W. (1999). Post-process filtering techniques to improve yield map accuracy. *ASAE/CSAE-SCGR Annual International Meeting,* Toronto, Ontario, July 18-21.

Begienbing, S., Bach, H., & Waldmann, D., & Mauser, W. (2005). Analyses of spaceborne hyperspectral and directional CHRIS data to deliver crop status for precision agriculture. In *Proceedings of the 5th European Conference on Precision Agriculture,* Uppsala, Sweden, pp. 227 - 234.

Blackmore, S. (1998). A yield map primer. In *Proceedings from the Conference on Precision Farming in Japan, USA, and Europe (October, 1998).* Hokkaido, Japan.

Blackmore, S. (2000a). The interpretation of trends from multiple yield maps. *Computers and Electronics in Agriculture, 26*(1), pp. 37 - 51.

Blackmore, S. (2000b). Developing the principles of Precision Farming. In *Proceedings of the ICETS 2000.* China Agricultural University, Beijing, China.

Blackmore, S. (2003). *The Role of Yield Maps in Precision Farming.* (Ph.D. dissertation, Cranfield University at Silsoe, 2003). Retrieved from: *www.cpf.kvl.dk/Papers/SIB_PhD.pdf*

Blackmore, S. & Larscheid, G. (1997). Strategies for managing variability. In *1st European Conference on Precision Agriculture (September 8-10, 1997).* Warwick, United Kingdom.

Blackmore, S., & Marshall, C. (1996). Yield mapping: errors and algorithms. *3rd International Conference on Precision Agriculture.* June 23 - 26. Minneapolis, MN.

Bouma, J. (1997). Precision agriculture: introduction to the spatial and temporal variability of environment quality. pp. 5 - 17. In CIBA Foundation (1997). *Precision Agriculture: spatial and temporal variability of environment quality.* CIBA Foundation Symposium 210. New York: John Wiley and Sons.

Bouma, J., Stoorvogel, J.J., van Alphen, B.J., & Bootlink, H.W.G. (1999). Pedology, precision agriculture and the changing paradigms of agricultural research. *Soil Science Society of America Journal, 63,* pp. 343 - 348.

Brase, T. (2006). *Precision Agriculture.* Thomson Delmar Learning.

Brenning, A., Piotraschke, H., & Leithold, P. (2008). Geostatistical analysis of on-farm trials in precision agriculture. In Ortiz, J.M., & Emery, X (Eds.). *GEOSTATS 2008, Proceedings of the 8th International Geostatistics Congress,* December 1-5, 2008, Santiago, Chile, pp. 1131 - 1136.

Breunig, M.M., Kriegel, H.P., Ng, R.T., & Sander, J. (2000). Identifying local outliers. In *Proceedings of PKDD '99, Prague, Czech Republic, Lecture Notes in Computer Science,* pp. 262 - 270, Springer Verlag.

Bullock, D.S., Swinton, S., & Lowenberg-DeBoer, J. (2002). Can precision agricultural technology pay for itself? The complimentarity of precision agriculture technology and information. *Spatial Data Analysis Workshop of the American Agricultural Economics Association Meetings,* Chicago, IL.(August 4, 2001).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. ACM Computing Surveys, *41*(3)

Chawla, S., & Sun, P. (2006). SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems, 9*(4), pp. 412 - 429.

Chen, D., Lu, C-T., Kou, Y., & Chen, F. (2008). On detecting spatial outliers. *Geoinformatica, 12,* pp. 455 - 475.

Corwin, D.L., & Lesch, S.M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture, 46,* 11 - 43.

Corwin, D.L., & Lesch, S.M. (2010). Delineating site-specific management units with proximal sensors. In Oliver, M.A. (Ed.). *Geostatistical Applications for Precision Agriculture,* pp. 139 - 165. Netherlands: Springer.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology, 17,* pp. 563 - 586.

Cressie, N. (1993). *Statistics for spatial data.* Wiley Interscience.

Davis, G., Casady, W., & Massey, R. (1998). *Precision Agriculture: an introduction*. University Extension, University of Missouri.

