Measuring the Spatial Accuracy of the Spatial Scan Statistic

Simon Read[a], Peter Bath[a], Peter Willett[a], Ravi Maheswaran[b]

[a] *Information School, University of Sheffield, Sheffield S1 4DP, UK*
[b] *ScHARR, University of Sheffield, Sheffield S1 4DA, UK*
E-mail: simon.read@sheffield.ac.uk

Abstract

The spatial scan statistic is well established in spatial epidemiology. However, studies of its spatial accuracy are infrequent and vary in approach, often using multiple measures which complicate the objective ranking of different implementations of the statistic. We address this with three novel contributions. Firstly, a modular framework into which different definitions of spatial accuracy can be compared and hybridised. Secondly, we derive a new single measure, Ω, which takes account of all true and detected clusters, without the need for arbitrary weightings and irrespective of any chosen significance threshold. Thirdly, we demonstrate the new measure, alongside existing ones, in a study of the six output filter options provided by SaTScan™. The study suggests filtering overlapping detected clusters tends to reduce spatial accuracy, and visualising overlapping clusters may be better than filtering them out. Although we only address spatial accuracy, the framework and Ω may be extendible to spatio-temporal accuracy.

Keywords: spatial accuracy, spatial scan statistic, benchmark testing, performance measures, Bernoulli, omega

1. Introduction

The spatial scan statistic, hereafter SSS, is a widely used tool in spatial and spatio-temporal epidemiology. Introduced by Kulldorff and Nagarwalla (1995) and Kulldorff (1997), the purpose of the SSS is to detect the presence and location of clusters within spatial and spatio-temporal data sets. Implemented within the freely available SaTScan™ software (`www.satscan.org`), it has been used in well over one hundred published scholarly studies; see list in Kulldorff (2009).

Whilst the capacity of the SSS to accurately detect the presence of clusters has been widely studied, much less so its capacity to accurately detect their location. This should be of some concern. As Kulldorff (1997) states:

> ... the scan statistic has the ability to identify the zone responsible for rejecting the null hypothesis, and if we fail to detect the real cluster, it is of little comfort if the null hypothesis is rejected based on an untrue cluster in another part of the study area.

The reason that fewer studies consider spatial (or spatio-temporal) accuracy, may be because it is not immediately obvious how to measure it. The literature presents a patchwork of different measures and nomenclatures, with the most suitable scheme dependent on the type of data used and the aims of the study. The first objective of this paper is to explore a modular framework into which different measures of spatial accuracy can be classified, the aim being to ease the comparison and hybridization of different measures. This is presented in Section 2.

One additional complication is that most measures of spatial accuracy have multiple output parameters. This is problematic if one wishes to rank cluster detection systems in terms of spatial accuracy, without making an arbitrary choice about the relative weighting of these parameters. A second complication is that most existing measures of spatial accuracy are dependent upon an arbitrary choice of significance threshold. For non-spatial performance measures, a solution to both these problems already exists in the form of the two-alternative-forced-choice (hereafter 2AFC) test. The 2AFC test forms the basis of the area under curve (AUC) measure used with receiver operating characteristic (ROC) curves, providing a single performance measure that is independent of significance threshold, or the relative weighting of sensitivity and specificity.

The second objective of this paper is thus to provide a novel measure of spatial accuracy, based on a intuitively

straightforward 2AFC test, which is customised for the spatial scan statistic. We provisionally call this Ω; a full definition and derivation are presented in Section 3.

The final objective of this paper is to provide a brief, but useful demonstration, of the new Ω measure. To this end, the six different output filtering options provided by SaTScan™ are evaluated in terms of their effect on spatial accuracy. This is done using the Bernoulli version of the spatial scan statistic, applied to synthetic point data sets containing one (or several) spatial clusters. This is presented in Section 4.

A conclusion, and discussion of future work, is provided in Section 5. Note that although this paper only concerns spatial accuracy, it may well be feasible to extend the work to spatio-temporal studies.

2. A framework for measures of spatial accuracy

Consideration for the spatial accuracy of the SSS dates back to its inception. However, in subsequent assessments of its performance, the ability to determine where a cluster is (termed *spatial accuracy*) has not been studied as widely as the ability to detect whether a cluster is actually present (loosely termed *power*). This is exacerbated by the lack of a universally accepted definition or measure of spatial accuracy. The aim of this section is to present a framework which, with reference to example studies from the literature, allows existing definitions to be compared on similar terms. Note that this framework only covers spatial accuracy at present; the temporal dimension is somewhat more complicated, especially in real-time surveillance where the present has special importance. That said, the framework can be used for spatio-temporal studies in which time is effectively just an additional spatial dimension.

The framework considers the measurement of spatial accuracy as a *metafunction*, a term used (loosely) here to describe a collection of processes acting together as a single measuring tool. The input of this metafunction is the study region itself, the data contained in each benchmark data set, and information about that data set (e.g. details of any injected clusters, and the process by which the data were generated, if synthetic). The output of the metafunction is one or more scalar values, each indicating how successful the detection system has been, in some regard, in identifying the locations of any injected clusters, either within a single data set or across a batch of data sets. Here, we use the term *batch* to refer to a collection of hundreds or thousands of benchmark data sets, all generated using a similar underlying model, and usually based on the same study region. For example, it is common to have a batch representing the null hypothesis (no cluster present) and one or more batches containing data sets into which one or more clusters of some kind have been injected. By aggregating performance measures over all the data sets in a batch, one can detect even relatively small differences in performance between detection methods.

The framework present in this section has five levels, listed below. Levels 1 to 4 concern individual data sets, and Level 5 concerns the aggregation of results at batch level. Each level is discussed, with reference to the literature, in Sections 2.1 to 2.5 respectively.

Level 1: Spatial support

Level 2: Data function

Level 3: Sub-regions

Level 4: Performance measures

Level 5: Aggregation

2.1. Level 1: spatial support

Consider any location s within a study region R. This could be a point, a ZIP code, a census area, or any spatial reference one can conceive of. By specifying s at Level 1, we are free to use it generically in Levels 2 to 5, where it can represent any type of location.

