Journal of Archaeological Science: Reports xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

journal homepage: http://ees.elsevier.com/jasrep

An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships

Elisavet Charalambous [a,*], Maria Dikomitou-Eliadou [b], Georgios M. Milis [c], Georgios Mitsis [a,d], Demetrios G. Eliades [e]

[a] Department of Electrical and Computing Engineering, University of Cyprus, Nicosia, Cyprus
[b] Archaeological Research Unit of the Department of History and Archaeology, University of Cyprus, Nicosia, Cyprus
[c] G.M. EuroCy Innovations Ltd, Nicosia, Cyprus
[d] Department of Bioengineering, McGill University, Montreal, QC, Canada
[e] KIOS Research Center for Intelligent Systems & Networks, University of Cyprus, Nicosia, Cyprus

ARTICLE INFO

Article history:
Received 4 November 2014
Received in revised form 22 June 2015
Accepted 13 August 2015
Available online xxxx

Keywords:
Experimental design
Ancient pottery analysis
Classification
Compositional analysis
Statistical testing

ABSTRACT

This paper proposes an experimental design for the compositional classification of 177 ceramic samples deriving from domestic and tomb contexts in Cyprus dated to the Early and Middle Bronze Age. In this design, ceramic sample classification is achieved with three well-known methods, a standard statistical learning method termed k-Nearest Neighbours (k-NN), a method using Decision Trees (C4.5) and a more complex neural network based method known as Learning Vector Quantisation (LVQ). It is shown that the examination of classification patterns through confusion matrices allows the exploitation of inter-class relationships and the ability to provide extra information to the researcher about the compositional categorisation of samples; which could not be grouped (with certainty) into classes with the employment of ceramic petrography. Due to the compositional heterogeneity of ceramics, the effectiveness of classification using only chemical elements with mean concentrations lower than 0.1% is also evaluated to illustrate their potential significance. The developed design follows a systematic approach and well-established methods, such as bootstrapping with replacement and the 5×2 cross validation (paired t-test and F-test) tests, to ensure that the results are statistically significant.

1. Introduction

Archaeology ultimately aims at investigating social causation through the examination of gathered residue evidence (Barceló, 2008). Pottery analysis, in particular, has been proven cross-culturally an indispensable tool for indirectly approaching past people and societies through their cultural remnants, allowing inferences about their technology, and their interaction with their surrounding physical and social environments. For this reason, compositional (mineralogical and chemical) and microstructural analyses have become an integral part of interdisciplinary archaeological research, underlining the importance of compositional and technological comparative studies. Nonetheless, any pottery analysis is not a straightforward process, and there are various parameters (i.e. contextual, spatial, chronological, compositional, technological) that the researchers need to consider while defining their research design, their sampling strategy, and later while evaluating their research results. Among these parameters, the inherent heterogeneity characterising ceramic composition sets significant challenges when trying to utilise the greatest amount of information possible, especially considering that, generally, the most highly variable elements have the greatest of the impacts on the multivariate data ensemble and that they do not necessarily depict elements with high concentrations (Reimann et al., 2012).

Multivariate statistical methods have a long track in archaeometric data analysis of which the most common are cluster analysis and dimensionality reduction techniques (both supervised and unsupervised). The analysis of archaeometric data imposes problems that are not easily handled, if at all, by classical methods. Due to this, over the past couple decades there has been a great interest in alternatives to the standard statistical methods of analysis (Baxter, 2006). The size of the produced datasets coupled with the multiple influencing analysis and contextual factors impose both analytical and computational problems, while it is important to note that the selection of the most effective analysis method depends on the characteristics of the data.

Statistical analysis on data may allow the study of their internal structure and reveal interesting technological and compositional patterns. As such, archaeological data classification is concerned with the application of classification methods on archaeological data, while also taking into account the characteristics of the artefacts under study.

* Corresponding author.

http://dx.doi.org/10.1016/j.jasrep.2015.08.010
2352-409X/

Classification aims at identifying to which element of a set of categories, a new uncategorised artefact belongs, on the basis of a training set of artefacts the class/type of which is known.

Classification results are not unique; they depend on the deployed classification method, their parameterisation, as well as the manner in which the raw data are treated to form the input dataset. Evaluation of the validity and the plausibility of classification results is not only necessary, but critical. Statistical hypothesis testing methods allow the inference of a hypothesis ensuring that the predicted result is unlikely to have occurred by chance alone, according to a pre-determined threshold probability (Coolican, 1999).

The roots of classification analysis of archaeometric data are traced back multiple decades ago with the contribution of Kowalski in 1972 (Kowalski et al., 1972) being an early landmark. In subsequent years, classification methods have been used in a number of studies (Mussumarra et al., 1995; Fermo et al., 2008; Kowalski and Bender, 1973; Baxter, 1994; Lopez-Molinero et al., 2000). A clear milestone in the analysis of archaeometric data is Baxter's work in 2006, where he reviews the application of classification methods (among others) on the chemical composition of glass artefacts (Baxter, 2006). The effectiveness of a variety of classification methods was evaluated; among them also the three methods deployed in this paper. However, despite the similarities between Baxter's review and this paper, the results of the two works cannot be straightforwardly comparable due to the different experimental data and deployed methodology.

