2011

Statistical Geocomputing: Spatial Outlier Detection in Precision Agriculture

Chu Su, Peter

Chu Su, P. "Statistical Geocomputing: Spatial Outlier Detection in Precision Agriculture". Canadian Association of Geographers Annual Meeting and Conference, University of Calgary, Calgary, Alberta, May 31 - June 4, 2011.
http://hdl.handle.net/1880/48670
conference proceedings

STATISTICAL GEOCOMPUTING: SPATIAL OUTLIER DETECTION IN PRECISION AGRICULTURE

PETER CHU SU

Department of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

The collection of crop 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). This explosive growth and widespread use of spatial data has challenged the ability to derive useful spatial knowledge, emphasizing the need for better data pre-processing. In addition, outlier detection remains a challenge because the technique and the definition of spatial neighbourhood remain non-trivial; false positives, false negatives, and the concept of region outlier remain unexplored in spatial data via quantitative assessments. The overall aim of this study is to evaluate different spatial outlier detection techniques in terms of correctness and computational efficiency.

With simulated crop yield 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 technique as a classifier. Performance assessment is evaluated with the area and partial area under the ROC curve at different detection rates. Results indicate that for point outlier scenario, spatial autocorrelation techniques are more superior to standard 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 outlier situations, standard techniques are superior in all performance aspects because they are less affected by masking and swamping effects.

KEYWORDS

Spatial outlier, geostatistical simulation, spatial neighbourhood, area under ROC curve, sensitivity-specificity, precision agriculture

1. INTRODUCTION

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. Particularly, spatial yield datasets obtained by combine harvester mounted with ground-based yield 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.

Yield errors are commonly dealt with expert filtering programs. Most common systematic error sources have been well defined and described in the literature. They are dealt objectively with the expert filters. On the other hand, stochastic errors from mostly unknown source, commonly referred as yield surges or spatial outliers, are dealt according to the discretion of the analyst. In most cases, these errors can be either completely ignored or incorrectly removed.

To deal with yield surges, the precision agriculture community utilizes local neighbourhood statistics, while the data mining community implement different 'S-Outlier' algorithms. While both research communities implement similar techniques, outlier detection in spatial data remains a challenge for various reasons. First, the choice of algorithm is non-trivial. Sudduth & Drummond (2007) state that no standard method for cleaning yield surges exists. Numerous spatial outlier techniques have been proposed to supersede previous techniques, but no knowledge exists whether new algorithms are better, and no quantitative evidence has been provided to support this claim. The current approach at assessing spatial outlier techniques is by ranking the top spatial outliers identified by each technique for a particular spatial dataset. 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, "+". Noack et al.

(2003) neighbourhood is similar to a letter "H", where the vertical lines correspond to the neighbouring harvest tramline. And neighbourhood from Bachmaier (2010) resembles a butterfly. In all cases, the number of neighbouring observations is left to the analyst's discretion.

Lastly, false positives and false negatives effects are not properly explored or treated. In other words, most of the study on spatial outliers has been focused on detecting single point outliers. However, region outliers, spatial outliers that are clustered together causing instances of false positive and false negatives, remain largely unexplored in spatial data.

2. METHODOLOGY

The proposed approach involves utilizing a spatial dataset with known characteristics and known errors in advance. Unlike previous studies, a real dataset should not be used to determine algorithm effectiveness because spatial outliers are really not known. In real datasets, spatial observations whose non-spatial attribute significantly deviate from their spatial neighbours can be either real spatial outliers, observations in a spatial framework that were indeed produced by a differing mechanism, or simply the inherent (natural) variability of the spatial data. Spatial outlier algorithms cannot distinguish such data properties.

2.1. Spatial Data Generator

An unconditional sequential Gaussian simulation is utilized to generate spatial point data, and is 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.

Yield point measurements for a hypothetical on-farm experiment are simulated with three treatments on a Gaussian random field $Z(x) = m(x) + \varepsilon(x)$ on a rectangular 40-hectare farm (400m by 1,000m). 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. The sampling interval is set to one sample every one metre, and the separation distance between strips is set to 20 metres.

