Multivariate Statistic Analysis of the Relationship between Archaeological Sites and the Geographical Data of their Surroundings. A Quantitative Model

José A. Esquivel, José A. Peña, M. Oliva Rodríguez-Ariza

Abstract

In the last years, the analysis of settlement patterns has been accomplished from different quantitative points of view, including statistic analysis, GIS, etc. This work is focused on obtaining a quantitative model of index parameters based in the UTM coordinates, using a 1:10,000 scale map. The set of indexes is studied by means of statistical analysis (Kruskal-Wallis, principal components analysis, cluster analysis and Voronoi tesselations), providing the chronocultural settlement patterns and the relationships between the archaeological site, the surrounding area and the chronological period of the site.

1 Introduction

The study of factors that determine the selection of archaeological sites has been approached from different standpoints. Particularly the parameters characterizing the selection of settlement sites in relationship to the environment have originated several attempts to establish some standards of selection. Some analyses are affected by bad definition and incorrect use of the variables (subjective variables, poorly defined variables, irrelevant data, etc.), leading to erroneous or irrelevant conclusions.

In this work, we attempt to characterize the patterns of settlement using the topographical information in the region surrounding the site and the spatial data in the archaeological deposit by means of cartographical variables. We establish a quantification through metric-index variables, including geographical concepts connecting the characteristics of each archaeological site with the shape of the relief that surrounds it. Each variable quantifies a well-defined and objective concept, providing consistent results under similar conditions ("universality" property).

The settlement patterns were analysed by means of multivariate statistical analysis using topographical variables only, and an equidistance analysis using UTM coordinates of the sites. These analyses show the relationship between the location of settlements, the shape of relief in the environment of each site and the chronocultural information.

This model is applied to the data obtained from archaeological surface surveys carried out in the Orce-Huéscar region (Granada, Spain) by one of the authors (Rodríguez-Ariza), comprising settlements of the Neolithic period, Copper and Bronze Ages, Protoiberian and Iberian periods, and the Roman and Medieval periods.

2 Previous framework

Many authors have attempted to study the location of archaeological settlements by the quantification of a set of variables, essentially focused on establishing a predictive model. In these works the variables can be: a) topographical only (Carmichael 1990), b) topographical and drainage (Kvamme 1992), and c) a model mixing topographical variables, soil variables, climatic variables, etc. measured at quantitative level, ordinal level and qualitative (categorical) level (Warren 1990). Though "the obtained results are not at all satisfactory" (sic) (Carmichael 1990), due in part to the very large scale of the maps used, these results constitute a preliminary basis for subsequent studies.

Other variables used that have major problems are the following:

1. Lithic resources. They are based on present data but frequently we do not have the knowledge the prehistoric people had about such resources. Furthermore, we do not know whether the use of the resources remained the same during the cultural changes, or whether the prehistoric resources exist at present-day etc.
2. Distance to water resources being based in the knowledge about present-day rivers and other water resources. This data is not appropriate
because the hydrological conditions have changed over the last 6000-7000 years.

3. **Distance to forest, agricultural and hunting resources.** The distribution of such resources is based in actual soil cartographies that frequently do not have the same relationship as the distribution of resources in other periods. Furthermore, we do not know to what extent the human action altered the environment (shepherding, fires, ploughing, etc.).

Undefined concepts such as **geomorphological unit** defined in the form: "is the geographical unit whose limit is the place where the level curves change in spacing or in inflexion" (sic) (Nocete 1994) allow as many results as researchers investigating the problem and completely lack objectivity.

3 The set of variables

The conceptual design is made using the information obtained from archaeological sites and the relief surrounding each site, using archaeological considerations and concepts from geographical analysis to evaluate the general physical environment of the site, and the relationships between the site and its closest physical environment. This design is defined by means of the following concepts:

1. the site is defined as the area with archaeological remains on the surface.
2. the neighbourhood is defined as the circular area \( C_{500} \) defined by the circle of 500 meters radius and centered on the site; this area is considered to be the closest space to the site.