In this paper, an experimental design is proposed for the classification of chemical compositional data obtained from a sample of utilitarian pottery. The aim of the experiment is neither to achieve perfect classification, nor to discriminate the origin of each artefact. The target is rather to develop a plausible, unbiased and statistically valid methodology for classification, which takes into consideration the idiosyncrasies of archaeometric data, in general, and chemical compositional in particular, and also to examine the validity of the produced categorisation. The proposed methodology is subsequently used to differentiate a series of ceramic specimens based on their fabric, and investigate the degree of similarity between discriminated types. For demonstration purposes and for the needs of this paper, classification is achieved with three well-known methods, a standard statistical learning method termed k-Nearest Neighbours (k-NN) (Duda et al., 2012), a method based on Decision Trees (C4.5) (Quinlan, 1993) and a more complex structure, based on neural networks, known as Learning Vector Quantisation (LVQ) (Kohonen, 2001). The selection of these three algorithms was driven by the need to test the effectiveness of different types of algorithms on the analysis of archaeological data in an effort to exploit different artefact attributes. Despite our selection of classification methods, the proposed design may be realised in combination with any classification method.

The deployment of established methods allows the evaluation of the validity of the results through the use of a special form of cross validation testing. The developed design follows a systematic approach and well-established methods, such as bootstrapping with replacement (Efron and Tibshirani, 1986) and the 5×2 cross validation (paired t-test and F-test) tests in order to ensure that the results are statistically significant. The proposed scheme is tested with the use of a sample of Early and Middle Bronze Age utilitarian pottery from Cyprus. The statistical experiment involved two analytical datasets deriving from the mineralogical and chemical characterisation of 177 ceramic samples, with the respective employment of petrography on ceramic thin sections and ED-XRF on pressed-powder pellets (Dikomitou, 2012).

1.1. Archaeometric analysis of ceramic data

Archaeological data is often characterised as complex data due to the large number of involved influencing factors during the analysis procedures. Much attention is given upon the gathering of the archaeological artefacts during excavations and their subsequent micro-structural, mineralogical and/or chemical analysis. However, many parameters influence the reliability of the produced data. Different archaeologists implement the same procedures in different ways, thereby increasing the within-class variance. This problem intensifies by taking into account that apart from variations generated due to the human factor, the acquired variability is also caused due to the deterioration of the source material because of its natural ageing as well as the environment of deposition.

The analysis of archaeological data is not a straightforward task. The aim of the archaeologist is to make inferences by taking into consideration as many parameters as possible. Ceramic classification remains the principal approach to the study of pottery in identifying patterns in the data. The most common way to categorise pottery is primarily based on the macroscopic observation of technological attributes and morphological types; extra attention is given to the shape, size and surface treatment. An alternative to this method is the characterisation based on their chemical composition by isolating ceramic ground of similar chemical profiles and statistically testing the validity of those groups (Garcia-Heras et al., 2001). The use of techniques achieving a clearer separation in different groups also results in increasing their interpretability. The validity of the emerging groupings can be further evaluated through typological and potentially mineralogical comparisons with data bearing known fingerprints, so as to address different aspects of ancient ceramic production and distribution.

2. Classification of chemical compositional archaeological data

2.1. Chemical compositional analysis of ceramics in archaeology

Chemical compositional data are defined as vectors of strictly positive components, usually expressed as percentages or parts-per-million (ppm), with constant sum — a restriction not always maintained. Quantitative chemical analysis is not involved in measuring, but in enumerating, or counting, the number of each type of atoms in a sample (Buxeda, 2008). Chemical compositional data do not vary independently and concentration based approaches to data analysis can lead to misleading conclusions (Reimann et al., 2012).

Chemical compositional data lay in the constrained Simplex Space (Aitchison et al., 1982; Buxeda, 2008), where correlation analysis and the Euclidean distance are not mathematically meaningful concepts (Reimann et al., 2012). Furthermore, graphical depiction of raw or log-transformed data should only be used in an exploratory data analysis sense, to detect unusual data behaviour or candidate subgroups of samples (Acton, 2013).

The chemical constituents of an archaeological artefact, or any other object, can be categorised into major and trace elements. Major elements comprise large proportions of the artefact under analysis, while trace elements are present in concentrations less than 0.01%. As ceramics are heterogeneous in composition with the majority of their major elements present in most artefacts, the discrimination of objects into groups makes necessary the utilisation of trace elements in determining the fingerprint of a deposit (Mirti et al., 1994).

2.2. The classification problem

Classification is a procedure that aims to assign items to a number of (possibly pre-known) target categories or classes based on statistical/machine learning principles (Bishop, 2006). It is an instance of supervised/unsupervised learning, whereby the former assumes that a training set of correctly identified observations is available (Bishop, 2006). Classification of archaeological ceramics deals with the categorisation of ceramic specimens of similar chemical profiles, given a number of artefacts of known fabric identify.

The supervised classification problem may be broken down into two separate stages: 1) the inference stage where training data is used to learn a model and 2) the decision stage in which the trained model is

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010

used to make class assignments (Bishop, 2006). The classifier will only assign data to the provided class labels, and therefore good classification requires a sample representative to the population of the data. Classification in archaeology is very important since it makes possible the identification of a newly found artefact based on already known information. Suppose we have an input vector $\boldsymbol{x}$ together with a corresponding vector t of target variables (the class labels). The goal is to predict t given a new value for x (Bishop, 2006). The individual observations x are analysed into a set of quantifiable features (chemical components in the case of compositional data). The goal of classification is to accurately predict the target class for any data vector $\boldsymbol{x}$.

