A homogeneity test for spatial point patterns

M.V. Alba-Fernández
University of Jaén
Paraje las lagunillas, s/n
B3-053, 23071, Jaén,
Spain
mvalba@ujaen.es

F. J. Ariza-López
University of Jaén
Paraje las lagunillas, s/n
A3-336, 23071, Jaén,
Spain
fjariza@ujaen.es

Abstract

It is proposed to analyze spatial point patterns (2D or 3D), theoretical or observed, by means of a linearization of the problem using space filling curves followed by the application of a multinomial law. This way, the problem to test the similarity between sampled spatial distributions is equivalent to the problem of testing the homogeneity of two multinomial distributions where the proportions are the account of cases on each cell of the grid derived by the space filling curve. In the application of this approach we have to deal with the existence of many cells of the grid without any count. This fact may affect the test statistic considered for testing the homogeneity test. Our proposal overcomes this problem by applying specific test statistic belonging to the power divergence family and two strategies to collapse cells. Some simulations are performed in order to analyze the behavior of the proposal in relation to the order of the curve, number of elements. Results indicate some rules for obtaining better results when the application of the proposal. An application to a real data set is included.

Keywords: spatial point patterns similarity, homogeneity test, multinomial distribution, power divergence family

1 Introduction

The understanding of spatial point patterns is one of the major challenges of geographical analysis and has interest in many sciences (e.g. biogeography, crop sciences, ecology, geology, etc.). It is usual that an estimate of an attribute or property (e.g. concentration of a mineral, positional accuracy, presence, etc.), is derived from a sample of points under some spatial distributions (theoretic or observed), for instance, different geological structures can determine presence/absence and concentration of a mineral. We think that, previous to any consideration about an attribute or property estimation, we must confirm the underlying hypothesis about the location of points events (e.g. the similarity between two given point patterns obtained by sampling).

A spatial point pattern has been defined as a set of locations, irregularly distributed within a region of interest, which have been generated by stochastic mechanisms –point process– (e.g. complete spatial randomness, aggregative, repulsive processes, etc.). However, the method we are proposing is independent of the true point process, and it doesn't matter if it is known or unknown.

Our proposal is very different to other techniques such as the Ripley's K function (Ripley, 1976). For the correct application of these functions it is needed to accept several assumptions that not necessarily are fully met in reality. To overcome these problems, we have developed an area-based test centered on the counting of positional events, which does not assume a theoretical model, it is spatial distribution free. Our proposal is more similar to the application of a quadrat-counts test (`quadrat.test` in spatstat package, Baddeley et al. (2017)). However, some important differences between both approaches have to be highlighted.

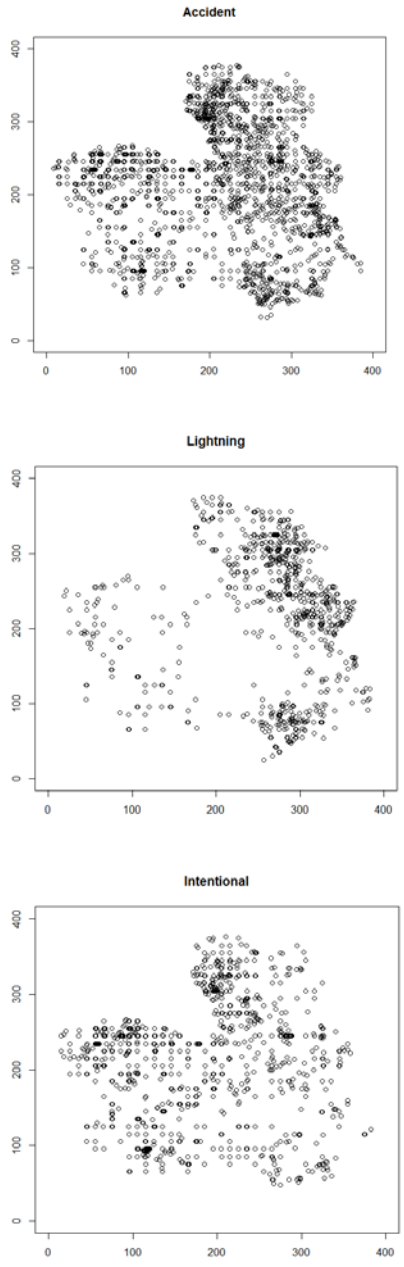
The first one is the null hypothesis to be tested. In Baddeley et al. (2017) the null hypothesis is the complete spatial randomness of a given point pattern, whereas our null hypothesis is the similarity of two given point patterns (e.g. two control test samples, two field works, presences at two different times, etc.), independently of the point process that generate them.