The Gaussian random field consists of spatially correlated error term $\varepsilon(x)$ and a deterministic trend modelled as:

$$m(x) = a + \beta_1 f_1(x) + \beta_2 f_2(x) + \gamma_1 t_1(x) + \gamma_2 t_2(x)$$

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. 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. f_1, f_2, and ε 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 ε 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 the average crop yield of the farmer's standard treatment approach, which is set to 76 bushels per acre. γ_1 and γ_2 represent two innovative site-specific management practices. Environmental variable 1 is set to be slightly correlated with crop yield, while innovative practice 1 was is set to increase crop yield by 3 units. Environmental variable 2 and innovative practice 2 are set to have no effect on crop yield. Therefore, the yield model equates to:

$$Z(x) = 76 + 3f_1(x) + 0f_2(x) + 3t_1(x) + 0t_2(x) + \varepsilon(x)$$

The first outlier scenario is the addition of single spatial outliers, which are random points in the field that are contaminated. A percentage of the population is randomly selected, and these points are further randomly divided into two groups. One group adds an error term to the yield measurement while the other subtracts from it. If any of these resulting contaminated yield measurements are more than the maximum original yield value or less than the minimum original value, then they are labelled as global outliers, otherwise, they are spatial outliers. The error term comes 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.

The second scenario involves the addition of region outliers, which are a groups of outliers clustered together in random locations. For population N, given a set size *G* for the amount of spatial outliers in a region, random seeds are selected from the $N^{th} - G + 1$ observations. For each seed, the seeded observation and the subsequent *G* observations are set as outliers by adding or subtracting the error term to the entire region.

2.2. Spatial Methods

Nine of the most popular statistical spatial outlier algorithms are tested on the simulated spatial data with known spatial outliers in advance. These include five standard algorithms, and four algorithms that account for spatial autocorrelation. These nine algorithms include:

Standard	Spatial Autocorrelation
Spatial Z (Shekhar et al., 2003)	Inverse Distance Weighting
Median Z (Chen et al., 2003)	Kriging Interpolation
Scatter Plot (Shekhar et al., 2003)	Weighted Z (Kou et al., 2006)
Local Area Mean	Averaged Difference (Kou et al., 2006)
Spatial Local Outlier Measure (Chawla & Sun, 2006)	

All spatial outlier algorithms are based on similar principles. Spatial neighbourhood for observation x, or $N(x)$, is defined as k nearest neighbours (KNN). KNN is computed by finding the Euclidean distance between location x and all its remaining neighbours. After defining $N(x)$, the aggregate function, $f_{aggr}(x)$, is computed to summarize the non-spatial attribute values of $N(x)$. Such function can be classified as distributive, algebraic, or hollistic (Han and Kamber, 2001). 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. Finally, $f_{diff}(x)$ is normalized by finding the centre and spread of f_{diff}.

2.3. Quantitative Assessment

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. Performance measures available in classification problems can be utilized. Two very popular analytical tools that encompass such performance measures are a confusion matrix and the receiver operating characteristics (ROC) curve. The ROC curve is utilized to assess the algorithm performance because it is more suitable and far more superior to a single confusion matrix.

The best method to compare algorithms is to reduce the information contained in the ROC curve down to a single convenient scalar value that represents the classifier performance. The most popular value is the AUC, the area under the ROC curve. The AUC is a scalar that

summarizes across all thresholds, reflecting the overall quality of the classifier. However, the AUC may be a misleading because total area is not a perfect measure of 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 situation, researchers may only be interested in a few situations rather than all of them. 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. The novel approach would be to incorporate the partial area under the ROC curve, $PAUC(t_0, t_1) = \int_{t_0}^{t_1} ROC(t)dt$, at a fixed FPR and TPR.

The 5% false positive rate (FPR) and 80% true positive rate (TPR) are chosen as the performance thresholds. This is because spatial algorithms with high TPR and 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. R package *ROCR* provides the tools to construct ROC curves along with performance measures such as the AUC, and PAUC at a fixed FPR and TPR.

