Developing Reproducible Methods for Defining and Evaluating Ceramic Compositional Groups Derived from NAA: A Jamaica Example

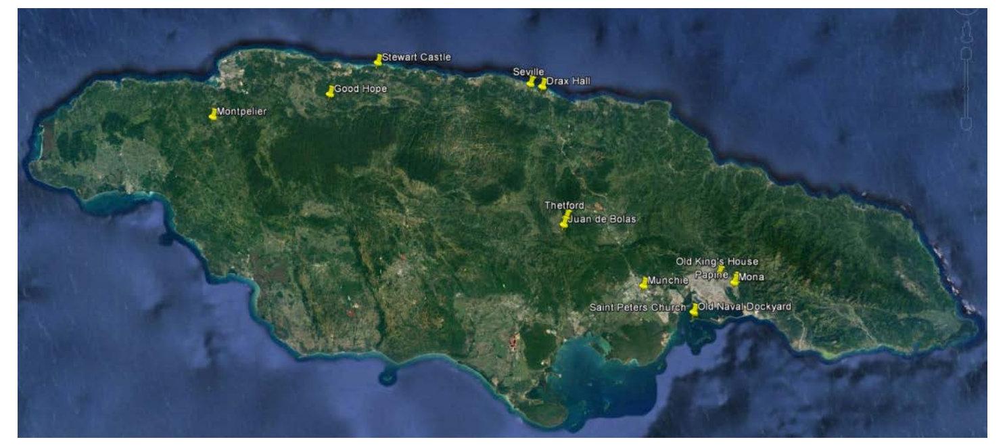
Fraser D. Neiman[1], Lindsay Bloch[2], Jillian Galle[3], Jeffrey Ferguson [4]
1) Monticello, 2) Florida Museum of Natural History, 3) DAACS, Monticello, 4) Archaeometry Laboratory at the University of Missouri Research Reactor

1. Introduction

Earthenware ceramics offer archaeologists a means to investigate the evolution of slave societies in the early-modern Atlantic. A major research focus is the spatial scale of their production and distribution, ranging from household production for consumption within a single local community to specialist production and distribution across many communities via an integrated colony-wide market.

Above: Market, Falmouth Jamaica, c. 1844.

We build on pioneering work by Hauser *et al.* (2008), who identified two major compositional groups in a sample of 51 sherds, presumed to be locally made, from sites in Jamaica. Sherds from both groups were found at sites on both coasts, suggesting an island-wide system of manufacturing and distribution. Our sample comprises 448 sherds, including Hauser *et al.*'s 51. The new samples were collected by DAACS researchers and their partners in the DAACS Research Consortium. DAACS sampled *all unidentified* coarse earthenwares, so our larger sample may include both Jamaican and British-made ceramics. All data reported here were generated at the Archaeometry Lab at MURR.



