The Effect of Diagnostic Misclassification on Spatial Statistics for Regional Data

by

Christopher Scott

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada

ABSTRACT

The Effect of Diagnostic Misclassification

on Spatial Statistics for Regional Data

Christopher Scott
University of Guelph, 2014

Advisors:
Dr. J. Horrocks
Dr. O. Berke

Spatial epidemiological studies which assume perfect health status information can be biased if imperfect diagnostic tests have been used to obtain the health status of individuals in a population. This study investigates the effect of diagnostic misclassification on the spatial statistical methods commonly used to analyze regional health status data in spatial epidemiology. The methods considered here are: Moran's I to assess clustering in the data, a Gaussian random field model to estimate prevalence and the range and sill parameters of the semivariogram, and Kulldorff's spatial scan test to identify clusters. Various scenarios of diagnostic misclassification were simulated from a West Nile virus dead-bird surveillance program, and the results were evaluated. It was found that non-differential misclassification added random noise to the spatial pattern in observed data which created bias in the statistical results. However, when regional sample sizes were doubled, the effect from misclassification bias on the spatial statistics decreased.

Acknowledgments

I would like to thank my advisors Dr. Julie Horrocks and Dr. Olaf Berke and my committee members Dr. Zvonimir Poljak and Dr. Paul McNicholas. I would also like to thank my family and friends.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

This work investigates the effect of diagnostic misclassification on statistics used in spatial epidemiology. For spatial epidemiological studies, interest lies in the relationship between geographical risk factors and the variation in health status of spatially distributed individuals. With the use of spatial statistics, this relationship may be analyzed by investigating the trend, clustering, and clusters of the health status of individuals in a geographical area. This analysis is important for identifying risk factors associated with the etiology of a disease or for finding regions in which a population is at high risk of contracting a disease. However, if the health information of a population is biased due to diagnostic misclassification, the results obtained from spatial statistical analysis may be misleading or biased as well. Hence, spatial epidemiological investigations which assume perfect health status information can be biased if imperfect diagnostic tests have been used to obtain the health status of individuals in the population under consideration.

The investigation presented here is a simulation study based on real-world data. The data come from a surveillance program for West Nile virus (WNv) in dead birds found in southern Ontario, Canada [6]. In this surveillance program, dead birds were collected in various public health regions in southern Ontario and tested for

WNv. The diagnostic test used to classify the dead birds as positive or negative for WNv in this program was not perfect; there are likely misclassified cases and controls in this data set. However, for the purpose of the simulation study presented in this paper, the number of dead birds that tested positive and negative for West Nile virus in the surveillance program are taken to be the true number of cases and controls of the disease. That is, the surveillance data is considered to have no misclassified cases and controls. This is done for two reasons. The first is simply that, even with the assumption of perfect classification, the data from the surveillance program represents a real-world distribution of a disease. Hence, the results from this work may be useful when considering diseases which are similar in nature. The second is that the statistical results obtained from the surveillance data set are compared to results from the simulated data sets with misclassified cases and controls. This determines how diagnostic misclassification affects the spatial statistical methods considered in this study.

Simulated data sets are produced by randomly misclassifying cases and controls in the WNv surveillance data based on varying degrees of diagnostic misclassification. Statistical methods used to estimate trend, clustering, and clusters of the disease are subsequently carried out on the simulated data sets in order to determine whether or not the results differ from those found from analysis on the original WNv surveillance data. If there are differences between the estimates from the simulated data and the original WNv data, they can be attributed to diagnostic misclassification. Typically, simulated data would be produced by generating new data sets from a Gaussian random field, and then misclassifying cases and controls in this new

data. For the simulations in this study, however, data is produced by only misclassifying cases and controls in a single data set. The simulations here were purposely designed this way in order to target the variability in the spatial statistical estimates due to misclassification only. Had data been simulated in the typical way, the effect of misclassification would have been obscured by additional variability due to sampling error.

In geostatistics, 'trend' is used to indicate the mean response over the spatial area of the study. In this simulation study, the trend is assumed constant and, hence, is taken to be the global prevalence of West Nile virus among dead birds in southern Ontario. 'Clustering' is the amount of spatial dependence among dead birds positive for West Nile virus. 'Clusters' are groups of cases within the population at risk which are more concentrated than what would be expected. The statistical methods used to detect these spatial patterns are described in section 2.4.

In Berke and Waller (2010) [5] it was found that, for large samples, diagnostic misclassification did not seriously affect the common statistics used to analyze the spatial patterns in geographically distributed health status data. However, it was suggested that for smaller, finite samples some effects could be present. The investigation presented in this work attempts to quantify the effects of diagnostic misclassification, if any, based on finite samples. Two scenarios are considered to investigate this issue. The first scenario is based on the WNv surveillance data where the number of dead birds that tested positive and negative in the public health regions in southern Ontario are assumed to be the true number of cases and controls. In the second scenario the numbers of regional cases and controls have been doubled.

This second scenario is considered in order to determine whether or not the effects of diagnostic misclassification on spatial statistics can be reduced if regional sample sizes are increased.

1.1 Spatial Epidemiology and Health Status Data

For spatial statistical methods to be applied, the health status data must be georeferenced. With respect to binary case-control data, this can be done by assigning a precise location to each of the cases and controls (point data), or alternatively, by aggregating case-control counts to specific regions of the study area (regional data). For regional data, it is important that the regional boundaries in the study area are clearly defined. This is necessary since a change in regional boundaries may require a re-aggregation of the cases and controls, which could ultimately lead to a detection of different underlying spatial patterns in the data [20]. This study is concerned only with aggregated regional data, where the regions are defined by 30 public health units (PHUs) in southern Ontario.

1.2 Diagnostic Tests

In order to obtain the health status of individuals in a population, a diagnostic test must be performed to indicate whether an individual is positive or negative for a disease. A diagnostic test that classifies cases and controls without error is called a perfect diagnostic test. In practise, however, these tests do not exist and imperfect diagnostic tests are used [11]. This is especially true when dealing with

emerging diseases, where diagnostic tests have yet to be developed, or in situations where a cheaper or faster diagnostic test is preferred to an alternative test which is more expensive or takes longer to apply [5].

In light of this, it is generally assumed that a diagnostic test carries with it a degree of misclassification. For binary health status data, the amount of misclassification inherent in a diagnostic test is defined by its sensitivity (SE) and specificity (SP). The probability that a diagnostic test will produce a positive result when it is administered to individuals that truly have the disease is defined as the sensitivity. The probability that it will produce a negative result when it is administered to individuals that are truly disease-free is defined as the specificity [14]. Sensitivity and specificity can thus be written as:

$$SE = P(T^+|D^+); \quad SP = P(T^-|D^-)$$

where T^+ denotes that an individual tests positive; D^+ denotes that an individual is positive for the disease; T^- denotes that an individual tests negative; and D^- denotes that an individual is negative for the disease [14].

To determine the sensitivity, the diagnostic test is administered to individuals who truly have the disease, and the ratio of the number of individuals who test positive to the number of individuals who actually have the disease is calculated. Similarly, specificity is determined by administering the test to individuals which are truly disease-free and calculating the ratio of the number of individuals who test negative to the number of individuals who do not have the disease:

$$SE = \frac{TP}{TP + FN}; \quad SP = \frac{TN}{TN + FP} \tag{1.1}$$

where TP is the number of individuals that test positive and have the disease; TN is the number of individuals that test negative and do not have the disease; FP is the number of individuals that do not have the disease but test negative (false positives); and FN is the number of individuals that have the disease but test negative (false negatives). Table 1.1 illustrates how sensitivity and specificity relate to diagnostic testing and disease status.

As mentioned, data from the WNv surveillance program are used as a reference by which the number of dead birds that tested positive and negative for WNv are taken as the true number of diseased and non-diseased birds, respectively. Various imperfect diagnostic tests are applied to these data in order to create new data sets which contain fixed numbers of false positives and false negatives. The methodology used to determine and distribute the false positives and negatives for the simulated data sets is described in section 2.2. Spatial statistics for trend (prevalence estimation), clustering (Moran's I and semivariogram estimation), and clusters (Kulldorff's spatial scan test) are then applied to these data sets for the purpose of investigating the effect of diagnostic misclassification on the results obtained from these statistics.

1.3 Raw and Smoothed Data

Before applying statistical methods to regional data, it is important to keep in mind the possibility of heterogeneity in the standard errors of regional prevalence

Table 1.1: Sensitivity (SE) and specificity (SP) as they relate to disease status (D^+/D^-) and diagnostic testing (T^+/T^-)

	$\mathbf{D^+}$	$\mathbf{D^-}$	**Total**
$\mathbf{T^+}$	TP	FP	TP + FP
$\mathbf{T^-}$	FN	TN	FN + TN
	TP + FN	TN + FP	N

$$\mathbf{SE = TP/(TP+FN)} \quad \mathbf{SP = TN/(TN+FP)}$$

estimates. This is particularly important to note when certain regions have low sample size, and adding or subtracting cases to the regions may change the prevalence estimates considerably. A common way to deal with this volatility in spatial variance is to smooth the data [5, p. 118]. With smoothing, it is possible to stabilize the standard errors of regional prevalence estimates and thereby reduce the impact of local extreme values on statistical tests for clustering [5]. Furthermore, by reducing volatility, it increases the chance of identifying any underlying trend in risk [26, p. 112].

Smoothing is also important with respect to choropleth mapping [3]. Choropleth maps are commonly used as a way to visualize regional prevalence or other measures such as cluster locations. By smoothing the data, it may be possible to reduce misleading visual cues brought on by unstable regional counts. For example, in a scenario where regional prevalence is high due to a single case among a low population count, a choropleth map would show a density equal to that of a region where a similar prevalence is the result of many cases among a high population count. With smoothing, the high regional prevalence associated with the low regional population

would be reduced, and hence could provide a better representation of risk distribution. Smoothing is applied to the various data sets in this study before estimating trend and clustering, since interest lies in comparing the spatial patterns which arise from misclassification in regions with both low and high sample sizes.

1.4 Spatial Dependence and Cluster Detection

The spatial patterns of concern in spatial epidemiology are the trend, clustering, and cluster locations in the geographic distribution of disease occurrence. Clustering is defined as the amount of spatial dependence (or spatial autocorrelation) in the data. Two notable statistics have been developed in order to quantify the strength of clustering in spatial data: Moran's I and Geary's C. For Moran's I [19], clustering is measured by the strength of correlation between defined neighbourhoods. Geary's C [13], also an indicator of clustering, measures the dissimilarity between neighbourhoods and is used to determine whether the prevalence estimates of neighbouring regions are random or spatially dependent [13, p. 115]. Before applying either of these statistics, however, it is important to consider any underlying spatial trends that may be present in the data. That is, these measures of spatial dependence should only be used in cases where trend is not considered to be present, or where it has been corrected for [13, p. 116]. For simplicity, only Moran's I is used to quantify clustering (i.e. spatial dependence) in the various data sets analyzed in this study.

To investigate trend effects, generalized linear models (GLMs) such as Pois-

son and logistic regression models may be applied. However, if there is spatial autocorrelation (i.e. clustering) in the data, there may be extra-variation in the spatial risk beyond what would be expected. This extra-variability is known as overdispersion and may cause GLMs to produce biased standard error estimates for the coefficients, since these models assume independence between observations [2, p. 1106]. In order to account for overdispersion, a generalized linear mixed model (GLMM) may be fit to the data. This model is an extension of a GLM, where the linear predictor contains both random effects and the usual fixed effects [2]. In this study a Gaussian random field model (GRFM) is fit to the smoothed data using maximum likelihood estimation. The trend component of this model is used to estimate prevalence and the sill and range parameters of the stationary Gaussian process are used to estimate the semivariogram [16]. The semivariogram models the dependence structure of the data, where the sill represents the overall variance in the data and the range describes the overall distance in the study region up to which regional observations (such as prevalence) are dependent on one another.

Clusters are defined as groups of cases which are too concentrated within the population at risk to have occurred by chance. That is, they are collections of cases which exceed the amount that would be expected within regions of the study area. Cluster identification is important in situations which necessitate disease containment, whereby high risk zones may be identified in order to take preventive action against the disease. Bayesian cluster detection [25], Besag-Newell [7], and the Kulldorff [17] spatial scan test have all been developed in recent years to identify clusters and are available for use in R [23]. In order to reduce computational time, only the Kulldorff

spatial scan test is used for cluster identification in this study.

1.5 Objectives

This investigation examines the effect of misclassification due to imperfect diagnostic tests (sensitivity and / or specificity of less than 100%) on the performance of statistical methods which quantify these spatial patterns of trend, clustering, and clusters. In particular, Monte Carlo simulation experiments are designed and constructed by distributing false positives and false negatives across 30 public health units in southern Ontario in accordance with varying (reduced) measures of sensitivity and specificity. The effect of diagnostic misclassification on prevalence estimation, the semivariogram, Moran's I, and Kulldorff's spatial cluster detection method is then investigated by examining their respective results from data with varying degrees of misclassification bias and with varying regional sample sizes. The estimation bias in this study is considered to be bias due to misclassification only (i.e. without variability due to sampling error). In other words, the bias here is the difference in the estimates of prevalence, semivariogram, Moran's I and cluster locations from the simulated data and the 'true' values corresponding to the original data, which can be attributed to diagnostic misclassification alone.

Chapter 2

Materials and Methods

2.1 Study Area and WNv Data

In 2000, the Public Health Agency of Canada implemented a surveillance program to monitor the prevalence of West Nile virus (WNv) in humans, birds, mosquitoes, and horses in 30 PHUs in southern Ontario, Canada [6]. The data for WNv infections among dead birds for the year 2005 is used as the foundation for this simulation study. The number of birds that tested positive for WNv in this data set is 272 and the number that tested negative is 745. With respect to the sample size, these counts are taken to be the true number of cases and controls in the study area, and correspond to a diagnostic test with 100% sensitivity and 100% specificity. That is, these counts, as well as the regional counts, are taken to represent the true distribution of the disease. Thus, the estimates for prevalence, Moran's I, sill, range, and cluster locations from this data set are considered to represent the true spatial pattern of WNv among the 30 PHUs in southern Ontario.

The data includes coordinates of the centroids for the 30 PHUs, in addition to the aggregated counts of dead birds and infected dead birds found in each PHU in southern Ontario. The dead birds found in each PHU were submitted for diag-

Table 2.1: Original 2005 WNv data. Sample size corresponds to the number of dead birds found in each public health unit (PHU ID); the number of dead birds which tested positive for WNv is denoted by cases.

PHU ID (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
sample size (n_i)	35	69	8	35	9	44	41	50	47	10	12	32	16	18	19
cases (m_i)	11	8	2	2	1	3	20	15	3	4	9	4	8	1	6
PHU ID (i)	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
sample size (n_i)	21	23	51	30	75	12	63	29	16	45	35	43	15	54	60
cases (m_i)	3	5	6	8	33	5	22	1	2	4	11	4	12	22	37

nostic testing to the Canadian Cooperative Wildlife Health Centre (CCWHC) by the general public and by public health unit personnel [6]. The dead birds submitted in 2005 were tested for WNv using the VecTest, which had a sensitivity of 83.3% and 95.8%. A second test, reverse transcriptase polymerase chain reaction (RT-PCR), was then applied to the positive dead birds. The RT-PCR was used to test for WNv ribonucleic acids in order to reduce the number of false positives [6]. It should be reiterated, however, for the purpose of the simulation study, the reduced sensitivity and specificity of the VecTest is ignored, and the number of dead birds which tested positive and negative for WNv are taken to be the true number of cases and controls of the disease.

The region of southern Ontario spans a range of approximately 600 kilometres (km) from the southwest to the northeast. Each of the 30 PHUs in this region differs in size as well as the number of dead birds and the number of cases found in each unit. Table 2.1 shows the total number of birds sampled and the number of WNv positive birds found per PHU.

2.2 Data Simulation

In order to investigate spatial misclassification bias due to imperfect diagnostic tests, scenarios in which the 2005 surveillance data contain false positives and false negatives are produced. These scenarios are simulated by randomly assigning a false positive status to a certain number of controls and a false negative status to a certain number of cases in the original data set. Each of the controls designated as a false positive and each of the cases designated as a false negative is then identified by PHU, and relabelled as positives and negatives, respectively. The number of false negatives and false positives are determined by the sensitivity and specificity of the imperfect diagnostic test under consideration. The total number of dead birds positive for WNv (D^+) and negative for WNv (D^-) in the simulated data sets are taken to be the number of cases and controls as identified by the (assumed) perfect diagnostic test used for the original data set.

A total of 24 different scenarios are investigated in this study. These scenarios are based on the application of 24 imperfect diagnostic tests which correspond to different combinations of sensitivity and specificity. The values of sensitivity and specificity considered here are SE = (100%, 95%, 90%, 85%, 80%) and SP = (100%, 95%, 90%, 85%, 80%), where (SE, SP) = (100%, 100%) corresponds to the perfect diagnostic test used to obtain the reference data. Using Table 1.1 and Equation 1.1, these different combinations of sensitivity and specificity may be used to determine the number of dead birds which are WNv positive and test positive (TP), WNv positive but test negative (FN), WNv negative and test negative (TN), and WNv negative

but test positive (FP). Specifically, since D^+ and D^- counts are taken to be the 272 cases and 745 controls of the WNv surveillance data, it is known (from Table 1.1) that $TP + FN = 272$ and $TN + FP = 745$. By rearranging Equation 1.1, this means that $TP = SE(TP+FN) = SE(272)$ and $TN = SP(TN+FP) = SP(745)$. The number of FNs and FPs can then be determined by $FN = 272 - TP$ and $FP = 745 - TN$.