Using topographical maps to 1:10,000 scale we increased the precision and minimized errors (Esquivel and Peña, in press, show the short precision obtained using a 1:50,000 scale map). The variables used in the quantification were:

1) X-Y UTM coordinates of the archaeological site.
2) \( H_{\text{max500}} = \) upper altitude in \( C_{500} \).
3) \( H_{\text{min500}} = \) lower altitude in \( C_{500} \).
4) \( D_{500} = \) lowest map distance between upper and lower altitude points.
5) \( S_p = \) direction and orientation of maximum slope in \( C_{500} \), being the angle between the slope from upper altitude to lower altitude in \( C_{500} \) and the UTM North in base to 3D dip (see Fig. 1; Billings 1972):

![Figure 1. Direction and orientation of maximum slope in C500. 3D geological dip.](image)

6) \( P_1, P_2, P_3, P_4, P_5, P_6 \) define a usual concept in geophysical studies established by Griffin to obtain the residuals and remove the regional component (Dobrin 1976). These variables are the altitude of six points located in the circle with angular separation of 60°, \( P_1 \) being oriented to UTM North (Fig. 2).

![Figure 2. The Griffin circle centered on the UTM location of the archaeological site.](image)

These variables approximate the morphology of relief in the \( C_{500} \) circle (flatland, sheer, landing, etc.) showing the relationship between the site itself and the circle (located in a flatland, in a hill, etc.).

7) \( H_{\text{max}} \) and \( H_{\text{min}} \) represent the highest and the lowest altitude in the area with superficial material remains.

8) \( H_{\text{site}} = \frac{H_{\text{max}} + H_{\text{min}}}{2} \)

\( H_{\text{site}} \) is defined as the average altitude between the highest and the lowest altitude at the archaeological site.
9) \( D_{site} = \) linear map distance between the highest and the lowest altitude points in the site.

10) \( S_{site} = \) direction and orientation of maximum slope in the site being the angle between the slope from highest to lowest altitude in the site and the UTM North.

11) \( L_1 \) and \( L_2 \) are respectively the maximum axis and the perpendicular axis in the delimited area with material archeological remains.

12) \( S_1 = \) angle between the maximum axis with material remains \( L_1 \) and the UTM North.

13) \( D_{river} = \) is defined as the minimal linear map distance between the site and the present-day substantial rivers or water resources.

14) \( H_{river} = \) altitude of the river nearest the settlement.

15) A qualitative (categorical) variable CHRONO with a fundamental archaeological focus: the chronocultural attachment of the sites. This variable ranges from Neolithic period (including from Last Neolithic to Initial Copper), Copper Age (including Complete and Final Copper), Bronze Age (including Complete and Final Bronze), Iberian period, Roman period (including Upper and Lower Empire) and the Middle Ages.

3.1 The dataset

The previous quantification and the statistical analysis are applied to the Baza-Huéscar Basin that occupies the Orce, Galera, Castillejar and Huéscar highland plains, a very arid and dry area (350 mm. of annual rainfall only) \( 530 \leq X \leq 550, 4170 \leq Y \leq 4190 \) S30 UTM coordinates. At present, the population is located near the river valleys and the small well-drained fluvial plains to take advantage of the water in the rivers.

The data was acquired by means of a systematic archaeological surface survey with the following characteristics:

1. The survey was carried out by means of parallel transepts separated by 50-75 meters, using the 1:10,000 topographical maps created by the Andalusian Cartographic Institute.

2. The data field was used to obtain the X-Y UTM coordinates of the sites, the delimited area with archaeological remains and the chronocultural analysis and control of archaeological materials. The other variables were calculated from the 1:10,000 topographical maps.

3.2 Quantitative settlement index parameters

The Pearson correlations analysis, carried out using the previous variables, show the following results:

1. the only significative relationships \( (r>0.78) \) were between \( H_{max500} \) with \( H_{max} \) and the Griffin variables, between \( H_{max} \) and \( H_{min} \) and, finally, between \( D_{site} \) and \( H_{max} \) and the Griffin variables.

2. the remaining variables did not correlate with another variable (the major value for \( r \) is 0.37).

