Juan A. Barceló – Alfredo Maximiano

Some Notes Regarding Distributional Analysis of Spatial Data

Abstract: The purpose of geostatistics and other quantitative spatial analysis methods is the characterization of the processes having generated the spatial distribution of archaeological data. In this paper[1] we investigate whether such methods can be used to distinguish the regularity or randomness of the social event or events having generated the observed spatial distribution. Our hypothesis is that only statistically significant deviations from spatial randomness can be interpreted as intentional clustering. Archaeological data distributions are best characterized in terms of spatial processes which are symmetrical around a central mean.

Introduction

Spatial data can be defined in two different ways. Distance-based data are given as series of bidimensional coordinate points. Frequency-based data are given as sums of points at discrete spatial regions. In fact, frequency-based data can be seen as a transformation of an original distance-based distribution, just by overlying a well defined grid and counting the number of points within each grid. In this paper we consider only the case of distance-based, that is, data points with coordinates, where each point represents the spatial location of an archaeological entity.

In any case, both point patterns and grid counts are a measure of spatial frequency. We consider the *spatial frequency* aspect of archaeological data when we describe them as an *accumulation* of some material items on the ground surface where the action took place, or as the *intensity* of the action. Obviously, this is not the only way spatial data can be analyzed. We have considered shape and interfacial boundaries elsewhere (Barceló 2002; idem 2005; Barceló / Maximiano 2007; Barceló / Maximiano / Vicente 2005; Barceló et al. 2003; Maximiano 2005; Vicente 2005), consequently we restrict here to the analysis of spatial frequencies.

Formally, spatial densities may be thought of as consisting of a set of locations (s_1, s_2, etc.) in a defined "study region", R, at which the material consequences of some social action performed in the past (archaeological event) have been recorded. The use of the vector s_i, referring to the location of the i_{th} observed event, is simply a shorthand way of identifying the 'x' coordinate, s_{i1}, and the 'y' coordinate, s_{i2}, of an event.

We can assume that the probability that a social action occurs at a specific location should be related someway to the frequency of its material effects (the archaeological record) at nearby locations. Therefore, when the frequency of the archaeological feature at some locations increases, the probability that the social action was performed in its neighborhood will converge towards the relative frequency at adjacent locations. Then, assuming that a measure of spatial density is a function of the probability an action was performed at that point, we will say that the area where spatial density values are more continuous is the most likely place where a social action was performed (Barceló / Maximiano 2007). This can be easily computed by estimating the spatial probability density function associated to each location. If we know the relationship between the social action and its archaeological descriptor, the density

[1] This research has been funded by the Spanish Ministry for Education and Research (Research Grant HUM2006-01129/HIST). Alfredo Maximiano also acknowledges his grant from the Program of Formation of Investigators F.I. 2007, managed by AGAUR (Generalitat de Catalunya). We also thank our research colleagues from Consejo Superior de Investigaciones Científicas and Universitat Autonoma de Barcelona. This paper constitutes a part of an ongoing joint project on the Archaeology of Coastal and Marine Environments between both institutions funded by the Spanish Ministry for Education and Research, and the Catalan Government Commision for Research (Research Grant GRS 00829 awarded to the AGREL Research group).

