ISPRS Int. J. Geo-Inf. **2014**, *3*, 1039-1057; doi:10.3390/ijgi3031039

OPEN ACCESS

ISPRS International Journal of

Geo-Information

ISSN 2220-9964

www.mdpi.com/journal/ijgi/

Article

Geographical Variation of Incidence of Chronic Obstructive Pulmonary Disease in Manitoba, Canada

Mahmoud Torabi * and Katie Galloway

Department of Community Health Sciences, University of Manitoba, 750 Bannatyne Ave., Winnipeg, MB R3E 0W3, Canada; E-Mail: gallowa3@cc.umanitoba.ca

* Author to whom correspondence should be addressed;
E-Mail: Mahmoud.Torabi@med.umanitoba.ca; Tel.: +1-204-272-3136; Fax: +1-204-789-3905.

Received: 3 March 2014; in revised form: 15 July 2014 / Accepted: 21 July 2014 / Published: 29 July 2014

Abstract: We aimed to study the geographic variation in the incidence of COPD. We used health survey data (weighted to the population level) to identify 56,944 cases of COPD in Manitoba, Canada from 2001 to 2010. We used five cluster detection procedures, circular spatial scan statistic (CSS), flexible spatial scan statistic (FSS), Bayesian disease mapping (BYM), maximum likelihood estimation (MLE), and local indicator of spatial association (LISA). Our results showed that there are some regions in southern Manitoba that are potential clusters of COPD cases. The FSS method identified more regions than the CSS and LISA methods and the BYM and MLE methods identified similar regions as potential clusters. Most of the regions identified by the MLE and BYM methods were also identified by the FSS method and most of the regions identified by the CSS method were also identified by most of the other methods. The CSS, FSS and LISA methods identify potential clusters but are not able to control for confounders at the same time. However, the BYM and MLE methods can simultaneously identify potential clusters and control for possible confounders. Overall, we recommend using the BYM and MLE methods for cluster detection in areas with similar population and structure of regions as those in Manitoba.

Keywords: bayesian computation; chronic obstructive pulmonary disease; geographic epidemiology; prediction; random effects; spatial cluster detection

---

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a lung disease defined by continuous airflow limitation caused by small airway disease (obstructive bronchiolitis) and parenchymal destruction (emphysema). The small airways narrow in response to chronic inflammation. As well, inflammatory processes cause the deterioration of the lung parenchyma, which leads to a decrease in the elastic recoil of the lung. As a result of these changes, the airways have a decreased ability to remain open during expiration [1]. The biggest and most widely known risk factor of COPD is cigarette smoking [2]. Other risk factors of COPD include occupational or environmental exposure to dust and hazardous gases, for example when burning biomass fuel [3]. A family history (*i.e.*, genetics), low socioeconomic status, poor nutrition, asthma, and recurrent lung infections can also be risk factors for COPD [1,4]. Therefore, COPD can be the result of a gene-environment interaction [1].

The impact of COPD is often underestimated by health authorities and government officials [5]. In Canada, one of the most overlooked chronic conditions is COPD. Patients suffering from a degenerative lung disease are often misdiagnosed as having bronchitis, a cough or a respiratory tract infection [6]. In 2008, COPD was the leading cause of hospitalizations in Canada. As well, 18% of COPD patients were readmitted to a hospital once within the year and 14% were re-admitted twice within the year. These readmission rates were higher than any other chronic illnesses [6,7]. According to a Canadian article [8], for severe COPD exacerbations or attacks, the average length of a hospital visit was 10 days with an estimated cost of $10,000. Within a single year, the estimated cost of moderate and severe COPD exacerbations exceeds $730 million. This number is expected to nearly double by 2015 [8].

There are various treatments for COPD including antibiotics and chest physiotherapy. However, early detection of COPD is crucial for a positive outcome [9]. Therefore, it is important to identify trends in COPD incidence that may suggest further epidemiological studies to identify risk factors and identify any changes in important factors. Trends may occur over a region and the focus of our paper is to examine geographical variation in the number of people diagnosed as having COPD during 2001 to 2010 in the province of Manitoba, Canada.

A spatial cluster is defined as a limited area within the entire study region which has a high proportion of disease cases [10]. Possible factors related to diseases may be determined by discovering disease clusters which may lead to an improved understanding of etiology. In fact, the identification of clusters may lead to further analyses to study how exposures and disease interventions are connected [11].

Spatial cluster detection methods can be classified into two statistical approaches, a focused approach or a non-focused (general) approach. The methodology of focused cluster detection approaches is to locate regions with an excess number of disease cases in an area near a possible cause (*i.e.*, a toxic waste site) [12,13]. On the other hand, non-focused cluster detection methods typically use various ways in order to discover areas with a high number of disease cases in the entire study region [14–16]. The circular spatial scan statistic (CSS) [17], flexible spatial scan statistic (FSS) [18], and Bayesian disease mapping (BYM) [14] are all considered to be focused cluster detection methods, whereas, the Besag and Newell (BN) [19,20] test and the maximizing excess event test (MEET) [21] are classified as non-focused cluster detection procedures. Non-focused tests are used to detect potential clusters in the study area, while focused tests are used to test the null hypothesis of no spatial cluster against the alternative hypothesis that a spatial cluster exists. In other words, the purpose of

focused tests (CSS, FSS, BYM) is to find possible clusters in an area of interest and the aim of non-focused tests is to discover any significant cluster without determining a specific area of interest. These approaches were compared by analyzing childhood cancer data in the province of Alberta, Canada [22]. Recently, a frequentist approach based on the maximum likelihood estimation (MLE), via data cloning (DC) [23,24], was also proposed to obtain possible clusters [25] in an area of interest. Another cluster detection method is the local indicator of spatial association (LISA) [26]. This method is simple and easy to implement.

This paper is based on the focused cluster detection methods. In particular, the aforementioned focused approaches (CSS, FSS, BYM, MLE, and LISA) are used to analyze a real dataset of COPD cases in the province of Manitoba, Canada, from 2001 to 2010.

2. Methods

2.1. Study Subjects

This study was based on the Canadian Community Health Survey (CCHS) [27] from Statistics Canada. The CCHS is a cross-sectional survey, which gathers information from the Canadian population regarding health status, health care utilization and health determinants. The CCHS collects health related data from individuals aged twelve and older in order to provide reliable estimates at the health region level [27]. The information from the CCHS used in this study was the number of COPD cases in the province of Manitoba, Canada, from 2001 to 2010. Eleven Regional Health Authorities, which are further divided into 67 Regional Health Authority Districts (RHADs) are in charge of delivering health care services to individuals in Manitoba. The RHADs are the geographic units used in our models and all of the data used in the study are related to these RHADs which are labeled 1, 2, …, 67 for simplicity. As well, a population-based centroid was provided for each RHAD, however, these centroids were not necessarily geographic centres. Since the data used in the study was from a survey, appropriate weights (see Section 2.2 for more details) established by Statistics Canada [27] were applied to the data, which was then aggregated over the study period from 2001 to 2010.

The population was stable in Manitoba from approximately 1.15 million people in 2001 to 1.20 million people in 2010. Region 38 had the smallest average population size of 920 people while region 62 had the largest average population size of 91,633 people. The mean and median population sizes across the regions were 17,471 and 9466, respectively. The total number of COPD cases in Manitoba was 56,944 with a mean of 850 and median of 504 cases. These observations are based on the weighted results of COPD cases across the 67 regions.

The observed number of COPD cases and the expected number of COPD cases as well as the population size of each region are important requirements for focused spatial cluster detection methods. Adjustments may be made when the expected number of cases varies by different factors such as year, age, and gender. The expected number of disease cases was then adjusted by year (1–10), age group ((0–5), (6–20), (21–40), (41, 88), (89+)) and gender (male, female). A review of the CSS, FSS, BYM, MLE, and LISA spatial cluster detection methods is given in the Appendix.

