Journal of Information Science
XX (X) pp. 1-13

DOI: 10.1177/016555150000000
jis.sagepub.com
SAGE

New developments in the spatial scan statistic

Read S, Bath PA, Willett P.
Information School, University of Sheffield

Maheswaran R.
School of Health and Related Research, University of Sheffield

Abstract
The quantity and variety of spatial data have increased over recent years, and the variety and sophistication of tools for analysing this type of data have also increased. One such tool is the spatial scan statistic, which is freely available (www.satscan.org) and has been the subject of much scholarly research since its introduction in 1995 due to its numerous applications in epidemiology, criminology and other fields. This paper provides readers with a non-technical introduction to the spatial scan statistic, together with an overview of associated research which focuses particularly on work conducted at the University of Sheffield's Information School, in collaboration with the School of Health and Related Research. This work falls into three main areas. Firstly, we provide an examination of the probability of obtaining false alerts when using the statistic, and ways in which this can be managed. Secondly, we describe the development of a definitive way of measuring the spatial accuracy of the statistic. Thirdly, and potentially the most important in terms of impact, we discuss a means of substantially increasing the detection capability of the statistic by placing a realistic constraint on the strength of any cluster which is likely to be present in the data. The paper also provides a discussion of potential future research directions.

Keywords
Spatial scan statistic; spatial epidemiology; geospatial data; case-control study.

1. Introduction

Through the efforts of quantitative geographers and statisticians, over the past half-century there have been dramatic developments in the field of spatial[1] data analysis. Simultaneously, information technology has facilitated the collation and processing of data in all walks of life, including spatial data. More recently, the inclusion of satellite positioning devices in cameras and mobile phones presents hitherto unimaginable opportunities for gathering spatial data, as well as justifiable concerns about how such data may be used. Interested readers can find discussions of some important topics in spatial data analysis in Maheswaran and Craglia (eds.) [1].

Within the field of spatial data analysis lies spatial cluster detection, which provides tools that find application in a variety of different areas, such as epidemiology, criminology and forestry and which provides the focus for this paper. Specifically, we describe recent work at the University of Sheffield on a popular method of spatial cluster detection: the Spatial Scan Statistic (SSS). The paper summarises these developments for a non-specialist audience in a form that avoids the use of mathematical expressions and overly technical language. Section 2 provides a primer which may be helpful for those new to handling spatial data, whilst Section 3 describes cluster detection and the SSS. Our studies are presented in Section 4, and the paper concludes with a discussion of potential future research directions. Further details of the research are provided by Read *et al*. [2-4].

[1] Although we only discuss spatial data in this paper, it should be noted that many of these concepts can be applied to spatio-temporal data.

2. A primer: basic concepts in spatial statistics

To discuss current research in spatial data analysis, it is helpful to be able to think about spatial data in the way that statisticians do. In this section, we include a primer which may be helpful for those new to the subject.

2.1. Location

By spatial data we mean any data set where one (or more) of the characteristics associated with each object or case is a *location*. This seems simple enough, but what we mean by location is not always obvious, and usually requires some thought.

When the subjects are immobile, such trees in a forestry study, location seems obvious. Even here, one must decide whether to give each subject a point location (e.g. GPS reading) or whether to divide the study region into areas (e.g. grid squares) and count the number of objects in each area. The former is called *point data*, the latter *areal data*; this is one of the most important demarcations in spatial statistics. When the objects of the research/cases are mobile, e.g. people, matters become considerably more complicated. Location can be fixed, for example, a residential address, or a workplace address, or the address of a shop where the subject obtained a particular good or service relevant to the study. Alternatively, the 'location' may be a route (e.g. a commute to work) rather than a fixed position [5]. It may also be a spatially vague location, e.g. "central London".

To simplify our discussion, in this paper we will only consider fixed point and areal locations, whilst bearing in mind that these basic principles can be extended to more complicated definitions of location.

2.2. Data quality

Data quality is a limiting factor in spatial data analysis, as it adds an additional element of "random noise" to spatial processes that inherently contain a high degree of randomness (see Section 2.5). Although we do not directly address this issue in our research into the SSS, it is important to be aware of spatial data quality in order to minimize any adverse effect it has on the veracity of results. To this end, Haining [6] presents a useful four-dimensional framework for spatial data quality issues:

1. *Accuracy*: the level of error within the variables obtained
2. *Resolution*: the level of detail to which locations can be specified
3. *Consistency*: compatibility in data between samples, and the suitability of the manner in which they are compared
4. *Completeness*: the presence of sufficient indicators, besides location, to complete a successful analysis

How this framework applies depends on the area of application. Accuracy, for example, may relate to the quality of geocoding, which can have a significant effect on the SSS [7]. Resolution may be related to confidentiality issues, such as when data are spatially aggregated to provide anonymity; this also can reduce the power of the SSS [8]. Consistency could, to give an epidemiological example, relate to differences in diagnosis criteria between hospitals. Completeness may be related to missing data; Kulldorf *et al.* showed that missing data can lead to an increased number of false alerts when used with the SSS.

These issues are so pervasive that it may be difficult to find data that do not have quality issues, in some regard. However, Mandle et al. make an important point, that even low quality data can be sometimes be useful if there are sufficient of them [9]; this is especially true of spatial statistics which usually require a much larger sample size than non-spatial statistics (see Section 2.5).

2.3. Stochastic processes

To understand best the material presented in this paper, it is helpful to think of data in the way that statisticians do: that outcomes we observe in the real world can often be usefully modelled by simple random processes. Such processes are not strictly random, rather they have a kind of 'guided' randomness, where some outcomes are more probable than others. This form of random process is called a ***stochastic process***, and the probability of each different outcome can usually be calculated by some mathematical formula. This allows us to use powerful mathematical tools to analyse our data, and make predictions about future data.