3. RESULTS

3.1. Point Outliers

Figure 3.1 shows the area under ROC curve for each algorithm under different number of nearest neighbour (NN) used to compute f_{aggr}. All standard algorithms, *Spatial, Median, Local, Scatter,* and *SLOM* falter against the sensitivity of NN. As the number of neighbours increases, the AUC decreases rapidly.

Algorithms that account for spatial autocorrelation are less influenced by the change of NN. This is primary because each spatial autocorrelation algorithm assign different weight to each neighbour during the computation of f_{aggr}. AUC for *Weighted, Kriging, AvgDiff,* and *SOTest* decrease slightly as NN increases. However, AUC increases for *Inverse Distance Weighting* as NN increases.

Figure 3.1: AUC Sensitivity analysis

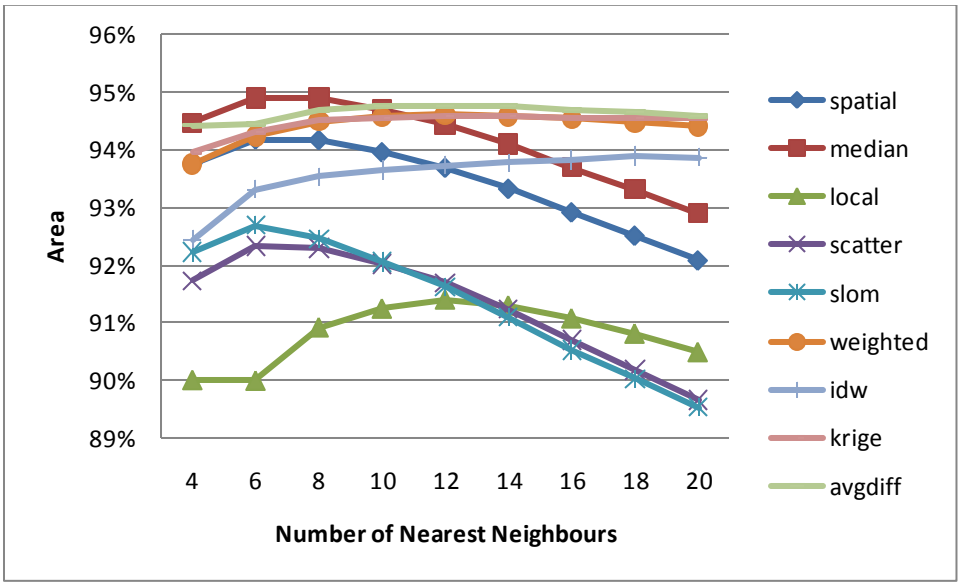


Figure 3.2 provides the partial area under ROC curve from 80% true positive rate for each algorithm against the sensitivity of NN. Given that TPR is restricted at 0.8, the maximum area that can be obtained is 0.2 or 20%. Spatial autocorrelation algorithms obtain higher PAUC than standard algorithms. 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% or more.