To discern whether any given s is part of the true cluster, a detected cluster, or some combination thereof, is to implicitly invoke some function whose domain is the study region itself. The *support* of this function is those parts of the study region where this function is defined.This may be a very limited set of points, e.g. Read et al. (2009) and Savory et al. (2010), where one is only concerned with the exact centre of the true and detected clusters. It may be a more extensive set of points, e.g. home addresses as used by Huang et al. (2007), or a set of area centroids, e.g. the county centroids used by Jacquez (2009). Potentially it may even represent a set of continuous geospatial areas.

Spatial support need not cover the entire study region, i.e. $\bigcup s_i = R$ is *not* required. However, if one wishes to use clearly delineated sub-regions (see Section 2.3) there should be no overlap (i.e. $s_i \cap s_j = \oslash \; \forall \; i \neq j$).

2.2. Level 2: Data function

When calculating spatial accuracy, sometimes one is only interested in counting the number of locations correctly (or incorrectly) detected; examples are Jacquez (2009), Neill (2009), Que et al. (2008), and Waller et al. (2006). However, some measures of spatial accuracy

take into account data associated with each location, e.g. the use of the counts of affected individuals in each areal unit by Jung et al. (2007) and Olson et al. (2006). In either case, it is helpful to express these counts as a (non-parametric) function of s, which we call here the *data function*, or $f(s)$ for brevity. For instance, if one is measuring spatial accuracy by counting the number of individuals correctly (or incorrectly) identified, then one should define $f(s)$ as the number of affected individuals at location s. If one is only interested in counting locations, it is convenient to set $f(s)$ equal to the indicator function[1] $I(s \in S)$, where S is some subset of the study region that interests us (this is discussed further in Section 2.3). These examples require only simple data functions, and there may be considerable scope here for novel developments in the measurement of spatial accuracy.

For reasons that will become apparent in Section 3, it is very useful to define the data function as being proportional to the probability density function of s, when s is the output of a spatial Poisson process, equivalent to the uniformly random selection of an element of the data set. For example, consider a spatial accuracy measure of the type used by Jung et al. (2007): if one selects an affected individual in the study uniformly at random and notes the associated location s, then $f(s)$ is proportional the to probability that any given s will be selected. This constraint is entirely compatible with most of the spatial accuracy studies cited in this paper, thus is a very mild condition.

2.3. Level 3: sub-regions

It is useful to have shorthand for referring to different parts of the study region, independent of the data function or the spatial support. An excellent example is the a, b, c, d notation used by Jacquez (2009), where areal units into one of four types (illustrated in Figure 1):

a) Inside both true and detected cluster(s)
b) Inside true but outside detected cluster(s)
c) Inside detected but outside true cluster(s)
d) Outside both true and detected cluster(s)

A different, and slightly more succinct, subdivision method is used in papers introducing new versions of the SSS: Huang et al. (2007), Jung et al. (2007) and Jung et al. (2010). This is shown below, the relationship to the notation above given in brackets:

[1]The indicator function I evaluates to 1 when the expression following in brackets evaluates true, 0 otherwise.

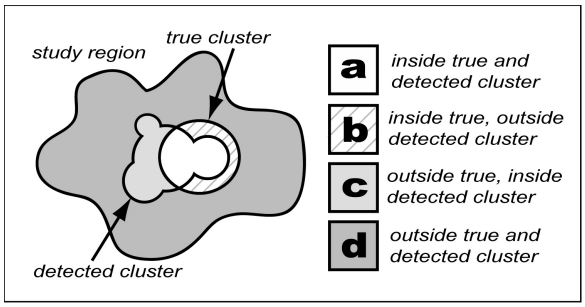

Figure 1: Example sub-regions

- The intersection of the true and detected cluster(s) ($\equiv a$)
- The true cluster(s) ($\equiv a \cup b$)
- The detected cluster(s) ($\equiv a \cup c$)

To use this kind of shorthand in the measures discussed in Section 2.4, one must first define scalar values to be associated with each sub-region[2]. This is simply the numerical integration of the data function $f(s)$ across each s within the sub-region concerned. For any subregion S, one express the its associated scalar value as:

$$\int_S f(s \in S)\, ds$$

For example, consider Que et al. (2008). Here s are postal code area centroids and (amongst other things) the authors measure the count of areal units included in both true and detected clusters. Using Levels 1 to 3 of this framework, one would define the area of overlap between true and detected clusters as a subregion (say a, for compatibility with Figure 1), and with this associate a value equal to the sum of $I(s \in a)$ for all postal code centroids s within a. It is important to note that these sub-regions, and their associated scalar values, are dependant not only on the size and shape of the detected clusters, but on the significance threshold used to screen out unlikely clusters. For example, if one uses a typical significance threshold of 0.05, then 'detected clusters' means only those scan windows produced that have a p-value of ≤ 0.05. This means that spatial accuracy, as it is measured in the studies cited here, is dependent on the exact choice of significance threshold. Section 3 explores this issue in more depth.

[2]Within this paper, sub-region notation such as a, b, c, d refers to both subsets of the study region (when used in set expressions), and the values associated with them (when used in scalar arithmetic). The usage should be clear from the context.

The next section explains how the scalar values associated with each sub-region are used to produce measures of spatial accuracy.

2.4. Level 4: Performance measures

The first three levels of the framework give a modular way of considering location and data within each data set. The fourth level brings these together to produce actual values representing the spatial accuracy of a particular detection method applied to a particular data set. This is where the difference between studies is greatest.

Read et al. (2009) and Savory et al. (2010) conducted benchmark studies where both the true and detected clusters have a clearly defined centre. Here the Euclidean distance between these centres provides a single, easily understandable measure of spatial accuracy. The advantage of a single measure is that it is straightforward to rank different detection algorithms in terms of their spatial accuracy. Unfortunately, the presence of multiple true and detected clusters within one data set complicates this approach, except when the correspondence of each true and detected cluster is very obvious, as in Savory et al. (2010). Also, the distance measure approach does not take account of the size and shape of the true and detected clusters.

