GEOSTATISTICAL MODELLING OF CHEMICAL RESIDUES ON ARCHAEOLOGICAL FLOORS IN THE PRESENCE OF BARRIERS

Joan NEGRE (corresponding author)

Centro Austral de Investigaciones Científicas
Consejo Nacional de Investigaciones Científicas y Técnicas
B. Houssay, 200 E-9410 Ushuaia (Tierra del Fuego, Argentina)

Laboratori d'Arqueologia Quantitativa
Departament de Prehistòria, Universitat Autònoma de Barcelona
Edifici B - Fac. Filosofia i Llletres, 08193 Bellaterra (Barcelona), Spain
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

Carla LANCELOTTI

CaSEs Research Group
Department of Humanities, Universitat Pompeu Fabra
C/Trias Fargas 25-27, 08005 Barcelona, Spain
carla.lancelotti@upf.edu

ABSTRACT

Maps representing the distribution of chemical residues over anthropogenic floors are the main diagnostic tools used by archaeologists for addressing the identification of geochemical signatures of past actions. Geostatistics allows producing these maps from a sample of locations by modelling the spatial autocorrelation structure of these kind of phenomena. However, the homogeneity of the prediction regions is a strong assumption in the model. The presence of barriers, such as the inner walls of domestic units, introduces discontinuities in prediction areas. In this paper, we investigate how to incorporate information of a geographical nature into the process of geostatistical prediction. We propose the use of cost-based distances to quantify the correlation between locations, a solution which has proved to be a practical alternative approach for archaeological intrasite analysis. The cost-based approach produces more reliable results avoiding the unrealistic assumption of the homogeneity of the study area. As a working example, a case study of the distribution of two specific chemical signatures in domestic floors is presented within a controlled ethnographical context in Northern Gujarat (India). On a broad disciplinary scale, the benefits of using our approach include improved estimates in regions with complex geometry and lower uncertainty in the kriging predictions.

KEYWORDS

non-Euclidean geostatistics, cost-based distance, spatial heterogeneity, kriging, housefloors

HIGHLIGHTS

- Mapping the distribution of certain combinations of chemical elements allows us to understand the activities that were developed in these areas.
- Geostatistical methods are usually used to model the results of this kind of analyses and hence facilitate interpretation.
- One of the main limits of the use of geostatistical methods for that purpose is the assumption of a homogeneous, unrestricted space of analysis.
- We propose the use of cost-based distances to quantify the correlation between sampling locations.

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 combinations of chemical elements allows us to understand the activities that were developed in the areas under study. This is based on the idea that different social actions of production, consumption or distribution are the cause of the variations observed in the material consequences detected through fieldwork. In this case, the variability of chemical soil composition is considered to be a reliable marker in order to detect, identify and analyse different activities in domestic contexts (Rondelli et al. 2014, Salisbury 2013, Middleton et al. 2010). The model connecting the concentration of particular residues (proxies) with the specific activities inferred from different information sources (archaeological experimentation, ethnoarchaeological reasoning, etc.) is defined as an anthropic activity marker. Nonetheless, the reading of chemical differential concentrations in archaeological floors is not exempt of critical reflections about its limitations (Vyncke et al. 2011, Dore and López Varela 2010, Wells 2010, Terry et al. 2004).

Geostatistical methods are increasingly used to model the results of geochemical analyses, hence facilitating the interpretation. These techniques provide a set of statistical tools specifically designed for spatial problems, in which predictions of missing values are 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. One of the main limits of the use of geostatistical methods for this purpose is the assumption of a homogeneous, unrestricted space of analysis (López-Quilez and Muñoz 2009). This premise fails when we consider the spatial demarcations and topography that affect the distribution of our phenomena over the study region. The analysis of chemical

residues in a domestic unit floor is a classic example of this kind of situation, where the walls of the house affect the distribution of the chemical elements.

In this work, we propose to overcome this problem by using cost-based distances to quantify the correlation between sampling locations (López-Quílez and Muñoz 2009). Thus, we present a case study on the distribution of chemical elements in domestic floors within a controlled ethnographical context in North Gujarat (India) (Rondelli et al. 2014). This paper explores the relative spatial variability of residues, taking into consideration spatial demarcations, to provide a method for the detection and interpretation of specific areas of activity. Our technique, therefore, can substantially improve the identification of both clustering patterns and different processes of floor maintenance and postdepositional dynamics considered as background noise (Rondelli et al. 2014, Pecci et al. 2013, Barba 2007, Lloyd and Atkinson 2004).

2. ON THE USE OF NON-EUCLIDEAN DISTANCES IN GEOSTATISTICS

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 twentieth century. There are several applications of geostatistical methods in a wide range of disciplines that share the problem of modelling a stochastic process over a continuous spatial region from a partial group of observations. This process of interpolation is commonly assumed to be Gaussian, isotropic and intrinsically stationary (Cressie 1993). Geostatistical modelling is based on the principle of spatial dependence, which states that near events are more related than distant ones. Nevertheless, what does *near* mean and how do we calculate it?

Interpolation techniques 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 irregularity 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 the simple euclidean distance between two points, which therefore cannot account for the correlation by itself.

There are more general situations where barriers are not absolute, but regions that are either harder or easier to cross depending on a series of relevant factors. For example, microtopography of the study region, soil texture and composition or the relationship of different anthropic activities between them are important restrictions that should be taken into consideration. All kind of heterogeneities in the surface in which chemical elements spread might be modeled with a cost surface, representing how hard it is to cross a given

portion of area. And accordingly, the correlation between two locations should be associated with the minimum-cost path connecting them. A cost surface presenting every relevant factor affecting correlation is, therefore, an efficient tool to deal with the distribution of chemical signatures in all kind of surfaces. In this framework, the standard geostatistical assumptions of an homogeneous region is a particular case where the Cost surface is a constant 1-valued surface. Therefore, the minimum-cost path between two given locations is the straight line connecting them, hence the Cost-Based distance equals the Euclidean distance. Also, the more general situation with barriers in the working region is another particular case where the Cost surface takes the value 1 over non-barrier areas and the value ∞ over barrier areas, therefore the Cost-Based distance equals the minimum distance needing to be traveled without crossing any barriers, as was required (López-Quílez and Muñoz 2009).

Methodologically, the first step in classical 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 (Cressie 1993). Our approach here is to use maximum likelihood methods that fit the mean value and the parameters of the semivariogram function. Once fitted, the main analytical interest lies in obtaining spatial prediction. Kriging assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. This technique is one of the most used approaches to this problem, in which a weighted average of the sample values is applied to generate the prediction. That is, sample points near the prediction's location are given larger weights than those far away. The general formula for the interpolator is formed as a weighted sum of the data:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$$

where $z(s_i)$ is the measured value at the □th location, □□ an unknown weight for the measured value at the □th location, □$_0$ the prediction location and □ the number of measured values.

Kriging determines these weights calculating them according to the value of the semivariogram, which is a function of the Euclidean distance (López-Quílez and Muñoz 2009). 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 or pseudo-Euclidean ones, represent the distance argument r of the semivariogram function more naturally.

