Visualization and Automatic Typology Construction of Pottery Profiles

Laurens van der Maaten[1], Guus Lange[2], and Paul Boon[1]

[1] TiCC, Faculty of Humanities, Tilburg University. Tilburg, The Netherlands.
[2] Cultural Heritage Agency. Amersfoort, The Netherlands.

Abstract

Since the second half of the last century, computers became available to archaeologists. Since then, many have used various multivariate analysis techniques in the classification of archaeological objects with more or less success. Over the last decades, the popularity of multivariate analysis seems to have seized, partly because of the frequently disappointing performance of techniques such as principal components analysis and factor analysis. In particular, these techniques are hampered by their linear nature and by their emphasis on retaining global data structure.

In this paper, we discuss a recently proposed technique for multivariate analysis, called t-SNE, which overcomes many of the weaknesses of principal components analysis and factor analysis. We illustrate the strong performance of t-SNE in experiments on a dataset of pottery profile drawings, in which we combine t-SNE with a shape matching technique based on shape contexts.

Next to our discussion on t-SNE, we discuss a recently proposed affinity-based clustering technique, called affinity propagation. We show how affinity propagation can be used for the automatic construction of a typology of the shape profile drawings.

***Key words**: Multivariate analysis, pottery profiles, typologies, shape matching, data visualization, clustering.*

1 Introduction

Multivariate analysis is a subfield of statistics that is concerned with the automatic analysis and interpretation of (large) datasets that contain high-dimensional data, i.e., data that contains a large variety of variables or measurements[1]. Multivariate analysis is very relevant to archaeology, as it allows for, e.g., the detection of hidden patterns in data gathered from excavations, the generation of hypotheses, and the presentation of objective evidence for hypotheses.

In the last half of the previous century, techniques for multivariate analysis received much attention in the archaeological field[2]. In particular, many archaeologists have used 'traditional' techniques for multivariate analysis developed in other fields, such as principal components analysis[3], factor analysis, or cluster analysis[4] in the analysis of their data. However, over the last decade, the interest of archaeologists in multivariate analysis seems to have faded. One of the main reasons for the reduced interest in multivariate analysis is probably the often disappointing performance of the 'traditional' statistical techniques mentioned above. The poor performance of these techniques can often be explained from either the linear nature of the techniques, or from problems in the objective functions (typically sums of squared errors) that the techniques use. In the statistics and machine learning communities, a large effort has been made to overcome these problems. This effort

[1] A.R. Feinstein. *Multivariable Analysis*. New Haven, CT: Yale University Press, 1996.

[2] Viz. Clarke 1962, Doran and Hodson 1975, Orton 1980, Lange 1990, Adams and Adams 1991.

[3] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philiosophical Magazine,* 2:559-572, 1901.

[4] J.A. Hartigan. *Clustering Algorithms*. Hoboken, NJ: Wiley, 1975.

has resulted in the development of a large number of new techniques for multivariate analysis with often remarkable performances compared to traditional statistical techniques.

In this paper, we discuss two novel techniques for multivariate analysis: (1) a technique that visualizes high-dimensional data through nonlinear dimensionality reduction and (2) a techniques for the identification of clusters of similar objects. We combine the two techniques with a sophisticated technique that measures the similarity between shapes, and apply the techniques on a collection of pottery profile drawings. Specifically, we use the visualization technique to make two-dimensional plots that reveal the underlying structure of the objects in the collection, and we use the clustering technique to automatically construct typologies for the objects in the collection.

The outline of the remainder of this paper consists of five main parts. First, we discuss the shape matching technique that forms the basis for our experiments. Second, we present the technique that is used to visualize the similarities between objects in a two-dimensional plot. Third, we address the clustering technique that we use to automatically construct typologies from the collection of pottery profile drawings. Fourth, we present the setup and results of the experiments we performed on the pottery profile drawing collection. Fifth and last, we discuss the results of our experiments, and we present conclusions and direction for future work.

2 Shape comparison

The computation of the similarity of two shapes is a well-studied problem in computer vision. Frequently used techniques include statistical moments[5], curvature-based measures[6] and shape contexts[7]. In this study, we opt for the use of the latter technique. The shape context method is briefly outlined below. Extensive descriptions of the algorithm are presented by Belongie, Malik, and Puzicha (2002). The main steps of the computation of dissimilarities between shapes based on shape contexts are illustrated in Figure 1.