De Veaux, R.D., Velleman, P.F., & Bock, D.E. (2005). *Stats: Data and Models* (2nd ed). Addison Wesley.

Diker, K., Heermann, D.F., & Brodahl, M.K. (2004). Frequency analysis of yield for delineating yield response zones. *Precision Agriculture, 5,* pp. 435 - 444.

Dobos, E., & Hengl, T. (2009). Soil mapping applications. In Hengl,T., & Reuter, H.I. (Eds.), *Geomorphometry: concepts software, applications*. Developments in Soil Science, vol. 33. Elsevier: Amsterdam, Netherlands, pp. 461-479.

Dodd, L.E. & Pepe, M.S. (2003). Partial AUC Estimation and Regression. *Biometrics, 59,* pp. 614 - 623.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), pp. 861 - 874.

Fleming, K.L., Heermann, D.F., & Westfall, D.G. (2004). Evaluating soil color with farmer input and apparent soil electrical conductivity for management zone delineation. *Agronomy Journal, 96,* 1581 - 1587.

Franzen, D.W., Hopkins, D.H., Sweeney, M.D., Ulmer, M.K., & Halvorson, A.D. (2002). Evaluation of soil survey scale for zone development of site-specific nitrogen management. *Agronomy Journa, 94,* pp. 381 - 389.

Fridgen, J.J., Kitchen, N.R., & Sudduth, K.A. (2000). Variability of soil and landscape attributes within sub-field management zones. In Robert, P.C. (Ed.). *Proceedings of the 5th International Conference on Precision Agriculture (July 16-19, 2000).* Bloomington, MN. ASA, CSSA, and SSSA, Madison, WI.

Fridgen, J.J., Kitchen, N.R., Sudduth, K.A., Drummond, S.T., Wiebold, W.J., & Fraisse, C.W. (2004). Management zone analysis (MZA): software for subfield management zone delineation. *Agronomy Journal, 96,* pp. 100 - 108.

Fountas, S. (2004). *System Analysis of Precision Agriculture.* (Doctoral dissertation, The Royal Veterinary and Agricultural University, 2004). Retrieved from *www.cpf.kvl.dk/Papers/Spyros_Fountas_PhD.pdf*

Gebbers. R., & de Bruin, S. (2010). Application of Geostatistical simulation in precision agriculture. In Oliver, M.A. (Ed.). *Geostatistical Applications for Precision Agriculture,* pp., 269 - 303. Netherlands: Springer.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation.* Oxford University Press.

Griffin, T.W. (2009). *Farmers' use of yield monitors.* University of Arkansas Division of Agriculture Factsheet FSA36.

Griffin, T.W. (2010). The spatial analysis of yield data. In Oliver, M.A. (Ed.). *Geostatistical Applications for Precision Agriculture,* pp. 89- 116. Netherlands: Springer.

Griffin, T.W., Dobbins, C.L., Vyn, T., Florax, R.J.G.M., & Lowengberg-DeBoer, J. (2008). Spatial analysis of yield monitor data: case studies of on-farm trials and farm management decision-making. *Precision Agriculture, 9,* pp., 269 - 283.

Haak, D. (2010). (unpublished data from Farm Environmental Management Survey (2006). AAFC, StatsCan)

Hadi, A.S., Rahmatullah Imon, A.H.M., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(1), pp. 57 - 70.

Han, J. & Kamber, M. (2001). *Data Mining: concepts and techniques.* (2nd ed.). Morgan Kaufman Publishers.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E.R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics, 36*(6), pp. 822 - 830.

Hatfield, J.L. (2000). *Precision agriculture and environment quality: challenges for research and education.* The National Arbor Day Foundation. Retrieved from: *www.arborday.org/programs/papers/PrecisAg.pdf*

Havlin, J.L., & Heiniger, R.W. (2009). *A variable-rate decision support tool. Precision Agriculture, 10(4), pp. 356 - 369.*

Hawkins, D. (1980). *Identification of outliers.* Chapman and Hall.