2.1 COST-BASED DISTANCES

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, allowing to create a pseudo-Euclidean framework on which the analysis can be performed (Løland and Høst 2003; Negre 2015). A fast Fourier Transform has also been explored for integrating moving-average functions that may be used to generate a large class of valid, flexible variogram models. This transform allows to both compute the cross-variogram

on a set of discrete lags and to interpolate it for any continuous lag (Ver Hoef et al. 2004). In this same direction, recent works also propose the use of 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 cost-based distances is to define the least cost path to reach a known point from each cell location in the original raster dataset. The calculation algorithms present the length of the irregular vectors formed by a spatial distribution using the shortest weighted distance; that is, the path with least accumulated cost. In order to perform these calculations, first we need to create a cost surface, the purpose of which is to assign an impedance value to each cell of a raster layer, in other words, the facility with which it can be crossed. A lot of different approaches have been proposed in order to fulfill the modelling of this kind of surfaces (van Leusen 1999, Waller and Gotway 2004, Awaida and Westervelt 2006). Formally, the resulting cost-weighted model can be defined as a function f, 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 of a displacement $\mathrm{d}x$ at the point x is $f(x)\mathrm{d}x$. From this function, the cost of any path in A can be calculated as the integration of every cell in the model, which has been gone through (Muñoz 2012: 55). Thus, the cost of a path α between points $s_1, s_2 \in A$ will be

$$\int_0^1 f(\alpha(t))\alpha'(t)\mathrm{d}t$$

These distances also maintain the same general properties as their metric counterparts (Waller and Gotway 2004: 321):

- non-negativity $(d(x, y) \geq 0)$
- symmetry $(d(x, y) = d(y, x))$
- triangle inequality $(d(x, z) \leq d(x, y) + d(y, z))$

This implies that these measurements could be used into geostatistical functions. Nevertheless, the results might not be, in some cases, statistically significant. The more homogeneous is the surface under study, the less significant are the changes with respect to the use of Euclidean measures. To obtain mathematical validity, the resulting covariance matrix of the observations must be positive definite. This condition requires that for any number, set of locations $s_1, \ldots, s_n$ and complex set of coefficients $\propto_1, \ldots, \propto_n$, the R function verify the next relationship:

$$\sum_{i,1=1}^{n} \overline{\alpha_i}\alpha_j R(d(s_i, s_j))$$

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

The adaptation of geostatistical computation to these metrics is present in three major stages: 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 holding the distances between observation points, a symmetric square matrix. The second one containing the distances between the observation points and the prediction location(s), thus a *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 rest on the cost-based distance values given in the corresponding matrix in order to make the initial classification, 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 based on observation data only. It is typically accomplished through testing several possible combinations iteratively and keeping the one that maximises the likelihood function.. This implies computation of the covariance matrix for each combination being tested. All that is needed is to ensure that the covariance matrix is computed based on the cost-based distances provided by the previously calculated matrix. The final step is the kriging prediction. At this point, the covariance model is assumed to be known. 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 cost-based distance matrix is to be used.

2.2 METHODOLOGICAL OVERVIEW

2.2.1 Computing the cost-surface

The first operation is to encode the spatial heterogeneity of the working area into one *cost-surface*. This implies some modelling decisions and assumptions, which are not technical but scientific in nature. In our case, we assume that the soil is homogeneous except for the areas with solid sunken structures. These structures will completely interrupt the continuity of the area, limiting the dispersion of substances.

This conceptual model yields a cost-surface with a constant value of 1 everywhere, except over the structures where it takes an infinite value. In practice, any value larger than the diameter of the region will suffice. Alternatively, it can be more practical to work in the inverted scale of a *conductivity* surface. In this case, the values would be 1 for regular conductivity and 0 for no conductivity, or infinite cost. Any of these alternative surfaces can be easily produced from a digital representation of the region with a GIS software, or with other spatially capable software like R (R Core Team, 2015).

In our case, we imported the ESRI shapefiles describing the geometry of the structures into spatial classes defined in the R package sp (Pebesma and Bivand, 2005). Then we used the function $rasterize()$ from the raster package (Hijmans, 2015) to produce a discretized surface

with constant value 1 over the region of interest, and 0 over the solid structures. We used a resolution of 20 pixels/m.

2.2.2 Computing the cost-based distances

The cost/conductivity surface is the object representing our model of the region, and from where the distances between locations can be computed. Specifically, two matrices of cost-based distances are required: one $n \times n$ matrix with distances among the n observations, and one $n \times m$ matrix with the distances between each observation to each of the m prediction locations.

We used the centroids of the conductivity raster cells as prediction locations to simplify mapping, although any set of prediction locations can be used. The computation of the distance matrices can also be performed using a GIS software or directly within R. López-Quílez and Muñoz (2009) use the first approach with the help of a specific script $v.costdist.mat$ (Muñoz 2015b) for GRASS GIS (GRASS Development Team, 2010). For this study, we used the R package $geoRcb$ (Muñoz 2015b) instead.

This package provides the function $distmatGen()$, which automatically computes the two cost-based matrices given the coordinates of the observations and the conductivity surface. Internally, it leverages the package $gdistance$ (van Etten, 2015) for efficient computation of least-cost paths, while attending to all the technical details. Ultimately, the cost-based distance is computed using the well-known Dijkstra's algorithm for finding shortest paths between nodes in a network.

2.2.3 Using cost-based geostatistical algorithms

The R package $geoRcb$ extends some functions from the $geoR$ package (Ribeiro and Diggle, 2015), in order to work with cost-based distances. Specifically, the functions $variog$ and $likfit$ feature an additional argument *dists.mat* which takes a symmetric matrix of distances between observation locations. These functions are respectively used to compute empirical variograms and to fit variogram models. Finally, the alternative function *krige.conv* performs the cost-based spatial prediction through conventional kriging by taking the required distance matrices as the additional arguments *dd.disst.mat* and *dl.disst.mat*.

2.2.4 Presentation of results

We use the *viridis* colour palette (Garnier 2015) for all the maps in the present paper. This palette is perceptually uniform and is also designed to be perceived by readers with the most common forms of colorblindness.

3. CHEMICAL RESIDUES ANALYSIS ON ARCHAEOLOGICAL FLOORS IN THE PRESENCE OF BARRIERS

3.1 CONTEXT AND DATA DESCRIPTION

The data used in this methodological presentation were obtained from a fieldwork campaign that took place in a domestic compound in Jandhala, a village in the Patan district of North

Gujarat, India (Rondelli et al. 2014). The region of North Gujarat is a natural corridor that connects the Indus delta with the Indian subcontinent. The alluvial plains of North Gujarat extend NE-SW from the foot of the Aravalli Hills to the coast of the Little Rann of Kachchh, which is a marsh area between the lowlands of North Gujarat and the Kachchh peninsula. To the west, the alluvial plains reach the boundary of the Thar Desert. To the east, the limits are the Sabarmati river catchment and the Nal Sarovar depression. This landscape is characterized by fossilized sand dunes and interdune areas that can become seasonal lakes during the monsoon and post-monsoon seasons. As in other parts of South Asia, many of these interdune depressions were converted into village ponds or irrigation water tanks (Conesa et al. 2015). North Gujarat is a sensitive, monsoonal-dependent, semi-arid region in which Indian summer monsoon patterns cause significant variations in seasonal precipitation at the regional and local level. Extreme climatic shifts can generate severe droughts or floods affecting resource availability (Conesa et al 2013). Current paleoclimatic models suggest monsoonal stability throughout the mid-Holocene (Balbo et al. 2014). As a consequence, the cycles of activities, both at a domestic and public level, is still largely dependent and affected by the variations in the monsoon patterns and intensity.

