Multi-classifier majority voting analyses in provenance studies on iron artefacts

Grzegorz Żabiński, Jaroslaw Gramacki, Artur Gramacki, Ewelina Miśta-Jakubowska, Thomas Birch, Alexandre Disser

Journal of Archaeological Science 113 (2020) 105055

Contents lists available at ScienceDirect

Journal of Archaeological Science

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

Multi-classifier majority voting analyses in provenance studies on iron artefacts

Grzegorz Żabiński [a,*], Jarosław Gramacki [b], Artur Gramacki [c], Ewelina Miśta-Jakubowska [d], Thomas Birch [e], Alexandre Disser [f]

[a] Institute of History, Jan Długosz University in Czestochowa, Poland
[b] Computer Center, University of Zielona Góra, Poland
[c] Institute of Control and Computation Engineering,University of Zielona Góra, Poland
[d] National Centre for Nuclear Research, Świerk, Otwock, Poland
[e] School of Culture and Society - Centre for Urban Network Evolutions, Aarhus University, Denmark
[f] Laboratoire Métallurgies et Cultures, IRAMAT UMR 5060 CNRS, 90010, Belfort, France

ARTICLE INFO

Keywords:
Archaeological iron
History of metallurgy
Provenance studies
Slag inclusions
Multivariate statistics
Classification

ABSTRACT

The main objective of this paper is to propose an approach for identification of provenance of archaeological iron artefacts making use of major oxides and trace elements. For this purpose, seven classifiers were built on the basis of the following techniques: Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forests (RF), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Recursive Partitioning and Regression Trees (RPART) and Kernel Discriminant Analysis (KDA). A final assignment of a given observation to a regional class was carried out on the basis of results provided by all classifiers using a majority voting technique. The proposed approach was first tested on experimental slag and then it was applied to actual archaeological data. It is hoped that this method can become part of a new integrated approach which will consider all available types of data, such as major and trace elements and isotopic ratios.

1. Introduction

The aim of this paper is to offer a new approach to the problem of provenance of archaeological iron on the basis of major oxides and trace elements. A significant novelty of the proposed method consists in the fact that it combines several different classification algorithms (parametric and non-parametric ones). Then, the final result is produced using a majority voting technique. Individual classifiers which are employed in this study have their advantages and disadvantages (see below) and they may perform differently on different data assemblages. It has therefore been assumed that a good way to obtain a credible classification it to first aggregate partial results provided by each method and then combine it into a final provenance assessment.

The proposed approach was first tested on experimental smelting data (six smelting experiments altogether, in which both identical and very different ores were used). In this way, strengths and weaknesses of each classification method could be demonstrated and assessed. As the final classification produced by the majority voting technique was reasonable, the approach was applied to actual archaeological data that has been previously analysed in Disser et al. (2017). It turned out again that individual classifiers produced results that in some cases may be at variance with each other. This seems to reinforce a claim that a method that allows to aggregate them and then to propose a final majority-voted classification is recommended.

In 2012 Charlton et al. proposed a method of identification of smelting slag inclusions in iron artefacts with the use of multivariate statistics. The identification model uses $-log$ transformed subcompositional ratios of six oxides: MgO, Al_2O_3, SiO_2, K_2O, CaO and TiO_2. The first step is a Principal Component Analysis (PCA) and Agglomerative Hierarchical Clustering (AHC, preferably Euclidean distance and average linkage agglomeration) in order to identify groups of slag inclusions of different origin (Charlton et al., 2012, pp. 2281–2283). In order to propose a provenance of artefacts, series of training sets are obtained by means of analyses of large samples of smelting slag from known locations. Data on the chemical composition of smelting slag ($-log$ subcompositional ratios of MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2 and

* Corresponding author.
E-mail addresses: g.zabinski@ujd.edu.pl (G. Żabiński), j.gramacki@ck.uz.zgora.pl (J. Gramacki), a.gramacki@issi.uz.zgora.pl (A. Gramacki), Ewelina.Mista@ncbj.gov.pl (E. Miśta-Jakubowska), t.birch@cas.au.dk (T. Birch), disser.alex@gmail.com (A. Disser).

https://doi.org/10.1016/j.jas.2019.105055
Received 3 June 2019; Received in revised form 10 November 2019; Accepted 19 November 2019
Available online 17 December 2019