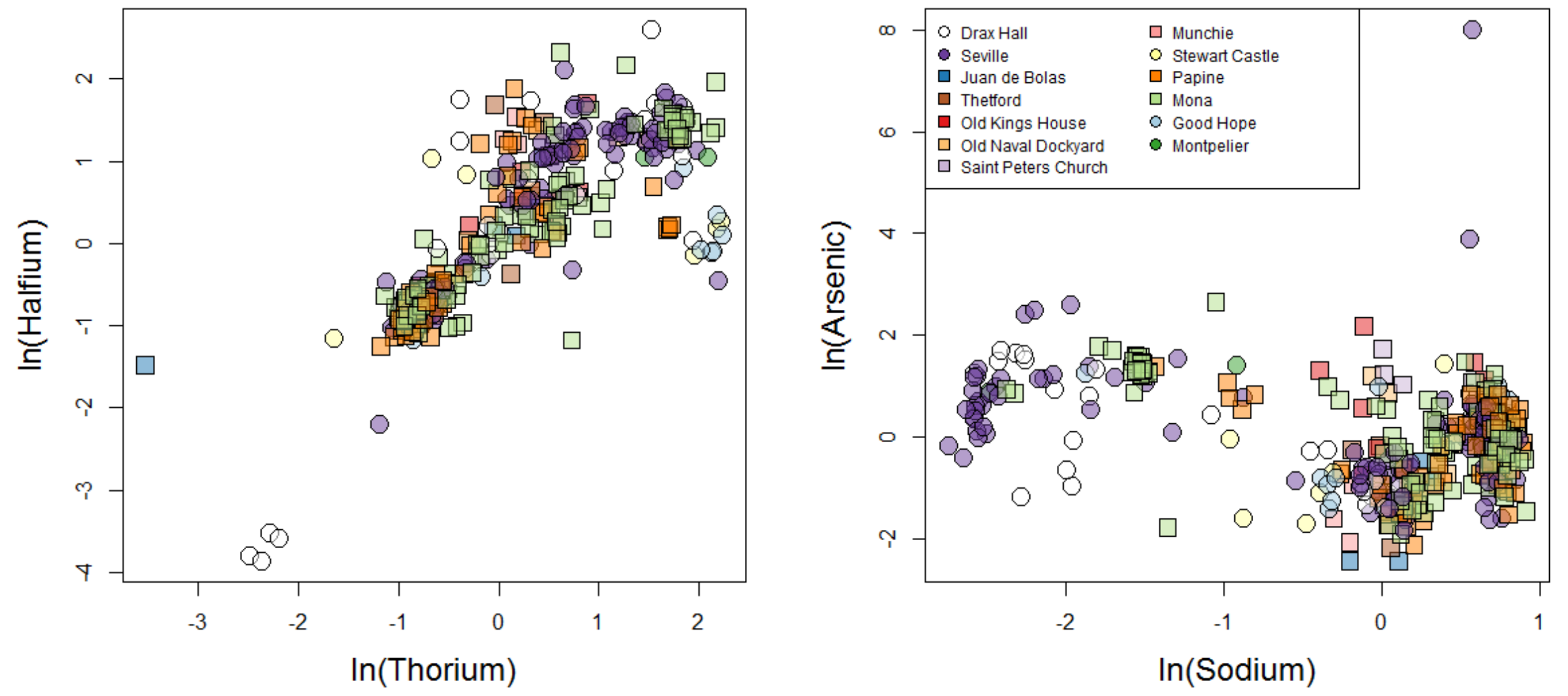
Left: 448 sherds from 13 sites on both the north and south coasts are included in this study.

R GitHub

2. Reproducibility

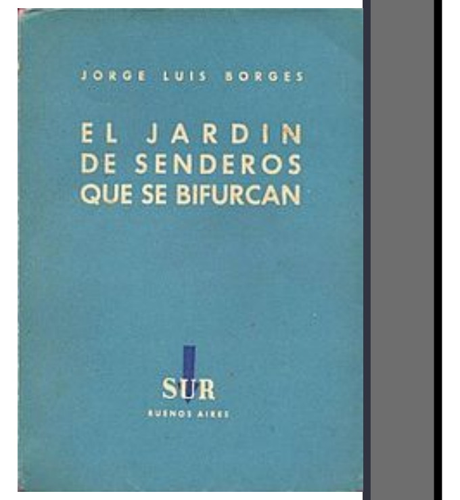
A second goal is to offer a useful, preliminary example of how compositional data analysis can be made more open and reproducible (*e.g.*, Marwick *et al.* 2017). **You can find our data and R code here:** https://github.com/fneiman/jamaica-naa
Compositional data analysis is a two-phase process. An exploratory phase aims to identify trial compositional groups. A confirmatory phase aims to evaluate them. Our knowledge of the past can only benefit by making both phases reproducible.

The identification of trial compositional groups is often the least reproducible step. A modal approach is eyeballing groups in a few of the ((p*p)-p)/2 possible pairs of scatter plots. Each plot is a step in the "*garden of forking paths*" (Gelman and Loken 2014). Groups are contingent on a small subset of the data, with the multiplicity of plots guaranteeing a few plots with good-looking groups can be found, even in random data.



Above: Two scatterplots (out of 496!!) . How many good-looking groups? Are good looks a result of lots of looking? Sites are coded by fill colors. *Circles* represent north coast sites and *squares* are south coast sites.

Identifying trial groups using a clustering algorithm would seem to be a more transparent and reproducible approach, although it is not without its own drawbacks, including its own set of forking paths! We explore this approach here.



