Archeologia e Calcolatori
26, 2015, 77-94

TAXICAB CORRESPONDENCE ANALYSIS OF ABUNDANCE DATA IN ARCHAEOLOGY: THREE CASE STUDIES REVISITED

1. Introduction

Recent publications (Baxter, Cool 2010; Alberti 2013a, 2013b; De Leeuw 2013; Siegmund 2014) show that correspondence analysis (CA) has become a popular method for the analysis of sites by artifacts abundance data in archaeology. CA is a factor analytic or dimension reduction method for exploratory visualization of non-negative data, such as counts or abundances. CA has body and soul: the body resides in the production of maps, mostly two-dimensional; the soul resides in the interpretation of the maps. We consider the maps and their interpretation as the essential aspects of CA.

This text exposes three specific illustrative examples displaying the essential aspects of CA, that is, the two-dimensional maps and their interpretation. As a corollary to Gould (1996, 106) who states that «science is rooted in creative interpretation», it follows that CA is a scientific method: in archaeology, CA maps are scientifically valuable if their interpretation produces meaningful results that allow inferring and describing past societies according to their social, cultural or economic activities. Otherwise, the CA maps are not interpretable, because they are contaminated by outliers.

Some authors who have analyzed abundance data in archaeology emphasize the fact that CA is sensitive to outliers, which may have disruptive impact on the maps (see for instance among others, Bølviken *et al*. 1982, 56-57; Lockyear 2000; Baxter, Cool 2010, section 4). Abundance data with outliers are described as "messy" or "noisy", resulting in CA maps that are not meaningfully interpretable. In these cases, the researcher either reduces the size of the data set or applies non-linear transformations with the aim of obtaining interpretable maps.

Reduction of the data set is done either by eliminating some sites (rows) or artifacts (columns) – see CA of the second data set taken from Alberti 2013a – or by grouping the types into major groups – see CA of the third data set taken from Bølviken *et al*. 1982. It should be emphasized that by reducing the size of the original data set to have interpretable results via CA, some useful information may inadvertently be thrown out. In the case of a non-linear transformation, a square-root or cubic-root is applied to the counts to reduce the influence of outliers before applying CA (see for instance Lockyear 2000 or Greenacre 2010). Non-linear transformation can be considered a re-coding of the data, frequently employed by the French school of data analysis developed by Benzécri.

How to identify and tackle outliers in CA of an abundance table is an unresolved open problem; for instance, Rao (1995) stressed the influence of rare observations (rows or columns that have relatively small weights compared to others) and proposed an alternative to CA based on Hellinger distance (a square-root transformation of the abundances). Greenacre (2013) highlighted the adverse influence of a zero-block structure in a data set and suggested its suppression from the data set. Similar to Greenacre's observation, earlier Novak and Bar-Hen (2005) observed that a zero-block structure in an abundance table disturbs CA results, but argued against the suppression of the zero-block. A third kind of outlier occurs when there are few relatively high valued cells. The last two data sets, that will be reconsidered in this paper, are heavily influenced by a combination of the three kinds of outliers: rare observations, zero-block structures and few relatively high valued cells.

In this paper we use a sturdy-robust-resistant variant of CA, named taxicab correspondence analysis (TCA). The aim is to show that this new method can smoothly handle any kind of abundance data in archaeology, and produce satisfactory meaningful results in the presence of outliers. Using TCA, we have reanalyzed many data sets previously analyzed by CA in published articles. We observed that if a data set does not have outliers, then CA and TCA maps will be very similar, resulting in similar interpretation (see example 1). However, if a data set has outliers, then CA and TCA maps will be different, and the interpretation of TCA maps will usually be clearer because of its robustness, as will be seen by the analysis of two data sets (see examples 2 and 3). Our experience has shown that only the comparison of CA and TCA maps of a data set will show if both maps are similar or not. For this reason, we suggest the parallel joint use of both methods. We have chosen three representative data sets: for the first data set CA and TCA produce similar results; for the second and third data sets CA and TCA produce different results. Further, these three data sets have been studied quite in detail by CA from an archaeological point of view; so we know at least what to expect from TCA.

This paper is organized as follows: in section 2, we present a brief non-mathematical descriptive review of CA and TCA; in section 3, we present the three case studies; and in section 4, we conclude with some general remarks.

For the interested reader, a reference for a mathematical exposition of the theory of CA in an archaeological context is the excellent paper by De Leeuw (2013). Since 2006, Choulakian and coauthors have published more than ten papers, where they studied mathematical properties of TCA applied to many kinds of non-negative data (in particular, TCA of contingency tables and their comparison with CA are studied in the following papers: Choulakian 2006; Choulakian *et al.* 2006; Choulakian 2008; Choulakian, Simonetti, Gia 2014). The recent book written by Beh and Lombardo (2014) presents a panoramic review of CA and related techniques.

2. Correspondence analysis

As stated in the introduction, CA and TCA are multivariate statistical methods, which summarize the essential aspects of a data set by projecting the multivariate data on two-dimensional maps. As an analogy for understanding what CA and TCA do to a data set, consider the act of taking a selfie with a smart phone. A selfie is a two-dimensional projection of a three-dimensional body by a smart phone; furthermore, a selfie can be taken from many different viewing angles, such as lateral, frontal, etc. In this analogy, the data set corresponds to a three-dimensional body; the smart phone corresponds to the multivariate method CA or TCA; selfies are the maps. Here, there are only two viewing angles: Euclidean by CA and Taxicab by TCA. Furthermore, these two viewing angles are based on some optimal mathematical properties.

Suppose that an abundance table is composed of n=30 sites by p=20 artifact types. There are three kinds of symmetric maps produced by CA or TCA: the map of the sites, which displays only n=30 sites, the map of the artifacts that displays only p=20 artifacts and the superimposed map – named biplot – which displays both n+p=30+20=50 sites and artifacts. How do we interpret these maps? For the first two kinds of maps, we use geometry. Let us consider the map of the n=30 sites: generally sites which are closer to each other in the space of artifacts described by the data will have projected points on the maps also closer to each other. Similarly, if we consider the map of the p=20 artifacts, artifacts which are closer to each other in the space of sites given by the data will have projected points on the maps also closer to each other. On the other hand, the interpretation of the biplot, the superimposed map of the n+p=50 sites and artifacts, can be done in three ways: the first two are based on geometry as described above; the third one is based on looking at the collection of sites and the collection of artifacts which are close to each other, where closeness of sites to artifacts means positive association among them. More details on the nature of these associations can be obtained by inspection of the abundances in the data set.