An operational definition of classification of archaeological chemical data can be stated as follows: given a set of n archaeological artefacts and a vector t indicating the label of each artefact find a model which successfully assigns new samples to the appropriate class. Consider a set of n archaeological observations $\boldsymbol{O}=\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n\}$. The analysis of the actual-tangible artefact o_i with the use of ED-XRF or any other chemical method of ancient pottery analysis will produce the chemical representation of o_i which has the form $\boldsymbol{x_i}\in[0,1]^p$ where p is the number of analysed chemical elements and vector x_i consists the chemical compositions of the artefact. Therefore the set $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x_n}\}$ represents the dataset of the artefacts' chemical composition. We assume that there are groups (subsets) of similar artefacts in O, the class of which is determined by the labels in $\boldsymbol{T}=\{\boldsymbol{t}_1,\ldots,\boldsymbol{t_n}\}$, and $t_i\in J$ where J denotes the possible (known) class labels. Each artefact is represented in the dataset with the pair $\boldsymbol{o_i}=(\boldsymbol{x_i},\boldsymbol{t_i})$, during parameter training the parameters Θ of the classifier are obtained by $\Theta=g(O)$, and the class (t_i) of an uncategorised set of artefacts $\boldsymbol{X}=\{\boldsymbol{x_{n+1}},\ldots,\boldsymbol{x_{n+l}}\}$ can be obtained by $t_i=f(x_i,\Theta)\,\forall\,\mathrm{n}<i\leq n+l$.

2.3. The archaeological dataset

This statistical study involves the compositional analysis of a small elemental dataset obtained from ED-XRF analysis (carried with a Spectro X-Lab 2000) of Early and Middle Bronze Age ceramics in the form of pressed-powder pellets from Cyprus (Dikomitou, 2012). The archaeological samples derive from the successive occupational phases of the Early and Middle Bronze Age settlement of Marki *Alonia* (2400–1700 BC, Frankel and Webb, 1996, 2006) in central Cyprus, and include the two predominant wares recorded at the site, i.e. Red Polished Philia[1] pottery from the first occupational phases of the settlement (c. 2400–2200 BC) and Red Polished pottery from the Early and Middle Bronze Age (c. 2300–1700 BC). Red Polished Philia pottery was also selected from other contemporary sites across Cyprus, in order to assess the degree of compositional and technological homogeneity among pottery assemblages that exhibit a significant degree of stylistic uniformity across the island. Therefore, the final 177 samples under study come from eight different sites across the island (Dikomitou, 2012). Their analytical study aimed at their compositional and technological characterisation in order to assess ceramic production, distribution and social interaction in Early and Middle Bronze Age Cyprus (Dikomitou, 2012; Dikomitou-Eliadou, 2013, 2014).

The samples were divided into two datasets. The first dataset includes the Red Polished Philia samples from various Philia sites, while the second dataset includes all the samples from Marki *Alonia*, including both the Red Polished Philia and Red Polished samples from the settlement. The statistical experiment was particularly challenging due to its small size (177 samples), as well as due to the relatively large number of one-member classes (21 samples), which could not be included in one of the predefined mineralogical groups, identified with the employment of petrography, either because they lack discriminating petrology, or because their fabric was dissimilar to those of the clustered samples. The remaining samples were labelled into 15 fabric groups (Marki fabrics I–XIII and Philia fabrics I and IV). Considering the fact that the identification of one-member classes is equally important for the assessment of ceramic compositional and technological variability, adding to fabric variability, all one-member classes were also included in the experiment. Each one-member class formed a separate standalone class, resulting in a total of 36 different fabric classes. The consideration of one-member classes served a twofold purpose: 1. To test the robustness of classification on complicated and highly overlapping data, and 2. To assess whether post-classification analysis could allow one-member class categorisation to one of the predefined fabric groups (a task that could not be solved with certainty beforehand due to the absence of discriminating petrology in the ceramic thin sections). Finally, the statistical experiment was conducted in order to explore other methods of statistical analysis that are not yet widely known in the field of archaeological sciences and investigate the relations among fabric groups within the two datasets that are suggested by petrography to be identical or very similar, with the ultimate objective to test the correspondence between the mineralogical and chemical compositional data.

The dataset was pre-processed as follows before statistical analysis: All elements were converted into oxide compounds with stoichiometry, the composition of each pottery artefact was normalised (i.e. force the sum of each row to be 100). Also, it is typical with archaeological data to exclude certain features upon processing. Trace elements with elemental concentration below 10 ppm were omitted along with sulphur trioxide (SO_3), chlorine (ClO) and lead oxide (PbO) concentrations due to analytical reasons. Sodium oxide (Na_2O), phosphorus pentoxide (P_2O_5), cobalt (Co_3O_4) and cerium oxides (CeO_2) were also omitted from multivariate statistics due to inconsistencies in their values and poor reproducibility in successive analytical runs (Dikomitou, 2012). The chemical compounds used for analysis are: MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, V_2O_5, Cr_2O_3, MnO, Fe_2O_3, NiO, CuO, ZnO, Ga_2O_3, Rb_2O, SrO, Y_2O_3, ZrO_2, and BaO.

