A method for testing the similarity of spatial samples

M. Virtudes alba-Fernández[1], Francisco J. Ariza-López[2], José Rodríguez-Avi[1]

[1]Dpt. Ststistics and O.R. University of Jaén, Sapin
[2]Dpt. Cartographic engineering, geodesics and Photogrammetry. University of Jaén, Spain

*Corresponding author: mvalba@ujaen.es

Abstract

The purpose of this work is to propose a method to analyse the similarity between two spatial point patterns. Such similarity should be understood in the sense that both point patterns come from the same spatial point distribution. For this aim, first, we make use of a space-filling curve as a tool to linealize the space, with independence of its dimension, second, we model the count of points in the resultant grid by means of the multinomial law, and after that we test the homogeneity of both multinomial distributions by using the negative of the Matusita's affinity. Finally, we evaluate the performance of the procedure by means of a simulation study.

I 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.). Spatial patterns and spatial statistical sampling is also a major issue in spatial data quality assessment because sampling is a very common procedure in order to derive estimates or perform tests (Ariza-Lpez, 2002). For these reasons the evaluation of the spatial similarity of two observed samples can be of interest in many cases.

It is usual that an estimate of an attribute or property (e.g. concentration of a mineral, positional accuracy, etc.), is derived from a sample of points under some spatial distributional hypothesis for such sample (e.g. following a theoretic or an observed spatial pattern). We think that, previous to any error or uncertainty consideration about the attribute or property estimation, we must confirm the underlying hypothesis about the position of samples, in our case: the spatial distribution of the taken sample in relation to other observed spatial pattern.

Two common methods in use to investigate discrete point events are the distance- and area-based tests. In the case of distance-based tests, we use whatever distance (e.g. Euclidean) between two events in order to determine a random, clustered, or uniform spatial pattern to the points (Bailey and Gatrell, 1995). On the other hand, in the case of area-based tests, we count the number of point events within a predefined spatial area (e.g. a quadrant, a census unit, etc.) (Andresen, 2009).

Our proposal is very different to other techniques such as the Ripley's K function (Ripley, 1976). The K function is a distance-based test that characterizes point processes at many distance scales and allows the detection of different behaviors (e.g. random. clustering, inhibition) (Freeman and Ford, 2002); and it can be used to test different specific patterns (e.g. homogeneous Poisson process -complete spatial randomness-, Matern hard-core process, Strauss process, etc.) (Dixon, 2012). For the correct application of these functions it is needed to accept several assumptions that not necessarily are fully met in reality, and if conditions are not met, the output may be incorrect (Bolibok, 2008).

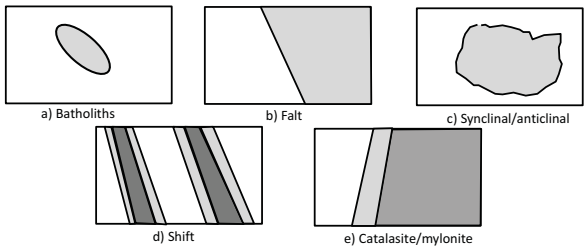

Figure 1: Examples of some geological structures that can influence spatial patterns: a) batholiths, b) falt, c) Synclinal/anticlinal, d) shift, e) cataclasite/mylonite

In this setting, we develop an area-based test centered on the counting of positional events (not attributes), which does not assume a theoretical model for the spatial distribution of the events (spatial distribution free). So, it can be applied for the comparison of two spatial samples (e.g. two control test samples, two field works, presences at two different times, etc.), with independence of sample sizes. For instance, the Figure 1 shows different geological structures that can determine the presence of point events in the space or the distribution of control samples in the space, and this method allows the comparison in such cases, and any other.

This method follows two steps; the first one is to make use of the space-filling curves (Sagan, 1994) as a tool to order the space, with independence of its dimension. The space-filling curves provide a partition of the space for a fixed level of neighborhood. The second one is to model the count of points in the resultant grid by means of the 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. For doing this, we considered as a test statistic the negative of Matusita's affinity (Matusita, 1967).

On the other hand, the spatial distribution of samples may have an effect on the representativeness of the sample. This is especially true in complex situations, although, at the end of the day, for many spatial data properties, the spatial pattern can be considered uniform. To evaluate the usefulness of the proposal, we carried out a simulation study in order to analyze the effect of the space-filling curves and the test statistics used, as well as, the effect of the sample sizes and the level of iterations in the curves. In particular, we took the Hilbert's curve and the Peano's curve, and we analyze the uniform pattern and for other spatial patterns, specially, those patterns associated with some well-known geological structures. Our findings suggest the methodology is able to detect the similarity of two spatial samples of points for all the tried cases.