The guiding element of a stochastic process is provided by two types of values which appear in the mathematical formula: ***input variables***, which contain observable real-world values (e.g. population density) that influence the

probability of each outcome; and *parameters*, which are underlying values, typically unknown that also influence the process (e.g. disease risk). The number (or numbers) produced by the formula are the *output variables.*

Consider a simple example, illustrated in Figure 1.Here the incidence count of some hypothetical disease in some hypothetical town is modelled as a Poisson random variable [2], where the mean of the variable is specified as the at-risk population in the town multiplied by the disease risk in the town. If we have many years of disease incidence records for the town (the output variable), and the at-risk population in the town for each year can be estimated with reasonable accuracy (the input variable), then using this model we can work backwards to guess the level of disease risk (the parameter). We can also estimate the accuracy of our guess.

Now if we have two towns, say A and B, we can model the disease incidence counts in each town as separate stochastic processes, guessing the parameters separately for each process. We can then compare our guesses about disease risk in the two towns, and decide whether or not they are significantly different. What if towns A and B were close together, or far apart: would this affect our conclusions? We discuss these issues in the next section.

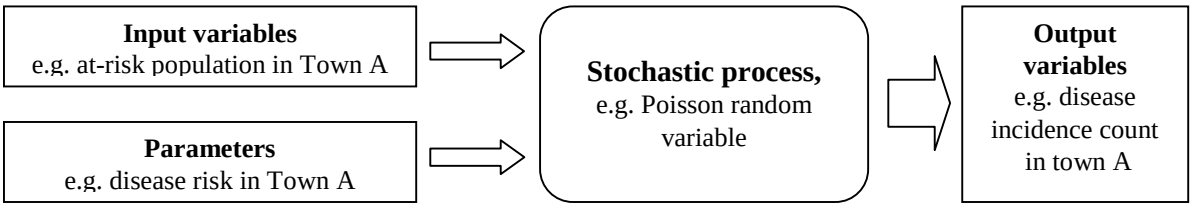

Figure 1. Symbolic representation of a stochastic process.

2.4. Stochastic processes in spatial data

How do stochastic processes relate to spatial data? To explain this, let us continue with the hypothetical examples given in Section 2.3 above. Rather than a single disease risk for a whole town, let us assume that we are interested in determining if there are variations in disease risk between different districts within a town. We could model each district separately, using the process above, and compare the resulting estimates of disease risk (i.e., what we did in Section 2.3 when considering towns A and B). However, most real world processes do not respect administrative boundaries. If district X has a high rate of disease, then the adjacent district Y may also have an increased rate too since they share overlapping neighbourhoods. Furthermore, changes in processes due to geographic location often occur gradually, as stated by Tobler's first law of geography [10], which may be paraphrased as: *close locations tend to be more related than distant locations*. By considering each district as a separate process, we are not taking account of the spatial relationship between districts.

For example, let us imagine that three neighbouring districts on the eastern side of town A have slightly elevated levels of estimated disease risk, none of which are statistically significant when considered separately. However, if these three districts are combined and modelled as a single process, then the elevated level of disease risk may well be statistically significant due to the increased sample size. This is a simple example of how taking location into account when analysing data can increase the power of the analysis; and this form of aggregation is a basic tool in spatial statistics. However, knowing which areas (or points) to combine, and being able to do this without compromising statistical integrity, is not a trivial task. In Section 3 we will explain a particular method of aggregation that has proved very popular in spatial analysis: the SSS; however, before so doing, we first explain some important issues in the next section.

2.5. Some important characteristics of spatial data

As discussed in Section 2.4, taking the spatial location of data into account can increase our ability to detect certain phenomena. However, it is important to understand the limitations of spatial data analysis. These are best explained using visual examples.

[2] A Poisson random variable generates integer numbers in a manner which realistically reflects the distribution of event counts in many real world processes, such as disease incidence, traffic flow, IT failures, call volumes to telephone lines, etc. To control the probability of this variable generating any particular number, one only need specify the mean average of the variable's output.

First, consider the two squares in Figure 2, each containing a spatial distribution of dots. Clearly there are patterns in each distribution: in some parts of the square there are no dots at all, in others the dots appear to fall along lines, usually curves. Now note the numbers on the left and bottom of each rectangle, from 0 to 100. Each dot in each rectangle was generated by selecting a pair of random numbers between 0 and 100, and using these as the horizontal and vertical coordinates. This is a process known as ***complete spatial randomness*** (often CSR for short). There is no underlying process responsible for generating the patterns we observe in Figure 1, they are ***random artefacts.*** Distinguishing between random artefacts and the patterns caused by genuine phenomena is the great challenge of spatial statistics, and it is what differentiates spatial statistics from descriptive spatial data analysis.

Second, consider the two diagrams in Figure 3. Each 'box' contains a distribution of dots, all of which lie in a plane parallel to the 'lid' of the box. The coordinates of each dot were generated using a stochastic process, controlled by a parameter whose value is represented by the surface shown at the bottom of the box: the height of the surface indicates how likely it is that a dot will occur over that part of the surface. For example 'peaks' in the surface will attract more dots than 'troughs'. Going back to our example of disease incidence, one could imagine this surface as representing the spatial variation in disease risk, with peaks representing areas of highest risk. Indeed, such a surface may be called a *risk surface*. If one compares the two diagrams in Figure 3, one will notice they are identical, save for a large peak at the centre of the right-hand surface. Now consider the distribution of dots in the two diagrams. Does the distribution of dots in the right-hand diagram clearly allude to the presence of the peak in risk at the centre? Possibly, in the sense there is some gravitation of cases towards the centre, but it is certainly not obvious from the dots what the underlying risk surface is. The number of data in each diagram (approximately 50 dots) is clearly an inadequate number for such a spatial analysis, despite being a perfectly respectable sample size for many non-spatial statistical purposes (for example, Cohen suggested that for a medium or large effect size[3] in multiple linear regression with two variables, a sample size between 30 and 67 gives 80% power at the standard significance level of 0.05 [11]). The reason for this is the *curse of dimensionality* [12], which means that as one adds extra dimensions to one's analysis, the sample size required grows exponentially. Put simply, spatial statistical analysis is best reserved for large data sets.