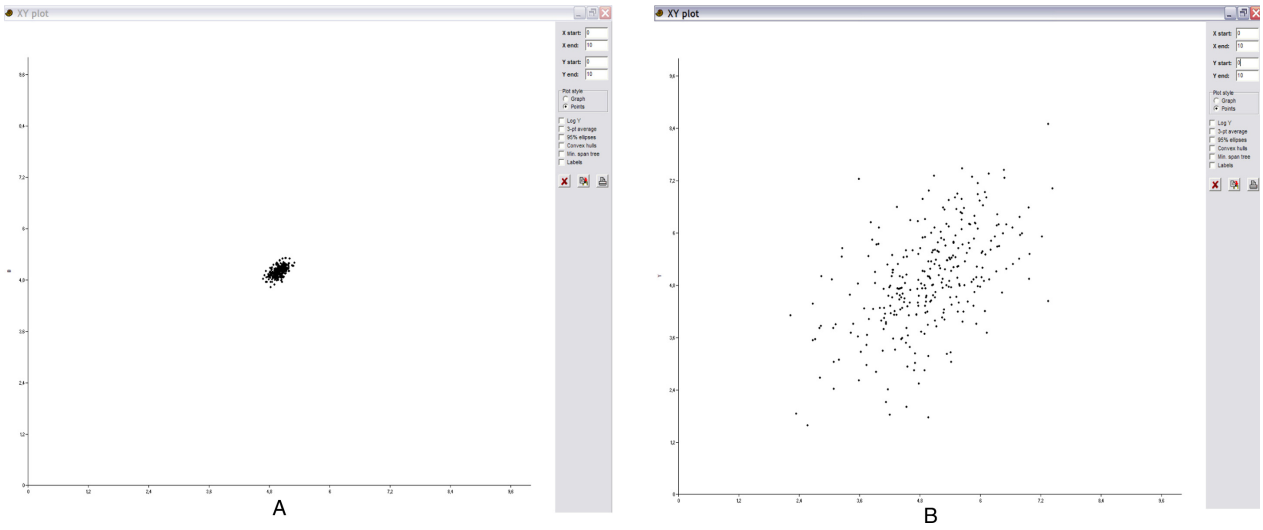

Fig. 1. Two of the simulated bivariate normal archaeological distributions. 1A: St. Dev = 0.1; 1B: St. Dev = 1.

probability function for the location of archaeological artifacts can be a good estimator for the spatial variability of the social action.

To infer the cause (social action performed at the household level) from the effect (the frequency of material evidences measured at some finite set of locations), we have to rebuild the real frequency that was generated in the past by the social action. This theory forms the underpinnings of geostatistics. Geostatistics applies the theories of stochastic processes and statistical inference to spatial locations. It is a set of statistical methods used to describe spatial relationships among sample data and to apply this analysis to the prediction of spatial and temporal phenomena (Fotheringham / Brunsdon / Charlton 2000; Haining 2003; Lloyd / Atkinson 2004).

The question that also arises is whether the spatial process displays any systematic spatial pattern or departure from randomness. Spatial questions of interest to archaeologists include:

- Is the observed clustering due mainly to natural background variation in the population from which intensities arise?
- Over what spatial scale does any clustering occur?
- Are clusters merely a result of some obvious a priori heterogeneity in the region studied?
- Are they associated with proximity to other specific features of interest, such the location of some other social action or possible point sources of important resources?
- Are frequencies that aggregate in space also clustered in time?

Distributional Analysis

We need tools and methods to differentiate the specific spatial ways that an action can be performed at different places. In archaeology, we can speak about two spatial modalities for the material effects of social action to be distributed: *regularity vs. randomness.* In some way, intentionality at the spatial level produces the regular spatial distribution of the material effects of the social action, whereas, non-intentionality generates random patterns of locations. These are the opposite extremes of the global range of spatial modality.

We can apply the theories of stochastic processes and statistical inference to analyze spatial modalities. The theoretical bivariate normal distribution can be used to represent the formation of spatially regular modalities of social action (Mardia 1970; Rose / Smith 1996; Kotz / Balakrishnan / Johnson 2000). Note that the multivariate normal distribution is not a mere composite of univariate normal distributions. Previous tests suggest that in order an observed distribution fits a bivariate normal, x and y must be moderately correlated.

Is the bivariate normal distribution the best way to describe archaeological spatial distributions? Obviously, this is a theoretical model, and only in ideal circumstances, observed data fit the model predictions. Such a theoretical distribution allows us to predict the probability of locating some material effects of an action at different distances from the place where the action was hypothetically performed. If and only if, the spatial modality of a social action performed in the past was *regular*, and its material