Hengl, T., Heuvelink, G.B.M., & Rossiter, D.G. (2007). About regression-kriging: from equations to case studies. *Journal of Computers & Geosciences, 33*(10), pp. 1301 - 1315.

Hollander, M., & Wolfe, D.A. (1973). *Nonparametric Statistical Methods.* New York: John Wiley & Sons.

Kerry, R., Oliver, M.A., & Frogbrook, Z.I. (2010). Sampling in precision agriculture. In Oliver, M.A. (Ed.). *Geostatistical Applications for Precision Agricutlure,* pp. 35 - 63. Netherlands: Springer.

Kleinjan, J., Chang, J., Wilson, J., Humburg, D., Carlson, G., Clay, D., & Long, D. (2002). Cleaning yield data. *SDSU Publication.*

Knorr, E., & Ng, R. (1997). A unified notion of outliers: properties and computation. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining,* pp. 219 - 222.

Knuth, D. (1998). *The Art of Computer Porgramming,* 2nd Edition, Massachusetts, Addison-Wesley.

Kou, Y., Lu, C-T., & Chen, D. (2006). Spatial weighted outlier detection. In *Proceedings of the SIAM Conference on Data Mining.*

Kou , Y., Lu, C-T., & Dos Santos, R.F. (2007). Spatial outlier detection: a graph-based approach. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence,* pp., 281 -288.

Kühn, J., Brenning, A., Werhan, M., Koszinski, S., & Sommer, M. (2009). Interpretation of electrical conductivity patterns by soil properties and geological maps for precision agriculture. *Precision Agriculture, 10,* pp. 490 - 507.

Lambert, D.M., & Lowenberg-DeBoer, J. (2000). *Precision agriculture profitability review.* Purdue University, West Lafayette, IN.

Lambert, D.M., Lowenberg-DeBoer, J., & Bongiovanni, R. (2003). Spatial regression models for yield monitor data: a case study from Argentina. In *Proceedings of the Agricultural Economics Association Annual Meeting,* Montreal, Canada, July 27 - 30, 2003.

Lee, W.S., Shueller, J.K., & Burks, T.F. (2005). Wagon-based silage yield mapping system. *Agricultural Engineering International: The CIGR Journal, 7,* Manuscript IT 05 003, pp. 1-14.

Liu, X., Lu, C-T., & Chen, F. (2008). An entropy-based method for assessing the number of spatial outliers. *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,* San Jose, California, Novermber 2 - 5, 2008.

Liu, Y., Swinton, S.M., & Miller, N.R. (2006). Is site-specific yield response consistent over time? Does it pay? *Americn Journal Agricultural Economics, 88*(2), pp. 471 - 483.

Long, D. (1998). Spatial autoregressive modelling of site-specific wheat yield. *Geoderma, 85,* pp. 181 - 197.

Lowenberg-DeBoer, J. & Swinton, S.M. (1997). Economics of site-specific management in agronomic crops. In Pierce, F.J., & Sadler, E.J. (Eds). *The state of site-specific management for agriculture,* pp. 369 - 396. Madison, WI: ASA-CSSA-SSSA.

Lu, C-T., Chen, D., & Kou, Y. (2003). Algorithms for spatial outlier detection, In *Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003.

Manchanda, M.L., Kudrat, M., & Tiwari, A.K. (2002). Soil survey and mapping using remote sensing. *Tropical Ecology, 43*(1), pp. 61 - 74.

McBratney, A.B., & Pringle, M.J. (1999). Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture, 1,* pp. 125 - 152.

McBratney, A.B., Odeh, I.A.O., Bishop, T.F.A., Dunbar, M.S., & Shatar, T.M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma, 97*, pp. 293-327.

McBratney, A.B., Mendoça Santos, M.L., & Minansy, B. (2003). On digital soil mapping. *Geoderma, 117*, pp. 3- 52.

McBratney A.B.., & Lagacherie, P. (2004). *Global Workshop on Digital Soil Mapping,* Montpellier.

