Archeologia e Calcolatori
27, 2016, 7-25

AGGLOMERATIVE CLUSTERING USING COSINE AND JACCARD DISTANCES: A COMPUTATIONAL APPROACH TO ROMAN VESSEL TAXONOMY

1. Introduction

When it comes to the comparison of ceramic finds from different archaeological projects, categorization can be problematic. Researchers incorporating older data into current questions are faced with the task of comparing finds which may have been classified or typed according to different criteria. After all, different methods, even applied to the same body of evidence, potentially result in different conclusions. The development of new classes and typologies over time also means that older data will inevitably have to be related to the current system in place through a series of concordances, which can take considerable effort to produce, especially for large collections of material.

In this paper, I offer a computational method to the categorization of vessels finds as an efficient way to deal with the problem of discrepant systems of classes and types. Each vessel artifact (or artifact-type) is described as a set of its component semantic attributes. The semantic distance between these sets can then be measured, and proximate sets clustered into new categories. The result is a framework of algorithmically derived categories to which vessel-types can be assigned, thereby ensuring standardized cross-project comparison.

This method is inspired by Latent Semantic Analysis (Deerwester *et al*. 1990), and is based on the creation of semantic sets of vessel-types, which are then compared to associate similar sets to one another, using two different ways to assess their associative distance. The first metric is the Jaccard index, which directly measures the degree of similarity between two sets. The second is the cosine similarity, which measures the angular distance between two different vectors, in this case, produced from the vessel sets. These two methods are used as a check on one another, to assess the integrity of the synthetic categories found by hierarchical clustering with Ward's minimum variance method. To summarize, this paper presents a method to effectively and flexibly relate different vessel classes and types to one another, creating a baseline for inter-project comparison of finds that are categorized under different rubrics.

By way of an example, this algorithm is applied to a small dataset containing 1,492 entries of vessel-types, drawn from various excavations in Italy, focusing on but not limited to the period of the Roman Republic. In general, the algorithm successfully assigned vessels to new groups, and

final adjustments to the cluster assignment could be undertaken manually. In this regard, visual inspection of hierarchical dendrograms is a useful step in obtaining a sound re-categorization. In addition, post-procedural analysis of categorical ambiguity was also performed to assess ambiguity in cluster assignment.

This article presents the pilot dataset and algorithm of the open-source project synthkat, an initiative aimed at computational means to construct synthetic categories of archaeological ceramic and glass vessels, written in Python (version 3.4.3). The project files and data used to illustrate this method are available at https://github.com/scollinselliott/synthkat/. Initial data processing was expedited by way of a pilot script in Python that processed raw .csv files as input, and then produced an output of documents in .json, .csv, and .txt format. The synthkat script was based on the topic-modeling package gensim (Řehůřek, Sojka 2010), as well as numpy and scipy. Plots were made using the package matplotlib for Python (Hunter 2007). Dendrograms were plotted with the assistance of code written by Jörn Hees[1]. Information with links to these libraries may be found on the project website.

2. Semantics and vessel taxonomy

The study of ceramic vessels of Roman Italy abounds in classifications and typologies, most of which have developed organically from the nineteenth century onward (Gandolfi 2005). While the boundaries between certain classes are fairly clear, such as lamps and transport amphorae, those between other classes can be less clear. Common ware(s), for example, represents one of the most difficult classes to treat, a category manifested in large part in negative terms and vaguely associated with cookware (Bats 1996; Cortese 2005; Santoro Bianchi 2005). Sometimes, unslipped wares are divided by the composition of the clay, inclusions, and other mineral or chemical factors. At other times, the distinction is functional: the first division is between vessels which were used in the consumption or storage of food (*ceramica da mensa* or *da dispensa*) and those which were used in its cooking (*ceramica da fuoco* or *da cucina*). Mortaria and other large storage containers in *opus doliare* may at times be distinguished from other common ware. A concordance is necessary to relate these different vessel classes with one another.

After all, there will always be slight differences in the organization and arrangement of classes that require steps to be taken to ensure concordance. For example, in the case of comparing en masse the vessels finds from two excavations published five years apart – those at Fiesole on the Via Marini (de

[1] https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/.

Marinis 1990) and those at Settefinestre (Ricci 1985) – finds would have to be re-classified: classes such as *ceramica acroma depurata* and *ceramica acroma grezza* from Fiesole are congruent with the *ceramica comune* of Settefinestre (Table 1). The task of translation is even more obvious when working across languages, considering the comparison of "kitchen ware," "coarse ware," or "domestic ware" (Dyson 1976), with *ceramica da cucina*, *da fuoco*, *grezza*, *da mensa*, or *da dispensa*. In sharp contrast to the ordinary and quotidian use of these vessels in their actual, systemic context, the semantics of their archaeological definition seem profound.

Moreover, tracing the progress of research will show that classes and types come into and fall out of use over time. Grey-gloss ceramics, for example, have only been recognized as a class separate from black gloss since 1980 (Giardino 1980; Vittoria 2011). Black-gloss Campanian B, as originally articulated by Nino Lamboglia, is an example of subclass in crisis. At first it was thought to have been produced only in Etruria, but currently it appears that the production of Campana B ceramics was widespread throughout Italy and the western Mediterranean (Lamboglia 1952, 140-142; Morel 1998a, 237-239; Morel 1998b, 12-15; Cibecchini, Principal 2004; Olcese 2006, 528). "Pre-sigillata" is a term now deprecated, though the present and future cannot erase its use in past publications (Ettlinger *et al*. 1990, 4). Even though the way in which older materials have been classified and typed might be "outdated," they should not be disregarded on account of obsolescence. Archaeology is, after all, the recovery of obsolete materials, and it would be both ironic and sad to discard the results of prior generations.

Thus, anyone who wishes to compare finds from one project with another has to become a sort of translator, who would have to develop a series of concordances, catch-as-catch-can, that relate different classes and types to one another (Table 1). This approach is business as usual, in re-identifying and re-typing archaeological finds, of relating similar to similar. However, the larger the dataset, the more prohibitive the rewards of this exercise would be. This is where the use of a computational method becomes helpful. But in order to deal computationally with ceramic vessel taxonomy, it is necessary to put that meaning in a language a computer can understand: mathematics.