Figure 3.2: PAUC at 80% TPR sensitivity analysis

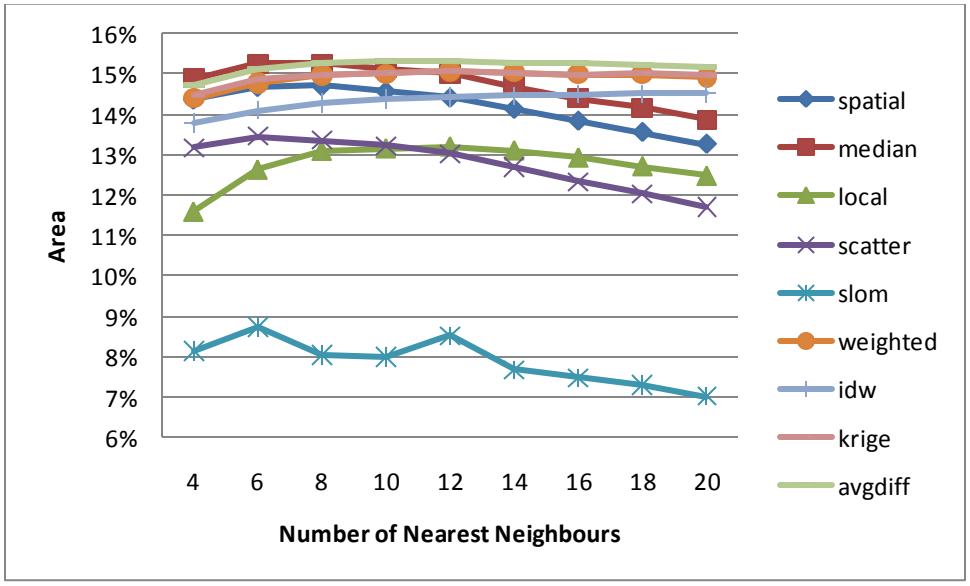


Figure 3.3: PAUC at 5% FPR sensitivity analysis

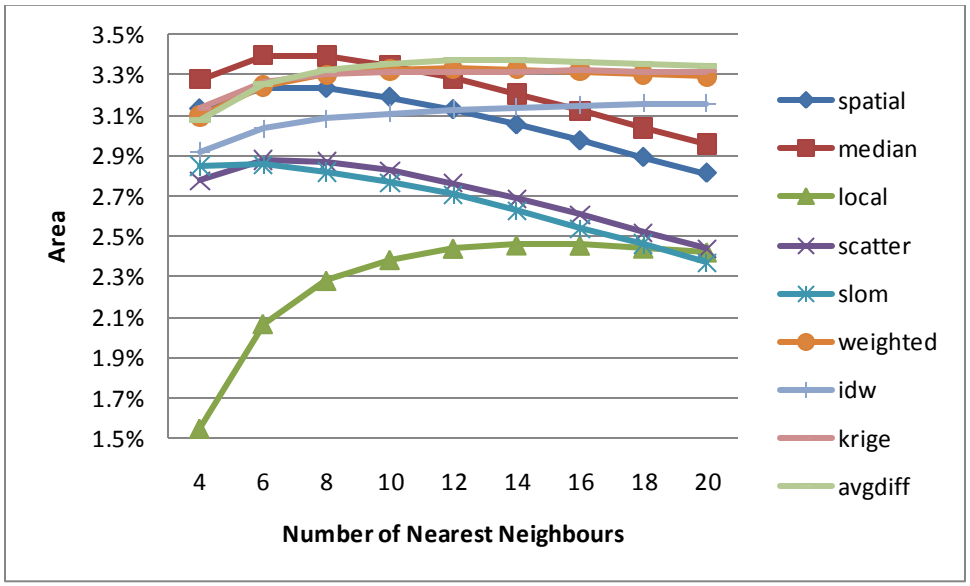


Figure 3.3 provides the PAUC at 5% false positive rate for each algorithm against the sensitivity of NN. The maximum area that can be obtained is 5% given that the FPR is restricted at 0.05. In this case, obtaining a large PAUC by implies that an algorithm obtains a relatively high true positive rate given a false positive of 5% or less. Figure 3.3 shows spatial autocorrelation algorithms obtain the highest PAUC, much similar to the AUC in Figure 3.1. *Local* is the algorithm with poorest performance, especially when NN is small, which means that *Local* obtains the smallest true positive rate with a false positive rate of 5% or less.

3.2. Region Outliers

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 (false positive), masking effects (false negative) affecting true spatial outliers is present in situations of region outliers.

Figure 3.4 presents the AUC performance given region outlier of size 2 (3.4a) and size 5 (3.4b). No much difference exists between Figure 3.4a and Figure 3.4b. This evidence suggests that the size of the region size has the same influence on the performance of all algorithms. However, a clear distinction is that that all spatial autocorrelation algorithms, except for *AvgDiff,* perform

worse than standard algorithm. In particular, 3.4a depicted *Kriging, SOTest,* and *Weighted* obtaining higher AUC performance than *Scatter*. At region outlier size of 5, *Scatter* outperforms *Kriging, SOTest,* and *Weighted.* As the region outlier size increases, standard algorithm outperform spatial autocorrelation ones.