McBratney, A.B., Minasny, B., & Whelan, B.M. (2005). Obtaining 'useful' high-resolution soil data from proximally-sensed electrical conductivity/resitivity (PSEC/R) surveys. In Stafford, J.V. (Ed.). *Precision Agriculture '05,* pp. 503 - 510. Wageningen, Netherlands: Wageningen Academic Publishers.

McBratney, A.B., Whelan, B., & Ancev, T. (2005a). Future directions of precision agriculture. *Precision Agriculture, 6,* pp. 7 - 23.

McClish, D.K. (1989). Analyzing a potion of the ROC curve. *Medical Decision Making, 9,* pp. 190 - 195.

Ministry of Agriculture Food & Rural Affairs. (2009). *Winter wheat production by county.* Retrieved from http://www.omafra.gov.on.ca/english/stats/crops/ctywwheat09.htm

Moran, M.S., Inosue, Y., & Barnes, E.M. (1997). Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment, 61,* pp. 319 - 346.

Murphy, P.A., Shung, E., Haneklaus, S. (1994). Yield mapping - a guide to improved techniques and strategies. In: *Site-specific Managament for Agricultural Systems,* Robert, Rust, Larson (Eds.), ASA, CSSA, SSSA, Madison, WI, pp. 33.

National Research Council. (1997). *Precision Agriculture in the 21st Century.* Washington, DC: National Academic Press.

Noack, P.H., Muhr, T., & Demmel, M. (2003). An algorithm for automatic detection and elimination of defective yield data. *In Precision Agriculture '03: Proceedings of the 4th European Conference on Precision*

*Agriculture*. Stafford, J.V., & Werner, A. (Eds.), Wageningen Academic Publishers, Wageningen, Netherlands, pp. 445 - 450.

Nolan, S.C., Haverland, W., Goddard, T.W., Green, M., Penney, D.C., Henriksen, J.A., & Lachapapelle, G. (1996). Building a yield map from geo-referencered harvest measurements. *In Proceedings of the 3rd International Conference on Precision Agriculture*. Minneapolis, MN, ASA, CSSA, SSSA, Madison, WI, June 23 - 26, pp. 885 - 892.

Odeh, I.O. A.,, Chittleborough, D.J., & McBratney, A.B. (1992). Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. *Soil Science Society of America Journal, 56,* pp. 505 - 516.

Oliver, M.A. (2010). An overview of geostatistics and precision agriculture. In Oliver, M.A. (Ed.). *Geostatistical Applications for Precision Agriculture,* pp. 1 - 34. Netherlands: Springer.

Pebesma, E.J. (2004). Multivariate geostatistics in S: the gstat package. *Computer & Geosciences, 30,* pp. 683 - 691.

Pedersen, S.M. (2003). *Precision farming: technology assessment of site-specific input application in cereals* (Ph. D. dissertation, Technical University of Denmark, 2003). Retrieved from: http://www.cpf.kvl.dk/Papers/SMPedersen_Thesis/SMPedersen_PhD.pdf.

Pedersen, S. M., Fountas, S., Blackmore, B.S., Gylling, J.L., & Pendersen, J.L. (2004). Adoption and perspective of precision farming in Denmark. *Acta Agriculturae Scandinavica, Section B - Plant Soil Science, 54*(1), pp., 2 - 8.

Pfost, D., Casady, W., & Shannon, K. (1998). Precision Agriculture: Global Postioning System (GPS). University Extension, University of Missouri.

Ping, J.L., & Dobermann, A. (2005). Processing yield data. *Precision Agriulture, 6,* pp. 193 - 212.

Preparata, F., & Shamos, M. (1988). *Computational Geometry: An Introduction*. Springer Verlag.

R Core Development Team (2010). *R: A language environment for statistical computing, reference index version 2.12.1.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org

Rands, M. (1995). The development of an expert filter to improve the quality of yield mapping data. Unpublished Msc. Thesis, Silsoe College, Cranfield University.

Robinson, T.P., & Metternicht, G. (2005). Comparing the performance of techniques to improve the quality of yield maps. *Agricultural Systems, 85,* pp. 19 - 41.

