Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Statistics & Data Analysis 48 (2005) 679–684

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Power comparisons for an improved disease clustering test[☆]

Al Ozonoff*, Marco Bonetti[1], Laura Forsberg, Marcello Pagano

*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue,
Boston, MA 02115, USA*

Received 26 January 2004; received in revised form 15 March 2004; accepted 18 March 2004

Abstract

The current note presents the power comparisons for disease clustering tests, as originally reported by Kulldorff et al. (Comput. Statist. Data Anal. 42 (2003) 665). A minor improvement to the implementation of the M-statistic, motivated by that work, results in dramatically higher power to detect clusters of disease.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Spatial statistics; Power; Spatial epidemiology; Hypothesis testing; Cluster detection

1. Introduction

This current note is intended to present the results first reported in this journal comparing the power of various statistics to detect simulated clusters (Kulldorff et al., 2003). The original paper compared three statistics: Kulldorff's spatial scan statistic (Kulldorff, 1997), Tango's MEET statistic (Tango, 2000) and the M-statistic of Bonetti and Pagano (Bonetti and Pagano, 2004). At the time of writing, the M-statistic was relatively in early stages of development. In particular, proper implementation for simulations and power comparisons was in an early stage. Since then some basic improvements have been made to the algorithmic implementation of the M-statistic, and some fundamental methodological improvements have also enhanced the power of the

[☆] Partially funded by grants from the National Institutes of Health, RO1AI28076, and the National Library of Medicine, RO1LM007677.

* Corresponding author. Tel.: +1-6174322416; fax: +1-6177391781.
E-mail address: aozonoff@hsph.harvard.edu (A. Ozonoff).

[1] Also at Dana-Farber Cancer Institute, Boston, MA 02115, USA.

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.03.012

M-statistic to detect simulated clusters. The results presented here are our most recent effort to accurately record the power of detection as it was earlier described (Kulldorff et al., 2003).

2. Methodology

A brief summary of the conceptual basis for the M-statistic follows, which will help us make clear the changes that led to the new results.

Bonetti–Pagano's M-statistic is a non-parametric general test for clustering. It represents and compares the spatial distributions of two populations (here the null and alternative data sets) via the interpoint distance distribution. From a collection of n cases, we can calculate the $\binom{n}{2}$ interpoint distances between cases and consider the distribution of these distances. A resampling procedure on the control population is used to generate a baseline (or null) distribution. The interpoint distances between the cases is calculated, and both distributions are binned into histograms, each of which can be represented as a vector. The test statistic is then a Mahalanobis-like distance between the two vectors, weighted by an estimate of the covariance between histogram bins.

More formally, repeated resampling from the null data sets is used to estimate the distribution of distances under the null hypothesis. Binning these distances and taking the mean over all iterations gives expected counts for each bin of the histogram. The number of bins, k, needed to achieve optimum sensitivity and specificity is being studied; experience with this method suggests that the optimal number of bins grows roughly on the order of $\sqrt{n}$, where n is the number of cases being assessed (see also, Mann and Wald, 1942). Let e denote the vector of expected bin counts. Then a resampling procedure also allows us to estimate the covariance of e, which we will write as S, a k by k square matrix.

The interpoint distances for the disease cases are calculated, binned, and written as a k-dimensional vector o, the observed counts. Then the M-statistic is

$$M = (o - e)' S^{-} (o - e),$$

where S^{-} is a generalized inverse of the sample covariance matrix S. Thus we calculate the difference between the expected (under the null hypothesis of no clustering) bin counts and the observed bin counts of the disease cases, inversely weighed by the covariance estimator. As S^{-} is a positive semi-definite matrix, $M \geqslant 0$.

The general asymptotic distribution of M is found in Bonetti and Pagano (2004). In practice, we can use the resampling procedure to calculate the distribution of M for the null population. Comparing the calculated value of the test statistic to the null distribution gives a Monte Carlo p-value that can be interpreted as the probability that the spatial distribution of the alternative data differs from the null by chance alone.

The relevant changes in the implementation of the M-statistic for these results lie in the binning procedure used to approximate the interpoint distance distribution. In the original power comparison Kulldorff et al. (2003), the collection of interpoint distances was binned into 40 equispaced bins, as in an ordinary histogram. Since that time, further investigation has indicated that higher sensitivity may be achieved by using equiprobable bins. Under this procedure, we estimate quantiles of the distribution of distances under the null hypothesis. These quantiles are used as the breakpoints for binning, thus resulting in equal expected counts under the null. We do not need to consider any particular alternative in order to determine the quantiles under the null, hence this method does not make any reference to the alternative model.

A typical unimodal distribution, such as we often find for the interpoint distances under the null, has the majority of its probability mass near the mode, and relatively little mass in the tails. Hence, equispaced bins will have expected counts that are high near the mode and small near the tails. In order to maintain stable estimates for the covariance matrix of bin proportions, we must avoid bin counts near zero and thus must keep an appropriately small number of equispaced bins.