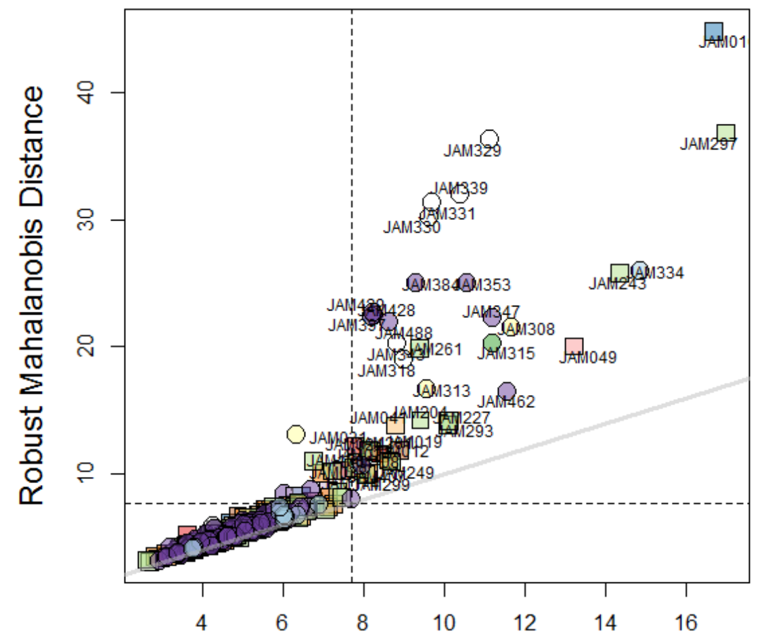
3. Initial Steps

MURR captured data on 33 elements. After log transforming the data, we imputed 10 missing values for U and Sr, using the EM algorithm (Honaker *et al.* 2012). Because over half its values were missing, we deleted Ni. We converted all variable values to *z* scores.