Once the FN and FP counts have been determined, simple random sampling is used to determine which true cases and controls from the original data set are to be designated as false negatives and false positives. These are then identified by PHU and relabelled accordingly. This, in turn, establishes the number of T^+ and T^- within each PHU and consequently the simulated data set. For a given PHU, the number of T^+ and T^- counts would be:

$$T_i^+ = m_i + FP_i - FN_i$$

$$T_i^- = l_i + FN_i - FP_i$$

where m_i is the number of cases in each PHU and $l_i = n_i - m_i$ is the number of controls. For the entire study region ($N = 30$ PHUs) this entails:

$$T^+ = \sum_i^N (m_i + FP_i - FN_i) = \sum_1^{30} m_i + \sum_1^{30} FP_i + \sum_1^{30} FN_i = 272 + FP - FN$$

$$T^- = \sum_i^N (l_i + FN_i - FP_i) = \sum_1^{30} k_i + \sum_1^{30} FN_i + \sum_1^{30} FP_i = 745 + FN - FP$$

It should be noted that, for a given combination of SE and SP, the number of T^+ and T^- in the PHUs are the counts used for statistical analysis. This is done because, in actuality, an investigator only observes or has knowledge of the T^+ and T^- counts produced by the diagnostic test used in the investigation. Table 2.2 illustrates how the TPs, FNs, TNs, and FPs are determined for the entire study area.

Table 2.2: Illustration of how sensitivity (SE) and specificity (SP) are used to determine the number of false negatives (FN) and false positives (FP) that are distributed among the WNv data

	$\mathbf{D}^+$	$\mathbf{D}^-$	
$\mathbf{T}^+$	TP = SE(272)	FP = 745 - SP(745)	272+FP-FN
$\mathbf{T}^-$	FN = 272 - SE(272)	TN = SP(745)	745+FN-FP
Total	272	745	1017

Since the sensitivity and specificity determine the number of false positives and false negatives that are randomly distributed among the original data set, each scenario contains a different number of misclassified cases and controls. As an example, consider a diagnostic test with 90% sensitivity and 95% specificity. From the 745 controls and 272 cases identified in the original data set, this would correspond to 37 dead birds falsely identified as positive for WNv and 27 dead birds falsely identified as negative for WNv. This example is shown in Table 2.3.

Further exploration into the effects of diagnostic misclassification is carried out by considering the situation in which the number of cases and controls in each PHU from the original data set is doubled. This is done in order to investigate whether or not increasing regional sample size could mitigate any impact that diagnostic misclassification might have on the results of the spatial statistical methods considered in this study. Data sets are simulated using the same method described above; however, the doubled case and control counts of the WNv surveillance data are used as the true number of cases and controls (i.e. are considered to be obtained by a perfect diagnostic test). Data sets are then produced by using an imperfect diagnostic test to determine a number of false positives and false negatives, and randomly distributing

Table 2.3: Example of how FP, FN, TP, and TN are determined using a diagnostic test with 90% sensitivity and 95% specificity

	$\mathbf{D^+}$	$\mathbf{D^-}$	
$\mathbf{T^+}$	TP = .90(272) = 245	FP = 745 - .95(745) = 37	282
$\mathbf{D^-}$	FN = 272 - .90(272) = 27	TN = .95(745) = 708	735
Total	272	745	1017

these false positives and false negatives among the data. Here, a diagnostic test with 90% sensitivity and 95% specificity would correspond to 75 dead birds being falsely identified as positive for WNv and 54 dead birds being falsely identified as negative for WNv, based on the 1490 controls and 544 cases in this data set.

A few comments should be made with regard to these simulations. First, simple random sampling is used to misclassify dead birds in the original data set(s). That is, each dead bird negative for WNv has an equal chance of being assigned a false positive status. Likewise, each dead bird with a positive WNv status has an equal chance of being classified as a false negative. In terms of a diagnostic test with 90% sensitivity and 95% specificity applied to the original data set, this would mean that 37 birds from the 745 controls would be sampled, without replacement, and assigned a false positive status, and 27 birds from the 272 cases would be sampled, also without replacement, and be assigned a false negative status. In addition, the number of false negatives and false positives identified in each simulation cannot exceed the number of cases and controls, respectively. Also, by relabelling the false negatives and false positives within the PHUs, it is ensured that the original sample size of each PHU is conserved. That is, for a given diagnostic test, the number of T^+ and T^- within

each PHU will sum to the original sample size. This is important with respect to the application of Kulldorff's spatial scan test, which depends on a population size parameter to detect cluster locations. Finally, though two diagnostic tests with the same sensitivity and specificity will produce the same number of false positives and false negatives, the distribution of these false positives and false negatives will likely be different for each simulated data set. This is due to the fact that for each simulation the FPs and FNs are randomly distributed among the PHUs.

2.3 Smoothing

Empirical Bayes Smoothing [18][26] is applied to the data because of spatially varying sample sizes in the WNv surveillance data. This method can be viewed as a technique that internally standardizes local disease rates and stabilizes local variances [4, p. 8-9]. It takes into consideration local and global observed disease rates in order to estimate regional rates with less variability than the observed. As a result, these estimates can be viewed as a better representation of the underlying spatial variability in disease rates [12, p. 623]. Stabilizing regional variability via smoothing is useful when estimating trend and the semivariogram since the impact of local extreme values on parameter estimation can be mitigated [5, p. 118]. However, it should be acknowledged that, though smoothing reduces the impact of local extreme values on parameter estimation, there may be a reduced ability to detect high regional rates [5, p. 118].

The technical details of Empirical Bayes Smoothing are explained in Mar-

tuzzi and Elliott (1996) [18]. For the 2005 WNv surveillance data, the details are as follows: first, since the crude overall prevalence of WNv in dead birds is $\frac{\sum m_i}{\sum n_i} \approx 26.7\%$ it is considered a non-rare disease and, hence, is assumed to follow a binomial distribution. That is, for the number of cases, m_i, and regional sample size n_i, of the i^{th} of N regions, $m_i \sim B(\pi_i, n_i)$. Here the goal is to estimate the parameter π_i, the prevalence of the i^{th} region in southern Ontario, where $p_i = m_i/n_i$ is the crude prevalence (observed proportion) of the i^{th} region and the maximum likelihood estimate of π_i.

It is assumed that π_i follows a unimodal and regular distribution with $\mu = E(\pi_i)$ and $V = var(\pi_i)$. Since m_i follows a binomial distribution, this gives:

$$E(p_i|\pi_i) = \pi_i; \qquad V(p_i|\pi_i) = \frac{1}{n_i}\pi_i(1-\pi_i)$$

and

$$E(p_i) = \mu; \qquad V(p_i) = E(var(p_i|\pi_i)) + var(E(p_i|\pi_i)) = V + \frac{1}{n_i}[\mu(1-\mu) - V]$$

An unbiased estimate of μ can be defined by $k = \frac{\sum m_i}{\sum n_i}$. If k is considered an error-free estimate of μ, for the weighted sample variance $s^2 = \sum n_i(p_i - k)^2 / \sum n_i$ this yields:

$$E(s^2) = \left(1 - \frac{1}{\bar{n}}\right) V + \frac{1}{\bar{n}}(k(1-k))$$

and V can be estimated by:

$$\frac{s^2 - \frac{1}{\bar{n}}(k(1-k))}{1 - \frac{1}{\bar{n}}}$$

With these approximations, the posterior estimates of p_i which minimize the total squared error loss can be calculated via:

$$\hat{\pi}_i = \rho_i p_i + (1 - \rho_i)k$$

where:

$$\rho_i = var(\pi_i)/var(p_i) = \frac{n_i s^2 - \frac{n_i}{\bar{n}}(k(1-k))}{(n_i - 1)s^2 + \frac{\bar{n} - n_i}{\bar{n}}k(1-k)}$$

The posterior estimates of p_i, $\hat{\pi}_i$, are used as the smoothed regional prevalences. ρ_i is between 0 and 1 and the smoothed estimator $\hat{\pi}_i$ adjusts the crude regional prevalence (p_i) toward the global mean k. It can be seen that, when heterogeneity is low, ρ_i approaches zero and the estimate for π_i is closer to the global mean k. Conversely, if there is a large amount of heterogeneity, s^2 is large and, hence, ρ_i approaches 1, which leads to estimates for π_i's which are closer to the regional prevalences.

Smoothing is applied to the surveillance data and each of the simulated data sets before estimating the parameters of the Gaussian random field model (section 2.4.1). In addition, an empirical Bayes index modification of Moran's I (EBI) [1] is used to assess clustering in the data, which takes into account the smoothed data when estimating Moran's I, and is described in section 2.4.2.

2.4 Statistical Methods

This section gives a brief overview of the statistical methods used to detect spatial patterns for trend, clustering and clusters. The methods used here are: a Gaussian random field model to estimate trend and the semivariogram, Moran's I to quantify the amount of clustering in the data, and Kulldorff's spatial scan test to identify clusters. In particular, multiple simulations are run for several different values

of sensitivity and specificity of an imperfect diagnostic test and, for each simulation, estimates for the prevalence, sill and range of the semivariogram, Moran's I, and cluster locations are obtained.

2.4.1 Trend and Semivariogram Estimation

For prevalence and semivariogram estimation, a Gaussian random field model without a nugget effect is fit to the smoothed data. This model gives an estimate for prevalence via the intercept for the regression coefficient (β), and an estimate of the semivariogram via the sill (σ^2) and range (ϕ) parameters of the stationary Gaussian process. Maximum likelihood is used to estimate β, σ^2, and ϕ. This model is given by:

$$Y(x) = \mu(x) + S(x) + e$$

where Y is the smoothed prevalence, x is the spatial location given by the midpoint coordinates of a PHU, $\mu(x) = X\beta$ is the trend (mean) component of the model, and $S(x)$ is the stationary Gaussian process which is described by a correlation function parameterized by ϕ (range) with variance σ^2 (sill). The residual process, e, has a nugget variance parameter τ^2. The nugget describes the amount of variation which is unaccounted for in the area under study (i.e. measurement error for PHU prevalence estimates). Since the data are smoothed prior to semivariogram estimation, the nugget is set to zero for the analysis in this study.

In this investigation the regional prevalence is modelled only as a function of its geographic location and no further covariates are used. That is, only the PHU

coordinates, as well as their case and control counts, are considered, and no other data is taken from the 2005 surveillance study. Hence, the β estimate is simply the intercept for this model and is interpreted as the global or baseline prevalence of WNv in southern Ontario. If other covariates were of interest and included in the model, the β estimate would be a vector, not a point estimate, and would quantify the average contribution of the covariates to the overall prevalence based on the PHU coordinates provided by x.

The sill parameter (σ^2) of this model is an estimate of the overall spatial variation in PHU prevalence, and the range (ϕ) is seen as the average distance for which there is dependence between observations [5, p. 118]. The semivariogram can thus be interpreted as a measure of spatial dependence and can provide information about disease clustering [24, p. 151].

It should be noted that the residuals for this model must be checked for normality in order to determine whether or not a Gaussian distribution model is a good fit for the data. Preliminary analysis showed that the residuals were approximately normally distributed. Hence, this model was considered appropriate for estimating baseline trend and the semivariogram.

2.4.2 Clustering

Another measure of spatial dependence is Moran's I (I), which assesses the amount of correlation in rates between nearest neighbours [19, p. 21-22]. This statistic produces a coefficient which gives information on the strength of spatial dependence in regional data, and also a test to indicate the significance of disease clustering. It

can be thought of as the spatial equivalent to Pearson's correlation coefficient, where a value near -1 indicates strong negative spatial correlation (or regularity) and a value near $+1$ indicates strong positive spatial autocorrelation. If the data exhibits a spatially random pattern, the expected value of I is $E(I) = \frac{-1}{N-1}$, where N is the number of regions [26, p. 227].

The neighbourhood structure used in Moran's I can be defined by adjacency (e.g. bordering PHUs) or by distance [20, p. 55]. In this investigation, the Euclidean distance between the midpoints of each PHU is used. In order to determine which PHUs are neighbours, the range parameter estimated from the original data set is used. The semivariogram range for the original data set is approximately 100 km. A neighbour is thus defined by any PHU midpoint that falls within a circle with radius 100 km centred at the midpoint of any particular PHU. This method of using the estimated semivariogram range to define neighbourhoods is also used in Berke and Waller (2010) [5]. In this study an empirical Bayes index modification (EBI) of Moran's I as described by Assunção and Reis (1999) [1] is used to quantify the amount of spatial clustering since smoothed regional prevalence estimates are also used in semivariogram estimation. This is given by:

$$EBI = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N}(z_i - \bar{z})^2}$$

where:

$$z_i = \frac{\frac{m_i}{n_i} - \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i}{\sqrt{v_i}}$$

$$v_i = s^2 - \frac{\sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} n_i / N} + \frac{\sum_{i=1}^{N} m_i}{n_i \sum_{i=1}^{N} n_i}$$

$$s^2 = \frac{\sum_{i=1}^{N} n_i (m_i/n_i - \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i)^2}{\sum_{i=1}^{N} n_i}$$

Here N corresponds to the number of observations (i.e. the number of PHUs), m_i is the number of cases in the i^{th} PHU with sample size n_i, and w_i is defined as the spatial weight given to its PHU. The weights given to the PHUs are determined by the neighbourhood structure of the study area. The neighbourhoods are determined using the *dnearneigh* function from the *spdep* package [8]. For the statistical analysis in this investigation, EBI is used exclusively to assess the amount of clustering in the data and will henceforth be referred to as simply Moran's I.

2.4.3 Cluster Detection

For cluster detection, the Kulldorff spatial scan test is used by implementing the *kulldorff* function from the *spdep* package in R [9]. This function has been developed in R in order to perform the spatial scan test as presented by Kulldorff (1997) [17]. This method finds clusters by creating different sized neighbourhoods from the midpoint of any specific PHU and calculating the likelihood for each constructed neighbourhood. The neighbourhoods are constructed by sequentially aggregating neighbouring PHUs until a maximum population limit is reached. The likelihood of each neighbourhood is then calculated using the binomial likelihood. Alternatively, the Poisson likelihood could be used. This would be done in situations where a disease is rare (i.e. prevalence is low). The statistic reports the neighbourhood which is the

most likely cluster, where the significance of a cluster is then determined by Monte Carlo testing [9, p. 11].

Certain parameters must be determined before applying the Kulldorff spatial scan test. These include the maximum cluster population size, the α-level used to determine the significance of a cluster, and the number of Monte Carlo simulations used to produce the p-values. In this investigation, a maximum cluster size of 40% of the total population, an α-level of 0.05, and 999 Monte Carlo simulations are chosen. Finally, the binomial likelihood is used to determine the likelihood of each zone since the overall observed prevalence of the 2005 surveillance data is approximately 27%.

This method will report both a primary cluster location and a secondary cluster location, provided they are both significant. If two cluster locations are found to be significant (p-value $< \alpha = 0.05$), the cluster location with the most evidence is considered to be the primary cluster. If there is no evidence to support a secondary cluster location, only a primary cluster location will be reported. Similarly, if there are no significant clusters, neither a primary cluster nor a secondary cluster will be reported.

2.5 Statistical Analysis and Simulations

In order to identify any bias that may be the result of diagnostic misclassification from an imperfect diagnostic test, analysis must first be performed on the data sets with no diagnostic misclassification. That is, estimates must be obtained for Moran's I, prevalence and the semivariogram, and cluster detection from both the

original data set and the situation where the case and control counts of the original data set have been doubled. These estimates are to serve as benchmark estimates, since they are based on case and control counts which were found using a presumed perfect diagnostic test (i.e. sensitivity and specificity are 100%). Benchmark estimates are obtained for both the 2005 surveillance data and the situation where the regional cases and controls of the surveillance data have been doubled. Once these estimates have been obtained, data sets may be simulated based on imperfect diagnostic tests, and the statistical methods described above may be applied to these data sets in order to determine whether or not the results differ from the benchmarks.

Data sets are simulated based on a variety of reduced values of sensitivity and specificity as described in section 2.2. Specifically, for any particular combination of reduced sensitivity and specificity, 3000 data sets are simulated by randomly distributing the false positives and negatives among the original data sets. For example, in the situation where the original WNv data is used as the reference data, 3000 data sets are produced for each of the 24 combinations of reduced sensitivity and specificity. This translates into $3000 * 24 = 72000$ simulated data sets. This is similarly done for the situation in which the doubled case and control counts of the WNv data is used as the reference data.

For each simulated data set, estimates are obtained for Moran's I, prevalence and the semivariogram, and cluster locations. The mean of the empirical distributions of the estimates for Moran's I, prevalence and the semivariogram from the simulated data are then compared to the benchmark estimates obtained from the original (reference) data sets. For cluster detection, cluster locations from the simulated data are

compared to those found on the reference data. If there are any notable discrepancies between estimates from the simulated data and the benchmark estimates, they can be attributed to diagnostic misclassification.

It should be noted that in order to allow for a comparison between estimates from the data sets with misclassification and estimates from the reference data sets with no misclassification, the parameter choices for each of the statistical tests were the same as those used for the analysis on the reference data sets; for Moran's I, a neighbourhood of 100 km was used; for prevalence and semivariogram estimation, a Gaussian random field model without spatial trend or nugget effect was fit to the data, and the intercept, sill, and range were estimated via maximum likelihood; and for the Kulldorff cluster detection method, a circular scanning method was used with a possible maximum cluster size defined as 40% of the overall population, as well as an α-level of 0.05.

Chapter 3

Results

3.1 Reference Data and Benchmark Estimates

3.1.1 WNv Surveillance Data

This section provides the statistical results for the original data from the 2005 WNv surveillance program. This example was used for analysis in Berke and Waller (2010) [5]. As mentioned previously, the results in this section are taken to represent the true underlying spatial patterns in the data and are used as a benchmark in order to compare the results from analysis on data which includes diagnostic misclassification (section 3.2.1 and 3.2.2).

The original data set consists of 272 WNv positive and 745 WNv negative dead birds. This translates into an overall prevalence of 26.7% for WNv infected dead birds in the entire study area of southern Ontario. For the 30 PHUs, raw regional prevalence ranged from 3.5% to 80.0% and had a mean of 27.8% with a standard deviation of 20.8%. In order to smooth these raw regional prevalence estimates, empirical Bayes smoothing was applied using a binomial model. This resulted in a mean regional prevalence of 27.2% with a standard deviation of 16.7%, where the

smoothed regional prevalence ranged from 7.1% to 66.0%.

