Noname manuscript No.
(will be inserted by the editor)

Supplement Material: Irregular Shape Small Nodules Detection Using a Robust Scan Statistic

Ali Abolhassani · Marcos O. Prates · Safieh Mahmoodi

Received: date / Accepted: date

SM-1 New validity indexes

In this section we call the validity index presented in Equation (5) (of the main text) as "Zhou validity index (ZVI)". We also define two other validity indexes to compare with ZVI. They are defined as follow:

1- Changing the denominator of ZVI to the

$$\text{mean(distance between sub-partitions)} + \sqrt{\text{Var(distance between sub-partitions)}}.$$

We call modified validity index as MSE-type validity index.

2- Changing the denominator of ZVI to the

$$\min_{i,j} |\lambda_{C^i} - \lambda_{C^j}|^2. \tag{1}$$

For this modification, the validity index will be minimum, when its denominator, i.e. equation (1), becomes maximum. We call this modified validity index as Minimax validity index. Based on our

Ali Abolhassani
Department of Applied Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran.
E-mail: ali.abolhassani@azaruniv.ac.ir
Marcos O. Prates
Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, MG-Brasil.
Tel.: +55 31 34095932
E-mail: marcosop@est.ufmg.br
Safieh Mahmoodi
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran.
E-mail: mahmoodi@cc.iut.ac.ir

simulation, we do not detect considerable difference in the performance of these three validity index in cluster detection problems.

SM-2 Generating random numbers from Bell distribution

Castellares et al. (2018) introduced the Bell distribution with one parameter under the regression framework and showed that it is suitable to deal with over-dispersed data. Because, for a random variable $V \sim \text{Bell}(\theta)$, where $\theta > 0$

$$E(V) = \theta e^{\theta},\; Var(V) = \theta(1+\theta)e^{\theta},$$

hence the index of dispersion is $I_d = \dfrac{Var(V)}{E(V)} = 1 + \theta > 1$, which means that this distribution covers over-dispersion.

To generate random numbers from Bell distribution, consider the following assumptions. Suppose that $U = X_1 + \ldots + X_N$, such that X_i's are independent zero-truncated Poisson(θ) and $N \sim \text{Poisson}(e^{\theta} - 1)$. Also, consider that, N is independent of X_i's. Then, one can prove that U follows a Bell(θ). Using this property we can generate random numbers from Bell distribution.

SM-3 Binomial maps

The framework of this scenario is similar to the Poisson maps in the main paper. Just, instead of the Poisson distribution, we use the binomial distribution to generate maps.

In the first step, $\text{Bin}(1000, 0.012)$ is used to generate the number of cases inside cluster areas and a $\text{Bin}(1000, 0.01)$ for outside. Four criteria: recall, precision, biassness, and $F1$, are used to compare three different scans. The results are shown in Figure 1. Since the figure for recall and precision are almost similar to scenario Poisson(10)-Poisson(12), its interpretation is the same. Bias value for Ir-Bell and Ir-binomial are similar and they show superiority in comparison with the bias value for Ir-Poisson, where its value is further from 1. The value of $F1$ for Ir-binomial is better than the two others. As before, we plot precision and recall for the first 50 iterations of simulations in Figure 2. In this figure, recall for Ir-binomial is considerably higher than two other scans. The precision of the three scans is about the same. Since recall for Ir-binomial is bigger, it means that this scan detects clusters better for the maps generated by the binomial distribution. In the case of the Ir-Bell scan, fluctuation of the recall and precision are very similar and stable. This means that the Ir-Bell scan does a fair job of detecting the true clusters in binomial maps.

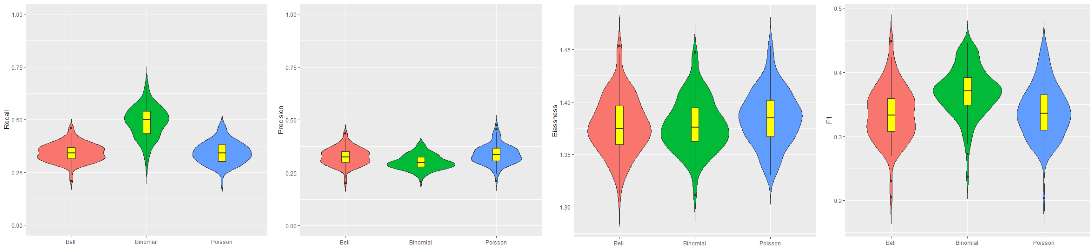

Fig. 1: Data for the map are generated from the binomial with parameter $(1000, 0.012)$ and $(1000, .01)$ respectively inside and outside cluster. From left to right: the violin plot for the recall, precision, bias, and $F1$. The number of iteration is 200. The red, green, blue colors are respectively the Ir-Bell, Ir-binomial, Ir-Poisson scan.