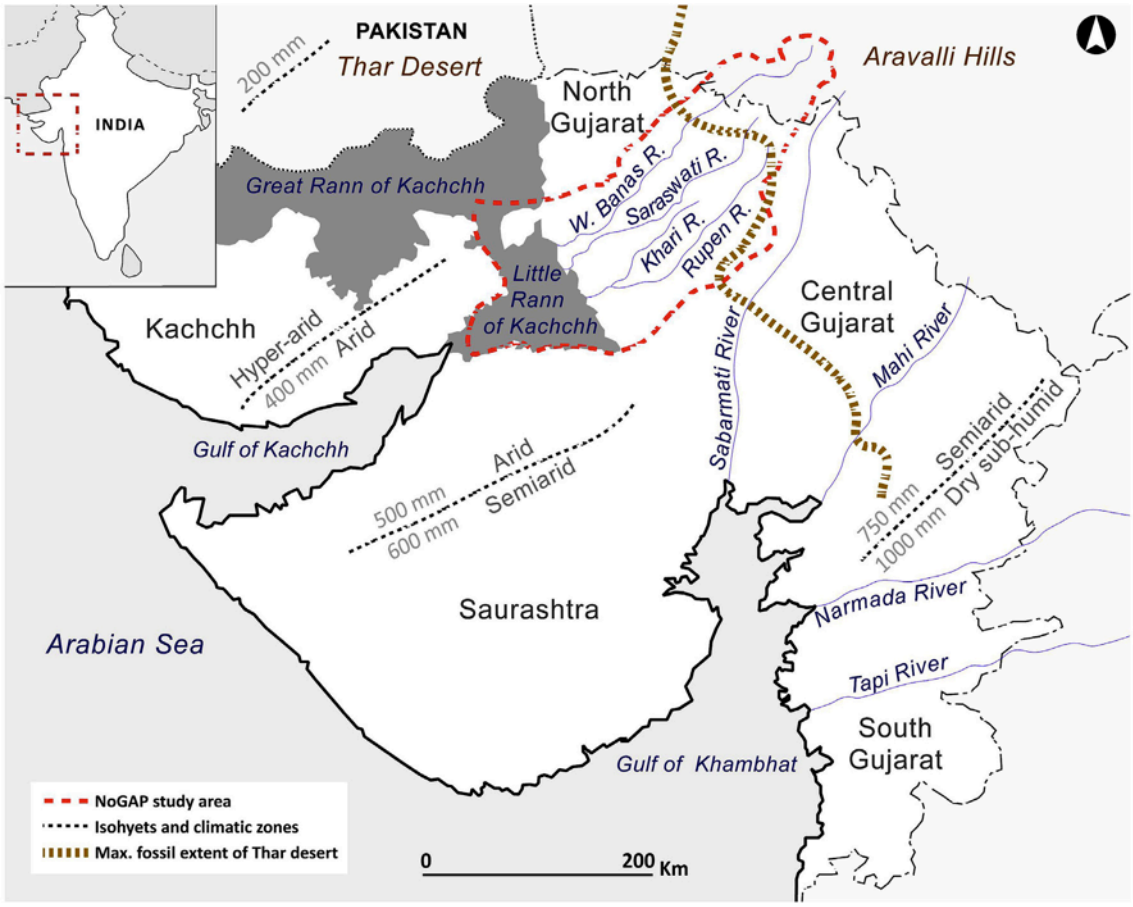

Fig. 1 Location map

The Patan compound is a non-mechanised farmer's residence built using local materials (locally-sourced clay, sand and animal dung) and traditional techniques (wattle-and-daub, plastering and wooden structures). A topographical and architectural survey of the whole complex was carried out in form of vectorial files, georeferenced in order to be taken as study area. A complete ethnoarchaeological study was developed in order to define the types and cycles of activities performed in the compound, as well as their spatial distribution within the courtyard, veranda and inner house. One of the two houses that composed the compound (located within a common open area) was systematically sampled on a regular lattice grid (50 cm), using a hollow metal pipe with a diameter of 5cm. Micromorphological floor samples were collected and thin section produced in order to understand the construction techniques and the formation processes of the house floors. The results show that the loose samples collected correspond to several years of floor use, thus representing the accumulation of repeated activities (something akin to what can be most probably sampled in an archaeological context). These analyses were followed through with Inductively Coupled Plasma - Atomic Emission Spectrometry (ICP-AES), providing quantitative data on chemical elements. Finally, a complete study of the use of space based on floor chemical analysis was published (Rondelli et al. 2014).

The ethnological observation of the daily activities in the house, as well as the study of micromorphological floor samples, allowed the original researchers to frame the main activity areas in the house and veranda and their floor formation processes. Understanding the cycles of floor maintenance through interviews and archaeological approaches was paramount to evaluate the significance of the results obtained from the geochemical analysis (Rondelli et al. 2014: 486). The ethnoarchaeological approach led to identifying two cooking areas, one in the inner house and one in the veranda, with a total of three fireplaces (2 inside and one in the veranda). In addition, the inner space was composed by a sleeping area, a food production and consumption area and a storage area (Fig. 2). The main goal of the work was, therefore, to test the hypothesis about the correlation of certain chemical signatures with the specific domestic activities previously framed.