To investigate the presence of clustering, an empirical Bayes index modification of Moran's I (I) was used for detecting spatial autocorrelation in the regional prevalence rates. This produced an estimate of $\hat{I} = 0.37$ with a p-value of 0.01, which gives strong statistical evidence of moderate positive clustering in the smoothed regional prevalence rates.

A Gaussian random field model without spatial trend or nugget effect was fit to the smoothed data and maximum likelihood was used to estimate the parameters of the model. The intercept (β) of this model is considered to be an estimate of the mean regional prevalence and was estimated at $\hat{\beta} = 25.0\%$ with a standard error of 7.6%. The sill (σ^2) and range (ϕ) from the stationary Gaussian process were estimated to be $\hat{\sigma}^2 = 0.0334$ and $\hat{\phi} = 100161.65$ (≈ 100.2 km). This indicates that there is spatial dependence in the regional prevalence estimates for distances up to approximately 100 km in the study area.

Finally, the Kulldorff spatial scan test was applied to identify the presence of clusters and their locations in the original data set. The results of the spatial scan test showed a significant primary cluster location containing the PHUs 30, 20, 29, and 7, and another significant secondary cluster containing the PHUs 11, 13, and 28. A choropleth map of the smoothed prevalence rates along with the identified primary and secondary cluster locations is shown in Figure 3.1.

The estimates for this data set are henceforth referred to as 'true values' of the original WNv surveillance data and are written as $I = 0.37$, $\beta = 25.0\%$ (0.2502), $\sigma^2 = 0.0334$, and $\phi = 100161.65$ (100.2 km), since they are assumed to be the results

from analysis on data with no diagnostic misclassification.

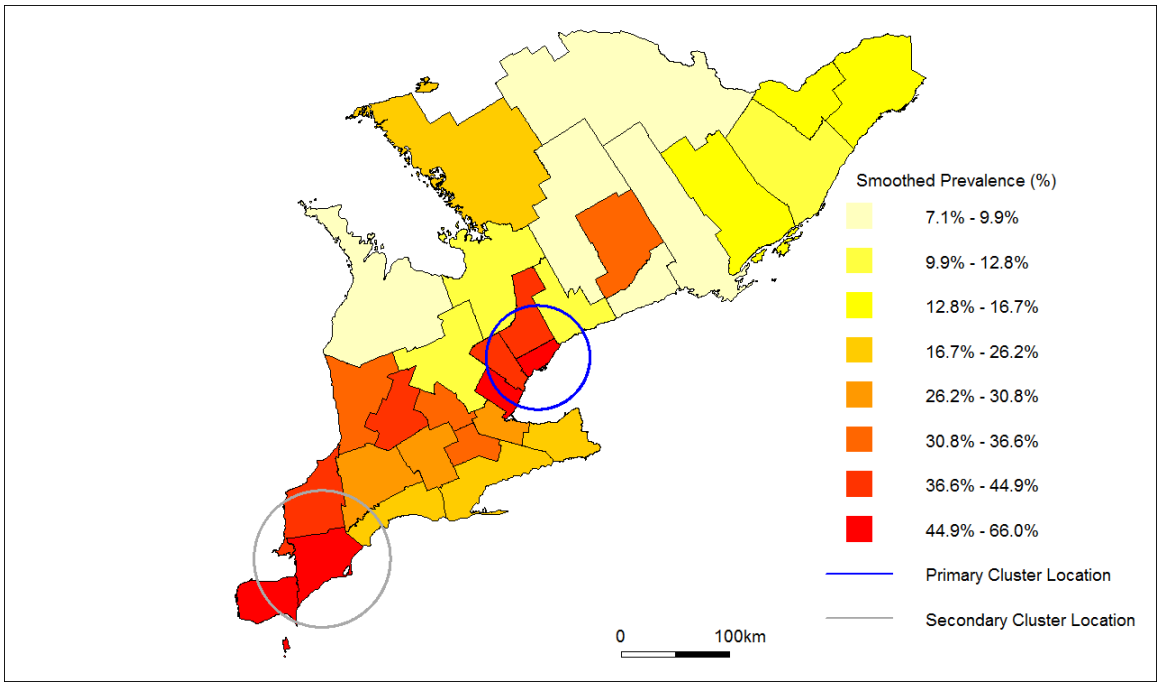


Figure 3.1: Choropleth map for smoothed prevalence rates and cluster locations (2005 WNv Data). The blue circle denotes the primary cluster location and contains the midpoints for the PHUs 30, 20, 29, and 7. The grey circle corresponds to the secondary cluster location and contains the midpoints for the PHUs 11, 13, and 28.

3.1.2 WNv Surveillance Data - Double Regional Cases and Controls

Here, the data used as a reference (i.e. the data in which there is no diagnostic misclassification) is created by simply doubling the number of cases and controls per region in the original data set. This preserves the regional prevalence from the original data set, as well as the inherent spatial patterns in the data, and allows for an investigation into the impact of misclassification on the performance of Moran's I, prevalence and semivariogram estimation, and Kulldorff's spatial scan test with respect to an increase in sample size.

For this reference data, Moran's I test statistic was estimated at approximately $\hat{I} = 0.39$ and had a p-value of 0.01, which gives strong statistical evidence for spatial clustering. The estimate for the intercept of the Gaussian random field model was $\hat{\beta} = 25.55\%$ and had a standard error of approximately 9.2%. The sill and the range were estimated at $\hat{\sigma}^2 = 0.043$ and $\hat{\phi} = 113854.2$, respectively, which gives a further indication of spatial dependence among observations in the WNv data with doubled regional cases and controls. Finally, the Kulldorff spatial scan test determined a primary cluster location containing the PHUs 30, 20, 29, and 7 and a secondary cluster location of 11, 13, and 28, which were the same locations found from analysis on the original 2005 WNv data set.

The estimates for this data set are henceforth referred to as 'true values' of the WNv surveillance data with doubled regional cases and controls, and are written as $I = 0.39$, $\beta = 25.55\%$ (0.2555), $\sigma^2 = 0.043$, and $\phi = 113854.2$ (113.9 km), since they are assumed to be the results from analysis on the doubled data with no diagnostic misclassification.

3.2 Simulated Data

3.2.1 Simulated Data: WNv Surveillance Data

For each combination of sensitivity and specificity, 3000 different data sets were simulated. For each data set, estimates for Moran's I, prevalence, the sill and range of the semivarioagram, and cluster locations were obtained and the means of these estimates (with the exception of cluster locations) were calculated. The means

of these estimates are represented by $\bar{\hat{I}}$, $\bar{\hat{\beta}}$, $\bar{\hat{\sigma}}^2$, and $\bar{\hat{\phi}}$.

In addition to the means, the standard deviations (SD) of $\hat{I}$ and $\hat{\beta}$ were found, as well as the average standard error ($\bar{SE}_{\hat{\beta}}$) for the $\hat{\beta}$s. Furthermore, the power of Moran's I test for clustering was calculated. Finally, the percentage of simulations in which the Kulldorff scan statistic had identified the same primary and secondary cluster locations as the benchmark cluster locations was determined. The results show that when sensitivity is fixed and specificity decreases, there is a decline in the mean estimated value of Moran's I, sill, and range, and an increase in the mean overall prevalence estimates. It is also shown that the scan test begins to misidentify cluster locations when the quality of the diagnostic test is reduced. These results can be seen in Table 3.1 and are discussed in sections 3.3.1, 3.4.1, and 3.5.1.

3.2.2 Simulated Data: WNv Surveillance Data - Double Regional Cases and Controls

For the WNv data with doubled regional cases and controls, the number of data sets simulated for each of the 24 combinations of sensitivity and specificity was again 3000. The mean estimates of Moran's I, prevalence, sill, and range for these simulations can be seen in Table 3.2, along with the percentage of simulations which accurately identified the primary and secondary cluster locations and the power of Moran's I. With regard to a decrease in sensitivity and specificity, the tendencies of these estimates are similar to that of those in Table 3.1. Specifically, when sensitivity is fixed and specificity decreases, there is a systematic decline in the mean estimated

Table 3.1: Mean estimates for Moran's I ($\bar{\hat{I}}$), prevalence ($\bar{\hat{\beta}}$), sill ($\bar{\hat{\sigma}}^2$) and range ($\bar{\hat{\phi}}$) of the semivariogram, as well as the power of Moran's I (Power(I)) and percentage of simulations which accurately identified the primary (%PC) and secondary (%SC) cluster locations found in the original WNv data set. $SD_{\hat{I}}$ and $SD_{\hat{\beta}}$ represent the standard deviations of the estimates for $\hat{I}$ and $\hat{\beta}$. $\bar{SE}_{\hat{\beta}}$ is the average standard error for the β estimates given by the Gaussian random field model. Sensitivity and specificity correspond to the quality of the diagnostic test used to simulate the data. Estimates when sensitivity = 100% and specificity = 100% correspond to the benchmark estimates found on the original WNv data set. 3000 data sets were simulated per combination of sensitivity and specificity.

Sensitivity	Specificity	$\bar{\hat{I}}$ ($SD_{\hat{I}}$)	Power(I)	$\bar{\hat{\beta}}$ ($\bar{SE}_{\hat{\beta}}$, $SD_{\hat{\beta}}$)	$\bar{\hat{\sigma}}^2$ ($\bar{\hat{\phi}}$)	(%)PC	(%)SC
100%	100%	0.3706 (-)	-	0.2502 (0.0760, -)	0.0334 (100161.65)	-	-
	95%	0.3500 (0.0351)	99.97%	0.2854 (0.0662, 0.006)	0.0280 (90847.77)	96.47%	31.77%
	90%	0.3288 (0.0504)	99.83%	0.3213 (0.0577, 0.0071)	0.0234 (82694.79)	88.63%	49.5%
	85%	0.3079 (0.0618)	98.43%	0.3588 (0.0498, 0.0077)	0.0193 (75424.83)	79.03%	32.9%
	80%	0.2867 (0.0725)	93.67%	0.3960 (0.043, 0.0077)	0.0159 (68589.92)	68%	21.07%
95%	100%	0.3629 (0.0283)	100%	0.2368 (0.0697, 0.0057)	0.0293 (96190.52)	99.67%	72.87%
	95%	0.3415 (0.0468)	99.9%	0.2723 (0.0603, 0.0074)	0.0244 (86949.17)	88.67%	66.7%
	90%	0.3183 (0.0602)	98.33%	0.3082 (0.0515, 0.008)	0.0199 (77903.91)	79.13%	35.57%
	85%	0.2926 (0.0703)	95.33%	0.3463 (0.0441, 0.0079)	0.0162 (70143.82)	68.7%	27.13%
	80%	0.2696 (0.0809)	88.03%	0.3837 (0.0374, 0.0075)	0.0130 (63258.25)	60.87%	13.97%
90%	100%	0.3549 (0.039)	99.97%	0.2242 (0.064, 0.0073)	0.0258 (92121.73)	96.17%	67.73%
	95%	0.3299 (0.0563)	99.33%	0.2598 (0.0544, 0.0079)	0.0210 (82077.4)	82.77%	53.73%
	90%	0.3069 (0.0686)	97.03%	0.2964 (0.0464, 0.008)	0.0172 (73485.53)	72.03%	38.77%
	85%	0.2786 (0.0791)	90.5%	0.3349 (0.0389, 0.0078)	0.0137 (65133.77)	60.93%	24.5%
	80%	0.2553 (0.0906)	82.33%	0.3723 (0.0327, 0.0075)	0.0110 (57610.87)	51.9%	9.1%
85%	100%	0.3481 (0.0494)	99.8%	0.2110 (0.0583, 0.0083)	0.0223 (88472.19)	88.53%	63.5%
	95%	0.3181 (0.0638)	98.63%	0.2469 (0.0487, 0.0082)	0.0179 (77498.91)	75.23%	43.97%
	90%	0.2921 (0.0766)	93.17%	0.2841 (0.0407, 0.0079)	0.0143 (67915.07)	64.37%	31.27%
	85%	0.2629 (0.0868)	85.47%	0.3223 (0.0333, 0.0076)	0.0111 (59120.22)	53.13%	13.13%
	80%	0.2352 (0.0935)	76.2%	0.3608 (0.0275, 0.0073)	0.0086 (51773.98)	45.7%	6.03%
80%	100%	0.3388 (0.0567)	99.33%	0.1990 (0.0529, 0.0089)	0.0193 (84540.29)	81.93%	53.13%
	95%	0.3100 (0.0699)	96.5%	0.2348 (0.0431, 0.008)	0.0150 (72291.3)	68.4%	41.8%
	90%	0.2811 (0.0835)	90.1%	0.2723 (0.0359, 0.0075)	0.0119 (63618.13)	56.73%	23.53%
	85%	0.2453 (0.092)	79.6%	0.3112 (0.0289, 0.0071)	0.0091 (54173.31)	49.4%	8.67%
	80%	0.2150 (0.1028)	66.9%	0.3499 (0.0232, 0.0068)	0.0069 (46105.29)	40.87%	3.03%

Table 3.2: Doubled data mean estimates for Moran's I ($\bar{\hat{I}}$), prevalence ($\bar{\hat{\beta}}$), sill ($\bar{\hat{\sigma}}^2$) and range ($\bar{\hat{\phi}}$) of the semivariogram, as well as the power of Moran's I (Power(I)) and percentage of simulations which accurately identified the primary (%PC) and secondary cluster (%SC) locations. $SD_{\hat{I}}$ and $SD_{\hat{\beta}}$ are the standard deviations of the estimates for $\hat{I}$ and $\hat{\beta}$. $\bar{SE}_{\hat{\beta}}$ is the average standard error for the β estimates of the Gaussian random field model. Sensitivity (SE) and specificity (SP) correspond to the quality of the diagnostic test used to simulate the data. Estimates when SE = 100% and SP = 100% correspond to the benchmark estimates found on the WNv data set with doubled regional cases and controls. 3000 data sets were simulated per combination of SE and SP.

Sensitivity	Specificity	$\bar{\hat{I}}$ ($SD_{\hat{I}}$)	Power(I)	$\bar{\hat{\beta}}$ ($\bar{SE}_{\hat{\beta}}$, $SD_{\hat{\beta}}$)	$\bar{\hat{\sigma}}^2$ ($\bar{\hat{\phi}}$)	(%)PC	(%)SC
100%	100%	0.3866 (-)	-	0.2555 (0.0921,-)	0.0434 (113854.2)	-	-
	95%	0.3762 (0.0256)	100%	0.2907 (0.0834, 0.0057)	0.0375 (107950.75)	99.57%	56.67%
	90%	0.3623 (0.0371)	100%	0.3263 (0.0745, 0.0071)	0.0320 (101103.54)	97.3%	55.37%
	85%	0.3507 (0.0457)	99.9%	0.3630 (0.0668, 0.0078)	0.0272 (95551.22)	91.83%	42.93%
	80%	0.3351 (0.0538)	99.63%	0.3987 (0.0592, 0.0083)	0.0229 (89326.39)	85.83%	43.17%
95%	100%	0.3819 (0.02)	100%	0.2422 (0.0858, 0.0056)	0.0386 (111153.78)	100%	78.07%
	95%	0.3703 (0.0338)	100%	0.2776 (0.0769, 0.0074)	0.0329 (104597.21)	97.7%	63.87%
	90%	0.3566 (0.0442)	99.93%	0.3133 (0.0683, 0.0082)	0.0278 (97887.13)	92.47%	53.9%
	85%	0.3422 (0.0527)	99.73%	0.3496 (0.0606, 0.0084)	0.0233 (91778.3)	86.7%	50.83%
	80%	0.3268 (0.0606)	99.23%	0.3860 (0.053, 0.0083)	0.0193 (85243.79)	80.23%	46.17%
90%	100%	0.3775 (0.0279)	100%	0.2287 (0.0795, 0.0074)	0.0340 (108269.85)	99.6%	75.7%
	95%	0.3652 (0.0395)	99.97%	0.2641 (0.0708, 0.0083)	0.0288 (101513.91)	94.9%	63.13%
	90%	0.3481 (0.05)	99.87%	0.3000 (0.0621, 0.0087)	0.0240 (93758.31)	87.83%	54.2%
	85%	0.3323 (0.0586)	99.3%	0.3368 (0.054, 0.0086)	0.0196 (86804.17)	80.87%	49.53%
	80%	0.3117 (0.0675)	97.33%	0.3730 (0.0468, 0.0084)	0.0161 (79696.59)	73.2%	42.6%
85%	100%	0.3736 (0.0354)	100%	0.2157 (0.0736, 0.0086)	0.0298 (106269.81)	96.97%	71.33%
	95%	0.3579 (0.0474)	99.8%	0.2505 (0.0644, 0.0087)	0.0247 (97765.82)	90.57%	63.83%
	90%	0.3407 (0.0576)	99.4%	0.2867 (0.0558, 0.0088)	0.0203 (89828.54)	82.93%	54.8%
	85%	0.3194 (0.0674)	97.8%	0.3235 (0.0479, 0.0087)	0.0164 (81627.31)	74.27%	45.2%
	80%	0.2994 (0.0748)	94.97%	0.3597 (0.0407, 0.0079)	0.0131 (74012.8)	65.53%	37.57%
80%	100%	0.3679 (0.0413)	99.93%	0.2019 (0.0673, 0.0091)	0.0258 (102402.37)	94%	67.6%
	95%	0.3494 (0.0542)	99.4%	0.2372 (0.058, 0.0091)	0.0210 (93287.62)	85.2%	58.37%
	90%	0.3294 (0.0649)	96.33%	0.2739 (0.0497, 0.0088)	0.0170 (84911.03)	77.4%	49.4%
	85%	0.3071 (0.0738)	89.7%	0.3106 (0.0417, 0.0083)	0.0134 (76277.22)	68.3%	44.1%
	80%	0.2843 (0.0825)	75.4%	0.3477 (0.0349, 0.0073)	0.0104 (68534.67)	58.73%	31.73%

value of Moran's I, sill, and range, and a systematic increase in the mean overall prevalence estimates. In addition, it is shown that the scan test again begins to misidentify clusters when sensitivity and specificity are reduced. However, it can be seen in Table 3.2 that the effect of diagnostic misclassification on the results is reduced when compared to the results in section 3.2.1 (Table 3.1), which is discussed in sections 3.3.2, 3.4.2, and 3.5.2.