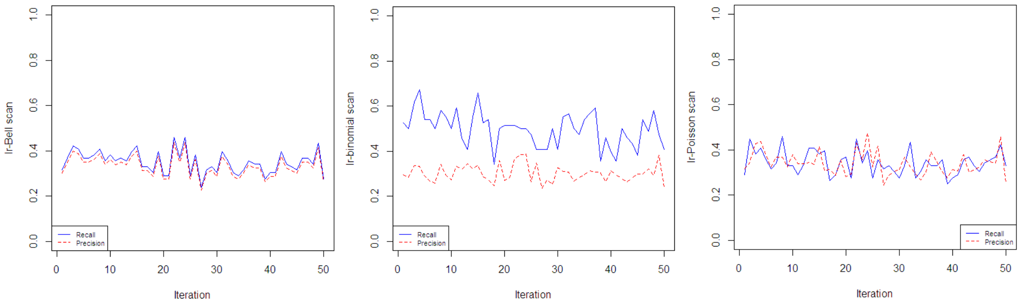

Fig. 2: Variation of the recall and precision in the first 50 iteration of simulation study for scenario binomial$(1000, 0.01)$-binomial$(1000, 0.012)$. From left to right: the Ir-Bell, Ir-binomial and Ir-Poisson scans.

Increasing the parameter inside the cluster from 0.012 to 0.02 leads us to obtain Figure 3. The Ir-binomial and Ir-Bell have better value for $F1$. Even though the bias value for the Ir-Poisson scan is slightly better, the three scans have their bias value very close to 1. As before, to have a better vision on the precision and recall, we plot the first 50 of them pointwise in Figure 4. As it can be seen, precision for the Ir-Bell and Ir-binomial is higher than for the Ir-Poisson. Since the values of the precision and recall in Ir-Bell is high, the detected cluster by the Ir-Bell is a really good approximation of the true cluster.

In the last step of this scenario, we increase the parameter from 0.02 to 0.04. The detection of clusters by the three scans is almost perfect and the four criteria are close to 1.

References

A. Abolhassani, M. O. Prates, F. Castellares, S. Mahmoodi, Zero-inflated bell scan: A moreflexible spatial scan statistic, Spatial Statistics (2020) 100433.

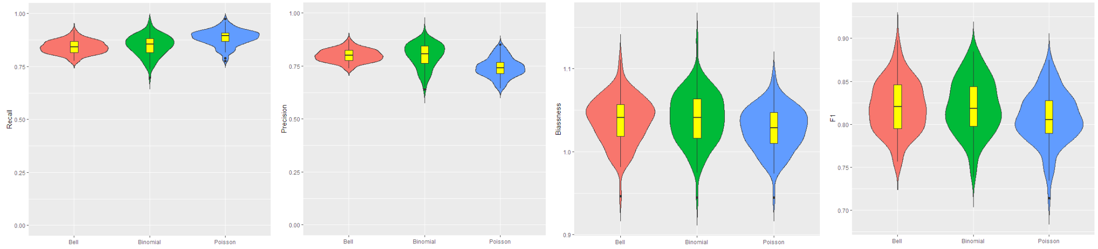

Fig. 3: Data for the map are generated from the binomial with parameter $(1000, 0.02)$ and $(1000, 0.01)$ respectively inside and outside cluster. From left to right: the violin plot for the recall, precision, bias, and $F1$. The number of iteration is 200. The red, green, blue colors are respectively for the Ir-Bell, Ir-binomial, Ir-Poisson scans.

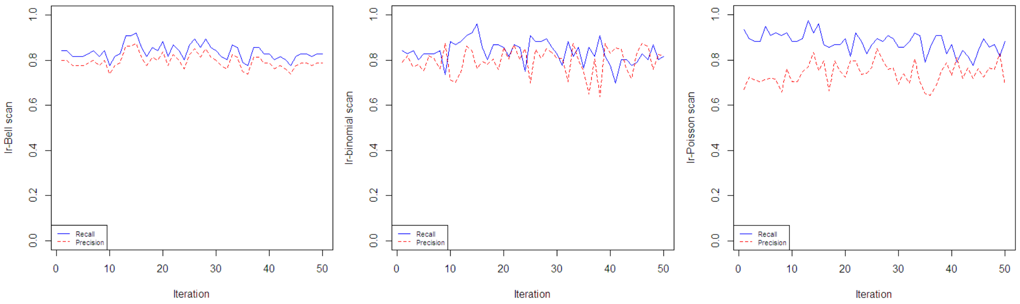

Fig. 4: Variation of the recall and precision in the first 50 iteration of simulation study for scenario binomial$(1000, 0.01)$-binomial$(1000, 0.02)$. From left to right: the Ir-Bell, Ir-binomial and Ir-Poisson scans.

F. Castellares, S.L., Ferrari, A.J., Lemontec, On the Bell distribution and its associated regression model for count data, Applied Mathematical Modelling (2018) 56:172-185.