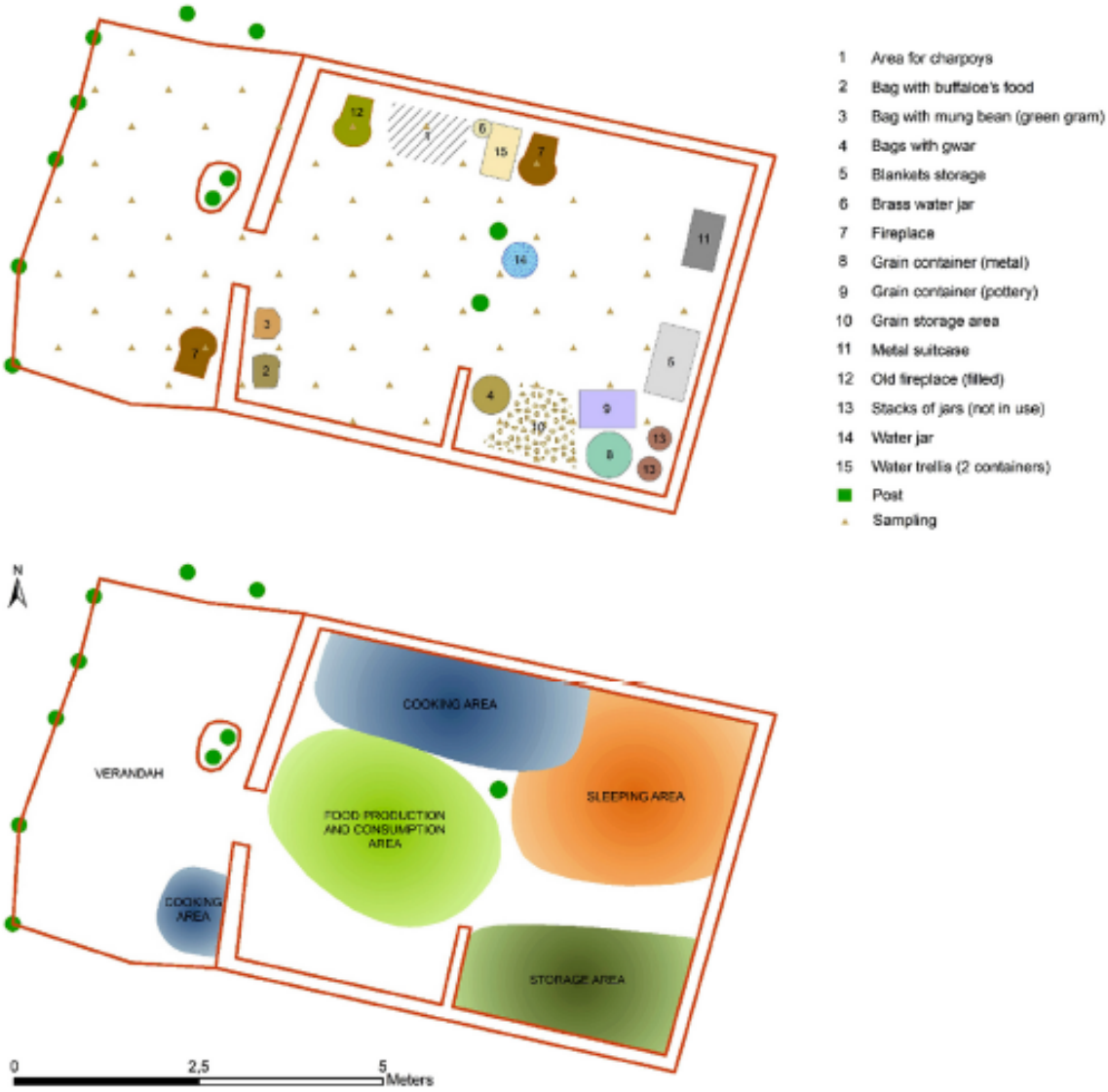

Fig. 2 House activity areas

It is important to clarify that in the present work, we are not trying to analyse in depth the specific use of space in this case study, but to confront the previous geostatistical results with the ones provided using our technique. In order to expound these differences two different examples were chosen : the distribution of Calcium residues (main chemical proxy for enclosed spaces) and a combination of several chemical elements (Calcium, Phosphorus, Potassium, Magnesium and Strontium) which can account for the signature of deposited food remains areas (according to Milek 2007 and Middleton and Price 1996). Seventy measurements were taken in various points distributed homogeneously over the house and its veranda to sample the chemical composition of these soils.

All the measurements have a weak gaussian behaviour, due mainly to outliers values in all the data collections, which introduce background noise in the models. Normality of input data is one of kriging premises in order to offer reliable predictions so the values have been

depurated, taking out the outliers which introduced errors in fitting the variograms and the cross-validation. In the case of the Calcium, values go from a minimum of 0.66% to a maximum of 5.70%, being its mean value 3.11% and its median 2.95%. In the case of deposited food remains, all the chemical elements presents in the anthropic marker were normalised on a scale 0 to 1 before being combined. Values go from a minimum of 1.05 to a maximum of 3.94 in this relative scale, being its mean value 2 and its median 2.01. The walls of the house are considered non-transparent barriers for the diffusion of chemical residues on the archaeological floors. For this reason, cost-based distances from each sampling point to everywhere else have been calculated. Figure 4 displays the cost-based distance maps to four selected points, showing how these distances honour the geometry of the region.

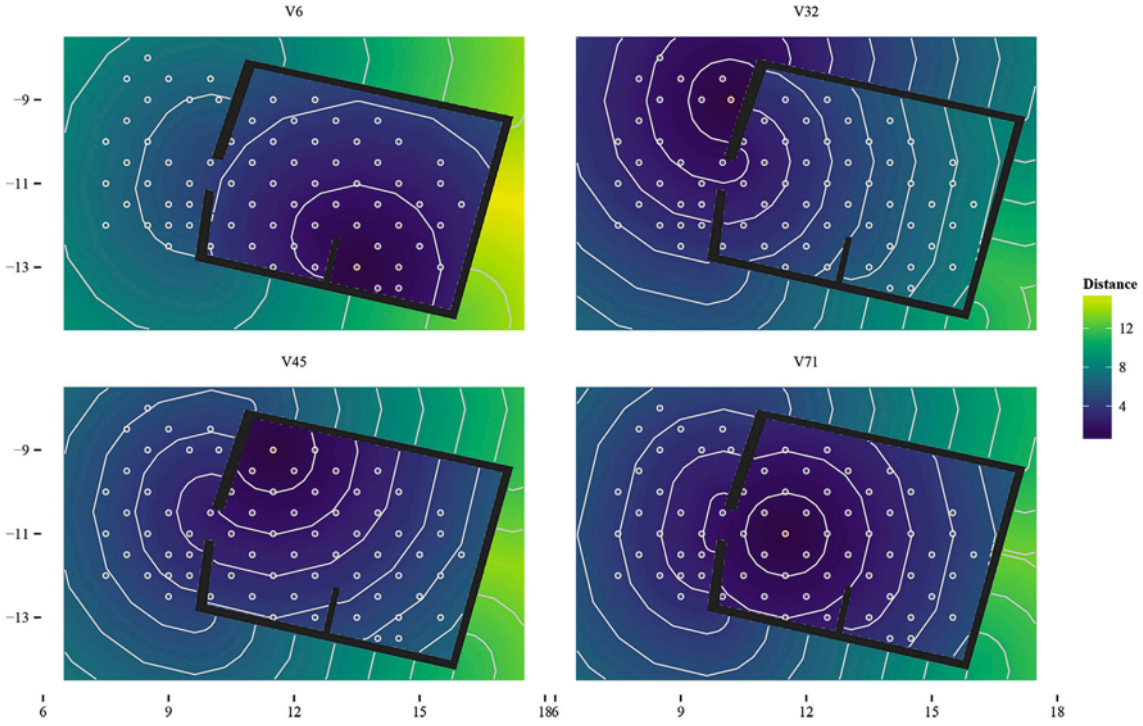

Fig.3 Example Cost-based surfaces for four sampling points

3.2 VARIOGRAM AND FITTED MODEL

Variogram features (i.e., the estimated statistical parameters) include the nugget (the modeled discontinuity of the variogram at a distance of zero, which can represent measurement error or variation at distances too small to be captured with the current spatial design), the sill (the value at which the variogram stops increasing, an estimate of variance), and the range (the distance where the sill is reached, meaning the distance at which samples are no longer spatially correlated).

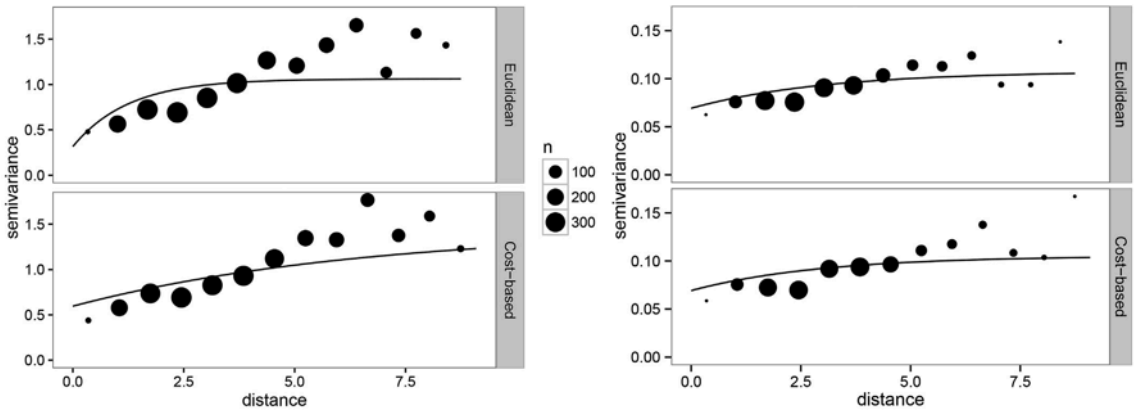

Fig.4 Euclidean and Cost-based empirical (points) and fitted (lines) variograms for the Calcium (left) and the deposited food remains (right) data

In both case studies we fitted an exponential variogram model using both Euclidean and cost-based distances. For Calcium, the main difference was that the cost-based approach yielded a larger range (Fig. 5; Table 1). This is usually the case, as incorporating the geometry into the analysis helps making sense of what was before interpreted as unstructured noise. Also the nugget was higher in the estimated cost-based variogram, which resulted in increased smoothing. On the other hand, the estimated variogram for deposited food remains showed similar values for both approaches, with a slightly higher practical range for the Euclidean model.