The results obtained using absolute variables were affected by the size factor and enabled us to define a set of index parameters to obtain morphometric relationships to quantify fundamental archaeological and geographical concepts, as follows:

1) \( I_1 \) measures the highest altitude in \( C_{500} \) relative to lowest altitude:

\[
I_1 = H_{max500} - H_{min500}
\]

2) \( I_2 \) measures if the relief is abrupt or not by means of the theoretical slope in \( C_{500} \):

\[
I_2 = \frac{H_{max500} - H_{min500}}{D_{500}}
\]

3) \( I_3 \) quantifies the relationship between the Griffin circle and the \( C_{500} \) area:

\[
I_3 = H_{pir} - H_{min500}
\]

4) \( I_4 \) quantifies the highest altitude in the site relative to lowest altitude in the \( C_{500} \) area:

\[
I_4 = H_{max} - H_{min500}
\]

5) \( I_5 \) measures the theoretical slope at the site:

\[
I_5 = \frac{H_{max} - H_{min}}{D_{site}}
\]

4 Statistical analysis

The analysis of each individual index shows the non-normality of each of these variables (rejection of null
hypothesis about normality). Furthermore, a nonparametric Kruskall-Wallis rank test (Hollander and Wolfe 1973) gave the following results (Fig. 3):

<table>
<thead>
<tr>
<th>Period</th>
<th>( I_1 )</th>
<th>( I_2 )</th>
<th>( I_3 )</th>
<th>( I_4 )</th>
<th>( I_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neolithic</td>
<td>45.9</td>
<td>44.4</td>
<td>35.6</td>
<td>43.1</td>
<td>32.9</td>
</tr>
<tr>
<td>Copper Age</td>
<td>48.5</td>
<td>57.5</td>
<td>53.9</td>
<td>41.4</td>
<td>48.3</td>
</tr>
<tr>
<td>Bronze Age</td>
<td>42.2</td>
<td>38.8</td>
<td>36.8</td>
<td>45.8</td>
<td>52.6</td>
</tr>
<tr>
<td>Iberian Period</td>
<td>29.6</td>
<td>29.3</td>
<td>26.5</td>
<td>30.9</td>
<td>22.5</td>
</tr>
<tr>
<td>Roman Period</td>
<td>29.9</td>
<td>28.6</td>
<td>31.4</td>
<td>30.9</td>
<td>31.5</td>
</tr>
<tr>
<td>Middle Age</td>
<td>40.1</td>
<td>43.</td>
<td>46.1</td>
<td>37.9</td>
<td>37.6</td>
</tr>
<tr>
<td>( t )</td>
<td>9.008</td>
<td>16.754</td>
<td>11.319</td>
<td>5.907</td>
<td>14.722</td>
</tr>
<tr>
<td>( p )</td>
<td>0.109</td>
<td>0.005</td>
<td>0.045</td>
<td>0.315</td>
<td>0.0116</td>
</tr>
</tbody>
</table>

Figure 3. Numerical results provided by the nonparametric Kruskal-Wallis rank test.

This results point out the classification effect in the CHRONO variable, showing the average ranks in the indexes. It is clear that the Neolithic, Copper and Bronze Age sites stand out upper ranks in all variables, the Iberian and Roman sites have the lowest ranks, and there are significative reasons (with a 0.05 signification level) to accept that the \( I_2 \), \( I_3 \) and \( I_5 \) sample values do not come from the same statistical populations in each chronocultural periods. Therefore, these indexes are different in each period and this identifies them as the most important discriminative parameters.

The principal components analysis is a powerful tool widely used in Archaeology (Doran and Hodson 1975; Esquivel and Contreras 1984; Dijndjian 1991; Baxter 1994). In the present work, we use the quantification provided by the indexes \( I_1 \) to \( I_5 \) for two main reasons: 1) to avoid the effects of type "size" that are obtained with the original variables, mainly using the absolute altitudes (see Hair, Anderson, and Tatham 1987 and Krzanowski 1988 to a complete revision), and 2) to analyze the relationships that exist between the shape of the relief that surrounds the site and the settlement; this relationship is independent of the absolute altitudes (Esquivel and Peña, in press).