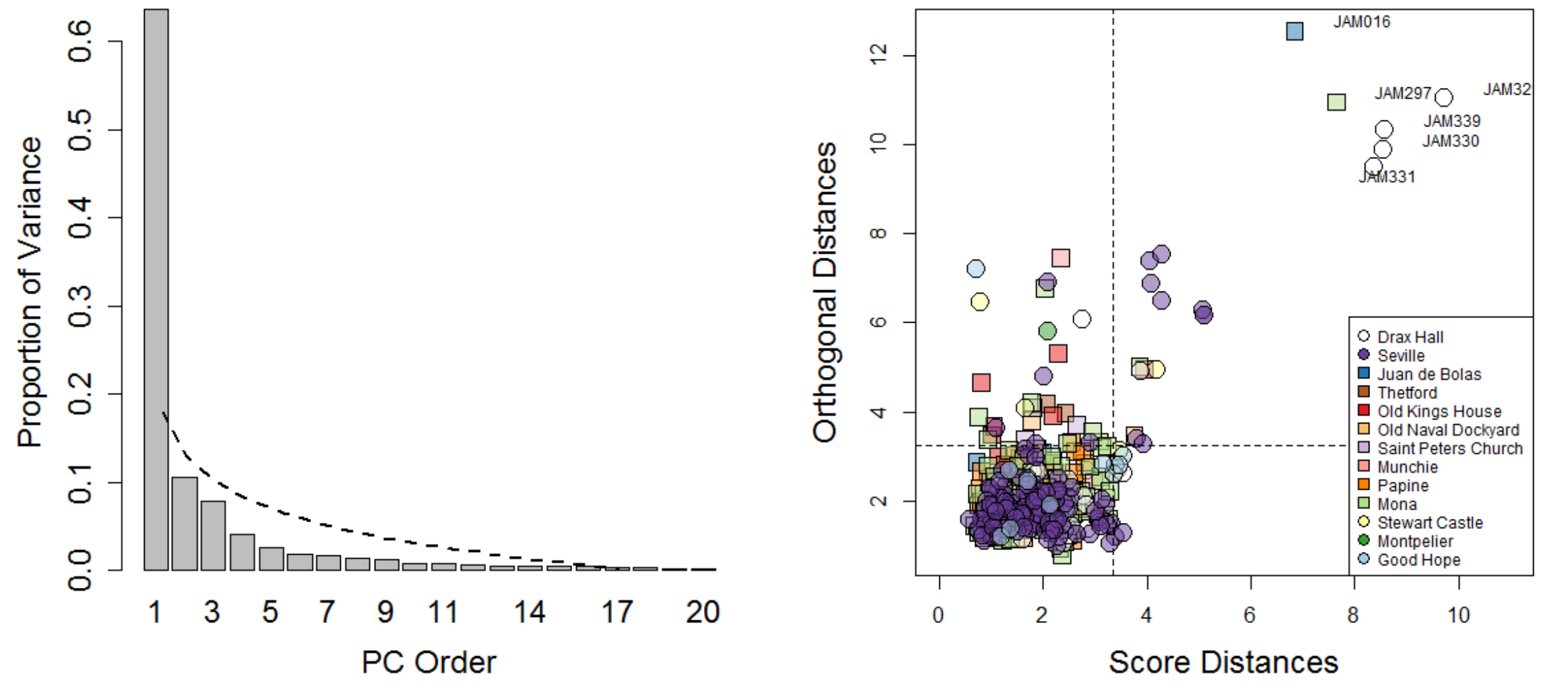
4. Outliers

Clustering results for data that contain multivariate outliers typically emphasize the contrast between the outliers and the bulk of the data, obscuring the structure of the latter. To mitigate this, we identify and remove outliers from the data. We used two approaches to detect outliers.

Right: Outlier detection with robust and classical Mahalanobis distances.



In the first, we plot classical Mahalanobis distances (MD) of our samples from the center of the data against a robust version of MD, computed using the minimum covariance determinant (MCD) method (Filmozer *et al.* 2005) .



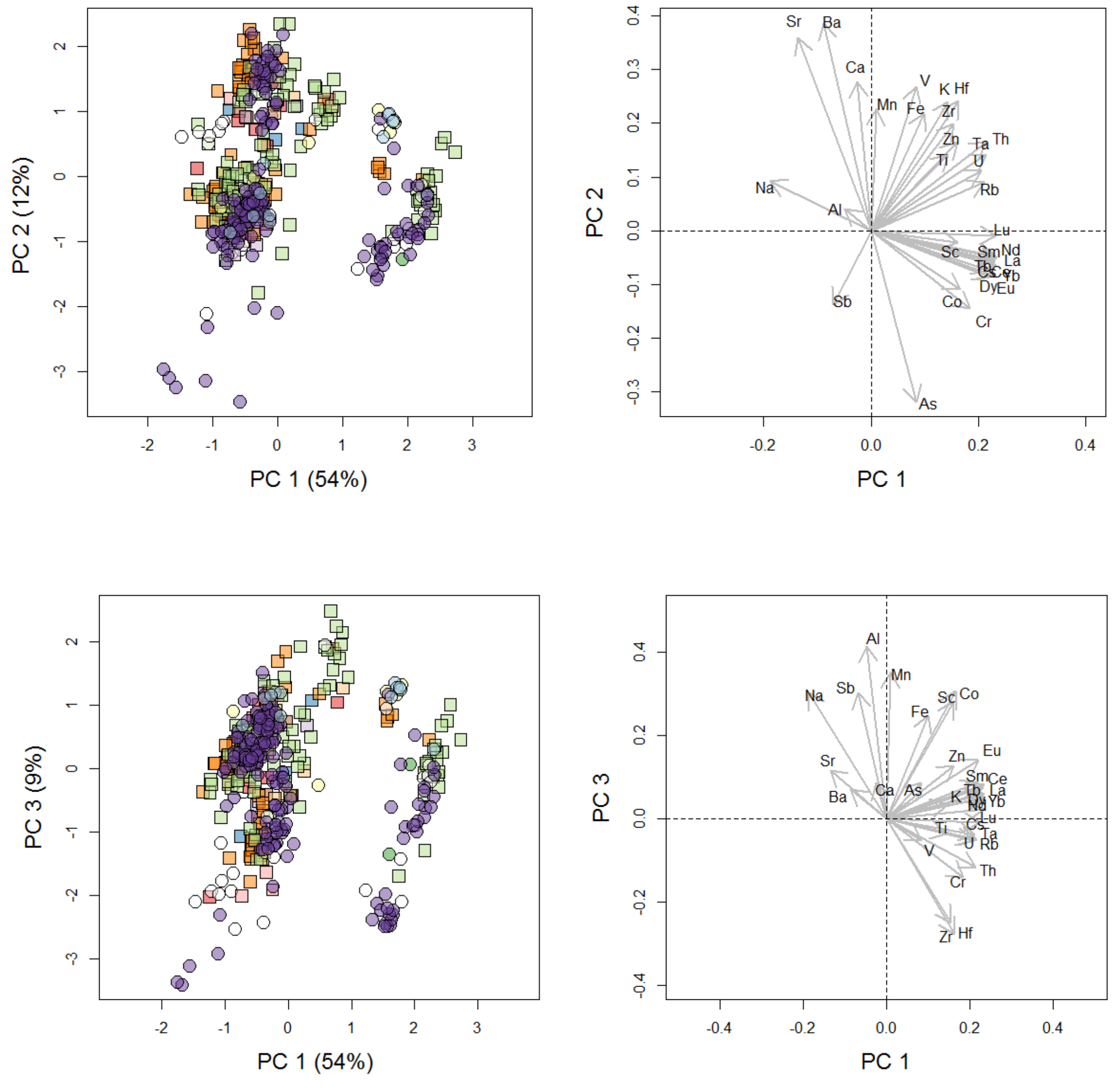
Left: Scree plot of variance accounted for by Robust PCs**. Right:** Score distances and orthogonal distances from 4 PC solution.
.