Figure 3.4: AUC performance

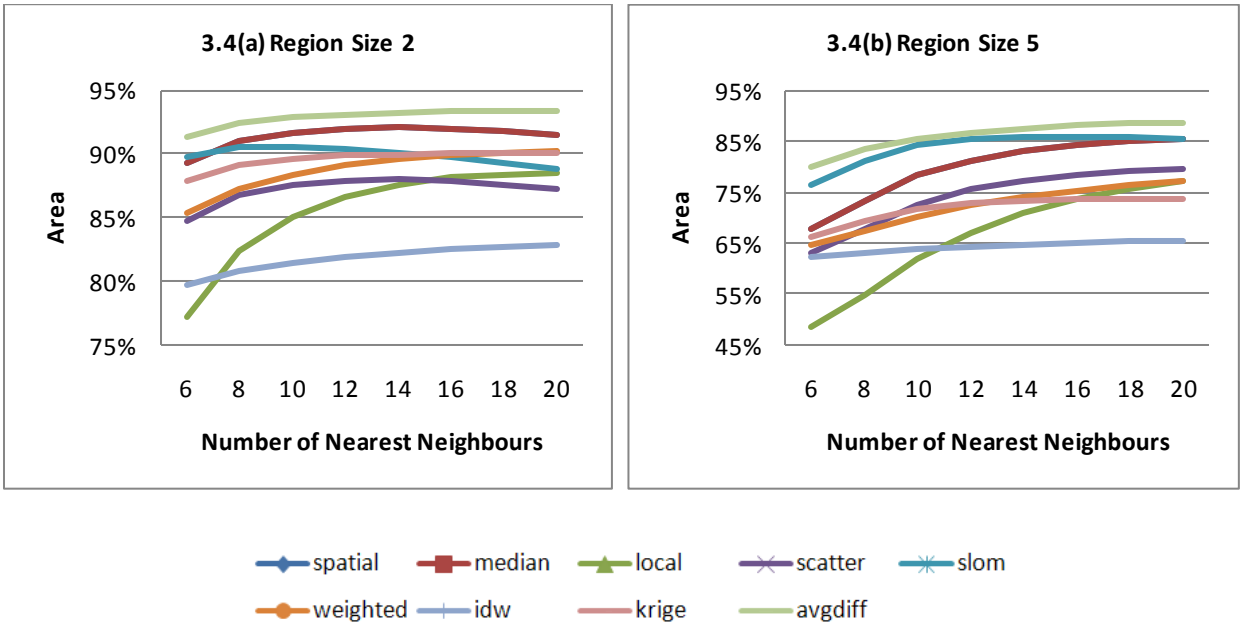


Figure 3.5a shows the PAUC 80% TPR for the detection of region outliers of size 2. This figure resembles the AUC performance of the same region outlier size. The only difference is the performance of *SLOM.* Unlike AUC performance whereby *SLOM* obtained the fourth highest performance, here *SLOM* obtains the worst performance among all algorithms, which is the same trend evidenced for the single outlier situation. This suggests *SLOM* performs relatively well on all decision thresholds with the exception of decision thresholds that achieve high sensitivity of 80% or more.

Figure 3.5a shows the PAUC 80% TPR for region outliers of size 5. Although all algorithms have decreased performance, it is depicted that *AvgDiff* is substantially superior to all other algorithms. For instance, *AvgDiff* performance is about 5% higher 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 such, *AvgDiff* is able to obtain the lowest FPR when obtaining 80% or higher TPR for all region outliers, as compared to other algorithms. This performance gap increases with increasing size of the region.