Rossiter, D.G., & Hengl, T. (2002). *Creating geometrically-corrected photo-interpretation, photomosaics and base maps for a project GIS.* Technical note. ITC, Department of Earth System Analysis, Enschede, NL.

Scottish Natural Heritage. (2009). *Boom Fertilizer Spreaders.* Retrieved from: http://www.snh.org.uk/tibre/section2_3_6.htm

Searcy, S.W., Schueller, J.K., Base, Y.H., & Stout, B.A. (1989). Mapping of spatially variable yield during grain combining. *Transactions of the ASAE, 32*(3), pp. 826 - 829.

Seelan, S.K., Laguette, S., Casady, G. M., & Seielstad., G.A. (2003). Remote sensing applications for precision agriculture: a learning community approach. *Remote Sensing of Environment, 88,* pp. 157 - 169.

Shapiro, S.S., & Wilk, M.B. (1965). The analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), pp. 591 - 611.

Shekhar, S., Lu, C-T., & Zhang, P. (2001). A unified approach to detecting spatial outliers. *In Department of Computer Science and Engineering, University of Minnesota, Technical Report TR 01-045,* Retrieved https://www.cs.umm.edu/tech_reports/listing/?year=2001

Shekhar, S., Lu, C-T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *Geoinformatica, 7*(2), pp. 139 - 166.

Shekhar, S., Zhang, P., & Huang, Y. (2005). Spatial data mining. In Maimon, O., & Rokach, L. (Eds.). *The Data Mining and Knowledge Discovery Handbook.* pp. 833 - 851. Springer.

Schumacher, J.A., Lindstrom, M., & Schumacher, T. (2000). An analysis of tilage and water erosion over a complex landscape. In *Proceedings of fifth International Conference on Precision Agriculture (CD),* July 16 - 19, Bloomington, Minnesota.

Simbahan, A., Dobermann, A., & Ping, L. (2004). Screening yield monitor data improves grain yield maps. *Agronomy Journal, 96*(4), pp. 1091 - 1102.

Song, X., Wang, J., Huang, W., Liu, L., Yan, G., & Pu, R. (2009). The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture, 10,* pp. 471 - 487.

Song, X., Wu, M., Jermaine, C., & Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering, 19*(5), pp. 635 - 645.

Stafford, J.V., Ambler, B., Lark, R.M., & Catt, J. (1996). Mapping and interpreting the yield variation in cereal crops. *Computers and Electronics in Agriculture, 14*(2), pp. 101 - 119.

Stafford, J.V. (2000). Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research, 76*(3), pp., 267 - 275.

Strassen, V. (1969). Gaussian elimination is not optimal. *Numer. Math, 13,* pp. 354 - 356.

Sudduth, K.A, & Drummond, S.T. (2007). Yield editor: software for removing errors from crop yield maps. *Agronomy Journal, 99,* pp. 1471 - 1482.

Swinton, S.M., & Lowenberg-DeBoer, J. (2001). Global adoption of precision agriculture technologies: who, when and why? In *Proceedings of the 3rd European Conference on Precision Agriculture,* edited by Grenier, G., & Blacmore, S., Agro Montpellier, Montpellier, France, pp. 557 - 562.

Thylen, L., Algerbo, P.A. & Giebel, A. (2000). An expert filter removing erroneous yield data. *In Precision Agriculture 2000 [CD-ROM]: Proceedings of the 5th International Conference,* edited by Robert et al., ASA, CSSA and SSSA, Madison, WI, 2001

Thylen, L., & Murphy, D.P. (1996). The control of errors in momentary yield data from combine harvesters. *Journal of Agricultural Engineering Research, 64*(4), pp., 271 - 278.

Tobler, W. (1970). A computer movie simulating urban growth in Detroit region. *Economic Geography, 46(*2), pp., 234 - 240.

Top Crop Manager. (2007). *How to do on-farm trials.* Retrieved from: http://www.topcropmanager.com/index.php?option=com_content&task=view&id=1465

van Alphen, B.J., & Stoorvogel, J.J. (2000). A functional approach to soil characterization in support for precision agriculture. *Soil Science Society of America Journal, 64,* pp. 1706 - 1713.