The use of two of more scalar measures provides more flexibility, especially where there may be multiple (possibly overlapping) true and detected clusters, any of which could be highly irregular in shape and varying in size. Rather than considering distance, these studies consider the amount of overlap between true and detected clusters, and the amount of overlap between the detected clusters and those parts of the study region outside the true clusters. Examples of this approach can be found in Huang et al. (2007), Jacquez (2009), Jung et al. (2007), Jung et al. (2010), Neill (2009), Olson et al. (2006), and Que et al. (2008). All of the measures used in these studies are, implicitly or explicitly, based upon scalar values calculated for each of the subregions shown in Figure 1. For example, consider the definitions of *spatial sensitivity* and *spatial positive predictive value* (PPV) used in Huang et al. (2007), Jung et al. (2007) and Jung et al. (2010). In an individual data set level, these can be expressed as $= a/(a+b)$ and $= a/(a+c)$, respectively. To understand spatial accuracy measures calculated from the scalar values associated with a, b, c and d, it is perhaps easiest to present them in the familiar 2×2 table used for calculating non-spatial measures (see Table 1).

The variety of measures used in various studies is shown in Table 2. Note that nomenclature varies from study to study, even when referring to what is essentially the same thing. A particularly interesting measure is highlighted by Neill (2009), who uses the terminology of information retrieval: $recall = a/(a+b)$ and $precision = a/(a+c)$. Recognising the value of having a single scalar measure of spatial accuracy when ranking different methods, Neill uses an established method of combining these: the F-measure (van Rijsbergen, 1979), which is the harmonic mean of recall and precision. The only drawback is that the F-measure requires an assumption (implicit or explicit) about the relative weighting of recall and precision (Rennie, 2004). This issue is discussed further in Section 3.

	Inside true	Outside true
Inside detected	a	c
Outside detected	b	d

Table 1: Adaption of standard 2×2 table for classifying the subregions used in calculating spatial accuracy, after Jacquez (2009)

2.5. Level 5: Aggregation

Each measure in Level 4 has a scalar variable as its output. This means that within each data set, spatial accuracy is represented by one or more (typically two) scalar values. However, benchmark testing involves manifold data sets, and one needs measures that represent spatial accuracy at batch level, rather than individual data set level. The obvious choice is to take the arithmetic mean for each measure across all data sets in the batch. This is the approach taken by Huang et al. (2007), Jung et al. (2007), Jung et al. (2010), Neill (2009), Olson et al. (2006) and Que et al. (2008).

An alternative approach is taken by Waller et al. (2006), who take the mean across only those data sets where at least one cluster is detected with a p-value at or below the chosen significance threshold. This is a crucial choice which can make a significant difference to value of the aggregated measures, in difficult benchmarks tests where the power of the SSS is low, or if a very strict significance threshold is applied. Here only a small proportion of data sets may have detected clusters, and taking the mean across these data sets alone could give a volatile result. However, calculating the mean spatial accuracy across all data sets in a batch could be paradoxical: as one would then be including results from data sets where, statistically speaking, nothing has been detected (in which case the measures in Level 4 usually default to a certain value, e.g. zero).

This raises a fundamental question about the approach to measuring the performance of the SSS in a benchmark test: is the intended end user of the SSS asking this question:

- "Does this data set contain any true clusters, if so where are they?"

or two separate questions:

- "How certain am I this data set contains any true clusters?"
- "Where are any true clusters in this data set most likely to be?"

If one is certain it is the first, then clearly one only need consider data sets where a statically significant cluster is detected. Otherwise, one might consider all (or some) of the other data sets. Ironically, this is not a major issue in many existing studies as they generate data sets using models that result in true clusters that are not too difficult to detect.

Table 2 summarises all the studies mentioned in this section, outlined in terms of this framework. It can be seen that, due to the modular nature of the framework, a hybrid approach can be used to select aspects of spatial accuracy measurement from existing studies that are most suitable to the application concerned.

3. A unified measure of spatial accuracy

As discussed in Section 2.4, except in limited cases one needs at least two measures of spatial accuracy: one quantifying the amount of the true cluster that has been correctly detected, and one quantifying the amount of the study region outside the true cluster that has been incorrectly detected. However, if one wishes to rank different implementations of the SSS in terms of spatial accuracy, one needs to combine these functions into one scalar value. For the purposes of this discussion let us call this *dimension reduction*. Neill (2009) achieved dimension reduction using the F-measure; however this implicitly requires an assumption about the weighting of the two dimensions being combined (in this case spatial precision and spatial recall). It would be advantageous to have a single measure of spatial accuracy that obviates the need to weight the dimensions concerned.

Furthermore, most of the measures of spatial accuracy discussed in Section 2 are dependent on an arbitrary choice of significance threshold. This is because each cluster detected by the SSS has an associated p-value, and when comparing detected and true clusters (e.g. when delineating a and c in Figure 1) one would normally exclude detected clusters with unconvincingly high p-values. Thus, one must specify a significance threshold. One may then have a situation where one detection algorithm produces better spatial accuracy at one threshold, and another algorithm performs better at another threshold. It would be advantageous to have a measure of spatial accuracy that covers all significance threshold levels simultaneously.

This section derives such a measure. The starting point is a two-alternative-forced-choice (hereafter 2AFC) test, similar to that used in the derivation of the *area under curve* (AUC) measure[3]. AUC combines sensitivity and specificity to produce a single scalar value, which happens to equal the area under the corresponding receiver operating characteristic (ROC) curve. The AUC is equivalent to the probability that, when faced with one 'true' sample and one 'false' sample, the detection method under consideration can correctly identify which is which. The use of this 2AFC obviates the need to make a decision about the weighting of sensitivity and specificity, or the significance threshold of the test.

The 2AFC test used in this section is a forced choice between two randomly selected locations, s_1 and s_2, both within the study region. Let s_1 lie somewhere inside the true cluster, and s_2 somewhere outside[4]. Let s_1 be generated by a spatial Poisson process, such that any location inside the true clusters may be chosen with a probability density function proportional to the data function $f(s)$ (see Section 2.2). Similar for s_2. Let us define a measure called Ω representing the probability that when one is presented blindly with these two locations, and using only information provided by the SSS algorithm, one can correctly determine which is s_1, and which is s_2. It is important to note that one will never actually need to generate the locations s_1 and s_2, it is only the probability density functions of their potential locations that is of interest in calculating Ω.