Figure 3.5: PAUC 80% TPR performance

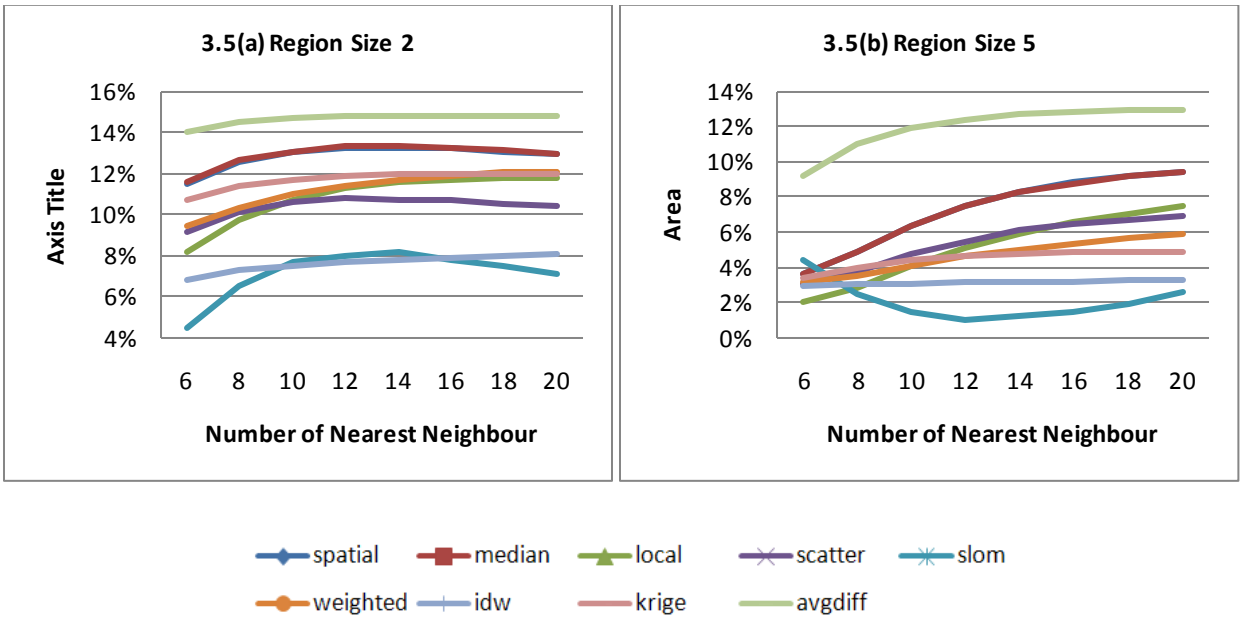


Figure 3.6: PAUC 5% FPR performance

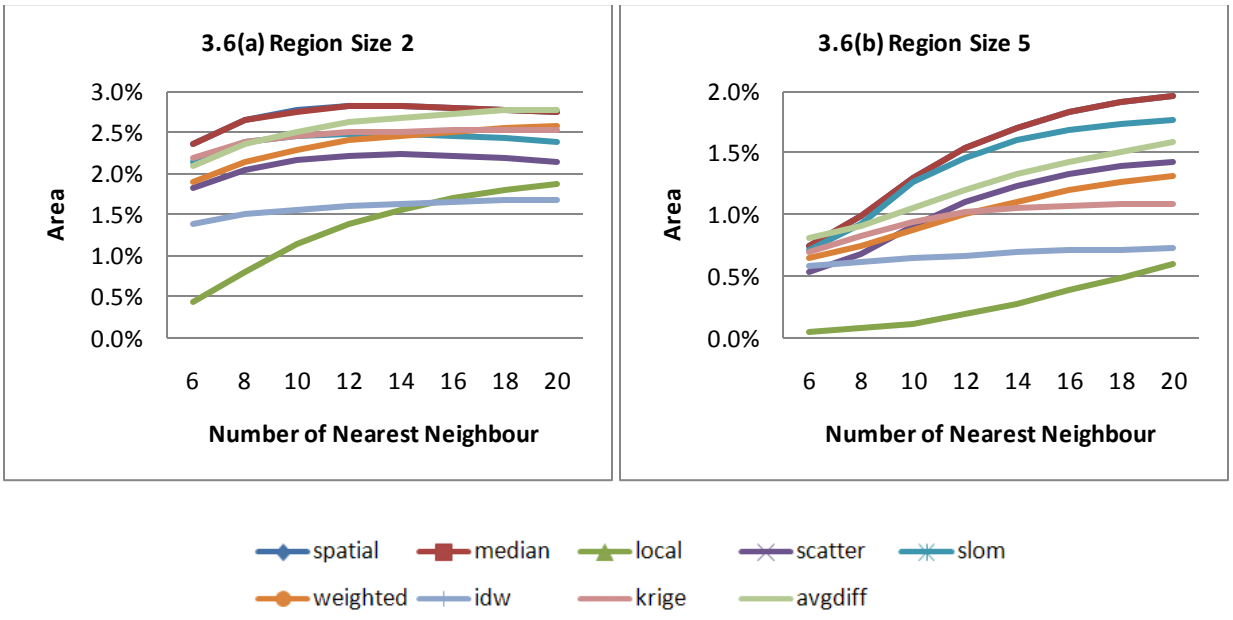


Figure 3.6a shows the PAUC at 5% FPR for the detection of region outliers of size 2. The results depict the same similarities as the AUC and PAUC TPR performance. The only difference is that *AvgDiff* obtains the third highest PAUC FPR behind *Spatial* and *Median, whereas AvgDiff* obtained the highest AUC and PAUC TPR performance overall. This suggests that *AvgDiff* performs best on all decision thresholds with the exception of decision thresholds that achieve

80% sensitivity or more. Notice that *Local* and *IDW* are significantly inferior to all other algorithms. Additionally, *SLOM* and *Kriging* are very similar.

Figure 5.18 provides the PAUC performance at 5% FPR for detecting region outliers of size 5. Not of much surprise, the performances of all algorithms 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.

4. DISCUSSION

Spatial and *Median* obtain similar ROC performance. However, *Median* is statistically superior to *Spatial* in all ROC aspects. *Median* is an upgrade of *Spatial* by computing f_{aggr} with the $median$ and standardizing f_{diff} with $median$ and the $Median\ Absolute\ Deviation\ (MAD)$. *Median* is more robust regarding variation in neighbourhood values. Thus, effects of masking and swamping are more properly suppressed when f_{aggr} is computed with $median$ than the $mean$.

In terms of region outlier detection, *Median* and *Spatial* obtain identical AUC, PAUC TPR, and PAUC FPR, suggesting that both algorithms obtain identical outlier scores. This is because of 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 have similar values. Therefore, f_{aggr} will be identical or very similar when computed either by $mean$ or $median$.

Scatter obtains poor ROC performance measures most because it requires the computation of the slope, m, between f and f_{aggr}, which requires calculating $mean$ of f and $mean$ of f_{aggr} both which are sensitive to outliers. *Scatter* is further exasperated by the fact that the intercept, b, requires the same computation of $mean$ of f and $mean$ of f_{aggr}. As such, estimates of slope, intercept, and of course the outlier scores, are not accurate, especially when masking and swamping is present.