MnO) is processed using PCA, Linear Discriminant Analysis (LDA) and Kernel Density Estimation (KDE) in order to define boundaries of training set fields. Then, data concerning the same oxides from smelting-derived slag inclusions in artefacts is projected with the use of PCA, LDA and KDE onto the fields defined by the chemistry of the training sets (Charlton et al., 2012, pp. 2284–2289). This approach was first tested on experimental smelting slag, and then these researchers applied it to archaeological slag data provided by Buchwald (2005). They maintained that they had obtained a reasonable separation between individual regions (Charlton et al., 2012, pp. 2289–2291); for earlier attempts at provenancing with the use of oxide ratios in slag see, e.g. Blakelock et al. (2009); Buchwald (2005); Buchwald and Wivel (1998).

Charlton et al. also raised several reservations concerning the use of this analytical approach. Identifications of provenance sources must not be considered in absolute sense, but should rather be seen as hypotheses which can be verified with other methods. Furthermore, training sets do not always contain actual sources of metal in examined artefacts and sometimes it is not possible to sufficiently discriminate between training sets. Another problem may be posed by the lack of comparative data on iron sources (Charlton et al., 2012, pp. 2290–2292); see also Charlton et al. (2013) for an overview of provenancing methods). One possible way of coping with these shortcomings is a combination of the major oxide analysis with examinations of other minor and trace elements (Charlton et al., 2012, p. 2291).

In the recent years it has become increasingly clear that major elements alone are not sufficient for provenance studies unless there are significant differences between iron ores. Analyses with the use of various sets of major and trace elements, sometimes combined with Pb isotopic ratios were carried out by several researchers (Serneels, 1995; Schwab et al., 2006; Coustures et al. 2003, 2006; Desaulty et al., 2008). In some cases, test sets were first filtered with regard to the contents of MnO and P_2O_5) (Leroy et al., 2012; Pryce et al., 2014). L'Heritier et al. (2016, pp. 213–230) discussed a method of quantification of major oxides and trace elements in slag inclusions with the use of Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS). In their multivariate analysis of SI from medieval iron bars from a shipwreck, Birch and Martinón-Torres highlight the importance of relating SI chemistry to the microstructures (i.e., steel or iron) in which they are embedded. They also highlight the issue of the dilution effect on multivariate statistical methods, imparted by dominant compounds such as FeO (Birch and Martinón-Torres (2015)).

A study on construction iron from the Carolingian bridge in Dieulouard in France made use of both major and trace elements in order to define the chemical signature of slag inclusions. After the identification of 'zones of interest' in the metal using metallography, smelting slag inclusions were isolated and the contents of the aforementioned elements were measured. Furthermore, a two-step PCA and AHC analysis was implemented in order to compare the chemical signatures of the examined artefacts to those of ores and smelting slags from Lorraine. However, a match was obtained in no case (Disser et al., 2016, pp. 149–159). A similar approach was applied by Disser et al. to study the provenance of iron for the construction of Metz Cathedral in France (Disser et al. (2017); see also Bauvais et al. (2017)).

A provenance study carried out by Dillmann et al. (2017) on Early Iron Age iron artefacts from the North Alpine region made use of previous developments, combined with the analyses of osmium isotopic ratios in the metallic matrix (as proposed by Brauns et al. (2013)). It was found out that many ores varied significantly with regard to their $^{187}Os/^{188}Os$ ratios, although some overlapping could also be seen. Another important aspect of this study was that both major and trace elements were used for provenance studies. Their contents for production areas (ores and slag) and slag inclusions in artefacts are first separately normalised using a log-ratio transformation. The results are processed in a multi-stage PCA-AHC approach. Observations are then compared with the results of isotopic analyses (Dillmann et al., 2017, pp. 108–115). The final outcomes demonstrated an existence of a complex system of iron production and exchange (Dillmann et al., 2017, pp. 115–122).

Leroy et al. (2017, pp. 1–21) studied construction iron from three medieval masonry complexes in Angkor in Cambodia. For the sake of isolation of production sites, these researchers preferred the use of LDA and AHC. An interesting novelty was the use of Multiple Correspondence Analysis (MCA).

Bearing in mind all the discussed recent developments which convincingly suggest (especially Brauns et al. (2013) or Dillmann et al. (2017)) that an integrated approach (major and trace elements, isotope ratios) is indispensable in archaeological iron provenance studies, the authors of this paper believe that the method proposed by Charlton et al. (2012) is still worth working on, especially due to the fact that their LDA-based discrimination approach proved to be valuable (as shown in Leroy et al. (2017)). It can be further refined and confronted with other multivariate methods. Furthermore, such a refined method can be successfully applied for repeated analyses of datasets from earlier works, where only data on major oxides is available.