The definition of Ω given above is necessarily technical for the purposes of the following proof. However, the idea of the probability of correctly choosing between two randomly selected locations, one inside the cluster and one outside, is straightforwardly intuitive. This is useful for non-technical readers of benchmark testing literature, who simply wish to know how well a detection method is likely to perform when faced with real data.

To calculate Ω, it is necessary to define some additional terms:

[3] Green and Swets (1966), revisited by Hanley and McNeil (1982)

[4] This necessitates that there is a least one true cluster present in the data set. This is acceptable as one cannot measure spatial accuracy when no true cluster is present, except perhaps to measure the amount of the study region incorrectly identified in any detected clusters. In this case a simple one-dimensional measure, specificity ($= d/(c + d)$), already exists.

Study	Framework level (see text)				
	spatial support	*data function*	*sub-regions*	*performance measures*	*aggregation*
Huang et al. (2007)	s = address	$f(s)$ = no. affected individuals at s	$S = a, b$ or c See Figure1	sensitivity $= a/(a+b)$ PPV $= a/(a+c)$	Arith. mean over all data sets generated under same model
Jung et al. (2007)	s = census tract				
Jung et al. (2010)	s = postal code area				
Read et al. (2009)	s_t = loci(centre) of the true cluster s_d = loci of most likely detected cluster	n/a	n/a	$\lvert s_t - s_d \rvert$	
Savory et al. (2010)	s_t = loci(centre) of each true cluster s_d = loci of each matching detected cluster	n/a	n/a	Arith. mean of $1/\lvert s_t - s_d \rvert$	Arith. mean and coeff. of variance over all data set gen. under same model
Waller et al. (2006)	s = census tract	$I(s \in S)$	$S = a, b$ or c where: a = centre of true cluster b = centres of all sig. detected clusters c = all detected clusters	Measure 1 = $I(a \cap b \neq \oslash)$ Measure 2 = $I(a \cap c \neq \oslash)$ where I is the indicator function	Measure 1: Sum over all data sets where $a \neq \oslash$ Measure 2: Sum over all data sets
Neill (2009)	s = grid square	$I(s \in S)$	$S = a, b$ or c See Figure1	recall $= a/(a+b)$ precision $= a/(a+c)$ F-measure $= 2a/(2a+b+c)$	Arith. mean over all data sets generated under same model
Jacquez (2009)	s = county centroids	$I(s \in S)$	$S = a, b, c$ or d See Figure1	power $= a/(a+b)$ false neg. $= b/(a+b)$ false pos. $= c/(c+d)$ specificity $= d/(c+d)$ detection acc. $= a/(a+c)$	n/a
Que et al. (2008)	s = postal code area	$I(s \in S)$	$S = a, b$ or c See Figure1	sensitivity $= a/(a+b)$ PPV $= a/(a+c)$	Arith. mean over all data sets generated under same model
Olson et al. (2006)	s = address *or* postal code area *or* census tract	$f(s)$ = no. affected individuals at s	$S = a, b$ or c See Figure1	Measure 1: $= I(a \geq b)$ Measure 2: $= c$	

Notes: $I(*)$ is the indicator function, where $I(*) = 1$ if * true, $I(*) = 0$ otherwise
PPV is positive predictive value
$\lvert s_t - s_d \rvert$ is the distance between points s_t and s_d

Table 2: Table showing the framework proposed in Section 2 as applied to various studies

- Let $A(\subset R)$ be the locus of all true clusters in R
- Let $A_C = R - A$, i.e. the locus of all R outside A
- Let $\{Z, \alpha\}$ represent a detected cluster, where $Z \subseteq R$ is the locus of the detected cluster and the α is the p-value, i.e. probability[5] that Z is a random artefact.
- Let $\{\alpha_1, \alpha_2 \ldots \alpha_X\}$ be the set of all unique p-values associated with the detected clusters.
- Define $\alpha_0 = 0$ and $\alpha_{X+1} = 1$
- Let $\mathcal{Z}_x$ be the union of all Z with associated p-value $\leq \alpha_x$, where $1 \leq x \leq X$.
- Define $\mathcal{Z}_0 = \oslash$, and $\mathcal{Z}_{X+1} = R$
- Let $a_1 = \mathcal{Z}_1 \cap A$
- Let $a_x = [\mathcal{Z}_x \cap A - \bigcup_{i=1}^{x-1} a_i]$ for $2 \leq x \leq X$
- Define $a_0 = \oslash$ and $a_{X+1} = A - \bigcup_{i=1}^{X} a_i$
- Let $c_1 = \mathcal{Z}_1 \cup A_C$
- Let $c_x = [\mathcal{Z}_x \cup A_C - \bigcup_{i=1}^{x-1} c_i\ ]$ for $2 \leq x \leq X$
- Define $c_0 = \oslash$ and $c_{X+1} = A_C - \bigcup_{i=1}^{X} c_i$

Put more intuitively: $\mathcal{Z}_x$ represents an amalgamation of the SSS output for each unique p-value, where all detected clusters with a p-value equal to or less than α_x are merged. Also $a_1 \cup c_1$ represents the locus of the most likely detected clusters, whilst $a_x \cup c_x$ represents the locus of the x^{th} most likely detected cluster, excluding locations included in the loci of more likely clusters. Note that a_{X+1} is equivalent to b in Figure 1, with c_{X+1} equivalent to d (this simplifies the expression for Ω). An example of this new spatial set notation is shown in Figure 2; note that a_3, a_4 and c_1 all happen to be null in the example illustrated.

As described in Section 2, one can numerically integrate the data function across each sub-region $a_1 \ldots a_{X+1}$ and $c_1 \ldots c_{X+1}$, to obtain scalar values representing the probability of s_1 and s_2 being located in each, respectively. Using these values, one can calculate Ω using the following formula:

$$\Omega = \frac{\sum_{y=1}^{X+1} \left[\sum_{k=0}^{y-1} a_k + \frac{1}{2}\, a_y \right] \cdot c_y}{\sum_{k=0}^{X+1} a_k \cdot \sum_{k=0}^{X+1} c_k} \qquad (1)$$

[5]For all but the most likely cluster, this probability is known to be slightly conservative (Kulldorff, 1997).