	Calcium		Deposited food remains	
	Euclidean	Cost-based	Euclidean	Cost-based
Intercept	3.12	3.17	1.95	1.94
Nugget	0.32	0.60	0.07	0.07
Partial sill	0.75	0.85	0.04	0.04
Phi	1.25	6.53	3.04	2.94
Pract. range	3.75	19.56	9.12	8.81
Log-likelihood	−89.25	−89.82	−12.84	−12.77

Table 1 Estimated parameters of the Euclidean and Cost-based variograms for the Calcium data and deposited food remains

A natural question that arises is how different the prediction values from the two types of distances are. Figure 5 plots a pointwise comparison of predictions, showing remarkable differences between both methods.

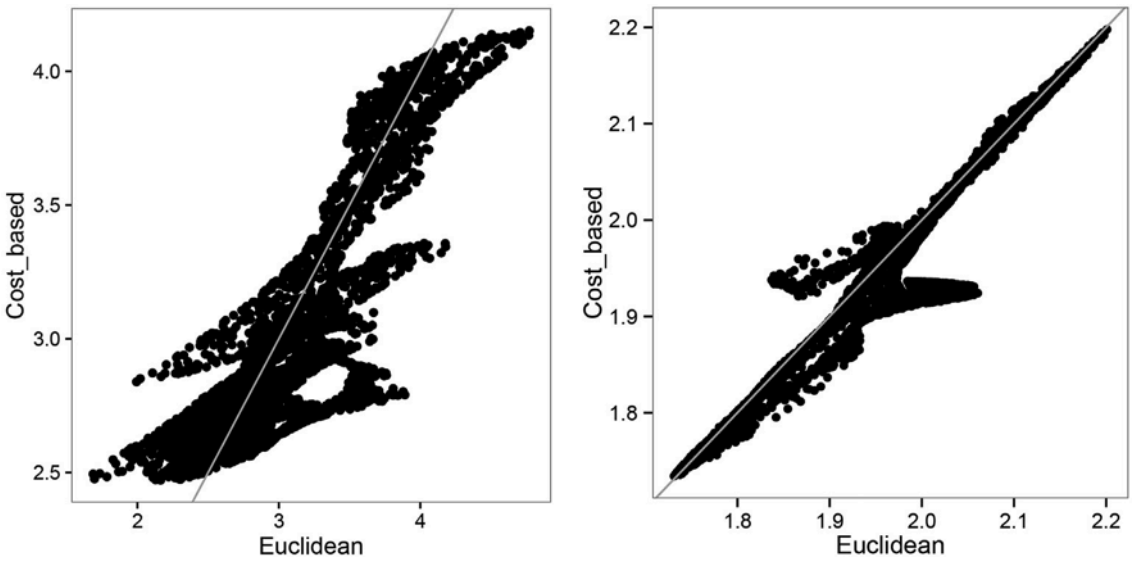

Fig.5 Pointwise comparison of Euclidean and Cost-based predictions for the Calcium data (left) and deposited food remains (right)

3.3 COMPARISON OF KRIGING ESTIMATES AND PREDICTED ERRORS

3.3.1 CALCIUM DISTRIBUTION

Figure 6 shows the prediction in Calcium distribution for each location using Euclidean and Cost-based distances, and their difference. There are three remarkable aspects to be noted. First, the highest differences in prediction happen in the outer unsampled area, especially just across the wall from the most extreme observation. The Euclidean prediction in that area is strongly influenced by the observed values inside, in contrast to the average value predicted by the cost-based approach. Second, the cost-based prediction is clearly smoother, as expected given the parameter estimates of the variogram model. This produces significant differences in the neighbourhoods of the most extreme observations. Finally the walls of the house modify the prediction even in the sampled area. For example, in the corner near the central inner wall and across from the highest measurement, the Euclidean method predicts higher values, undoubtedly under the influence of that specific observation. On the contrary, the Euclidean predictions are lower in the veranda, influenced by the observations in the interior. In the proximities of all observations, the cost-based approach has a larger prediction error due to its increased estimation of the nugget (i.e. short-range variance). In the main area, the prediction errors are practically the same with both approaches. Behind the walls, the Euclidean prediction error is unrealistically low. Leave-one-out Cross Validation (LOOCV) yielded similar error values for both approaches.

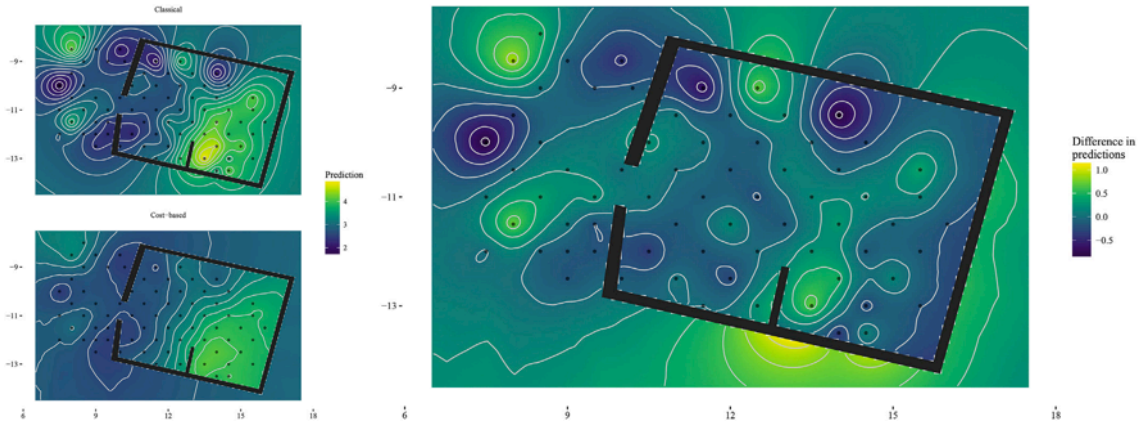

Fig.6 Euclidean and Cost-based predictions (left) and comparison of the results (right) for the Calcium

3.3.2 COMBINED CHEMICAL SIGNATURE DISTRIBUTION

Figure 7 shows the prediction in the combined chemical signature of Calcium, Phosphorus, Potassium, Magnesium and Strontium for each location using Euclidean and Cost-based calculations, and their difference. As in the previous case study, the differences between both approaches are highest in the outer unsampled area. A similar situation is repeated also in the north-west corner of the house, where the very low values of the sampling points in the veranda are trespassing their influence into the inner house predictions. The drop of about -5% prediction variance for the Euclidean prediction in this sector (as well as circa -2% droppings all along this separation wall) are a consequence of incorrectly assuming that observations at the other side of the wall are *close*. In the main area, the prediction errors are practically the same with both approaches. In most of the area, the Cost-based approach is slightly more accurate (median 1%) achieving up to 1.4% of improvement in accuracy in the rightmost corners. Leave-one-out Cross Validation (LOOCV) present similar error values for both approaches.