3.3 Moran's I

3.3.1 WNv Surveillance Data

The results in Table 3.1 show that when sensitivity is fixed and specificity decreases, there is a systematic decline in the mean estimated value of Moran's I. In particular, the mean estimates from the 3000 simulations for each combination of sensitivity and specificity are strictly less than the true value of $I = 0.37$ from the original 2005 WNv surveillance data. This indicates an apparent decline in spatial clustering as sensitivity and specificity are reduced. However, Table 3.1 also shows that the power of Moran's I test for clustering (Power(I)) decreases as sensitivity and specificity are reduced. Thus, the chance of committing a type II error increases. That is, for a decline in sensitivity and specificity, there is an increased chance of Moran's I failing to reject the null hypothesis when there is indeed spatial autocorrelation in the data.

Since clustering was found in the benchmark data, it is assumed that the alternative hypothesis is true (i.e. there is spatial autocorrelation in the data). Hence,

the power of Moran's I can be calculated by determining the percentage of simulations which correctly reject the null hypothesis. The results show that the power of Moran's I can be as low as 67% (SE = 80% and SP = 80%) which indicates nearly a third of the tests on these simulated data sets failed to reject the null hypothesis.

Boxplots for the estimates of Moran's I for each combination of sensitivity and specificity can be seen in Figure 3.2. This figure shows that, for a decrease in specificity, the variability of the estimates increases along with the decline in the mean estimated value of Moran's I. The notches surrounding the median in this figure denote an approximate 95% confidence interval for the median empirical estimate of Moran's I. It can be seen that none of the 95% confidence intervals contain the true value of $I = 0.37$, which gives an added indication of a strict decline in spatial clustering as the quality of diagnostic tests decreases.

3.3.2 WNv Surveillance Data - Double Regional Cases and Controls

For the doubled data, the results in Table 3.2 show that the mean estimates for Morans I were all below the true value of $I = 0.39$. Again, this suggests a decline in apparent spatial dependence as sensitivity and specificity are reduced. However, in comparison to the results shown in Table 3.1, the decrease in the Moran's I coefficient is not as severe. This suggests that increasing regional sample size mitigates the effect of diagnostic misclassification with respect to the estimation of the amount of spatial clustering as determined by Moran's I.

Since the p-value for the true value of $I = 0.39$ was 0.01, the null hypothesis was rejected and it is assumed that there is spatial autocorrelation in the data

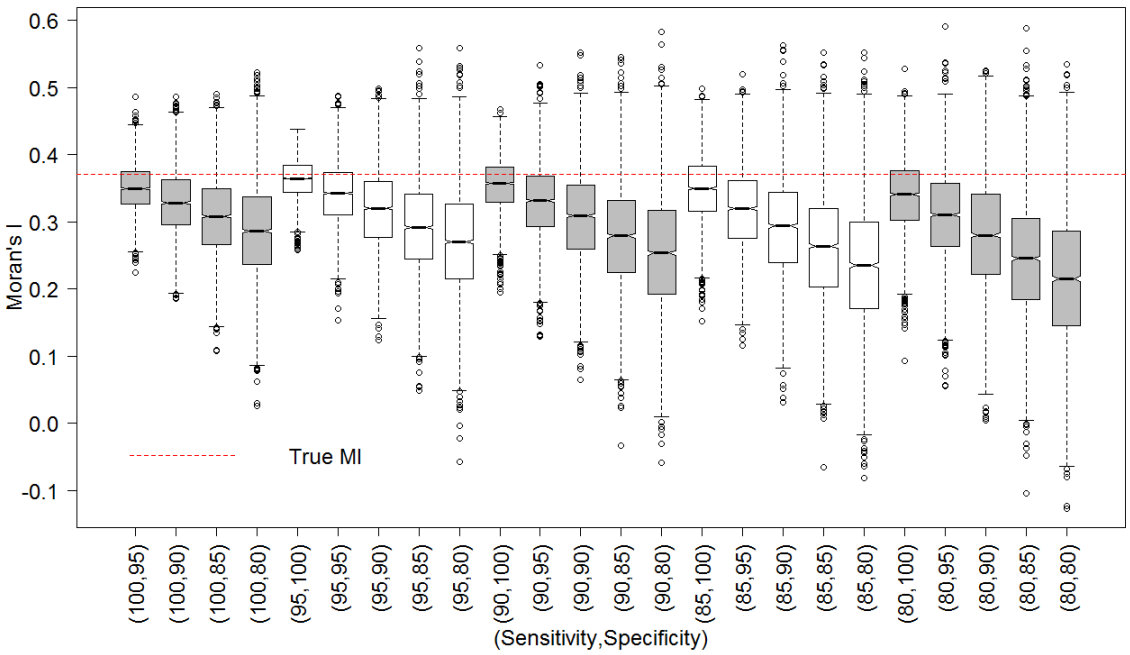


Figure 3.2: Boxplots for Moran's I ($\hat{I}$) from simulated data which was obtained from the application of imperfect diagnostic tests on the original WNv data set. The notches surrounding the median represent approximate 95% confidence intervals for the median estimates of $\hat{I}$. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

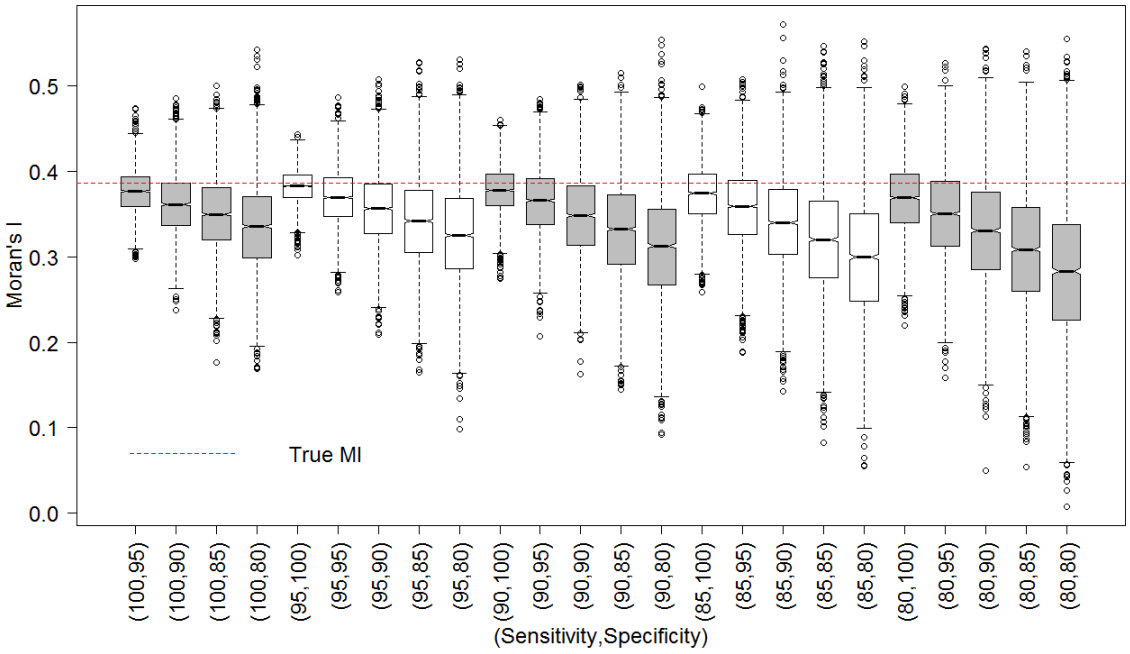


Figure 3.3: Boxplots for Moran's I ($\hat{I}$) from simulated data which was obtained from the application of imperfect diagnostic tests on the WNv data set with doubled regional cases and controls. Notches surrounding the median represent approximate 95% confidence intervals for the median estimates of $\hat{I}$. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

with doubled regional cases and controls. Here, the power of Moran's I was again calculated by determining the number of simulations in which the test for clustering correctly rejected the null hypothesis. It can be seen from the results in Table 3.2 that the power of Moran's I is greater for lower values of sensitivity and specificity in contrast to the results in Table 3.1. This gives a further indication that increasing regional sample size can reduce the effect of diagnostic misclassification on the Moran's I test for clustering.

Boxplots for the mean estimates of Moran's I for the doubled data can be seen in Figure 3.3. This figure shows that none of the 95% confidence intervals for the median contain the original estimate of 0.39. It can be seen from this figure that the amount of variability between estimates (for each combination of sensitivity and sensitivity) is smaller in comparison to the results in 3.2. This is also shown by the standard deviation estimates of Moran's I in Table 3.2.

3.4 Gaussian Random Field Model

3.4.1 WNv Surveillance Data

Further evidence of a decline in spatial autocorrelation as sensitivity and specificity are reduced is given by the mean estimates for the sill and range parameters. As with Moran's I, these estimates decrease with respect to the true values of $\sigma^2 = 0.0334$ and $\phi = 100.2$ km as sensitivity and specificity decrease. Furthermore, as the range becomes smaller, so does the overall variance (i.e. the sill). In particular, the estimates in Table 3.1 show that for any fixed level of sensitivity, as specificity

is decreased from 100% to 80%, the estimated range at which prevalence estimates are correlated is reduced by approximately 45 km. This is also accompanied by a reduction in the sill. Semivariogram plots can be seen in appendix A.1.1.

Prevalence estimates are given by the maximum likelihood estimates of the intercept from the Gaussian random field model. The estimates in Table 3.1 show that when sensitivity is fixed and specificity decreases, the mean prevalence estimates increase. This is due to the fact that a decrease in specificity will generate more false positives in the data when sensitivity is fixed, and hence increase the estimated prevalence. Similarly, more false negatives are generated in the data when sensitivity decreases, leading to a decrease in estimated prevalence. As a result, the spatial dependence among observations decreases, and the prevalence estimates from the Gaussian random field model become more stable. This is evidenced by the $\bar{SE}_{\hat{\beta}}$ estimates in Table 3.1, which show that the standard errors of the prevalence estimates decreases when sensitivity and specificity are reduced.

The results show that $\bar{SE}_{\hat{\beta}}$ and $SD_{\hat{\beta}}$ are quite different from one another. This is due to the nature of the simulated data in this study. Since $\bar{SE}_{\hat{\beta}}$ is the average of the model based standard errors for $\hat{\beta}$, this includes the variability due to sampling error (from generating new data sets) and misclassification. That is, each SE estimate provided by the Gaussian random field model corresponds to the variability in $\hat{\beta}$ that would be expected in a typical simulation. On the other hand, since $SD_{\hat{\beta}}$ corresponds to the empirical standard deviation of the simulated $\hat{\beta}$'s, this only includes the variability in $\hat{\beta}$ due to misclassification. Hence, $SD_{\hat{\beta}}$ is smaller than

$\bar{SE}_{\hat{\beta}}$ because it does not include variability due to sampling error.

Boxplots for the prevalence estimates can be seen in Figure 3.4. This shows that 6 combinations of sensitivity and specificity lead to an underestimation of the true prevalence, while the rest of the combinations overestimate the prevalence. Furthermore, none of the approximate 95% confidence intervals for the median of the β estimates, again denoted by notches, contain the true value of $\beta = 0.2502$.

3.4.2 WNv Surveillance Data - Double Regional Cases and Controls

For the doubled data, additional evidence of clustering is given by the mean estimates for the sill and range of the semivariogram. As was the case for the simulated data presented in section 3.4.1, these mean estimates show spatial clustering in the study area for up to around 110 km, specifically for high values of specificity. The results in Table 3.2 also demonstrate that when the sensitivity of a diagnostic test is fixed, and the specificity is lowered, the range at which there is spatial dependence decreases, as well as the amount of overall variance given by the sill. In particular, the mean estimated range decreased by approximately 30 km when specificity was reduced from 100% to 80%, and the overall variability in prevalence drops by nearly 2%. These results are similar to those presented in section 3.4.1, though in this case both the sill and range have higher mean estimates for each combination of sensitivity and specificity. Plots of the semivariograms for these simulations can be seen in A.1.2.

Since these simulations were based on reference data which had double the case and control counts found in the original data, the mean estimates for the beta parameter (i.e. the prevalence) for the simulated data sets were very similar to the

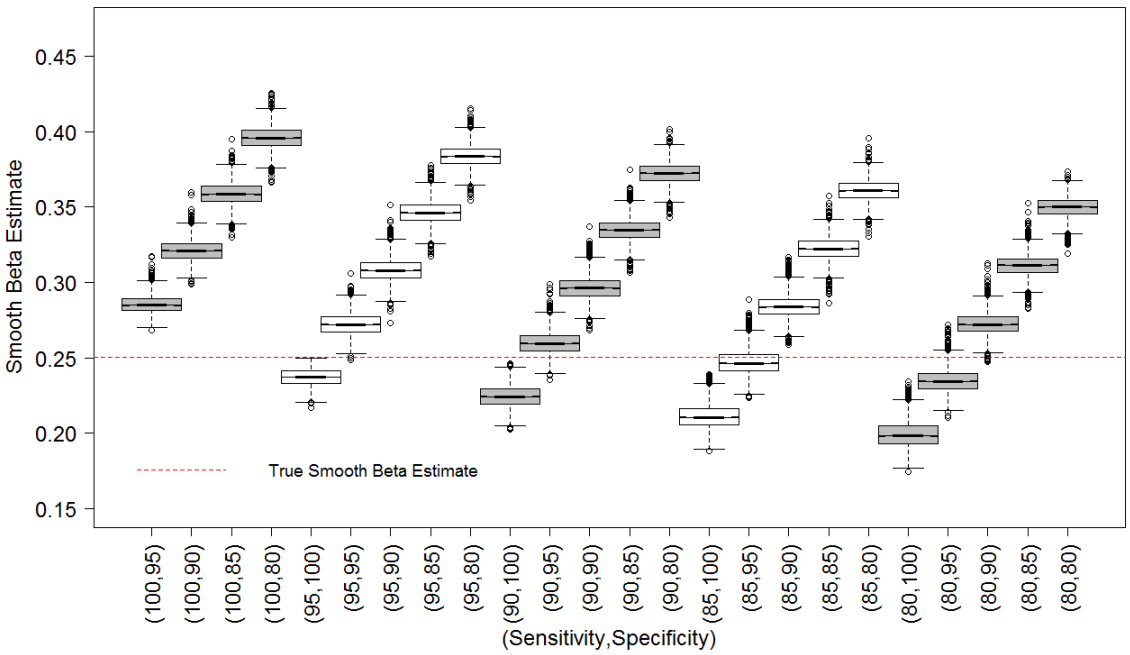


Figure 3.4: Boxplots for prevalence estimates ($\hat{\beta}$) from simulated data which was obtained from the application of imperfect diagnostic tests on the original WNv data set. The notches surrounding the median represent approximate 95% confidence intervals for the median estimates of $\hat{\beta}$. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

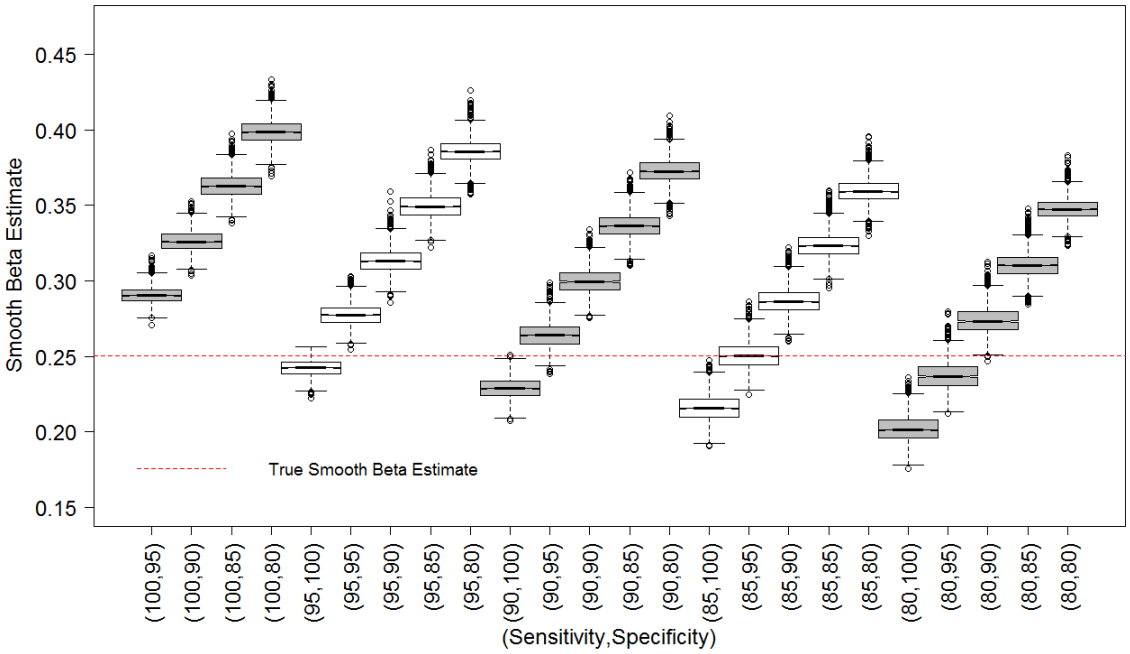


Figure 3.5: Boxplots for prevalence estimates ($\hat{\beta}$) from simulated data which was obtained from the application of imperfect diagnostic tests on the WNv data set with doubled regional cases and controls. Notches surrounding the median represent approximate 95% confidence intervals for the median estimates of $\hat{\beta}$. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

results in the previous section. The results here again show a steady increase in the mean estimated prevalence when sensitivity is fixed and specificity is reduced. However, these estimates are slightly higher than the estimates in section 3.2.1, which may be the result of less smoothing due to higher regional sample size. Furthermore, it is shown that the $\bar{SE}_{\hat{\beta}}$ estimates are slightly higher than those presented in 3.2.1. This can be attributed to the reduced impact of false positives and negatives on the estimation of spatial dependence. In other words, by increasing the sample size, it is correctly predicted that there exists clustering in the data, which gives rise to variability in the estimation of prevalence.