Our preference goes to the biplot, i.e. the superimposed map, which is richer; however, readability of a map plays a key role in its choice. Given that the biplot displays both the sites and the artifacts identified by their labels, the biplot might be cluttered, and the labels not readable. In this case, we use the sites map and the artifacts map separately. Indeed, the use of a representative labelling of sites and artifacts simplifies the interpretation of the map and makes it easier to see facts and associations among different characteristics of the points. We will also attempt to identify the three kinds of outliers that a table can have as outlined above: rare observations, zero-block structure and relatively high valued cells. In the case of rare observations, we shall identify them in the labelling.

3. Case studies

We shall compare CA and TCA results on three representative data sets: as already said, for the first data set CA and TCA produce very similar results; for the second and third data sets CA and TCA produce different results.

3.1 *Ksar Akil data*

Tab. 1, copied from Alberti 2013b, who cites Shennan (1997, 355-357) as his source, describes the abundances of 5 lithic types excavated from 10 levels at the Palaeolithic cave in Ksar Akil (Lebanon). Fig. 1 displays both the CA and TCA maps, where adjacent levels have been joined by a line. We note that, in both maps, the positions of the corresponding points are almost identical; so both maps have the same interpretation. Alberti provides a much detailed interpretation of the biplot; here, we provide two main points using his text:

a) The first dimension opposes the first six levels (1-6) to the last four levels (7-10). Furthermore, levels 7-10 are associated with lithic types blades and flake blades; while levels 1-6 are associated with lithic types partially cortical, non cortical and bladelets.
b) The 10 levels display a "slight parabolic curve".

3.2 *Punta Milazzese of Panarea data*

Tab. 2, copied from Alberti 2013a, presents the abundances of 31 artifacts found in 19 huts, excavated at the Middle Bronze age settlement at Punta Milazzese on the island of Panarea (Aeolian Archipelago, Italy). Two complementary pieces of information on the huts are also available from Alberti 2013a: first, their geographic locations displayed in Fig. 2 (upper diagram); second, their surface areas in m^2, which we include in the labelling in Tab. 2. The labelling of the huts and the artifacts is explained below. Further archaeological references, maps and detailed CA results are given in Alberti 2013a, which we shall reuse in this text. The aim of using CA and TCA techniques is to help discover meaningful patterns and clusters of the huts and the artifacts, from which some useful information may be inferred on the past activities of the settlement.

First, we note that the data set has a lot of zeros. One way to measure the sparsity of the data set is to compute the percentage of the zero abundances, which is 58%; further, given the great number of zero abundances, one can see the presence of many zero-blocks in the structure of the data set. Second, there are two relatively high abundance values of 12, one of them is an outstanding outlier as will be seen in subsection 3.2.1. Third, looking at marginal sums of the abundances, found in the last column and the last row of the data set, we can identify the presence of some rare observations, such

Levels	Partially cortical	Non cortical	Flake blades	Blades	Bladelets	Sum
1	2	12	6	12	4	**36**
2	16	44	14	6	4	**84**
3	72	105	54	55	69	**355**
4	111	87	114	148	115	**575**
5	35	40	48	47	55	**225**
6	60	74	76	53	56	**319**
7	62	51	206	127	66	**512**
8	24	50	80	67	31	**251**
9	52	177	344	205	75	**853**
10	21	81	138	31	22	**293**
Sum	**455**	**721**	**1080**	**751**	**496**	**3503**

Tab. 1 – Ksar Akil data.

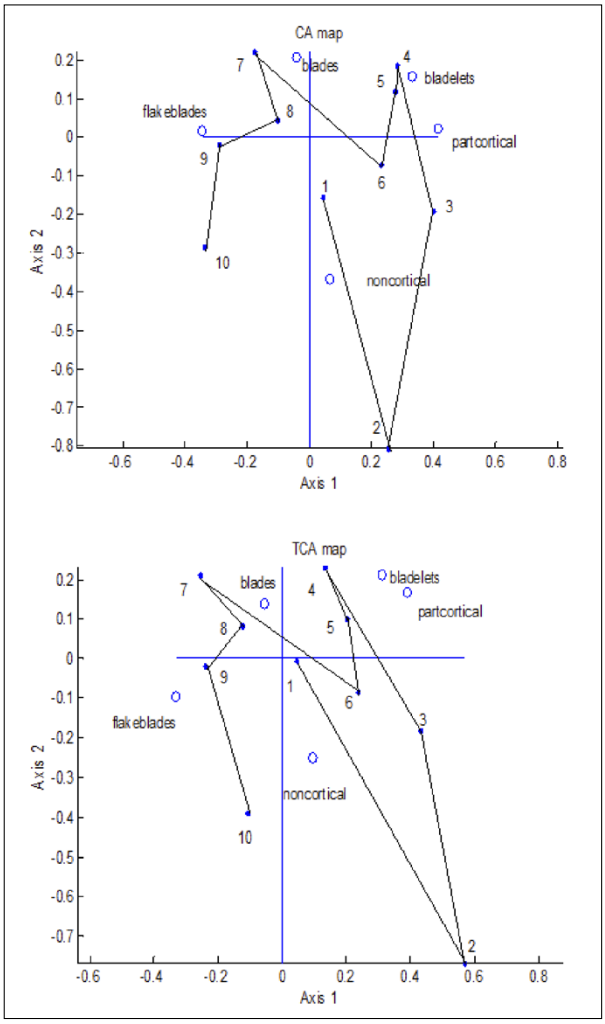

Fig. 1 – CA and TCA maps of Ksar Akil data.