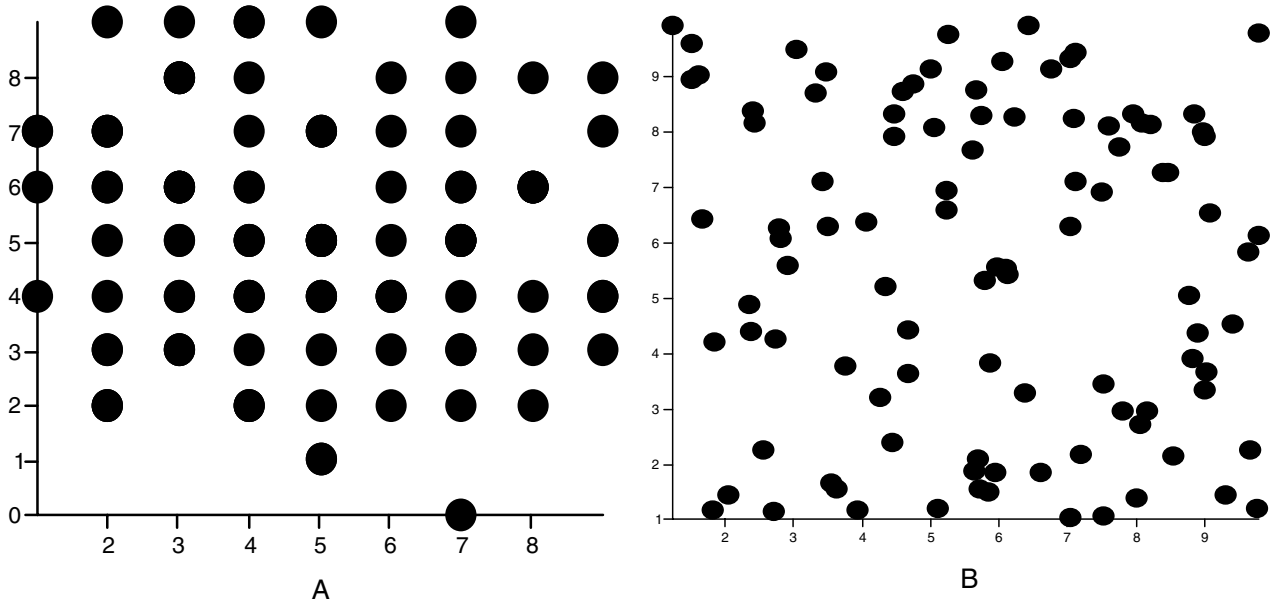

Fig. 2. Non-bivariate normal simulated archaeological distributions. 2A: Uniform; 2B: Random.

consequences have not been altered in a significant way by post-depositional processes, archaeologically measured spatial frequencies will fit the bivariate normal distribution. The basic law in geography, Tobler's Law is the basis for such an assumption: near things appear to be more related than distant things, when an action has been intentionally performed at a precise location. Bivariate normal distributions offer a reference model to test the degree of regularity and hence of spatial intentionality of social action.

When spatial analysis methods were applied in archaeology in the 70s and 80s (Hodder / Orton 1979; Blankholm 1991), archaeologists began to look for spatial clusters and groups assuming that archaeological data were always regular. The purpose of this paper is to insist in the necessity of distributional analysis to assert the quality of spatial data, and the relevance of resulting spatial classifications as a model of social action in space.

We have randomly generated a series of different bivariate normal populations[2] of locations using the same mean and different standard deviations (*Fig. 1*). Here the mean refers to the place the action was performed, and the standard deviation estimates the intensity of distance differences in locating the material effects of such an action. In this case, we have used the same correlation coefficient for all data sets. Only the standard deviation varies, generating different concentration patterns within the same regular modality. Additionally, non bivariate normal distributions have been generated for comparison purposes. *Fig. 2* shows a uniform and a random distribution. Note the difference between both. A uniform pattern is a regular pattern, but without the characteristic aggregated pattern of bivariate normal distributions. Here locations are equally distributed all along the studied area. Although statistical literature uses the terms uniform and random distribution synonymously we wanted to experiment with different social processes generating different spatial patterns. Our hypothesis is that intentionality in space produces non-random distributions ranging between aggregated (bivariate normal) and uniform distributions. Therefore the three types of spatial modalities should be compared.

Spatial regularity can be tested using Mardia's skewness and kurtosis multivariate test (Mardia 1985). Testing bivariate normality conditions in a distribution of observed locations allows us to distinguish between two regular patterns (aggregated, uniform) and one general random pattern (Cox / Small 1978; Smith / Jain 1988; Curran / West / Finch 1996). In these data, kurtosis decreases proportional to the increase in standard deviation of the distribution. That means that as soon as spatial entropy increases, concentration decreas-

[2] Simulated data were generated using the Stats4U package (http://www.statpages.org/miller/openstat/Stats4U.htm).

