Journal of Archaeological Science 40 (2013) 4744–4753

Contents lists available at SciVerse ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

Applying bootstrapped Correspondence Analysis to archaeological data

Kris Lockyear*

Institute of Archaeology, University College London, 31–34 Gordon Square, London WC1H 0PY, UK

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
16 August 2012
Accepted 18 August 2012

Keywords:
Correspondence Analysis
Bootstrapping
Stability
Multivariate analysis

ABSTRACT

This paper examines the usefulness of bootstrapping in Correspondence Analysis when applied to archaeological data. By simulating and displaying possible variation within the data sets, bootstrapping provides us with a means to assess the stability of our CA maps and influences the interpretations we can place upon them. Five real data sets are examined and the results discussed. The paper concludes that bootstrapping is a useful and powerful way of examining the results of CA and should be employed on a regular basis.

1. Introduction

Correspondence Analysis (CA) is a technique which allows the investigation of tables of non-negative integer data via 'maps' (scattergrams where x and y are plotted to the same scale) and associated diagnostic statistics (Greenacre, 1984). One aspect of CA is that "it is a purely deterministic, algebraic technique, so that there is little indication of the strength, or otherwise, of any apparent relationship" (Ringrose, 1992, p. 615). One method which enables us to examine the stability of the maps from CA, and hence the strength of the relationships between objects or variables, is bootstrapped CA (Greenacre, 2007, pp. 193–7). Despite the desirability of examining the stability of maps (Lockyear, 2007, p. 178) and the early description of the technique within archaeology (Ringrose, 1988, 1992) it appears not to have seen much application. Indeed, Baxter (1994, 2003), was only able to cite the papers by Ringrose. A rare exception is the recent paper by Peeples and Schachner (2012). It is likely that one factor which has restricted the adoption of the technique was the lack of easily obtained and user friendly software. Freely available code is now, however, available for the R statistical package (R Development Core Team, 2012).[1] The aim of this paper is to examine the usefulness, or otherwise, of bootstrapped CA by applying the technique to five archaeological data sets. The relationship between sample size, diagnostic statistics and the results of the bootstrapping is also examined.

2. Correspondence Analysis and bootstrapping

The mathematical basis of CA has been described many times elsewhere (e.g., Greenacre, 2007) and need not be repeated here. The output from CA consists of various diagnostic statistics as well as scores for the objects and variables in the analysis which can be plotted as maps. For the sake of clarity in the following discussion the various diagnostics are briefly explained. The inertia of a contingency table is the χ^2 value for that table divided by the total number of items included, n. The inertia is decomposed in a variety of ways. Each axis derived from the analysis also has an inertia. The sum of the inertias from all the derived axes will equal the total inertia. From this, the percentage variation "accounted for" or "explained" by each axis can be calculated. There then follows two tables of information, one for the objects and one for the variables in the analysis. An example is given in Table 1 which shows the output for the variables from the analysis of Romano-British site assemblages discussed below (Section 3.1). The mass (mas) of each variable/object is simply its abundance in the analysis expressed as a permill. Similarly, the inertia (inr) is the permill of the total inertia "accounted for" by that variable/object. The scores used for plotting

* Tel.: +44 20 7679 4568; fax: +44 20 7383 2572.
E-mail address: noviodunum@hotmail.com.

[1] The code for bootstrapping given by Greenacre (2007, pp. 250–2) is available from http://www.carme-n.org. This code requires the use of the R package ca (Nenadic and Greenacre, 2007). The code for the technique outlined by Ringrose (2012) used in this paper is available via email. R is readily available from http://www.r-project.org. A more general guide to CA in archaeology using R has recently been published (Baxter and Cool, 2010).

0305-4403/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jas.2012.08.035

Table 1
Decomposition of inertia for the variables in the analysis presented in Section 3.1.

Period	qual	mas	inr	$k = 1$	cor	ctr	$k = 2$	cor	ctr
I	620	18	46.82	558	444	58	−351	176	43
II	582	24	78.59	419	197	44	−585	385	157
III	594	17	53.62	518	311	46	−494	283	78
IV	777	41	57.4	513	667	109	−209	110	33
V	652	30	45.4	518	638	82	−76	14	3
VI	607	28	31.59	429	604	53	−26	2	0
VII	532	32	27.58	364	531	43	17	1	0
VIII	441	25	34.69	401	431	41	61	10	2
IX	331	14	28.4	406	306	24	117	25	4
X	374	29	30.22	305	324	27	119	50	8
XI	356	19	27.43	354	307	24	141	49	7
XII	266	21	33.05	328	242	23	102	24	4
XIII	114	111	27.14	−6	1	0	86	113	15
XIV	87	103	38.91	−65	41	4	69	47	9
XV	282	35	31.87	−35	5	0	264	278	46
XVI	394	56	26.29	−70	38	3	214	356	48
XVII	525	146	54.31	−172	295	44	152	230	63
XVIII	489	85	55.56	−287	469	72	59	20	6
XIX	416	89	62.42	−274	402	68	53	15	5
XX	183	15	32.26	−297	147	13	−146	35	6
XXI	904	61	186.74	−595	424	222	−634	481	464

the results are given in the columns $k = 1$, $k = 2$,…, $k = n$. The absolute contribution (cor) is the contribution of that variable/axis or object/axis to the total inertia. The first column (or columns) represents the quality (qual) of the representation of that variable/object on the map, and is the sum of the two absolute contributions for the relevant axes. The relative contribution (ctr) is the permill of the inertia explained by that variable/object for that axis. These last three columns are repeated for as many axes as the analyst wishes to examine up to $k - 1$ where k is the lesser of the total number of objects or variables.