3. Hypothesis testing

It should be clarified that the two different datasets, the "Marki" and "Philia", were studied in the same period of time and by the same researcher, therefore they are characterised by a high degree of analytical consistency, as all samples were first collected and then analysed by the same person in the same laboratory, using the exact same procedures. The combination of the data into a single dataset could, however, reveal compositional and/or technological relationships between types/classes of ceramics, as well as links emerging either due to their context of production or recovery and/or their technology of production.

The null hypothesis behind the classification problem stated that the classification algorithms, k-Nearest Neighbour, C4.5 (based on Decision Trees) and Learning Vector Quantization (LVQ Networks) perform equally well for the dataset of interest when the performance is assessed with the classification accuracy. Upon rejection of the null hypothesis, the alternative hypothesis tests whether any of the three algorithms outperforms the other using pairwise comparisons.

Each of the three selected algorithms operates based on different principles, each corresponding to a different category of statistical/machine learning approaches. Doing so allows solving the classification problem from different perspectives, hoping that the differences in the results may disclose new information about the data.

4. The experimental design

There is no universal solution for every problem. The design of the methodology to tackle a problem has to serve its specific needs while considering limitations and constraints. The choice of deployed methods accounts for the non-uniformity in the distribution of classes, as well as the size of the dataset, which is relatively small compared to the number of classes.

[1] The so-called Philia phase is a transitional cultural phase between the Chalcolithic and the Early Bronze Age in Cyprus (*c.f.* Webb and Frankel, 1999).

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010

Classification of small datasets (with regard to the number of classes) requires the use of resampling methods. For the needs of this study, bootstrapping with replacement is used to allow the generation of datasets of 177 samples. The choice of resampling adds the assumption that the dataset is representative to the population, in other words, the sample of collected ceramics yields a good representation of the ceramic population at the sampled sites, during the specified time period.

Moreover, the performance of the algorithms is evaluated using the classification accuracy, i.e. the number of correct predictions from all predictions (Sokolova and Lapalme, 2009). The calculation of the accuracy for a problem with more than two classes also accounts the instances in which the classifier correctly decides not to assign an artefact to those fabrics, to which it does not belong. The classification result is also evaluated using the Jaccard index – an external cluster validity index – which is calculated as the number of correctly classified samples over the number of samples that exist either in the true or estimated classification (Tan et al., 2005); this index does not account for instances in which the algorithm correctly did not assign a sample to a specific class. The calculation of both indices is done in an effort to assess their robustness and emphasise the significance of their choice.

It is not expected that an algorithm will perform equally well with any dataset. The results are highly dependent on the parameterisation of the algorithm, as well as the structure and complexity of the data. Since the evaluation of classification performance forms part of the question, it was necessary to consider some fine-tuning of each algorithm's parameters for each bootstrapped dataset. However, the operation of the three algorithms is not the same and the fine-tuning of each algorithm's parameters cannot be assumed to be identical. For instance, the LVQ algorithm requires the parameterisation of more parameters than the other two and its performance is more likely to suffer.

4.1. k-NN

The k-NN algorithm is a non-parametric approach used for classification, operating in the belief that a sample will more likely belong to the class of its closest already classified samples (Duda et al., 2012). k-NN is among the simplest and most intuitive machine learning algorithms. For each uncategorised artefact, its distance to all classified samples is measured, the k closest samples are selected and the artefact is categorised to the class to which most of its k neighbours belong; k is a positive integer. If $k = 1$, the object is simply assigned to the class of its immediate nearest neighbour, according to some distance metric. The input consists of the k closest training examples in the feature space and the output is a class membership. The valid implementation of the algorithm for compositional data in the Simplex space requires measuring the distance with the Aitchison distance metric (Eq. (1)) $d(x,y)$, where x and y are points in the Simplex space (Aitchison, 1986).

$$d(x,y) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2} \quad (1)$$

4.2. C4.5 algorithm

The C4.5 is an extension of the earlier ID3 algorithm and performs classification by generating a decision tree. Decision trees are generated incrementally by breaking down a dataset into smaller and smaller subsets. C4.5 builds decision trees from a set of training data, using the concept of information entropy. At each tree node, the chemical element/feature that most effectively splits the dataset into subsets is selected while the attribute with the highest normalised information gain is chosen to make the decision. The idea is to refine T (the tree) into subsets of samples that are heading towards one-member class collections of samples. An appropriate test is chosen, based on a single element that has one or more mutually exclusive outcomes (Thakur and Mann, 2014). The decision tree T consists of a decision node identifying the test and one branch for each possible outcome, an example of which is given in Fig. 1. The C4.5 algorithm then recurs on the smaller sub-lists. The decision trees generated by C4.5 can be used for classification.

4.3. LVQ

Learning Vector Quantization is a special case of an artificial neural network which deploys the winner takes it all learning-based approach. An LVQ system is represented by prototypes $W = \{w_1, \ldots, w_m\}$ where m is the number of classes defined in the feature space of observed data. Algorithms based on this approach, assign each data sample to the label of the prototype that is closest to it, according to a given distance measure. The position of this so-called winner prototype is then adapted, i.e. the winner is moved closer if it correctly classifies the data point or moved away if it classifies the data point incorrectly.