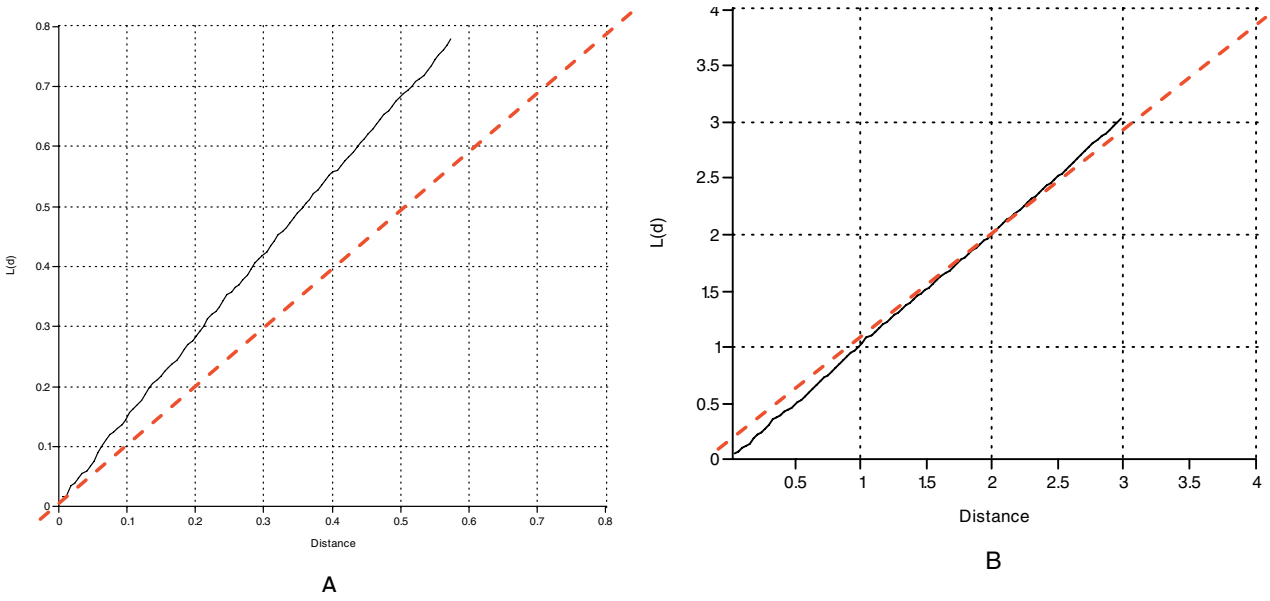

Fig. 3. Ripley's L(d) function. (A) Simulated bivariate normal data; (B) Simulated random data.

es, without affecting the regular modality of that distribution.

Years ago, instead of bivariate normality tests, spatial regularity was investigated using nearest neighbor tests (Clark / Evans 1954; Hodder / Orton 1979; Hammer / Harper 2006). However, these tests were soon rejected because results can vary depending on the way the area delimited by the most eloigned points has been measured. To solve this problem, we have fixed the dimensions of the analyzed area using different geometrical approaches: the convex hull and the smallest rectangle. In this way, the studied area has always well defined limits, and the decision of its extent is not left to the analyst. Preliminary ethnoarchaeological observations suggest that socially defined areas (huts, houses, etc.) coincide with this geometrically defined boundaries. Using our test data, this corrected version of the traditional nearest neighbor tests concurs with the spatial normality test: bivariate normal distributions deviate strongly from the random assumption.

Ripley's *L(d)* function has been used to compare the aggregated point pattern with point patterns generated by a random process (Orton 2005; Schabenberger / Gotway 2005). This procedure compares the number of points within any distance to an expected number for a spatial random distribution (Conolly / Lake 2006, 166–168). The empirical count is transformed into a square root function, called *L*. The distance at which the estimated counts are above the random expectation (in *Fig. 3* it has been represented as a dashed line) defines the extent of the clustering. Here, our bivariate normal simulated data are significantly non-random; the data appear clustered much more than expected under Complete Spatial Randomness. Even more, with Ripley's *L(d)* function, the aggregation becomes more and more evident when increasing the distance, at least for scales below 1 m, which is a logical assumption in intra-site analysis. This result is obvious given that the bivariate normal data we have simulated consisted of 300 points in an area of 56 m^2, and standard deviation between points was fixed at 1.5.

Once the non-randomness of the spatial distribution of archaeological finds has been tested, we can proceed to examine its relationship with spatial clustering. We have added four different bivariate normal processes with some minor variations in their respective mean, building a spatial distribution that can be clustered into four different groups. Are those spatial clusters a random effect or can they be defined as differentiated areas within the global distribution?

