Evaluating the Evaluation Metrics for Spatial Disease Cluster Detection Algorithms

Raphaella Carvalho Diniz
Universidade Federal de Minas Gerais
raphaella.diniz@dcc.ufmg.br

Pedro O.S. Vaz-de-Melo
Universidade Federal de Minas Gerais
olmo@dcc.ufmg.br

Renato Assunção
Universidade Federal de Minas Gerais
assuncao@dcc.ufmg.br

ABSTRACT

We show that the usual evaluation metrics used in machine learning are not appropriate to measure the performance of spatial disease cluster detection algorithms. We demonstrate that the usual recall and precision metrics give a distorted evaluation of the algorithms. To solve this problem, we propose new metrics based on probability predictive rules. We evaluate the performance of the main spatial disease cluster algorithms with these new metrics. Our analysis and experiments offer insights into when the usual metrics are not appropriate and also show that our proposal is very effective at eliminating the bias from the usual metrics.

CCS CONCEPTS

• **Information systems** → **Clustering**; • **Mathematics of computing** → **Cluster analysis**.

KEYWORDS

cluster detection, scan statistics, spatial statistics

ACM Reference Format:
Raphaella Carvalho Diniz, Pedro O.S. Vaz-de-Melo, and Renato Assunção. 2020. Evaluating the Evaluation Metrics for Spatial Disease Cluster Detection Algorithms. In *28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20), November 3–6, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397536.3422251

1 INTRODUCTION

A major problem for public health authorities is the need for investigating suspected cancer clusters and responding to community concerns. According to Thun and Sinks [22], annually, more than 1,000 inquiries about suspected cancer clusters must be responded by state and local health departments. A cancer cluster is defined as a greater than expected number of cancer cases that occurs within a group of people in a geographic area over a defined period of time. The first step to evaluate a suspected cluster is the comparison between the observed number of cancer cases in the community with the expected number, taking into account the population size and its age-sex distribution. This comparison is made by means of a statistical hypothesis test to verify if the difference between the observed and expected numbers could happen by mere chance when there is no increased risk in the region. Classical solutions for these comparisons were plagued by the multiple testing problem, where a very large number of comparisons leads to a substantial number of false-positive zones. Kulldorff spatial scan statistic [9, 11] solved this problem in an elegant and simple way. First, the comparison should consider initially the zone leading to the worst-case over all these candidate cluster zones. Second, the adoption of a Monte Carlo method to carry out a principled statistical test for the significance of this worst-case zone. This test delivers a valid p-value for the null hypothesis of *no cluster in the map* taking into account the multiple testing and the overlapping potential candidate clusters. This idea has been widely adopted and extended in many ways [3, 8, 15, 19]. For the spatial cluster detection, the classical scan statistics algorithm had a spatial restriction: it scans only circular shaped potential clusters. Several papers tried to overcome this shape restriction by searching over a larger set of potential cluster zones [2, 4, 10, 21].

The many alternative algorithms for the spatial disease detection cluster problem have been evaluated using the common metrics from machine learning, such as recall, precision, and F1-score. However, the usual evaluation metrics used in machine learning are not appropriate to measure the performance of spatial disease cluster detection algorithms. In short, regions with a small population are easily prone to have very high or very low incidence rates with only a few cases more or less and, because of that, should receive less weight in the evaluation metric. The usual recall and precision metrics give a distorted evaluation of the algorithms.

2 RELATED WORK

A recent trend in Machine Learning is asking researchers to increase their standards for empirical analysis and empirical rigor across the field [5, 14, 17]. Gunawardana and Shani [6] demonstrated how using an improper evaluation metric can lead to the selection of an improper algorithm for three tasks associated with recommender systems. To the best of our knowledge, we are the first to revisit the evaluation metrics traditionally used in the task of spatial disease cluster detection. However, we acknowledge that Han et al. [7] discussed an associated task in their paper. These authors pointed out that the spatial information a single point event contributes is limited to a binary 0 or 1 depending on either being inside or outside a region. To overcome this problem, they proposed a method that makes using a continuous scan statistic that generalizes the spatial scan statistics by allowing the point contribution to be determined by a Gaussian kernel. In our case, we handle a similar limitation by proposing evaluation metrics that reduce the importance of regions sensitive to small random fluctuations in the observed data.