Setting aside the deep but venerable nuances of the terms "class", "type," or "group" (Steward 1954; Shepard 1956; Rice 1987, 274-288; Read 2007; Reynolds 2008, 82), mathematically, the basic concept at work is one of sets: what classification and typology have in common is the essential act of grouping elements (here, the vessels or vessel fragments) into sets (a class or type), on the basis of a criterion. That criterion has been, is, and will be subject to variation, predicated now on physical properties or characteristics

Fiesole: major ceramic classes (de Marinis 1990).	
Ceramica da mensa e da dispensa	ceramica a vernice nera, ceramica a vernice rossa ceramica grigia ceramica d'impasto chiaro granuloso ceramica a pareti sottili terra sigillata italica terra sigillata tardo-italica terra sigillata africana terra sigillata chiara italica ceramica a vernice rossa tarda ceramica dipinta tarda ceramica tardoromana a superficie lisciata ceramica acroma depurata
Ceramica da cucina	ceramica a vernice rossa interna, ceramica africana da cucina, ceramica acroma grezza
Mortaio	mortaio
Grande contenitore	bacini dolia
Contenitore da trasporto	anfora
Lucerna	lucerna
Settefinestre: major ceramic classes (Ricci 1985).	
Contenitori da cantina e da trasporto	anfora
Suppellettile da cucina e da dispensa	opus doliare coperchio, ceramica comune, ceramica a vernice rossa interna, ceramica africana da cucina
Suppellettile da mensa	ceramica comune ceramica a vernice nera sigillata italica sigillata tardo-italica decorata sigillata sud-gallica decorata ceramica verniciata con sovradipinture a spugna ceramica africana di produzione A ceramica africana di produzione A/D ceramica africana di produzione C ceramica africana di produzione D sigillata orientale di produzione A sigillata orientale di produzione B ceramica invetriata ceramica a pareti sottili ceramica tipo aco
Suppellettile da illuminazione	lucerna
Recipienti per il lavaggio e incensari	ceramica comune
Strumenti per la preparazione di sostanze	mortaio

Table 1 – Different taxonomic systems for ceramic vessel assemblages from Fiesole (de Marinis 1990) and Settefinestre (Ricci 1985).

of the clay or slip, now on morphology, as derived either from traditional or common-sense traditional categories or aesthetic-rational aspects of the vessel design. In the case of Roman ceramics in Italy, at least, classification has tended to describe the technical-functional divisions among vessel manufactures, while typology has tended to refer to morphological-functional variations within classes. Yet, any process of classification or typology is, in effect, a reduction of the object down to its constituent semantic components,

Element	Description	Element	Description
$a_{f,0}$	VN.P3	$a_{w,0}$	ceramica a vernice nera
$a_{f,1}$	Morel 2255	$a_{w,1}$	vernice nera
$a_{f,2}$	Lamboglia 5	$a_{w,2}$	black gloss
$a_{f,3}$	piatto	$a_{w,3}$	tableware
$a_{f,4}$	plat	$a_{w,4}$	fineware
...	...	...	...

Table 2 – Example of semantic connotations of the type VN.P3, from Fiesole (DE MARINIS 1990, 105), expressed as a set attributes related to form (*f*) and ware (*w*).

to compare it with another object's semiotic system, the notion of the type or class to which it is attributed.

Thus, typing or classing vessels is nothing more than establishing links between the object itself and an ideal class or type, itself formed by comparisons with other artifacts. Breaking down this process mathematically in terms of set theory, the process of identifying a sherd representing a vessel-type as belonging to a particular class or type can be construed as the creation of a set A with any number n of attributes a_i. For example, from excavations at Fiesole, Via Marini, the type VN.P3 is related to the Morel 2255/Lamboglia 5 black-gloss plate (DE MARINIS 1990, 105). Expanding this record's connotations across different languages, one can derive an explicit set of attribute characteristics (Table 2). Increasing the connotations of the object renders it all the more relatable to other types or sets which share those attributes. This process is basically doing with meaning what chemical analysis does with the raw material of the vessel. Indeed, the chemical composition of the vessel can be construed as another attribute of the vessel itself.

I utilized two domains of semantic sets, one for vessel morphology, or form, and another for ware. The division of these two domains was necessary, as combining the two tended, in practice, to result in confused comparisons: experimentation with the algorithm and the high dimensionality of the dataset (see below) coerced dimensional reduction, which could be accomplished by limiting the use of certain terms to either vessel form or ware. Thus, separate synthetic categories were developed for these two different aspects of the vessel-type.

There are any different number of ways in which the composition of a semantic set has been articulated in past scholarship. It was a central point of discussion in processual archaeology, such as with James Deetz's factemes and formemes, or David Clarke's hierarchical taxonomy of cultural entities down to their constituent attributes (DEETZ 1967; CLARKE 1978; PREUCEL 2006, 101-109; cfr. also RAMAZZOTTI 2010). Unlike Clarke's definition, however,

that attributes are the smallest unit, a «logically irreducible character of two or more states, acting as an independent variable within a specific frame of reference» (Clarke 1978, 156), it must be recognized that attributes cannot by themselves be "logically irreducible". They are bound by the very process of interpretation and identification, which are logically reducible, since attribute identification must be subject to definition and explanation.

In other words, culture does not have a smallest particle. Attributes are not atomic. The frame of reference shifts. Even though the system of culture can be analogized as information systems, as Clarke sees it, it still does not make sense to reduce cultural information – ideas – to the level of a bit (Clarke 1978, 88-91). Even ideas that one could argue are maximally reduced, like a color or a shape, or carination on a vessel wall still have a depth of meaning that is dependent on other factors within a cultural system. No single idea can exist in isolation, just as the definition of an attribute is inextricably bound to the meaning of the entire cultural system, and will always be logically reducible against the backdrop of some context. In this way, attributes that are used to identify the basic components of an artifact can be reducible or referable to other cultural components. Semiotics is an inescapable and self-perpetuating act; the full explicit articulation of the meaning of a vessel, or vessel fragment, *ad absurdum*, will never end.