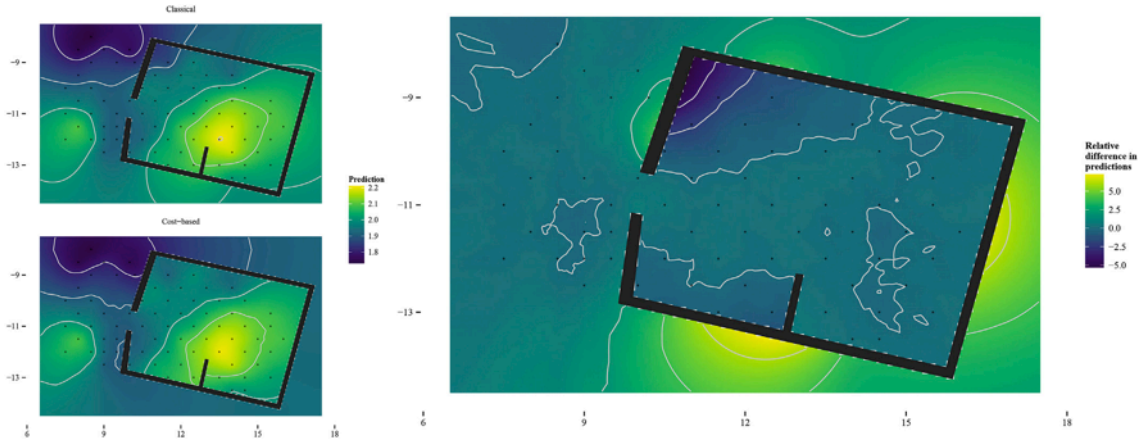

Fig. 7 Euclidean and Cost-based predictions (left) and comparison of the results (right) for the food remains

By taking into account the heterogeneities of the working region through the use of cost-based distances, our geostatistical analysis presented differences in predictions that were remarkable in some locations. The discrepancies reached values near 1% (Ca) and 5% (Ca+P+K+Mg+Sr). We

believe that the cost-based approach produces more reliable results, avoiding the assumption of certain unrealistic premises about the homogeneity of the study area. If these remarkable differences are being detected by our cost-based kriging in a domestic unit with very simple inner divisions (a partial wall in the inner area and the separation between this and the veranda), critical variability of predictions is expected to happen in more complex buildings.

3.3.3 ETHNOARCHAEOLOGICAL INTERPRETATION

Archaeological reading of chemical signatures in domestic floors is highly related to the combination of diverse element residues, such as different combinations of Potassium, Calcium, Phosphorus, Magnesium and Strontium for identifying deposited food remains, living room sediments, enclosed spaces or burning areas (Rondelli et al. 2014: 488). Calcium has long been regarded as a good indicator of human activity, usually correlated with Strontium, suggesting similar chemical pathways. This element are normally interpreted as a good indicator of the integrity of anthropogenic deposits (Cook and Heizer 1965).

The previous identification of the activity areas through ethnoarchaeological methods had allowed the proposition of a working hypothesis regarding the concentration of this marker specifically around the storage area, located in the south-east sector of the house. From this point of view, and taking into account only the house and the veranda areas, cost-based approach presents a good-fit model of the chemical distribution of Calcium, which literature propose relate with covered, enclosed spaces (Middleton and Price 1996: 679). Therefore, our preliminary working example, taking into consideration this proxy, is as suited to test the robustness of the method than to provide inferences about human behaviour. This way, it is clear that the cost-based approach provides a much refined definition of the activity areas inside the house. In this case, it restricts the concentration of Ca in the south-east area close to the inner wall. Not only the area of concentration is better defined but also the false positive areas, such as the external perimeter of the house or the veranda are excluded from the map. This is of much assistance in archaeological contexts where the physical definition of the space might not always be conserved, e.g. when the structures are build in perishable material, such as the present ethnographic case.

In the presence of several geochemical indicators related at the same time, a perfect delimitation of their distribution areas is crucial for a valid inference. In our work, the combination of different elements has allowed to identify a food production and consumption area, where chemical traces of deposited food remains and ashes from the hearth were detected (Middleton and Price 1996: 678, Milek 2007: 342). This signature, in combination with other proxies, such as concentrations of highly fragmented charcoal or the absence of macro-remains (lithic and zooarchaeological record), configure a specific anthropic marker in order to frame areas of food processing, handling and consumption. In this case, nevertheless, we confronted spatial distribution of the geochemical signature against ethnoarchaeological observations and interpretation of the use of social space.

The working hypothesis regarding food processing and consumption areas proposed its location in the central place of the inner area, in particular in the west half of it. Both predictors hit the target in framing this activity near the central part of the house, being the cost-based approach slightly more precise in defining the preference for the use of the west

half of the house against the right half. That might be clearly observed in the inner wall propinquity, where only cost-based interpolation could discern between the observed food consumption and handling area (west side of the wall) and the storage area (east side of the wall). Another important difference between both methods must be highlighted in the north-west corner of the house. While Euclidean prediction allow the negative values of the Veranda's north side (where food-related activities were not performed) to influence the inner house predictions, only cost-based ones are capable to frame correctly the extension of cooking activities in all the extension of the north wall. Finally, both approaches identify correctly a external area of food-related activities, but only cost-based one is capable of detecting a slightly rupture in the continuity of the values near the external west wall of the house, just where the external fireplace was located. Again, Euclidean interpolation failed to model correctly the spatial distribution of chemical signatures in the presence of barriers.

4. DISCUSSION

Despite the specific results of our case study, the proposed method is widely applicable to any interpolations in any archaeological intrasite example. As previously indicated, the more heterogeneous the surface under study (especially regarding the presence of barriers) the more useful this approach will be. The most important aspect of this work, therefore, lies in the general methodology for overcoming the geostatistical restriction on the homogeneity of the prediction region. In the study of geochemical signatures on archaeological floors, that limitation is a major problem for obtaining significant interpretation of the sampling data, especially in the presence of areas where different activities are performed. Thus, spatial distribution maps of chemical residues benefits from this methodology, since walls in domestic contexts are relevant restrictions in their distribution. Furthermore, the possibility of applying geostatistical techniques enables us to obtain results based on statistical models, providing reliable predictions together with estimations of uncertainty, which commonly used deterministic methods cannot provide. Altogether, all this information allows the correct interpretation of the archaeological data, both the distribution of the chemicals and the postdepositional effects over them, linked among other factors to the intensity of the uncertainty measure.

This kind of studies have also several fieldwork aspects to be remarked. As a predictive model, the interpolation surfaces created by the means of kriging approaches allow to identify possible interest areas to be tested through excavation methods. Moreover, in archaeological sites, these models can lead to the assessment of future campaigns, informing the decision of new areas to be excavated. It is also noted that this approach could be very beneficial in order to design the best-fitted sampling strategy. The analysis of kriging variance, for example, is a function of the form of the variogram, the sample configuration and the sample support. The coefficients of the model fitted to the variogram might be used to ascertain the maximum punctual kriging variance for different sample spacings, enabling the researcher to choose the maximum sample spacing possible to achieve a particular precision (Lloyd and Atinson 2004: 160).