SIGSPATIAL '20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8019-5/20/11.
https://doi.org/10.1145/3397536.3422251

3 DEFINITIONS AND PROBLEM STATEMENT

The input data is a region partitioned into N small areas indexed by $i = 1, \ldots, N$. Each area counts a certain number y_i of cases or events in a given period of time among a risk population n_i. We assume that $y_i \sim \text{Poisson}(\theta_0 n_i)$. The parameter θ_0 is the per capita risk. The cluster C is a relatively small subset of *adjacent areas* that has a higher risk than θ_0. If $i \in C$, we have $y_i \sim \text{Poisson}(\theta_c n_i)$ with $\theta_c > \theta_0$. The objective is to detect the presence of a zone with high risk given **only** (i) the cases y_i and (ii) the risk population n_i.

3.1 Evaluation Using Synthetic Data

The standard way is to compare what the algorithm retrieves to a true and known high-risk cluster synthetically imposed on a map [12, 13, 16, 18, 20]. A benchmark dataset created by [12, 13] is used by virtually all the algorithms being proposed for the spatial disease cluster detection problem, which is known as *Northeastern USA Benchmark Data, Purely Spatial*. This database consists of 245 counties in the Northeastern United States. The population of the study region represents the female population registered by the 1990 demographic census, totaling around 29 million individuals. It has a large variety of patterns and regions with small populations and others with very large populations.

A large number of synthetic datasets are generated by simulating disease cases and spreading them among the areas. The first set of 100 thousand random maps are generated with no cluster present. The risk is constant and equal to θ_0 for all areas on the map and cases are randomly assigned to the areas with probabilities proportional to the population sizes n_i. For the scenarios where clusters are present, one set C of adjacent areas has a risk θ_c larger than θ_0. Disease cases are again distributed, but now the spatial clusters areas have a larger probability of receiving cases than simply determined by their population size.

3.2 Spatial Detection Algorithms

The algorithms proposed to detect spatial disease clusters establish a large collection $\mathcal{Z}$ of candidate clusters $z \in \mathcal{Z}$. Each potential cluster z is a connected region composed of a subset of areas. The classical spatial scan statistic [9] builds $\mathcal{Z}$ by reducing the potential clusters to a circular shape. It is based on the maximum likelihood procedure indexed by the parameter triplet $(z, \theta_c(z), \theta_0(z))$, where $z \in \mathcal{Z}$, the parameter $\theta_c(z)$ is the risk inside the zone z, and $\theta_0(z)$ represents the risk outside z. The null hypothesis is that $H_0 : \theta_c(z) = \theta_0(z)$ for all $z \in \mathcal{Z}$ and the alternative hypothesis is that there is one $z \in \mathcal{Z}$ such that $\theta_c(z) > \theta_0(z)$. The spatial scan statistic finds the zone $\hat{z}$ that maximizes the likelihood function $L(z, \theta_z, \theta_0)$ under the restriction that $\theta_z > \theta_0$:

$$\hat{L} = L(\hat{z}, \hat{\theta}_c(\hat{z}), \hat{\theta}_0(\hat{z})) = \sup_{z \in Z, \theta_c > \theta_0} L(z, \theta_c(z), \theta_0(z)) \qquad (1)$$

The null hypothesis distribution of $\hat{L}$ is obtained through a Monte Carlo method with a corresponding single valid p-value for the most likely cluster.

There have been many extensions and improvements of the classical scan statistics. One direction is the use of a more flexible set $\mathcal{Z}$ of cluster candidates. Duczmal and Assunção [4] were the first to propose a flexible scan statistics allowing any connected subgraph based on the adjacency graph in the collection $\mathcal{Z}$. They used a simulated annealing algorithm to maximize the likelihood, but the results are plagued by overfitting issues, with the detected zone showing odd shapes, with elongated arms as an octopus [1, 21]. The elliptic scan statistic [10] detects clusters with circular and elliptic shapes. Tango and Kunihiko proposed FleXcan [21], which composes $\mathcal{Z}$ by all sets of connected subgraphs based on boundary adjacency up to certain maximum number K of areas. Costa et al. [2] proposed two other flexible shaped cluster detection methods, namely Mlink and Double, which greedily search for a connected adjacency sub-graph $\hat{z}$ that maximizes the likelihood $L(z, \theta_c(z), \theta_0(z))$.

