**GEOSTATISTICAL MODELLING USING NON-EUCLIDEAN DISTANCES AND SPATIAL DEMARCATIONS. A CASE STUDY FROM THE ANALYSIS OF CHEMICAL SOIL COMPOSITION IN ARCHAEOLOGICAL FLOORS IN TEOTIHUACÁN (MÉXICO).**

Joan NEGRE PÉREZ (corresponding author)

Austral Centre for Scientific Research

National Research Council for Scientific and Technological Research

B. Houssay, 200 E-9410 Ushuaia (Tierra del Fuego, Argentina)

negreperez@gmail.com

Facundo MUÑOZ

UR0588 Unité Amélioration Génétique et Physiologie Forestières,

INRA Val de Loire Orléans, 2163 avenue de la Pomme de Pin,

CS 40001 Ardon, Orléans Cedex 2, F-45075, France

facundo.munoz@orleans.inra.fr

*ABSTRACT*

*KEYWORDS*

*HIGHLIGHTS*

**1. INTRODUCTION**

The analysis of chemical soil composition in archaeological domestic floors is becoming increasingly considered as an important topic for historical research. Mapping the distribution of certain elements allows us to understand the activities that were developed in these areas. This assumption comes from the idea that different social actions of production, consumption or distribution are the cause of variations observed in the material consequences detected through fieldwork. In this case, the variability on chemical soil composition is a valuable marker in order to detect, identify and analyse different activities in domestic contexts. Thus, if the real value of archaeological observations comes from our ability to extract meaningful information from them (Barceló 2015: 6), methodological improvements must be encouraged. One of those focuses on surpassing the current use of deterministic and geostatistical methods that assume a homogeneous, unrestricted space of analysis. The main goal of this paper, therefore, is both to consider this major limitation of current interpolation techniques and present a new framework in which develop more accurate analyses.

Geostatistics provides a set of statistical tools specifically designed for spatial problems, in which prediction is required over a region of interest where some observations have been taken. Predictions are based on an underlying statistical model that can take additional information into account as explanatory variables. In addition, the prediction error can be estimated based on propagation of uncertainty. The main drawback with these tools is that they consider, as above mentioned, an undifferentiated surface that is easy to generalise. This assumption fail when we consider the spatial demarcations (barriers, walls, etc.) and topography (slope, roughness, etc.) that affect the distribution of our phenomena on the study region.

In this work, we develop a methodology for overcoming this problem. We propose the use of cost-weighted distances to quantify the correlation between locations. In this way, we take into consideration the heterogeneous configuration of the inside of a domestic area.

**2. GEOSTATISTICS AND NON-EUCLIDEAN METRICS**

Geostatistics is a branch of statistics that encompasses the techniques that apply to geographical analysis. We owe its origins to the works of D.G. Krige (1951) and G. Matheron (1963) in the central decades of the 20th century. There are several applications of geostatistical methods in a very wide range of disciplines, with the common problem of modelling a Stochastic Process over a continuous spatial region from a partial group of observations. This process is commonly assumed to be Gaussian, isotropic and intrinsically stationary. Geostatistical modelling is based on the principle of spatial dependence or autocorrelation, this is near events are more related than distant ones. But what means near in this context and how we calculate it?

These methods assume that the correlation between the elements of a group of observations is a function of the Euclidean distance between them. In other words, stationarity is often accepted to mean that the spatial point process has constant intensity and uniform correlation depending only on the lag vector between pairs of points (Møller and Toftaker 2012). Considering the inherent heterogeneity of geographical terrain, either the presence of barriers or the difficulty to cross a region are presented as a major problem for this technical requirement. Imagine two locations at a given (Euclidean) distance such that they are significatively correlated, because of underlying relevant factors affecting both of them. Now put a barrier between them that blocks or absorbs the effect of the underlying factors. This obviously pulls the correlation down. Therefore, when some kind of barriers exist, the correlation depends on something other than this kind of metrics, which therefore cannot account for correlation by itself.