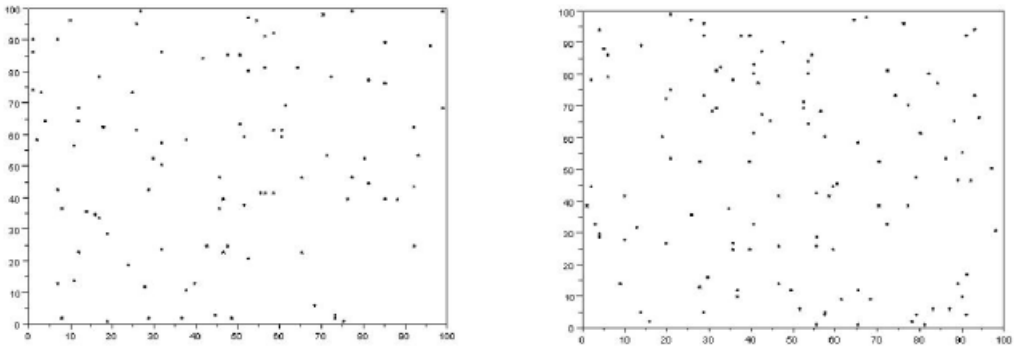

Figure 2. Two homogeneously random distributions of points.

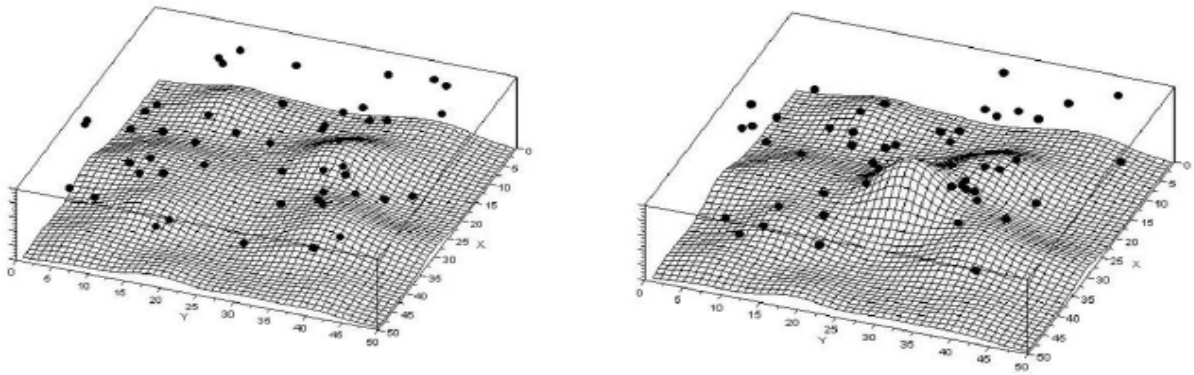

Figure 3. Example of control data points (left) and case data points (right), with differing risk surfaces shown.

[3] Meaning a value of 0.15 (medium) and 0.35 (large) for the value of $R^2/(1-R^2)$, this expression being equivalent to the sum of squares due to the regression divided by the sum of squares due to the error terms. See [11] for full details.

3. Research background

Having provided a general introduction to spatial statistics, we now focus on a specific branch of the subject, cluster detection, and then on a particular method of cluster detection, the SSS.

3.1. Cluster detection

In Section 2.5 we explained that when using statistics it is very difficult to interpret *specific* patterns in spatially distributed data. That is, while it is always possible to observe detailed patterns in spatial data (such as can be seen in Figure 2) these are likely to be random artefacts without strong evidence to the contrary. However, it is underlying trends in the distribution of spatial data that usually concern spatial statisticians. The simplest of such trends is the tendency for results to cluster, i.e. to spatially aggregate. Ironically, despite being an intuitively simple concept, there is no generally accepted definition of what spatial clustering is and, consequently, no generally accepted optimum method of spatial cluster detection. There are, in fact, numerous different spatial cluster detection tests, with very little objective evidence for their relative efficacy.

There is, however, a common, fairly unambiguous three-part categorization of cluster detection techniques: global, focused and local.

Global refers to a method that detects the presence of clusters somewhere within the study region, but does not point to where the coalescence is. One example is Ripley's K-function [13], which examines the cumulative distribution of distances between point locations, and another is Cuzick and Edward's test [14].

Focussed refers to a method which tests for the presence of a cluster in some part of the study region which has been identified *a priori*, as a result of some external hypothesis. An example might be the area around some putative pollution source. Exclusively focussed methods are less common than global and local methods, but one example is Stone's test [15].

Local refers to methods that, as with global, attempt to identify clusters anywhere within the study region, but also attempt to specify the location(s) concerned. The SSS, which we discuss in the next section, is a prime example of a local method. Other well known examples are Openshaw's Geographic Analysis Machine [16] and Turnbull's test [17] (which are both predecessors of SSS), as well as Besag and Newell's test[18].

3.2. The spatial scan statistic

The SSS was first introduced by Kulldorff [19, 20] as an extension of Naus' one-dimensional [21] and two-dimensional scan statistic [22] and has been the subject of much recent study.. In this section we point the reader to a representative (but by no means exhaustive) selection of this research.

As mentioned in Section 3.1, the SSS is local cluster detection method, the aim of which is to detect the presence and location of clusters within spatially distributed data[4]. However, the term 'spatial scan statistic' does not refer to a single mathematical formula; rather, it refers to a collection of formulae, each suitable for use with different types of data, but all sharing a similar method of application. In its broadest sense, the term also refers to the computer algorithm which performs the application. In this sense it has a parallel with the word SaTScan™, which is the name of the free software package (www.satscan.org), overseen by Kulldorff since 1997, which provides a user friendly interface for many of the different versions of the SSS.