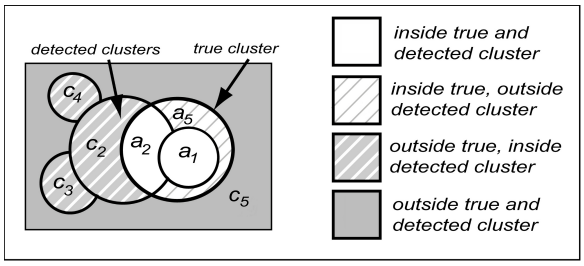

Figure 2: Example of subdivisions used in defining Ω

This gives a value from 0 to 1, with $\Omega = 1$ representing perfect spatial accuracy, where all detected clusters lie within the true clusters, and all true clusters lie within the detected clusters. As with the AUC measure, if one obtains $\Omega = 0$, i.e. perfect spatial inaccuracy, one could simply invert the detected clusters to achieve $\Omega = 1$. Hence for practical purposes, $\Omega = 0.5$ is the worst case. This means the detection system has provided no useful information in distinguishing which location is s_1 and which s_2, and the probability of guessing correctly is the same as if one were tossing a coin. The proof of the formula is as follows.

Proof. First, let i and j be the indices of the most significant $\mathcal{Z}$ to contain s_1 and s_2 (respectively), with 1 being the most significant and X being the least. If either s_1 or s_2 fall outside of $\mathcal{Z}_X$, let i or j (respectively) $= X + 1$. Therefore:

$$i = \begin{cases} i : s_1 \in \mathcal{Z}_i, s_1 \notin \mathcal{Z}_{i+1} & \text{if } 1 \leq i \leq X \\ i = X + 1 & \text{otherwise} \end{cases}$$

$$j = \begin{cases} j : s_2 \in \mathcal{Z}_j, s_2 \notin \mathcal{Z}_{j+1} & \text{if } 1 \leq j \leq X \\ j = X + 1 & \text{otherwise} \end{cases}$$

With regard to the 2AFC, if one is presented blindly with locations s_1 and s_2, and one's decision is only informed by the SSS output, then there are just three possibilities:

- P1: if $i < j$, one will answer correctly with probability 1
- P2: if $i > j$, one will answer incorrectly with probability 1
- P3: if $i = j$, one will have no useful information and be forced to guess, with probability 0.5 of being correct.

Using ρ to denote the probability, then Ω by definition equals:

$$\begin{aligned}\Omega &= \rho(\mathbf{P1}) + 0.5\,\rho(\mathbf{P3}) \\ &= \rho(i < j) + 0.5\,\rho(i = j) \\ &= \rho(i < y \mid j = y) + 0.5\,\rho(i = y \mid j = y)\end{aligned}$$

Where y is some integer. Taking marginal probabilities over possible values of y gives us:

$$\begin{aligned}\Omega &= \sum_{y=1}^{X+1} \rho(i < y) \cdot \rho(j = y) \\ &\quad + 0.5 \sum_{y=1}^{X+1} \rho(i = y) \cdot \rho(j = y) \\ &= \sum_{y=1}^{X+1} \left[\rho(i < y) + 0.5\,\rho(i = y)\right] \cdot \rho(j = y)\end{aligned}$$

Each probability can be expressed as follows:

$$\rho(i < y) = \frac{\sum_{k=0}^{y-1} \rho(s_1 \in a_k)}{\sum_{k=0}^{X+1} \rho(s_1 \in a_k)} \tag{2}$$

$$\rho(i = y) = \frac{\rho(s_1 \in a_y)}{\sum_{k=0}^{X+1} \rho(s_1 \in a_k)} \tag{3}$$

$$\rho(j = y) = \frac{\rho(s_2 \in c_y)}{\sum_{k=0}^{X+1} \rho(s_2 \in c_k)} \tag{4}$$

Because Ω is being defined in conjunction with the framework described in Section 2, we can draw upon the assumption made in Level 2 (see Section 2.2). This states that the data function $f(s)$ must be proportional to the probability density function of s, when s is the output of a spatial Poisson process. As s_1 and s_2 are randomly generated under such a process, one can write:

$$\begin{aligned}\rho(s_1 \in a_k) &\propto \int_{a_k} f(s)\,\mathrm{ds} \\ \rho(s_2 \in c_k) &\propto \int_{c_k} f(s)\,\mathrm{ds}\end{aligned}$$

When we are discussing $a_1 \ldots a_{X+1}$ and $c_1 \ldots c_{X+1}$ in terms of their scalar values, we can write the above expressions simply as:

$$\begin{aligned}\rho(s_1 \in a_k) &\propto a_k \\ \rho(s_2 \in c_k) &\propto c_k\end{aligned}$$

Now to obtain exact values for these probabilities, one would divide a_k by $(a_1 + a_2 + \cdots + a_X)$ and c_k by $(c_1 + c_2 + \cdots + c_X)$. However, as these denominators are the same for all a_k and c_k respectively, we can discard them when inserting the above values into Expressions 2, 3 and 4, which then become:

$$\rho(i < y) = \frac{\sum_{k=0}^{y-1} a_k}{\sum_{k=0}^{X+1} a_k}$$

$$\rho(i = y) = \frac{a_y}{\sum_{k=0}^{X+1} a_k}$$

$$\rho(j = y) = \frac{c_y}{\sum_{k=0}^{X+1} c_k}$$

Now inserting these into our previous expression for Ω gives Expression 1:

$$\Omega = \frac{\sum_{y=1}^{X+1} \left[\sum_{k=0}^{y-1} a_k + \frac{1}{2}\, a_y\right] \cdot c_y}{\sum_{k=0}^{X+1} a_k \cdot \sum_{k=0}^{X+1} c_k}$$

□

4. Example application

This section provides an example application of the framework presented in Section 2 and the Ω measure presented in Section 3. The number of candidate detected clusters produced by implementations of the SSS can be considerable. To aid users, SaTScan™ provides six choices of filter option (here called F1 to F6, listed below). These filters have no effect in power studies, as true and false positive rate depend only on the p-value of the most likely cluster, which is never filtered out. However, they almost certainly affect spatial accuracy. To the best of our knowledge, these filters have not been examined in the literature, save a basic overview in Kulldorff (2009), and there has been no published study to date concerning the pros and cons of each option.