In the second, we employ Hubert's ROBPCA algorithm (Hubert *et al.* 2005). ROBPCA uses the MCD method to estimate iteratively a specified number of PCs. We chose 4, based on a scree plot. The PCs yield two measure of outlyingness: *Score distances* measure distance from the centroid in the PC space. *Orthogonal distances* measure distance from the centroid outside it.

Both plots identify 6 outliers which we exclude from the analysis. Four of them are from Drax Hall, on the north coast.

5. Compositional Variation among Sites

Principal components analysis (PCA) offers a way to visualize the extent to which sherds from the same sites have similar compositions, as we might expect with household production and community distribution.



Above: Classical PCA of the data, with 6 outliers removed. Sites are coded by fill colors. *Circles* represent north coast sites and *squares* are south coast sites.

Plots of sherds on the first three PCs hint at distinct compositional clusters or groups. But samples from sites on the north and south coasts are found in each of the major clusters. This result favors the idea that most of the pots from which these sherds were derived were produced by specialists and moved in island-wide markets.

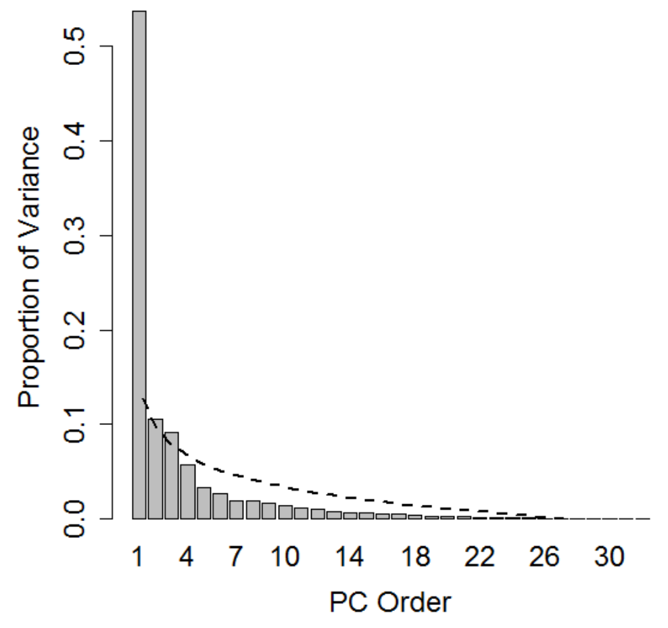
6. Identifying Compositional Groups: PAM

How might we more precisely define the compositional clusters or groups? One option is to eyeball them from the PCA plots. Another is cluster analysis. Here we use PAM.

PAM (Partitioning Around Medoids) is a robust version of the k-means algorithm so beloved by quantitatively savvy archaeologists. It uses typical observations ("medoids") to define clusters rather than means

For a specified value of *k* (the number of groups), the algorithm finds *k* trial medoids and then assigns observations to clusters based on which medoid it is closest to. The process iterates until the medoids stabilize. PAM requires two inputs: a value for *k* and the variables on which the clusters are to be based (Kaufman and Rouseeuw 1990).