Label	Functionality	Object type	10M1	14M2	17M3	20M4	16M5	19M6	11M8	12M9	M10	21M16	20M18	14M20	M4A	16M11	11M12?	17M13	M14?	10M15	M19	SUM
29cookT	cooking	tray	2	5	3	0	1	0	2	4	2	1	1	2	1	3	0	0	1	0	1	**29**
5cookV	cooking	vessel	0	0	0	0	0	0	0	1	0	1	0	3	0	0	0	0	0	0	0	**5**
2cookS	cooking	stand	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**2**
1cookT	cooking	Ae-tray	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	**1**
10cookH	cooking	hook	0	0	2	1	0	0	0	1	0	1	1	0	3	0	1	0	0	0	0	**10**
21dinS	dinner	stand	0	1	4	2	2	1	1	0	0	3	2	1	0	2	0	0	0	1	1	**21**
29dinV	dinner	A-open vessel	0	1	2	1	1	0	1	1	4	3	2	2	0	6	0	3	1	1	0	**29**
5dinV	dinner	Ae-open vessel	0	0	0	0	0	0	0	0	1	1	1	0	0	2	0	0	0	0	0	**5**
33dinB	dinner	fine-ware bowl	1	3	1	1	1	2	2	5	4	2	1	3	0	2	0	2	0	2	1	**33**
10dinB	dinner	coarse-ware bowl	0	0	0	0	0	0	0	2	1	0	1	2	1	0	0	0	2	0	1	**10**
9dinV	dinner	A-closed vessel	1	1	2	0	1	0	0	1	2	0	0	1	0	0	0	0	0	0	0	**9**
47dinJ	dinner	fine-ware jug	1	3	4	3	2	2	2	4	4	2	5	5	1	7	0	0	0	1	1	**47**
8dinB	dinner	big bowl	1	0	0	2	1	0	0	2	0	0	0	0	1	0	0	0	0	1	0	**8**
19dinB	dinner	small bowl	1	3	2	1	0	0	2	1	2	1	0	1	1	1	1	1	0	0	1	**19**
5dinB	dinner	small bowl	1	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**5**
3pouJ	pouring	coarse-ware jug	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	**3**
29proMP	processing	mortar/pestel	0	2	2	2	0	0	1	12	3	1	0	2	1	2	0	0	0	1	0	**29**
16proMH	processing	millstone/handstone	0	4	1	0	0	0	1	3	1	0	0	6	0	0	0	0	0	0	0	**16**
10spinW	spinning	spindle wholrs	0	0	0	0	0	0	0	2	5	0	0	0	0	0	0	2	0	0	1	**10**
21storV	storing	big vessel	0	2	2	1	0	1	2	2	2	4	0	1	1	1	0	0	0	1	1	**21**
8storV	storing	Ae-closed vessel	0	1	1	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	1	**8**
27storV	storing	small vessel	2	2	1	2	1	1	3	4	2	1	2	3	0	1	0	1	0	0	1	**27**
3storV	storing	small vessel	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	**3**
11storV	storing	small vessel	2	2	1	1	0	1	0	1	1	0	0	1	0	0	0	0	0	0	1	**11**
20storV	storing	vessel	2	3	1	1	1	0	2	2	1	1	1	0	1	0	1	1	0	1	1	**20**
13workST	working	stone tool	0	2	1	0	1	0	2	3	4	0	0	0	0	0	0	0	0	0	0	**13**
53workL	working	lithic core/flake	0	4	7	0	1	6	1	7	8	5	5	6	0	0	0	3	0	0	0	**53**
1covL	covering	A-lid	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
5workL	working	lid	0	0	1	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**5**
18workV	working	miniature vessel	1	0	0	3	0	2	0	0	1	1	0	3	0	4	0	1	1	0	1	**18**
18workP	working	pebble	0	0	0	0	0	**12**	1	0	0	0	0	0	0	**5**	0	0	0	0	0	**18**
		SUM	**16**	**43**	**39**	**23**	**14**	**28**	**25**	**60**	**53**	**29**	**23**	**44**	**11**	**37**	**3**	**14**	**5**	**9**	**13**	**489**

Tab. 2 – Punta Milazzese of Panarea data.

as huts 12? (it has 3 artifacts) and 14? (it has 5 artifacts). The question mark indicates that we consider them as rare observations.

To facilitate the interpretation of CA and TCA maps, we use a special row labelling in Tab. 2, where the row label has 3 parts: the first part of a label represents the *marginal abundance* of the artifact, given in the last column; it is followed by the *function* of the artifact, described in the 2nd column; then it is followed by its *type*, described in the 3rd column. For instance, consider

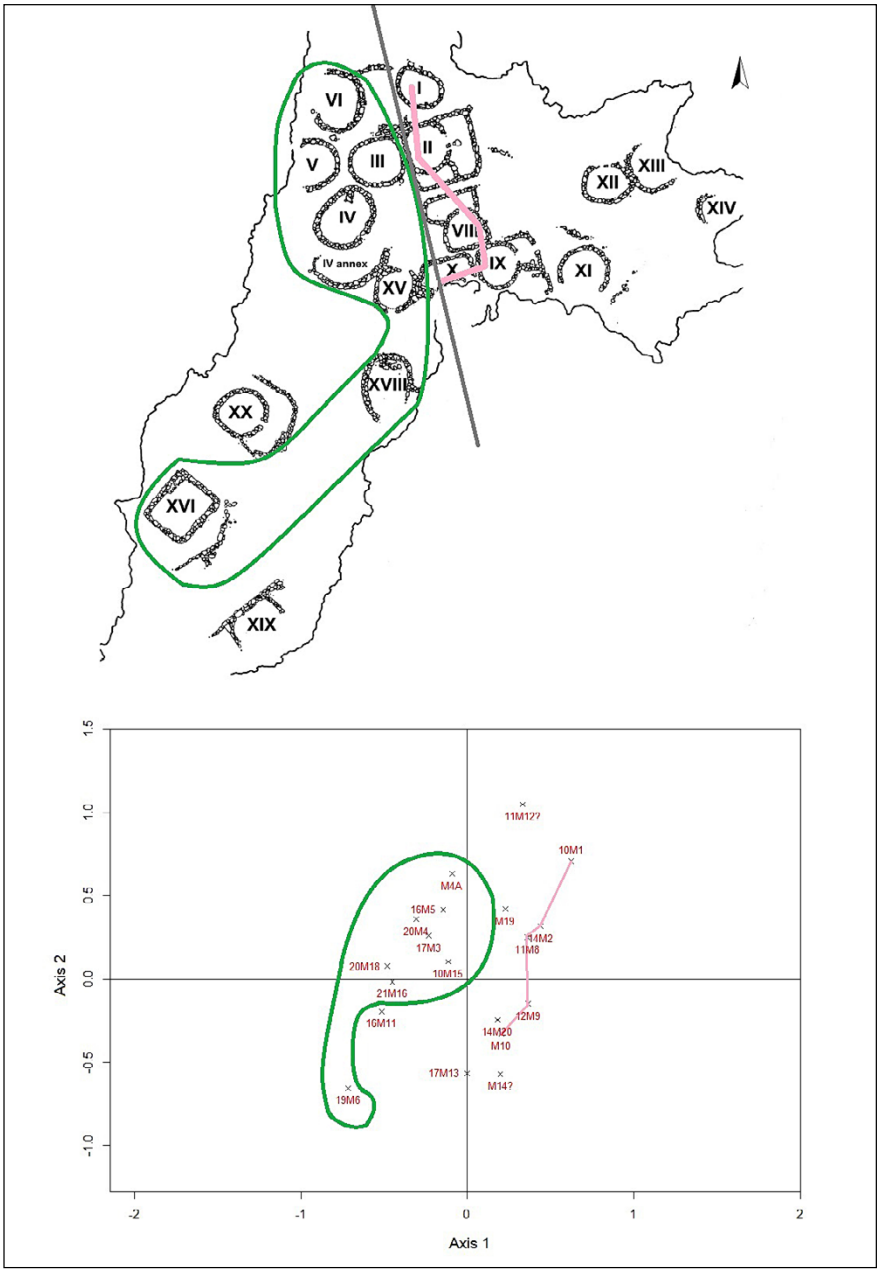

Fig. 2 – Geographical map of the studied huts (upper) and TCA display of the huts from the Punta Milazzese dataset (lower).