We have tested the bivariate normality and the spatial randomness of the entire population. As we would expect, the global distribution is significantly non multivariate normal, and it is also not random. Each individual spatial class is, however, bivariate normal.

The discovery of spatial clusters should be based on detecting the spatial influence each observation has on its neighbors and also on the global spatial variance within the study area. The idea is to investigate the possibilities of relevant discontinuities in the general distance pattern. If such discontinuities

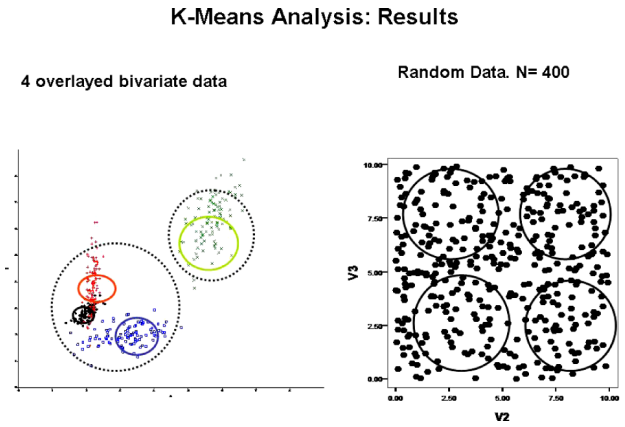

Fig. 4. K-means analysis of four overlying bivariate samples (left) and one single random sample (right).

exist and each one defines a regularly distributed group of spatial observations, then we would conclude that spatial clustering is an effect of the causal event. In most occasions, randomness should be related with the inconsistencies of archaeological observation and spatial location measuring.

We have used *k-means* analysis for detecting spatial clusters (Kintigh / Ammerman 1982; Blankholm 1991) *(Fig. 4)*. To test the efficiency of the method, we have generated a random distribution of points with the same mean and standard deviation. In the first case, the clustering algorithm correctly generates two differentiated spatial areas, and effectively subdivides the first one into other three sub-areas. In the other, when using the k-means on a random distribution of points, the algorithm does not detect the random nature of the data and tries to impose four groups, which only resume the total variance in four equally distributed clusters.

The consequence is obvious. We should restrict the use of k-means analysis to non-random data, and the analysis of spatial normality is a necessary prior condition before subsequent spatial interpretations of archaeological data.

Conclusions

Different social actions can have the same spatial modality, and the same actions can be spatially performed in different ways at different moments or different places. Therefore, testing regularity and randomness in archaeological field data is not the only approach to interpret archaeological field data, but they become a necessary previous requirement before more sophisticated interpretations.

The main conclusion addressed by this paper is that randomness at the spatial level should be detected before social action at the spatial level is explained. Spatial normality tests and nearest neighbor statistics can be used for this purpose. These tools are well known in the archaeological literature, but the modern fashion of GIS visualization has neglected the previous examination of data quality and necessary assumptions prior to interpretation.

Obviously, bivariate normal distributions are not the only possibility for representing spatial modalities. We are experimenting with other assumptions, like bivariate exponential distributions, which can be used to simulate cleaning patterns; or multimodal distributions, which can be used to simulate social interaction patterns. In any case, the importance of the bivariate normal assumption lies in the fact that intentional social processes are best characterized in terms of symmetrical spatial distributions around a central mean. The idea is that an event took place at a specific location, where the social event material effects are concentrated, and around this central point, the spatial frequency of other material effects diminishes gradually. Spatial frequency decreases proportional to distance. Non-intentional processes are best characterized in terms of random distributions, where each location has the same frequency and no central point can be identified.

We are also studying whether spatial randomness can be the result, in some limited circumstances, of intentional social activity. Much more work on the spatial modalities of social action at a household level is still necessary. We think that geostatistical analysis of ethnoarchaeological data can be useful in this task.

In this paper, we have restricted our investigation to the analysis of spatial frequencies. The analysis of shape and interface boundaries spatial data requires other approaches that have been published elsewhere.

References

Barceló 2002

J. Barceló, Archaeological Thinking: between space and time. Archeologia e Calcolatori 13, 2002, 237–256.

Barceló 2005

J. Barceló, Multidimensional Spatial Analysis in Archaeology: Beyond the GIS Paradigm. Paper presented at the GIS Symposium "Reading the Historical Spatial Information in the World": Studies for Human Cultures and Civilizations based on Geographic Information System, Kyoto, Japan, February 7–11, 2005

(Kyoto 2005). http://antalya.uab.es/prehistoria/Barcelo/publication/Kyoto2005.pdf [30 Sep 2008].