3.3 Usual Evaluation Metrics

All methods are said to detect a zone $\hat{z}$ if the test is statistically significant or, equivalently, if its associated p-value is smaller than a certain threshold, typically 0.05 or 0.01. Otherwise, no cluster is detected. To establish notation, we denote by C the true known set of areas selected to be a cluster in the simulations. We use $\hat{C}$ to represent the detected cluster (that is, $\hat{C} = \hat{z}$ when the algorithm has a statistically significant result, and $\hat{C} = \emptyset$, otherwise).

The most basic metric is the **statistical power**, the probability π that a cluster is detected when one is indeed present in the data: $\pi = \mathbb{P}(\hat{C} = \hat{z} | C \neq \emptyset)$. Note that we can have $\hat{C} \cap C = \emptyset$. **Recall** or sensitivity is denoted by R and there is more than one possible definition. The most common is $\mathbb{P}(C \cap \hat{C} \neq \emptyset \mid C \neq \emptyset)$: the probability that we detect at least one of the areas that make up the cluster C. Another traditional metric is precision: $\mathbb{P}(C \cap \hat{C} \neq \emptyset \mid \hat{C} \neq \emptyset)$. Given that the algorithm returns some statistically significant cluster $\hat{C}$, it is the probability that we detect at least one of the areas in C.

3.4 Evaluation Difficulties

The main problem of these measures is that all regions are treated equally by them. This is problematic because a region belonging to cluster C is not a guarantee that the observed data reveals the presence of a greater risk in that area. In fact, surprisingly, the *opposite* can be true. The situation becomes more complicated when we assess the likelihood of a zero disease cases region belonging to the cluster. First, even belonging to the cluster and having a higher risk than in the rest of the map, we can have zero disease cases as the more probable value to be observed in a certain area. If it has a small population, it would be necessary a huge relative risk to likely produce even a single case in the area. Second, and complicating further the analysis, the probability of observing zero cases can be *larger under the null hypothesis than under the alternative hypothesis*. That is, even belonging to the cluster and having an underlying higher risk than in the rest of the map, zero cases is more compatible with the *not in the cluster* situation than with the *in the cluster* situation: $\mathbb{P}(y_i = 0 \mid i \notin C) > \mathbb{P}(y_i = 0 \mid i \in C)$.

4 NEW EVALUATION METRICS

Each particular area $i \in C$ has a different probability of being detected depending on (i) the disease incidence rates, (ii) what disease cases are instantiated in that area and (iii) on the cluster detection algorithm adopted. In the usual recall metric calculation, we take the average (over simulations) of $\#(\hat{C} \cap C)/\#(C)$. To count equally

all the areas from C in this denominator is not adequate as some of them, most of the time, will not provide any data-evidence of belonging to a high-risk cluster. This will unduly benefit some algorithms such as those that *a priori* search only for circular clusters. At the same time, these same algorithms will be unduly penalized in the precision metric.

Let $D_i = 1$, if area i is detected (or, equivalently, if $i \in \hat{C}$) and $D_i = 0$, otherwise. Let Z_i be a binary indicator that area i belongs to the spatial disease cluster C and denote by $\pi_i = \mathbb{P}(Z_i = 1)$ the prior probability that the i-th area belongs to a cluster C. We have

$$\mathbb{P}(Y_i = y_i | Z_i = 1) = f(y_i; \theta_c)$$

and

$$\mathbb{P}(Y_i = y_i | Z_i = 0) = f(y_i; \theta_0) \, .$$

Using the Bayes rule, we obtain $\rho_i = \mathbb{P}(Z_i = 1 | Y_i = y_i)$, the probability that the i-th area belongs to the cluster conditioned on the observed number y_i of disease cases:

$$\rho_i = \frac{f(y_i; \theta_c)\pi_i}{f(y_i; \theta_c)\pi_i + f(y_i; \theta_0)(1 - \pi_i)} \, . \tag{2}$$