the first artifact label *29cookT*: its first part is its total abundance number of *29 (it is not a rare observation)*, which is also found in the last column; the second part is *cook*, which represents the cooking function of the object described in the 2nd column; its third part is the letter *T*, which conveys its description as a tray given in the 3rd column. For instance, artifacts *5cookV, 2cookS* and *1cookT* are rare observations, because their first two digits (marginal abundances) are 5, 2 and 1, which are small. This labelling aids us in seeing the effect of rare observations on the diagram, which according to Rao (1995) might produce difficulties in its interpretation.

Similarly, the column label of a site is composed of 2 parts: the first represents the surface area of the hut, followed by its index. The index of the hut corresponds to the Roman numeral used in the geographical map in Fig. 2. For instance, consider *10M1* the label of the first hut: it means that hut number 1 has surface area of 10 m^2. The question mark found in the labels of the huts 11M12? and M14? indicates that we consider them as rare observations, their total abundances being too small (3 and 5). Further, four surface areas of huts are missing, their labels are: M4A (A means annex), M10, M14 and M19.

3.2.1 Correspondence analysis results

Figs. 3 and 4 respectively display the CA maps of the 31 objects and 19 sites separately. The first dimension in Fig. 3 is dominated by the artifact *18workP* (18 working pebbles); the artifact *18workP* characterizes the two huts 19M6 and 16M11, which dominate the first axis of Fig. 4. Looking at the data, we find in the second last row of Tab. 2 that 12 working pebbles were recovered at hut 19M6, and 5 working pebbles at hut 16M11. Therefore, clearly the abundance value of 12 is an outlier cell with adverse influence. Alberti continued his CA analysis by deleting 7 rows (rows 4-5, 15, 28-31) and 7 huts (the last seven huts to the right); for further details refer to his published article: Alberti 2013a.

3.2.2 Taxicab correspondence analysis results

Fig. 5 presents the TCA biplot of Tab. 2, where we see the huts' distribution on two seemingly parallel lines. Given that the biplot in Fig. 5 is quite cluttered and labels are not clearly readable, we represent only the TCA map of the huts in Fig. 1 (lower diagram). We were able to interpret the first two principal axes, which explain 31% of the taxicab dispersion, the contribution of the first axis being 17.66% and that of the 2nd axis being 13.34%. The first axis in the TCA map of the huts has very clear interpretation: in fact it represents two different but complementary aspects of the huts, as follows (Tab. 3):

a) Consider the geographical location of the 19 huts displayed in the upper part of Fig. 2. The oblique line divides the settlement into two clusters, eastern (9 huts) and western (10 huts). In Tab. 3, the list of the huts of these two clusters is shown using the original Roman numerals. Now, we consider the

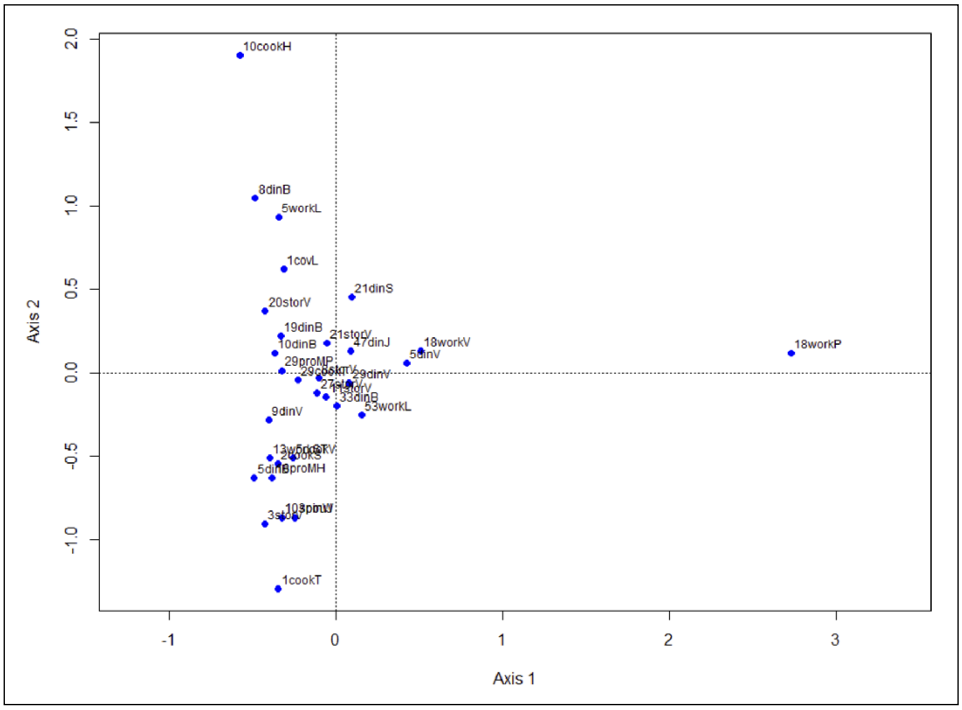

Fig. 3 – CA map of the 31 object types. In each label, the number represents the abundance of the object type, followed by the function of the object, then by its type. Refer to Table 2 for the details of the labels.

Western			**Eastern**		
Area	Roman numerals	Negative 1st axis coordinates	Area	Roman numerals	Positive 1st axis coordinates
19	VI	19M6	10	I	10M1
16	V	16M5	14	II	14M2
17	III	17M3	11	VIII	11M8
20	IV	20M4	12	IX	12M9
	IV annex	20M4A	10	X	M10
10	XV	10M15		XI	
20	XVIII	20M18	11	XII	11M12?
14	XX		17	XIII	17M13
21	XVI	21M16		XIV	M14?
	XIX				
		16M11			14M20
					M19

Tab. 3 – List of the huts divided in the two visible clusters.