Taking the term 'spatial scan statistic' in its broadest sense then, we can consider the algorithm as having two main parts, each of which contains several steps:

1) Delineating and evaluating a large, manifoldly overlapping, set of potential cluster locations:
   a) Drawing a series of *scan windows* over the study region, of various sizes (and possibly different shapes).
   b) Assessing the probability (in this sense called the *likelihood*) that the data contained within each scan window is, in some sense, a cluster.
   c) Ranking the likelihood values from each scan window, taking particular note of the *most likely scan window to contain a cluster*
2) Calculating the probability that the strongest potential cluster is a random artefact. Typically (after Dwass [23]), this involves:
   a) Spatially randomising the data in a manner that removes any possibility of a genuine cluster being present, and repeating the first part of the test a large number of times, each time recording the likelihood value of the most likely scan window

[4] It can also be used to detect temporal and/or spatio-temporal clusters, although we do not address this here.

b) Ranking the likelihood value of the 'real' most likely scan window, obtained in part 1, step (c), amongst the likelihood values recorded in the step (a) above.
c) If the 'real' value is amongst, say, the top 5% of these values, we may come to the conclusion that the cluster represented by the 'real' most likely scan window is too prominent to be a random artefact, and therefore decide that there is a cluster present in the data, whose location coincides with the most likely scan window.
d) We may also wish to consider other scan windows, with 'real' likelihood values close to that of the most likely scan window. These may represent secondary clusters, although we may chose to disregard those that overlap considerably with the most likely scan window.

For the sake of accuracy, the description above is couched in quasi-technical language, and an analogy may help the reader to grasp exactly what the SSS does. Let us imagine ourselves as a baker, cutting raisin cookies from a sheet of dough that our assistant has prepared. Looking at the distribution of raisins within the dough, we suspect the assistant has been lazy, and not mixed the raisins thoroughly into the dough, resulting in clusters of raisins forming at one or more places on the dough sheet. The assistant denies this, claiming the raisins were thoroughly mixed, and any clusters are just random artefacts (as we saw in Figure 2). Before admonishing the assistant, we decide to put her/his claim to the test. By moving a cookie-cutter over the sheet in a systematic way, we can count the raisins appearing within the cookie-cutter "scan window", and take note of the highest count recorded. This is analogous to part 1 of the algorithm. Next, we gather up the dough, mix it thoroughly so we are certain the raisins are randomly distributed, and re-roll a new sheet. We then perform the same process with the cookie cutter, and record the highest number of raisins observed within any scan window. This re-rolling and re-scanning is repeated many times. If the highest count in the assistant-prepared dough is exceeded by only a small proportion (say, 5%) of the highest counts in the re-mixed dough, then the likelihood that the assistant was telling the truth is quite small. Moreover, we can point to the part of the original dough which contains the evidence of the deceit. This is analogous to part 2 of the algorithm. When one replaces dough with a geospatial area, and raisins with cancer cases, one can see how such an algorithm can have far less trivial applications. And of course, all that work of "cookie-cutting" and "rerolling" is done by computer!

Most of the research conducted into the SSS since 1997 has focussed on part 1 of the algorithm, specifically steps 1(a) or 1(b).Following the original proposal of a strictly circular scan window [19], others have proposed different schemes for different ways of selecting scan windows, e.g. elliptical[24], echelon-based [25] or free-form[26].Following Kulldorff's derivation of the Bernoulli and Poisson SSSs (for case-control point data and areal count data, respectively), many other versions have been proposed, e.g. for normally distributed data[27], survival data [28], and ordinal data [29].

Less attention has been paid to part 2 of the algorithm, from which much of the computational cost of the process arises. The method of repeatedly randomising data and recalculating a statistic is commonly called Monte Carlo testing (the name alluding to the element of chance involved), and there is a large body of technical research on this method, summarised in books such as that by Lui [30]. However, due to the combinatorial complexities involved in analysing spatial data, it is not straightforward to apply such methods directly to the SSS. One exception is sequential Monte Carlo (e.g. see [31]), which has already been implemented in SaTScan™. Another method specific to the SSS is *Gumbel approximation*, as proposed by Abrams *et al*. [32] and discussed in more detail in Section 4.2. A totally different approach to part 2 is the Bayesian approach suggested by Neill [33], which has connection to the work in Section 4.3.

4. Recent work in Sheffield

This section presents an overview of our recent work on the SSS [2-4]. Section 4.1 concerns how the performance of the SSS is measured, and explains a novel measure we have developed that has several advantages over existing ones. Section 4.2 concerns a new method for improving the computational performance of the SSS, and explains a study we have conducted which significantly extends the understanding of the method. Finally, Section 4.3 presents a novel version of the SSS and explains some of the advantages this delivers.

4.1. Measuring spatial accuracy

This section addresses the measurement of spatial accuracy for the SSS. Before discussing this, we must briefly explain the broader context of performance measurement for the statistic. As mentioned in Section 3.1, the SSS is a local cluster detection test. This means it has two objectives: to detect the presence of a cluster; and to detect the location of

a cluster. Correspondingly, there are two performance measures associated with the SSS, relating to each of these objectives, respectively: Power (the term being used here somewhat loosely[5]); and spatial accuracy

Measuring power is relatively straightforward, and most studies of the SSS include some form of power study using benchmark data. A classic example is Song and Kulldorff's study [34]. If one has a large number of benchmark datasets into which an artificial cluster has been injected, and a large number of datasets in which it is safe to assume no cluster is present, then one can apply the SSS to every dataset and record the results in a 2x2 table such as that shown in Table 1. From this one can obtain standard test performance measures such as *sensitivity* (a/(a+c)), *specificity*(d/(b+d)), *positive predictive value*(a/(a+b)) and *negative predictive value*(d/(c+d)).Sensitivity is equivalent to recall in information retrieval research and positive predictive value is equivalent to precision [35].

Table 1. The standard contingency table for measuring test performance

	Cluster present in dataset	**Cluster not present in dataset**
Test positive	Number of true positives (a)	Number of false positives (b)
Test negative	Number of false negatives (c)	Number of true negatives (d)

Measuring spatial accuracy is not so straightforward. This is principally because, unlike power, there is no widely accepted definition of spatial accuracy. Many studies of the SSS do not consider spatial accuracy at all, and those that do use *ad-hoc* measures that, although entirely appropriate for the study concerned, are not comparable with the measures used in other studies. This makes meta-studies particularly difficult.