The five focused spatial cluster detection procedures (CSS, FSS, BYM, MLE, and LISA) have different assumptions. Although the CSS, FSS, and LISA approaches are distribution free,

it is assumed that the number of disease cases follow a Poisson distribution in the BYM and MLE methods. As well, while the number of regions to be included in the cluster must be specified for the CSS and FSS methods, this is not a requirement for the BYM and MLE approaches. For the model-based cluster identification methods (BYM and MLE), if the model does not fit the data well, the result can be misleading. So, the *deviance residual* [28] should also be checked. While the expected number of disease cases or the population of each region is required for the above methods, they are not a requirement for the LISA method.

The University of Manitoba's Research Data Centre approved the study, and Statistics Canada approved administrative data access. ArcGIS version 10.0 (Environmental Systems Research Institute, Redlands, CA, USA) was used to produce choropleth maps of risks.

2.2. Weighting Process

The weighting was completed by Statistics Canada using a detailed weighting process [27]. A brief summary of this procedure is given here. First, the weighting depends on the sampling method (area frame *vs.* telephone frame) used in each region. In the area frame an initial weight is assigned based on the Labour Force Survey (LFS). Out-of-scope units (*i.e.*, Dwellings that are under construction, vacant, seasonal or secondary and institutions) are removed from the sample. As well, sub-clusters (*i.e.*, Sub-sampling within a selected dwelling), larger sample sizes and non-response units are adjusted for in the weighting process. In the telephone frame (the survey is conducted by telephone) an initial weight is assigned as the probability that phone number will be selected, which depends on the number of units sampled and the number of units available to be sampled. In this method, samples are drawn every two months therefore, an adjustment factor is applied to reduce the weights of each two-month sample so that the total sample is representative of the population only once. Similar to the area frame method, out-of-scope numbers (*i.e.*, Businesses, institutions, out-of-scope dwellings or numbers that are not in service) are removed from the sample. Also, non-response units and dwellings with multiple phone numbers are adjusted for in the weighting process [27].

The weights common to the area frames and telephone frames need to be integrated using an adjustment factor α ($0 < \alpha < 1$). Then a person-level weight is created by taking the inverse of the probability a person in the selected dwelling will be selected, which depends on the number of people in the household and the ages of those people. After the appropriate adjustment is made, a "winsorization" trimming method is used to decrease any extreme weights that occur. Finally, a calibration approach is used to ensure the weights are representative of the population estimates for the different age groups and genders in each health region [27].

2.3. Specific Hypotheses

We specify the alternative hypotheses for the CSS, FSS, BYM, and MLE approaches. We consider multiple alternatives that are tested separately. Further, let RR_i indicate the relative risk for the i-th region within a cluster when compared with the region outside a cluster; the latter has $RR_i = 1$. For example for cluster X, the RR_i is given by

$$RR_i = \begin{cases} 3 & i \in X \\ 1 & \text{otherwise.} \end{cases}$$

3. Results

The results of five different cluster detection techniques when applied to a COPD dataset in the province of Manitoba, Canada, from 2001 to 2010 are shown and compared in this section.

Based on the 67 regions, four different clusters were tested: (1) a case of no clusters (called A); (2) seven regions from the northern part of the province (called B); (3) seven regions from south-central part of the province (called C); and (4) 12 regions which comprise the Winnipeg region (called D). For A, no region was specified as a potential cluster. Moreover, the regions belonging to clusters B, C, and D are: B = {31, 33, 34, 36, 38, 40, 41}, C = {27, 28, 29, 30, 50, 51, 52}, and D = {56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67}. Since the LISA method does not depend on the expected number of observations, it could only be applied to cluster A as the other clusters require the adjustment of the expected number of disease cases for those regions inside the specified cluster.

The areas that are statistically significant (potential clusters) are shown for each cluster and each method separately (Figures 1–4). The summary of cluster A, no region specified as a potential cluster, is presented in Table 1. For the CSS and FSS procedures, the regions that are most likely to constitute a disease cluster are presented, as well as the regions that are second and third most likely to be considered as a cluster. For the BYM and MLE methods, a region is considered (and ranked) to be a significant cluster if the lower limit of the credible/prediction interval follows the specified criteria. For example, in the BYM method region 10 is most likely to be considered as a cluster and region 61 is least likely to be considered as a cluster under the criteria that the lower bound of RR is greater than one. For the LISA method, a region is determined (and ranked) to be significant if the p-value is less than 0.1.

The FSS method identified more regions as potential clusters than the CSS approach for cluster A, although, the regions with potential clusters that were detected by the CSS method were also identified by the FSS approach. The CSS approach detected regions {10, 43, 45, 61, 62} as potential clusters, and the FSS method identified the same regions as the CSS method, as well as regions {1, 11, 12, 13, 14, 20, 21, 27, 46, 50, 51, 54, 56, 60, 64, 65}. The BYM and MLE methods identified regions {1, 3, 6, 10, 11, 12, 20, 21, 24, 27, 43, 45, 50, 54, 61, 62, 64, 65} as possible clusters. The only difference between the results of these two procedures was the order of significance for the potential clusters. As well, most of the regions identified using these two approaches were also identified by the FSS approach and the regions identified by the CSS technique were also detected by the BYM and MLE approaches. Note that by evaluating the criterion of the RR values from greater than 1 to 1.5 or even 2, the number of potential clusters decreases (Table 1). Based on the deviance residual plots for both the BYM and MLE methods, we found that there is no serious lack of fit in the model. The LISA method found regions {2, 7, 16, 24, 43, 56, 57, 58, 60, 62, 64, 67} to be possible clusters of COPD. This approach identified some different regions to be potential clusters as compared to the other methods.

For the case of cluster B, none of the methods were able to detect all the regions in cluster B as a potential cluster. However, the CSS method identified regions 10, 43, and 62 as a potential cluster while the FSS method detected the same regions as the CSS method in addition to regions {11, 12, 13, 14, 20, 21, 27, 31, 45, 46, 50, 51, 54}. The BYM approach could identify regions {1, 3, 6, 10, 11, 12, 20, 21, 24, 27, 31, 43, 45, 50, 54, 61, 62, 64, 65} as potential clusters. The MLE method was also able to identify the same regions as the BYM method in addition to region 19.

Figure 1. The order of most likely clusters of COPD for the CSS, FSS, and LISA (based on the *p*-value) methods, and the special effects of the regional COPD risks for the BYM and MLE methods; in the case of cluster A. Major urban centre (Winnipeg region) is incorporated as an inset. (**a**) CSS; (**b**) FSS; (**c**) BYM; (**d**) MLE; (**e**) LISA.