The variables we chose were the standardized scores of the observations (0 mean, unit variance) on the first four classical PCs of the standardized data matrix, with outliers removed. These account for 80% of the variance in the data.

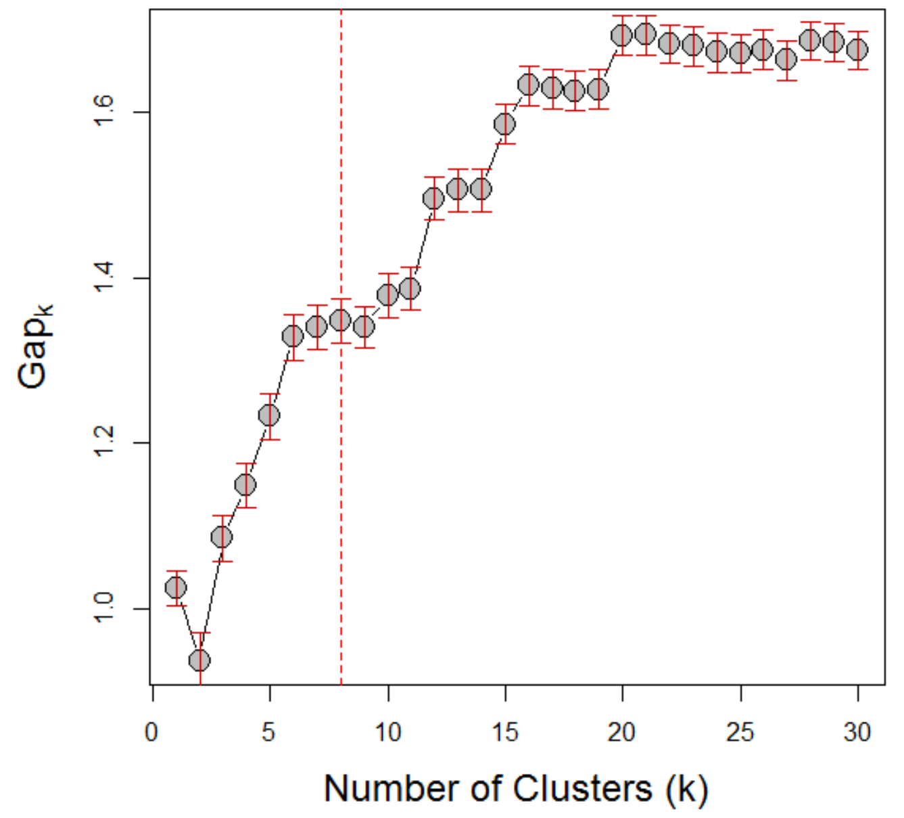


Above: Scree plot of classical PCA of the data. The dotted line is the expectation under the random broken-stick model. We use the first 4 PCs for PAM, hoping they capture the signal in the data, while the high-order PCs capture noise.

Using standardized PC scores means that the Euclidean distances used in the PAM algorithm are Mahalanobis distances. In theory, using the leading PCs has two additional effects: 1) reducing noise in the data and 2) reducing noise effects of high dimensionality on the clustering algorithm (Ronan *et al.* 2016).

For guidance on the choice of *k* (the number of clusters) we used the gap statistic (Tibshirani *et al.* 2001). The gap statistic compares the average actual within-cluster dispersion for the *k* clusters (W_k) to the expected within -cluster dispersion $E^*(W_k)$ under a null model in which the *n* observations are uniformly distributed. So $Gap_k = \log(E^*(W_k)) - \log(W_k)$. Local maxima will occur for *k* values at which within-cluster dispersion is much *lower* than expected, producing a larger $Gap_{k.}$