The second one is the way to overcome the requirement of the χ^2 test related to the number of points in each quadrant (be greater than 5). Although in both cases, it is used the multinomial distribution of the vector of quadrant counts, for the testing problem in Baddeley et al. (2017) this fact can be obviated by using the Monte Carlo test, however, for testing the homogeneity of two multinomial laws, this solution is not always possible for the χ^2 test, especially when the sample sizes are small and empty cells are observed. So, it is necessary to analyse new strategies to solve this drawback. Here we study if collapsing cells in the order in which we scan the space together with a test statistic more appropriate than the χ^2 test is able to provide a formal procedure to determine the similarity of two point patterns.

For this purpose, some simulation experiments are carried out and an illustration of the application to a real data set is included. In particular, we have used a data set from Baddeley et al. (2017) that comprises the distribution of fires by different causes (accident, intentional and lightning) in the region of "Castilla La Mancha" (Spain). Figure 1 shows the distribution of fire locations. Our proposal is oriented to answer this question: Are these spatial distributions similar?

And to answer this question we do not need any statistical assumption about the underlying processes.

Figure 1: Example of fires distribution by causes in the region "Castilla La Mancha" (Spain).



2 Description of the methodology

The method follows the proposal stated by Alba-Fernández et al. (2016) consisting in a procedure with two steps:

i) use of a space-filling curves to order the space and model the count of points in the resultant grid by means of the multinomial law,

ii) carry out an homogeneity test for multinomial populations.

This way, the problem to test the similarity between sampled spatial distributions is equivalent to the problem of testing the homogeneity of two multinomial distributions where the vector of observed probabilities observed is equal to the relative frequencies of counts on the grid.

One of the problems pointed out in the work of Alba-Fernández et al. (2016) is the treatment of empty cells, which is a very common situation in Geography because many features (e.g. homes, trees, wells, lakes, etc.) are disperse by nature, generating cells with zero counts or very low accounts. In this way, as stated by Baddeley et al. (2017) and many others, an important requirement of the χ^2 test is that the counts in each quadrat be greater than 5. This paper is a step forward in order to solve this problem, the advance we present consist of two strategies to collapse cells and the application of a test statistic belonging to the Power-Divergence family.

So, let $X_{i1}, X_{i2}, \dots, X_{i\,n_i}, i = 1,2$ be two independent samples of points from two spatial patterns with sizes n_i, respectively. Without loss of generality, let us suppose that both samples take values in the unit-square $S=[0,1]^2$. The application of a particular space-filling curve induces a partition on S with $M = 2^v \times 2^v$ squares, where υ represents the number of iterations in the space-filling curve construction. The order of the curve induces in the space implies that the sampled points can be grouped into M classes, $C_1,C_2,\dots,C_M$, or equivalently, taking values in $\Gamma=(1,2,\dots,M)$. We assume the distribution of point falling into the M cells of the grid can be modelled by a multinomial law.

Let $\pi_i = (\pi_{i1}, \dots, \pi_{iM})'$ be the cell probabilities associated with each multinomial distribution, that is, $\pi_{im}=P[X_i=m]$, for $i=1,2$, and $m=1,\dots,M$. The application of a particular space-filling curve to both samples of points, provide us the observed frequencies on each cell for both samples, hence, the maximum likelihood estimator of π_i can be obtained as $\hat{\pi}_{im} = \frac{n_{im}}{n_i}, i = 1,2,\ m = 1, \dots, M$.