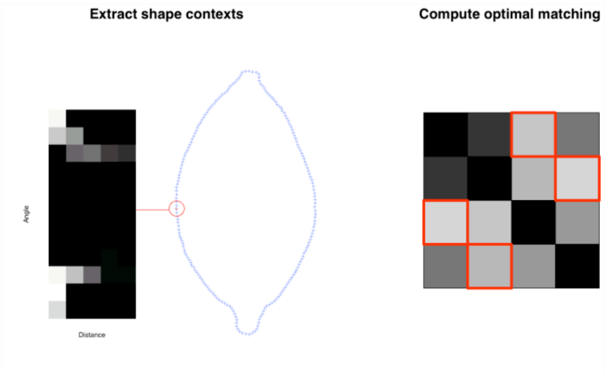

Figure 1. Illustration of shape comparison using shape contexts.

The key idea behind shape contexts is to sample a set of points from the shape contour and to describe these points with local descriptors – the shape contexts – that measure the measure the relative angle and distance to the other points that were sampled from the shape contour. The advantage of the use of local descriptors, i.e., descriptors that assign large weight to the neighborhood of the point at hand, is that the descriptors are robust against noise and minor deformations of the shape. The relative distance and angle measurements give rise to descriptors that are invariant under rotation and rescaling of the shape images. The dissimilarity between two shape contexts, i.e., between two collections of local descriptors corresponding to two shapes, is computed in the following two steps.

First, the minimal assignment costs between the two collections of local descriptors are computed. The distance between two local descriptors in the assignment problem is a standard Euclidean distance. The solution of the assignment problem defines a one-to-one mapping from points on the first shape contour to points on the second contour.

[5] J. Ricard, D. Coeurjolly, and A. Baskurt. Generalizations of angular radial transform for 2D and 3D shape retrieval. *Pattern Recognition Letters* 2614:2174–2186, 2005.

[6] F. Mokhtarian, S. Abbasi, and J. Kittler. Efficient and robust retrieval by shape content through curvature scale space. In *Proceedings of the International Workshop on Image Databases and Multimedia Search*, pages 35-42, 1997.

[7] S. Belongie, J. Malik, and J. Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 244:509-521, 2002.

Second, a thin plate spline warping[8] is computed based on the correspondences between the shapes in such a way, as to minimize the difference between the two shapes. Herein, the difference between the two shapes is defined as the sum of the lengths of the correspondence edges. The dissimilarity between the two shapes is defined as a linear combination of the energy of the thin plate spline warping and the residual difference between the first shape and the second (warped) shape.

3 Visualization

Data visualization can be performed using techniques for dimensionality reduction. Dimensionality reduction techniques model objects by points in a low-dimensional space (typically of two dimensions to facilitate easy visualization) in such a way, that the original pairwise (dis)similarities between the objects are preserved as good as possible in the low-dimensional space. For instance, principal components analysis finds a linear mapping of the original high-dimensional data that minimizes the sum of squared errors between the pairwise Euclidean distances in the high-dimensional and the low-dimensional space. The main problem of principal components analysis is that, as a result of the squared error criterion, it mainly focuses on retaining large pairwise distance in the low-dimensional space. However, it is generally accepted that preservation of small pairwise distances is much more important in data visualization.

Recently, van der Maaten and Hinton (2008) proposed an alternative to PCA, called t-SNE, that performs remarkably well in the visualization of high-dimensional data. The key idea behind t-SNE is to focus on retaining local data structure (i.e., small pairwise distances) instead of on the global data structure. The technique consists of three main steps, which are briefly described below. The three steps are illustrated in Figure 2.

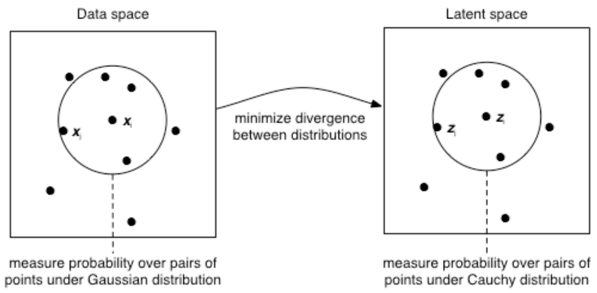
Data space
Latent space
x_j
x_i
minimize divergence between distributions
z_j
z_i
measure probability over pairs of points under Gaussian distribution
measure probability over pairs of points under Cauchy distribution

Figure 2. Illustration of visualization using t-SNE.