Both *SLOM* and *Local* are the two algorithms with the lowest ROC performance. The former can only obtains a high TPR by obtaining a high FPR, while the latter cannot obtain a high TPR with a low FPR. This is implicative that *SLOM* over-classify observations by identifying more observations as outliers in hopes to obtain a high performance. However, by over-classifying, *SLOM* also obtains a high FPR. In contrast, *Local* under-classify by identifying less observation as outliers. This allows *Local* to obtain a low FPR but also a low TPR.

Unlike all other spatial outlier algorithms, *SLOM* and *Local Area Mean* require the calculation of two values that are influenced by local structure of the dataset. The issue is that there may be

instances where the one of those two values may not be able to distinguish between spatial outliers by adding confusion to the outlier scores. For example, *SLOM* is the product between a difference function and an oscillation parameter. The problem would be that the oscillation parameter can add additional error to the outlier score.

IDW has a quadratic distance decay function, distance weight calculated with the power of 2, which allocates more importance to closer neighbours for the computation of f_{aggr} than *Weighted's* linear distance decay function. Masking problems are exacerbated in *IDW* 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 a biased estimate of the neighbourhood average value. This is evidenced in the region outlier situations where *Weighted* substantially outperforms *IDW* in all ROC performance measures.

Similarly, a noisy local neighbourhood with substantial variability would introduce more confusion to outlier scores in *IDW* than in *Weighted*. This is particularly evidenced in region outlier scenarios where *IDW* obtains significantly inferior performance than Weighted, which is implicative that *IDW* is the algorithm most susceptible to masking effects. However, the disadvantage of both *IDW* and *Weighted* is the assumption that spatial autocorrelation changes uniformly in space.