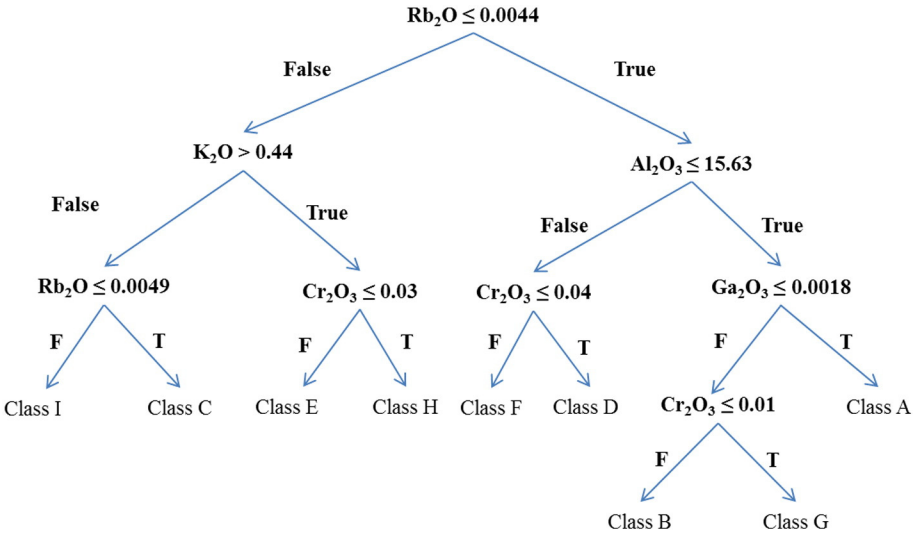

Fig. 1. Decision Tree example based on the values of five chemical elements. For example, if the concentration of Rb_2O is less than or equal to 0.0044 and AL_2O_3 is greater than 15.63 and Cr_2O_3 is greater than 0.01 then the artefact is assigned to Class B.

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010

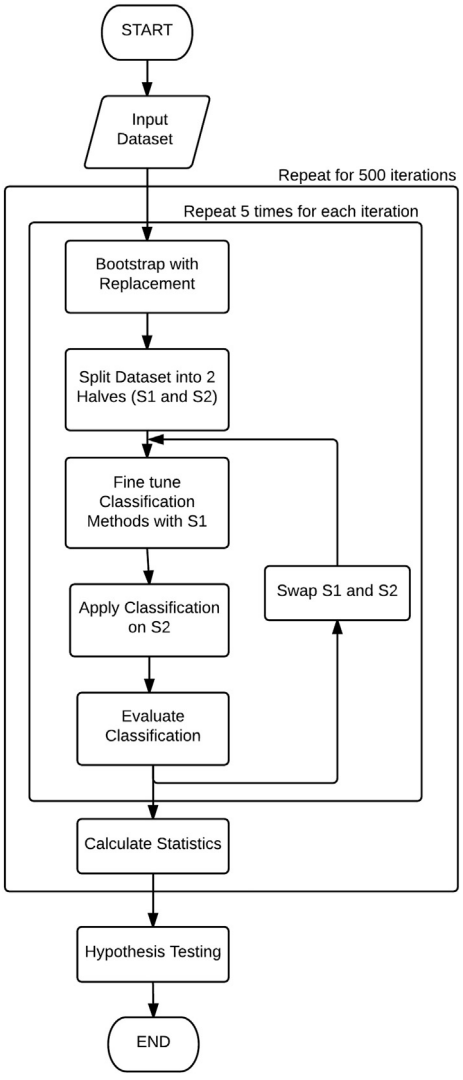

Fig. 2. Diagrammatic representation of the experimental design; any classification method that seems to be appropriate may be used with this design.

An advantage of LVQ is that it creates prototypes that are easy to interpret. A trained LVQ network allows the visualisation of the map of prototypes which gives insight on which classes are closer to others. In archaeological context this may provide information as of how "close" fabrics are to each other. Even though the algorithm does not restrict the dimensionality of the map, it is usually implemented as a 2-dimentional map.

4.4. Fine tuning of algorithms

The deployment of the discussed algorithms for each bootstrapped dataset required some fine-tuning of the algorithms. However, exhaustive fine-tuning might easily become very time consuming especially when a large number of re-sampled datasets need to be processed.

The set of categorised samples is used for the parameterisation of the algorithm and it is further divided into 2 smaller sets named the training tuning and testing tuning sets. The parameters maximising the classification performance of the algorithm on the testing set are selected for processing the bootstrapped dataset. Considering the very restrictive size of the training and testing tuning datasets, it is expected that the selected parameters might not be the optimal – since insufficient amounts of data may not allow enough training (Duda et al., 2012).

The C4.5 algorithm did not become subject to parameterisation or pruning. Pruning is a way of reducing the size of the decision tree, for this experiment, the parameter determining the pruning stage was set to 0 corresponding to no pruning. The k-NN method only requires the specification of the parameter k. For each dataset the algorithm was tested for integer values of $k = 1$ to 10; 96% of the time the values of k maximising the classification accuracy were between 1 and 4, with $k = 1$ being the most frequently occurring value scoring 73%. LVQ required more time consuming training; the configuration of the network requires specifying the learning rate and map size in each dimension; for convenience this study was limited to 2 dimensional square maps — the most common implementation (Kohonen, 2001). The tuning of an LVQ network has a significant computational cost, and due to this the parameterisation was limited to a small range of possible values.