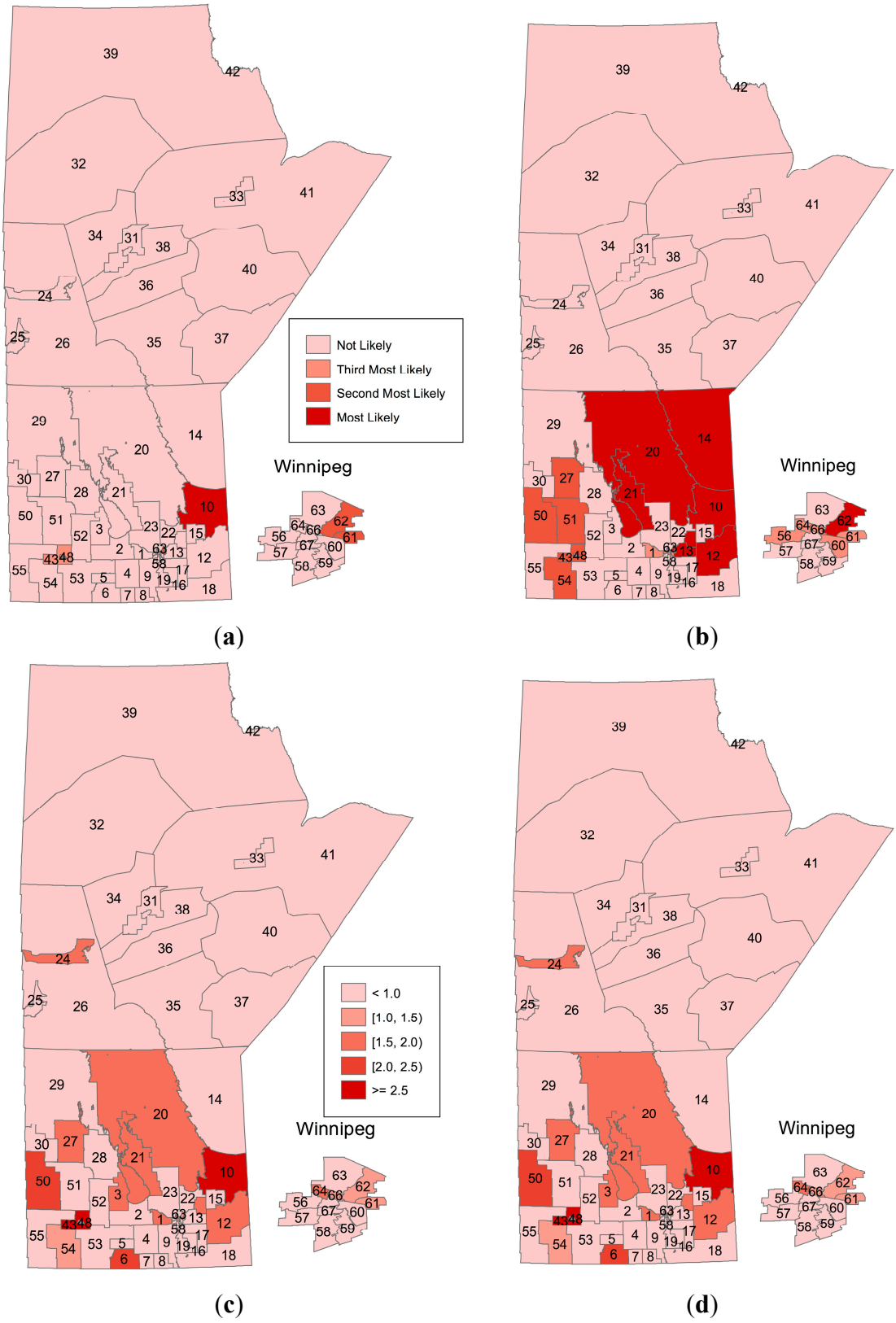
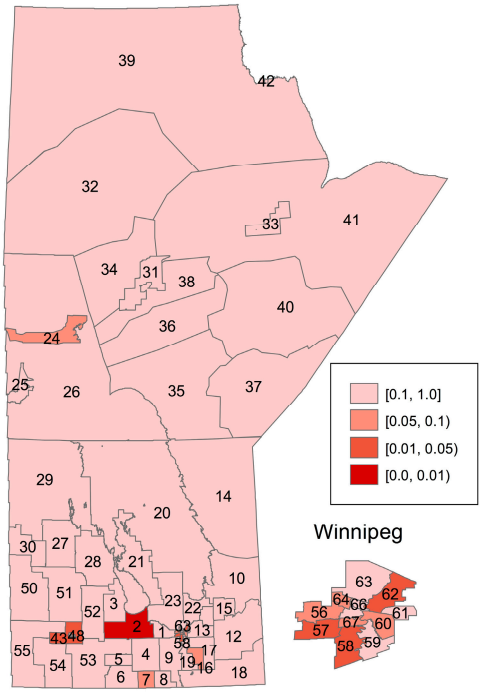
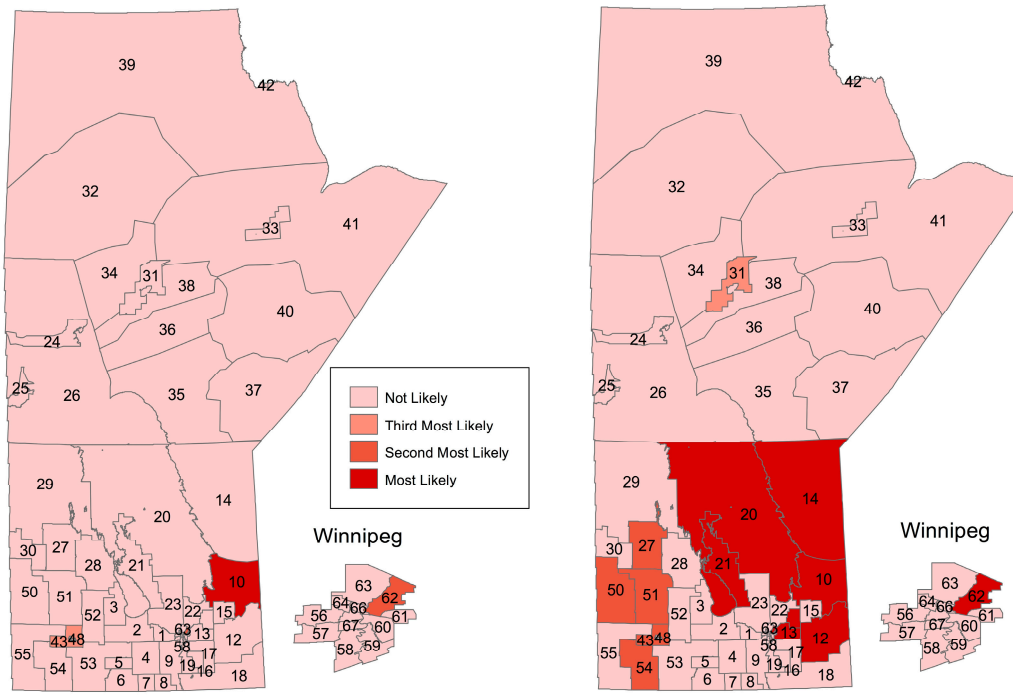


Figure 1. *Cont.*



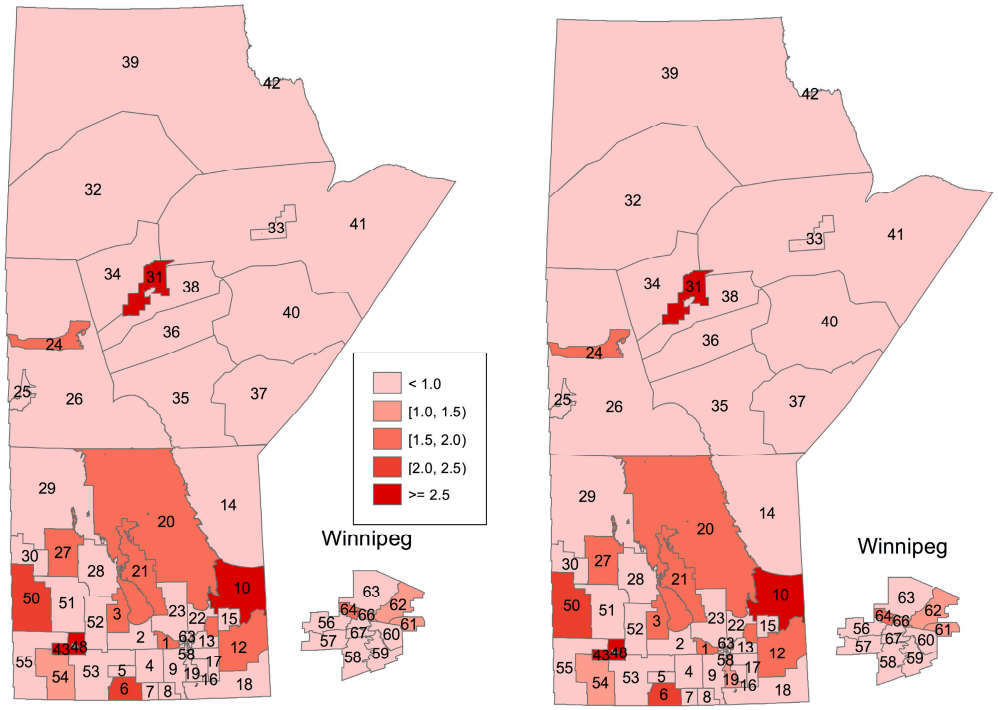
(e)