The approach proposed in this paper consists in using as many as seven different classification techniques for the purpose of provenancing of analysed observations. These techniques are: Linear Discriminant Analysis (LDA), Supported Vector Machine (SVM), Random Forests (RF), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Recursive Partitioning and Regression Trees (RPART) and Kernel Discriminant Analysis (KDA). Then, the quality of each classifier was tested using reclassification, holdout, leave-one-out cross-validation and k-fold cross-validation methods. A final identification of observations was carried out on the basis of results provided by all classifiers using a majority voting technique. It is hoped that experience gained in the course of this study can be used in the future for proposing a new integrated method. Such a method will take into consideration all types of data which can be used for discrimination and provenancing.

All calculations for the needs of this paper were done in the R software, Version 3.4.4 (R Core Team, 2019). Source codes are attached in Electronic Supplements. The attached source files allow to reproduce all figures included in the paper and all data provided in the tables.

2. Data and research methods

Data from three smelting experiments (marked in the paper as XP27, XP61, and XP90, or collectively XPA) conducted by Peter Crew and analysed by Thomas Birch (Birch (2014)) was used for the purpose of testing and presenting the proposed approach. This assemblage includes two groups: smelting slag (SLAG) and slag inclusions (SI) in the iron (see Electronic Supplements), which were analysed for their major element oxide and trace-element composition using a combination of SEM-EDS and LA-ICP-MS. The analytical methodology, accuracy and precision and a full dataset (Birch (2014)) are available online (Aberdeen University Library, https://digitool.abdn.ac.uk) or from the author. Major element oxides investigated were Non-Reduced Compounds. Concerning trace elements, strictly lithophile ones were selected. A final selection of variables was made after a series of LDA-based experiments with major oxides alone, different combinations of trace elements and a "full set", containing both groups of data (see Birch (2014)). Concerning the smelting experiments themselves, although operating conditions were similar, there were some significant differences between the rich bog ores that were used. For XP27 ore from the River Perry in Shropshire was used (3% MnO, 3% P_2O_5), XP61 was carried out with ore from Crawcwellt West (3% MnO, 0.4% P_2O_5), while for XP90 a manganese-rich bog ore blend from Crawcwellt South was used (11% MnO, 0.1% P_2O_5). The XP27 and XP90 blooms were processed to currency bars while the XP61 bloom remained unrefined (on the nature of these experiments see Crew (2013) and references therein).

Bearing in mind the fact that the ores that were used in XP27, XP61 and XP90 strongly differed from one another and thus they may not be

perfectly suitable to fully demonstrate the classification ability of the proposed approach, it was decided to test the classifiers again using datasets from another three experiments (XP17, XP23 and XP26, called collectively XPB). These datasets have already been used for testing purposes in Charlton et al. (2012) and Blakelock et al. (2009). For XP17, Blaenafon siderite ore (1.0% MnO, 0.4% P_2O_5) was used, while XP23 and XP26 shared both design features as well as the ore (South African Sishen hematite, MnO below detection, 0.1% P_2O_5). For these experiments, however, only major oxide data was available. Therefore, the following oxides were taken into consideration for provenance identifications: MgO, Al_2O_3, SiO_2, $K_2$2O, CaO, TiO_2, and MnO.

Then, the proposed approach was applied to actual archaeological data (Disser et al., 2017). This data has been defined by identifying consistent chemical groups by unsupervised approaches and then the groups have been discussed by recontextualising each individual one according to its archaeological context. Some words must also be said on the strategy of the selection of variables to be considered in the analysis. Concerning major elements, Non-Reduced Compounds were selected, i.e., those which are never reduced in the smelting stage of the bloomery process and thus display constant ratios in smelting-related slag. As regards trace elements, the measurement accuracy was below 12% in most cases. The elements that behave as siderophile or partition between the metallic and lithic phases were discarded, following the Ellingham approximation (based on system enthalpy). Following the methodology initiated by Peter and Vincent Serneels, elements that tend to be polluted during smelting were eliminated. These elements were identified by performing, when possible, experimental smelting of the ores which were dealt with. Thus, only such trace elements were selected that are strictly lithophile and not prone to pollution. Furthermore, the scattering of the elemental ratios, mainly by plotting the data, was examined in order to identify potential biases (due to measurement, sampling and the like). Sometimes the contribution of the variables in multivariate analyses were also made use of in order to discard those that carry almost no information (see also (Disser et al., 2017, pp.500–501). Eventually, the results were compared with those obtained in the aforementioned paper.