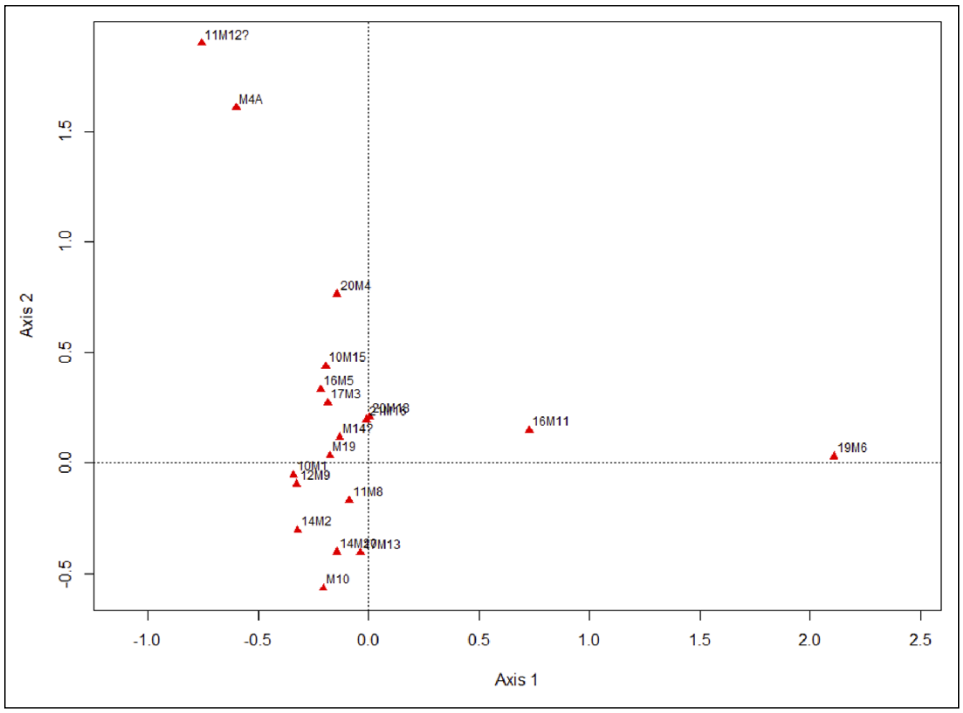

Fig. 4 – CA map of the 19 huts. In each label, the number represents the surface area in m² of a hut followed by its index.

TCA map of the huts displayed in the lower part of Fig. 2. The first axis divides the huts into two parts according to the sign of their coordinates on the first axis; these are also displayed with their labels in Tab. 3. The comparison of the four columns, two by two, shows that the majority of the huts in the western cluster have negative first axis coordinates, and the majority of the huts in the eastern cluster have positive first axis coordinates. The first axis misclassifies only three huts (16M11, 14M20 and M19) out of 19 that is 15.8%, which is acceptable. So we interpret the first dimension of the TCA map as an East-West contrast between two clusters of huts: huts of the eastern cluster which have mostly positive coordinates on the first axis, and huts of the western cluster which have mostly negative coordinates on the first axis. To make this assertion visually clearer, we have encircled the common western huts in the upper geographical map and in the lower TCA map in Fig. 2.

b) Here, by looking at the surface areas of the huts in Tab. 3 we provide another interpretation of the first dimension as follows: 8 out of 9 huts with negative first axis coordinate have surface areas larger than 15 m²; while only

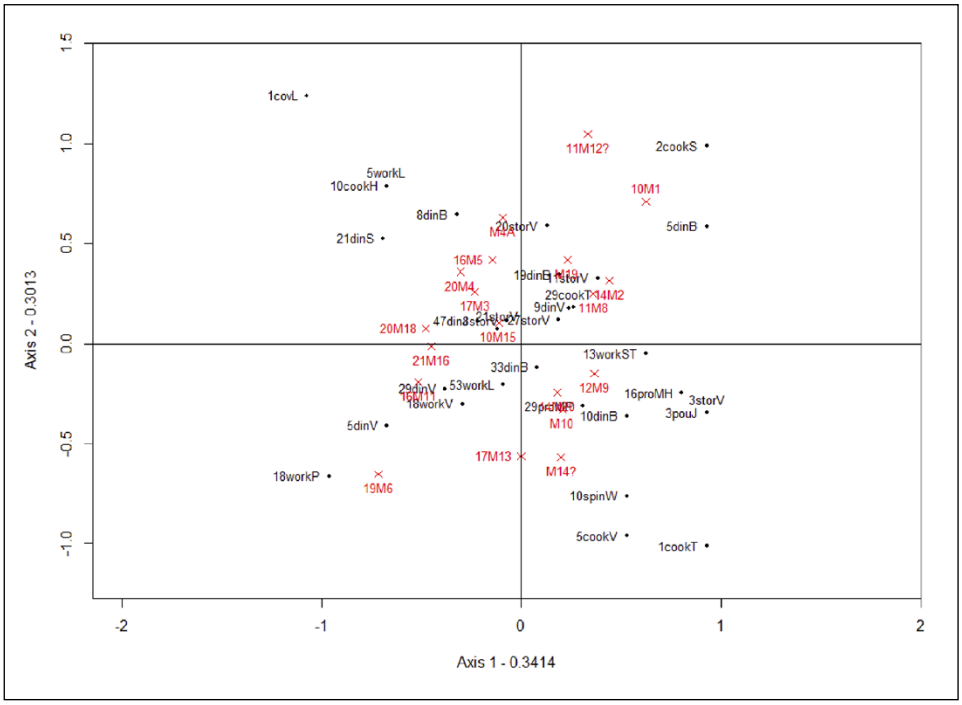

Fig. 5 – TCA map of the 31 object types and the 19 huts of Punta Milazzese.

1 (17M13) out of the 10 huts with positive first axis coordinate has known surface area larger than 15 m^2. However, the hut M13 with a surface area of 17 m^2, has almost 0 first axis coordinate, so it may also belong to the left bundle of huts. So, the first dimension opposes huts with surface area larger than or equal to 15m^2 to huts with surface area smaller than 15m^2.
c) Alberti 2013a suggested two functional uses of the huts, residential or utilitarian; and based on some hypotheses, he identified only 5 residential huts: 21M16, 20M18, 19M6, 17M3 and 16M11. In Fig. 2, these 5 huts have negative first axis coordinates and are found in the western sector.