The analysis was carried out normalizing the variables (\( I_1 \) to \( I_5 \) to mean 0 and variance 1 to avoid the size effect due to the range value of variables, using the correlation matrix as basis of analysis (Doran and Hodson 1975; Esquivel and Contreras 1984). Therefore, the value \( I_i=1 \) enables us to determinate the lower number of principal components that summarize the overall information in the data set. In this case, it is sufficient for further analysis to take the first two principal components that include 85% of overall variation in the data (Fig. 4):

The first component takes the form:

\[ y_1=0.95I_1+0.85I_2+0.91I_3+0.83I_4+0.53I_5 \]

showing a great weight in the \( I_1 \), \( I_2 \), \( I_3 \) and \( I_4 \) indexes while the settlement slope has less importance. This component is a "neighbourhood factor" that includes the neighbourhood indexes (they refer to upper altitude, theoretical slope and the average altitude in \( C_{500} \)) and the upper settlement altitude, while the theoretical slope of settlement has a small weight.

The second component has an important weight in \( I_5 \) only, discriminating between settlements with upper and lower slope; consequently, this component is a "slope factor in the settlement" that discriminates between settlements with greater and lesser slope.

The graphic display using the first two components shows a trend with respect to the cultural and chronological attachment that increases from left to right (Fig. 5):

Figure 5. Principal Components Analysis. The two first factors showing the trends in the settlements. The CHRONO variable was not included in the analysis and only appears in the display symbols.
1. the Copper and Bronze Age sites without further settlement follow a main trend (first quadrant) with settlements located in a high relief (relative to the minimum altitude in C500) and situated in an upper relative altitude. The sites in the quadrant IV show further settlement during Iberian and Roman Periods and are located in a high neighbourhood but have low slope at the site.

2. the Roman Period sites are located generally in a low neighbourhood and have a low slope at the settlement (quadrant II and III). The Roman Empire completely dominated the Iberian Peninsula and they were not concerned with seeking a defensible site: they lived in the better agricultural lands.

3. the Iberian sites have the same trend that the Roman settlements and each Iberian site was subsequently occupied by a Roman settlement.

4. the small quantity of Neolithic and Medieval period sites makes any conclusion about them negligible.

The further cluster analysis applied to the data set using the indexes as variables only confirms the previous results obtaining chronological groups, characterizing these clusters and obtaining the pattern of settlement in the region. The applied cluster analysis method is a hierarchical, agglomerative, non-overlapping and non-weighted technique, using as similarity measure the Euclidean distance and the single-linkage clustering algorithm (Doran and Hodson 1975; Esquivel and Contreras 1984). The variables were standarized to mean 0 and variance 1, to avoid the great weight of the variables with wide range of variation. In addition, we used the empirical SSE_2 rule of thumb as a homogeneity index to find the appropriate number of clusters (Johnson 1972). The resulting dendrogram shows the following clusters (Fig. 6):

Cluster 1. Contains the sites named "castellated" with great slopes with no flat surfaces or small ones resulting from human efforts at levelling. All the sites belong to the Argaric Bronze Age except one Roman settlement that is not a flat site. The neighbourhood is high in altitude, with both flat and sloped areas, and the site is higher with a steeper slope than the mean of neighbourhood.

Cluster 2. The sites in this cluster belong to Chalcolithic Age and have the same characteristics as the previous cluster (neighbourhood with a steep slope, high altitude and the site located higher than the neighbourhood mean) but the site shows a very low slope. The settlement is located on a large, high plateau. In this group there is a Roman site, because it was not situated in the flatlands.

Cluster 3. These sites belonging to the Chalcolithic Age are located at a lower altitude than the high surrounding neighbourhood, in a narrow river valley between steep banks. The settlements in this group have a steep slope, and the chalcolithic areas are located on low hills above the surrounding neighbourhood.