Equiprobable binning allows for a higher number of bins to be used, since computational problems with sparse bin counts near the tails can be controlled as the number of bins increases. Because the bins are of variable width, the bins near the tails are much wider than the bins near the mode. When calculating the test statistic, the heteroskedasticity associated with these variable bin widths is accounted for via the covariance matrix. Equiprobable binning allows one to use many more bins and hence, in most cases, to achieve greater sensitivity to potential disturbances in the distribution of distances. This greater sensitivity is reflected in the improved results below.

For the results reported here, 200 equiprobable bins were used, with quantiles estimated from the null data sets. To make the differences in binning clear, with 600 cases there are $\binom{600}{2} \sim 180,000$ interpoint distances. This gives an expected value of roughly 900 distances per bin across all 200 bins. To contrast this with the case of 40 equispaced bins, for the null data sets used in this study the expected counts for the last three bins were only 2.8, 1.6, and 0.4.

3. Results

Tables 1 and 2 present the results for the M-statistic under the hot spot and global chain (GCC) models, respectively. The column labelled M_{es} contains the results of the M-statistic under equispaced binning, as computed in the original paper (Kulldorff et al., 2003). The column M_{ep} contains the new results for the M-statistic under equiprobable binning. Only the results for the data sets with 600 cases were recomputed. Figures for the scan statistic and MEET are also from the original paper.

For the single hot spot alternatives (in particular urban and rural clusters), the M-statistic showed considerable gains in power after implementing the new binning procedure. For this class of alternatives, the M-statistic still underperforms the scan statistic, as expected for a single hot spot model. The GCC models with small fixed and exponential distances also showed marked improvement. In particular, the M-statistic is now the most powerful test for GCC alternatives with zero or short exponential distances.

Table 1
Estimated power of the spatial scan statistic, the M-statistic and the MEET for 35 different alternative models with different hot-spot clusters, for 600 simulated cases and for significance levels 0.05 and 0.01

Counties		$\alpha = 0.05$				$\alpha = 0.01$			
		Scan	M_{ep}	M_{es}	MEET	Scan	M_{ep}	M_{es}	MEET
Rural (edge)	1	0.998	0.816	0.355	0.196	0.992	0.653	0.127	0.057
	2	0.991	0.753	0.406	0.221	0.986	0.546	0.154	0.072
	4	0.973	0.428	0.292	0.229	0.946	0.194	0.082	0.064
	8	0.971	0.293	0.241	0.213	0.937	0.094	0.058	0.055
	16	0.969	0.204	0.197	0.229	0.936	0.053	0.041	0.062
Mixed (corner)	1	0.936	0.885	0.909	0.925	0.871	0.759	0.757	0.833
	2	0.939	0.853	0.883	0.896	0.871	0.704	0.703	0.771
	4	0.937	0.767	0.815	0.838	0.873	0.578	0.590	0.654
	8	0.941	0.692	0.794	0.817	0.876	0.472	0.567	0.599
	16	0.949	0.602	0.745	0.832	0.886	0.372	0.484	0.602
Urban (central)	1	0.922	0.907	0.342	0.941	0.818	0.805	0.115	0.870
	2	0.903	0.859	0.397	0.920	0.823	0.722	0.154	0.830
	4	0.892	0.905	0.711	0.961	0.794	0.799	0.428	0.902
	8	0.913	0.855	0.844	0.983	0.824	0.705	0.619	0.951
	16	0.926	0.738	0.777	0.986	0.836	0.527	0.504	0.950
Rural and mixed	1	1.000	0.992	0.980	0.964	0.999	0.974	0.916	0.910
	2	0.999	0.986	0.970	0.952	0.997	0.954	0.894	0.871
	4	0.997	0.934	0.931	0.930	0.987	0.814	0.804	0.793
	8	0.996	0.862	0.915	0.931	0.986	0.689	0.741	0.772
	16	0.996	0.774	0.827	0.941	0.982	0.535	0.590	0.804
Rural and urban	1	1.000	0.994	0.709	0.970	0.998	0.980	0.400	0.923
	2	0.999	0.987	0.644	0.962	0.996	0.949	0.334	0.895
	4	0.992	0.975	0.811	0.971	0.974	0.924	0.538	0.912
	8	0.991	0.939	0.884	0.977	0.968	0.830	0.667	0.936
	16	0.987	0.835	0.776	0.975	0.947	0.634	0.481	0.915
Mixed and urban	1	0.987	0.997	0.964	0.998	0.950	0.990	0.868	0.995
	2	0.984	0.990	0.950	0.995	0.950	0.967	0.829	0.984
	4	0.966	0.986	0.954	0.991	0.901	0.954	0.830	0.969
	8	0.954	0.954	0.970	0.990	0.871	0.876	0.873	0.960
	16	0.935	0.873	0.929	0.984	0.811	0.696	0.742	0.935
Rural, mixed and urban	1	1.000	1.000	0.991	0.999	0.999	1.000	0.958	0.997
	2	1.000	0.999	0.981	0.998	0.999	0.997	0.920	0.992
	4	0.996	0.996	0.979	0.994	0.981	0.984	0.895	0.973
	8	0.992	0.981	0.980	0.989	0.964	0.932	0.901	0.952
	16	0.977	0.908	0.929	0.983	0.916	0.755	0.744	0.918