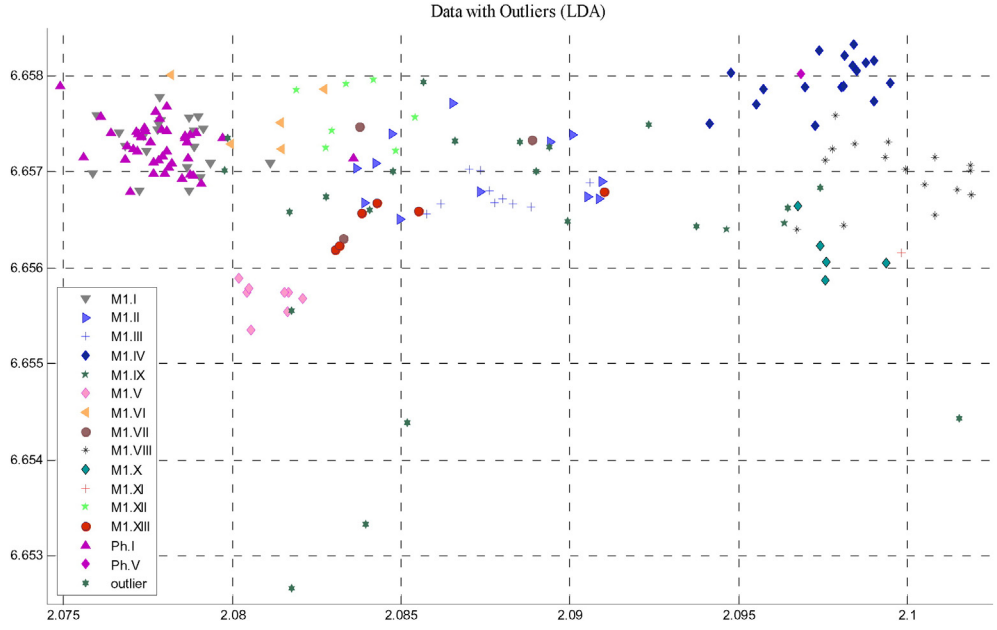

Fig. 3. Two dimensional plot of the sample generated with LDA. The fabrics are not distinct in the state space, many classes overlap.

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010

Table 1
The estimated accuracy and Jaccard index of each algorithm. The scores represent the mean score of all iterations.

	Classification accuracy (%)			Jaccard index (%)		
	Mean	Max	Min	Mean	Max	Min
k-NN	72.1	79.4	64.2	56.7	70.1	42.7
C4.5	68.5	77.2	61.7	49.1	63.7	38
LVQ	55.8	65.2	46.2	30.3	38.8	21

4.5. Statistical testing

The 5 × 2 cross validation paired *t*-test (Dietterich, 1998) and the 5 × 2 cross validation *F*-test (Alpayd, 1999) were deployed to statistically test the significance of the classification results and to evaluate their robustness. Benchmarks on significance testing propose that cross validation testing methods are more robust when dealing with small datasets where reproducibility of the experiment is not an issue (Dietterich, 1998). The 5 × 2 cross validation method was selected to allow large enough datasets for testing while ensuring that no further dependencies of overlapping training and testing sets are introduced when cross validation is used (Salzberg, 1997).

The diagrammatic representation of the designed scheme is shown in Fig. 2. The experiment was allowed to run for 500 iterations to allow the generation of valid statistics and the significance of the results was calculated at level 0.05 ensuring that a 95% confidence level for the results of statistical testing.

5. Results and discussion

Fig. 3 shows a plot obtained by Linear Discriminant Analysis (a dimensionality reduction method) (Bishop, 2006) of the original dataset. Many classes are overlapping and the discrimination of classes is not trivial. Table 1 shows the performance, in terms of classification accuracy and the Jaccard index. The scores are calculated as the mean scores of all iterations. As expected, the values scored measuring the classification accuracy are higher than the values scored by the Jaccard index. In archaeological ancient pottery analysis, it is important to measure a classifier's robustness in assigning artefacts to the correct fabric. Since classification accuracy accounts the instances in which the classifier correctly does not assign an artefact to fabrics to which it does not belong, it served better the needs of the problem compared to the Jaccard coefficient.

The scores suggest that the k-NN algorithm consistently scores the highest compared to the other algorithms. Significance testing between the algorithms, in a pairwise fashion, has illustrated that the performance of k-NN and C4.5 outperformed the LVQ Network, while the Null hypothesis between the k-NN and C4.5 (corresponding to non-significantly different performance) is accepted. In other words, k-NN and C4.5 score better than LVQ, however, k-NN is not significantly better than C4.5. It is also important to note that both statistical tests, the 5 × 2 cv paired t-test and the 5 × 2 cv F-test, confirmed the same hypothesis results.

The experiment does not show that LVQ performs worse; it is rather shown that its implementation with very limited fine-tuning and the examined dataset, results in lower performance compared to the other classification methods. LVQ is admittedly a more complex algorithm and its parameterisation needs to be handled with care, especially when dealing with datasets of very limited size containing a large number of classes. The potentially poor selection of parameters during fine-tuning may hinder the classification results; something that became apparent in the performance of the LVQ. When one deploys any classification approach, they should take into account the operation of the algorithm and its appropriate configuration.