The costs of the time spent in incorporating hierarchical systems of classification or typology, such as Jean-Paul Morel's magisterial *Céramique campanienne* (Morel 1981), or along the lines of the Chronotype developed by Timothy E. Gregory and Nathan Myers for the Sydney Cyprus Survey Project (Meyer 2003; Moore 2008), proved to be excessive. Hierarchy after all mandates the ordering of the attribute set. For example, just taking VN.P1 (a_0) as a subset of black gloss (a_1) would necessitate the creation of a tuple – a nested ordered pair – expressed as $(a_0,a_1) = \{\{a_0\},\{a_0,a_1\}\}$, within the semantic set. Constructing hierarchies involves the introduction of yet more complex criteria, which, for the purposes of constructing new ceramic categories, would hinder comparison. Keeping each set at the level where all attributes are unnested, $\{a_0,a_1\}$, rather than nested, e.g., $\{\{a_0\},\{a_1\},\{a_0,a_1\},\{\{a_0\},\{a_0,a_1\}\}\}$, lowers the costs of time and energy spent in data entry and computational work necessary to deal with these more complex entities.

3. A computational approach to synthetic classification

Proceeding from the creation of semantic sets for each vessel-type, the next step is to find a means of comparing them to one another and grouping like with like, building categories from the bottom up. Cluster analysis is used ultimately to locate and identify groupings, but there are many different methods available, and not all produced useful results. Initially, I focused on

	PompeiVI.5CE1001	PompeiVI.5CE2255	SanSilvestroCT7	FVMVN.P3	Settefinestre35.2	Settefinestre35.3
Pompei VI.5 CE 1001	1	0	0	0	0	0
Pompei VI.5 CE 2255	0	1	0	0	0	0
San Silvestro CT 7	0	0	1	0	0	0
FVM VN.P3	0	0	0	1	0	0
Settefinestre 35.2	0	0	0	0	1	0
Settefinestre 35.3	0	0	0	0	0	1
Morel 2250	0	0	0	0	1	0
Morel 2255	0	0	0	1	0	0
Morel 2270	0	0	0	0	0	1
Morel 2942	1	0	0	0	0	0
Lamboglia 5	0	0	0	1	0	0
Pompei VI.5 Olla 2c	0	0	1	0	0	0
Luni II Group 35a	0	1	0	0	0	0
coppa	1	0	0	0	0	0
cup	1	0	0	0	0	0
bowl	1	0	0	0	0	0
schale	1	0	0	0	0	0
coupe	1	0	0	0	0	0
olla	0	1	1	0	0	0
piatto	0	0	0	1	1	1
plat	0	0	0	1	1	1
teller	0	0	0	1	1	1
platter	0	0	0	1	1	1
assiette	0	0	0	1	1	1
plate	0	0	0	1	1	1
platte	0	0	0	1	1	1
open	1	0	0	1	1	1
closed	0	1	1	0	0	0
forma aperta	1	0	0	1	1	1
forma chiusa	0	1	1	0	0	0
vernice nera	1	0	0	0	1	1
ceramica a vernice nera	1	0	0	1	1	1
black gloss	1	0	0	1	1	1
ceramica grezza	0	1	0	0	0	0
ceramica depurata	0	1	0	0	0	0
ceramica comune	0	1	1	0	0	0
ceramica comune tirrenica	0	0	1	0	0	0
tableware	1	0	0	0	1	1
fineware	1	0	0	1	1	1
unslipped	0	1	1	0	0	0
slipped	1	0	0	1	1	1
ceramica da mensa	0	0	0	1	0	0
ceramica da dispensa	0	0	0	1	0	0

Table 3 – "Term-document"-style matrix for six vessel entries selected from the pilot dataset of project synthkat, with attributes belonging to ware and form combined.

replicating results analogous to topic identification using Latent Semantic Analysis (LSA), which is used in large-scale comparisons of textual corpora (Deerwester *et al.* 1990). In this regard, the first step consisted of the creation of a term-document matrix, in which each term represented a semantic attribute given to each vessel, and in which each document represented the vessel-type itself (Table 3). Following LSA in a banausic and unreflective way, though, posed certain problems. For example, term-document matrices are typically subjected to term frequency-inverse document frequency weighting (tf-idf), which weights the elements of the matrix proportional to the number of times it appears in each document, so rare terms are given more prominence than they would otherwise receive. Since each attribute appears only once in each semantic set for each vessel type, performing tf-idf would not seem to hold any particular theoretical validity.

Furthermore, singular value decomposition, which lies at the heart of LSA, did not elicit a particularly useful representation of the data (Fig. 1). Multiple attempts were made to select different vector spaces which would be amenable to a density-based clustering algorithm, such as DBSCAN and OPTICS (Ester *et al.* 1996; Ankerst *et al.* 1999; Kriegel *et al.* 2011). Yet, manipulation of the data at the level of the term-document matrix, selection of different vector spaces, and variation in the scale of the epsilon-neighborhood in each of the clustering methods, repeatedly failed to provide groupings that made sense as archaeological categories. These poor results were probably obtained due to the high-dimensional space of a sparse matrix (10,384 of 2,584,144 entries in the term-document matrix generated from the pilot dataset were non-zero), as well as the overall shape of the semantic sets.

Therefore, I sought out alternative approaches that could still serve to measure the distance between the semantic sets and produce archaeologically meaningful clusters, notwithstanding the high dimension of the data matrix. Two methods were assessed that could be used to create a distance matrix, cosine similarity and the Jaccard index. Both methods were used as a check on the integrity of the results obtained from clustering, as each measures distance differently (Table 3).

Cosine similarity is a good measure for high-dimensional spaces, and thus is particularly appealing for the case of large datasets (cosine similarity is also germane to LSA, and is included as a feature in gensim). Taking two vectors in an *n*-dimensional space, the distance between vector *A* and *B* is measured as the cosine of the angle between the two:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

The cosine similarity will have a value of 1 where two vectors have precise the same angle, and a value of 0 when they are orthogonal to one

