A Bayesian approach to the Bernoulli Spatial Scan Statistic

A working paper by Simon Read

Information School, University of Sheffield

simon.read@sheffield.ac.uk / simonread14@hotmail.com

September 11, 2011

Abstract

This document describes a novel approach to finding localised clusters in spatially distributed, binary labelled point data. The frequentist spatial scan statistic, introduced by Martin Kulldorff in 1995, was developed into a Bayesian spatial scan statistic for areal data by Daniel Neill, circa 2006, where computationally expensive Monte Carlo testing is replaced by the use of historical data and expert judgement. Following Neill's approach, I present here my derivation of a Bayesian spatial scan statistic for binary labelled point data. I have also developed a method for replacing historic data with expert judgement, by using a prior probability distribution of relative risk. Please note this document describes work in progress, and content may be subject to revision.

1 Introduction

First introduced by [1] and [2], the spatial scan statistic (hereafter SSS) is widely used in spatial epidemiology, and other fields. The SSS is an umbrella term for a range of statistics which share a common purpose and similar method of application, but vary in the nature of the data to which they can be applied. In this document I consider the Bernoulli version, applicable to spatially distributed binary labelled point data, such as that used in a geospatial case control study.

The frequentist version (described in [2]) uses Monte Carlo testing to obtain statistical inference, which can be computationally expensive for large data sets, especially if real time surveillance of data is required. [3] developed the Bayesian SSS, based on the Poisson SSS and suitable for areal data, which obviates the need for Monte Carlo testing.

In this document I use the Bayesian approach set out in [3], and apply it the the Bernoulli SSS. Additionally, I show how the use of historic data suggested in [3] can, with some relatively mild assumptions and approximations, be replaced by expert knowledge concerning the probability distribution of relative risk level of any cluster which may occur.

2 Derivation of a Bayesian Bernoulli spatial scan statistic

Preliminaries:

- R = study region.
- D = set of points within R, some of which are *cases*, some of which are *controls*.
- $N = |D|$.
- C = number of cases in D, always $\leq N/2$.
- Z = a subset of R. For the SSS there are typically thousands of different Z generated for a given R, by some automated process.
- D_{in} = the data points in D that lie within Z.
- D_{out} = the data points in D that lie within $R - Z$.
- $n = |D_{in}|$.
- c = numbers of cases in D_{in}.
- H_0 = null (no clustering) hypothesis where the probability of any point being a case (q_{all}) is uniform across D.

- H_A = Alternate hypothesis, where the probability (q_{in}) of any point in D_{in} being a case is different (usually higher) than the probability (q_{out}) of any point outside D_{out} being a control.

As stated in [3], the aim of the Bayesian SSS is to find $P(H_0|D)$[1]. If it is below a certain threshold one may wish to declare an anomaly present at whatever Z has the largest value of $P(Z|D, H_A)$.

Using the same starting point as Neill's derivation in [3] one has (from Bayes' Theorem):

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)} \tag{1}$$

and

$$P(H_A|D) = \frac{P(D|H_A)P(H_A)}{P(D)} \tag{2}$$

Following on from this, I have derived the following. First consider $P(D|H_0)$. Although Z is normally only considered in association with H_A, we are in fact also free to consider Z in association with H_0 (after all, if H_0 is true, the choice of Z is irrelevant). So we have:

$$P(D|H_0) = \sum_{\forall Z} P(D|H_0, Z)P(Z) \tag{3}$$

Following [3], we assume *a priori* that all Z are equally likely, so $P(Z) = \frac{1}{|Z|}$. Furthermore, as D_{in} and D_{out} are considered as series of independent Bernoulli trials:

$$P(D|H_0, Z) = \int_{x=0}^{x=1} P(D_{in}|q_{all} = x)P(q_{all} = x)dx \times \int_{x=0}^{x=1} P(D_{out}|q_{all} = x)P(q_{all} = x)dx \tag{4}$$

In a real case/control study the ratio of cases to controls is always a parameter, never a variable (i.e C, as well as N, is a value always given in advance). Hence, as we are using Bernoulli trials to model D, it seems appropriate to assume q_{all} takes