The second axis does not have a clear-cut interpretation like the first axis. However, we note that for the eastern cluster of huts, the five huts bordering the oblique line are ordered on the second axis. This is represented in both diagrams of Fig. 2 by joint segments. So there is a hint of North-South opposition on the second axis partially for the eastern cluster of huts.

Finally, the artifacts associated with huts in the TCA map (Fig. 5) show a certain support to the hypothesis of spatial separation according to activities. The first and fourth quadrants' huts show a predominance of work, dinner,

cooking and storing artifacts, while the artifacts found in the second and third quadrants' huts seem to point toward a more diversified range of activities, like working, processing, dinner, spinning, but also to storing, cooking and dinner. These are evidently global associations.

Here, we discuss specific local associations by examining the distribution of the artifacts in the third quadrant, where the huts 19M6, 16M11, 17M13 and 20M18 are mainly associated with the artifacts *18workP* (18 working pebbles), *53workL* (53 working lithic core/flakes), *18workV* (18 working miniature vessels), *29dinV* (29 dinner A-open vessels) and *5dinV* (5 dinner Ae-open vessels). Some remarks are to be made: together, 19M6 and 16M11 form almost 95% of the weight of *18workP*, which was very influential in CA. Also, *18workP* is closer to 19M6 because it accounts by itself for 12 out of the 18 occurrences of this artifact type. 17M13 is situated exactly on the second axis, because by examining the data in Tab. 2 we see that the eight artifacts that characterize 17M3 are equally distributed in each side of the first dimension.

Another interesting feature is that generally rare observations appear on the periphery of the TCA map; in particular this is true for the two huts (11M12? and M14?) considered as rare observations. Recall that "?" in the label of a hut signifies "rare", because hut M12 contains three artifacts and hut M14 five artifacts.

3.3 *Iversfjord data*

Tab. 4, copied from Bølviken *et al.* (1982, Tab. 1), presents the abundances of 37 artifacts found at 14 Late Stone Age house sites near Iversfjord, Arctic Norway. More details regarding archaeological references, maps and detailed CA results are given in that article. The 37 artifact types belong to 9 general function categories that we symbolize as: p=points, sb=scrapers/burins, ct=core tools, k=knives, ns=net sinkers, tm=tool manufacture, sf=slate fragments, uf=utilized flakes and ps=perforated stones. A visual inspection of Tab. 4 reveals that: first, the data set has a lot of zeros: 60.04% of abundances are null; second, given the great number of zero abundances, many zero-blocks are seen to be present in the structure of the data set; third, looking at the last column and the last row of the data set, which display the marginal sums of the abundances, some rare observations may be identified, such as huts 12 (it contains 8 artifacts) and 13 (it contains 10 artifacts).

The row labelling of the artifacts in Tab. 4 is composed of two parts: an abbreviation of the function category of the artifact, followed by its total abundance. For instance consider the label p110: the letter p signifies points (artifact type) and the number 110 reproduces the marginal abundance (found in the last column of Tab. 4). A column label is composed of 2 parts: the last

	161h1	62h2	36h3	284h4	152h5	260h6	92h7	208h8	26h9	23h10	18h11	8h12	10h13	27h14	SUM
p110	14	6	3	9	5	28	2	22	4	6	9	1	0	1	**110**
p20	4	2	2	3	2	2	0	3	0	2	0	0	0	0	**20**
p16	0	0	0	3	0	6	1	6	0	0	0	0	0	0	**16**
p3	0	0	0	0	1	0	0	2	0	0	0	0	0	0	**3**
p5	1	0	0	0	0	0	0	2	1	1	0	0	0	0	**5**
sb107	4	1	0	26	13	29	11	14	0	0	1	1	4	3	**107**
sb85	2	1	0	33	11	20	7	6	1	0	0	2	1	1	**85**
sb50	2	0	0	11	3	17	6	10	0	0	0	0	1	0	**50**
sb119	9	2	0	42	15	16	16	11	2	0	0	2	0	4	**119**
sb11	0	0	0	4	3	2	2	0	0	0	0	0	0	0	**11**
sb19	1	0	0	9	0	3	2	3	0	1	0	0	0	0	**19**
sb3	0	0	0	0	0	2	1	0	0	0	0	0	0	0	**3**
sb4	0	0	0	2	1	1	0	0	0	0	0	0	0	0	**4**
sb4	1	0	0	0	1	1	0	1	0	0	0	0	0	0	**4**
sb2	0	0	0	0	0	1	0	1	0	0	0	0	0	0	**2**
sb47	3	3	0	9	6	7	2	12	2	1	0	1	0	1	**47**
sb1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**1**
ct8	1	0	0	4	1	0	0	0	0	0	1	0	1	0	**8**
ct2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	**2**
ct2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	**2**
ct1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**1**
k31	15	2	0	2	5	3	0	4	0	0	0	0	0	0	**31**
k8	2	0	0	0	2	4	0	0	0	0	0	0	0	0	**8**
k1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**1**
ns37	19	5	3	0	5	2	0	2	0	0	1	0	0	0	**37**
tm334	38	22	19	60	39	51	14	64	2	6	3	1	1	14	**334**
tm40	18	4	1	4	4	5	1	2	0	0	0	0	0	1	**40**
tm10	0	0	0	2	3	4	0	0	0	0	0	0	0	1	**10**
tm8	0	0	0	2	1	0	0	4	1	0	0	0	0	0	**8**
tm11	0	0	0	2	0	2	0	3	3	1	0	0	0	0	**11**
sf33	6	2	2	2	3	2	3	8	4	0	1	0	0	0	**33**
sf24	4	6	0	1	7	1	0	3	2	0	0	0	0	0	**24**
sf13	0	0	4	1	1	0	0	0	2	5	0	0	0	0	**13**
uf188	14	5	1	48	18	49	24	23	2	0	1	0	2	1	**188**
uf3	0	0	0	2	0	0	0	1	0	0	0	0	0	0	**3**
uf7	2	0	1	3	1	0	0	0	0	0	0	0	0	0	**7**
ps2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	**2**
SUM	**161**	**64**	**36**	**284**	**152**	**260**	**92**	**208**	**26**	**23**	**18**	**8**	**10**	**27**	**1369**

Tab. 4 – Iversfjord data.

number following the letter "h" indicates the house site index, which varies from 1 to 14; the number preceding "h" indicates the total sum of artifacts excavated at the particular house site, identified by the number given in the