Figure 2. The order of most likely clusters of COPD for the CSS and FSS methods, and the special effects of the regional COPD risks for the BYM and MLE methods; in the case of cluster B. Major urban centre (Winnipeg region) is incorporated as an inset. (**a**) CSS; (**b**) FSS; (**c**) BYM; (**d**) MLE.



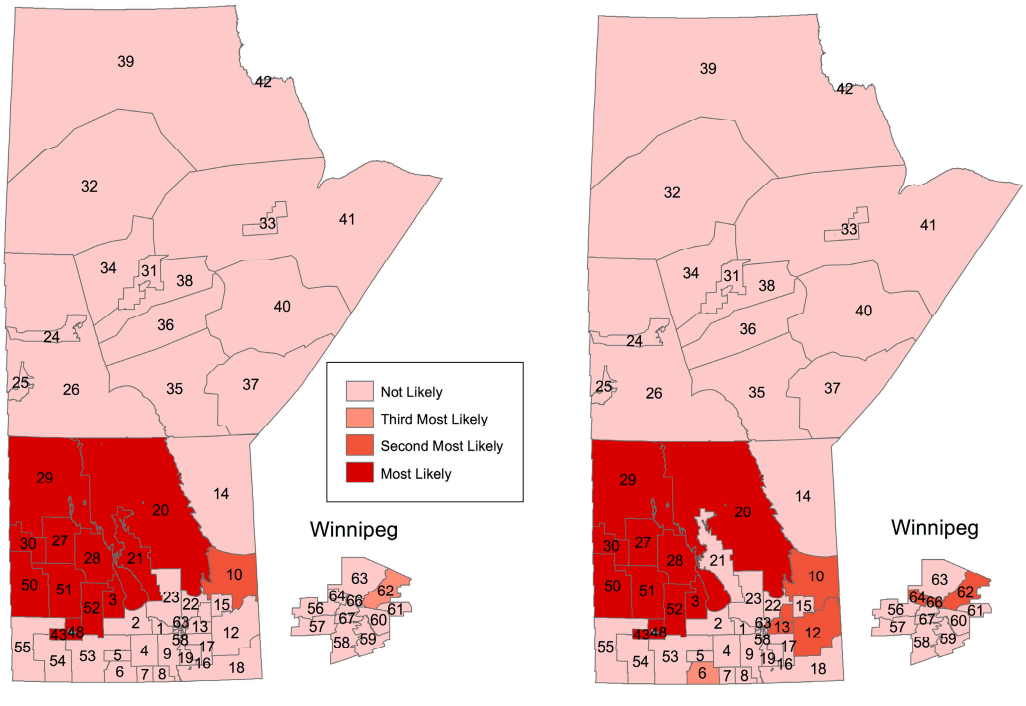
(a) **(b)**

Figure 2. *Cont.*



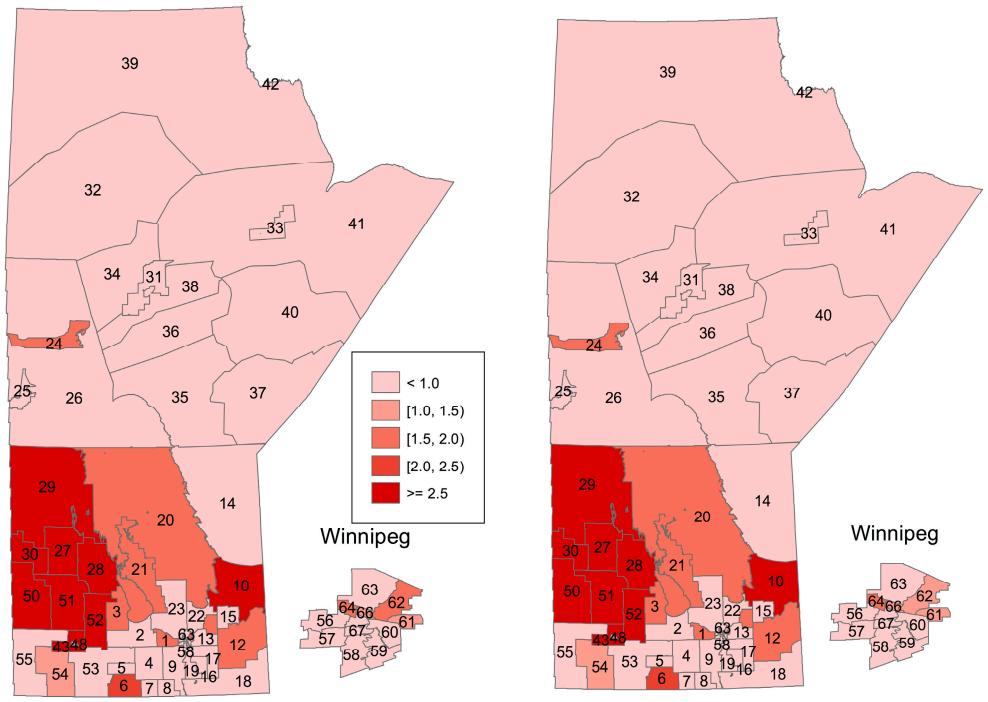
(c) **(d)**

Figure 3. The order of most likely clusters of COPD for the CSS and FSS methods, and the special effects of the regional COPD risks for the BYM and MLE methods; in the case of cluster C. Major urban centre (Winnipeg region) is incorporated as an inset. (**a**) CSS; (**b**) FSS; (**c**) BYM; (**d**) MLE.