First, the local structure of the input data is evaluated. This is done by centering a normal (i.e., Gaussian) distribution over each data point, and measuring the density of all datapoints under that normal distribution. After normalization, this process results in probabilities p_{ij} that are proportional to the similarity of the datapoints i and j.

Second, we generate a random set of points (which corresponds to the input data) in the two-dimensional scatter plot, and we define similar in the scatter plot. The only difference from the similarity measurements described above is that, now, a Student-t distribution with a single degree of freedom (i.e., a Cauchy distribution) is employed instead of a normal distribution. The probabilities in the two-dimensional scatter plot are denoted by q_{ij}.

Third, the points in the two-dimensional scatter plot are iteratively moved around in order to minimize the difference between the probabilities over the data p_{ij} and the probabilities over the map q_{ij}. The difference between p_{ij} and q_{ij} is measured by the so-called Kullback-Leibler divergence[9], which is a natural measure for the difference between two probability distributions. The ‘moving around’ of the points in the scatter plot (in order to represent the similarities of the data points faithfully) is performed using gradient descent.

[8] F. Bookstein. Principal warps: Thin-plate splines and decomposition of transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 116:567-585, 1989.

[9] S. Kullback and R.A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics* 221:79-86, 1951.

4 Clustering

The aim of clustering is to identify groups of objects that are similar under the specified similarity measure (in our case, the shape context distance). This aim can be achieved by, e.g., minimizing the sum of the distances between points and their corresponding cluster centers, as is done in *k*-means clustering[10]. An important disadvantage of *k*-means clustering, and most other clustering techniques, is that they require the user to specify the number of clusters *k* beforehand. In the context of automatic typology construction, this implies that the archaeologist needs to specify the number of types that the typology should comprise, even though he has no clear insight into the underlying similarity structure of the data. Recently, Frey and Dueck (2008) presented a technique, called affinity propagation, that overcomes this limitation by automatically selecting the number of clusters. In our work, we adopt the affinity propagation technique, which is described below.

Affinity propagation aims to maximize the sum of the pairwise similarities and their corresponding exemplars. Supposing we are given a similarity matrix S with elements s_{ij}, such as a matrix of negative shape context distances, affinity propagation performs the following maximization:

$$\max \sum_{i\in E} \sum_{j\in I_i} s_{ij} - \partial_{ij}$$

Herein, the term ∂_{ij} is included to prevent the technique from selecting illegal clustering solutions, i.e., from clustering solution in which data points are assigned to more than one exemplar.

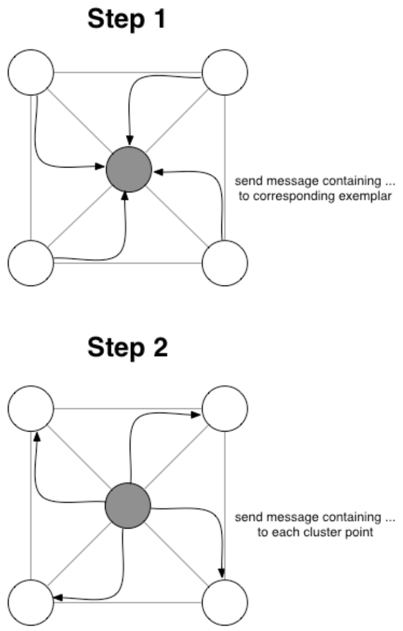
Step 1
send message containing ...
to corresponding exemplar
Step 2
send message containing ...
to each cluster point

Figure 3. Illustration of the message-passing algorithm underlying affinity propagation.

The maximization of the function is performed using a simple message-passing algorithm[11], which is illustrated in Figure 3.

As mentioned above, affinity propagation automatically selects the number of clusters in the data. It does so by introducing a small penalty for each point that is selected to be an exemplar. As a result, the technique will only add an additional exemplar (i.e., cluster) if this exemplar is truly required to model the cluster structure of the data (if the cluster structure does not require the addition of an additional exemplar, affinity propagation will not add an exemplar because of the penalty). Hence, in the automatic typology construction setting, affinity propagation automatically determines the optimal number of types that is required to describe the structure in the data.

[10] J.A. Hartigan. *Clustering Algorithms*. Hoboken, NJ: Wiley, 1975.

[11] The message passing algorithm can be derived by performing belief propagation in a factor graph corresponding to the objective function, see Frey and Dueck, 2008.