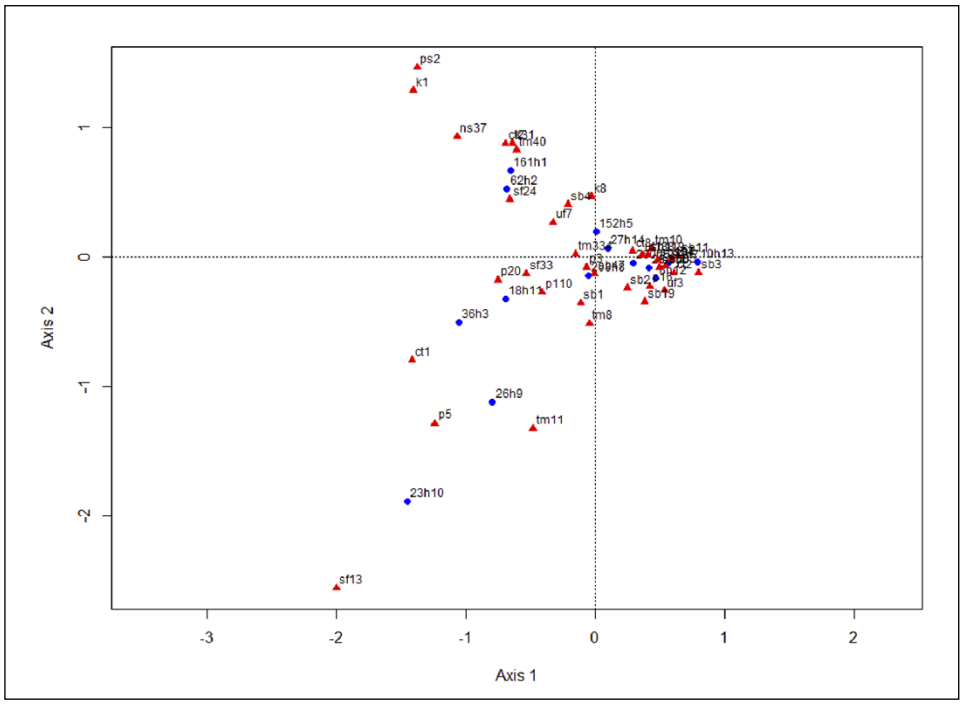

Fig. 6 – CA map of the Iversfjord data.

last row of Tab. 4. For instance, 161h1 means that there were 161 artifacts excavated at house site 1.

3.3.1 Correspondence analysis results

According to Bølviken *et al.* (1982), the main aim of their study was to find clusters of the sites that are characterized by different kinds of economic activities. So, they performed CA to the data set in Tab. 4 and obtained the CA biplot displayed in Fig. 6. Fig. 6 is funnel-shaped because of the particular structure of the data set; they did not find it to be meaningfully interpretable. So, the authors collapsed the original data set into an abundance table of size 9×14, by grouping the 37 artifacts into 9 general tool categories as described in the introduction. They interpreted the CA biplot (not shown) of the collapsed data set in the following meaningful terms:

a) House sites 8, 9 and 10 are associated with projectile points and worked slate fragments, which reflect sea-mammal hunting activities.
b) House sites 1, 2 and 3 are associated with knives, net sinkers and perforated stone, which reflect fishing activities.

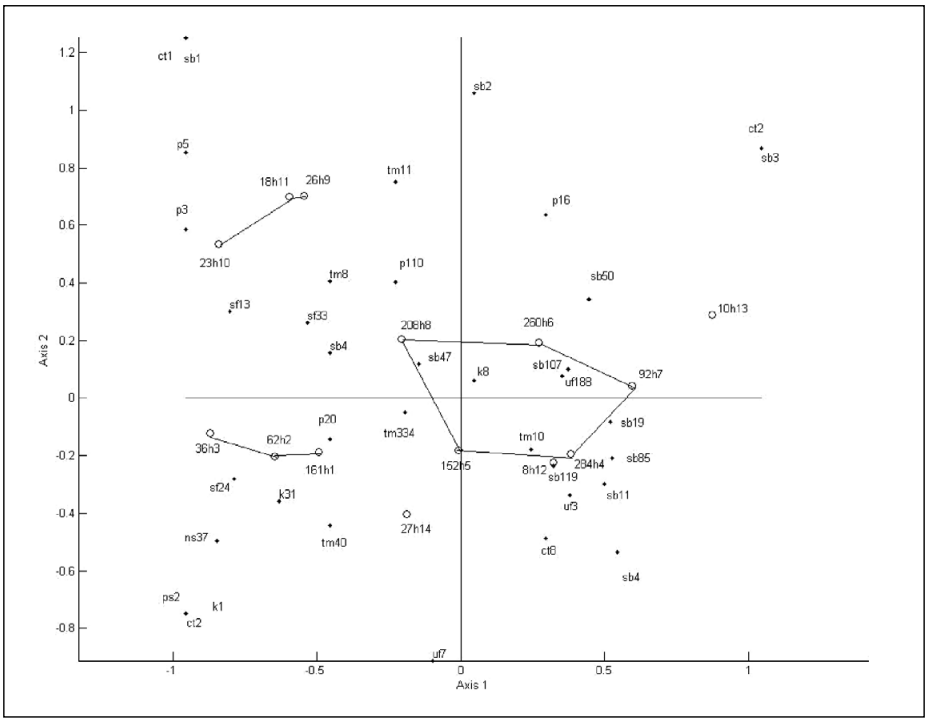

Fig. 7 – TCA map of the Iversfjord data.

c) The remaining house sites are associated with tool maintenance and scrapers/burins and utilized flakes, which reflect maintenance activities.

Consequently, the authors discovered three distinct clusters of sites, each cluster being characterized by a specific economic activity. In the next subsection it will be seen that these results may be obtained directly without collapsing the original data set into a smaller one.