Ver Hoef, J. & Cressie, N. (2001). Spatial statistics: analysis of field experiments. In Sheiner, S.M. & Gurevitch, J. (eds.). *Design and Analysis of Ecological Experiments*, 2nd edition, pp., 289 - 307, Oxford University Press.

Virin, T., Koko, J., Piron, E., Martinet, P., & Berducat, M. (2008). Optimisation-based approach for better centrifugal spreading. *International Journal of System Science, 39*(9), pp. 913 - 924.

Vrindts, E., Mouazen, A.M., Reyniers, M., Maertens, K., Maleki, M.R., Ramon, H., & de Baerdemaeker, J. (2005). Management zones based on the correlation between soil compaction, yield and crop data. *Biosystem Engineering, 92*(4), pp. 419 - 428.

Wang, Z.Q., Wang, S.K., Hong, T. Wan, X.H. (2004). A spatial outlier detection algorithm based multi-attribute correlation. *Proceedingof the 3rd International Conference on Machine Learning and Cybernetics,* Shanghai, Augugst 26 - 29.

Webster, R. & Oliver, M.A. (2007). *Geostatistics for environmental scientists.* John Wiley and Sons.

Werner, A., Jarfe, A., Roth, R., & Pauly, J. (1999). Precision agriculture, a new technology in crop production - will it enhance sustainable development in land use? In Olejnik, J et al. (Eds.). *Sustainability in Land Use: Proceedings of an International Conference,* Poznan, Poland, November 17 - 20, pp. 327 - 342.

Whitley, K.M., Davenport, J.R., & Manley, S.R. (2000). Difference in nitrate leaching under variable and conventional nitrogen fertilizer management in irrigated potato systems. In *Proceedings of fifth International Conference on Precision Agriculture (CD),* July 16 - 19, Bloomington, Minnesota.

Xue, A., Yao, L., Ju, S., Chen, W., & M, H. (2008). Algorithm for fast spatial outlier detection. *The 9th International Conference for Young Computer Scientists.* pp. 1872 - 1877.

Yakushev, V.P., Vure, V.M., & Yakushev, V.V. (2008). Methodology and tools for analyzing on-site data in precision agriculture. *Russian Agricultural Sciences, 34*(6), pp. 431 - 434.

Zhang, X., Shi, L., Jia, X., Seielstad, G., & Helgason, C. (2010). Zone mapping application for precision-farming: a decision support tool for variable rate application. *Precision Agriculture, 11,* pp. 103 -114.

# Appendix A:

# Shapiro-Wilk Test

Analysis of variance test requires the observations, in this case the ROC performance scores, to be normally distributed. The Shapiro-Wilk test of normality tests whether the null hypothesis that a sample came from a normally distributed population (Shapiro & Wilk, 1965). The test statistic is as follows (Shapiro & Wilk, 1965):

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $y_i$ is the *i*th–smallest ROC value in the sample, $\bar{y}$ is the mean ROC value, and $a_i$ is a constant given by:

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

where $m^T = (m_1, \dots, m_n)$ is a vector of the expected value of standard normal order statistics, and $V = v_{ij}$ is the corresponding $n \times n$ covariance matrix.

If the test statistic, $W$, is small enough, the null hypothesis that the sample comes from a normally distributed population is rejected. Table 7.1 through 7.3 provides the test statistic for the Shapiro-Wilk test.