The results of classification can be useful in verifying the initial distinction of the samples into fabric groups and one-member classes. Most importantly, further analysis on the classification results has shown that classification may provide more information to the archaeologist. During each iteration, a matrix with the correctly and wrongly classified artefacts for each class (i.e. the confusion matrix) is generated. Systematic study of the misclassified artefacts has shown that some elements are not misclassified randomly. For instance, the misclassified elements of class M1.II were (or would be assigned) to class M1.III; the same holds for all classes shown in Table 2. Indeed the comparative study of the fabric groups suggested by petrography and the highlighted elemental relations suggested by our statistical experiment confirmed that fabrics M1.I and Ph.I are essentially the same fabric distributed across Cyprus during the Philia cultural phase, including the settlement at Marki (Dikomitou, 2012; Dikomitou-Eliadou, 2013, 2014), while M1.II and M1.III share many mineralogical characteristics, having been made with raw materials deriving from a similar geological environment (Dikomitou, 2012). Interestingly, the algorithms proposed a link between the two most igneous fabrics, i.e. M1.VIII and M1.IV from Marki. Actually the two fabrics were used for the production of cooking pots and other utilitarian pottery throughout the lifespan of the settlement at Marki, and fabric M1.VIII actually replaced M1.IV after some period of co-existence.

Another important finding is that the analysis of misclassified artefacts suggested a possible class (fabric group) to some one-member class samples. This analysis might become useful for the classification of samples of unknown class with a degree of confidence (analogous to the classification accuracy). The result of this process will return a possible class for each of these samples which when interpreted may lead to their definite categorisation.

The inability of the algorithm to classify the specimen to their true class reveals possible relationships between certain classes. Furthermore, it is not anticipated that all misclassified artefacts imply correlations between classes; it is expected that classifiers may yield non-zero error rates. The specimens were analysed for a number of chemical elements, some of which in very small concentrations (<0.1%). The heterogeneous composition of ceramics needs to be accounted during classification. Trace elements may concur more characteristically in determining the fingerprint of a deposit (*Mirti et al.*, 1994), making important the evaluation of their discriminating abilities. Due to this, the experiment as discussed previously was repeated *two* more times, using the chemical elements with mean concentration > 0.1% (MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, MnO, Fe_2O_3, BaO) and using only the chemical elements with mean concentration < 0.1% (V_2O_5, Cr_2O_3, NiO, CuO, ZnO, Ga_2O_3, Rb_2O, SrO, Y_2O_3, ZrO_2); in both cases the data rows were normalised to sum 100. The results of the experiments bring to our attention some interesting data properties summarised in Table 3. The exclusive use of elements with mean concentration > 0.1% allows the equivalent discrimination of artefacts when using all available information (see Table 1). Table 3 also shows that despite not utilising 99.8% of the measured information (when <0.1%), the majority of characteristics

Table 2
Each table column shows potential relationship between Class 1 and Class 2. Upon the classification of samples belonging to Class 1 if they are unsuccessfully classified they would more likely be allocated to Class 2. An OMC sample is essentially a one-member class.

Class 1	M1.I	M1.II	M1.II	M1.II	M1.III	M1.IV	M1.IV	M1.V	M1.X	M1.XII	M1.XII	Ph.I
Class 2	Ph.I	M1.III	M1.VII	M1.OMC10	M1.II	M1.VIII	M1.OMC17	M1.XIII	M1.VIII	M1.VII	M1.OMC6	M1.I
Sample ID				14604			16513				12372	

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010

Table 3
Classification accuracy and Jaccard index scores when classification is performed on elements with mean concentration > 0.1% and <0.1%.

	Elements used	Classification accuracy (%)			Jaccard index (%)		
		Mean	Max	Min	Mean	Max	Min
KNN	>0.1%	73.2	79.5	67.6	52.9	64.8	40.5
	<0.1%	66.6	75	58.2	47.8	57.7	35.7
C4.5	>0.1%	67.8	75.1	62.7	43.2	56.8	32.8
	<0.1%	64.5	71.2	55.8	46.1	58.1	35.6
LVQ	>0.1%	57.3	67	48.8	29.3	36.6	20.9
	<0.1%	59.1	66.2	51.4	40.2	52.2	26.1

that allow the discrimination of the specimen into their categories is maintained. This finding allows us to hypothesise that the use of trace elements during classification needs to be studied further.

6. Conclusions and future work

The analysis of archaeological ceramic artefacts through means of classification may assist the recognition of compositional, technological, stylistic, and even social patterns, and the identification of possible categorisation mistakes. This paper stresses the significance of using systematic and statistically valid methods, through the development of an analysis procedure, to allow the deployment of multiple classification methods in order to answer archaeological questions. It has also been illustrated that the inference of classification results through the generation of confusion matrices allows the exploitation of inter-class relationships and the ability to provide extra information to the expert for the categorisation of samples that could not be grouped (with certainty) into one of the classes.

The classification accuracy was used to evaluate the performance of the algorithms while the Jaccard index was also calculated in effort to observe to which degree the two indices are in accordance. Over the years, a wide range of classification evaluation metrics have been developed with the accuracy being one of the most widely used ones. However, further investigation is required so that the appropriate metric, taking into consideration commonly occurring issues in archaeological data, is found, for use with multi-class data. Part of our future work is the evaluation of performance while taking into consideration that certain classes are highly correlated.

Acknowledgements

This project is supported by the European Union under the 7th Framework Programme “FP7-PEOPLE-2010-ITN”. Grant agreement number 265010 — “New Archaeological Research Network for Integrating Approaches to Ancient Material” (NARNIA-ITN).