Barceló / Maximiano 2007
J. Barceló / A. Maximiano, The Mathematics of domestic spaces. Paper presented at the Archaeology of the Hosehold Workshop (Barcelona 2006). http://antalya.uab.es/prehistoria/Barcelo/publication/mathdomspaces.pdf [30 Sep 2008].

Barceló / Maximiano / Vicente 2005
J. Barceló / A. Maximiano / O. Vicente, La Multidimensionalidad del Espacio Arqueológico: Teoría, Matemáticas, Visualización. In: I. Mira (ed.), La Aplicación de los SIG en la Atqueología del paisaje (Alicante 2005) 29–40.

Barceló et al. 2003
J. Barceló / O. De Castro / D. Trevet / O. Vicente, A 3D Model of an Archaeological Excavation. In: M. Doerr / A. Sarris (eds.), The Digital Heritage of Archaeology. CAA 2002. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 30th CAA Conference, Heraklion, Crete, April 2002 (Athens 2003).

Blankholm 1991
H. Blankholm, Intrasite Spatial Analysis in Theory and Practice (Aarhus 1991).

Clark / Evans 1954
P. Clark / C. Evans, Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. Ecology 35,4, 1954, 445–453.

Conolly / Lake 2006
J. Conolly / M. Lake, Geographical Information Systems in Archaeology (Cambridge 2006).

Cox / Small 1978
D. Cox / N. Small, Testing multivariate normality. Biometrika 65,2, 1978, 263–272.

Curran / West / Finch 1996
P. Curran / S. West / J. Finch, The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological Methods 1, 1996, 16–29.

Fotheringham / Brunsdon / Charlton 2000
A. Fotheringham / C. Brunsdon / M. Charlton, Quantitative geography: perspectives on spatial data analysis (London 2000).

Haining 2003
R. Haining, Spatial data Analysis: Theory and Practice (Cambridge 2003).

Hammer / Harper 2006
O. Hammer / D. Harper, Paleontological data analysis (Oxford 2006).

Hodder / Orton 1979
I. Hodder / C. Orton, Spatial Analysis in Archaeology (Cambridge 1979).

Kintigh / Ammerman 1982
K. Kintigh / A. Ammerman, Heuristic approaches to spatial analysis in archaeology. American Antiquity: Journal of the Society for American Archaeology 47, 1982, 31–63.

Kotz / Balakrishnan / Johnson 2000
S. Kotz / N. Balakrishnan / N. Johnson, Bivariate and Trivariate Normal Distributions. In: W. A. Shewhart / S. S. Wilks (eds.), Continuous Multivariate Distributions 1: Models and Applications (New York 2000) 251–348.

Lloyd / Atkinson 2004
C. Lloyd / P. Atkinson, Archaeology and geostatistics Journal of Archaeological Science 31, 2004, 151–165.

Mardia 1970
K. Mardia, Families of bivariate distributions (London 1970).

Mardia 1985
K. Mardia, Mardia's Test of Multinormality. In: S. Kotz / N. Johnson (eds.), Encyclopedia of Statistical Sciences 5 (New York 1985) 217–221.

Maximiano 2005
A. Maximiano, Métodos geocomputacionales aplicados al análisis espacial en arqueología (Barcelona 2005).

Orton 2005
C. Orton, Point pattern analysis revisited. Archaeologia e Calcolatori 15, 2005, 299–315.

Rose / Smith 1996
C. Rose / M. Smith, The Multivariate Normal Distribution. The Mathematica Journal 6, 1996, 32–37.

Schabenberger / Gotway 2005
O. Schabenberger / C. Gotway, Statistical Methods for Spatial Data Analysis (Boca Raton 2005).

Smith / Jain 1988
S. Smith / A, Jain, A test to determine the multivariate normality of a dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence 10, 1988, 757–761.

Vicente 2005
O. Vicente, La aplicación de las nuevas tecnologías de visión computacional en el registro y modelización de yacimientos arqueológicos (Barcelona 2005).

Juan A. Barceló
Alfredo Maximiano

Universitat Autonòma de Barcelona
Departament de Prehistoria
Edifici B
08193 Bellaterra (Cerdanyola del Vallès)
Barcelona
Spain
amaximiano77@msn.com