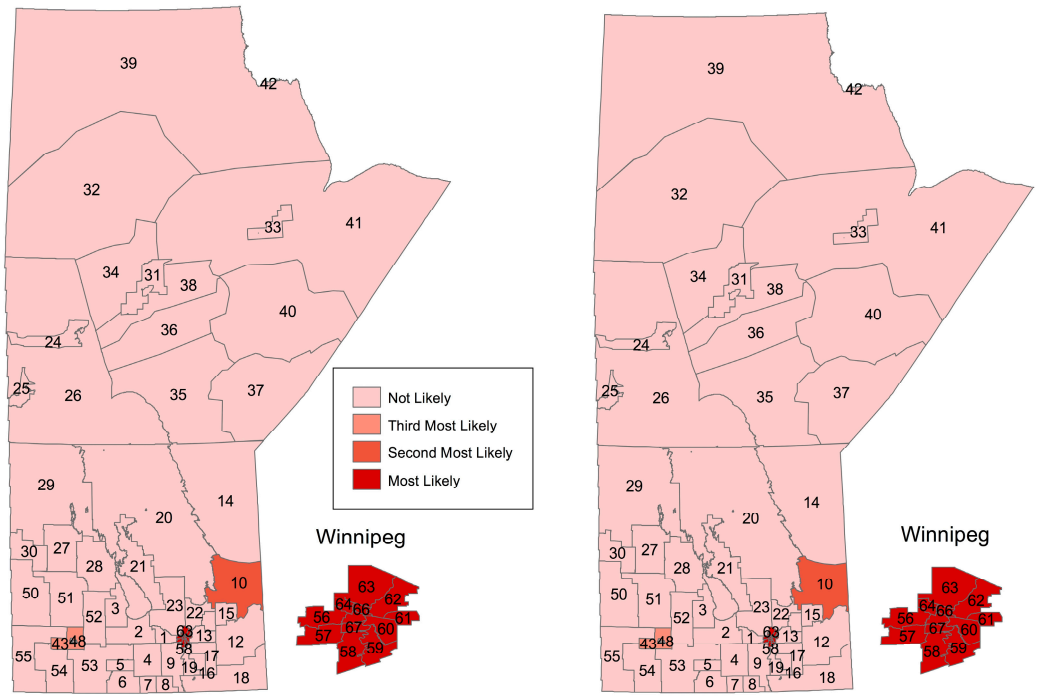
(a) **(b)**

Figure 3. *Cont.*



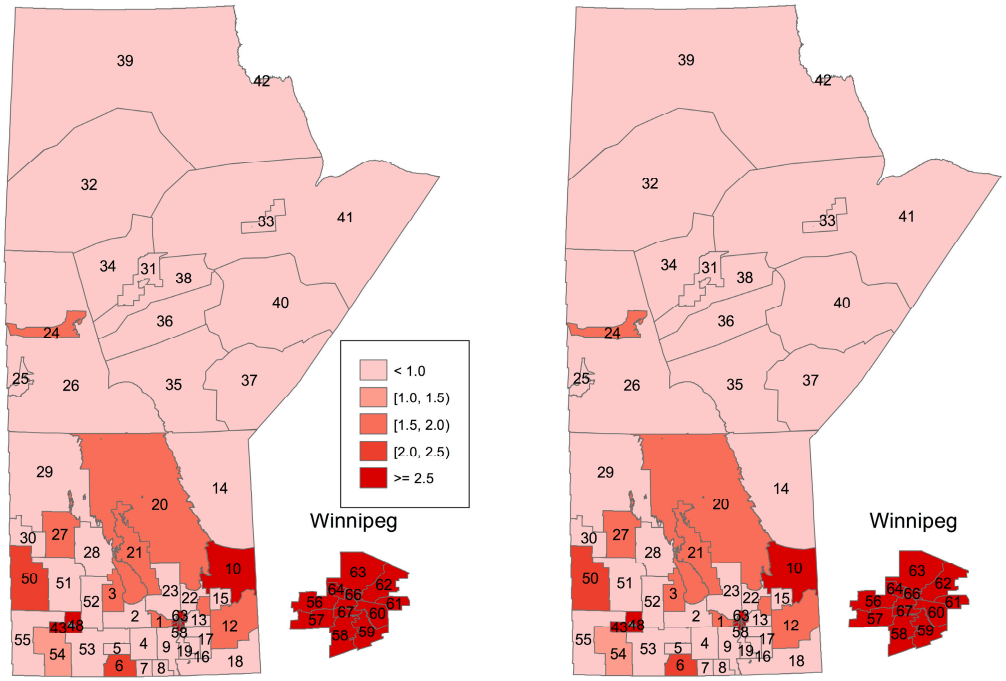
(c) **(d)**

Figure 4. The order of most likely clusters of COPD for the CSS and FSS methods, and the special effects of the regional COPD risks for the BYM and MLE methods; in the case of cluster D. Major urban centre (Winnipeg region) is incorporated as an inset. (**a**) CSS; (**b**) FSS; (**c**) BYM; (**d**) MLE.



(a) **(b)**

Figure 4. *Cont.*



(c) **(d)**

Table 1. The order of significant regions for the LISA, CSS, FSS, BYM, and MLE methods for cluster A.

Region	Method								
	LISA	CSS	FSS	RR > 1.0		RR > 1.5		RR > 2.0	
				BYM	MLE	BYM	MLE	BYM	MLE
1	-	-	3	13	11	-	-	-	-
2	1	-	-	-	-	-	-	-	-
3	-	-	-	11	13	-	-	-	-
6	-	-	-	5	5	5	5	-	-
7	12	-	-	-	-	-	-	-	-
10	-	1	1	1	1	1	1	1	1
11	-	-	1	15	15	-	-	-	-
12	-	-	1	10	10	-	-	-	-
13	-	-	1	-	-	-	-	-	-
14	-	-	1	-	-	-	-	-	-
16	9	-	-	-	-	-	-	-	-
20	-	-	1	4	4	4	4	-	-
21	-	-	1	14	14	-	-	-	-
24	7	-	-	6	6	6	6	-	-
27	-	-	2	12	12	-	-	-	-
43	2	3	2	2	2	2	2	2	2
45	-	3	2	7	8	-	-	-	-
46	-	-	2	-	-	-	-	-	-
50	-	-	2	3	3	3	3	-	-

Table 1. *Cont.*

Region	Method								
	LISA	CSS	FSS	RR > 1.0		RR > 1.5		RR > 2.0	
				BYM	MLE	BYM	MLE	BYM	MLE
51	-	-	2	-	-	-	-	-	-
54	-	-	2	17	17	-	-	-	-
56	6	-	3	-	-	-	-	-	-
57	4	-	-	-	-	-	-	-	-
58	5	-	-	-	-	-	-	-	-
60	8	-	3	-	-	-	-	-	-
61	-	2	3	18	18	-	-	-	-
62	3	2	1	8	7	-	-	-	-
64	11	-	3	9	9	-	-	-	-
65	-	-	3	16	16	-	-	-	-
67	10	-	-	-	-	-	-	-	-

For cluster C, all four methods were able to detect all the regions of cluster C as a potential cluster. Furthermore, the CSS method also identified regions {3, 10, 20, 21, 43, 45, 49, 62} as potential clusters while the FSS method detected those regions identified by the CSS method in addition to regions {6, 11, 12, 13, 64, 65}. Both the BYM and MLE methods identified regions {1, 3, 6, 10, 11, 12, 20, 21, 24, 43, 45, 54, 61, 62, 64, 65} in addition to the regions in cluster C as potential clusters.

For cluster D, all four methods detected the regions belonging to the D cluster as a potential cluster. In addition to the regions in Winnipeg (cluster D), the BYM and MLE approaches were also able to detect some neighbours of Winnipeg (14 regions) as potential clusters. However, the CSS and FSS methods only detected two regions, 10 and 43 as a potential cluster in addition to cluster D.

4. Discussion and Conclusions

We used five popular approaches in spatial epidemiology to identify potential clusters of COPD cases in the Canadian province of Manitoba, Canada. These five methods have been used extensively in the literature and are relatively comprehensive. These methods use different approaches (non-parametric to parametric) to test for significant clusters.

We considered four different alternative hypotheses to compare the results of the CSS, FSS, BYM, and MLE methods. All four methods did a good job of identifying potential clusters for dense populations (clusters C and D) but not for a dispersed population (cluster B). In general, the CSS method identified a lower number of regions combined as a potential cluster compared to the FSS method, due to the non-circular shape of some regions in the province of Manitoba. A disadvantage of the LISA method is that the results do not depend on the expected number of cases or the population in each region. This is concerning since regions with high populations will likely have higher observed numbers of disease, however, this is not taken into account when using the LISA method. Therefore, the LISA method could only be applied to cluster A as cluster B, C, and D require the expected numbers to be adjusted for the respective regions in a cluster. This may explain why the LISA approach identified some different regions as potential clusters when compared with the other procedures.