3.3.2 Taxicab correspondence analysis results

Applying TCA to the data set of Tab. 4, it was possible to interpret the first two principal axes, which explain 64.14% of the taxicab dispersion; the part of the first axis being 47.92% and the part of the 2nd axis being 16.22%. Fig. 7 displays the TCA biplot of the 37 artifacts and the 14 huts, for which some details may be provided. In fact, we identify four clusters of house sites as follows:

a) House sites labeled 26h9, 23h10 and 18h11, located in the 2nd quadrant of Fig. 7, are associated with projectile points (p110) and worked slate frag-

ments (sf13 and sf33); the abundances in Tab. 4 support these associations. Additionally, it is worth noting that the weight of these three hunting house sites is very small, 4.89% = (26+23+18)/1369. So huts 9, 10 and 11 form a peripheral cluster characterized by hunting activities.
b) House sites labeled 161h1, 62h2 and 36h3, located in the 3rd quadrant of Fig. 7, are associated with knives (k31, k1), net sinkers (ns37) and perforated stone (ps2); the abundances in Tab. 4 support these associations. The weight of these three house sites is 18.92% = (161+62+36)/1369, which is almost 4 times larger than the weight of the three hunting sites. Note that even though site 14 (27h14) is located in quadrant 3, it is not associated with the fishing house sites (1, 2 and 3), because its abundances do not point in this direction: 14 out of 27 (more than 50%) artifacts found at site 27h14 belong to the type tool manufacture (tm334). So huts 1, 2 and 3 form also a peripheral cluster characterized by fishing activities. However, because of its weight, this fishing cluster is much more important than the previous hunting cluster.
c) The central five house sites labeled 284h4, 152h5, 260h6, 92h7 and 208h8, form a pentagon, whose weight is 72.75%=(284+152+260+92+208)/1369; they are characterized by tool maintenance (tm334, tm10), scrapers/burins (sb4, sb11, sb19, sb47, sb50, sb85, sb107, sb119), and utilized flakes (uf188). So huts 4, 5, 6, 7 and 8 form the core cluster characterized by maintenance activities; around which the other clusters are found.
d) We identify a fourth cluster of three dispersed house sites labeled 8h12, 10h13 and 27h14 with considerably small weight: 3.29%=(8+27+10)/1369. We consider them as rare observations, probably abandoned mainly maintenance house sites.

We can summarize our interpretation by an ordering of the four clusters of huts according to their weights: central cluster of 5 huts for maintenance activities (weight=72.75%), peripheral cluster of 3 huts for fishing activities (weight=18.92%), peripheral cluster of 3 huts for hunting activities (weight=4.89%) and an abandoned cluster of 3 huts (weight=3.29%).

4. Conclusion

Correspondence analysis has been gaining popularity among archaeologists in the past few years and is often applied to archaeological abundance data. Sometimes data sets are sparse, where the degree of sparsity of a data set is defined as the percentage of zero abundances. For sparse data sets, three kinds of potential outliers may be identified: rare observations, zero-block structure and relatively high valued cells. Often CA is very sensitive to a combination of the aforementioned three kinds of outliers. In those cases, we suggest the use of both methods CA and its robust version TCA. Each method

sees the data from a particular angle; sometimes the views are similar, other times different. The use of both methods is enriching and useful as shown by the reanalysis of three well-known data sets in this paper.

Solène Mallet Gauthier

Département des Sciences Historiques
Université Laval, Québec, Canada
solene.mallet-gauthier.1@ulaval.ca

Vartan Choulakian

Département de Math/Statistique
Université de Moncton, NB, Canada
vartan.choulakian@umoncton.ca

Acknowledgements

NSERC of CANADA financed both authors for this research. The authors thank the editor Dr. Paola Moscati, the associate editor Dr. Alessandra Caravale, two anonymous reviewers and Pr. F. Ashkar for their constructive comments.

REFERENCES

Alberti G. 2013a, *Making sense of contingency tables in archaeology: The aid of correspondence analysis to intra-site activity areas research*, «Journal of Data Science», 11, 479-499.

Alberti G. 2013b, *An R scipt to facilitate correspondence analysis. A guide to the use and the interpretation of results from an archaeological perspective*, «Archeologia e Calcolatori», 24, 25-53.

Baxter M.J., Cool H.E.M. 2010, *Correspondence analysis in R for archaeologist: An educational account*, «Archeologia e Calcolatori», 21, 211-228.

Beh E., Lombardo R. 2014, *Correspondence Analysis: Theory, Practice and New Strategies*, New York, Wiley.

Bølviken E., Helskog E., Helskog K., Holm-Olsen I.M., Solheim L., Bertelsen R. 1982, *Correspondence analysis: An alternative to principal components*, «World Archaeology», 14, 41-60.

Choulakian V. 2006, *Taxicab correspondence analysis*, «Psychometrika», 71, 333-345.

Choulakian V. 2008, *Taxicab correspondence analysis of contingency tables with one heavyweight column*, «Psychometrika», 73, 309-319.

Choulakian V., Kasparian S., Miyake M., Akama H., Makoshi N., Nakagawa M. 2006, *A statistical analysis of synoptic gospels*, in *8es Journées Internationales d'Analyse Statistique des Données Textuelles*, *JADT 2006 (Besançon 2006)*, Besançon, Presses Universitaires de Franche-Comté, 281-288.

Choulakian V., Simonetti B., Gia T. 2014, *Some new aspects of taxicab correspondence analysis*, «Statistical Methods and Applications», 23, 401-406.

De Leeuw J. 2013, *Correspondence analysis of archaeological abundance matrices*, in C.R. Nance, J. De Leeuw, K. Prado, D. Verity (eds.), *Correspondence Analysis and West Mexico Archaeology: Ceramics from the Long-Glassow Collection*, Albuquerque, University of New Mexico Press, 67-100.

Gould S.J. 1996, *The Mismeasure of Man*, 2nd ed., New York, W.W. Norton and Co.

Greenacre M. 2010, *Log-ratio analysis is a limiting case of correspondence analysis*, «Mathematical Geosciences», 42, 129-134.

Greenacre M. 2013, *The contributions of rare objects in correspondence analysis*, «Ecology», 94, 1, 241-249.

Lockyear K. 2000, *Site finds in Roman Britain: A comparison of techniques*, «Oxford Journal of Archaeology», 19, 397-423.

Novak E., Bar-Hen A. 2005, *Influence function and correspondence analysis*, «Journal of Statistical Planning and Inference», 134, 1, 26-35.

Rao C.R. 1995, *A review of canonical coordinates and an alternative to correspondence analysis*, «Qüestiió», 19, 23-63.

Shennan S. 1997, *Quantifying Archaeology*, Edinburgh, Edinburg University Press.

Siegmund F. 2014, *Tutorial for archaeologists: How to perform a correspondence analysis - A practitioners guide to success and reliability* (http://www.storia-culture-civilta.unibo.it/it/risorse/files/regolamento/eventi/vi-seminario-ornamenta-materiale-didattico/tutorial; last access 07/06/14).

ABSTRACT

This paper compares the method of Correspondence Analysis (CA) for finding patterns in archaeological sites by artifacts abundance data, with a robust variant, named Taxicab Correspondence Analysis (TCA). We show that this comparison is useful, especially for sparse tables with outliers. We identify three kinds of outliers. Three well-known datasets are reanalyzed.