As mentioned in Section 3.2, the SSS is an umbrella term covering a variety of different statistics for application to a variety of different types of data; it is therefore hardly surprising that there is no commonly agreed definition of spatial accuracy, or method of measuring it. However, many studies share similarities in the way they measure spatial accuracy, and breaking down the measurement process into different elements makes it easier to compare different measures. For this purpose, in a recent paper [2], we introduced a five-level framework for measurements of spatial accuracy, to facilitate the comparison and hybridisation of different measures.

The framework presented in [2] is necessarily technical, and the details are somewhat beyond the scope of this paper. However, when viewed within this framework, two common limitations of existing spatial accuracy measures become apparent: they require the specification of an arbitrary detection threshold; and they produce two or more values. This is perhaps easiest to understand by considering Table 2, which is the spatial equivalent of Table 1.

Firstly one can generate many measures based on such a contingency table, similar to those described above, but these measures frequently used in pairs (such as recall and precision are in information retrieval). Secondly, as the SSS produces many candidate clusters with varying degrees of statistical significance, building such a contingency table necessitates choosing an arbitrary significance threshold, dictating what counts as the 'detected cluster'. This is analogous to the situation in information retrieval, where the application of a threshold to a ranking permits the calculation of both the recall and the precision.

Table 2. A contingency table for measuring spatial accuracy

	Study area inside actual cluster	**Study area outside actual clusters**
Study area inside detected cluster	Area of cluster correctly detected	Area outside cluster incorrectly detected as being part of a cluster
Study area outside detected cluster	Area of cluster remaining undetected	Area outside cluster correctly assumed not to be within a cluster

The first of these limitations is important because it means that the resulting spatial accuracy measure is also arbitrary. To explain, the detection threshold is typically the significance threshold, which is the maximum level one is prepared

[5] Strictly speaking, *power*, or even more strictly speaking *empirical power*, is a synonym for *sensitivity* as it is defined here.

to accept for the probability[6] that the test has incorrectly detecting a cluster when in fact none is present. A common choice of value for the significance threshold is 0.05, but it is not the only choice by any means, and one should always bear in mind that to use 0.05 is an arbitrary choice. For example, at one detection threshold 0.05, version A of the SSS might perform better than version B, and yet perform worse at other detection thresholds, e.g. 0.01, or 0.001. With existing measures, all one can do is use a variety of different detection thresholds and hope that the results are representative.

The second of the two limitations is important if one is seeking to rank objectively different versions of the SSS in terms of spatial accuracy. Many measures produce two values: one representing the proportion of the cluster that has been correctly identified as such, and another representing the proportion of the study region outside the cluster that has been correctly identified as not being part of the cluster. One version of the statistic may perform better in respect of the former, whilst another performs better in respect of the latter. How does one then say which one has performed better? Of course, one can combine the two measures using some formula, but the choice of this formula is unavoidably arbitrary and so, therefore, is the result.

Within the context of the framework presented in [2], we have developed an entirely new measure of spatial accuracy that overcomes both of these limitations, insofar as it does not require the specification of a detection threshold, or any other arbitrary parameter, and it produces a single value that has a tangible real-world meaning. We have provisionally called this measure Ω (pronounced 'omega'). For a technical definition of Ω we refer the reader to [2], however Ω also has a straightforward intuitive definition, which we present here.

Consider a spatial study containing one or more clusters, and consider two randomly chosen data from the study, one having a location somewhere inside the cluster(s), the other having a location somewhere in the study region outside the cluster(s). Imagine that the version of the SSS in question is applied to this data set. Ω is simply the probability that a rational observer, using only the information provided by the output of the SSS, can correctly determine which datum is which. Being a probability, Ω naturally has a value between 0 and 1, with 1 representing perfect spatial accuracy: i.e. the test has correctly ranked every point within the true cluster location over and above every point outside the true cluster location. An Ω value of 0 is the exact converse, i.e., perfect spatial inaccuracy. Ironically, Ω = 0 is not a poor result, as one can simply invert the output of the SSS to achieve Ω=1. In practice, an Ω value of 0.5 is the worst outcome, as it means the test has provided no useful information as to the spatial location of the cluster; in which case one might as well have tossed a coin to decide which datum is which.

An example application of the Ω measure is given in [2], where it is used to examine the difference among the six different ways in which the SaTScan™ software handles the reporting of secondary clusters (i.e. clusters that are unlikely to be random artefacts but that are not the most likely potential cluster). The Ω measure also finds application in the work discussed in Section 4.3, where its unique qualities make it ideal for the study described.

In terms of the volume of literature that has been reported to date, the measurement of the SSS comes a long way behind the development of new versions of the statistic. Indeed, so far as we are aware this is the first attempt to treat the measurement of spatial accuracy for the SSS as a subject worthy of consideration for its own sake. We believe that this work is important, not only because it introduces some intellectual rigour into a largely overlooked subject, but because a better understanding of how to measure the performance of the SSS is key to understanding how to improve it; after all, one cannot make something better unless one is clear about what 'better' actually means.

4.2. Gumbel approximation

This section concerns the efficacy of a new method for reducing the computational cost of the spatial statistic algorithm: we first describe briefly this new method (called Gumbel approximation) and we then explain extensions to the method.

As discussed in Section 3.2, the established SSS algorithm uses Monte Carlo simulation to estimate the probability of incorrectly detecting a cluster when none is present. This probability is often simply called the ***p-value*** of the test. Calculating the p-value involves repeating the test a large number of times (typically 999), which greatly increases the computation expense. Although negligible for small datasets (e.g. several minutes for a few hundred data points, running on 2GHz desktop machine), the computation time increases quadratically[7] with the size of the data. This means that for large datasets, the computational expense can become prohibitive, especially for routine scanning of data,

[6] This is equal to 1-specificity, i.e. b/(b+d), referring to Table 1.

[7] A quadratic relationship is one of the form $y=x^2$, meaning that as x (e.g. the size of the data) increases, y (e.g. the computation time) increases considerably more.

where the computation needs to be repeated at regular intervals. Furthermore, there is a computational cost associated with the accuracy of the p-value, as every extra decimal place requires up to a ten-fold increase in computational time.