4 Experiments

In order to evaluate the performance of the novel visualization and clustering techniques, we performed experiments on a dataset of pottery profile drawings. This section presents the results of our experiments, and consists of three main parts.

First, we introduce the pottery profile dataset that was used in the experiment. Second, we present the setup of the experiments. Third, we present the results of the experiments.

Dataset

The pottery profile dataset contains 1,087 drawings of pottery profiles that were scanned from 14 publications. All pottery was found in The Netherlands, but the majority of the pottery originates from the northern part of the country. The images were scanned to scale with a resolution of 200 dpi. The images were manually preprocessed in order to remove small scanning artifacts, and the resulting images were binarized. For parts of the dataset, typological classifications are available that follow the typology introduced by Taayke[12]. The dataset is described in more detail by Mom[13].

Experimental setup

The setup of our experiments consists of two main steps, which are described below.

First, we perform shape context matching on all pairs (i,j) of pottery profile drawings to compute their pairwise dissimilarity d_{ij}. In the experiments, we assume that the similarity of two shapes is symmetric, i.e., that $d_{ij} = d_{ji}$. The shape matching results in a pairwise dissimilarity matrix D with elements d_{ij}. We normalize the elements of matrix D in such a way that they lie between 0 and 1.

Second, we use the pairwise dissimilarity matrix D as input into t-SNE (for visualization) and affinity propagation (for automatic topology construction). As affinity propagation takes as input a collection of pairwise similarities, we define a similarity matrix S with elements $s_{ij} = 1 - d_{ij}$. We set the preference value s_{ii} that is used by affinity propagation to determine the number of clusters as the median of the pairwise similarities s_{ij} (with $i \neq j$).

The result of running t-SNE on the pairwise dissimilarity matrix D is a two-dimensional scatter plot, in which each point corresponds to a pottery profile. We visualize the result by plotting the pottery profile images on top of their corresponding points.

The result of running affinity propagation on the pairwise similarity matrix S is a subdivision of the pottery profiles into clusters. We visualize the clustering of the pottery profiles by plotting all profiles in a rasterized image, in which the profiles in a single column correspond to the same cluster.

Results

In Figure 4, we show the results of applying t-SNE onto the pairwise dissimilarities that were obtained through shape context matching. All shape profiles in the drawing are drawn to scale. The resulting visualization reveals a great deal of the underlying structure of the data. In particular, the results reveal the variety in properties such as scale, angle of the pottery's belly, thickness of the pottery, and the shape of the base in the entire pottery profiles collection.

In Figure 5, we show a visualization that was obtained in exactly the same way as the visualization in Figure 4, except now we only used the data that was gathered and presented by Taayke[14]. The pottery profiles presented by Taayke are all categorized accroding to this – traditionally made – typology. This allows us to color the profiles in the visualization based on Taayke's typology. This coloring is performed in Figure 5. The results presented in the figure show that the visualization captures much of the structure that Taayke aimed to capture when he developed his

[12] E. Taayke. *Die einheimische Keramik der nördlichen Niederlande, 600 v.Chr. bis 300 n.Chr.* Dissertation, University of Groningen, 1996.

[13] V. Mom. *Secanto: The Section Analysis Tool.* In Proceedings of the XXXIII Computer Applications in Archaeology Conference, pages 95–101, 2005.

[14] E. Taayke. *Die einheimische Keramik der nördlichen Niederlande, 600 v.Chr. bis 300 n.Chr.* Dissertation, University of Groningen, 1996.

typology: profiles with the same color (i.e., the same type) are often modeled fairly close together. On the other hand, the visualization also shows profiles that have a distinctively different color from their neighbors in the visualization, and are thus not modeled near the other profiles with the same type. There are four possible reasons for such 'errors': (1) an error occurred in the computation of the shape similarities, (2) an error occurred in the modeling of the shape similarities in the visualization, (3) important attributes have not been included in the analysis, for instance, as in our case, surface treatment and decoration, or (4) there are 'errors' or inconsistencies in the used typology. In the latter case, visualizations such as the ones presented in Figure 4 and 5 may help the archaeological expert to identify and resolve such inconsistencies[15]. Also the visualizations may point out to the expert which relevant variables still need to be included in the analysis.

In Figure 6, we show the results of applying affinity propagation on the pairwise dissimilarities obtained from the shape context matching. In the figure, each column corresponds to a cluster. As in Figure 5, the colors of the pottery profiles indicate the type of the pottery according to the typology that was constructed by Taayke. The number of clusters to be constructed was automatically determined by affinity propagation. The results presented in Figure 6 may be interpreted as an automatically constructed typology.