As a result, testing whether two spatial distributions are equal is equivalent to test whether two multinomial populations are equal. This fact is expressed as the following null hypothesis

$$H_0: \pi_1=\pi_2,$$
against
$$H_1: \pi_1\neq\pi_2.$$

Although several choices are possible as a test statistics for testing H_0, we have considered a test statistic belonging to the Power-Divergence family (PD) because this family contains as a particular cases several well-known test statistics, for example the χ^2 test statistic.

2.1 The test statistic

For testing H_0 we consider the following test statistic based on the PD family which is a subclass ϕf -divergence measure between two probability function (see Cressie and Read, (1984), Pardo et al., (1999), for additional information),

$$T_\lambda = 2(n_1+n_2)D(\hat{\pi}_1,\hat{\pi}_2) = 2(n_1+n_2)\sum_{i=1}^{M}\hat{\pi}_{2i}\phi\left(\frac{\hat{\pi}_{1i}}{\hat{\pi}_{2i}}\right) \quad (1)$$

where $\phi(x) = \frac{x^{\lambda+1}-x-\lambda(x-1)}{\lambda(\lambda+1)}$.

For $\lambda = 0, -1$, the test statistic is obtaining taking limit. The Pearson's chi-square, the log likelihood ratio statistic, the Freeman-Tukey statistics, the modified likelihood ratio statistic and the Cressie-Read statistic are all members of this family for λ= 1, 0, -0.5 and 2/3, respectively.

In Basu and Sarkar (1994), it is proven the test statistic T_λ has an asymptotic χ^2_{M-1} distribution under the null hypothesis of homogeneity between multinomial laws.

However, it is known (see, Kim, (2009), Alba-Fernández et al. (2009) or Jiménez-Gamero et al. (2014)) the chi-square approximation is rather poor for small and moderate sample sizes, and the approximation of the null distribution by means of a parametric bootstrap estimator behaves better than the asymptotic one.

From these results, we approximate the *p*-value by bootstrapping (see Alba-Fernández et al. (2009) for the theoretical properties of the bootstrap estimator).

3 Performance of the methodology

To evaluate the similarity of two spatial point patterns, we can proceed as follows:

- Given the point patterns, choose a space-filling curve and a value of *v* (and hence of *M*).
- Apply the space-filling curve and obtain $\hat{\pi}_i, i = 1,2$.
- Calculate $T_{\lambda,\text{obs}}$ the observed values of T_λ.
- Approximate the *p*-value by bootstrapping and by using the pooled sample to estimate π_1 under the null hypothesis.

The bootstrap algorithm to approximate the *p*-value for testing H_0 can be assessed as follows:

1. Calculate $T_{\lambda,obs}$, the observed values of T_λ.
2. For *b=1,…,B*, generate *2B* independent bootstrap samples,
$$\{X^{*b}_{1,j}\}, 1 \le j \le n_1, \{X^{*b}_{2,j}\}, 1 \le j \le n_2,$$
from the pooled multinomial distribution
$$M(n_1+n_2, \hat{\pi}_{01}, \hat{\pi}_{02}, \dots, \hat{\pi}_{0M})$$
where
$$\hat{\pi}_{om} = \frac{n_{1m}+n_{2m}}{n_1+n_2}, m = 1, \dots, M.$$
3. Calculate the values of T_λ for each couple of bootstrap samples, say T_λ^{*b}, *b=1,…,B*.
4. Approximate the *p*-value by means of
$$\hat{p} = \frac{card\left(b: T_\lambda^{*b} \ge T_{\lambda,obs}\right)}{B}.$$

In Alba-Fernández et al. (2016) it was studied the role of the space-filling curve. Their results suggest all the tried curves behave similarly and here we will only consider one of them, specifically, the z-order or Morton's curve. However, it is common to find empty cells, especially for small and moderate sample sizes of samples or when the number of cells increases.

In order to evaluate the performance of the methodology in these cases and also to analyze the usefulness of the proposed test statistics, several simulation experiments are carried out. On the one hand, to avoid empty cells, we try to collapse cells according to the order induced by the curve.

Two collapsing strategies are considered:

(1) strategy 1: assure at least one point in each new cell (collapsed cell),

(2) strategy 2: assure at least 5 points in each new cell (collapsed cell).