A region was identified as a potential cluster if the lower bound of credible/prediction interval of the estimated relative risk was larger than one for the BYM and MLE approaches. Different decision rules may be defined where the estimated relative risk (in terms of the credible/prediction interval) would be larger or smaller than one [29]. One could also consider the exceedance probability $\Pr(RR_i > b) > c$, where b can be 1, 2 or 3 and c may be a large value such as 0.90 [30]. For the LISA method, a region was defined to be significant if the associated p-value was less than 0.1. However, different decision rules could be used where the level of significance is smaller than 0.1.

Here, three important factors, age, gender and year were used to adjust the expected number of COPD cases in the province of Manitoba. Unlike the CSS, FSS, and LISA methods, we can extend the model A2–A3 in the Appendix, for both the BYM and MLE methods, to include other covariates directly, which may be required for some applications.

We also note that the methods have different settings and assumptions, which motivate our comparison. User-chosen settings are part of all cluster tests and different choices could lead to different results. All five methods have been proposed for local clusters. Under the null hypothesis, the number of COPD cases follows a Poisson distribution for the BYM and MLE methods, while the test statistic for the CSS and FSS methods has an asymptotically χ^2 distribution and the LISA method uses an empirical distribution. These features motivated us to consider these important methods and apply them to our COPD cases.

As limitations of this study, we assumed that our COPD cases are rare cases to be able to use the Poisson model in the BYM and MLE methods. Also, we used survey data (weighted to the population level) in our study. Strengths of the study include the evaluation of multiple cluster detection methods.

Overall, the potential clusters of COPD were located in the southern part of the province with the exception of region 24, which was identified by the BYM and MLE methods (cluster A). According to findings from Fransoo *et al.* [31], which are based on the Community Health Survey [27], binge drinking and smoking levels are higher than the Manitoba average in the south-central part of the province. As well, there are a higher percentage of people who consume low levels of fruits and vegetables in this region, although these differences are not statistically significant. Obesity levels are significantly higher than the Manitoba average in this region as well. In the south-eastern part of the province binge drinking, exposure to second hand smoke, and overweight and obesity levels are higher than the Manitoba average although the results are not statistically significant [31]. These are some possible health determinants that may explain the clusters of COPD in these regions. We found that the BYM and MLE methods are the best approaches in terms of identifying potential clusters and controlling for possible confounders (if any). These results may represent real increases in COPD or may be due to unmeasured covariates that need to be adjusted for in the model. Further investigation is needed to examine these findings, and also to explore the cause of these increases.

Acknowledgments

We would like to thank two referees for constructive comments and suggestions, which led to an improved version of the manuscript. This work was supported by grants from the Manitoba Health Research Council (MHRC) and the Natural Sciences and Engineering Research Council of Canada (NSERC) to M. Torabi. Disclaimer: The interpretations, conclusions and opinions expressed in this

paper are those of the authors and do not necessarily reflect the position of Statistics Canada. This study is based in part on data provided by the Research Data Centre at the University of Manitoba. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of Statistics Canada.

Author Contributions

The authors were both involved in the study concept and design, acquisition of data, interpretation of data, and the writing and critical revision of the manuscript for important intellectual content.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Circular Spatial Scan Statistic (CSS)

The spatial scan statistic is widely used in the field of epidemiology for a variety of purposes [32]. A circular window S is set on each region by the circular spatial scan statistic with the radius of the circle ranging from zero to a pre-specified maximum distance d or a pre-determined maximum number of regions J to be considered in the cluster. The window made up of the $(j\text{-}1)$-th nearest neighbours to region i is denoted by $S_{i:j}(j = 1,...,J)$. As well, $S_1 = \{S_{i:j}; i = 1,\ldots,m; j = 1,\ldots,J\}$ denotes the set of all windows to be scanned by the circular scan statistic. A likelihood ratio statistic based on the number of observed and expected cases inside and outside the circle is computed for each circle. Now, L_0 denotes the likelihood under the null hypothesis and $L_i(i = 1,\ldots,m)$ represents the likelihood under the alternative hypothesis. The null hypothesis states that there is no cluster in region i and is tested against the alternative hypothesis that a cluster exists in region i based on its j-th nearest neighbours. The likelihood ratio statistic is given by

$$\max_i \frac{L_i}{L_0} = \left(\frac{C_i}{E_i}\right)^{C_i} \left(\frac{N\text{-}C_i}{N\text{-}E_i}\right)^{N\text{-}C_i} I(C_i > E_i) \quad \text{(A1)}$$

where C_i denotes the observed number of cases inside a circle and E_i represents the expected number of cases inside a circle. Also, $(N\text{-}C_i)$ and $(N\text{-}E_i)$ denote the observed and expected number of cases outside the circle, respectively. The indicator function $I(C_i > E_i)$ is equal to 1 when $C_i > E_i$ and 0 otherwise. Circles with high likelihood ratios are identified as possible clusters [17].

Using SaTScan [33] or FleXScan [34] software, the CSS method can be applied. Generally, J is chosen to include at most 50% of the population at risk, however, we used the FleXScan default which is a maximum spatial cluster size of $J = 15$. In order for the region to be part of the circle, the region centroid had to be included in the radius of the circle.

Flexible Spatial Scan Statistic (FSS)

The flexible spatial scan statistic works in the same manner as the circular spatial scan statistic but now the shape of the potential cluster is flexible while still being restricted to a small neighbourhood of

each region. By connecting adjacent regions, the flexible scan statistic places an irregularly shaped window S on each region. For any region i, the set of irregularly shaped windows of length j, containing j connected regions including region i, can vary from 1 to the pre-determined maximum J, where J is the maximum length of a cluster. Moreover, to prevent unlikely cluster shapes, the joined regions are restricted to the subsets of the set of regions i and $(J\text{-}1)$-th nearest neighbours of region i. The set of all windows to be scanned by the flexible spatial scan statistic is then $S_2 = \{S_{i:j(k)}; i = 1,\ldots,m; j = 1,\ldots,J; k = 1,\ldots,k_{ij}\}$. The circular spatial scan statistic examines J circles for any region i and the flexible spatial scan statistic examines J circles plus all the sets of connected regions whose centroids are found within the J-th largest concentric circle. Subsequently, the size of S_2 is much larger than S_1 which is at most mJ. The test statistic used in the FSS method under the Poisson assumption is based on the likelihood ratio test given in Equation (A1). Now, the circle defined in Equation (A1) refers to S_2 instead of S_1. Using the FSS method, circles with high likelihood ratio values are considered to be potential regions of disease clusters [18]. The FSS method can also be applied using the FleXScan software [34] with the FleXScan default which is $J = 15$.

Bayesian Disease Mapping (BYM)

Identifying clusters can also be done through a Bayesian framework using Markov chain Monte Carlo (MCMC) methods [14,15,35,36]. First used by Besag *et al.* [14], Bayesian disease mapping (BYM) is a modeling approach consisting of two parts. In the first part of the model, it is assumed that the cases follow a Poisson distribution with an area specific parameter $\theta_i E_i$:

$$C_i \sim Poisson(\theta_i E_i) \tag{A2}$$

where the observed and expected number of cases in region i are given by C_i and E_i, respectively. The second part of the model is achieved through

$$\log(\theta_i) = \mu + \eta_i \tag{A3}$$

where the relative risk (RR) in region i is given by θ_i, μ represents the overall mean ratio over the entire region and η_i represents spatially correlated random effects. These spatial random effects are captured using the usual CAR model. A variety of CAR models can be used by obtaining a collection of mutually compatible conditional distributions $p(\eta_i|\eta_{-i}), i = 1,\ldots,m,$ where $\eta_{-i} = \{\eta_j: j \neq i, j \in \delta_i\}$ and δ_i refers to a set of neighbours for the i-th region [14]. We use the following general model for spatial effects η_i:

$$\boldsymbol{\eta} = (\eta_1,\ldots,\eta_m)' \sim N(0, \Sigma_\eta)$$

$$\Sigma_\eta = \sigma_\eta^2 (I_m - \lambda_\eta D)^{-1} P$$

where P is a $m \times m$ diagonal matrix with elements $P_{ii} = 1/e_i$; D is a $m \times m$ matrix with elements $D_{ij} = (e_j/e_i)^{1/2}$ if region i and j are adjacent and $D_{ij} = 0$ otherwise; e_i is the number of regions adjacent to region i; σ_η^2 is the spatial dispersion parameter; λ_η measures the spatial autocorrelation, $\lambda_{min} \leq \lambda_\eta \leq \lambda_{max}$, where λ_{min}^{-1} and λ_{max}^{-1} are the smallest and largest eigenvalues of $P^{-1/2} D P^{1/2}$; and I_m is an identity matrix of dimension m. We refer to [37] for details of this proper CAR model. Using vague

prior distributions and within the Bayesian framework (MCMC), the parameters may be estimated to produce posterior distributions for the parameters in the model [14].