Cluster 4. This contains the sites located near the alluvial flatland and, as a general rule, theses occupy small hills located on the elevated fluvial plates not very near to the water courses or water sources. The neighbourhood is flat with a slope of nearly zero, and each site is located higher than its neighbourhood.

Cluster 5. In this group are included two sites placed on hills with flat surfaces that form small plateaux in
the valley and are clearly differentiated from the environment. The other settlements are located on glacis (stratum river superior) far from the course of the river and present slightly inclined surfaces. The slope in the site is near zero.

Cluster 6. This is comprised almost exclusively of flat deposits relatively near the water courses or sources. The sites and their neighbourhoods are quite flat with slopes near zero, with prevalence of the Roman settlements that in some cases have equal distances between them.

Cluster 7. The deposits of this group are located on hills which are not long but very well differentiated, although the lower ones are similar in height to the general neighbourhood. Also they can appear in the higher parts of the valley, that is, with soft forms and broad valleys, though the distances to the rivers or sources of water are not very large. In this group appear the argaric sites denoted "of flatness".

The last statistical analysis was used to find the intersite relationships by means of the X-Y UTM coordinates. This analysis is called the "Voronoi tessellation" and divides a 2-dimensional scatterplot into regions whose boundaries equally divide the distance between adjacent data points; any point within a particular region is closer to the plotted point in that region than it is to any other plotted point.

Using the entire data set, there was no evident rule of settlement pattern at all, but each chronocultural period had its specific characteristics (we excluded the periods with five or less sites, these providing a negligible information). The main results are the following:

Figure 7. Voronoi tessellations showing a different settlement pattern respective to separation between the sites: (a) Copper Age, (b) Bronze Age, (c) Iberian period and (d) Roman period. The Neolithic and Middle Age sites are not displayed due to the low number of sites.
1. the Copper Age sites located in a high neighbourhood are situated with equal minimal distances between them, about 2.5 Km. (Fig. 7a). The neighbourhood of the other sites (two only, with subsequent occupation) are plain and very similar to the sites in the Roman period (they are clustered with the Roman sites) maintaining a minimal separation greater than that of the previous sites.

2. the Bronze Age sites show an intersite minimal distance of 2.5-3 Km., including a site with flat neighbourhood and also sites with topographic parameters near 0 (Fig. 7b). This result shows the main feature of these sites: the Bronze Age sites maintain an almost constant distance between them, including the sites (one site only in this region) located in a flat neighbourhood.

3. the Iberian sites do not show any pattern of distance between them (Fig. 7c).

4. the analysis of the Roman sites as a group, reveals no general pattern between them, but considering only the sites with X>537 and Y>4177.5, the separation is merely about 400-500 meters (Fig. 7d). This result suggests that the settlement is made in function to the agricultural land use.

5 Settlement, classification of relief and water resources

There are two important variables in an analysis of the pattern of settlement: the proximity of water resources and the type of relief surrounding the site. We studied the proximity of courses or water sources at present-day, find a strong settlement pattern and low discrimination of this variable: all the sites are near river courses or sources of water, with distances ranging between 200 and 900 meters, but these data are not discriminatory because the set of sites have similar values in this variable (Fig. 8 and 9).

Figure 8. Location in UTM coordinates of prehistoric sites in the Orce-Huéscar region showing the settlement sites located very near river courses or water sources (we consider only near to river courses (Fig. 8)

Figure 9. Location using UTM coordinates of historic sites in the Orce-Huéscar region. The settlements sites are all situated on alluvial plains at all (we consider only present-day rivers and water sources).
Figure 10. Geomorphologic classification of the relief in the Orce-Huéscar region showing the four main patterns of relief.

The types of the relief constitute highly important data to find settlement patterns. The relief in the Orce-Huéscar region is classified in four main non-subjective categories (Peña 1979): alluvial plains that include rivers banks, soft relief, glacis relief and steep slopes. The largest number of sites is situated in the alluvial plains and the soft relief in order to exploit the natural resources (only one settlement is located in the glacis and there are no sites located on steep slopes).