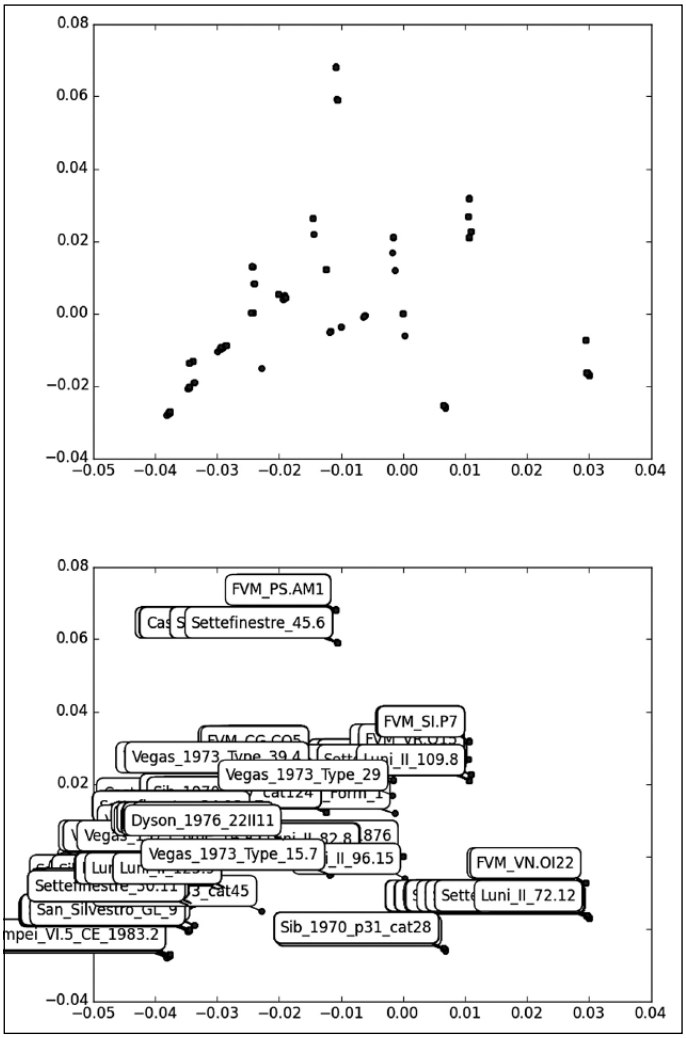

Fig. 1 – Scatter plot of vessel types using the second and third axes resulting from singular value decomposition of the term-document matrix generated from the pilot dataset, using only vessel ware attributes, with and without labels for each record.

another (as well as a value of -1 when opposite, but this will not happen since we are working with non-negative values such as here). Taking for example the six columns in the term-document matrix in Table 3 as six vectors in a 43-dimensional space (the number of rows in the matrix), calculating the

cosine similarity between each of the vectors will produce a 6 × 6 matrix of cosine similarities that records the angle between each vector. To obtain a distance matrix, the cosine similarity was subtracted from 1 such that 0 indicates similarity, and 1 represents distance[2]. The distance matrix will be of the form:

$$D = \begin{bmatrix} 0 & \dots & \dots & j_{1,n} \\ \vdots & 0 & \dots & j_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ j_{n,1} & j_{n,2} & \dots & 0 \end{bmatrix}$$

where $j_{x,y} = j_{y,x}$. Accordingly, the following distance matrix for the six vectors from Table 3 is:

$$D_{\cos(\theta)} = \begin{bmatrix} 0.00 & 1.00 & 1.00 & 0.63 & 0.50 & 0.50 \\ 1.00 & 0.00 & 0.41 & 1.00 & 1.00 & 1.00 \\ 1.00 & 0.41 & 0.00 & 1.00 & 1.00 & 1.00 \\ 0.63 & 1.00 & 1.00 & 0.00 & 0.26 & 0.26 \\ 0.50 & 1.00 & 1.00 & 0.26 & 0.00 & 0.12 \\ 0.50 & 1.00 & 1.00 & 0.26 & 0.12 & 0.00 \end{bmatrix}$$

where the column/row entries correspond to the artifact-types (here, Pompei VI.5 CE 1001, Pompei VI.5 CE 2255, San Silvestro CT 7, FVM VN.P3, Settefinestre 35.2, and Settefinestre 35.3). The diagonal entries are zero, which reflect the fact that the distance from an artifact-type to itself is zero, while a value of 1 indicates the maximal distance possible. The two entries from Settefinestre have the lowest cosine distances from one another (0.12), while the example from Fiesole (VN.P3) is slightly higher from either of the two (0.26).

The second measure used is the Jaccard index, which does not rely on creation of a term-document matrix, but is a direct measure between two sets. This measure of similarity was first used by botanist Paul Jaccard and has found use in taxonomy and information science (Jaccard 1907, 962; Levandowsky, Winter 1971). For two sets, *A* and *B*, the Jaccard index is defined as:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where Jaccard (*A*,*B*) = 1 indicates that the sets are identical, and Jaccard (*A*,*B*) = 0 indicates complete dissimilarity. The Jaccard distance, like the cosine distance, is equal to 1 – Jaccard (*A*,*B*). The number elements in each set does not have to be equal – e.g., it is possible to have one set with three elements

[2] Properly speaking, the cosine distance is not an actual metric, but since we are interested only in the ranking of the scalar values of the cosine similarity, it is acceptable in this context.

and another of ten elements. Just as with the cosine distance, the matrix of Jaccard distances will be symmetric, with zeros on the diagonal. Applying the Jaccard distance to the six semantic sets expressed in the term-document matrix (Table 3) will yield the matrix:

$$D_{Jaccard} = \begin{bmatrix} 0.00 & 1.00 & 1.00 & 0.71 & 0.68 & 0.68 \\ 1.00 & 0.00 & 0.58 & 1.00 & 1.00 & 1.00 \\ 1.00 & 0.58 & 0.00 & 1.00 & 1.00 & 1.00 \\ 0.71 & 1.00 & 1.00 & 0.00 & 0.32 & 0.32 \\ 0.68 & 1.00 & 1.00 & 0.32 & 0.00 & 0.21 \\ 0.68 & 1.00 & 1.00 & 0.32 & 0.21 & 0.00 \end{bmatrix}$$

which generally resembles the values of the cosine distance matrix. Given these two matrices, it only remains to apply a method of agglomerative clustering, to identify clusters of nearest vessel-types that would be most similar to one another; in other words, to regroup them into clusters that would correspond to new, synthesized categories. Accordingly, I chose agglomerative clustering using Ward's minimum variance method (Ward 1963; Murtagh, Legendre 2011). As defined in the documentation of the scipy library, Ward's method establishes linkages $d(u,v)$ recursively for a cluster u (newly formed by clusters s and t) and another cluster v,

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 + \frac{|v|}{T}d(s,t)^2}$$

where $T = |v| + |s| + |t|$[3]. Applied to the distance matrices given for the six artifact-types in Table 3, Ward's method produces a hierarchical dendrogram which, predictably, clusters together the four black-gloss types (FVM VN.P3, Pompei VI.5 CE 1001, Settefinestre 35.2, and Settefinestre 35.3) in one group and the two unslipped vessels (Pompei VI.5 CE 2255 and San Silvestro CT 7) into another group (Fig. 2). While the resulting measures were slightly different for the linkages from the Jaccard and the cosine distances (as indicated by the values of their matrices), the overall shape and groupings were identical.