We redefine the recall metric R in a way that an algorithm that misses small population areas belonging to a high-risk cluster but has small counts of cases are not unfairly penalized:

$$\begin{aligned} R \quad &\stackrel{\text{def}}{=} \quad \frac{\sum_{i \in C} \mathbb{P}(D_i = 1 \text{ and } Z_i = 1 \mid Y_i = y_i)}{\sum_{i \in C} \mathbb{P}(Z_i = 1 \mid Y_i = y_i)} \\ &= \quad \frac{\sum_{i \in C} \mathbb{P}(D_i Z_i = 1 \mid Y_i = y_i)}{\sum_{i \in C} \rho_i} \end{aligned} \tag{3}$$

The motivation is clear if we focus on the denominator. Rather than adding 1 irrespective of what is observed, each area of the cluster contributes with ρ_i, its posterior data-evidence of belonging to the cluster. The denominator will be a (non-integer) number in the interval $(0, \#C]$ and it is equal to the posterior expected number of areas in the cluster: $\mathbb{E}(\sum_{i \in C} Z_i | Y_i = y_i)$. This is the expected number of the true cluster areas that may be considered riskier than the baseline risk *given the specific observed instances* y_i. The numerator is the expected number of the true cluster that may be considered riskier than baseline and that is detected by the algorithm conditioned on the observed y_i.

Concerning precision, we consider a definition with a rationale similar to that used in (3).

$$P \stackrel{\text{def}}{=} \frac{1}{\#\hat{C}} \sum_{i \in \hat{C}} \mathbb{P}(Z_i = 1 \mid D_i = 1, Y_i = y_i) \tag{4}$$

These measures depend on the choice of the hyperparameter π_i. Typically, we adopt a constant value for all areas $i = 1, \ldots, N$. We say more about this choice in the next section.

5 EVALUATION

Our proposed metrics were evaluated using the "Northeastern USA Benchmark Data, Purely Spatial" dataset, described in Section 3.1. The experiments were conducted using an rural area cluster composed of fours areas with a small population and 600 cases in the map. The benchmark dataset works basically with circular shaped clusters. Hence, we created a second source of ground truth spatial clusters with irregularly shaped clusters but these results are not shown here due to lack of space.

The methods considered in our experiments were the Circular Spatial Scan Statistics [9], Double, Mlink [2], FleXScan [21] and Elliptic Spatial Scan Statistics [10] with three values for the penalty parameter: 0 (no penalty), 0.5 (medium penalty) and 1 (strong penalty). For all methods, we adopted the Poisson model for the likelihood function and a cluster size limit of 50% of the population. For FleXScan method, we set the maximum number of regions to be included in the cluster as 10, greater than the true cluster used in all the simulations.

We evaluated the scenarios for each method using the usual recall and precision metrics, $|C \cap \hat{C}|/|C| = |C \cap \hat{C}|/4$ and $|C \cap \hat{C}|/|\hat{C}|$, respectively. We also present our R and P proposed metrics, Bayesian Recall and Bayesian Precision, respectively. Our metrics were analyzed using two different values for π_i, the prior probability that an area is part of the spatial cluster. The first value is equal to $1/2$ and it represents a state of complete ignorance about the cluster. The second value is given by the number of areas in the true cluster divided by the total number of areas in the region of study. We do not favor this second choice for this parameter as it seems too pessimistic in terms of the prior knowledge of what the spatial detection algorithm will return. The results for these two values for π_i are denoted by $B(1/2)$ and $B(n)$. We also present the results using two simple-minded metrics R^* and P^* based on the sum of the populations. The significance level considered in all tests was equal to 0.05. Only significant results were included in the metrics calculation.