A method of reducing computational expense, whilst maintaining accuracy, was proposed by Abrams *et al*. [32].This makes use of fact that the SSS is a maximum value (see Section 3.2), by assuming that it follows the Gumbel distribution, a type of statistical distribution that is used for modelling extreme values [36]. This process involves running a Monte Carlo test with a relatively small number of iterations to obtain a rough sample of the true distribution of maximum likelihood values (again, see Section 3.2). A Gumbel distribution is then 'fitted' to this rough sample, and it is then this Gumbel distribution, rather than the actual values produced by the Monte Carlo simulation, that is used to estimate the p-value. For brevity we refer to this process as *Gumbel approximation.* The rationale for using Gumbel approximation is that, in principle, it reduces the amount of random variation in the p-value, making it substantially more accurate for little additional expense.

A limitation of a previous study by Abrams et al [32] is that it was based on a relatively small number of datasets, and concentrated on the Poisson version of the SSS (which is used for areal data). We chose to complement the work of Abrams *et al*. by conducting an extensive set of benchmark tests on the application of Gumbel approximation to the Bernoulli version of the SSS (this version is discussed in more detail in Section 4.3). Our study found the following:

1. Gumbel approximation produces substantially more accurate p-value estimates for the Bernoulli SSS than the conventional Monte Carlo method, for a given number of iterations. This confirms the previous finding [32].
2. Gumbel approximation tends to produce slightly more extreme p-value estimates than the conventional method for the Bernoulli SSS. This means it has lower specificity (see Section 4.1) than the conventional method making it more likely to generate false alerts.
3. In circumstances when Gumbel approximation has lower specificity for the Bernoulli SSS, it also has correspondingly higher sensitivity than the conventional method, so the overall detection capability is not impaired.
4. Over and above any reduction in computational expense, Gumbel approximation is superior to the conventional method for estimating very low p-values (e.g. < 0.0001).

Point 1 is encouraging, as it provides additional support for the recent introduction of Gumbel approximation, as an option, into the SaTScan™ software. However, points 2 and 3 reveal that Gumbel approximation is not necessarily the optimal choice for every application of the SSS, as some consideration needs to be given to an, albeit slight, increase in false alerts that can ensue from using it. For example, Fienberg and Shmueli [37] noted that epidemiological applications of the SSS required a high significance threshold to achieve a reasonably high level of sensitivity, resulting in a correspondingly high false alert rate. If the cost of a false alert is non-trivial (in terms of staff time and/or resources) then even a small increase in the false alert rate has a financial cost. For example, the Bernoulli SSS was part of the system designed to provide early warning of an outbreak of West Nile virus by monitoring reports of bird deaths; here the follow-up to an alert was the collection and examination of mosquitoes, which is labour intensive and thus costly [38]. So, for large datasets in certain applications it may be more cost effective to find additional computational resources to run the conventional method, rather than incur an increased probability of false alerts. Obviously, this depends on the specific application.

Point 4 applies to applications where evidence of clustering is very strong, but correspondingly a very high level of certainty is required to assume a cluster is present. In such applications, Gumbel approximation should be used in preference to the conventional method, since its tendency to overestimate p-values is more than counteracted by the increased accuracy of Gumbel approximation for very small p-values.Point 4 may be somewhat theoretical, as the use of such low p-value thresholds in spatial epidemiology have not yet been reported. However there are examples of p-value thresholds close to this, such as the study by Souris et al. [39] where a threshold of α=0.001 (0.1%) was stipulated.

Note that points 2 and 3 stem directly from a numerical characteristic of the Bernoulli SSS, originally identified in one of our early papers [40]. Although this characteristic also applies, in theory, to the Poisson SSS, Gumbel approximation for the Poisson version is much less likely, in practice, to produce an increase in false alerts. However, the space-time permutation scan statistic [41], for which Gumbel approximation is available in SaTScan™, shares this characteristic and *is* likely to have an increased false alert rate when it is used. These points are discussed in detail by Read *et al*. [3].

4.3. Beta-Bernoulli spatial scan statistic

In this last section, we detail a new version of the SSS which we have derived from first principles as an alternative to the existing Bernoulli version, which is useful for spatial (or spatio-temporal) case-control studies. If our findings are

correct then our beta-Bernoulli SSS gives substantially improved detection capability and spatial accuracy over the Bernoulli version. To explain the beta-Bernoulli version, it is necessary to understand something of how the existing Bernoulli version works, and exactly what is meant by a spatial case control study.

First, consider Figure 3. As well as showing how two similar risk surfaces can give rise to very different distributions of events (represented here by points), the two sides of this figure also provide an illustration of how a spatial case-control study works. If we consider the points on the left-hand side of Figure 3 to be *controls*, and those on the right-hand side to be *cases*, then the aim of a spatial case control study is to determine whether the underlying risk surface for the cases is significantly different from that which underlies the controls. Put another way, one is trying to identify areas within the study region where cases are more (or conversely, less) likely to occur than controls. One uses controls to account for clustering which *we already know to be present*; thus we are only looking for clustering in the cases which is *not* explained by the clustering of the controls.

The best way to understand this is by example, perhaps the best being Diggle's classic study of larynx cancer and lung cancer around a waste incinerator in North West England [42]. When viewed on a map, the residential location of lung cancer patients tends to cluster, chiefly due to spatial variations in population density, but also due to spatial variations in risk factors, such as tobacco consumption. Larynx cancer shares many risk factors with lung cancer, so when viewed on a map we would expect larynx cancer cases to cluster in similar locations to those for lung cancer. However, according to [42], larynx cancer is presumed to be more closely related to air pollution than lung cancer. So, if larynx cancer cases tend to cluster more strongly than lung cancer cases in the area downwind of a waste incinerator, this may be evidence of a link between the incinerator and larynx cancer.

The methodology used in [42] predates the Bernoulli SSS, but nonetheless the study emphasises that it is the spatial distribution of cases, relative to that of controls, that is of interest. As described in Section 3.2, the SSS places many thousands of scan windows over the study region and applies the statistic to each one in order to assess the likelihood of that scan window being the location of a cluster. For case control data, the Bernoulli SSS counts the number of cases and controls within each scan window, and compares this to the number of cases and controls in the rest of the study region.