F1: No geographical overlap

F2: No cluster centres in other clusters

F3: No cluster centres in more likely clusters

F4: No cluster centres in less likely clusters

F5: No pairs of centres both in each others clusters

F6: No restrictions, i.e. most likely cluster for each s

Here we briefly present the results of a benchmark study of filtering options, measuring spatial sensitivity and PPV (as defined in the sub-regions and performance measure columns of table 2), and Ω as defined in Section 3. The benchmark data sets used are similar to those in Read et al. (2009), where full details of the generation procedure is given. Each data set is a randomised distribution of 100 cases of a hypothetical disease, and 200 controls. Four batches were used; two generated using an homogeneous background density (CSR for short[6]); two using an inhomogeneous background density proportional to the 2001 population of the Trent region of the UK (TRENT for short). The background density is (effectively) the underlying spatial distribution of controls, i.e. the probability of a control occurring at a particular point is proportional to the background density at that point. E.g. in CSR data sets a control is equally likely to occur at any point with the study space.

The underlying distribution of cases follows that of controls, aside from the injection of one (×1TC) or three (×3TC) true clusters, i.e. one or three localised multiplicative increases in risk. These injections are Gaussian in shape (i.e. the increase is highest at the centre then tails off smoothly), isotropic, and uniformly randomly located[7]. The risk multiplier at the centre of each injection is termed the *maximum relative risk* (hereafter MRR). All MRR values were set to 15; which gave > 50% power in all batches for a standard 5% false alarm rate. Although a potential limitation of the study, this use of consistent sizes and shape of anomaly considerably reduces the number of different batches required, making this preliminary study feasible. The reference codes used in this study are given in Table 3.

For each data set in each batch, SaTScan™ was run six times on each data set, once for each choice of output filter (F1 - F6). Four batches of 1000 data sets, and six filters, gives 24,000 sets of results in total. A combination of Linux scripts and a MATLAB™ program was used to extract true and detected cluster information from each, and calculate spatial sensitivity and PPV, and Ω.

Batch code	Description
CSR×1TC	1000 data sets, each with: **one true cluster**, MRR=15 background density = CSR
CSR×3TC	1000 data sets, each with: **three true clusters**, MRR=15 background density = CSR
TRENT×1TC	1000 data sets, each with: **one true cluster**, MRR=15 background density ∝ pop. of Trent, UK
TRENT×3TC	1000 data sets, each with: **three true clusters**, MRR=15 background density ∝ pop. of Trent, UK

Table 3: Description of the four batches tested

Following the framework presented in Section 2, in similar layout to Table 2, the measures are shown in Table 4. Note the use a data function (Level 2) based on $\lambda(s)\times rr(s)$ rather than the count of events or areal units. $\lambda(s)$ is the background event rate at s, and $rr(s)$ is the multiplicative relative risk at s due to any true cluster located nearby; if s is unaffected by true clusters $rr(s) = 1$. This is more representative than the count of areal units, as it takes account of background population. It should also have lower variance than measures based on counts of events, which are inevitably subject to more random variation than the underlying risk value. For Level 5, two types of aggregation method were used. As Ω is intended to be independent of significance threshold, it is here averaged across all data sets. As sensitivity and PPV are linked to a particular significance threshold (here chosen as 0.05), they were averaged across only those data sets where the most likely cluster had a p-value of ≤ 0.05.

The results for each measure, for each batch and filter combination, for both scans, are shown in Table 5. As Mean Ω can be used for ranking the different filters in terms of spatial accuracy, a 95% confidence interval is included (shown as a ± in brackets, based on the standard error). It can be seen that, in terms of all three measures, filter performance falls into two loose groupings; F1, F2 and F4 tend to have lower Ω, lower sensitivity, and higher PPV; F3, F5 and F6 tend to have higher Ω, higher sensitivity, and lower PPV. The inverse relationship between sensitivity and PPV emphasises the need for a single measure when ranking different measures. Regarding Mean Ω, it can be seen from the confidence intervals that the overall difference between the two groups is highly unlikely be due to random chance.

[6]CSR stands for "complete spatial randomness", referring to the spatial distribution of controls in these data sets.

[7]With the exception of an exclusion area close to the border to avoid the need to consider edge effects.

Framework level (see text)				
spatial support	*data function*	*sub-regions*	*performance measures*	*aggregation*
s = point	$f(s) = \lambda(s) \times rr(s)$ where: $\lambda(s)$ = background event rate at s $rr(s)$ = relative risk at s attributable to true clusters (=1 outside clusters)	$\{a_i\}, \{c_i\}$	Ω	Arith. mean over all data sets
		a, b, c, d (at sig. threshold 0.05)	sensitivity $= a/(a+b)$ PPV $= a/(a+c)$	Arith. mean over all data sets with at least one detected cluster at sig. threshold 0.05

Notes: For examples of $\{a_i\}$ and $\{c_i\}$ see Figure 2. PPV is positive predictive value.

Table 4: Measures used in Section 4, in context of the framework presented in Section 2.

Figures 3a-d provide some explanation of this. Each figure represents the results of the six filter choices on a randomly selected data set from batch CSR×3TC. For each filter choice, the six charts within contain outlined and solid circles, with an Ω value above. The outlined circles delineate A in three (sometimes overlapping) parts, each being the outer limit[8] of a Gaussian shaped true cluster. The solid circles represent the detected clusters passed by the filter, shaded relatively from dark (low p-values) to light (high p-values); the shading is relative within each chart because Ω is only concerned with the relative p-values of different circles. Hence a shade in one chart may represent a different p-value to the same shade in a different chart. Note that more likely detected clusters overlap less likely ones, as per Figure 2. Note also that as $\lambda(s)$ is uniform in CSR datasets, Ω (and sensitivity) rewards correct detection of the centre of each true cluster much more than the periphery. Due to varying $\lambda(s)$, charts for TRENT data sets are harder to interpret and are not presented here.