Before we show the results, it is important to understand what we aim to show. We are not trying to find what is the best spatial cluster detection algorithm. What we aim is to show that the ranking among the different algorithms change if more appropriate evaluation metrics than the usual recall and precision are used for this task. A longer and more detailed study, considering additional proposals for this task and a larger set of cluster shapes and population sizes is required to properly evaluate the spatial detection methods. The point is that this evaluation should use metrics such as the ones we are proposing in this paper rather than the naive versions of precision and recall. We hope this paper motivates a discussion about the evaluation metrics and, having established what are the best way to measure performance in this task, to undertake the comparison between the spatial detection methods.

Table 1 show the metrics for the rural cluster used in this paper. The usual recall and precision are under the Regular column header while the Bayesian alternatives for $\pi_i = 1/2$ and $\pi_i = 7/245$ (the number of areas in the cluster divided by the total number of areas in the map) are shown under the headings $B(1/2)$ and $B(n)$, respectively. The population-count metrics are labelled as R^* and P^*.

We focus initially on the recall results. We see that the regular recall has smaller values for all algorithms when compared to our proposed recall metric with $\pi_i = 1/2$ ($B(1/2)$) and with R^*. This reflects the harsh penalty imposed on the regular recall metric when the small population area is not included in the detected cluster. Hence, the usual recall metric gives a much too pessimistic view of how the the algorithms perform. The increase in recall from the regular recall metric to $B(1/2)$ or R^* varies from 0.16 to 0.01. Comparing R^* and $B(1/2)$, we do not see much difference

Method	Recall				Precision			
	R^*	Regular	$B(1/2)$	$B(n)$	P^*	Regular	$B(1/2)$	$B(n)$
Circular	**0.964 (1)**	**0.959 (1)**	0.979 (2)	0.996 (2)	0.912 (2)	0.862 (2)	0.909 (2)	0.803 (2)
Double	0.931 (6)	0.871 (6)	0.929 (6)	0.990 (4)	**0.940 (1)**	**0.896 (1)**	**0.941 (1)**	**0.856 (1)**
Mlink	0.947 (4)	0.899 (5)	0.948 (5)	0.995 (3)	0.890 (5)	0.835 (5)	0.887 (5)	0.777 (5)
FleXScan	0.927 (7)	0.763 (7)	0.878 (7)	0.990 (4)	0.872 (6)	0.758 (6)	0.872 (6)	0.688 (6)
Elliptic(0)	0.944 (5)	0.910 (4)	0.955 (4)	0.995 (3)	0.816 (7)	0.734 (7)	0.827 (7)	0.642 (7)
Elliptic(0.5)	0.951 (3)	0.940 (3)	0.974 (3)	0.996 (2)	0.895 (4)	0.843 (4)	0.900 (4)	0.783 (4)
Elliptic(1)	0.960 (2)	0.955 (2)	**0.981 (1)**	**0.997 (1)**	0.906 (3)	0.857 (3)	0.907 (3)	0.800 (3)

Table 1: Rural circular cluster

and neither one dominates the other. The $B(n)$ metric has higher values than the other definitions, with all methods hosing very high recall. Regarding the ranking, the best and second best methods were swapped by $B(1/2)$ and $B(n)$. The same happened for the methods in the $5th$ and $6th$ position. So, although the ranking order is similar among the regular recall and the other metrics in this circular cluster scenario, some important changes were seen.

We focus now on the precision results, which are shown in the second vertical block of the previous tables. Similar to what was seen for the recall metrics, the Bayesian precision (with $\pi_i = 1/2$) increased the values measured by the regular metric in all cases but there is virtually no change in the relative ranking of the methods in all clusters. The P^* measure did not have a consistent result. For the rural and B clusters, the values of P^* were practically identical to the Bayesian metric while cluster A and D had P^* restored to the regular precision values. Cluster C had P^* values typically higher than Bayesian precision values.

6 CONCLUSIONS

This paper focuses on the task of detecting spatial clusters of high risk in maps composed of small areas. In each area, we observe a rate determined by the ratio between the number of events, disease or death cases, and the population size. We showed that the usual recall and precision metrics are not appropriate to evaluate the algorithms in synthetic datasets. The reason is the mismatch between the cluster specification, crystallized on the generative probabilistic model used to produce the data, and what the instantiated data shows. We develop a probabilistic argument to redefine recall and precision in such a way to provide a fair evaluation of the algorithms. We hope to start a debate about the specificities and idiosyncrasies of the evaluation measures in the spatial disease cluster detection problem and to have provided an adequate preliminary solution.