To understand how the comparison is made, we need to leave the concrete world behind and enter the abstract world of probability models. Leaving aside all the complexity of how the cases and controls in our study came into existence, imagine that the locations of each point in the study are pre-ordained, and that the status of each point (i.e. case or control) is simply the outcome of an unfair coin toss, the coin landing "case" side up with a certain probability and "control" side up with a certain probability (the two probabilities not being equal, hence it being an unfair coin). Such a two-outcome test is called a *Bernoulli trial*, hence the name of the statistic. For each scan window, the statistic calculates the ratio of two things:

1. The likelihood that the unfair coin used to decide the status of the points inside the scan window is *different* to the unfair coin used for the rest of the study region. By different, we mean the probabilities of landing on each side are different (but still not necessarily equal). Typically, it is an increased probability of a case that we are interested in.
2. The likelihood that the *same* unfair coin was used to decide the status of all the points in the study region (i.e. both inside and outside the scan window).

The first point is analogous to saying that the scan window is the location of a cluster of cases, relative to controls. The second is analogous to saying there is no clustering of cases relative to controls (we will use the term *null hypothesis* to refer to this statement). As described in Section 3.2, the remainder of the SSS algorithm then calculates the probability that the null hypothesis is correct, and if this probability is very low (say < 0.05) we may wish to assume a cluster is present at the scan window where the first likelihood above is highest.

However, no limitation is placed on the assumptions made about the coin in point 1 above. If one has a study region with 100 cases and 200 controls, and therein one finds a scan window containing 5 cases and only 1 control, the Bernoulli SSS, following the probability model above, assumes that outside the scan window the odds of a point being a case are 1:2, whereas inside they are5:1. The problem is that in real applications, such as epidemiology where one talks about the odds of an individual having a particular disease, one simply does not find such dramatic spatial variation in odds. Although we emphasise it is not yet proven, we think this may lead the Bernoulli SSS to promote small random artefacts (see Section 2.5) where the odds are unfeasibly high, ahead of larger clusters with lower, but more realistic, odds.

Our beta-Bernoulli SSS uses the same probability model, but places realistic constraints on the difference in the odds (called the *odds ratio*) inside and outside the scan window. It does this by taking into account the probability of the

data point being a case when calculating the likelihoods in bullet-points 1 and 2 above, more specifically it takes into account the ***probability that the*** probability has a certain value, this being known as a ***prior probability***[8]. For a Bernoulli trial, this prior probability follows a distribution known as the ***beta distribution***, hence the name of the statistic. We have developed a novel and time-efficient method for calculating the characteristics of this beta distribution from assumptions about the mean and variance of the odds ratio, which can be specified by the user and given values realistic for the application concerned (see [4] for more details).

We have presented empirical evidence that the beta-Bernoulli version outperforms the Bernoulli version in the detection of a variety of different shapes and sizes of cluster [4, 43]. Though abstract in nature, there is no reason why the findings should not be generalisable to more specific scenarios, such as the detection of spatial clusters in cancer incidence or child mortality. Moreover, we have discovered that to obtain this improvement one only need specify vaguely realistic values for the mean and variance of the odds ratio (meaning expert judgement is not required). The performance improvement is observed in terms of both spatial accuracy [4] and detection capability [43]. Although the latter is only a preliminary study, our results suggest that, for a given false alert rate, the sensitivity of the beta-Bernoulli SSS is in the order of 10-20% higher than that of the Bernoulli version. We stress these require further confirmation, but if correct then this represents a major improvement to a widely accepted technique in spatial statistics.

The only downside that we are so far aware of is that beta-Bernoulli SSS requires slightly more computation resources than the Bernoulli. It is difficult to quantify the time increase as it is proportional to the number of points in the study, but independent of the number of Monte Carlo iterations as the additional calculations only need to be performed once for any given dataset. As Monte Carlo testing is the main source of computational expense in the algorithm, the increase is therefore very modest, even for large datasets.

5. Conclusions

The SSS is one of the most important and most widely used tools for the important, fast-growing field of spatial data analysis. In this paper, we have provided a non-technical overview of research in the University of Sheffield that seeks to develop further SSS. Specifically, we have discussed the measurement of spatial accuracy, and the use of Gumbel approximation and of the beta-Bernouilli distribution to improve the efficiency and the effectiveness, respectively, of the method.

We believe the work outlined in this paper has the potential to make a significant impact to the field of research into the SSS, and thereby the wider field of spatial data analysis.

There are many potential directions which future research might take. These include: extending the studies in Sections 4.2 and 4.3 to different versions of the SSS; extending all the studies to spatio-temporal, as well as spatial, studies; and examining the theoretical properties of the beta-Bernoulli SSS. The first two points would be relatively straightforward, but require a significant amount of additional programming work. The purpose of the third point would be to seek mathematical proof of the improvements observed using the beta-Bernoulli SSS, providing confirmatory evidence for the empirical evidence reported here.

6. Acknowledgements

We thank the Medical Research Council for funding Simon Read, Martin Kulldorff of Harvard University for his kind and constructive feedback, particularly on the material in Section 4.2, and Bob Haining of Cambridge University and John Holliday of the University of Sheffield for helpful comments. Lastly, none of this work would have been possible without access to the University of Sheffield's high performance computing cluster[9], and we are grateful to Dr. Anthony Brookfield for keeping things running smoothly.

[8] We are not the first to use a prior probability in conjunction with the spatial scan statistic, as [33] used a gamma prior for the Poisson version. However [33] rely on historic data whereas we require only vaguely realistic assumptions about the odds ratio.

[9] http://www.shef.ac.uk/wrgrid/iceberg

7. References

[1] Maheswaran R, Craglia M (eds). GIS in Public Health Practice. Boca Raton: CRC Press 2004.

[2] Read S, Bath, PA, Willett, P and Maheswaran R. Measuring the spatial accuracy of the spatial scan statistic. *Spatial and Spatio-temporal Epidemiology* 2011; 2(2): 68-79.