Only power for the M-statistic has been revised.

Table 2
Estimated power of the spatial scan statistic, the M-statistic and the MEET for 26 global chain clustering models, for 600 simulated cases and for significance levels 0.05 and 0.01

Distance(r)		$\alpha = 0.05$				$\alpha = 0.01$			
		Scan	M_{ep}	M_{es}	MEET	Scan	M_{ep}	M_{es}	MEET
Twins									
No distance	0	0.791	1.000	0.860	0.990	0.513	0.995	0.616	0.945
Fixed distance	0.005	0.392	0.600	0.346	0.624	0.197	0.368	0.130	0.376
	0.01	0.285	0.269	0.163	0.406	0.131	0.123	0.044	0.201
	0.02	0.194	0.100	0.087	0.264	0.084	0.032	0.019	0.110
	0.04	0.124	0.052	0.060	0.174	0.049	0.012	0.014	0.068
	0.08	0.080	0.052	0.051	0.109	0.024	0.013	0.009	0.038
	0.16	0.055	0.053	0.050	0.059	0.014	0.012	0.009	0.014
Exponential distance	0.005	0.452	0.789	0.449	0.738	0.229	0.563	0.189	0.486
	0.01	0.351	0.550	0.304	0.556	0.165	0.319	0.106	0.299
	0.02	0.262	0.314	0.184	0.378	0.110	0.130	0.051	0.171
	0.04	0.185	0.162	0.114	0.250	0.073	0.055	0.027	0.096
	0.08	0.124	0.095	0.083	0.166	0.042	0.026	0.018	0.056
	0.16	0.080	0.061	0.059	0.107	0.023	0.015	0.010	0.029
Triplets									
No distance	0	0.995	1.000	0.996	1.000	0.949	1.000	0.969	1.000
Fixed distance	0.005	0.674	0.839	0.569	0.884	0.460	0.684	0.291	0.728
	0.01	0.491	0.400	0.253	0.646	0.309	0.229	0.087	0.415
	0.02	0.318	0.130	0.117	0.430	0.178	0.051	0.032	0.237
	0.04	0.189	0.063	0.070	0.265	0.094	0.018	0.018	0.135
	0.08	0.102	0.060	0.053	0.141	0.038	0.018	0.010	0.057
	0.16	0.046	0.061	0.049	0.050	0.010	0.017	0.011	0.015
Exponential distance	0.005	0.762	0.968	0.734	0.960	0.538	0.894	0.457	0.862
	0.01	0.610	0.800	0.497	0.826	0.388	0.602	0.232	0.615
	0.02	0.436	0.489	0.294	0.599	0.253	0.262	0.099	0.363
	0.04	0.289	0.226	0.162	0.390	0.144	0.089	0.043	0.202
	0.08	0.171	0.112	0.096	0.226	0.068	0.030	0.021	0.096
	0.16	0.091	0.064	0.062	0.115	0.027	0.014	0.013	0.036

Only power for the M-statistic has been revised.

The results suggest that the M-statistic is comparable to both the spatial scan statistic and the MEET statistic for the most alternative models of clustering. We see that the M-statistic's power was within 0.05 or greater than that of the spatial scan statistic on 16 of the 35 hot spot models, and 19 of the 26 GCC models. Likewise, the M-statistic was within 0.05 or greater than that of the MEET statistic on 23 of 35 hot spot models and 11 of 26 GCC models.

References

Bonetti, M., Pagano, M., 2004. The interpoint distance distribution as a descriptor of point patterns: an application to cluster detection. Statist. Med., submitted for publication.

Kulldorff, M., 1997. A spatial scan statistic. Commun. Statist. Theory Methods 26, 1481–1496.

Kulldorff, M., Tango, T., Park, P.J., 2003. Power comparisons for disease clustering tests. Comput. Statist. Data Anal. 42, 665–684.

Mann, H.B, Wald, A., 1942. On the choice of the number and width of classes for the chi-square test of goodness of fit. Ann. Math. Statist. 13, 306–317.

Tango, T., 2000. A test for spatial disease clustering adjusted for multiple testing. Statis. Med. 19, 191–204.