This paper is organized as follows, after this introduction follows the description of the approach, where we present the statistical basis and the procedure for its application; next a simulation experiment is developed using some spatial patterns and levels of the space-filling curves. Finally conclusions are presented.

II DESCRIPTION OF THE APPROACH

As said in the Introduction, the aim of this work is to propose a formal procedure to test whether two spatial distributions of points can be considered as similar or not. By Similar one should understand that coming from the same spatial pattern. To show how this approach works, let us consider the bivarate case, although for a general dimension, the procedure follows the same steps. So that, let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be two independent samples of points from two spatial patterns with size n_i, $i = 1, 2$; and without loss of generality, let us suppose that both samples take values in the unit-square $S = [0, 1] \times [0, 1]$. The application

of a particular space-filling curve induces a partition on S with $M = 2^\nu \times 2^\nu$ squares, where ν represents the number of iterations in the space-filling curve construction. This way, for a given space-filling curve and a fixed level ν, the sampled points can be grouped into M classes, $C_1, C_2, \ldots, C_M$, or equivalently, taking values in $\Upsilon_M = (1, 2, \ldots, M)$. The order in M is induced by the space-filling curve considered. Note that the degree of neighborhood is different for each space-filling curve, specially, when the number of iterations increases.

Let $\pi_i^t = (\pi_{i1}, \pi_{2m}, \ldots, \pi_{iM})^t$ be the cell probabilities associated with each multinomial distribution, that is, $\pi_{im} = P[X_i = m]$, for $i = 1, 2$, and $1 \leq m \leq M$, and where the superscript^t denotes transpose. The application of a particular space-filling curve to both samples of points, provide us the number of points falling into each class in both samples. That is to say, we obtain the observed frequencies on each cell and hence, the maximum likelihood estimator of $\hat{\pi}_i$, say, $\hat{\pi}_i = \frac{n_{im}}{n_i}$, $i = 1, 2$, $m = 1, \ldots, M$. As a result, the problem of testing whether two spacial distributions are equal is equivalent to test whether two multinomial populations are equal. So that, our objective is to test the following null hypothesis

$$H_0 : \pi_1 = \pi_2. \tag{1}$$

For this purpose, several measures able to discriminate between multinomial populations can be used. Here, we consider a f-dissimilarity measure between multinomial populations because the smaller these measure is, the harder it is to discriminate between them. Specifically, it is considered the negative of Matusita's affinity (Matusita, 1967) defined by

$$T = -\sum_{m=1}^{M} \sqrt{\hat{\pi}_{1m}\hat{\pi}_{2m}},$$

which is a member of a class of test statistics based on f-dissimilarity (see Zoografos, 1998; Alba et al., 2009) and references therein for further theoretical results).

To decide when to reject H_0, we need to know the null distribution of T, or at least an approximation to it. Following Zoografos (1998), under H_0, the asymptotic null distribution of the test statisticn is given by

$$8\frac{n_1 n_2}{n_1 + n_2}(1 + T) \xrightarrow{\mathcal{L}} \chi^2_{M-1}. \tag{2}$$

So, we reject H_0 if

$$8\frac{n_1 n_2}{n_1 + n_2}(1 + T_{obs}) \geq \chi^2_{\alpha, M-1},$$

where T_{obs} represents the observed value of the test statistic T, and $\chi^2_{\alpha,M-1}$ denotes the $1 - \alpha$ percentil of the chi-square distribution with $M - 1$ degrees of freedom, $0 \leq \alpha \leq 1$; or equivalently, we reject H_0 if the corresponding p-value is less than or equal to α, that is, if

$$p = P_{H_0}\left[\chi^2_{M-1} \geq 8\frac{n_1 n_2}{n_1 + n_2}(1 + T_{obs})\right] \leq \alpha,$$

where P_{H_0} stands for the probability law under the null hypothesis.

As observed among others in Kim, 2009, Alba et al., 2009 or Jiménez-Gamero et al., 2014, the χ^2 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).

The application of the procedure to study of similarity of two spatial samples of points follows the steps:

(1) Given the two spatial distributions of points $\{X_{1j}\}_{1\leq j\leq n_1}$ and $\{X_{2j}\}_{1\leq j\leq n_2}$, choose a space-filling curve and the number of the iterations in its construction, ν. These decisions will determine the value of M.
(2) Apply the space-filling curve and obtain $\hat{\pi}_i$, $i = 1, 2$.
(3) Calculate T_{obs}, the observed values of T.
(4) Approximate the corresponding p-value by bootstrapping.
(5) Conclude if the sample spatial distributions can be considered equal or not.