The first group of experimental data (XPA), which is used as training set in this paper, included 170 observations (XP27 – 20; XP61 – 70; XP90 – 80). The other group (used as test set) encompasses slag inclusions in bars and billets. The total number of observations in this group before verification of smelting-related slag was 104 (XP27 – 16; XP61 – 44; XP90 – 44). Concerning the other assemblage of experimental data (XPB), the relevant figures were the following: the training set – 225 observations (XP17 – 75; XP23 – 45; XP26 – 105); the test set before verification of smelting-related slag – 153 observations (XP17 – 64; XP23 – 44; XP26 – 45).

In order to isolate smelting slag inclusions in both groups, the approach proposed by Charlton et al. (2012) was used. Oxide values below detection limits were replaced with minimum values recorded in a given regional assemblage for a given variable. This replacement is necessary, as zeros or B/Ds (below detection) would render further data transformations impossible. Prior to running the PCA, it is necessary to transform the raw SI chemical data in order to remove the dilution effect imparted by non-modeled compounds and to give approximately equal weight to all compounds of interest. The dilution problem can be corrected by converting the original variables to subcompositional ratios, where a subcompositional ratio is equal to the measured composition of a compound divided by the sum of all compounds of interest (Birch and Martinón-Torres (2015)). Variables with large variances will dominate any PCA. In general, variables with the largest variances also tend to be those with the largest magnitudes. Therefore, subcompositional ratios of relevant groups of oxides were calculated and their – *log* values were taken (for other possible approaches see Charlton et al. (2012, pp. 2283–2284)). Then, the PCA (correlation type) was run on the assemblages (MgO, Al_2O_3, SiO_2, K_2O, CaO, and TiO_2 were taken into consideration) and PC scores were used for the AHC (dissimilarity type, Euclidean distance, weighted average agglomeration) in order to isolate inclusion groups. Then, biplots of PC scores were examined in order to find groups of smelting slag inclusions.

It must be mentioned here that a different approach to identification of smelting slag inclusions was proposed by Disser et al. In this case, the Ward method is used in the AHC and the identification of slag inclusion groups is additionally verified using biplots of contents of pairs of relevant oxides (Disser et al., 2014, pp. 322–326, Figs 8–10). Both methods of identification of smelting slag inclusions (i.e., proposed by Charlton et al. (2012) and Disser et al. (2014)) were compared and they yielded similar results, see Żabiński et al. (2018).

After the verification, the total number of observations in the first group of the test data (XPA) was 97 (XP27 – 15; XP61 – 40; XP90 – 42). In the next stage, positively verified smelting slag inclusions were analysed, taking the following variables into consideration: MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, MnO, BaO (major elements), as well as V, Cr, Rb, Sr, Y, Zr, Nb, Ce, Nd, Sm, Eu, Tb, Tm, Yb, Hf, Ta, Th, and U (trace elements). Concerning the other group (XPB), the number of verified observations was 110 (XP17 – 36; XP23 – 32; XP26 – 42). As for this group only data on major oxides was available, the following ones were included in further analyses: MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, and MnO.

In Table 1 we briefly summarise the data that we analyse. In Table 2 we state which variables we selected for further analyses. Let us draw attention here to a large disparity in the number of observations in classes in Disser's datasets (abbreviated as AD). For instance, the Min class has as many as 107 records, while the SaH class has only 6.

Table 1
Datasets used in the paper. The following abbreviations of regions were used: Barrois – Bar, Saint-Dizier-MA – SaM, FerFort – Fer, Bajocian-Bathonian – Baj, Minette – Min, Saint-Dizier-HMA – SaH, Bruche – Bru.