Finally, ethnoarchaeological analysis of chemical soil composition has also potential applications in the field of reading behavioural tendencies in the formation processes of floors. Interviews and direct observation were conducted in order to identify the types and cycles of activities carried out in the compound, with specific attention to the construction and maintenance of the floors. Re-plastering episodes of the entire floor take place on average four times a year together with ad-hoc re-plastering whenever the floor is damaged. Through micrompohology analysis this entire stratigraphic sequence of floor construction was evaluated in order to assess the temporal representativeness of the samples, allowing to detect micro-shapes of activity remains accumulations between the re-plastering episodes. The study of the use of space based on floor chemical analysis relied, therefore, on the concept that identified residues in these interstices were the sum of different activities carried out in a specific part of the space. Hence the samples in the original study represented the sum of cycles of use, being the chemical signatures average indicators of the accumulation of repetitive activities (Rondelli et al. 2014: 487). The identification in this ethnoarchaeological case study of this whole process of soils formation processes and chemical markers accumulation have important ramifications in future archaeological analyses, allowing to define certain aspects of floor maintenance imperceptible at first sight.

On the general objective of accurate readings on the spatial trends of chemical signatures, this proposal fit into the recent tendency on the application of quantitative methods to our discipline. Since the already classic work of Lloyd and Atkinson (2004) on the specific field of Geostatistics in Archaeology, a series of methodological milestones have taken place. Our work aims at joining this effort in developing by overcoming a recurrent limitation of these methods: the non-Euclidean characterisation of almost any geographical surface in the real world. As for macro-scale archaeological cases, this is not a minor aspect to take into consideration (Negre 2015). The treatment of heterogeneity and uncertainty, both in spatial and temporal spheres, has become the focus of most of the methodological step forwards to our discipline in the last years (Bevan et al. 2012; Deravignone et al. 2015; Fernández-López and Barton 2015; Crema 2015). The use of Anisotropic Geostatistics deals with both elements at the same time. The development of complex spatio-temporal reasoning in Archaeology is highly desirable when statistical, physical and chemical ground-breaking methodological advances are allowing a better understanding of our reality. Interdisciplinary work teams dealing with transversal problems are currently the most important path into new forms to understand and process archaeological information.

5. CONCLUSIONS

To sum up, this case study allows us to understand the main differences on the use of either non-Euclidean distances or their traditional metric counterpart. In this example, we tried to focus on two specific chemical signatures that we believe that are representative of a spatial pattern of repetitive activities. We are not trying to frame the *storage area* and *food-related activities* areas for good, since that would involve a series of other proxies, but in presenting the different geochemical information obtained depending on the methodology that we used. The sum of small errors or imprecisions in the process of modelling the distribution surfaces of

each element is erroneously accumulated when those are combined in a more complex chemical proxy. This is an important question in order to explain the distribution of chemical signatures and also for using that information in combination with other proxies. Regression Kriging using the combined trends that can be deduced from the statistical analysis of several chemical concentrations in each sampling point is also tied to this important limitation. When this kind of models is used to represent the general tendencies of geochemical residues, it is necessary to confront the interpolation of the resultant factor with the signatures of each individual sample for interpretation purposes. Even so, spatio-temporal descriptions, such as interpolation maps, are not explanations but are themselves something to be explained (Barceló et al. 2015: 35). When dealing with the modelling of regularities in the distribution of material consequences of past actions, every bit of precision is required in order to understand their causal structure. That plus of accuracy is what we look forward to providing with our approach.

ACKNOWLEDGEMENTS

We wish to thank Marco Madella and Ajithprasad P., co-directors of the North Gujarat Archaeological Project (NoGap), which allowed us to use its data in order to exemplify the benefits of our method. The present work benefited from the input of Juan Antonio Barceló, Head of the Quantitative Archaeology Laboratory and Professor at the Autonomous University of Barcelona, Giancarlo Macchi, Assistant Professor at the University of Siena, who provided valuable comments to the undertaking of the research presented here. This work was supported by *SimulPast Consolider Ingenio* project (CSD2010-00034) and *Experimentation and development of advanced artificial intelligence techniques for the computer simulation of social dynamics and historical evolution* project (HAR2009-12258), funded by the Spanish Ministry of Science and Innovation. Joan Negre Pérez has worked on this research on a postdoctoral fellowship from the Argentinian Ministry of Science, Technology and Productive Innovation (PICT2012-2148) and partially funded by research grant 2008UAB503 from the Autonomous University of Barcelona (Spain). Facundo Muñoz has worked on this research on a postdoctoral fellowship from the French National Institute for Agricultural Research and partially funded by research grant MTM2013-42323-P from the Spanish Ministry of Economy and Competitiveness and ACOMP/2015/202 from Generalitat Valenciana (Spain). Carla Lancelotti is part of the CaSEs Research Group, a recognised excellence group of the Generalitat de Catalunya (SGR2014-1417). Her work in this paper is partially conducted within the framework of the MoMarq Project (*Modelado y Simulación de Marcadores de Actividades Antrópicas: de lo etnográfico a lo arqueológico*) funded by the Spanish Ministry of Economy and competitiveness (HAR2014-55518-P).

REFERENCES

Awaida, A. y Westervelt, J. 2006. r.cost: cumulative cost computation for GRASS GIS. URL http://grass.osgeo.org/grass63/manuals/html63_user/r.cost.html.

Balbo, A.L., X. Rubio, B. Rondelli, M. Ramirez, C. Lancelotti, A. Torrano, M. Salpeteur, N. Lipovetzky, V. Reyes-García, C. Montañola and M. Madella 2014. Agent-based simulation of Holocene monsoon precipitation patterns and hunter-gatherer population dynamics in semi-arid environments. Journal of Archaeological Method and Theory 21: 426-446.

Barba, L. 2007. Chemical residues in lime-plastered archaeological floors. Geoarchaeology 22: 439-452.

Barceló, J.A., 2007. Arqueología y Estadística. Introducción al estudio de la variabilidad de las evidencias arqueológicas. Universitat Autònoma de Barcelona, Bellaterra.

Barceló, J.A., K.F. Achino, I. Bogdanovic, G. Capuzzo, F. del Castillo, V. Moitinho and J. Negre, 2015. Measuring, Counting and Explaining: An Introduction to Mathematics in Archaeology. In J.A. Barceló & I. Bogdanovic (eds.). Mathematics and Archaeology (pp. 3-64). Boca Ratón: CRC Press.

Bevan, A., J. Conolly, C. Hennig, A. Johnston, A. Quercia, L. Spencer and J. Vroom, 2012. Measuring Chronological Uncertainty in Intensive Survey Finds. Archaeometry 55: 318-328.

Buck, C. and S.K. Sahu, 2000. Bayesian models of relative archaeological chronology building. Applied Statistics 49: 423-440.

Cook, S.F. and R.F. Heizer, 1965. Studies on the chemical analysis of archaeological sites. University of California, Publications in Anthropology, Berkeley.

Conesa, F.C., N. Devanthéry, A.L. Balbo, M. MAdella and P. Ajithprasad, 2013. Identification of seasonally flooded areas in North Gujarat using radar satellite imagery: implications for archaeology. Journal of Multidisciplinary Studies in Archaeology 1: 344-355.