Boxplots for the $\hat{\beta}$ estimates can be seen in Figure 3.5. In comparison to the boxplots in 3.4, it can be seen that the same combinations of sensitivity and specificity underestimate the true prevalence. Also, similar results hold with regard to the 95% confidence intervals, where none contain the true value of $\beta = 0.2555$ of the reference data set used in these simulations.

3.5 Kulldorff Spatial Scan Test

3.5.1 WNv Surveillance Data

The two far right columns in Table 3.1 show the number of simulations which exactly identified the benchmark primary and secondary cluster locations. These quantities do not consider scenarios where cluster location estimates from simulated data are close to the benchmark locations, i.e. in scenarios where some or most (but not all) of the PHUs in the benchmark primary and secondary cluster locations are

accurately identified in the simulated data. This is indeed the case for the results simulated by a diagnostic test with 100% sensitivity and 95% specificity. Here, only 34% of the simulations predicted the exact location of the secondary cluster which contained the PHUs 11, 13, and 28. However, most of the location estimates in this scenario identified a secondary cluster location containing the PHUs 11 and 28. A bar plot for the number of unique misidentified primary cluster locations can be seen in Figure 3.6.

As an alternative, choropleth maps may be used to visualize the cluster locations predicted from the simulated results of each combination of sensitivity and specificity. These maps show the percentage of simulations in which each PHU was identified in a cluster location, and can be found in A.2.1. It should be noted, however, that cluster locations from the Kulldorff scan test are defined by neighbouring PHUs within a circular radius of a centroid for given PHU. Hence, it is important to keep in mind neighbourhood structures when inspecting the these choropleth maps.

3.5.2 WNv Surveillance Data - Double Regional Cases and Controls

Lastly, Table 3.2 shows that, when sample size is doubled, a better estimation of cluster locations is achieved. Specifically, in comparison to the results in 3.2, the number of unique misidentified clusters is reduced and the percentage of simulations in which the scan test correctly identified the true cluster locations increased. For the misidentified cluster locations, the results are nearly halved, which suggests that the number of misidentified cluster locations is inversely proportional to the increase in regional sample size. The barplot for this can be seen in Figure 3.7.

The choropleth maps for the PHUs in southern Ontario which were identified as part of cluster regions can be seen in A.2.2. In contrast to the results in section 3.2.1, these maps show that when regional sample size is doubled the Kulldorff spatial scan test provides more accurate results in terms of the percentage of simulations which correctly identified the PHUs belonging to the true cluster region.

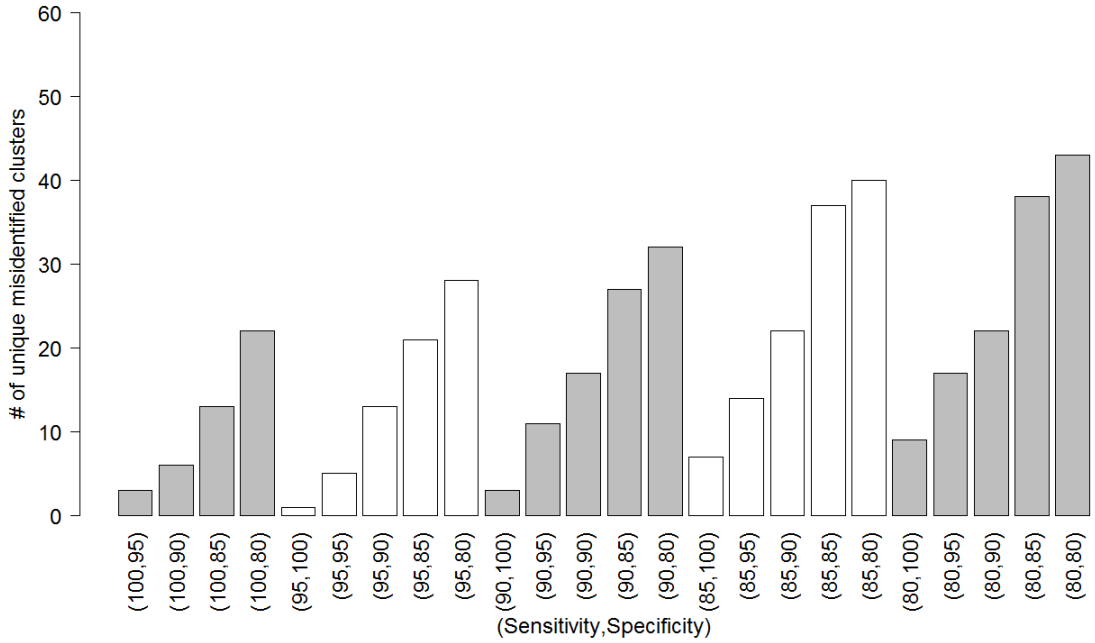


Figure 3.6: Number of (unique) misidentified cluster locations provided by the Kulldorff spatial cluster detection method when applied to simulated data based on the original WNv data set. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

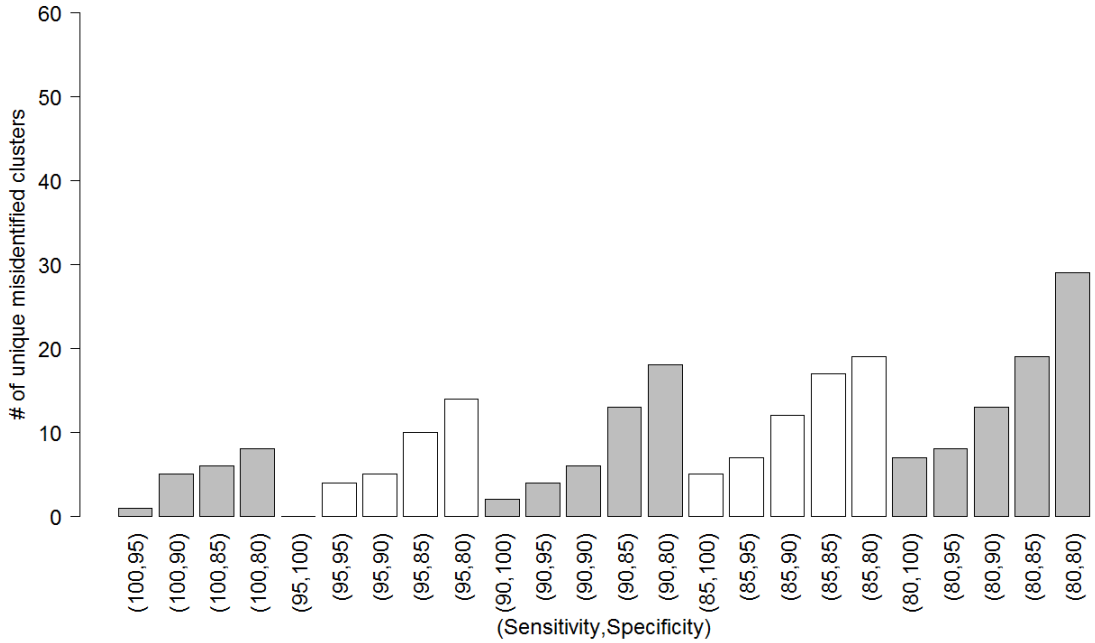


Figure 3.7: Number of (unique) misidentified cluster locations provided by the Kulldorff spatial cluster detection method when applied to simulated data based on the WNv data set with doubled regional cases and controls. The values for sensitivity (SE) and specificity (SP) denote the quality of the diagnostic test used for simulating the data. 3000 data sets were simulated per combination of SE and SP.

Chapter 4

Discussion

4.1 Summary of Results

This study investigates the effect of diagnostic misclassification on the results for statistics which measure spatial prevalence, clustering, and clusters in regional data. Specifically, when an imperfect diagnostic test is used to assess the disease status of regional populations, it is shown here that the results for Moran's I, and prevalence, sill and range for the Gaussian random field model are all biased with respect to the true values obtained via the application of these tests on error-free case-control data. In addition, the analysis shows that the Kulldorff spatial scan test overlooks cluster locations and identifies phantom clusters when the sensitivity and specificity of a diagnostic test are reduced.

Moran's I and the sill and range of the semivariogram are underestimated in the presence of misclassification. This is due to the spatial randomness which is added to the underlying distribution by implementing an imperfect diagnostic test. Within the context of the simulations, a reduction in sensitivity and specificity generates more false positive and negatives in the study region, which disrupts the spatial correlation structure since misclassified counts are distributed randomly. As sensitivity and

specificity approach 80%, the results in section 3.2.1 show that the power of Moran's I test for clustering is reduced, where nearly a third of tests on the simulated data sets committed a type II error. When regional case and control counts were doubled (section 3.2.2), however, it was shown that the effect of diagnostic misclassification on the power of Moran's I was reduced. Specifically, only two combinations of sensitivity and specificity (SE=80% and SP=85%; SE=80% and SP=80%) lead to a power of less than 90%. This indicates that increasing the number of regional samples may reduce the effect of misclassification on the detection of spatial clustering.

For the semivariogram of the Gaussian random field model, both the analysis in sections 3.4.1 and in 3.4.2 showed that the estimates for the sill and the range were reduced for a decline in sensitivity and specificity. Again, for the simulations considered here, a decline in sensitivity and specificity will, in turn, generate more false positives and false negatives among the data and hence increase the amount of spatial randomness. Thus, the estimates for the sill and the range will tend towards the null hypothesis (i.e. towards a spatial random pattern in the data). In some cases, the range of the semivariogram was estimated at around 40 to 50 km which, with regard to the study area, could suggest minimal to no clustering since some PHUs in southern Ontario are around 50 km in diameter.

On the other hand, prevalence estimates based on the intercept ($\hat{\beta}$) of the Gaussian random field model are affected by diagnostic misclassification in a different way. A decrease in specificity will increase the number of false positives distributed among the data and therefore cause prevalence estimates to rise. A decrease in sensitivity will decrease prevalence estimates due to an increased number of false

negatives being distributed among the regions. This is true for both the results in sections 3.4.1 and 3.4.2. Still, it is shown that the results differ, in that the mean prevalence estimates in 3.4.2 are higher than those in 3.4.1. This is due to the fact that less smoothing is required for the data with doubled cases and controls. Furthermore, the average standard error and standard deviation for the prevalence estimates were higher in section 3.2.2, which might be considered as the result of more spatial autocorrelation in the data.

In some instances, prevalence estimates are similar or near to the true overall prevalence. For example, for the original WNv surveillance data, this occurs for a diagnostic test with a sensitivity of 90% and a specificity of 95%. For this particular combination of sensitivity and specificity, the number of misclassified cases and controls are equal or comparable to one another. In this situation, the false positives and false negatives cancel each other out, and the overall observed prevalence is close to the true underlying prevalence. However, it is important to keep in mind that this is a measure of global prevalence. Local prevalence might be vastly different in this situation since the regions in the study area may contain different numbers of misclassified cases and controls. Hence, this should be considered before applying or interpreting tests for clustering or cluster locations.

For cluster locations, the Kulldorff spatial scan test performed well for high values of sensitivity and specificity, but began to degrade once the quality of the diagnostic test was reduced. When this occurs, the method either detects no clusters, partial clusters, or cluster locations which can be referred to as phantom clusters [15, p. 2]. The analysis here showed that, for fixed values of sensitivity, a decrease

in specificity corresponded with a decrease in the percentage of simulations which accurately identified true cluster locations as well as an increase in the number of unique misclassified cluster locations. By increasing regional sample size, however, a reduction in this effect was achieved.

To summarize, the results presented here can be viewed (in part) as a confirmation of the results found in Berke and Waller (2010) [5], in which it was reported that the spatial statistics were not seriously affected by diagnostic misclassification for large sample sizes [6, p. 121]. That is, while the results in section 3.2.1 show that the β, σ^2, ϕ estimates for the Gaussian random field model, Moran's I, and cluster location estimates were all affected by misclassification bias, the results in section 3.2.2 demonstrate that an increase in sample size reduces this effect.

4.2 Applications

Phantom clusters cause false alarms by either identifying larger locations which contain the true cluster locations of the disease or by identifying entirely different regions which do not contain the true cluster. This is important to keep in mind for analysis on emerging diseases with unknown etiology, whereby imperfect diagnostic tests are often used in order to provide a rapid assessment on the disease status of a population [5]. As shown here, poor quality diagnostic testing may result in a misrepresentation of spatial autocorrelation, prevalence estimation, or cluster locations. Thus, in these types of situations, where results from low quality diagnostic tests are relied upon for the purpose of swift analysis on the population, the

results provided by this study might be useful for making informed decisions about the quality of statistical analysis or what steps might be appropriate for preventative action (e.g. cluster containment).

As such, the simulation study in this investigation may be suitable for determining whether or not reliable statistical results are, or can be, acquired from analysis on data for which imperfect diagnostic tests were used to obtain case-control counts. For example, in a study provided by Cohen et al. (2009) [10] it was found that rapid diagnostic tests used to detect dengue in northern and northeastern Thailand had sensitivity which ranged from 7.6% to 21.5% and specificity that ranged from 87.7% to 98.9%. Here it is clear that spatial analysis relating to prevalence estimation, clustering, or clusters of the regional population might be unreliable due to the low sensitivity of these diagnostic tests. However, it is important to keep in mind that this investigation only considered diagnostic tests with sensitivity as low as 80%.

Another example for which this simulation study may be useful is a study presented by Poljak et al., where the Kulldorff spatial scan test was used to identify clusters of swine herds infected with H1N1 and H3N2 influenza in southern Ontario [21]. For H1N1 influenza, a diagnostic test with a sensitivity of 98.8% and specificity of 91.6% was used to determine the disease status of the herds, and for H3N2, a diagnostic test was used which had a sensitivity of 87.8% and specificity of 94.8%. With the high values of sensitivity and specificity of these two diagnostic tests, the identified cluster locations obtained in the Poljak study might be considered reliable when consideration is given to the results presented by the simulation study in sections 3.2.1 and 3.2.2. That is, with regard to primary cluster location, the results from the

simulation study show that, based on 3000 simulated data sets, a diagnostic test with sensitivity of 95% and a specificity of 90% will correctly identify the cluster location approximately 79% to 92% of the time (depending on sample size), and a diagnostic test with sensitivity of 85% and 95% will correctly identify the cluster location approximately 75% to 91% of the time (again depending on sample size). However, it should be noted that point data was used in Poljak et al., so the results in the simulation study may not be directly applicable.

In general, this simulation study is useful for two main reasons. One is for determining whether or not statistical analysis is appropriate based on the quality of diagnostic tests used for assessing regional disease counts. The other is for evaluating the quality of results from statistical tests for trend, clustering, and clusters on data which was found using an imperfect diagnostic test. The latter is particularly important when dealing with crisis management; for example, when response to cluster location is required. Since cluster response often requires immediate action, it is useful to have a variety of statistical tools at hand to provide complementary analysis and insight from alternative perspectives [22, p. 829]. Thus, where the analysis in this study is concerned, the results can be used as an additional tool for making decisions on whether the reported cluster locations are likely to be true or phantom clusters, if regional sampling should be increased, or whether better diagnostic tests are required.

4.3 Extrapolation and Assumptions

Before using the results in this study, it is important to note particular decisions that were made with regard to data simulation and statistical analysis. The first was that the case and control counts of the WNv data from the 2005 surveillance program (i.e. the original data set) were taken as true, i.e. as the result of a perfect diagnostic test. In actuality, this data was not actually obtained via a perfect diagnostic test. However, this decision was made for the purpose of comparison between data sets with varying degrees of misclassification. The disease, in this sense, is interchangeable; any other disease with similar prevalence would suffice with regard to this investigation. Having said that, the results here are less about WNv in southern Ontario and more about the effect of misclassification on spatial statistics for regional count data.

The second decision was that the sensitivity and specificity of the imperfect diagnostic tests used to produce the simulated data were considered as fixed proportions of observations that get misclassified. In practise, sensitivity and specificity are considered as expectations, so the proportions of misclassified observations would have some variation. This extra-variation may produce more uncertainty in the spatial statistical estimates. A simulation study similar to the one presented here, but which incorporated this characteristic of sensitivity and specificity, could be performed in order to investigate this.

Another decision was that misclassified cases and controls were chosen through simple random sampling. That is, true cases and controls in every region of the study

area had an equal probability of being misclassified. In this regard, any bias in statistical results may be attributed to non-differential misclassification. If interest were in investigating the effects of differential misclassification, however, a weighting scheme could be added to the simulations here, for which misclassified cases and controls were distributed among regions according to predefined weights.

It is also important to keep in mind the nature of the disease and study area before extrapolating from these results. For instance, this study may not be applicable to study regions which are less grid-like, as is the case for the southern Ontario region, or for regions in which edge effects are more of a factor. Furthermore, the prevalence of WNv here was approximately 25%. Applying the same method to a disease with less or greater overall prevalence may accentuate or mitigate the misclassification bias shown here.

Finally, an additional investigation into misclassification bias may be done by carrying out this simulation study with alternative spatial statistical tests; for clustering, the performance of Geary's C may be compared to Moran's I; for cluster detection, there may be interest in evaluating the performance of the Bayesian and Besag-Newell cluster detection methods; for modelling, if there is clustering in the data, any extension of GLMs to include random effects (be it in a Bayesian or frequentist context) may be used to account for spatial autocorrelation [2]. These alternative tests may provide further insight into the nature of diagnostic misclassification bias.

4.4 Conclusion

It is shown here that the use of imperfect diagnostic tests to assess the disease status of regional populations creates bias in the results for the spatial statistics used in this study. The statistics considered here were: the Moran's I coefficient to assess the amount of clustering in the data, the intercept of a Gaussian random field model to measure prevalence, the range and sill parameters of the semivariogram to characterize the spatial dependence structure of the data, and the Kulldorff cluster detection method to identify cluster locations.