[1]This is comparable with the p-value of the frequentist SSS

a value which maximises the probability of C occurring, i.e. $P(q_{all} = x) = 1$ when $x = \frac{C}{N}$, 0 otherwise. Thus we have:

$$P(D|H_0, Z) = P(D_{in}|q_{all} = \frac{C}{N}) \times P(D_{out}|q_{all} = \frac{C}{N}) = (\frac{C}{N})^C(1 - \frac{C}{N})^{N-C} \quad (5)$$

and similarly the marginal over all Z is:

$$P(D|H_0) = \frac{1}{|Z|}\sum_{\forall Z} P(D|H_0, Z) = (\frac{C}{N})^C(1 - \frac{C}{N})^{N-C} \quad (6)$$

which is an identical result to the frequentist Bernoulli SSS. Now for $P(D|H_A)$, we also have:

$$P(D|H_A) = \sum_{\forall Z} P(D|H_A, Z)P(Z) = \frac{1}{|Z|}\sum_{\forall Z} P(D|H_A, Z) \quad (7)$$

Now as D_{in} and D_{out} are considered as series of independent Bernoulli trials we have:

$$P(D|H_A, Z) = \int_{x=0}^{x=1} P(D_{in}|q_{in} = x)P(q_{in} = x)dx \times \int_{x=0}^{x=1} P(D_{out}|q_{out} = x)P(q_{out} = x)dx \quad (8)$$

Unlike q_{all}, we will permit some variation in q_{in} and q_{out}. It is known that for a series of independently, identically distributed Bernoulli trials, where the number of successes is known, the distribution of the probability of success follows the beta distribution. So we have

$$P(q_{in} = x) = \frac{x^{\alpha_{in}-1}(1-x)^{\beta_{in}-1}}{B(\alpha_{in}, \beta_{in})} \quad (9)$$

Where the beta function is:

$$B(\alpha_{in}, \beta_{in}) = \int_{t=0}^{t=1} t^{\alpha_{in}}(1-t)^{\beta_{in}}dt \quad (10)$$

Exactly the same applies for q_{out}, with similar parameters α_{out} and β_{out}. As the beta functions are independent of x, this leads to:

$$
\begin{aligned}
P(D|H_A, Z) &= \frac{1}{B(\alpha_{in}, \beta_{in})B(\alpha_{out}, \beta_{out})} \\
&\times \int_{x=0}^{x=1} x^{\alpha_{in}+c-1}(1-x)^{\beta_{in}+n-c-1} \\
&\times \int_{x=0}^{x=1} x^{\alpha_{out}+C-c-1}(1-x)^{\beta_{out}+N-n-C+c-1}
\end{aligned}
$$

Now as these integrals are themselves beta functions, $P(D|H_A)$ conveniently becomes:

$$
P(D|H_A, Z) = \frac{B(\alpha_{in}+c, \beta_{in}+n-c)B(\alpha_{out}+C-c, \beta_{out}+N-n-C+c)}{B(\alpha_{in}, \beta_{in})B(\alpha_{out}, \beta_{out})} \quad (11)
$$

Here one has a similar situation to [3], except where Neill has gamma functions with parameters $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out}, I have a beta functions. The selection of these parameters is discussed in the next section.

Regarding the other probabilities required to calculate $P(H_0|D)$, $P(D)$ is simply:

$$
P(D) = P(D|H_0) + \sum_{\forall Z} P(D|H_A)
$$

We also know that, as they are mutually exclusive, $P(H_0)+P(H_A) = 1$. However, as we have no other D over which to calculate these marginal probabilities, either $P(H_0)$ or $P(H_A)$ must be assumed from expert knowledge (see [3]). Thus, we consider either $P(H_0)$ or $P(H_A)$ to be a *tuning parameter*.

3 Selection of $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out}

Aside from the selection of the tuning parameter mentioned above, the main challenge in using the Bayesian approach to the Bernoulli SSS is selecting suitable values for $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out}. In [3], the selection of these values comes from fitting the gamma functions to historic data. In this historic data, it is assumed there is no clustering; thus $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out} are modified by a variety of arbitrary multipliers to simulate a clustering situation.