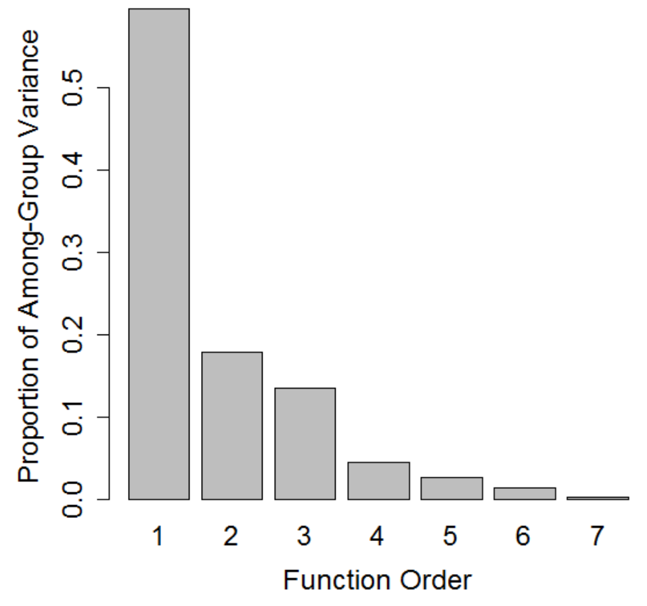
Right: As the value of *k* increases, a local maximum in the gap statistic indicates lower within-cluster dispersion than expected under a null model in which variables are uniformly distributed across the ranges observed data. Discarding an uninformative local maximum at *k=1*, we find a second at *k=8*.