In implementing Ward's method on the small pilot dataset that follows, dendrograms were used heuristically to find a cut-off distance for the optimal number of clusters, and it is anticipated that the selection of cluster size will persist in being a point of considerable debate and interpretation. The following example, which uses pilot data from synthkat, illustrates the way in which deciding on the number of clusters should be done manually to ensure that the cluster groups make sense for an archaeological categorization.

[3] http://docs.scipy.org/doc/scipy-0.17.0/reference/generated/scipy.cluster.hierarchy.linkage.html.

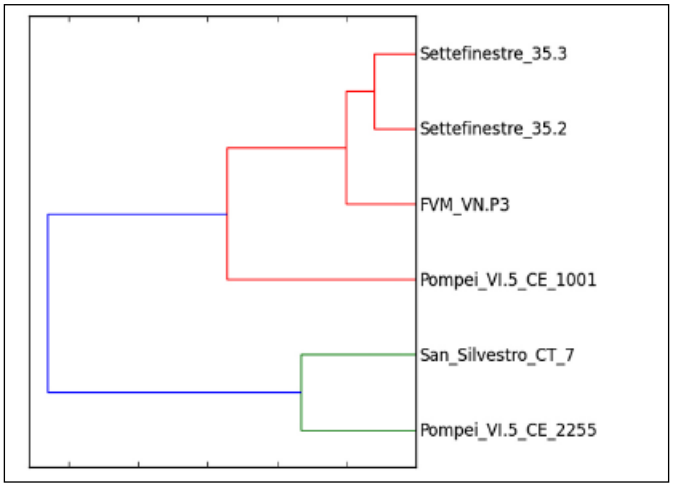

Fig. 2 – Dendrogram of the six vectors from Table 3 using Ward's method. The cosine distance matrix revealed identical groupings as the Jaccard distance.

4. Synthetic groups of Roman ceramics

The bulk of the pilot dataset used as a test case for the performance of the above method was collected from the published results of various projects (Guzzo 1970; Frova 1977; Bonghi Jovino 1985; Ricci 1985; de Marinis 1990; Milanese 1993; Bilde, Poulsen 2008). The collection of vessel-types were based principally on, but were not limited to, classes and types that pertain to the last two centuries BCE. Collection and definition of semantic sets of vessel-types continues, and datasets are uploaded to https://github.com/scollinselliott/synthkat/ as they become available. The pilot dataset fixed for this paper comprises a total number of 1,492 vessel-types. The total number of semantic terms used was 1,732, intentionally limited to increase inter-referentiality between the different document-vectors in the pilot dataset: terms and types that were thought to have broad referentiality were included, whereas those that tended toward particularity (such as descriptions of clays, slip quality, finer morphological details) were, for this pilot dataset, avoided (Table 4).

Measure	Form	Ware
Cosine distance	26	27
Jaccard distance	22	26

Table 4 – The number of clusters *k* by agglomerative hierarchical clustering using Ward's minimum variance method, when the maximum distance from clusters is set at a value of 10.

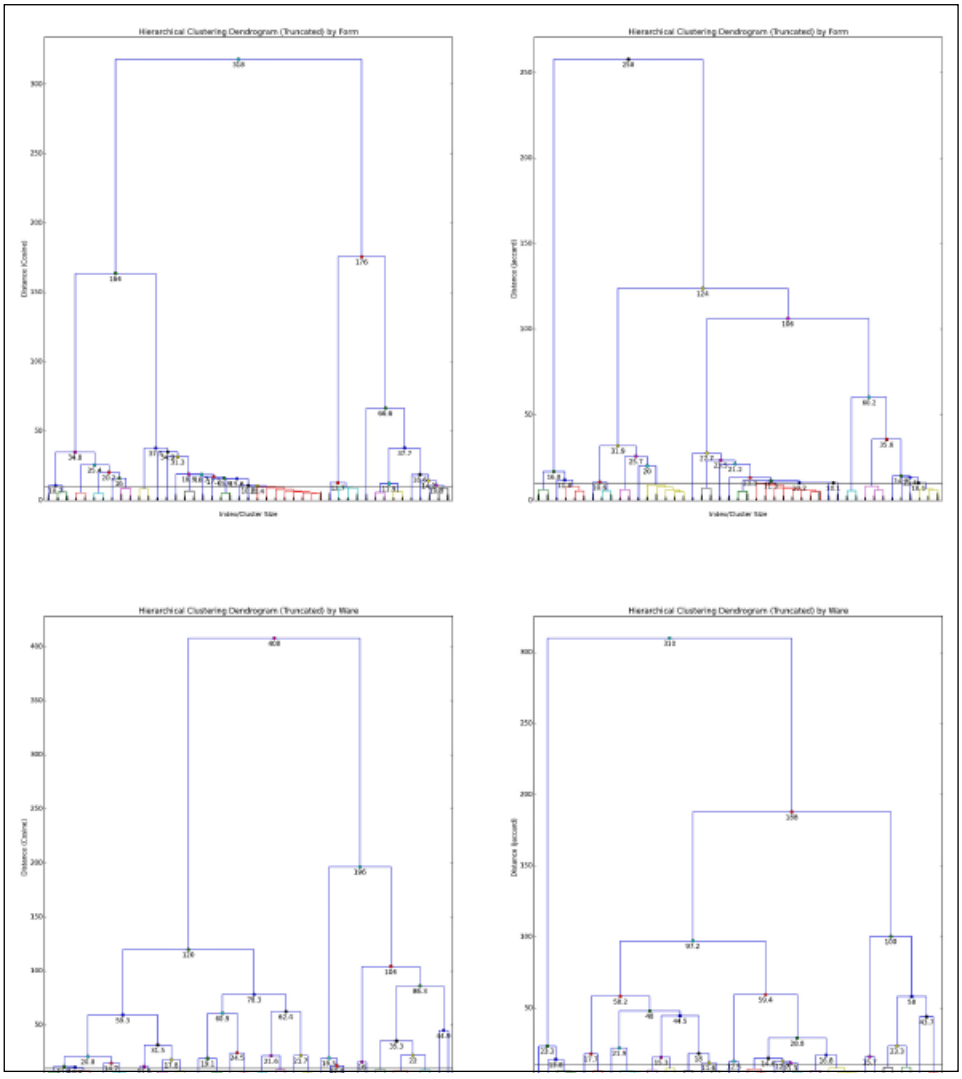

Fig. 3 – Dendrograms of agglomerative clusters using Ward's method (truncated), with a cut-off distance of 10. Vessel-type clusters in the domain of vessel form (above) and ware (below). Dendrograms on the left were based on cosine distances, those on the right on Jaccard distances.