The results from section 3.2.1 demonstrated that diagnostic misclassification will affect the results of the spatial statistics towards detecting more spatially random patterns, since the addition of false positives and negatives to the data will disrupt the spatial structure and hence add random noise. Furthermore, with reference to Berke and Waller (2010) it is expected that increasing sample size will mitigate the effect of misclassification bias. This was shown by the results in section 3.2.2 where the effect of diagnostic misclassification was reduced when regional sample sizes were doubled. Thus, caution should be used when interpreting results from data ascertained by low quality diagnostic tests, especially when regional sample sizes are low. Future simulation studies could be carried out, such as those dealing with rare disease, point data, or differential misclassification, in order to account for the effects of misclassification and to investigate ways to reduce bias.

Bibliography

[1] R. M. Assunção and E. A. Reis. A new proposal to adjust moran's i for population density. *Statistics in Medicine*, 18:2147–62, 1999.

[2] L. Beale, J. J. Abellan, S. Hodgson, and L. Jarup. Methodologic issues and approaches to spatial epidemiology. 116:1105–1110, 2008.

[3] O. Berke. Choropleth mapping of regional count data of echinococcus multilocularis among red foxes in lower saxony, germany. *Preventive Veterinary Medicine*, 52(2):119–131, 2001.

[4] O. Berke. Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(18), 2004.

[5] O. Berke and L. Waller. On the effect of diagnostic misclassification bias on the observed spatial pattern in regional count data - a case study using west nile virus mortality data from ontario, 2005. *Spatial and Spatio-temporal Epidemiology*, 1:117–122, 2010.

[6] H. Beroll, O. Berke, J. Wilson, and I. K. Barker. Investigating the spatial risk distribution of west nile virus disease in birds and humans in southern ontario from 2002 to 2005. *Population Health Metrics*, 3(5), 2007.

[7] J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1), 1991.

[8] Roger Bivand, with contributions by Micah Altman, Luc Anselin, Renato Assunção, Olaf Berke, Andrew Bernat, Guillaume Blanchet, Eric Blankmeyer, Marilia Carvalho, Bjarke Christensen, Yongwan Chun, Carsten Dormann, Stphane Dray, Rein Halbersma, Elias Krainski, Pierre Legendre, Nicholas Lewin-Koh, Hongfei Li, Jielai Ma, Giovanni Millo, Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Gianfranco Piras, Markus Reder, Michael Tiefelsdorf, and Danlin Yu. *spdep: Spatial dependence: weighting schemes, statistics and models*, 2012. R package version 0.5-53.

[9] Cici Chen, Albert Y. Kim, Michelle Ross, Jon Wakefield, and E. S. Venkatraman. *SpatialEpi: Performs various spatial epidemiological analyses*, 2012. R package version 1.0.

[10] Adam L. Cohen, Scott F. Dowell, Ananda Nisalak, Mammen P. Mammen Jr, Wimol Petkanchanapong, and Tamara L. Fisk. *Rapid diagnostic tests for dengue and leptospirosis: antibody detection is insensitive at presentation*, 12(1):47–51, 2009.

[11] N. Dendukuri, E. Rahme, P. Bélisle, and L. Joseph. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics*, 60:388–397, 2004.

[12] O. J. Devine, T. A. Louis, and M. E. Halloran. Empirical bayes methods for stabilizing incidence rates before mapping. *Epidemiology*, 5(6):622–630, 1994.

[13] R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146, 1954.

[14] B. Bert Gertsman. *Epidemiology Kept Simple: An Introduction to Traditional and Modern Epidemiology*. Wiley-Blackwell, San Jose, CA, 2013.

[15] G. M. Jacquez. Cluster morphology analysis. *Spat Spatio-temporal Epidemiology*, 1:19–29, 2009.

[16] Paulo J. Ribeiro Jr and Peter J. Diggle. *geoR: a package for geostatistical analysis*, 2001.

[17] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, (26):1481 – 1496, 1997.

[18] Marco Martuzzi and Paul Elliott. Empirical bayes estimation of small area prevalence of non-rare conditions. *Statistics in Medicine*, 15:1867–1873, 1996.

[19] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23, 1950.

[20] Dirk U. Pfeiffer, Timothy P. Robinson, Mark Stevenson, Kim B. Stevens, David J. Rogers, and Archie C. A. Clements Archie. *Spatial Analysis in Epidemiology*. Oxford Scholarship Online, New York, 2008.

[21] Z. Poljak, Catherine E. Dewey, S. Wayne Martin, Jette Christensen, Susy Carman, and Robert M. Friendship. Spatial clustering of swine influenza in ontario on the basis of herd-level disease status with different misclassification errors. *Preventive Veterinary Medicine*, 81:236–249, 2007.

[22] P. K. M. Quataert, B. Armstrong, A. Berghold, F. Bianchi, A. Kelly, and M. Marchi. Methodological problems and the role of statistics in cluster response studies: A framework. *European Journal of Epidemiology*, 15(9):821–831, 1999.

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[24] O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman & Hall, Boca Raton, Florida, 2005.

[25] J. Wakefield and A. Kim. A Bayesian model for cluster detection. *Biostatistics*, 2012.

[26] L. Waller and C. Gotway. *Applied spatial statistics for public health data*. Wiley, New York, 2004.

Appendix A

Supplemental Material

A.1 Semivariogram Estimation

A.1.1 Semivariogram Estimation for Section 3.4.1

This section provides the plots for the semivariograms based on data sets which were simulated from the original WNv data set with varying degrees of misclassification bias. For each combination of sensitivity and specificity, 3000 data sets were simulated. The σ^2 and ϕ parameter estimates from the spatial component of the Gaussian random field model were used to estimate the semivariogram.

It is shown here that, for fixed sensitivity, the semivariogram is systematically underestimated when specificity decreases. The same applies for fixed specificity and decreasing sensitivity, though the underestimation is not as severe.

Figure A.1: Semivariogram for sensitivity = 100% and specificity 95% - 80%

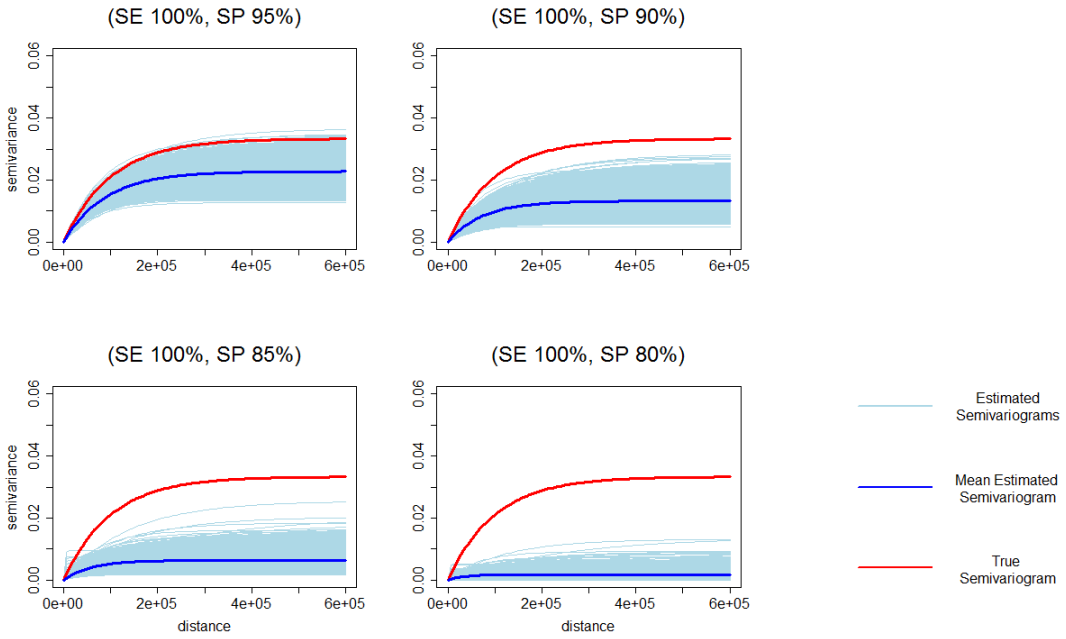


Figure A.2: Semivariogram for sensitivity = 95% and specificity 100% - 80%

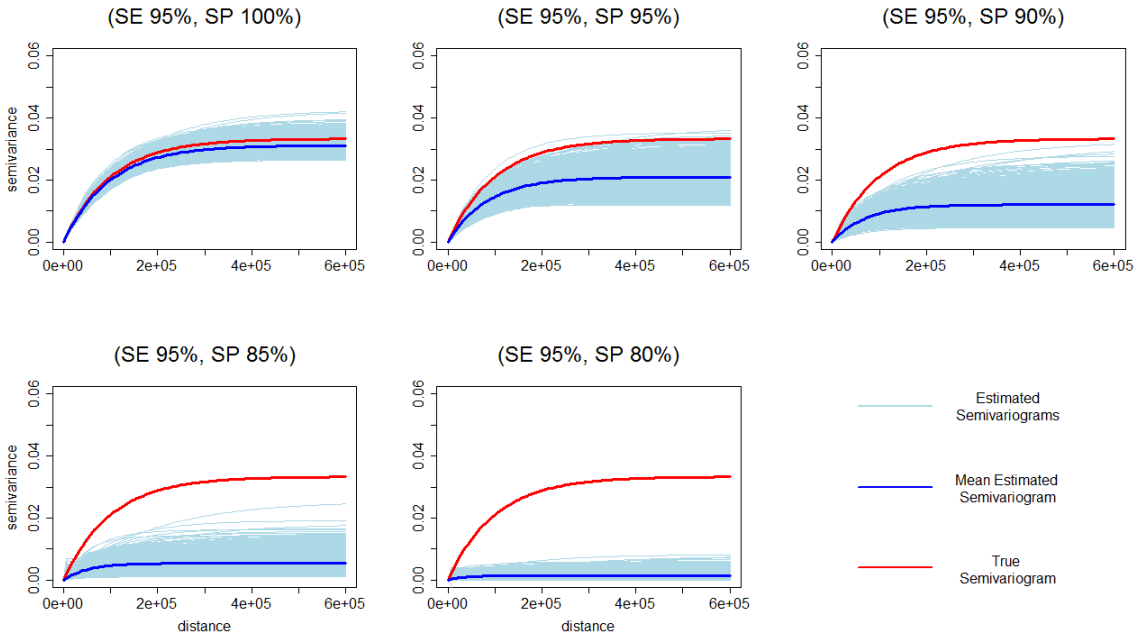


Figure A.3: Semivariogram for sensitivity = 90% and specificity 100% - 80%

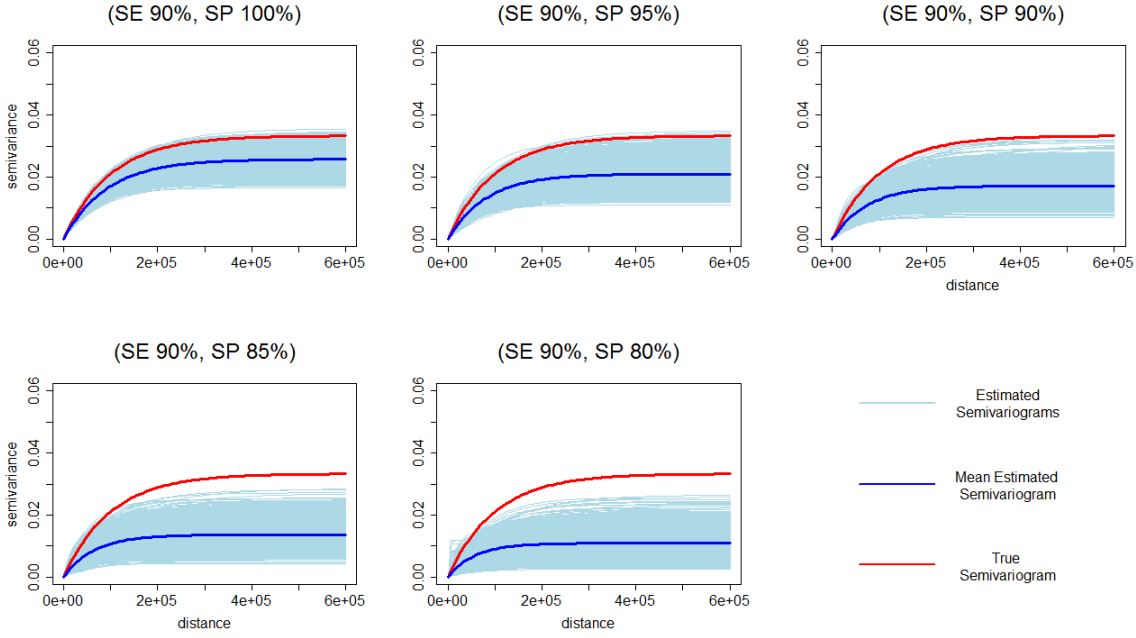


Figure A.4: Semivariogram for sensitivity = 85% and specificity 100% - 80%

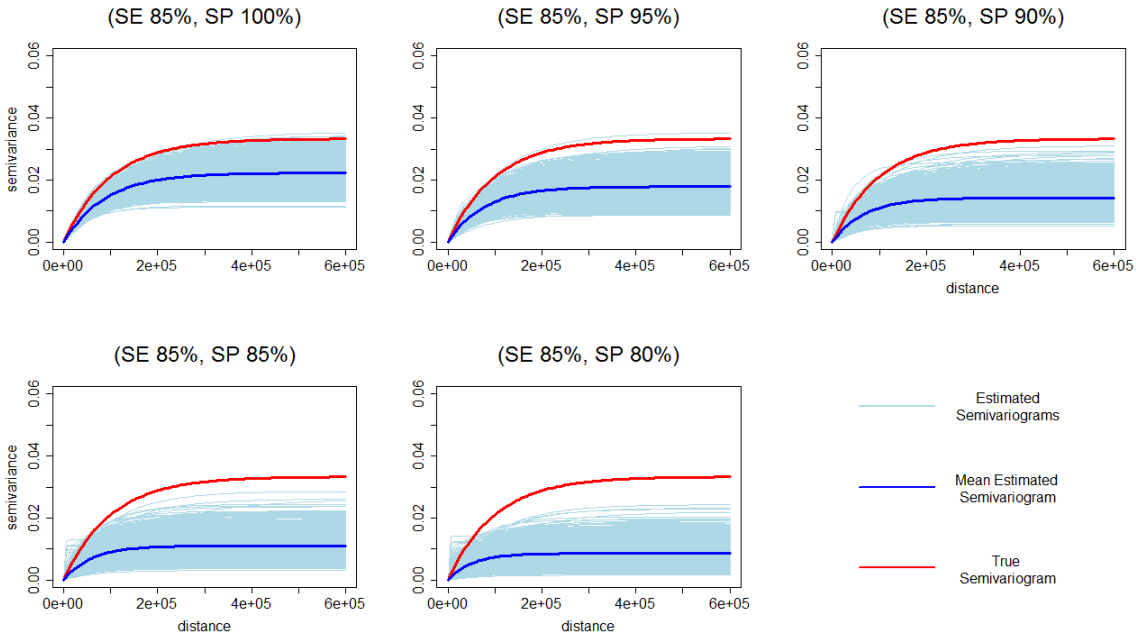
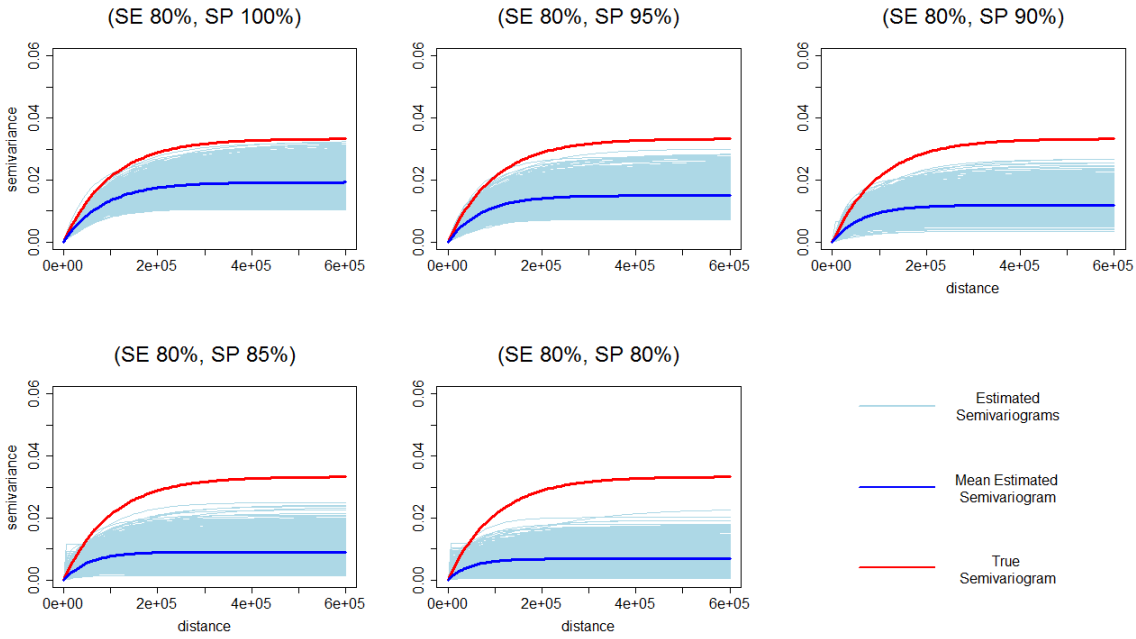


Figure A.5: Semivariogram for sensitivity = 80% and specificity 100% - 80%



A.1.2 Semivariogram Estimation for Section 3.4.2

This section provides the plots for the semivariograms based on data sets which were simulated from the WNv data set with doubled cases and controls with varying degrees of misclassification bias. For each combination of sensitivity and specificity, 3000 data sets were simulated. A Gaussian random field model was then fit to these data in order to obtain σ^2 and ϕ parameter estimates to estimate the semivariogram. With reference to section A.1.1, these plots illustrate that the effect of diagnostic misclassification is reduced when regional cases and controls are doubled.