**Table 7.1: p-values from AUC Shapiro-Wilk test**

| | Groups: Number of Nearest Neighbours | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Spatial | 0.85 | 0.91 | 0.85 | 0.60 | 0.64 | 0.65 | 0.96 | 0.97 | 0.66 |
| Median | 0.70 | 0.53 | 0.80 | 0.66 | 0.76 | 0.93 | 0.97 | 0.99 | 0.84 |
| Local | 0.74 | 0.49 | 0.76 | 0.84 | 0.58 | 0.28 | 0.64 | 0.81 | 0.91 |
| Scatter | 0.93 | 0.40 | 0.32 | 0.48 | 0.86 | 0.52 | 0.38 | 0.15 | 0.87 |
| SLOM | 0.02 | 0.01 | 0.38 | 0.32 | 0.34 | 0.15 | 0.03 | 0.02 | 0.03 |
| Weighted | 0.77 | 0.62 | 0.77 | 0.60 | 0.51 | 0.37 | 0.42 | 0.51 | 0.82 |
| IDWP | 0.92 | 0.92 | 0.84 | 0.79 | 0.80 | 0.80 | 0.78 | 0.76 | 0.75 |
| Krige | 0.32 | 0.24 | 0.26 | 0.19 | 0.17 | 0.16 | 0.12 | 0.17 | 0.92 |
| AvgDiff | 0.21 | 0.36 | 0.29 | 0.53 | 0.90 | 0.86 | 0.84 | 0.81 | 0.16 |
| SOTest | 0.35 | 0.22 | 0.14 | 0.08 | 0.12 | 0.18 | 0.19 | 0.17 | 0.62 |

**Table 7.2: p-values from PAUC TPR Shapiro-Wilk test**

| | Groups: Number of Nearest Neighbours | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Spatial | 0.78 | 0.80 | 0.99 | 0.18 | 0.98 | 0.51 | 0.23 | 0.71 | 0.66 |
| Median | 0.95 | 0.70 | 0.65 | 0.68 | 0.28 | 0.65 | 0.64 | 0.70 | 0.84 |
| Local | 0.98 | 0.39 | 0.24 | 0.99 | 0.07 | 0.60 | 0.47 | 0.15 | 0.91 |
| Scatter | 0.78 | 0.27 | 0.06 | 0.67 | 0.88 | 0.28 | 0.65 | 0.08 | 0.87 |
| SLOM | 0.42 | 0.07 | 0.09 | 0.58 | 0.09 | 0.11 | 0.10 | 0.18 | 0.03 |
| Weighted | 0.96 | 0.51 | 0.85 | 0.73 | 0.92 | 0.72 | 0.82 | 0.95 | 0.82 |
| IDWP | 0.50 | 0.41 | 0.04 | 0.38 | 0.45 | 0.44 | 0.29 | 0.26 | 0.75 |
| Krige | 0.70 | 0.34 | 0.05 | 0.37 | 0.96 | 0.19 | 0.32 | 0.52 | 0.92 |
| AvgDiff | 0.19 | 0.22 | 0.26 | 0.42 | 0.75 | 0.75 | 0.57 | 0.40 | 0.16 |
| SOTest | 0.40 | 0.40 | 0.53 | 0.27 | 0.24 | 0.34 | 0.34 | 0.27 | 0.62 |

**Table 7.3: p-values from PAUC FPR Shapiro-Wilk test**

| | Groups: Number of Nearest Neighbours | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Spatial | 0.65 | 0.71 | 0.89 | 0.85 | 0.77 | 0.47 | 0.27 | 0.30 | 0.66 |
| Median | 0.91 | 0.53 | 0.96 | 0.78 | 0.75 | 0.39 | 0.33 | 0.28 | 0.84 |
| Local | 0.95 | 0.56 | 0.37 | 0.26 | 0.12 | 0.04 | 0.04 | 0.55 | 0.91 |
| Scatter | 0.66 | 0.27 | 0.50 | 0.41 | 0.60 | 0.63 | 0.33 | 0.61 | 0.87 |
| SLOM | **0.00** | 0.19 | 0.11 | 0.23 | 0.83 | 0.46 | 0.66 | 0.64 | 0.03 |
| Weighted | 0.35 | 0.85 | 0.63 | 0.81 | 0.78 | 0.85 | 0.89 | 0.91 | 0.82 |
| IDWP | 0.04 | 0.07 | 0.08 | 0.09 | 0.11 | 0.11 | 0.14 | 0.17 | 0.75 |
| Krige | 0.55 | 0.60 | 0.56 | 0.69 | 0.76 | 0.83 | 0.88 | 0.86 | 0.92 |
| AvgDiff | 0.42 | 0.84 | 0.81 | 0.73 | 0.80 | 0.82 | 0.80 | 0.86 | 0.16 |
| SOTest | 0.93 | 0.52 | 0.54 | 0.58 | 0.52 | 0.43 | 0.47 | 0.39 | 0.62 |