Both situations are studied and, at the same time, it is tried a rule between the sample size and the number of cells. On the other hand, we have considered several members of the Power-divergence family and we have compared them.

The objective of the simulation experiment is to analyze if the methodology behave properly, in other word, we carried out a simulation study about the type I error associated with the homogeneity test H_0, that is to say, if the methodology is able to detect the similarity between two point patterns when really they come from the same point process.

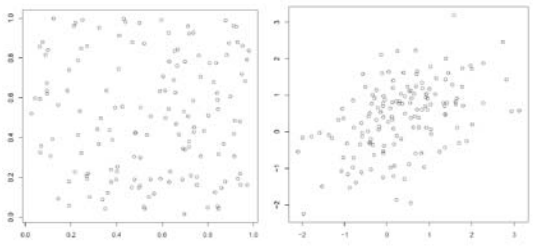
For this purpose, we have generated two uniform point patterns on the unit-square of sizes $n_1=n_2=5M$ (denoted PD1 in tables), we have applied the method described above with B=1000 bootstrap replications. We repeated this 5000 times and calculated the fraction of *p*-values less than or equal to 0.05, which is the estimated type I error probability for α=0.05.

We have repeated the whole experiment for $n_1=n_2=10M$, *15M*, *30M* and for a bivariate normal distribution with mean (0.5,0.5) and variance-covariance matrix Σ (denoted by PD2 in tables), with

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

These spatial distributions of points represent two situations quite different. The first one is used, because in most cases, it is considered that the underlying spatial pattern is uniform. The second one is taken because it represents an example of concentrated locations which is frequent in nature. Figure 2 shows a sample point pattern of size 150 of both spatial point distributions considered.

Figure 2: Spatial point distributions (PD1 on the left, PD2 on the right.



Finally, due to the number of iterations in the construction of the curve determines the level of neighborhood, the number of quadrants and the way to scan the space, three values of *v* are considered, say, *v=1,2,3*.

Tables 1 2, and 3 show the estimated type I error obtained for each value of *v*, respectively.

Table 1: Estimated type I error (*v=1*).

			Strategy 1		Strategy 2	
	n	λ	PD1	PD2	PD1	PD2
5M	20	-0.5	0.061	0.064	0.038	0.030
		0	0.056	0.060	0.041	0.028
		2/3	0.055	0.061	0.042	0.027
		1	0.045	0.057	0.043	0.030
10M	40	-0.5	0.052	0.051	0.034	0.039
		0	0.050	0.050	0.032	0.042
		2/3	0.051	0.051	0.033	0.046
		1	0.047	0.051	0.033	0.045
15M	60	-0.5	0.050	0.045	0.053	0.050
		0	0.048	0.046	0.052	0.049
		2/3	0.049	0.045	0.051	0.050
		1	0.049	0.045	0.052	0.049
30M	120	-0.5	0.053	0.052	0.048	0.051
		0	0.055	0.050	0.049	0.050
		2/3	0.053	0.052	0.049	0.048
		1	0.053	0.049	0.048	0.050

Table 2: Estimated type I error (*v=2*).

			Strategy 1		Strategy 2	
	n	λ	PD1	PD2	PD1	PD2
5M	80	-0.5	0.073	0.029	0.022	0.025
		0	0.074	0.031	0.020	0.024
		2/3	0.068	0.030	0.015	0.016
		1	0.064	0.030	0.012	0.015
10M	160	-0.5	0.050	0.042	0.024	0.024
		0	0.050	0.045	0.020	0.021
		2/3	0.048	0.048	0.016	0.018
		1	0.048	0.048	0.015	0.016
15M	240	-0.5	0.050	0.044	0.048	0.049
		0	0.053	0.045	0.047	0.046
		2/3	0.053	0.044	0.043	0.045
		1	0.050	0.045	0.044	0.045
30M	480	-0.5	0.049	0.049	0.056	0.049
		0	0.048	0.049	0.055	0.049
		2/3	0.048	0.048	0.055	0.048
		1	0.048	0.050	0.056	0.050

Table 3: Estimated type I error (*v=3*).