In single outlier scenario, *Kriging* obtains better ROC performance than *Weighted* and *IDW* by being able to model the changes in spatial autocorrelation. *Kriging* is unique in the calculation of distance weights, as it depends on the autocorrelation structure set by the semivariogram. Neighbours which are far away and not autocorrelated obtain negative weights. In contrast, algorithms *Weighted, AvgDiff,* and *IDW* 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}.

However, unlike all other spatial algorithms, *Kriging* requires additional input parameters. *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. For instance, Table 4.1 provides the ROC performance measures for *Kriging* interpolation with 8 NN under different semivariogram models.

Table 4.1: 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

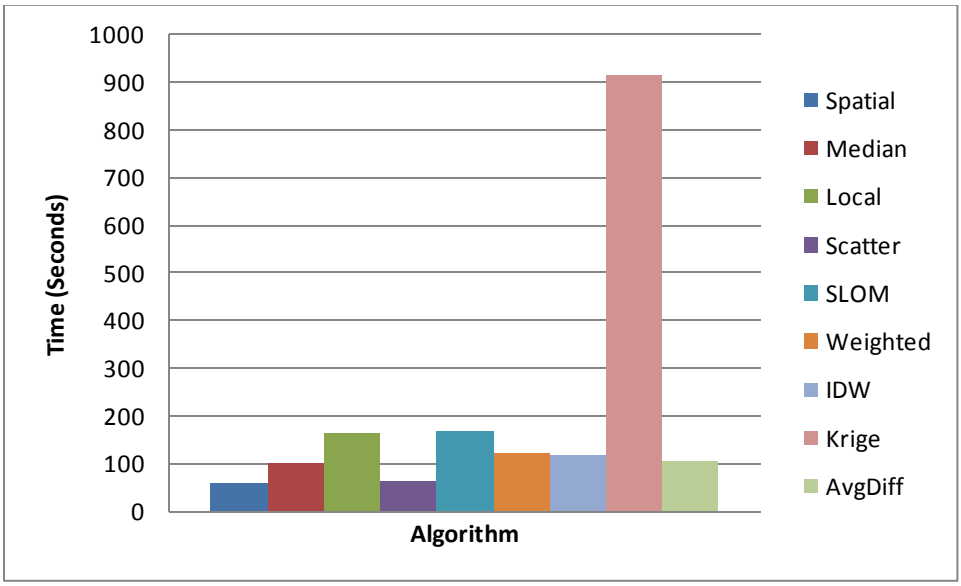
Another major drawback of utilizing *Kriging* is its computational complexity. Figure 4.1 provides the computation time for all ten algorithms. *Spatial* and *Scatter* obtain the fastest time, because they are mostly composed of basic operations. *Median* is more complex than *Spatial* as it takes about twice the computation time given that $median$ and MAD are more complex operations than $mean$ and $standard\ deviation$. Spatial autocorrelation algorithms are approximately twice the computation time of *Spatial* mainly because of the computation of distance weights 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.

The results suggest the best algorithm is *AvgDiff*. For single outlier scenario, *AvgDiff* obtains the highest AUC, PAUC at 80% TPR, and PAUC at 5% FPR at most NN settings. 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 best algorithm.

First, *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. 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, so $|0 - 50| = 50$ 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* $f_{diff}(x)$ value of 50. The first advantage of *AvgDiff* is its capability of properly adapting to the neighbourhood variance.

Figure 4.1: Computation time of spatial outlier algorithms



The second advantage of *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. Normalization adds 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.

5. CONCLUSION

Erroneous data and associated variability that results from inconsistent data collection practices can deviate the analysis of data 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 do the data mining algorithms are applicable for precision agriculture

applications, their algorithms far exceed the standard techniques used by the precision agriculture community. Three types of spatial algorithms have been utilized by the precision agriculture community: *Local, IDW*, and *Kriging.* The data mining community have developed the remaining algorithms.

Problem with both communities is that they 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 fail against determining true spatial outliers.

What has been determined here is the recommendation of using *AvgDiff* algorithm for cleaning yield surges and all other point datasets that exhibits spatial dependence. The nearest neighbour parameter for the neighbourhood aggregate function is still non-trivial. 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.

REFERENCES

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.

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.

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

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

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.

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

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

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

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

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

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.

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

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