Methodologically, the first step in geostatistical processing is to fit the data and its empirical semivariogram function to a known parametric model. There is a variety of methods for estimating this correlation. Our approach here is to use maximum likelihood methods that fits the mean function and the parameters of the semivariogram function. Once fitted, the main interest for analysing purposes is spatial prediction. Kriging is one of the most used approaches to this problem, in which a weighted average of the sample values is used to generate the prediction. This is sample points near the prediction's location are given larger weights than those far away. Kriging determines these weights based on the semivariogram function, calculating them according to the value of the semivariogram, which is a function of the Euclidean distance (López-Quílez and Muñoz 2008: 11-12). That seems to incur into the above mentioned error of assuming the validity of the spatial homogeneity premise. Thus, in certain cases, alternative measurements to Euclidean metrics, such as cost-based, Riemannian or pseudo-Euclidean ones, represent the distance argument r of the semivariogram function more naturally.

2.1 ALTERNATIVE MEASUREMENTS TO EUCLIDEAN METRICS

Alternative measures to Euclidean distances have been largely tested in several disciplines. A multidimensional scaled reconfiguration of the spatial distribution has proved to be very useful in some cases. This technique allows us to create a pseudo-Euclidean framework on which our analysis can be performed (Løland and Høst 2003; Negre 2015). A fast Fourier Transform has also been checked for integrating moving-average functions that may be used to generate a large class of valid, flexible variogram models. This transform allows us both to compute the cross-variogram on a set of discrete lags and to interpolate the cross-variogram for any continuous lag (Ver Hoef et al. 2004). Other approaches are still being tested for solving this kind of issues. For example, we can currently read recent works about using Riemannian metrics associated to cost-based distances and Banach algebra using Kuratowski immersion (Muñoz 2012: 118). For its relative simple implementation, the use of cost-based distances directly into the covariance matrix of the Kriging, has proved to be a practical and competitive option for our research topic.

From a methodological perspective, the main goal of this kind of cost measurements is to define the least cost path to reach a known point from each cell location in the original raster dataset. The cost-weighted distance algorithms, therefore, calculate the length of the irregular vectors formed by a spatial distribution using the shortest weighted distance; this is the path with least accumulated cost. There is a large variety of these algorithms in order to represent a friction or cost surface. The basic purpose of these is to assign an impedance value to each cell of a raster layer, that is, the ease with which it can be crossed. Formally, the resulting cost-weighted model can be defined as an *f* function, which describes for each cell of our model a real, positive value representing the difficulty to go through them, that is, its cost-weighted density. Therefore, the cost-weighted movement to the point is . From this function the cost of any path in A can be calculated as the integration of every cell in the model which is gone through (Muñoz 2012: 55). Thus, the cost of a path α between points will be

Once the use of relative distances proves to be an appropriated choice to solve the least cost path between enumerated settlements, it must be proved that this approach can also be applied to geostatistical functions. In any case, anisotropically cost-weighted distances maintain the same general properties than their metric counterparts (Waller and Gotway 2004: 321):

* non-negativity
* symmetry
* triangle inequality

Despite the fact that the application of cost-weighted distances in geostatistical functions involves a work of adaptation, the results might be, in some cases, statistically non-relevant. The more homogeneous is the surface under study, the less significant are the changes with respect to the use of Euclidean measures.

There are three major stages in classical geostatistical analysis computation that need to be adapted: empirical variogram computation, variogram model parameter fitting and the actual kriging prediction. Apart from observation data and prediction locations needed for standard kriging, we also need two Cost-Based distance matrices previously computed. One holds the distances between observation points, a symmetric square matrix. The second matrix holds the distances between observation points and the prediction location(s), so it is an *n* (observations) x *m* (locations) sized matrix. The empirical variogram is computed from the observation data only. It classifies pairs of observations into groups according to their distance, and then computes an estimator of the theoretical variogram value for that distance based on the differences between the observed values.