Conesa, F.C., M. Madella, N. Galiatsatos, A.L. Balbo, S.V. Rajesh and P. Ajithprasad, 2014. CORONA Photographs in Monsoonal Semi-arid Environments: Addressing Archaeological Surveys and Historic Landscape Dynamics over North Gujarat, India. Archaeological Prospection: 10.1002/arp.1498

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley Series in Probability and Statistics. Wiley-Interscience, New York.

Deravignone, L., H.P. Blankholm and G. Pizziolo, 2015. Predictive Modeling and Artificial Neural Networks (ANN): from model to survey. . In J.A. Barceló & I. Bogdanovic (eds.). Mathematics and Archaeology (pp. 335-351). Boca Ratón: CRC Press.

Dore, C., S. López Varela, 2010. Kaleidoscopes, palimpsests, and clay: realities and complexities in human activities and soil chemical/residue analysis. Journal of Archaeological Method and Theory 17: 279-302.

Ebert, D., 2002. The potential of geostatistics in the analysis of fieldwalking data. In: Wheatley, D, G. Earl and S. Poppy (Eds.), Contemporary Themes in Archaeological Computing, University of Southampton Department of Archaeology Monograph No. 3. Oxbow books, Oxford, pp. 82-89.

Fernández-López de Pablo, J. and C.M. Barton, 2015. Bayesian Estimation Dating of Lithic Surface Collections. Journal of Archaeological Method and Theory 22: 559-583.

Garnier, S. 2015. viridis: Matplotlib Default Color Map. R package version 0.2. https://github.com/sjmgarnier/viridis

GRASS Development Team (2010). Geographic Resources Analysis Support System (GRASS GIS) Software. USA: Open Source Geospatial Foundation. http://grass.osgeo.org/

Hijmans R.J. (2015). raster: Geographic Data Analysis and Modeling. R package version 2.4-15. http://CRAN.R-project.org/package=raster

Krige, D.G., 1951. A statistical approach to some basi mine valuation problems on the witwaterstrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa 52(6): 119-139.

Lancelotti, C. and M. Madella, 2012. The "invisible" product: developing markers for identifying dung in archaeological contexts. Journal of Archaeological Science 39: 953-963.

Lloyd, C.D. and P.M. Atkinson, 2004. Archaeology and geostatistics, Journal of Archaeological Science 31: 151-165.

Løland, A. & Høst, G. (2003). Spatial covariance modelling in a complex coastal domain by multidimensional scaling. Environmetrics, 14 (3), 307-321.

López-Quílez A. and F. Muñoz (2009). Geostatistical computing of acoustic maps in the presence of barriers. Mathematical and Computer Modelling 50(5-6):929-938.

Matheron, G., 1963. Principles of geostatistics, Economic Geology 58(8): 1246-1266.

Middleton, W.D. and T.D. Price, 1996. Identification of Activity Areas by Multi-element Characterization of Sediments from Modern and Archaeological House Floors Using Inductively Coupled Plasma-atomic Emission Spectroscopy. Journal of Archaeological Science 23: 673-687.

Middleton, W.D., L. Barba, A. Pecci, J.H. Burton, A. Ortiz, L. Salvini and R.R. Suárez, 2010. The study of archaeological floors: methodological proposal for the analysis of anthropogenic residues by spot tests, ICP-OES, and GC-MS. Journal of Archaeological Method and Theory 17: 183-208.

Milek, K.B. 2007. House and Households in Early Icelandic Society: Geoarchaeology and the interpretation of Social Space. PhD Thesis, University of Cambridge.

Møller, J. and H. Toftaker, 2012. Geometric anisotropic spatial point pattern analysis and Cox processes. Research Report Series, Department of Mathematical Sciences, Aalborg University.

Muñoz, F. M. (2012). Geoestadística en regiones heterogéneas con distancia basada en el coste. PhD Thesis: Universitat de València.

Muñoz F. (2015a). geoRcb: An Extension of Package geoR that Works with Cost-Based Distances. R package version 1.7-5. https://github.com/famuvie/geoRcb. DOI:10.5281/zenodo.23568

Muñoz F. (2015b). v.costdist.mat: A GRASS script for computing cost-based distance matrices. Zenodo. DOI:10.5281/zenodo.23546

Negre, J. (2015). Non-Euclidean Distances in Point Pattern Analysis: Anisotropic Measures for the Study of Settlement Networks in Heterogeneous Regions. In J.A. Barceló & I. Bogdanovic (eds.). Mathematics and Archaeology (pp. 369-382). Boca Ratón: CRC Press.

Pebesma E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews/

Pecci, A., M.A. Cau Ontiveros, C. Valdambrini and F. Inserra, 2013. Understanding residues of oil production: chemical analyses of floors in traditional mills. Journal of Archaeological Science 40 (2): 883-893.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ribeiro Jr P.J. and P.J. Diggle (2015). geoR: Analysis of Geostatistical Data. R package version 1.7-5.1. http://CRAN.R-project.org/package=geoR

Rondelli, B., C. Lancelotti, M. Madella, A. Pecci, A. Balbo, J. Ruiz Pérez, F. Inserra, Ch. Gadekar, M.A. Cau Ontiveros and P. Ajithprasad (2014). Anthropic activity markers and spatial variability: an ethnoarchaeological experiment in a domestic unit of Northern Gujarat (India), Journal of Archaeological Science 41: 482-492.

Salisbury, R.B., 2013. Interpolating geochemical patterning of activity zones at Late Neolithic and Early Copper Age settlements in eastern Hungary. Journal of Archaeological Science 40: 926-934.

Terry, R.E., F.G. Fernández, J.J. Parnell and T. Inomata, 2004. The story in the floors: chemical signatures of ancient and modern Maya activities at Aguateca, Guatemala. Journal of Archaeological Science 31: 1237-1250.

van Etten J. (2015). gdistance: Distances and Routes on Geographical Grids. R package version 1.1-7. http://CRAN.R-project.org/package=gdistance

van Leusen, M. 1999. Viewshed and cost surface analysis using GIS (Cartographic modelling in a cell-based GIS II). In: J.A. Barceló, I. Briz and A. Vila (eds.). New techniques for old times (pp. 215-223). Oxford, British Archaeological Reports.

van Overwalle, F. and D. van Rooy, 2013. A connectionist approach to causal attribution. In: S.J. Read and L.C. Miller (Eds.). Models of Social Reasoning and Social Behavior. (pp. 143-171). Routledge, New York.

Ver Hoef, J.M., N. Cressie and R.P. Barry, 2004. Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform (fft), Journal of Computational & Grafphical Statistics 13 (2): 265-282.

Vyncke, K., P. Degryse, E. Vassilieva and M. Waelkens, 2011. Identifying domestic functional areas. Chemical analysis of floor sediments at the Classical-Hellenistic settlement at Düzen Tepe (SW Turkey). Journal Archaeological Science 38: 2274-2292.

Waller, L. & Gotway, C. (2004). Applied Spatial Statistics for Public Health Data. Hoboken, Wiley.

Wells, E., 2010. Sampling design and inferential bias in archaeological soil chemistry. Journal of Archaeological Method and Theory 17: 209-230.

Wheatley, D. and M. Gillings, 2002. Spatial Technology and Archaeology: The Archaeological Applications of GIS. Taylor & Francis, London.