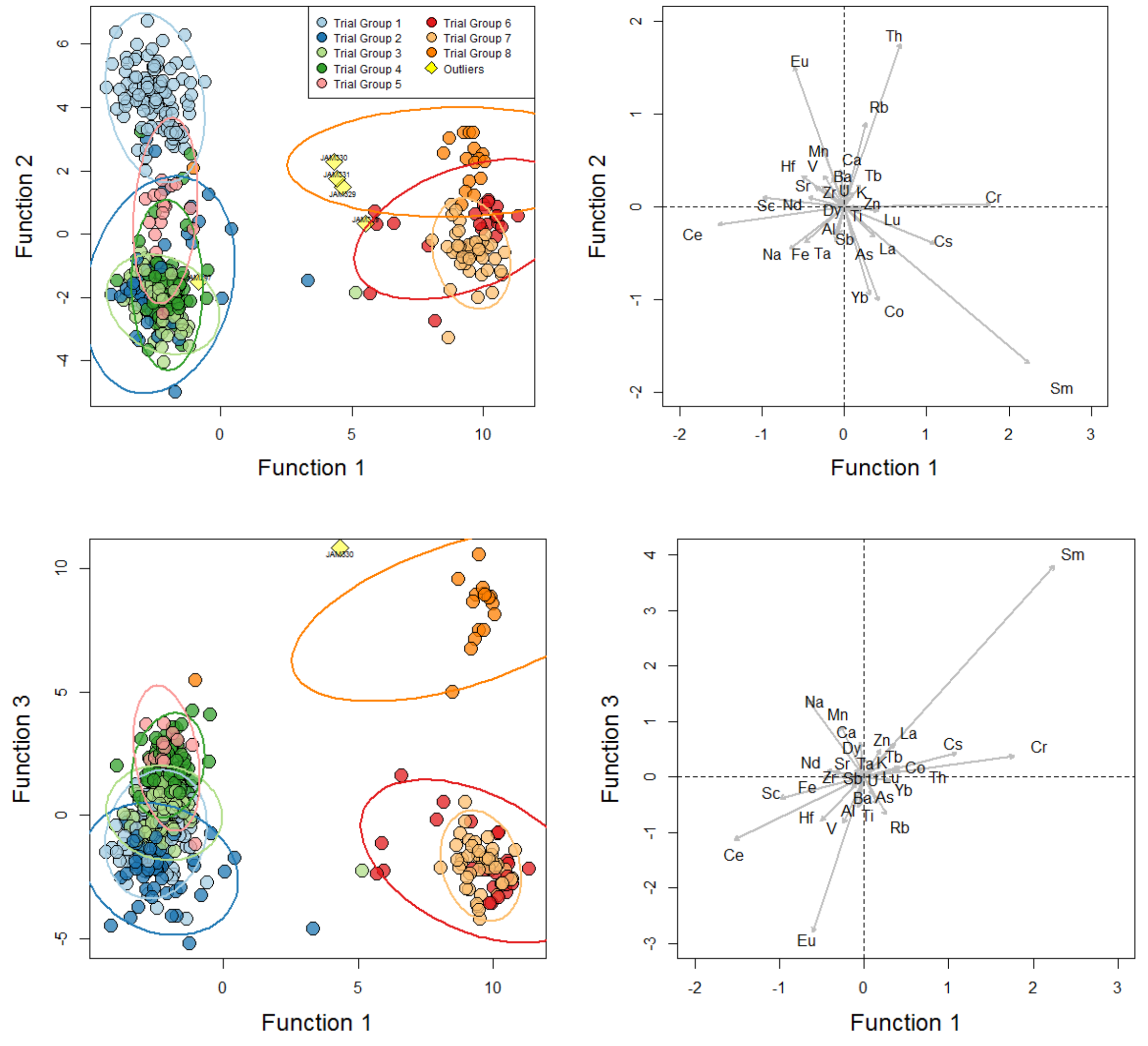


7. Predicting Group Membership: LDA

Just how distinctive are the *k=8* clusters from PAM? Linear discriminant analysis offers a visual answer in the form of plots of sherd scores on perpendicular axes designed to maximize among-group variance.

Right: Scree plot of proportion of among-group variation accounted for by successive discriminant functions





Above: Plots of the 8 trial groups (TG1-TG8) identified by PAM on the first 3 discriminant functions. Points are color coded by group. Yellow points are the outliers that were removed from the analysis but then scored on the functions. Function 1 separates Jamaican groups from English ones.

Predicting trial group (TG) membership for each sherd, based on LDA estimated without that sherd, yielded mean accuracy of 89%.

Trial Group	Juan de Bolas	Thetford	Saint Peters Church	Munchie	Old Kings House	Old Naval Dockyard	Mona	Papine	Montpelier	Seville	Stewart Castle	Drax Hall	Good Hope	Total
TG-1	3	5	2	3	4	0	18	19	0	35	1	9	0	99
TG-2	1	0	1	0	2	2	25	12	0	8	2	4	1	58
TG-3	0	0	3	0	4	2	28	25	0	38	0	4	3	107
TG-4	0	0	0	0	2	2	26	11	0	32	0	0	2	75
TG-5	1	1	0	0	0	0	11	3	0	0	0	1	0	17
TG-6	0	0	0	0	0	0	20	2	1	15	0	7	1	46
TG-7	0	0	0	0	0	0	1	4	0	3	3	1	5	17
TG-8	0	0	0	0	0	0	1	0	1	18	0	3	0	23
Total	5	6	6	3	12	6	130	76	2	149	6	29	12	442

Discriminant function 1 separates Jamaica-made sherds (TG1-TG5) from English ones (TG6-TG8), as shown in the companion poster. Hauser *et al.*'s two groups roughly correspond to TG1 and TG3. TG1 includes 3 sherds from pots made by Jamaican potter "Munchie" in Spanish Town. All 5 Jamaican groups are found on both the north and south coasts, as are all 3 English groups. Further work is needed to ensure TG identification is robust and to explore relationships between TGs and formal variation. We tentatively conclude that Jamaican ceramics were specialist-produced for sale in an island-wide market where they successfully completed with British imports.

References

Gelman, A. and E. Loken (2014) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
Hauser, Mark W., C. Descantes, M.D. Glascock (2008) Locating enslaved craft production: chemical analysis of eighteenth century Jamaican pottery. *Journal of Caribbean Archaeology* 2, 123-148.
Honaker, James, Gary King, and Matthew Blackwell (2012) AMELIA II: A Program for Missing Data, Version 1.6.2. https://r.iq.harvard.edu/docs/amelia/amelia.pdf
Hubert,Mia., PJ Rousseeuw, K Vanden Branden (2005) ROBPCA: a new approach to robust principle component analysis *Technometrics* 47 (1), 64-79.
Marwick, Ben, *et al.* (2017) Open science in archaeology. *SAA Archaeological Record*, 17(4), 8-14.
Ronan, Tom, Zhijie Qi, Kristen M. Naegle (2016) Avoiding common pitfalls when clustering biological data. Science Signalling 9(423). doi:10.1126/scisignal.aad1932
Rousseeuw, Peter.J. and van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association* 85, 633-639.
Tibshirani, Robert , Walther, G. and Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63: 411-423. doi:10.1111/1467-9868.00293

Th: Jefferson MONTICELLO
DAACS www.daacs.org
FLORIDA MUSEUM™
Archaeometry Laboratory University of Missouri Research Reactor