In order to make a Cost-Based empirical variogram it is enough to make the initial classification based on the Cost-Based distance values given in the corresponding matrix, rather than calculating Euclidean distances. Note that this modification produces a different grouping of observation pairs. Therefore, variogram estimates will be different. The variogram model parameter fitting is also made based on observation data only. It is typically accomplished through maximum likelihood methods, basically trying out many possible combinations iteratively and keeping the best. This implies computation of the covariance matrix for each combination being tested. All we need is to make sure that the covariance matrix is computed based on the Cost-Based distances provided by our previously computed matrix. Finally, there is the kriging prediction. At this point, the covariance model is assumed to be known. But here again, we need to make sure that the covariance matrix of the observations is computed with the Cost-Based distances. In addition, the covariance between observation points and prediction locations are to be computed in order to make predictions. So this is when the second of the Cost-Based distance matrices is to be used.

To obtain mathematical validity, not just argumentative, results obtained must be checked with the positive definition of the covariance matrix of the observations. This condition requires that for any number, set of locations and complex set of coefficients , the function verify the next relationship:

where represents the cost-weighted distance between their arguments (Muñoz 2012: 201). Ultimately, and provided verification of the above validation factors, a functional model can be described providing the best linear unbiased prediction.

**3. COST-BASED GEOSTATISTICS: METHODOLOGICAL OVERVIEW**

**4. ARCHAEOLOGICAL CHEMICAL SOIL ANALYSIS IN THE PRESENCE OF BARRIERS**

4.1 GEOSTATISTICS AND ARCHAEOLOGY: A STATE OF THE ART

Despite the fact that there are several reviews regarding the use of geostatistical tools in Archaeology (Ebert, 2002; Wheatley and Gillings, 2002), the work of Lloyd and Atkinson (2004) is considered as the the main reference on the topic. The two geographers tried in that paper to take a broad overview of the basic tools of geostatistics in archaeological contexts through different case studies, mainly using kriging methods. That text also works as a state of the art about this topic, reviewing the main geostatistical previous applications in Archaeology. In those pages is highlighted the importance of spatial autocorrelation measures to characterise spatial variation in archaeological variables. Specifically, the authors emphasise the importance of framing correctly the fieldwork of geostatistics in Archaeology: this is the analysis of a regionalised variable and its generalisation in a continuous surface. A classic archaeological example is, for example, the mapping of the chemical analysis samples of archaeological floors.

In the field of detecting and interpreting anthropic activity markers by using spatial interpolations, different approaches have been tested recently. In an ethnoarchaeological case study in Northern Gujarat (India), these techniques helped to explore the relative variability of floor chemical residues. People tend to recurrently use specific areas of their living space, producing an accumulation of evidences that represent the result of the activity performed. The possibility to identify and connect these evidences to the activity that generated the record is pivotal to our understanding of past human behaviour. Ethnoarchaeology and experimental archaeology drive the inferential reasoning that creates the models connecting the distribution of proxies with specific activities. These models are defined as 'anthropic activity markers', such as those coming from the analysis of chemical soil residues (Rondelli et al. 2014). Spatial interpolation of these proxies is the main technique in order to extrapolate a continuous surface from a group of samples or observations. In this way, the researcher can discern both singular clustering patterns -the anthropic activity markers- and different processes of floor maintenance and postdepositional dynamics considered as a background noise.

However, the use of this kind of approaches is still under development in different disciplines, also in Archaeology. As above pointed out, it is necessary to develop a general methodology to overcoming the geostatistical restriction on the homogeneity of the prediction region. The spatial distribution of chemical residues on domestic floors is highly influenced by the presence of barriers, such as walls. Therefore, these studies would benefit from a new methodology that avoid the interpolation of the observations both trough and under spatial demarcations. Our approach in this paper is thought to be a useful tool in order to surpass this major limitation. In the next paragraphs we investigate how to incorporate information of topographical nature into the process of geostatistical prediction using a cost-based kriging adaptation. Furthermore, the possibility of applying geostatistical methods enables us to obtain results based on statistical models, providing reliable predictions together with estimations of uncertainty, which commonly used deterministic methods cannot provide.

4.2 CASE STUDY: INTERPOLATION OF FLOOR CHEMICAL RESIDUES ANALYSIS IN TEOTIHUACAN

**5. SUMMARY AND CONCLUSIONS**