Figure A.6: Semivariogram for sensitivity = 100% and specificity 95% - 80% (double regional case-controls)

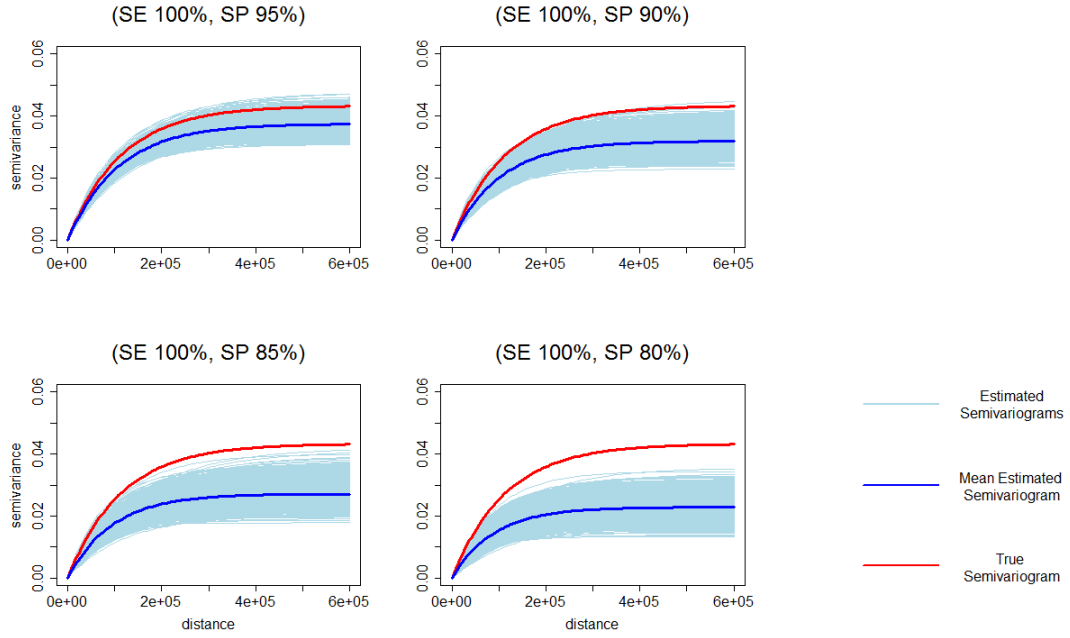


Figure A.7: Semivariogram for sensitivity = 95% and specificity 100% - 80% (double regional case-controls)

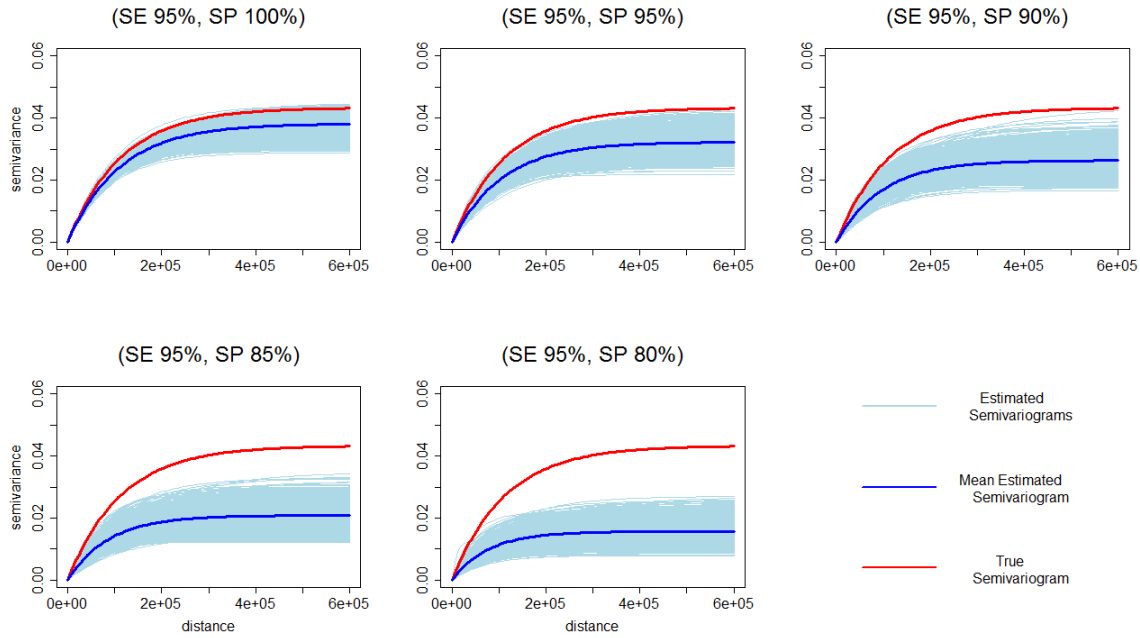


Figure A.8: Semivariogram for sensitivity = 90% and specificity 100% - 80% (double regional case-controls)

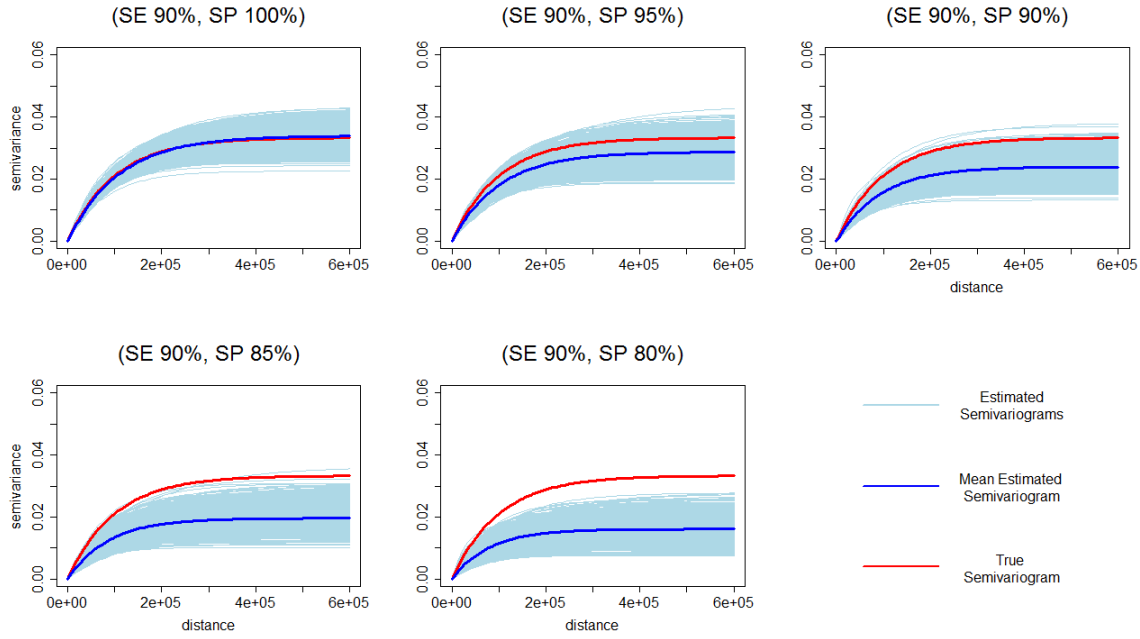


Figure A.9: Semivariogram for sensitivity = 85% and specificity 100% - 80% (double regional case-controls)

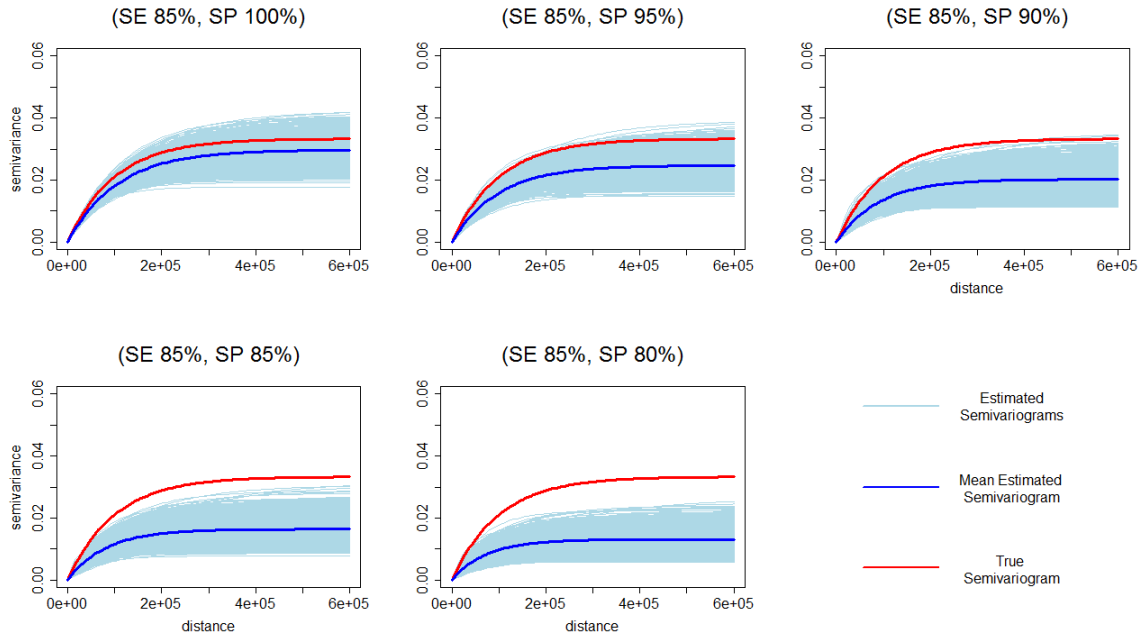
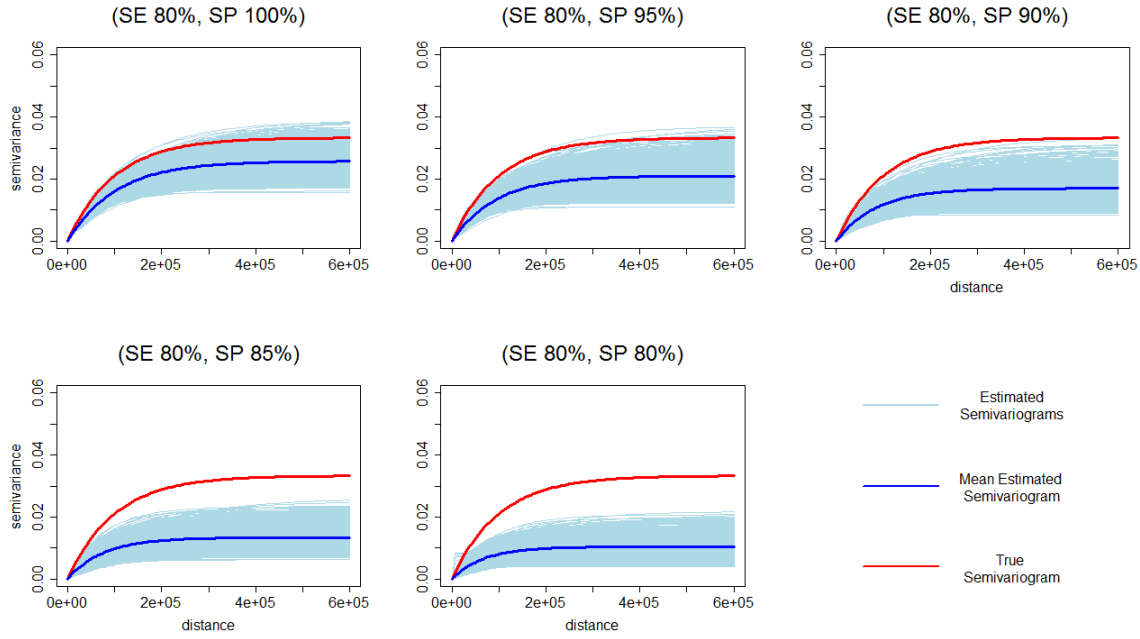


Figure A.10: Semivariogram for sensitivity = 80% and specificity 100% - 80% (double regional case-controls)



A.2 Choropleth Maps of Cluster Locations

A.2.1 Choropleth Maps of Cluster Locations for Section 3.5.1

This section provides the choropleth maps for the PHUs that were identified as part of a cluster location. These estimates are based on data sets which were simulated from the original WNv data set with varying degrees of misclassification bias. For each combination of sensitivity and specificity 3000 simulated data sets were generated. The Kulldorff spatial cluster detection was applied to these data sets in order to identify cluster locations. Any PHUs identified as part of a cluster location were then used in then plotted in a choropleth map corresponding to the diagnostic test that was used to simulate the data sets.

The maps here show that, for fixed sensitivity, the Kulldorff scan test begins to identify PHUs that do not belong to the true cluster location when specificity is decreased. Furthermore the percentage of simulations which identify the PHUs that belong to the true cluster locations also decreases for fixed sensitivity and decreasing specificity.

Figure A.11: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 100% and specificity 95% - 80%

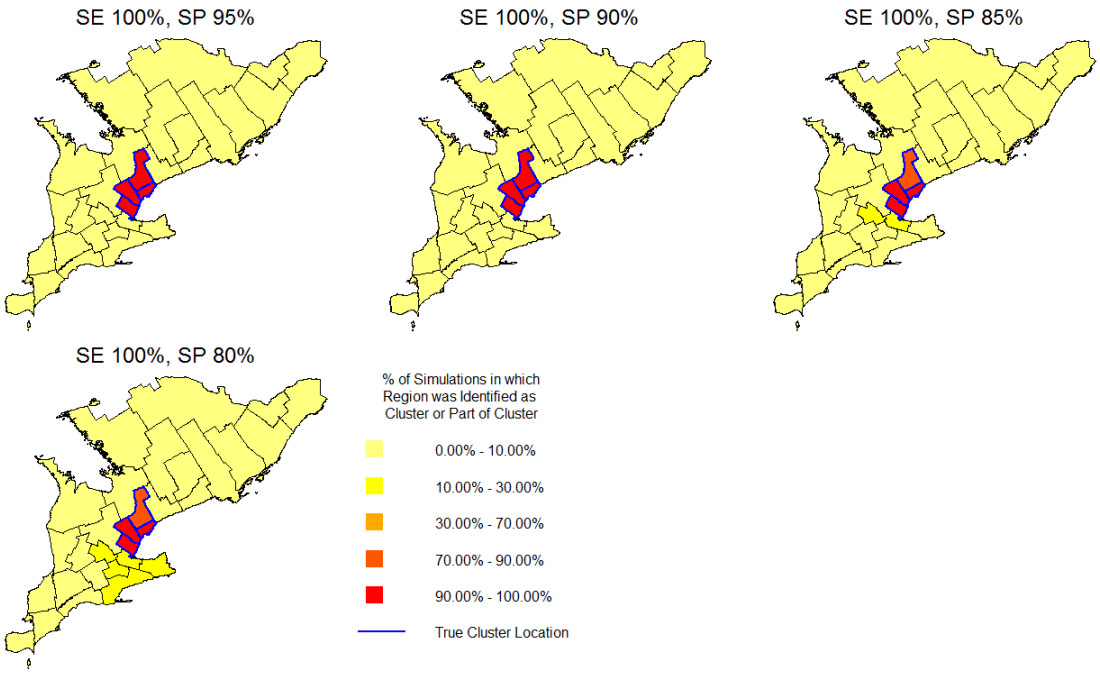


Figure A.12: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 95% and specificity 95% - 80%

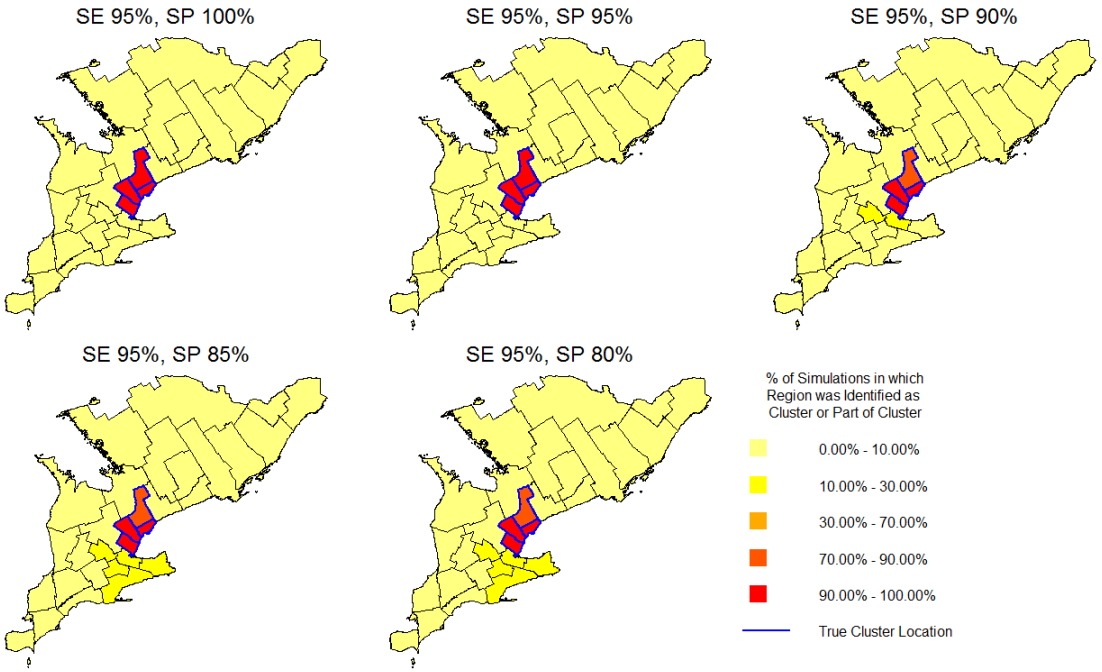


Figure A.13: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 90% and specificity 95% - 80%