A cluster is specified as a region where the estimated relative risk (in terms of the lower credible set) is significantly larger than one [38]. To apply this method, WinBUGS software [37] is used to calculate the relative risk values.

Frequentist Approach Using Maximum Likelihood Estimation (MLE)

The data cloning (DC) approach is a computational algorithm to obtain the MLE for hierarchical models [23,24]. This approach is based on the Bayesian computational method and is used for frequentist purposes. This method involves independently repeating the observations $\boldsymbol{C}=(C_1,\ldots,C_m)'$ for L different individuals. Subsequently, these individuals all have the exact same set of observations $\boldsymbol{C}$ which are represented by $\boldsymbol{C^{(L)}}=(\boldsymbol{C},\boldsymbol{C},\ldots,\boldsymbol{C})$. The posterior distribution of $\boldsymbol{\alpha}=(\mu,\lambda_\eta,\sigma_\eta^2)'$ conditional on the data $\boldsymbol{C^{(L)}}$ is then given by

$$\pi_L\left(\boldsymbol{\alpha}\middle|\boldsymbol{C^{(L)}}\right)=\frac{\{L(\boldsymbol{\alpha},\boldsymbol{C})\}^L\pi(\boldsymbol{\alpha})}{H(\boldsymbol{C^{(L)}})} \tag{A4}$$

where the prior distribution on the parameter space is $\pi(\boldsymbol{\alpha})$ and $H\left(\boldsymbol{C^{(L)}}\right)=\int\{L(\boldsymbol{\alpha},\boldsymbol{C})\}^L\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}$ is the normalizing constant. Also, $\{L(\boldsymbol{\alpha},\boldsymbol{C})\}^L$ represents the likelihood for L copies of the original data. As shown by Lele *et al.* [23,24], when L is large enough, $\pi_L(\boldsymbol{\alpha}|\boldsymbol{C^{(L)}})$ converges to a multivariate Normal distribution with the mean given by the MLE of the model parameters and variance-covariance matrix equal to $1/L$ times the inverse of the Fisher information matrix for the MLE. Hence, the sample mean vector of the generated random numbers from Equation (A4) acts as an estimate of the MLE and an estimate of the asymptotic variance-covariance matrix for the MLE $\hat{\boldsymbol{\alpha}}$ is given by L times the sample variance-covariance matrix of the generated random numbers from Equation (A4). Lele *et al.* [24] also provided various tests to determine the adequate number of clones L.

Prediction of Relative Risk:

The prediction of relative risk (random effects) can be fairly problematic, especially in the frequentist framework. One approach to estimate $\boldsymbol{r}$ using the data is to use $\pi(\boldsymbol{R}=\boldsymbol{r}|\boldsymbol{C},\hat{\boldsymbol{\alpha}})$ where $\boldsymbol{R}=(RR_1,\ldots,RR_m)'$. However, the variability introduced by the model parameters estimate is not captured in this approach. In the literature [39], it has been suggested to use the following density in order to take into consideration the variation of the estimator,

$$\pi(\boldsymbol{r}|\mathbf{y})=\frac{\int f(\boldsymbol{C}|\boldsymbol{r},\alpha_1)g(\boldsymbol{r}|\boldsymbol{\alpha_2})\phi(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},I^{-1}(\hat{\boldsymbol{\alpha}}))d\boldsymbol{\alpha}}{H(\boldsymbol{C})} \tag{A5}$$

where $\alpha_1=\mu$, $\boldsymbol{\alpha_2}=(\lambda_\eta,\sigma_\eta^2)'$, $f(\cdot)$ and $g(\cdot)$ are Poisson and Normal distributions, respectively, and $\phi(.,\xi,\Sigma)$ denotes a multivariate Normal density with mean ξ and variance-covariance Σ. In this paper, the prediction of the $\boldsymbol{r}$ was found using Equation (A5) through MCMC sampling. A disease cluster is defined as a region where the estimated relative risk (in terms of the lower prediction interval) is significantly larger than one. The dclone package [40] is utilized within the R software [41] in order to calculate the relative risk values.

Local Indicator of Spatial Association (LISA)

Another method for identifying spatial clusters is a local indicator of spatial association (LISA) statistic [26]. In general, for an observation, y_i in the i^{th} region, the LISA statistic is given by

$$L_i = f\left(y_i, y_{J_i}\right)$$

where f is a function and the values observed in the J^{th} neighbourhood of region i are given by y_{J_i}. In order to determine the statistical significance of the spatial association at region i, the following must be satisfied

$$\Pr(L_i > \delta_i) \leq \alpha_i$$

where a critical value is given by δ_i and α_i is a given level of significance. Another condition of a LISA statistic is the total of all LISA statistics in a region must be proportional to a global indicator of spatial association. In other words,

$$\sum_i L_i = \gamma\Lambda$$

where Λ is an indicator of the global indicator with a scale factor defined by γ. In order to test whether there is statistically significant spatial association over all the regions, the following statement must be true [26]

$$\Pr(\Lambda > \delta) \leq \alpha.$$

A general LISA statistic may be used to test the null hypothesis of no spatial association against the alternative hypothesis that spatial clustering exists across a region. However, the distribution of the general LISA may be hard to find. For this reason, conditional randomization or a permutation approach is used to find an empirical distribution. The randomization is done by holding the observed value (y_i) in region i constant and the remaining observed values across the entire study region are randomly permuted and the value of L_i is computed. This is done for each region in the study area. The result is an empirical distribution function, which expresses the extent to which each observation is considered to be extreme in comparison with the other observed values [26].

The LISA method is usually a simple method to apply, however, it is complicated by the fact that the LISA statistics for individual regions may be correlated. For example, when regions i and k are neighbours or have common elements in their neighbourhood sets, the corresponding LISA statistics, L_i and L_k will be correlated. Typically, it is extremely hard to derive the marginal distributions of each statistic and therefore, the significance levels must be approximated by Bonferroni inequalities or the method outlined by Sidák [42]. Using Bonferroni inequalities, the individual significance levels (α_i) are set to α/m and using Sidák's method, they are equal to $1 - \left(1 - \alpha\right)^{1/m}$, where the overall significance level is set to α and there are m comparisons. It has been suggested that m is taken to be the number of observations n. However, this may result in bounds that are too conservative and in fact very few observations may be deemed to be significant clusters [26]. Further investigation is being conducted to determine the best value for m. In our study this method is implemented in R [41] using the ncf package [43].