The attempt was to apply, as consistently as possible, the same level of semantic description to each vessel type. Following the above procedure, cosine and Jaccard distance matrices were calculated for both vessel form and ware. Exploratory dendrograms aided in determining the cut-off point for

Domain	$k_{cos(\theta)}$	$k_{Jaccard}$	Jaccard($k_{cos(\theta)}$,$k_{Jaccard}$)
Form	1	4	0.98
Form	2	4	0.01
Form	2	5	0.99
Form	4	7	0.84
Form	4	8	0.09
Form	5	8	0.14
Form	6	8	0.71
Form	11	12	0.66
Form	11	14	0.03
Form	12	14	0.01
Form	12	15	0.89
Form	13	14	0.10
Form	14	13	0.77
Form	14	14	0.02
Form	15	14	0.06
Form	16	14	0.05
Form	17	14	0.73
Form	18	1	0.97
Form	19	1	0.01
Form	19	2	0.27
Form	19	3	0.72
Form	21	18	0.15
Form	21	22	0.08
Form	22	18	0.85
Form	24	19	0.84
Form	24	22	0.09
Form	26	22	0.70
Ware	2	15	0.48
Ware	3	7	0.08
Ware	3	15	0.48
Ware	17	7	0.85

Table 5 – Categorical ambiguity. Instances where different clusters were formed by the cosine distance and the Jaccard distance, with the maximum distance of 10.

the number of clusters (Fig. 3). The methodology for choosing the number of clusters is a particular problem in hierarchical cluster analysis, and computational methods to determine cluster size can still defer to rules-of-thumb or *ad hoc* judgments, such as the "elbow method," used to find the optimal numbers of clusters with respect to variance.

In light of the nature of this exercise, the resulting number of cluster groups should be roughly equivalent to the number of categories that one began with. If there were too few clusters, the ceramic categories become one undifferentiated mass, but too many will result in micro-categories whose boundaries might be semantically irrelevant or misleading (creating sharp

distinctions between "plate" and "platter", for instance). Here, I tried to adhere as closely as possible to a number of clusters that was roughly equivalent to the number of classificatory and/or morphological categories which were used in individual projects at the outset, such as the number of classes in Table 1. Setting an arbitrary maximum distance of 10 between clusters (for both the Jaccard and cosine distances) resulted in an output of 22-27 categorical clusters for either form or ware (Table 4). The resulting dendrograms are given in Fig. 3. The full list of categorical assignments derived from the pilot dataset may be found on the project website.

Furthermore, it would be ideal if linkages produced from cosine and Jaccard distances resulted in equivalent groupings. In other words, the synthetic categorization should not be affected by the selection of the method of measuring distance. A check on the integrity of the resulting categories was made by seeing whether or not there was any overlap in the memberships of ceramic entries when either cosine or Jaccard distances were used, that is, if there were any clusters k – here denoted $k_{\cos(\theta)}$ or k_{Jaccard} depending on its measure – where $0 < \text{Jaccard}(k_{\cos(\theta)}, k_{\text{Jaccard}}) < 1$, that is, if there was any overlap in the categorical membership of an entry. This check revealed that, while most categories were exclusive or identical in either measure, there were some where membership was ambiguous. This tended to be the case more in instances of vessel form, rather than vessel ware. Table 5 lists the different categories which possessed some ambiguity in terms of vessel-type membership.

Merging categories is one effective way to resolve this ambiguity. For the example of vessel ware: $2_{\cos(\theta),w}$ and $3_{\cos(\theta),w}$ were formed by entries of *ceramica comune tirrenica*, *ceramica grezza di importazione extraregionale* (in this case, from Provence), and *grezza ligure*, all of which nomenclature were particular to the corpus of finds from the excavations in the cloister of San Silvestro in Genova (Milanese 1993). The cluster $17_{\cos(\theta),w}$ was made up of various entries of unslipped coarse or common wares. These elements were grouped differently by the Jaccard difference, into two different clusters, $7_{\text{Jaccard},w}$ and $15_{\text{Jaccard},w}$. Accordingly, equalizing these two categories according to the Jaccard measure and those three categories according to the cosine distance resolves that issue.

The case of vessel form is more problematic. Very often it is a matter of just one or two entries that is falling into either one category or another, given the frequent approximation of $\text{Jaccard}(k_{\cos(\theta)}, k_{\text{Jaccard}})$ to either 0 or 1. This would indicate that only slight adjustment would have to be made to the categorization. One case where ambiguity was greater involved the categories $11_{\cos(\theta),f}$…,$17_{\cos(\theta),f}$ and $12_{\text{Jaccard},f}$…,$15_{\text{Jaccard},f}$. Directly examining the assignments made to each of these categories reveals that the Jaccard distance failed to elicit a reasonable categorization. $14_{\text{Jaccard},f}$, for example, groups forms of