Bold denotes significant at a 1% critical level

# APPENDIX B:

# BROWN-FORSYTHE TEST

Analysis of variance requires the group variances are statistically equal. If this assumption is violated, then ANOVA's F-statistic is invalid. The Brown-Forsythe test of homogeneity tests for the equality of group variances by performing an ANOVA test on a transformation of the response variable (Brown et al., 1974). The test is as follows (Brown et al., 1974):

$$F = \frac{(N-p)\sum_{j=1}^{p} n_j\left(z_j - z_i\right)^2}{(p-1)\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(z_{ij} - z_{.j}\right)^2}$$

where $z_{ij}$ is the transformed ROC value, $z_{ij} = \left|y_{ij} - \tilde{y}_j\right|$, where $\tilde{y}_j$ is the median of group $j$. $p$ is the number of groups, $n_j$ is the number of observations in group *j*, and *N* is the number of total observations. If the test statistic is small enough, then the null hypothesis that the group exhibit equal variance is rejected. Table 8.1 provides the Brown-Forsythe p-values.

**Table 8.1: p-values from Brown-Forsythe test**

| Algorithm | AUC | PAUC 80% TPR | PAUC 5% FPR |
|---|---|---|---|
| Spatial | 0.997 | 0.991 | 0.992 |
| Median | 0.994 | 0.954 | 0.995 |
| Local | 0.987 | 0.850 | 0.793 |
| Scatter | 1.000 | 0.938 | 0.996 |
| SLOM | 0.977 | 1.000 | 1.000 |
| Weighted | 0.999 | 0.963 | 1.000 |
| IDWP | 1.000 | 0.998 | 1.000 |
| Krige | 1.000 | 0.340 | 1.000 |
| AvgDiff | 1.000 | 0.999 | 1.000 |
| SOTest | 1.000 | 1.000 | 1.000 |

# APPENDIX C:

# LIST OF ACRONYMS

AUC – area under ROC curve

DEM – digital elevation model

DGPS – differential global positioning system

DSM – digital soil mapping

EMI – electromagnetic induction

$EC_a$ – apparent soil electrical conductivity

ER – electrical resistivity

FPI – fuzzy performance index

FPR – false positive rate

GIS – geographic information systems

GLS – generalized least squares

GPS – global positioning system

MZ – management zones

NCE – normalized classification entropy

NN – nearest neighbour

NFSP – National Farm Stewardship Program

PA – precision agriculture

PCM – precision crop management

PF – precision farming

PAUC – partial area under ROC curve

ROC – receiver operating characteristic

RTK GPS – real-time kinematics global positioning system

SAR – spatial autoregressive model

SSCM – site-specific crop management

SSM – site-specific management

TDR – time domain reflectometry

TPR – true positive rate

VRT – variable rate technology

# APPENDIX D:

# LIST OF SPATIAL OUTLIER ALGORITHMS

*Averaged Difference (AvgDiff)*

*Graph-based*

*Inverse Distance Weighted to a Power (Inverse Distance Weighting, IDWP)*

*Iterative R (Iterative Ratio)*

*Iterative Z (Iterative Spatial Statistic Z)*

*Kriging Interpolation (Kriging, Krige)*

*Local Area Mean (Local)*

*Median Statistic Z (Median)*

*Moran Scatter Plot (Moran)*

*Scatter Plot (Linear Regression, Scatter)*

*Spatial Local Outlier Measure (SLOM)*

*Spatial Outlier Test (SOTest)*

*Spatial Statistic Z (Spatial, Spatial Z, Z algorithm)*

*Variogram Cloud*

*Weighted Z (Weighted, IDW)*