References

1. Global Initiative for Chronic Obstructive Lung Disease (GOLD), Global Strategy for the Diagnosis, Management and Prevention of COPD (updated 2013). Available online: http://www.goldcopd.org (accessed on 3 July 2013).
2. Eisner, M.D.; Anthonisen, N.; Coultas, D.; Kuenzli, N.; Perez-Padilla, R.; Postma, D.; Romieu, I.; Silverman, E.K.; Balmes, J.R. An official American Thoracic Society public policy statement: Novel risk factors and the global burden on chronic obstructive pulmonary disease. *Amer. J. Respir. Crit. Care Med.* **2010**, *182*, 693–718.
3. Sezer, H.; Akkurt, I.; Guler, N.; Marakoğlu, K.; Berk, S. A case-control study on the effect of exposure to different substances on the development of COPD. *Ann. Epidemiol.* **2006**, *16*, 59–62.
4. Burt, L.; Corbridge, S. COPD exacerbations. *Amer. J. Nurs.* **2013**, *113*, 34–43.
5. Lamprecht, B.; McBurnie, M.A.; Vollmer, W.M.; Gudmundsson, G.; Welte, T.; Nizankowska-Mogilnicka, E.; Studnicka, M.; Bateman, E.; Anto, J.M.; Burney, P.; *et al.* COPD in never smokers: Results from the population-based burden of obstructive lung disease study. *Chest* **2011**, *139*, 752–763.
6. Canadian Thoracic Society. The Human and Economic Burden of COPD: A Leading Cause of Hospital Admission in Canada; Canadian Thoracic Society: Ottawa, ON, Canada, 2010.
7. Canadian Institute of Health Information. Health Indicators 2008. Available online: https://secure.cihi.ca/free_products/HealthIndicators2008_ENGweb.pdf (accessed on 3 July 2013).
8. Mittmann, N.; Kuramoto, L.; Seung, S.J.; Haddon, J.M.; Bradley-Kennedy, C.; FitzGerald, J.M. The cost of moderate and severe COPD exacerbations to the Canadian healthcare system. *Respir. Med.* **2008**, *102*, 413–421.
9. Scheinfeld, M.H.; Maniatis, T.; Gurell, D. COPD? *Amer. J. Med.* **2006**, *119*, 839–842.
10. Lawson, A.B. *Statistical Methods in Spatial Epidemiology*, 2nd ed.; John Wiley & Sons, Ltd.: London, UK, 2006.
11. Jennings, J.M.; Curriero, F.C.; Celentano, D.; Ellen, J.M. Geographic identification of high gonorrhea transmission areas in Baltimore, Maryland. *Amer. J. Epidemiol.* **2005**, *161*, 73–80.
12. Elliott, P.; Briggs, D.; Morris, S.; de Hoogh, C.; Hurt, C.; Jensen, T.K.; Maitland, I.; Richardson, S.; Wakefield, J.; Jarup, L. Risk of adverse birth outcomes in populations living near landfill sites. *Brit. Med. J.* **2001**, *323*, 363–368.
13. Lawson, A.B.; Biggeri, A.; Williams, F.L.R. A review of modeling approaches in health risk assessment around putative sources. In *Disease Mapping and Risk Assessment for Public Health*; Lawson, A.B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J., Bertollini, R., Eds.; Wiley: New York, NY, USA, 1999; pp. 231–245.
14. Besag, J.E.; York, J.C.; Mollìe, A. Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Stat. Math.* **1991**, *43*, 1–59.
15. Clayton, D.; Bernardinelli, L. Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*; Elliott, P., Cuzick, J., English, D., Stern, R., Eds.; Oxford University Press: Oxford, UK, 1996; pp. 205–220.
16. Clayton, D.; Kaldor, J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **1987**, *43*, 671–681.

17. Kulldorff, M. A spatial scan statistics. *Commun. Statist. A—Theor. Method.* **1997**, *26*, 1481–1496.
18. Tango, T.; Takahashi, K. A flexibly shaped spatial scan statistic for detecting clusters. *Int. J. Health Geogr.* **2005**, *4*, 1–15.
19. Besag, J.E.; Newell, J. The detection of clusters in rare diseases. *J. R. Stati. Soc. Ser. A* **1991**, *154*, 143–155.
20. Torabi, M.; Rosychuk, R.J. Spatial event cluster detection using an approximate normal distribution. *Int. J. Health Geogr.* **2008**, *7*, 1–22.
21. Tango, T. A test for spatial disease clustering adjusted for multiple testing. *Stat. Med.* **2000**, *19*, 191–204.
22. Torabi, M.; Rosychuk, R.J. An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada. *Spat. Spatiotemporal Epidemiol.* **2011**, *2*, 321–330.
23. Lele, S.R.; Dennis, B.; Lutscher, F. Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* **2007**, *10*, 551–563.
24. Lele, S.R.; Nadeem, K.; Schmuland, B. Estimability and likelihood inference for generalized linear mixed models using data cloning. *J. Am. Stat. Assoc.* **2010**, *105*, 1617–1625.
25. Torabi, M. Spatial disease cluster detection: An application to childhood asthma in Manitoba, Canada. *J. Biom. Biostat.* **2012**, doi:10.4172/2155-6180.S7-010.
26. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *2*, 93–115.
27. Statistics Canada. *Canadian Community Health Survey User Guide (2001–2010)*; Statistics Canada: Ottawa, ON, Canada, 2010.
28. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman and Hall: London, UK, 1989.
29. Richardson, S.; Thomson, A.; Best, N.; Elliott, P. Interpreting posterior risk estimates in disease-mapping studies. *Environ. Health Perspect.* **2004**, *112*, 1016–1025.
30. Banerjee, S.; Gelfand, A.E.; Carlin, B.P. *Hierarchical Modeling and Analysis for Spatial Data*; Chapman and Hall: London, UK, 2004.
31. Fransoo, R.; Martens, P.; Burland, E. *The Need to Know Team*; Prior, H., Burchill, C., Eds.; Manitoba Centre for Health Policy, Manitoba RHA Indicators Atlas: Winnipeg, MB, Canada, 2009.
32. Fukuda, Y.; Umezaki, M.; Nakamura, K.; Takano, T. Variations in social characteristics of spatial disease clusters: Examples of colon, lung and breast cancer in Japan. *Int. J. Health Geogr.* **2005**, *4*, 1–13.
33. Kulldorff, M.; Rand, K.; Gherman, G.; Williams, G.; DeFrancesco, D. *SaTScan V2.1: Software for the Spatial and Space-Time Scan Statistics*; National Centre Institute: Bethesda, MD, USA, 1998.
34. Takahashi, K.; Yokoyama, T.; Tango, T. *FleXScan: Software for the Flexible Scan Statistic*; National Institute of Public Health: Nagoya, Japan, 2006.
35. Bernardinelli, L.; Montomoli, C. Empirical Bayes *versus* fully Bayesian analysis of geographical variation in disease risk. *Stat. Med.* **1992**, *11*, 983–1007.
36. Gilks, W.R.; Richardson, S.; Spielhalter, D.J. *Markov Chain Monte Carlo in Practice*; Chapman and Hall/CRC: London, UK, 1995.

37. Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. *WinBUGS Version 1.4 User Manual*; MRC Biostatistics Unit, Institute of Public Health: London, UK, 2004.
38. Aamodt, G.; Samuelsen, S.O.; Skrondal, A. A simulated study of three methods for detecting disease clusters. *Int. J. Health Geogr.* **2006**, *5*, 1–11.
39. Hamilton, J.D. A standard error for the estimated state vector of a state-space model. *J. Econometrics* **1986**, *33*, 387–397.
40. Sólymos, P. Dclone: Data cloning in R. *R J.* **2010**, *2*, 29–37.
41. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
42. Sidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.* **1967**, *62*, 626–633.
43. Bjornstad, O.N. ncf: Spatial Nonparametric Covariance Functions. R Package Version 1.1-5. 21 November 2013. Available online: http://cran.r-project.org/web/packages/ncf/index.html (accessed on 24 July 2014).