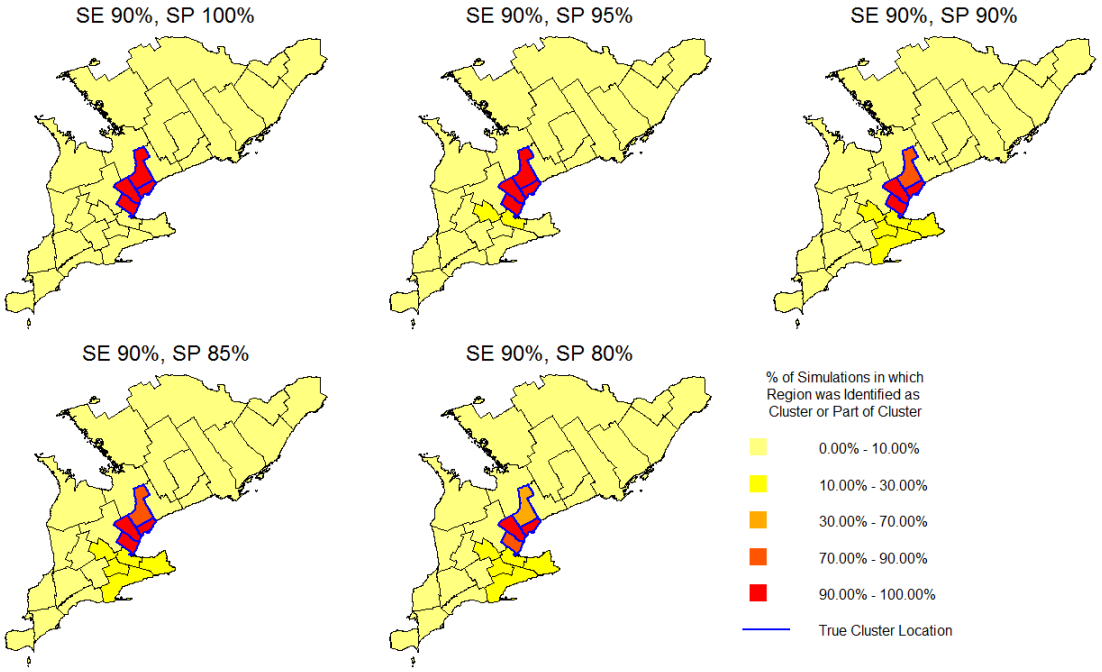


Figure A.14: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 85% and specificity 95% - 80%

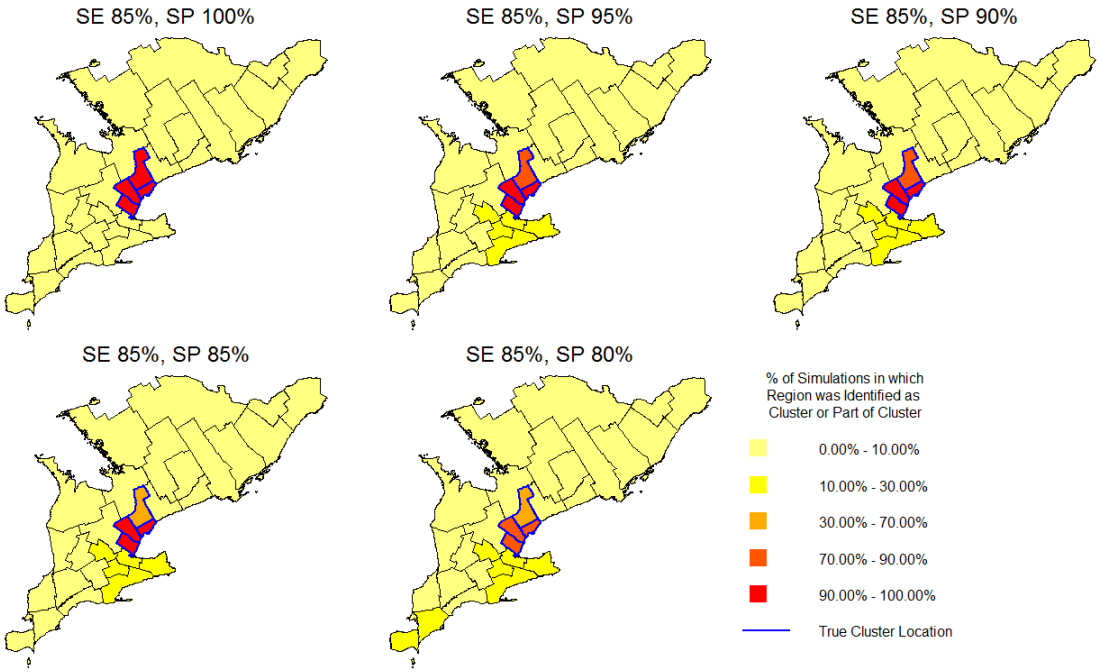
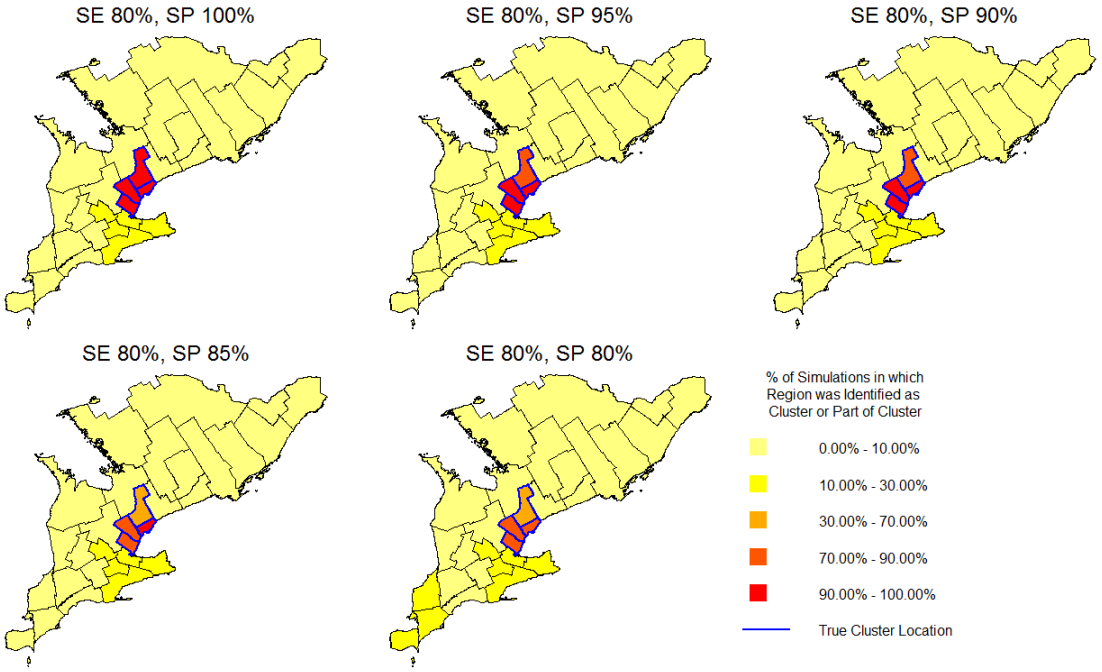


Figure A.15: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 80% and specificity 95% - 80%



A.2.2 Choropleth Maps of Cluster Locations for Section 3.5.2

This section provides the choropleth maps for the PHUs that were identified as part of a cluster location. These estimates are based on data sets which were simulated from the WNv data set with double regional cases and controls with varying degrees of misclassification bias. Again, 3000 simulated data sets were generated for each combination of sensitivity and specificity. In comparison to the results in section A.2.1, the maps here illustrate that the effect of diagnostic misclassification is reduced when regional cases and controls are doubled. However, there is still an effect; Kulldorff's scan test begins to misidentify cluster locations, and the percentage of simulations in which the scan test identifies the PHUs belonging to the true cluster location decreases as specificity decreases (fixed sensitivity).

Figure A.16: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 100% and specificity 95% - 80% (double regional case-controls)

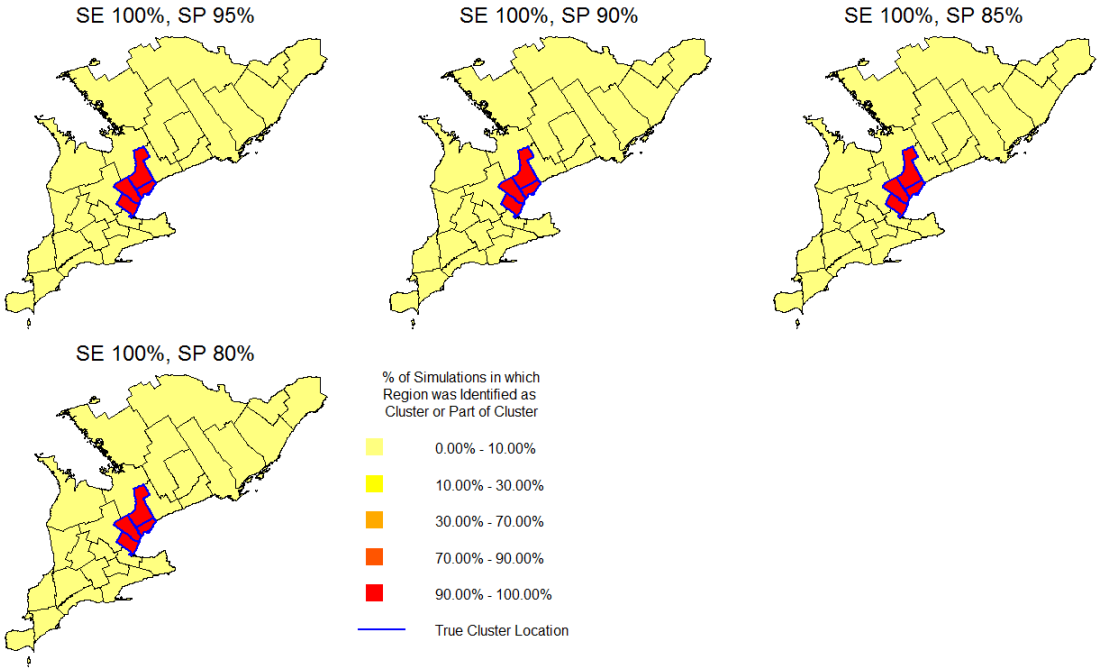


Figure A.17: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 95% and specificity 95% - 80% (double regional case-controls)

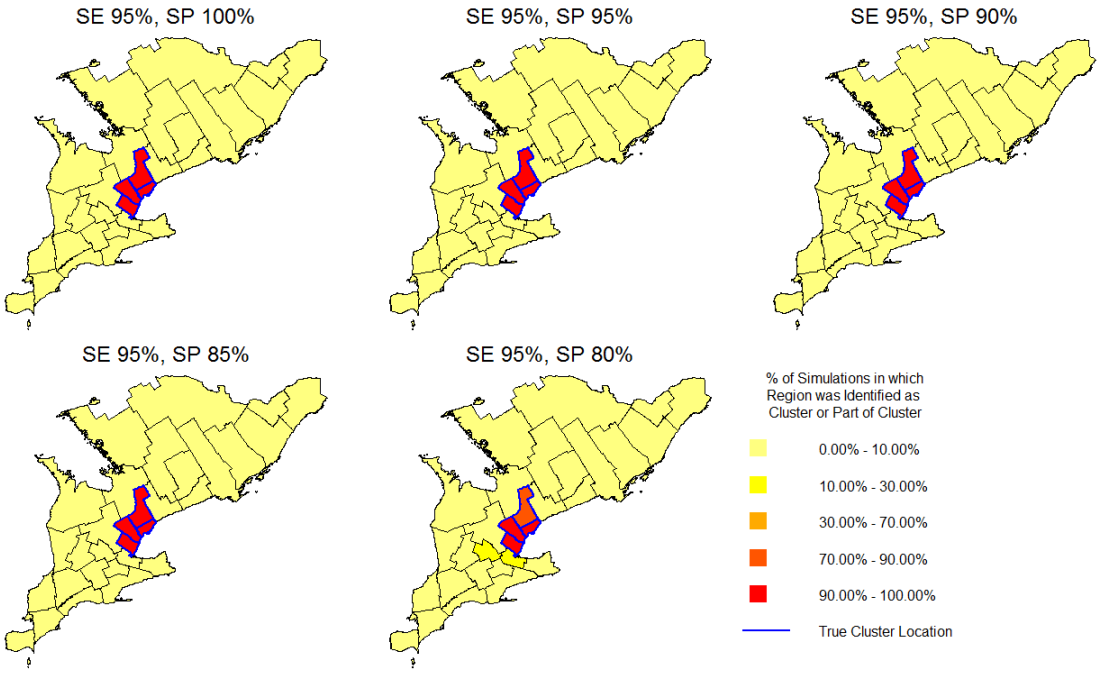


Figure A.18: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 90% and specificity 95% - 80% (double regional case-controls)

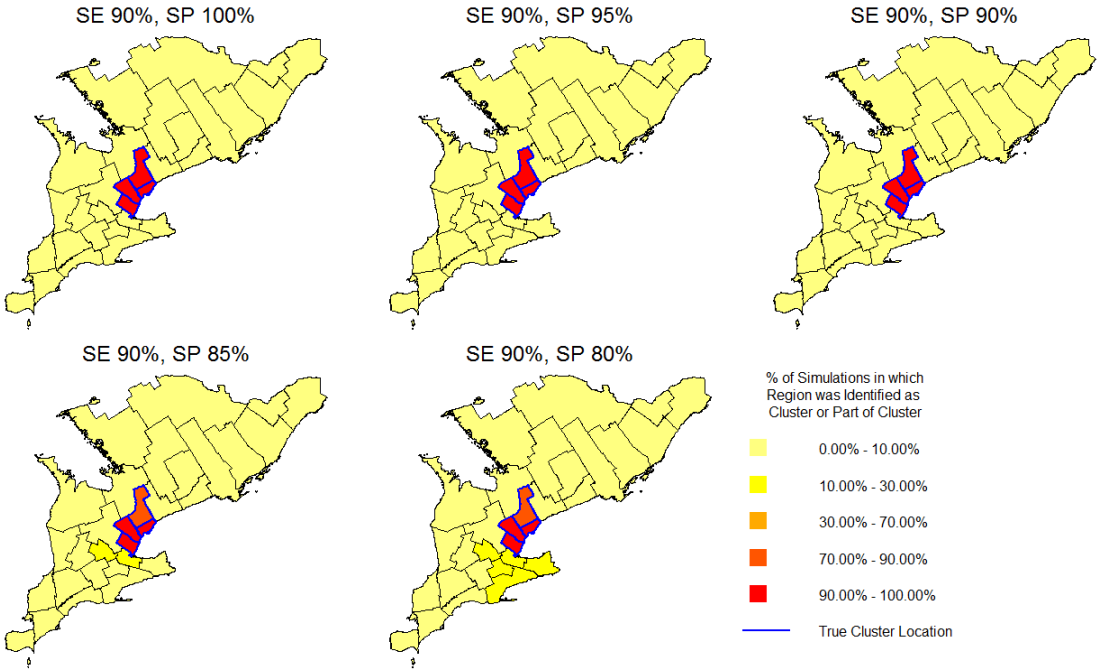


Figure A.19: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 85% and specificity 95% - 80% (double regional case-controls)

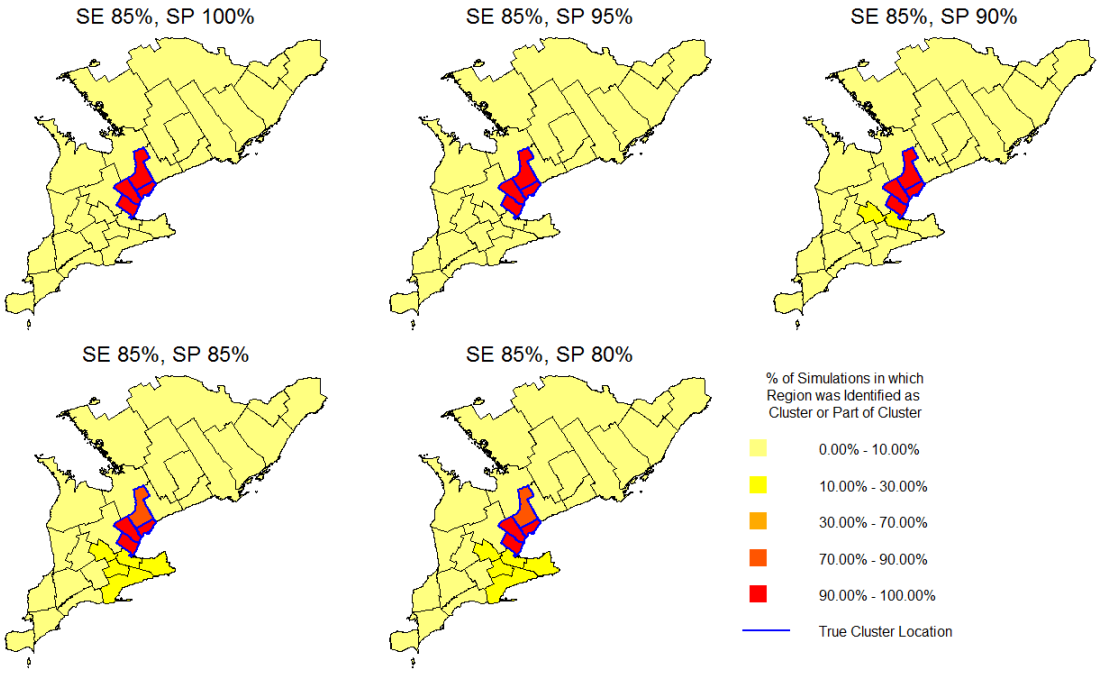


Figure A.20: Choropleth map for PHUs identified as a cluster or part of a cluster location with sensitivity = 80% and specificity 95% - 80% (double regional case-controls)