In addition, the bootstrap algorithm to approximate the p-value for testing (1) can be assessed as follows:

(i) Calculate T_{obs}, the observed values of T.
(ii) For $b = 1, \ldots, B$, generate $2B$ independent bootstrap samples, $\left\{X^{*b}_{1,j}\right\}_{1\leq j\leq n_1}$, $\left\{X^{*b}_{2,j}\right\}_{1\leq j\leq n_2}$ from the pooled multinomial distribution $M(n_1 + n_2; \hat{\pi}_{01}, \ldots, \hat{\pi}_{0M})$ where

$$\hat{\pi}_{0m} = \frac{n_{1m} + n_{2m}}{n_1 + n_2},\ m = 1, \ldots, M.$$

(iii) Calculate the values of T for each couple of bootstrap samples, say T^{*b}, $b = 1, \ldots, B$.
(iv) Approximate the p-value by means of $\hat{p} = card\left\{b : T^{*b} \geq T_{obs}\right\}/B$, respectively.

III SIMULATION EXPERIMENT

To evaluate the performance of this method, we have carried out a simulation study. The goal of this experiment is twofold, the first objective is to analyze the behaviour of the proposed methodology for small sample sizes with respect to the type I error for some space-filling curves and levels of sweep. In other words, if the procedure is able to conclude if two sample of points come from the same spatial pattern when they really do.

Together with this, the second task is to evaluate the power of the procedure, that is, if this method is able to detect two samples of point which have been generated by different spatial patterns. Next we briefly describe the simulation experiment and display the results obtained. All computations in this paper have been performed using scripts written in the R language (http://www.cran.r-project.org).

The methodology we propose is independent of the spatial pattern of sample of points, however, we are going to evaluate its behaviour by considering some spatial patterns related to the geological structures mentioned before (see figure 1). In particular, we will identify the uniform pattern with Pattern 1, the bivariate normal $N(\mu, \Sigma)$ with $\mu = (0.25, 0.25)$, $\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}$ simulates a synclinal/anticlinar(Pattern 2) and the bivariate normal $N(\mu, \Sigma)$ with $\mu = (0.5, 0.5)$, $\Sigma = \begin{pmatrix} 0.25 & 0.8 \\ 0.8 & 0.25 \end{pmatrix}$ reproduces a Batholiths (Pattern 3) (see Figure 2) for a scatter plot of a sample of size 150).

So, we have generated two uniform samples on the unit-square of sizes $n_1 = n_2 = 25$, we have applied the method following the steps (1)-(5) described above with $B = 1000$ bootstrap

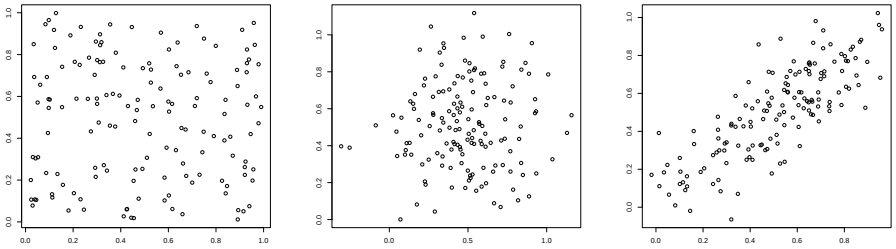

Figure 2: Example of Pattern 1 (left), Pattern 2 (center) and Pattern 3 (right).

replications. We repeated this 1000 times and we calculated the fraction of p-values less than or equal to 0.05 and 0.10 (denoted as $f05$ and $f10$ in tables), which are the estimated type I error probabilities for $\alpha = 0.05, 0.10$, respectively. We have repeated the whole experiment for the sample sizes $n_1 = n_2 = 50, 150$. Among the available space-filling curves, we have considered the Peano's and the Hilbert's curve (identified as n and h in tables). Note the value of M is related to the sample size and ν, and we try to find an agreement between them. Here, we try the values $\nu = 1, 2$. We have repeated this experiment for Patterns 2 and 3. The estimated type I error probabilities are shown in Table 1.