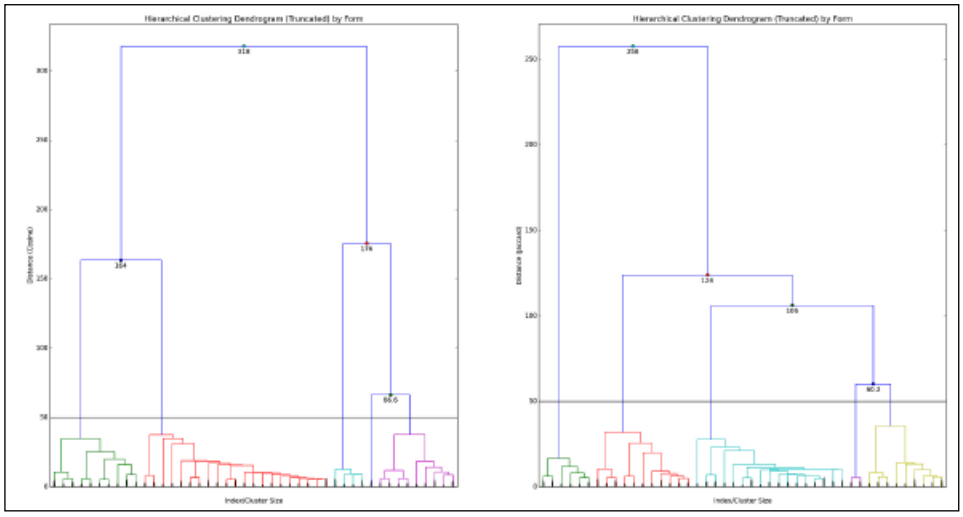

Fig. 4 – Dendrograms of agglomerative clusters using Ward's method (truncated), with a cut-off distance of 50, of vessel-type clusters based on form. Dendrogram on the left was based on cosine distances, that on the right on Jaccard distances.

beakers and saucepans together). The reason why such a clustering failed in this case is foremost that the quality of the semantic set for each of the elements in those sets was fairly poor, sometimes consisting of a single entry, which negatively impacted that vessel-type's referentiality to other forms. Richer description of their semantic sets is necessary, such that it contains more non-unique elements shared by other elements of the entire corpus. The output is only as good as the input.

Additionally, the cut-off point could be adjusted to alter the maximum distance between clusters. In this case of vessel form, setting the maximum distance from 10 to 50 results in identical groupings according to both the cosine and Jaccard distance (even if the way such clusters are positioned hierarchically differs – see Fig. 4). The creation of fewer categories could be justified given a significant enough degree of ambiguity, but improvement of the quality of the semantic sets should be the initial step.

5. Conclusions

The approach presented here admits of considerable flexibility and reverses the traditional approach to vessel classification and typology, creating categories computationally from the bottom up. Since the algorithm produces a new categorization for different datasets, there is no limit to the way in which semantic sets can be constructed, even to serve particular intra- or

sub-classificatory questions. Further work, beyond the scope of this article, includes the implementation of fuzzy sets (or soft categories) as a way to treat, computationally, the ambiguity in vessel categorization and comparison. LSA methods can likewise be explored to examine how well vessels finds may accord with *a priori* categories in existence. Additional measures, such as the Tanimoto distance (Rogers, Tanimoto 1960), can also be explored as a means to assess the differences between semantic sets.

While there should be a post-procedural check on categorical assignment, an algorithmic process can greatly expedite the assignment of ceramic finds to classificatory and morphological groups, enabling the accumulation of large-scale comparisons of ceramic vessel assemblages from multiple different projects that have used different taxonomic systems. The project synthkat offers a preliminary set of data, and the refinement and expansion of the semantic descriptions from the pilot dataset will improve the effectiveness of the clustering algorithm. Algorithmic means do not entail a complete replacement of traditional methods of vessel identification, but they assist in sorting and organizing a body of data at a level that would be impossible for a single individual or group to accomplish. It serves ultimately to help bridge the difficulty in working with older data, and in the inevitable difficulty that comes with translating and interpreting past data, since the knowledge obtained today will inevitably run the same risk of obsolescence.

Stephen A. Collins-Elliott
Department of Classics
University of Tennessee
sce@utk.edu

REFERENCES

Ankerst M., Breunig M.M., Kriegel H.-P., Sander J. 1999, *OPTICS: Ordering Points To Identify the Clustering Structure*, in S.B. Davidson, C. Faloutsos (eds.), *SIGMOD '99. Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data and Symposium on Principles of Database Systems*, New York, ACM Press.

Bats M. (ed.) 1996, *Les céramiques communes de Campanie et de Narbonnaise (1er s. av. J.-C.-IIe s. ap. J.-C.). La vaisselle de cuisine et de table*, Collection du Centre Jean Bérard 14, Naples, Centre Jean Bérard.

Bilde P.G., Poulsen B. 2008, *The Temple of Castor and Pollux, II.1. The Finds*, Occasional Papers of the Nordic Institutes in Rome 3, Roma, L'Erma di Bretschneider.

Bonghi Jovino M. (ed.) 1985, *Ricerche a Pompei: l'insula 5 della Regio VI dalle origini al 79 d.C.*, Roma, L'Erma di Bretschneider.

Cibecchini F., Principal J. 2004, *Per chi suona la campana B?*, in E.C. De Sena, H. Dessales (eds.), *Metodi e approcci archeologici: l'industria e il commercio nell'Italia antica,* Oxford, Archaeopress, 159-172.

Clarke D.L. 1978, *Analytical Archaeology*, London, Methuen & Co (2nd ed.).

Cortese C. 2005, *Le ceramiche comuni: problemi generali e criteri di classificazione*, in Gandolfi 2005, 325-338.

Deerwester T., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. 1990, *Indexing by Latent Semantic Analysis*, «Journal of the American Society for Information Science», 41, 391-407.

Deetz J. 1967, *Invitation to Archaeology*, Garden City, NY, Natural History Press.

Dyson S.L. 1976, *Cosa: The Utilitarian Pottery,* Memoirs of the American Academy in Rome 33, Rome, American Academy in Rome.

Ester M., Kriegel H.-P., Sander J., Xu X. 1996, *A density-based algorithm for discovering clusters in large spatial databases with noise*, in E. Simoudis, J. Han, U. Fayyad (eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Menlo Park, CA, AAAI Press, 226-231.

Ettlinger E., Hedinger B., Hoffmann B., Kenrick P.M., Pucci G., Roth-Rubi K., Schneider G., von Schnurbein S., Wells C.M., Zabehlicky-Scheffenegger S. 1990, *Conspectus formarum terrae sigillatae Italico modo confectae*, Bonn, Dr. Rudolf Habelt.

Frova A. (ed.) 1977, *Scavi di Luni II. Redazione delle campagne di scavo 1972-1973-1974*, Roma, L'Erma di Bretschneider.