Figs. 1 and 2 are the maps from this analysis (Lockyear, 2000). The variables in this analysis are the twenty-one issue periods for Roman coinage defined by Reece (1987), the objects are 136 assemblages of coins from Roman sites in Britain. The map of the periods (Fig. 1) shows a distinct curve from early to late periods with a reasonable ordering of the periods. Looking at Table 1 we can see that period IX is the rarest in the data set as it has the lowest mass. The total inertias, however, are relatively evenly spread with period XXI contributing the most, but no variable

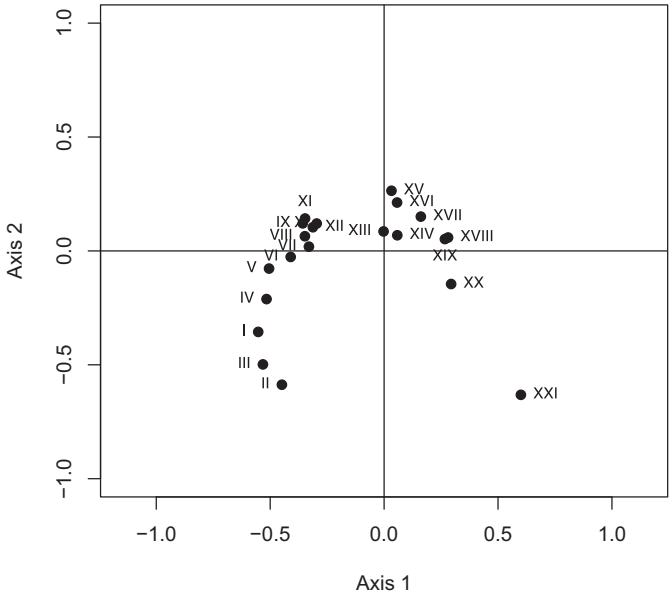

Fig. 1. Object map from CA of site finds from Britannia.

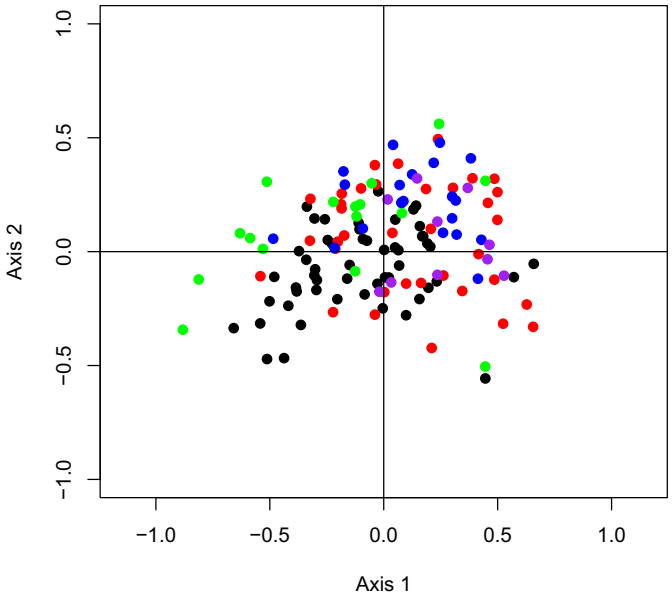

Fig. 2. Variable map from CA of site finds from Britannia. Black: *Civitas* capitals; red: rural sites; blue: villas; green: military; purple: temples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

having a particularly low contribution. Looking at the quality column we can see that period XIV is the least well represented on the map of axes 1 *v.* 2. Looking down the column for the relative contributions (ctr) for the first axis we can note that they seem relatively high for the early and late periods but low for periods XIII to XVI. The second axis is slightly more complicated with periods I–IV and XX–XXI having a high negative relative contribution and periods XV–XVII having a high positive contribution. The combination of these two axes creates a pattern which is quite common when there is a marked gradient in the data, and is known as a 'horseshoe curve' or the Guttman effect. This gradient is often related to time as here, but is potentially linked to space or social status. By examining the decompositions of inertia in this way, problematic objects or variables can be identified, and an assessment made as to how well a particular item is represented on the CA maps. Unfortunately, many users of the technique do not examine these data, relying on the interpretation of the maps alone. The map of the objects (Fig. 2) is much more confused, but careful use of symbols (as used in Lockyear, 2000) or colour can reveal patterns related to site types.

How are the results of a CA to be judged? For many users, analyses which exhibit patterns which are easily interpretable in archaeological terms, or meet the analyst's expectations, are considered 'successful.' Although formal testing of the first axis is relatively easy (Greenacre, 2007, pp. 198–200), testing of the lower order axes is more difficult. Additionally, the testing procedure is conservative, and does not utilise additional information that may be known about the data set such as the location or date of the samples.

Having derived the CA map, the analyst should then offer an interpretation of the pattern presented. In the example given in Table 1 the first axis can be interpreted as showing periods I–XII *v.* periods XVIII–XXI, and the second axis as contrasting presence of period XV–XVII coins *v.* period I–IV and/or period XX–XXI coins. Although the decompositions of inertia allow us to identify well-represented items, or to discount poorly-represented ones, we have no clear way of knowing how much weight to give the position of each point on our map, and therefore how much credence to give the inter-point distances. As noted above, one way of assessing