Although the clustering seems to contain a few 'errors', the clustering seems fairly consistent overall. Note the difference in the size of the clusters. Als notice that the automatically constructed typology was made based only on shape information, and not based on any other domain or contextual knowledge about the artifacts (we discuss this issue in more detail below).

[15] We contacted the original developer of the typology, Taayke, but, unfortunately, he did not comment on the visualization in Figure 2.

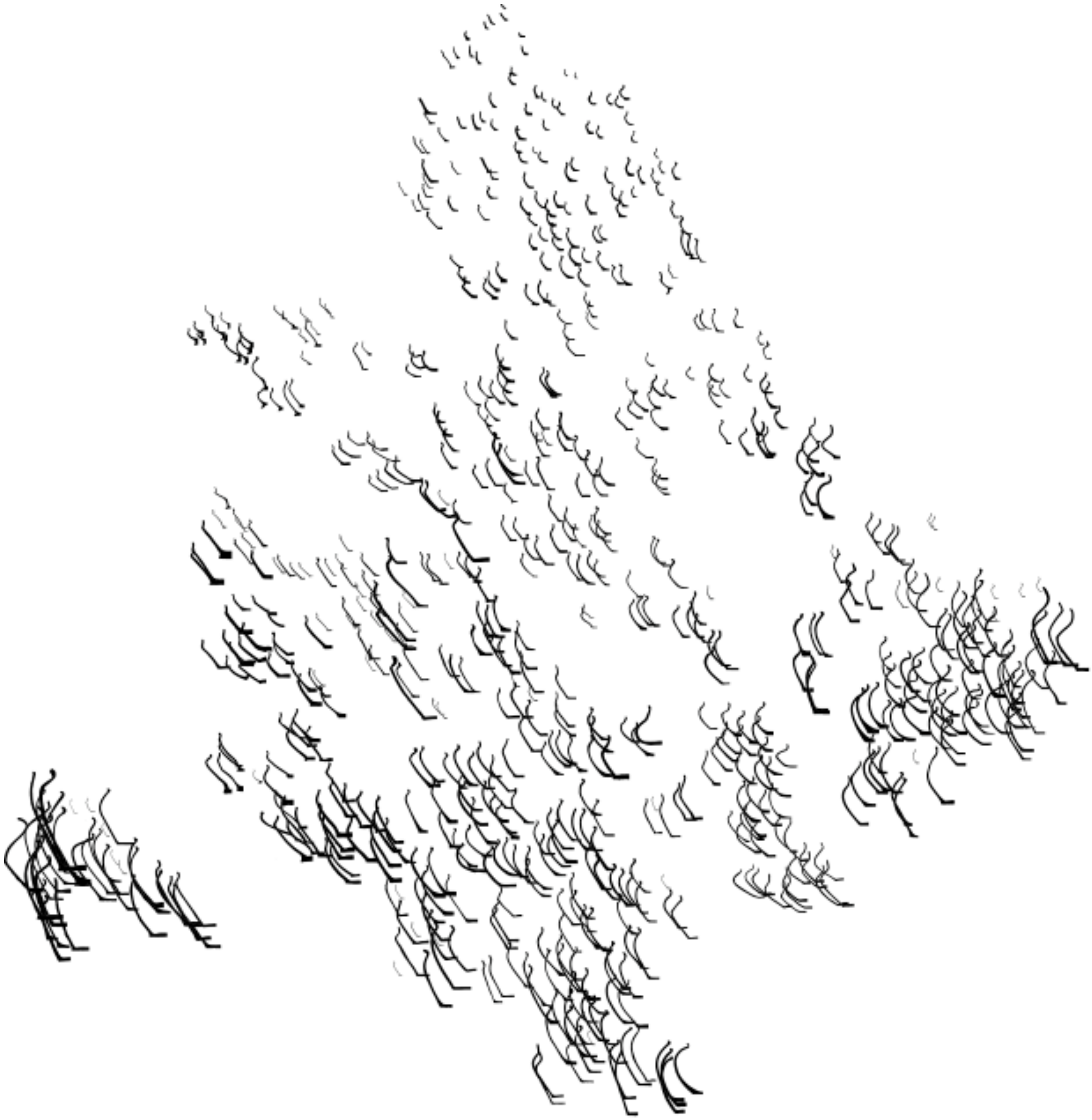

Figure 4. Visualization of the pottery profiles dataset using shape contexts and t-SNE. All profiles are drawn to scale.