Here I divert from Neill's approach. As C is a parameter rather than a variable, what really interests us about q_{in} and q_{out} is their ratio, rather than their individual values, as this strongly influences the likely distribution of c (recall, c is the number of cases in Z, and if c is high in proportion to n it is strong evidence of clustering). However, the size of n, N and C all influence the likely distribution of c, meaning that the parameters $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out} will vary for each different size of Z.

It is entirely reasonable, however, to assume that the prior distribution of relative risk, i.e. $\frac{q_{in}}{q_{out}}$, is the same for all Z, or at the very least not influence by N, C and n. After all, relative risk is an epidemiological concept relating to real disease processes, not some artefact of a spatial model, as q_{in} and q_{out} are. So, if we allow expert knowledge to influence our prior distribution of relative risk (hereafter RR), we can make a good guess at what parameters $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out} should be for any given Z.

To do this, let us consider q_{in} and q_{out}. Assuming both remain in the range 0 to 1, we can allow them vary whilst still maximising the likelihood of obtaining C cases in total; the same assumption used to work out $P(D|H_0)$). This is expressed as:

$$nq_{in} + (N - n)q_{out} = C \tag{12}$$

So we can express RR explicitly in terms of either q_{in} or q_{out}:

$$RR = \frac{(N - n)q_{in}}{(C - nq_{in})} = \frac{C - (N - n)q_{out}}{nq_{out}} \tag{13}$$

Now when a variable X follows the beta distribution, the expression $\frac{X}{1-X}$ follows a distribution called the *beta prime*, which has pdf:

$$P(RR = x) = \frac{x^{\alpha_{RR}-1}(1 + x)^{-\alpha_{RR}-\beta_{RR}}}{B(\alpha_{RR}, \beta_{RR})} \tag{14}$$

where B is again the beta function. Beta prime is used for modelling the distribution of odds ratios, and is controlled by selection of its two parameters (here I call them α_{RR} and β_{RR}). The above expressions for RR are both of a very similar form to $\frac{X}{1-X}$, and if one imposes a beta prime distribution on RR, it can be clearly seen

from plotting q_{in} and q_{out} that they follow distributions very close to beta[2] (which is what is required for use in Expression 8).

Now one can select $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out} such that q_{in} and q_{out} closely follow the 'true' distributions they would have for the chosen beta prime distribution of RR. In doing so one is, effectively, converting the two beta priors into a single beta prime prior, upon which real-world expert knowledge can be applied.

Although $\alpha_{in}, \beta_{in}, \alpha_{out}$ and β_{out} must currently be calculated using a fitting process, the 'true' distribution is so close to beta that the fitting is very straightforward, requiring only the mode of the beta prime pdf together with one other point on the pdf. The process is of linear time complexity, with the estimate converging to four decimal places in typically less than 10 iterations. Also, the fitting only need be done once for each unique value of n in the study.

Strictly speaking, as q_{in} and q_{out} are no longer independent one should replace their values in Expression 8 with formulae including only RR, and integrate over all values of RR. However, the resulting integral can, so far as I can tell, only be integrated numerically, which would add significantly to the computational expense. Allowing q_{in} and q_{out} to remain independent, but following distributions compatible with that RR, appears satisfactory in limited tests conducted so far (not yet published).

Using the Ω measure defined in [4], the spatial accuracy of the freqentist and Bayesian Bernoulli SSS have been shown to be compatible for sensible choices of beta prime (results currently pending publication as a conference paper). I hope to compare the raw detection capability of the two methods in the near future.

I would very much welcome feedback on the contents of this document from any interested parties.

[2]They may in fact be exactly beta distributed, but I have not yet shown this analytically.

4 Bibliography

References

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

[2] Kulldorff, M. (1997). Communications in statistics - theory and methods 26(6): 1481-1496

[3] Neill, D. B. (2006). Detection of spatial and spatio-temporal clusters. PhD thesis. School of Computer Science. Pittsburgh, Carnegie Mellon University

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