Taking into account the location of prehistoric and historic sites with reference to relief (see Fig. 10) we obtained the following results:

1. The prehistoric settlement was carried out in the soft relief surrounding the alluvial plains present-day rivers and water sources).
2. The historic settlement was established on the alluvial plains to use the agricultural resources. The main locations are in the widest alluvial plains (Fig. 9), with greatest number of settlements very nearby mutually (distance about 300 or 400 meters).

6 Conclusions

We have established a characterization of settlement patterns using topographical information in the region that surrounds the site and in the own site by means of the location using the UTM coordinates of site and cartographical variables of neighbourhood. For significative and accurate results, the 1:10.000 scale maps must be used in the data recording process (less detailed scale maps give highly subjective, erroneous and vague results).

The initial set of variables includes a great number of variables (used by other authors) concerning the neighbourhood of the archaeological site and the site itself. We establish the neighbourhood as a circle, named C500, with 500 meters of radius being a quantitative and objective concept that quantifies very well the topographic characteristics in the neighbourhood of site. A correlation analysis shows that the most correlated variables are the highest altitude in C500, the lowest altitude in C500, the lowest map distance between upper and lowest altitude points, the altitude of six points located in the circle with angle separation of 60° being the first point oriented to UTM North, the highest and the lowest altitude in the area with superficial material remains and the lineal map distance between the highest and the lowest altitude points into the site. These variables cannot reveal the settlement pattern because they are absolute variables and thus provide only the "size effect". Using these variables, we have established indexes of form to quantify patterns of relief, we have found that the pattern of relief is independent of the absolute altitude (by example a plain). These indexes (I1, I2, I3, I4 and I5) measure the greatest altitude in C500, the theoretical slope in C500, the average altitude of C500, the greatest altitude in the settlement and the theoretical slope of settlement.

The statistical multivariate analysis of these indexes points out the following results:

1. The quantification by means of the previous indexes reveals the trends underlying the data set and enables us to discriminate between sites, and associate the chronocultural period of each site and its topographical characteristics.
2. The set of indexes classifies very well the pattern of settlements, indicating the relationship between the forms of relief and the chronocultural periods of sites.
3. In order of the arrangement of territory, we can discriminate two main topographic models: A) a model that comprises the Copper and Bronze Ages sites characterized by a high and steep topography situated on the banks of fluvial valleys and comprises the 1-3 groups, and B) constituted by sites in the flattest areas or near to agricultural terrains; these are all the sites of the Iberian and Roman periods, the Argaric sites in cluster 7 and the Final Neolithic period sites.
4. Using the X-Y UTM distance in the 1:10,000 maps between sites, the Voronoi tessellation indicates the different settlement patterns in Prehistoric and Historic Ages, showing the main characteristics: the Argaric sites are placed in locations that maintain equal distances between them, while the Roman and Iberian sites are almost always situated in the wide alluvial plains. This pattern induces a specific settlement of the argaric people seating in sites far between them to fixed distances. Furthermore, taking into account the Roman sites located in the wide alluvial plains only, and these sites maintain equal distances between them of about 300-400 meters to include the influence of the agricultural area of each settlement.

**Bibliography**


Esquivel, J A and Peña, J A (in press), Estudio y caracterización de asentamientos arqueológicos mediante métodos estadísticos, *Proyectos de investigación del Instituto de Estadística de Andalucía*, Junta de Andalucía, Sevilla


Nocete, F, 1994 *La formación del Estado en las Campiñas del Alto Guadalquivir (3000-1500 a.n.e.)*, Monografías de Arte y Arqueología, Servicio de Publicaciones, Universidad de Granada


**Contact details**

José A. Esquivel¹, M. Oliva Rodriguez-Ariza²
Department of Prehistory and Archaeology
University of Granada, Campus Cartuja s/n
18071 Granada
SPAIN
¹e-mail: esquivel@ululaya.ugr.es
²e-mail: oliva@ululaya.ugr.es
José A. Peña
Andalusian Institute of Geophysics
University of Granada, Campus Cartuja s/n
18071 Granada
SPAIN
e-mail: pena@gea.ugr.es