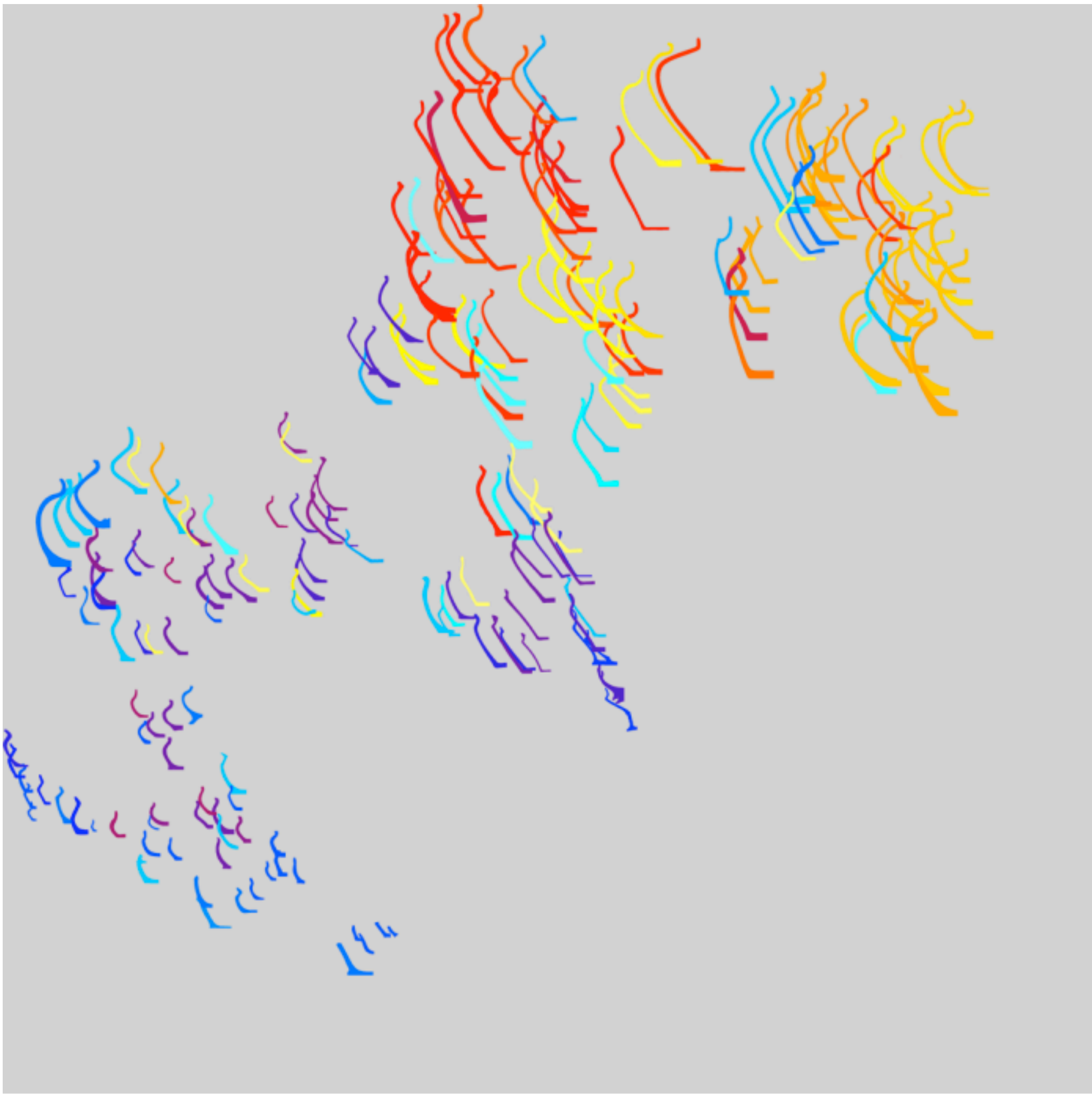

Figure 5. Visualization of all pottery profiles published by Taayke. The pottery profiles are colored according to the typology presented by Taayke.