References

Acton, A., 2013. Issues in Environmental Research and Application: 2013 Edition. ScholarlyEditions.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London–New York.
Aitchison, J., Society, S., Methodological, S.B., 1982. 44 (2), 139–177.
Alpayd, E., 1999. Combined 5 × 2 cv F test for comparing supervised. Neural Comput. 1892 (8), 1885–1892.
Barceló, J.A., 2008. Computational Intelligence in Archaeology, Henshey (NY), Information.
Baxter, M.J., 1994. Exploratory Multivariate Analysis in Archaeology. Edinburgh University Press, Edinburgh.
Baxter, M.J., 2006. A review of supervised and unsupervised pattern recognition in archaeometry. Archaeometry 48, 671–694.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning 4. Springer, New York.
Buxeda, J., 2008. Revisiting the compositional data. Some Fundamental Questions and New Prospects in Archaeometry and Archaeology.
Coolican, H., 1999. Research Methods and Statistics in Psychology. Hodder & Stoughton.
Dietterich, T., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.
Dikomitou, M., 2012. Ceramic production, distribution, and social interaction. University College London, An Analytical Approach to the Study of Early and Middle Bronze Age Pottery from Cyprus.
Dikomitou-Eliadou, M., 2014. Rescaling perspectives: local and island-wide ceramic production in Early and Middle Bronze Age Cyprus. In: Webb, J.M. (Ed.), Structure, Measurement and Meaning: Studies on Prehistoric Cyprus in Honour of David FrankelStudies in Mediterranean Archaeology. Åströms Förlag, Uppsala, pp. 199–211.
Dikomitou-Eliadou, M., 2013. Interactive communities at the dawn of the Cypriot Bronze Age: an interdisciplinary approach to Philia phase ceramic variability. In: Knapp, A.B., Webb, J.M., McCarthy, A., Stewart, J.R.B. (Eds.), An Archaeological LegacyStudies in Mediterranean Archaeology CXXXIX. Åströms Förlag, Uppsala, pp. 23–31.
Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1, 54–75.
Fermo, P., Delnevo, E., Lasagni, M., Polla, S., de Vos, M., 2008. Application of chemical and chemometric analytical techniques to the study of ancient ceramics from Dougga (Tunisia). Microchem. J. 88, 150–159.
Frankel, D., Webb, J.M., 1996. Marki Alonia: An Early and Middle Bronze Age town in Cyprus: Excavations 1990–1994. Studies in Mediterranean Archaeology CXXIII: 1. Åströms Förlag, Jonsered.
Frankel, D., Webb, J.M., 2006. Marki Alonia. An Early and Middle Bronze Age Settlement in Cyprus. Excavations 1995–2000. Studies in Mediterranean Archaeology CXXIII: 2. Åströms Förlag, Sävedalen.
Garcia-Heras, M., Blackman, M.J., Fernandez-Ruiz, R., Bishop, R.L., 2001. Assessing ceramic compositional data: a comparison of total reflection X-ray fluorescence and instrumental neutron activation analysis on Late Iron Age Spanish Celtiberian ceramics. Archaeometry 43 (3), 325–347.
Kohonen, T., 2001. Self-Organizing Maps, 3rd ed. 30. Springer Series in Information Sciences.
Kowalski, B.R., Schatzki, T.F., Stross, F.H., 1972. Classification of archaeological artifacts by applying pattern recognition to trace element data. Anal. Chem. 44, 2176–2180.
Kowalski, B.R., Bender, C.F., 1973. Pattern recognition. II. Linear and nonlinear methods for displaying chemical data. J. Am. Chem. Soc. 95 (3), 686–693.
Lopez-Molinero, A., Castro, A., Pino, J., Perez-Arantegui, J., Castillo, J.R., 2000. Classification of ancient roman glazed ceramics using the neural network of self-organizing maps. Fresenius J. Anal. Chem. 367, 586–589.
Mirti, P., Aruga, R., Appolonia, L., Casoli, A., Oddone, M., 1994. On the role of major, minor and trace elements in provenancing ceramic material. A case study: Roman terra sigillata from Augusta Praetoria. Fresenius J. Anal. Chem. 348, 396–401.
Mussumarra, G., Stella, M., Matteini, M., Rizzi, M., 1995. Multivariate characterization, using the SIMCA method, of mortars from two frescoes in Chiaravalle Abbey. Thermochim. Acta 269 (270), 797–807.
Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco.
Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., The GEMAS Project Team, 2012. The concept of compositional data analysis in practice—total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210.
Salzberg, S., 1997. On comparing classifiers : pitfalls to avoid and a recommended approach. Data Min. Knowl. Disc. 1 (3), 317–328.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437.
Tan, P.N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley.
Thakur, B., Mann, M., 2014. Data mining with big data using C4.5 and Bayesian classifier. International Journal of Advanced Research in Computer Science and Software Engineering 4 (8), 959–962.
Webb, J., Frankel, D., 1999. Characterising the Philia facies: material culture, chronology, and the origin of the bronze age in Cyprus. Am. J. Archaeol. 103, 3–43.

Please cite this article as: Charalambous, E., et al., An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relations..., Journal of Archaeological Science: Reports (2015), http://dx.doi.org/10.1016/j.jasrep.2015.08.010