REFERENCES

[1] Renato Assunção, M Costa, A Tavares, and S Ferreira. 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in medicine* 25, 5 (2006), 723–742.

[2] Marcelo Azevedo Costa, Renato Martins Assunção, and Martin Kulldorff. 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. *Computational Statistics & Data Analysis* 56, 6 (2012), 1771–1783.

[3] Jing Dai, Feng Chen, Sambit Sahu, and Milind Naphade. 2010. Regional behavior change detection via local spatial scan. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 490–493.

[4] Luiz Duczmal and Renato Assuncao. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 45, 2 (2004), 269–286.

[5] Jessica Zosa Forde and Michela Paganini. 2019. The Scientific Method in the Science of Machine Learning. In *ICLR Debugging Machine Learning Models Workshop*. 1–9. arXiv:1904.10922 http://arxiv.org/abs/1904.10922

[6] Asela Gunawardana and Guy Shani. 2009. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research* 10 (2009), 2935–2962.

[7] Mingxuan Han, Michael Matheny, and Jeff M. Phillips. 2019. The Kernel Spatial Scan Statistic. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, USA, 349–358. https://doi.org/10.1145/3347146.3359101

[8] Yan Huang and Jason W Powell. 2012. Detecting regions of disequilibrium in taxi services under uncertainty. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 139–148.

[9] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.

[10] Martin Kulldorff, Lan Huang, Linda Pickle, and Luiz Duczmal. 2006. An elliptic spatial scan statistic. *Statistics in medicine* 25, 22 (2006), 3929–3943.

[11] Martin Kulldorff and Neville Nagarwalla. 1995. Spatial disease clusters: detection and inference. *Statistics in medicine* 14, 8 (1995), 799–810.

[12] Martin Kulldorff, Toshiro Tango, and Peter J Park. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 42, 4 (2003), 665–684.

[13] Martin Kulldorff, Z Zhang, J Hartman, R Heffernan, L Huang, and F Mostashari. 2004. Benchmark data and power calculations for evaluating disease outbreak detection methods. *Morbidity and Mortality Weekly Report* (2004), 144–151.

[14] Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine-learning scholarship. *Queue* 17, 1 (2019), 1–15. https://doi.org/10.1145/3317287.3328534 arXiv:1807.03341

[15] Michael Matheny, Raghvendra Singh, Liang Zhang, Kaiqiang Wang, and Jeff M Phillips. 2016. Scalable spatial scan statistics through sampling. In *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems*. 1–10.

[16] Al Ozonoff, Marco Bonetti, Laura Forsberg, and Marcello Pagano. 2005. Power comparisons for an improved disease clustering test. *Computational statistics & data analysis* 48, 4 (2005), 679–684.

[17] D. Sculley, Jasper Snoek, Ali Rahimi, and Alex Wiltschko. 2018. Winner's curse? On pace, progress, and empirical rigor. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings* 2017 (2018), 1–4.

[18] Changhong Song and Martin Kulldorff. 2003. Power evaluation of disease clustering tests. *International journal of health geographics* 2, 1 (2003), 9.

[19] Roberto CSNP Souza, Renato M Assunção, Daniel B Neill, and Wagner Meira Jr. 2019. Detecting Spatial Clusters of Disease Infection Risk Using Sparsely Sampled Social Media Mobility Patterns. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 359–368.

[20] Kunihiko Takahashi and Toshiro Tango. 2006. An extended power of cluster detection tests. *Statistics in medicine* 25, 5 (2006), 841–852.

[21] Toshiro Tango and Kunihiko Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics* 4, 1 (2005), 11.

[22] Michael J Thun and Thomas Sinks. 2004. Understanding cancer clusters. *CA: A Cancer Journal for Clinicians* 54, 5 (2004), 273–280.