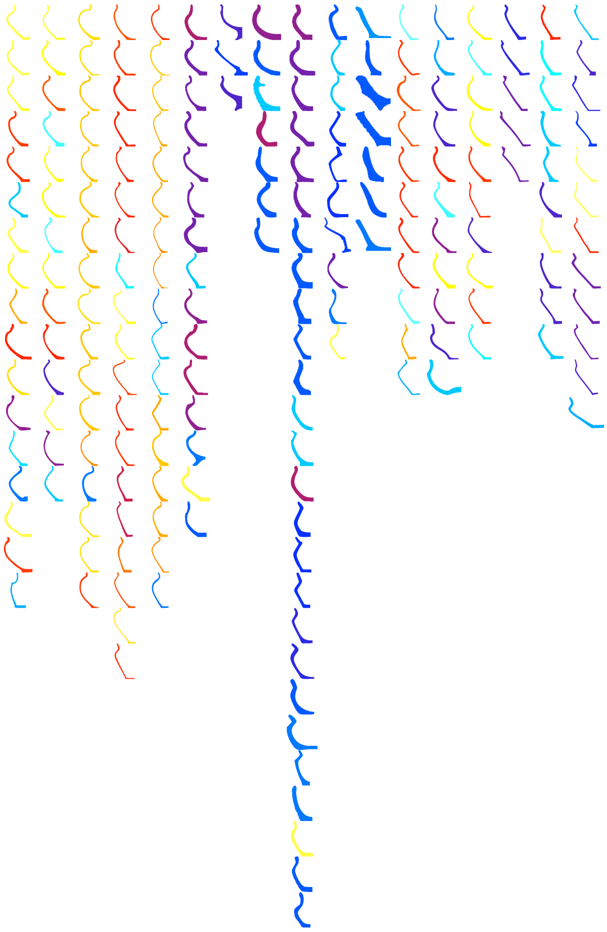

Figure 6. Clusters identified by affinity propagation based on shape context similarities. Each column in the figure corresponds to a cluster.

5 Discussion

The results presented in the previous section raise a number of new interesting questions, which could not be addressed before (due to the poor performance of traditional techniques for multivariate analysis). Below, we pose and discuss three such questions.

How to evaluate the results presented by statistical approaches?

An important question that quickly comes to mind when carefully inspecting our results is how we should assess such results and what we may conclude from them. This question is not easy to answer, as we cannot employ standard statistical tests to test hypotheses about whether, e.g., two profiles belong to the same group.

In general, visualizations and clusterings such as those presented in the paper should be considered as additional objective, meaning independent repeatable, evidence that can be used to support hypotheses we already developed. If a large number of different shape matching and visualization techniques (and the combinatorial number of combinations thereof) consistently model two profiles close together, we can use this information to support our hypothesis that the two profiles are member of the same group of profiles. On the other hand, we do acknowledge that the use of statistical evidence in this way is not completely satisfactory.

How to incorporate domain or contextual knowledge?

The approach to visualization of pottery profile drawings that we presented in this paper does only incorporate a fairly limited amount of domain knowledge. An example of domain knowledge that is included is that the shape of pottery is of relevance to its function, and hence, to its type. However, it is unlikely that this degree of domain knowledge is sufficient to obtain high-quality visualizations or clusterings of archaeological data. It is certainly possible to include more domain knowledge into approaches such as the one presented above. For instance, in the case of pottery, it is well-known that small changes in the width of the base have a large influence on the volume of the pottery, as a result of which base width is a relatively important variable in the determination of pottery. It is possible to incorporate this domain knowledge in the shape matching, e.g., by assigning more weight to the base of the profiles (as is done by Mom and Drenth elsewhere in this volume)[16]. In similar ways, contextual knowledge such as find location, dating (based on C14 or dendrochronology), etc., should also be incorporated in our approaches as well. For instance, the find location may be included in the construction of the visualization as an additional variable.

[16] Mom and Drenth, elsewhere in this volume.

How do the results of multivariate analysis relate to 'traditional' typologies?

We do not intend to rediscuss the theoretical pros and cons of the use of typologies. These pros and cons have received ample discussion in the literature cited above and more lately in Lange (ed) 2004[17]. What we hope to have demonstrated is that the automatic analysis of archaeological material is a valid and practical tool in visualizing the variation and the mutual relations in the 'unmanageable mass of individual units that form the basic archaeological record'[18].

Our results do raise archaeologically relevant questions with respect to the value of typologies if the multivariate analysis does not identify the presence of clusters. This can be observed in our visualizations in Figure 4 and 5, where most structure in the data seems to consist of gradual changes in, e.g., the thickness of the pottery, the size of the pottery, and the angle of the belly of the pottery. This raises the question whether for this kind of material the use of hard coded, traditionally constructed, typologies, is scientifically sound.