[3] Read S, Bath PA, Willett P and Maheswaran R. A study on the use of Gumbel approximation with the Bernoulli spatial scan statistic. [submitted for publication]

[4] Read S, Bath PA, Willett P and Maheswaran R. A spatial accuracy assessment of a beta-Bernoulli spatial scan statistic. [submitted for publication]

[5] Duczmal L and Buckeridge DL. A workflow spatial scan statistic. *Statistics in Medicine* 2006; 25(5): 743-754.

[6] Haining, RP. *Spatial data analysis: theory and practice*. Cambridge: Cambridge University Press, 2003.

[7] DeLuca PF and Kanaroglou PS. Effects of alternative point pattern geocoding procedures on first and second order statistical measures. *Journal of Spatial Science* 2008; 53(1): 131-141.

[8] Olson KL, Grannis SJ and Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health* 2006; 96(11): 2002-2008.

[9] Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin JA, Gesteland PH, Treadwell T, Koski E, Hutwagner L, Buckeridge DL, Aller RD and Grannis S. Implementing syndromic surveillance: a practical guide informed by the early experience. *Journal of the American Medical Informatics Association* 2004; 11(2): 141-150.

[10] Tobler WR. A computer movie simulating urban growth in the Detroit region. *Economic Geography (supplement: Proceedings. International Geographical Union. Commission on quantitative methods)* 1970; 46: 234-240.

[11] Cohen J. A power primer. *Psychological Bulletin* 1992; 112(1): 155–159.

[12] Bellman, RE. *Dynamic programming.* Princeton: Princeton University Press, 1957.

[13] Ripley BD. *Spatial statistics*. New Jersey: Wiley, 1981.

[14] Cuzick J and Edwards R. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B (Methodological)* 1990; 52(1): 73-104.

[15] Stone RA. Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine* 1988; 7(6): 649-660.

[16] Openshaw S, Charlton M, Mymer C and Craft AW. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1987; 1(4): 335-358.

[17] Turnbull BW, Iwano EJ, Burnett, WS, Howe, HL and Clark LC. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 1990; 132(supp1): 136-143.

[18] Besag J and Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 1991; 154(1): 143-155.

[19] Kulldorff, M and Nagarwalla, N. Spatial disease clusters, detection and inference. *Statistics in Medicine* 1995; 14(8): 799-810.

[20] Kulldorff, M. A Spatial scan statistic. *Communications in statistics - theory and methods* 1997; 26(6): 1481-1496

[21] Naus, JI. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association* 1965; 60(310): 532-538.

[22] Naus, JI. Clustering of random points in two dimensions. *Biometrika*, 1965; 52(1): 263-267.

[23] Dwass M. Modified randomization tests for non-parametric hypotheses. *The Annals of Mathematical Statistics* 1957; 28(1): 181-187.

[24] Kulldorff M, Huang L, Pickle L and Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; 25(22): 3929-3943.

[25] Patil GP and Taillie, C. Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science* 2003; 18(4): 457-465.

[26] Tango T and Takahashi K. A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters. *International Journal of Health Geographics* 2005; 4(1): 11.

[27] Kulldorff M, Huang L and Konty K. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 2009; 8(1): 58.

[28] Huang L, Kulldorff M and Gregorio D. A Spatial Scan Statistic for Survival Data. *Biometrics* 2007; 63(1): 109-118.

[29] Jung I, Kulldorff M and Klassen AC. A spatial scan statistic for ordinal data. *Statistics in Medicine* 2007; 26(7): 1594-1607.

[30] Lui JS. *Monte Carlo strategies in scientific computing.* Berlin: Springer-Verlag, 2004.

[31] Silva I, Assunção R and Costa M. Power of the Sequential Monte Carlo Test. *Sequential Analysis* 2009; 28(2): 163-174.

[32] Abrams, AM, Kleinman, K and Kulldorff, M. Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics* 2010; 9(1): 61.

[33] Neill DB, Moore AW and Cooper GF. A Bayesian spatial scan statistic. In: Weiss Y (ed.) Advances in Neural Information Processing Systems, volume 18. Cambridge, MA: MIT Press, 2006, pp. 1003-1010.

[34] Song C and Kulldorff M. Power evaluation of disease clustering tests. *International Journal of Health Geographics* 2003; 2(1): 9.

[35] Bath PA. Data mining in health and medical information. *Annual Review of Information Science and Technology*. 2004; 38:331-369.

[36] Gumbel EJ. *Statistics of extremes.* New York: Columbia University Press, 1967.

[37] Fienberg SE and Shmueli G. Statistical issues and challenges associated with rapid detection of terrorist attacks. *Statistics in Medicine* 2005; 24(4):513-529

[38] Mostashari F, Kulldorff M, Hartman JJ, Miller JR and Kulasekera V. Dead bird clusters as an early warning system for West Nile Virus activity. *Emerging Infectious Diseases* 2003; 9(6): 641-646

[39] Souris M, Gonzalez JP, Shanmugasundaram J, Corvest V and Kittayapong P. Retrospective space-time analysis of H5N1 Avian Influenza emergence in Thailand. *International Journal of Health Geographics* 2010; 9(1): 3

[40] Read S, Bath PA, Willett P and Maheswaran R. A power-enhanced algorithm for spatial anomaly detection in binary labelled point data using the spatial scan statistic. In: Setchi R et al. (eds.) Knowledge-based intelligent information and engineering systems, part II (Lecture Notes in Artificial intelligence 6277). Berlin: Springer-Verlag, 2010, pp. 163-172.

[41] Kulldorff M, Heffernan R, Hartman J, Assuncao R and Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. *Public Library of Science, Medicine* 2005; 2(3): 216-244.

[42] Diggle PJ. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society* 1990; 153(3): 349-362.

[43] Read S, Bath PA, Willett P and Maheswaran R. A pilot inference study for a beta-Bernoulli spatial scan statistic. Proceedings of the GIS Research UK 20th Annual Conference, Lancaster UK, April 11th-13th 2012. (Eds. Whyatt D and Rowlingson B). Volume 1: 67-71. Lancaster, UK: The University of Lancaster.