			Strategy 1		Strategy 2	
	n	λ	PD1	PD2	PD1	PD2
5M	80	-0.5	0.098	0.092	0.010	0.001
		0	0.095	0.091	0.009	0.009
		2/3	0.079	0.081	0.005	0.004
		1	0.073	0.076	0.003	0.002
10M	160	-0.5	0.051	0.058	0.033	0.016
		0	0.049	0.056	0.028	0.015
		2/3	0.042	0.050	0.018	0.008
		1	0.041	0.047	0.013	0.005
15M	240	-0.5	0.048	0.049	0.046	0.046
		0	0.050	0.048	0.042	0.044
		2/3	0.046	0.046	0.046	0.046
		1	0.045	0.046	0.044	0.044
30M	480	-0.5	0.048	0.050	0.049	0.051
		0	0.049	0.051	0.051	0.049
		2/3	0.050	0.051	0.050	0.049
		1	0.049	0.047	0.051	0.050

Looking at these tables, it can be concluded that the estimated type I error are close to the nominal one (α=0.5) for both tried strategies for large sample sizes. Note that in the simulation experiment, the sample sizes are taken by following a ratio *n/M*=5, 10, 15 and 30. According to this ratio and for the order of the curve tried, the results show that the strategy 1 works if the sample size in both point patterns are greater than or equal to 10 times the number of the cells of the grid. However, the strategy 2 provides acceptable behavior if the sample sizes are greater than or equal to 15 times the value of M. Note the point patterns are generated on the unit square and after applying the curve, we count the locations which falling into each cell, that is to say, we do not control the points we have in each cell.

Coming back to the example mentioned in the Introduction, due to the sample sizes in all cases are greater than 15 times *M* even for *v=3* (1786 for intentional fires, 4193 for accidental fires and 1256 for fires caused by lightning), we applied the proposed methodology for *v=3* and for both strategies.

The obtained *p*-values were 0 for all the values of λ, for both strategies and for all the possible comparisons two by two. So, it means that the spatial point patterns of the locations of fires in the region "Castilla La Mancha" can be considered clearly different attending to the cause of fires.

4 Conclusions

Our findings suggest: 1) the members of the Power-divergence family we have tried behaves similarly, 2) the methodology works well when the sample sizes are greater than 10 times the number of initial cells after assuring at least one point by each collapsed cell, 3) under these considerations, the proposed methodology is able to determine whether two point patterns can be considered similar in the sense of coming from the same spatial point distribution.

The application of the methodology to a real data set illustrates how it is possible to distinguish what point patterns can be considered similar and what of those no.

Next steps of our research will be addressed to reduce the sample sizes and try to apply a hierarchical approach in the homogeneity test.

Acknowledgements

This work has been partially supported by grant CMT2015-68276-R of MINNECO/FEDER,UE.

References

Alba-Fernández, V., Jiménez-Gamero, M.D. (2009) Bootstrapping divergence statistics for testing homogeneity in multinomial populations. *Mathematics and Computers in Simulations*, 79, 3375-3384.

Alba-Fernández, M.V., Ariza-López, F.J., Jiménez-Gamero, M.D., Rodríguez-Avi, J. (2016) On the similarity analysis of spatial patterns. *Spatial Statistics*, 18, 352-362.

Baddeley A. et al. (2017). Package 'spatstat': Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests.

Basu, A., Sarkar, B. (1994) On disparity based goodness-of-fit tests for multinomial models. *Statistics & Probability Letters*, 19, 307-312.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness of fit tests, *Journal of the Royal Statistical Society*. B, 46, 440-464.

Jiménez-Gamero, M.D., Alba-Fernández, V., Barranco-Chamorro, I., Muñoz-García, J. (2014). Two classes of divergence statistics for testing uniform association. *Statistics*, 48 (2), 367-387.

Kim, S.H., Choi, H., Lee, S. (2009). Estimate-based goodness of-fit test for large sparse multinomial distributions. *Computational Statistics and Data Analysis*, 53, 1122–1131.

Pardo, L., Pardo, M.C. and Zografos, K. (1999). Homogeneity for multinomial populations based onϕ *-divergences. Journal of the Japan Statistical Society*, 29, 213-228.

Ripley, B.D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13, 255-266.