It can be seen in Figures3a-d that filters F1, F2 and F4 generally pass fewer (and smaller) detected clusters than F3, F5 and F6. This explains the contrast between spatial sensitivity and PPV in the two groupings. On no occasion are the true clusters correctly delineated, but this is hardly surprising given that the SSS has only 300 points locations in total for each data set. As can be seen in Table 4, F6 performed unquestionably better in each batch, Ω wise, with the exception of TRENT×3TC where it is highest, but by a much smaller margin. This is counter intuitive, as F6 is the null option, i.e. no filtering. However, recall that Ω is the probability that one can correctly decide which of two randomly generated points lies inside the true clusters, based only on information provided by the detected clusters. If one views the unfiltered SSS output as a probability map of where the true clusters are more likely to be, rather than as a list of individually detected clusters, then it is hardly surprising that it provides the best source of information on spatial accuracy. However, if a more succinct list of detected clusters is desirable, then one should apply filters F1, F2 or F4, with filter F2 (no cluster centres in other clusters) performing best in terms of Ω in all four batches used in this study.

		Mean Ω over all data sets	Mean over all data sets sig. at 0.05	
			sensitivity	PPV
CSR×1TC	F1	0.7364 (±0.0037)	0.838245	0.596085
	F2	0.7659(±0.0040)	0.841861	0.591564
	F3	0.8684 (±0.0044)	0.877693	0.46603
	F4	0.7612 (±0.0039)	0.850334	0.594105
	F5	0.8725 (±0.0044)	0.882901	0.466492
	F6	0.8962 (±0.0044)	0.920411	0.380853
CSR×3TC	F1	0.7402 (±0.0028)	0.537794	0.572392
	F2	0.7704 (±0.0031)	0.545679	0.562416
	F3	0.8466 (±0.0030)	0.637829	0.388202
	F4	0.7654 (±0.0030)	0.548487	0.565982
	F5	0.8518 (±0.0029)	0.641853	0.387308
	F6	0.8631 (±0.0030)	0.681826	0.31248
TRENT×1TC	F1	0.7315 (±0.0038	0.858498	0.759631
	F2	0.7582 (±0.0041)	0.876176	0.710421
	F3	0.8746 (±0.0045)	0.947331	0.345964
	F4	0.7519 (±0.0040)	0.878359	0.73161
	F5	0.8779 (±0.0044)	0.948619	0.34458
	F6	0.8874 (±0.0045)	0.963441	0.264817
TRENT×3TC	F1	0.7124 (±0.0033)	0.695842	0.786179
	F2	0.7455 (±0.0036)	0.719602	0.735298
	F3	0.8430 (±0.0031)	0.841569	0.396941
	F4	0.7354 (±0.0035)	0.718791	0.76095
	F5	0.8477 (±0.0031)	0.84665	0.39312
	F6	0.8486 (±0.0032)	0.874754	0.30667

Table 5: Spatial accuracy results for the four batches, with output filters F1 to F6 applied.

5. Discussion and future directions

It would be useful, especially from the point of view of those wishing to conduct metastudies, if there were

[8]This is the loci of s where $rr(s) > 0.00001$. This limit is arbitrary, but varying this value by several decimal places either way has negligible effect on any of the measures.