Gandolfi D. (ed.) 2005, *La ceramica e i materiali di età romana. Classi, produzioni, commerci e consumi*, Quaderni della Scuola Interdisciplinare delle Metodologie Archeologiche 2, Bordighera, Istituto Internazionale di Studi Liguri.

Giardino L. 1980, *Sulla ceramica a pasta grigia di Metaponto e sulla presenza di alcuni bolli iscritti: studio preliminare*, Studi di antichità 2, Galatina, Congedo Editore, 247-287.

Guzzo P.G. 1970, *Le campagne 1960-1962 della Soprintendenza al Parco del Cavallo*, in *Sibari. Scavi al Parco del Cavallo (1960-1962; 1969-1970) e agli Stombi (1969-1970)*, «Notizie degli Scavi di Antichità», Suppl. 3, Roma, Accademia Nazionale dei Lincei, 24-73.

Hunter J.D. 2007, *Matplotlib: A 2D graphics environment*, «Computing in Science & Engineering », 9, 90-95.

Jaccard P. 1907, *La distribution de la flore dans la zone alpine*, «Revue générale des sciences pures et appliqués», 18, 961-967.

Kriegel H.-P., Kröger P., Sander J., Zimek A. 2011, *Density-based clustering*, «WIREs Data Mining and Knowledge Discovery», 1, 231-240.

Lamboglia N. 1952, *Per una classificazione preliminare della ceramica campana*, in *Atti del I Congresso Internazionale di Studi Liguri (Monaco-Bordighera-Genova 1950)*, Bordighera, 139-206.

Levandowsky M., Winter D. 1971, *Distance between sets*, «Nature», 234, 34-35.

de Marinis G. 1990, *Archeologia urbana a Fiesole: lo scavo di Via Marini-Via Portigiani*, Firenze, Giunti.

Meyer N. 2003, *Pottery strategy and chronotype*, in M. Given, A.B. Knapp (eds.), *The Sydney Cyprus Survey Project: Social Approaches to Regional Archaeological Survey*, Los Angeles, Cotsen Institute of Archaeology, 14-16.

Milanese M. 1993, *Genova romana. Mercato e città dalla tarda età repubblicana a Diocleziano dagli scavi del colle di Castello (Genova-S. Silvestro)*, Roma, L'Erma di Bretschneider.

Moore R.S. 2008, *A decade later: The chronotype system revisited*, in W.R. Caraher, L.J. Hall, R.S. Moore (eds.), *Archaeology and History in Roman, Medieval and Post-Medieval Greece. Studies on Method and Meaning in Honor of Timothy E. Gregory*, Aldershot, Ashgate, 137-151.

Morel J.-P. 1981, *Céramique campanienne: les formes*, BEFAR 244, Rome, École française de Rome.

Morel J.-P. 1998a, *La ceramica a vernice nera del Piemonte: tipologia, storia, cultura*, «Archeologia in Piemonte», 2, 235-252.

Morel J.-P. 1998b, *L'étude des céramiques à vernis noir, entre archéologie et archéométrie*, in P. Frontini, M.T. Grassi (eds.), *Indagini archeometriche relative alla ceramica a vernice nera: nuovi dati sulla provenienza e la diffusione (Milano 1996)*, Como, Edizioni New Press, 9-22.

Murtagh F., Legendre P. 2011, *Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm* (http://arxiv.org/abs/1111.6285).

Olcese G. 2006, *Ricerche archeologiche e archeometriche sulla ceramica romana: alcune considerazioni e proposte di ricerca*, in D. Malfitana, J. Poblome, J. Lund (eds.), *Old Pottery in a New Century. Innovating Perspectives on Roman Pottery Studies. Atti del Convegno Internazionale di Studi (Catania 2004)*, Monografie dell'Istituto per i Beni Archeologici e Monumentali-CNR 1, Roma, L'Erma di Bretschneider.

Preucel R.W. 2006, *Archaeological Semiotics*, Malden, MA, Blackwell.

Ramazzotti M. 2010, *Archeologia e semiotica. Linguaggi, codici, logiche e modelli*, Torino, Bollati Boringhieri.

Read D.W. 2007, *Artifact Classification: A Conceptual and Methodological Approach*,Walnut Creek, CA, Left Coast Press.

Řehůřek R., Sojka P. 2010, *Software framework for topic modelling with large corpora*, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* Valletta, Malta, ELRA, 45-50 (http://is.muni.cz/ publication/884893/en/).

Reynolds P. 2008, *Linear typologies and ceramic evolution*, «Facta: A Journal of Roman Material Culture Studies», 2, 61-87.

Ricci A. (ed.) 1985, *Settefinestre. Una villa schiavistica nell'Etruria romana. La villa e i suoi reperti*, Modena, Panini.

Rice P.M. 1987, *Pottery Analysis: A Sourcebook*, Chicago, University of Chicago Press.

Rogers D.J., Tanimoto T.T. 1960, *A computer program for classifying plants*, «Science», 132, 1115-1118.

Santoro Bianchi C. 2005, *La ceramica comune: ancora qualche riflessione*, in Gandolfi 2005, 349-352.

Shepard A.O. 1956, *Ceramics for the Archaeologist*, Publication 609, Washington, D.C., Carnegie Institution of Washington.

Steward J.H. 1954, *On the concept of types*, «American Anthropologist», 56, 54-57.

Vittoria E. 2011, *Grey ware*, in J.C. Carter, A. Prieto (eds.), *The Chora of Metaponto 3: Archaeological Field Survey, Bradano to Basento,* I, Austin, University of Texas Press, 271-301.

Ward J.H. Jr. 1963, *Hierarchical grouping to optimize an objective function*, «Journal of the American Statistical Association», 58, 236-244.

ABSTRACT

This paper addresses the issue of standardization in the cross-comparability of different vessel assemblages. It presents a computational method for building vessel categories from the bottom up, by comparing the specified attributes of a collection of vessel-types, and grouping like with like. Thus, it provides a platform for translating vessel data which may have been classified or divided by type using one taxonomy, bringing them into communication with those categorized by another. Two different methods of measuring the similarity among vessel-types – cosine similarity and the Jaccard index – are explored, toward providing a control on the resulting "synthetic" categories. An exploratory dataset, collected from published data of archaeological projects in Italy focusing on ceramic vessels of the last two centuries BCE, was used to test the performance of this approach. Project data and results are open source and are available online at https://github.com/scollinselliott/synthkat/.