Typologies define either hard boundaries between groups of objects, or they define a number of centroids – ideal/holo/archetypes – around which all similar others aggregate more or less closely. As a result, they often do not completely respect the gradual scales. However, they remain an essential tool in the classification of archaeological material that facilitate the generation of manageable descriptions of the material. We believe multivariate analysis and traditional typologies should be used together in order to generate descriptions of the data that are as complete as possible. In such a combined approach, the multivariate analysis can serve as objective evidence for the proposed typology of classification.

6 Conclusion

The paper discussed two new techniques for multivariate analysis, and combined these techniques with a sophisticated shape matching technique. The results of our experiments with this combination on a dataset of pottery profile drawings are encouraging, and may readily be used as objective evidence for typologies or classification that were constructed in traditional manners.

Future work primarily focuses on incorporating more domain knowledge into the developed techniques. For instance, it is well known that the base of potterys is of large influence to its volume, and is thus of high importance to its function. This domain knowledge could be exploited in the software, e.g., by assigning additional weight to differences in base width in the shape context matching. The same holds for the angle of the top of the pottery ('open' or 'closed'), which is very important as it determines whether pottery was used for cooking or storage.

[17] A.G. Lange. *De Horden near Wijk bij Duurstede: plant remains from a native settlement at the Roman frontier; a numerical approach*. ROB, Amersfoort, 1999.

[18] J.E. Doran and F.R. Hodson. *Mathematics and Computers in Archaeology* (p.158). Edinburgh University Press, 1975.

Acknowledgements

The authors thank Vincent Mom for generously providing the pottery profiles dataset. This research was supported by The Netherlands Science Foundation (NWO) in the context of the CATCH program, project RICH, under grant 640.000.002.

Bibliography

Adams, W.Y. and Adams, E.W. 1991. *Archaeological typology and practical reality. A dialectical apprach to artifact classification and sorting*. Cambridge.

Belongie, S., Malik, J., and Puzicha, J. 2002. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 244:509-521.

Bookstein, F. 1989. Principal warps: Thin-plate splines and decomposition of transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 116:567-585.

Clarke, D.L. 1962. Matrix analysis and archaeology with particular reference to British Beaker pottery. *Proceedings of the Prehistoric Society* 28, 371-382.

Doran, J.E. and Hodson, F.R., 1975. *Mathematics and Computers in Archaeology*. Edinburgh University Press.

Feinstein, A.R. 1996. *Multivariable Analysis*. New Haven, CT: Yale University Press.

Frey, B.J. and Dueck, D. 2008. Clustering by Passing Messages Between Data Points. *Science* 315:972–976.

Hartigan, J.A. 1975. *Clustering Algorithms*. Hoboken, NJ: Wiley.

Kullback, S. and Leibler, R.A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 221:79-86.

Lange, A.G. 1990. *De Horden near Wijk bij Duurstede: plant remains from a native settlement at the Roman frontier; a numerical approach*. ROB, Amersfoort.

Lange, A.G. (ed.) 2004. *Reference Collections: Foundation for Future Archaeology*. ROB, Amersfoort.

Mokhtarian, F., Abbasi, S., and Kittler, J. 1996. Efficient and robust retrieval by shape content through curvature scale space. In *Proceedings of the International Workshop on Image Databases and Multimedia Search*, pages 35-42.

Mom, V. 2007. Secanto: The Section Analysis Tool. In: A. Figueiredo / G. Leite Velho (eds.), The world is in your eyes. CAA. 2005. Computer Applications and Quantitative Methods in Archaeology. *Proceedings of the XXXIII Computer Applications in Archaeology Conference, Tomar, Portugal, March 2005,* 95–101. Tomar.

Orton, C.R. 1980. *Mathematics in Archaeology*. Cambridge University Press.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philiosophical Magazine,* 2:559-572.

Ricard, J., Coeurjolly, D., and Baskurt, A. 2005. Generalizations of angular radial transform for 2D and 3D shape retrieval. *Pattern Recognition Letters* 2614:2174–2186.

Spearman, C. 1904. General intelligence objectively determined and measured. *American Journal of Psychology* 15:206-221.

Taayke, E., 1996: *Die einheimische Keramik der nördlichen Niederlande, 600 v.Chr. bis 300 n.Chr.* Dissertation, University of Groningen.