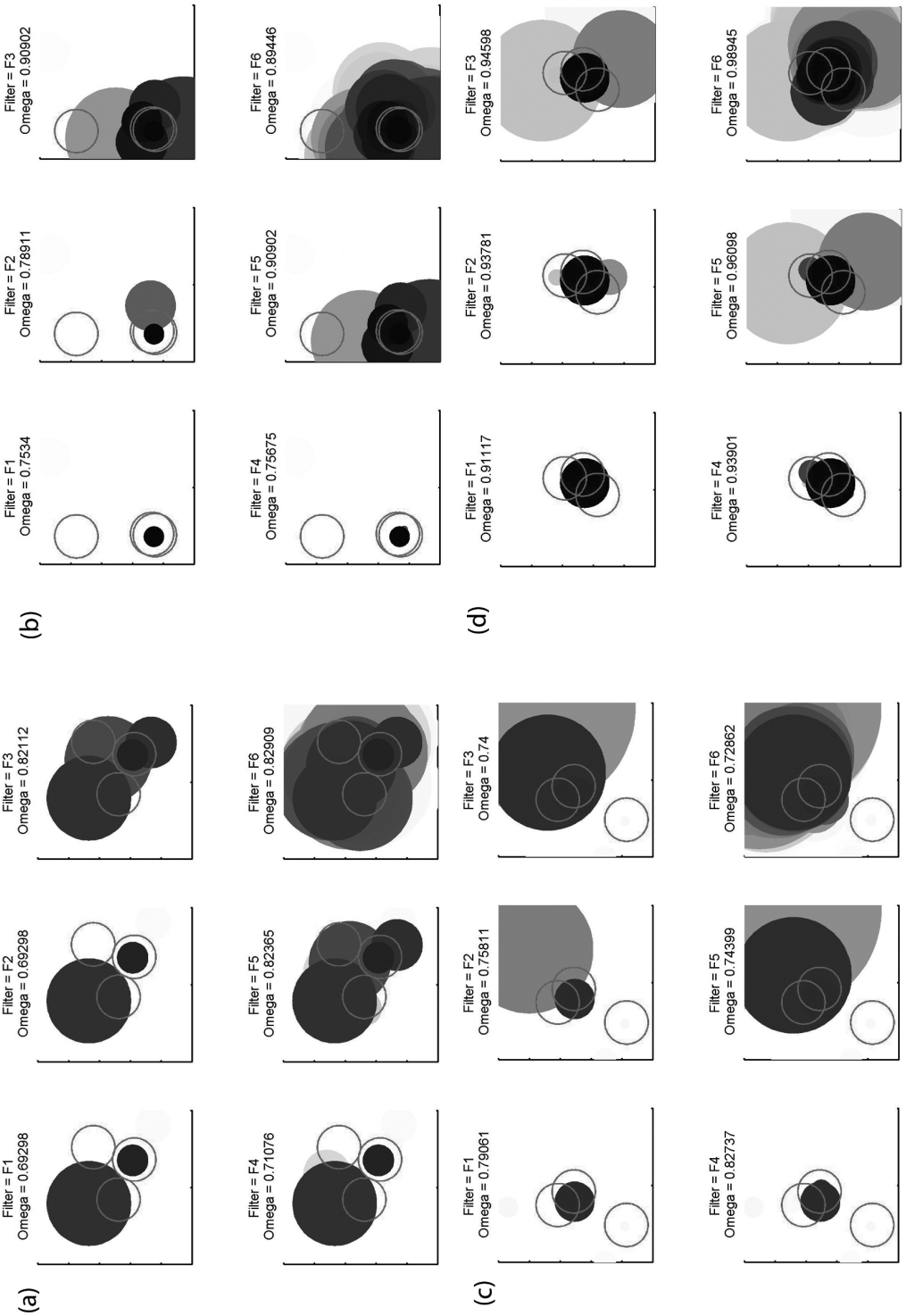

Figure 3: Examples of true (outlined) and detected (solid) clusters in four randomly selected data sets of batch CSRx3TC, using filter options F1 to F6. See text for details.

a universal definition of spatial accuracy for the SSS. Given the range of different measures, suited to different studies and different kinds of data, this seems unlikely. However a framework, such as the one presented here, may at least provide a means of comparing different methods, and help to avoid "reinvention of the wheel". The main limiting factor of the framework is that it is currently only suitable for measures of spatial, not spatio-temporal, accuracy. In studies where time can be considered as an extra spatial dimension, e.g. a one-off retrospective study, then the framework should be applicable. In contrast, with the detection of emerging clusters the direction and currency of time makes it different to space[9]. It may be feasible to extend this framework to cover spatio-temporal accuracy, and this could be a direction for future research.

The Ω measure presented in Section 3 provides a solution to the problem of arbitrarily specifying a significance threshold and the relative weighting of existing measures such as spatial sensitivity and PPV. As it fits within Levels 3 and 4 of the framework described in Section 2, Ω can be used with any combination of spatial support and data, not just the case/control data used in the example study. The chief drawback is that, for most existing studies, implementing Ω requires writing additional code and re-examining the SSS output files. If one is happy to specify a significance threshold, then easier to calculate (if somewhat cruder) single measures are available, based upon a similar 2AFC to Ω. Details available from the corresponding author.

Although limited in scope, the preliminary study of SaTScan™ filter options presented in Section 4 is, so far as we are aware, the first of its kind to be published. Despite being part of a particular software package, in some form or other these filters would be a natural part of any SSS implementation. The observation that Ω appears to be optimised by not applying a filter does not diminish their usefulness, but it does suggest that another means of presenting SSS output, beyond a simple list of detected clusters, could be beneficial. Studies in visualising SSS output already exist (e.g. Boscoe et al. (2003) Chen et al. (2008)), and one future research direction could be to investigate the spatial accuracy of different visualisation techniques, however that might be defined. We hope the material contained in this paper is of interest to those in the research community and welcome feedback.

[9]For the interested reader, an approach to measuring spatio-temporal accuracy (which treats time as something fundamentally different from space) is given in Fricker Jr. (2010)

6. Acknowledgements

We thank the Medical Research Council for funding Simon Read, the principal author, researcher and programmer on this project.

References

Boscoe FP, McLaughlin C, Schymura MJ and Kielb CL. Visualization of the spatial scan statistic using nested circles. Health & Place 2003;9(3):273-277.

Chen J, Roth R, Naito A, Lengerich E and MacEachren A. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. Int J Health Geogr 2008;7(1):57.

Fricker Jr. RD. Rejoinder: some methodological issues in biosurveillance. Retrieved July 16th, 2010, from: http://faculty.nps.edu/rdfricke/docs/Rejoinder%20-%20Issues%20in%20Biosurveillance.pdf. Monterey CA: Navel Postgraduate School; 2010.

Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley; 1966.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic curve. Radiology 1982;143(1):29-36.

Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival Data. Biometrics 2007;63(1):109-118.

Jacquez GM. Cluster morphology analysis. Spat Spattemporal Epidemiol 2009;1(1):19-29.

Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. Stat Med 2007;26(7):1594-1607.

Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Stat Med 2010;29(18):1910-1918.

Kulldorff M, Nagarwalla N. Spatial disease clusters, detection and inference. Statistics in Medicine 1995;14:799-810.

Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481-1496.

Kulldorff, M. SaTScan user guide for version 8.0. Cambridge MA: Harvard; 2009.

Neill DB. Expectation-based scan statistics for monitoring spatial time series data. Int J Forecast 2009;25(3):498-517.

Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. Am J Public Health 2006;96(11):2002-2008.

Que J, Tsui F-C, Espino J. A Z-Score based multi-level spatial clustering algorithm for the detection of disease outbreaks. Lect Notes Comput Sci 2008;5354:108-118.

Read S, Bath, PA, Willett P, Maheswaran R. A spatial accuracy assessment of an alternative circular scan method for Kulldorff's spatial scan statistic. In: Fairbairn D, editor. Proceedings of the GIS Research UK 17th Annual Conference, Durham University, 1st - 3rd April 2009. Durham: Durham University; 2009. p. 57-62.

Rennie JDM. Derivation of the F-Measure. retrieved July 16th, 2010, from: http://www.ai.mit.edu/people /jrennie/writing/fmeasure.pdf. Cambridge MA: MIT; 2004.

Savory D, Cox K, Emch M, Alemi F, Pattie D. Enhancing spatial detection accuracy for syndromic surveillance with street level incidence data. Int J Health Geogr 2010;9(1):1.

van Rijsbergen CJ. Information retrieval. London: Butterworths; 1979.

Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Med 2006;25(5):853-865.