Looking at this table, we can conclude that the estimated type I error probabilities are quite close to the nominal ones in all the tried cases and the simulations results do not show differences between the space-filling curves. On the other hand, we have also studied if the methodology is able to distinguish between spatial patterns, and for this task we generated a sample of points following the Patter 1 of size $n_1 = 25$ and other sample of points from Pattern 2 of size $n_2 = 25$. We applied the methodology in the same conditions as before. We repeated the whole experiment 1000 times and we calculated the fraction of $p's$ values less than or equal to $0.5, 0.10$, which are the estimated power for $\alpha = 0.5, 0.10$ (now, $f05$ and $f10$ in tables). We repeated the experiment for samples of points following the Patterns 3. The estimated powers are shown in Table 2. From these results, we can say the procedure is able to distinguish clearly between two different Patterns.

$n_1 = n_2$	ν	curve	Pattern 1		Pattern 2		Pattern 3	
25	1		f05	f10	f05	f10	f05	f10
		n	0.042	0.099	0.052	0.100	0.048	0.096
		h	0.047	0.102	0.052	0.096	0.050	0.096
50	1		f05	f10	f05	f10	f05	f10
		n	0.055	0.103	0.053	0.105	0.054	0.106
		h	0.053	0.101	0.053	0.101	0.053	0.107
150	1		f05	f10	f05	f10	f05	f10
		n	0.054	0.104	0.058	0.107	0.046	0.093
		h	0.057	0.099	0.059	0.104	0.045	0.098
150	2		f05	f10	f05	f10	f05	f10
		n	0.056	0.104	0.048	0.094	0.060	0.108
		h	0.057	0.103	0.048	0.093	0.057	0.107

Table 1: Estimated type I error probabilities.

$n_1 = n_2$	ν	curve	Pattern 1 vs. Pattern 2		Pattern 1 vs. Pattern 3		Pattern 2 vs. Pattern 3	
25	1		f05	f10	f05	f10	f05	f10
		n	0.429	0.548	0.409	0.534	0.570	0.665
		h	0.427	0.557	0.366	0.501	0.537	0.635
50	1		f05	f10	f05	f10	f05	f10
		n	0.747	0.837	0.719	0.802	0.886	0.938
		h	0.748	0.837	0.715	0.802	0.881	0.928
150	1		f05	f10	f05	f10	f05	f10
		n	0.994	0.995	0.998	0.999	0.999	1.000
		h	0.995	0.995	0.998	0.999	0.999	1.000
150	2		f05	f10	f05	f10	f05	f10
		n	0.987	0.994	1.000	1.000	1.000	1.000
		h	0.989	0.995	1.000	1.000	1.000	1.000

Table 2: Estimated power.

IV CONCLUSION

To sum up, the procedure introduced in Section 2 takes advantage of the use of space filling curves as a way to linearize spatial distributions, and following the order induced by the space filling curve for a fixed level of its construction, the count of the number of points falling into each grid may be modeled by a multinomial distribution. For testing the homogeneity of two multinomial laws, the negative of Matusita's affinity is considered. Finally, the results of the simulation experiment reveals that the proposed method provide us a statistical tool in order to decide about the similarity of spatial samples.

ACKNOWLEDGMENTS

Research in this paper has been partially funded by grant CTM2015-68276-R of the Spanish Ministry on Science and Innovation.

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.

Andresen M.A. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. Applied Geography 29, 333345.

Ariza-Lpez F.J. (2002). Calidad en la produccin cartogrfica. Madrid, ES, Ra-Ma.

Bailey T.C., Gatrell, A.C. (1995). Interactive spatial data analysis. Harlow, UK: Prentice Hall.

Bolibok L. (2008). Limitations of Ripleys k(t) function use in the analysis of spatial patterns of tree stands with heterogeneous structure. Acta Sci. Pol. Silv. Colendar. Rat. Ind. Lignar 7(1), 5-18.

Dixon P.M. (2012). Ripley's K function. Encyclopedia of Environmetrics, 2nd ed. John Wiley and Sons, Inc.

Freeman E.A., Ford E.D. (2002). Effects of data quality on analysis of ecological pattern using the k(t) statistical function. Ecology 83(1), 3546.

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.

Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. Annals of the Institute of Statistical Mathematics, 19,181–192.

Sagan H. (1994).Space-Filling Curves. Springer-Verlag.

Ripley, B.D. (1976). The second-order analysis of stationary point processes. Journal of Applied Probability 13, 255266.

Zografos, K. (1998). f-dissimilarity of several distributions in testing statistical hypotheses. Annals of the Institute of Statistical Mathematics, 50, 295-310.