Description	Dataset abbreviation	Elements	Number of observations	Number of observations (in parenthes) divided into classes
Smelting experimental data, courtesy Peter Crew and Thomas Birch (abbreviated as XPA datasets, M – major elements,T – trace elements)	XPA-SLAG-M	Major	170	XP27 (20) XP61 (70) XP90 (80)
	XPA-SLAG-T	Trace		
	XPA-SI-M	Major	97	XP27(15) XP61(40) XP90(42)
	XPA-SI-T	Trace		
Smelting experimental data used by Charlton et al. (2012), XP17, XP23, XP26 (abbreviated as XPB datasets, M – major elements)	XPB-SLAG-M	Major	225	XP17 (75) XP23 (45) XP26 (105)
	XPB-SI-M	Major	110	XP17(36) XP23(32) XP26(42
Iron for the construction of Metz Cathedral in France, data provided by Alexandre Disser (abbreviated as AD datasets, M – major elements, T – trace elements)	AD-SLAG-M	Major	188	Baj(14) Bar(15) Bru(21) Fer(13) Min(107) SaH(6) SaM(12)
	AD-SLAG-T	Trace		
	AD-SI-M	Major	92	not applicable
	AD-SI-T	Trace		

Table 2
Names and numbers of variables in particular datasets.

Datasets	Variable names	Number of variables
XPA-SLAG-M XPA-SI-M	MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, MnO, Ba	8
XPA-SLAG-T XPA-SI-T	V, Cr, Rb, Sr, Y, Zr, Nb, Ce, Nd, Sm, Eu, Tb, Tm, Yb, Hf, Ta, Th, U	18
XPB-SLAG-M XPB-SI-M	MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, MnO	7
AD-SLAG-M AD-SI-M	Mg, Al, Si, K, Ca, Mn	6
AD-SLAG-T AD-SI-T	Ce, Eu, Gd, Hf, La, Nb, Nd, Pr, Sm, Tb, Th, U, Y, Yb	14

2.1. Datasets transformations

In Electronic Supplements one can find a detailed specification of the datasets presented in Table 1. These four datasets are compositional data (or composition for short) Aitchison (1986); van den Boogaart and Tolosana-Delgado (2013); Pawlowsky-Glahn et al. (2015). There is no doubt that such data cannot be analysed directly. Two commonly used transformations are *log-ratio transform* (*lr*) and *centred log-ratio transform* (*clr*). These transformations are widely known, which is why we do not discuss them here in detail. Nevertheless, for the sake of completeness of the presentation, we provide relevant mathematical formulas in Electronic Supplements. There, we also make a brief comparison of the relevant transformations. In addition, we pay attention to a certain variant of transformations used in Charlton et al. (2012).

The differences between simpler transformations (like *lr*) and other more complex (like *clr*) are, in most cases, minor and qualitatively imperceptible. However, we would suggest using the *clr*, because it has become the most standard approach to compositional data analyses in archaeology, as suggested by references quoted in this paper.

2.2. Workflow for identifying SI provenance

From Table 1 we can see that ten different datasets are used in this work. Below, we briefly discuss how these datasets are processed in computational experiments.

In the first step, the M (major elements) and T (trace elements) datasets are combined into one set of M + T components. In other words, we do not carry out separate experiments for the M and T sets. Before the M and T sets are combined into one result set they must be transformed accordingly. Since the M and T sets are in different units (in percent and ppm respectively), they must be reduced to one unit (percentages were selected as target). Finally, the *clr* transform is applied, as presented in Electronic Supplements. The entire procedure can be summarised in the following 3 steps:

Step 1: $T_{ppm} \rightarrow T_{\%}$
Step 2: $T_{\%} + M_{\%} \rightarrow (T + M)_{\%}$
Step 3: $(T + M)_{\%} \rightarrow \mathrm{clr}(T + M)_{\%}$

We must also note that in case it is decided to use the *lr* transform, Step 2 should be followed by the z-score transformation (also known as z-values, normal scores, or standardized variables). This transformation is done in order to avoid the situation where large values in the data will dominate over small ones. However, if the *clr* transform is applied, z-score is not strictly needed. Otherwise, an additional step would appear, i.e. **Step 2a:** $(T + M)_{\%} \rightarrow z - \mathrm{score}(T + M)_{\%}$.

From now on, by using the XPA-SLAG, XPA-SI, XPB-SLAG, XPB-SI, AD-SLAG and AD-SI symbols, we mean data that underwent the transformations discussed above. If at any given time it does not matter whether we are dealing with SLAG or SI datasets (or both), we will simply use XPA, XPB, or AD abbreviations.

Two types of numerical experiments were performed. In order to make our discussion more comprehensible, an illustrative workflow diagram was prepared, see Fig. 1. Individual nodes have been marked with capital letters and in the text we refer to these markings. In Electronic Supplements, a script in the R language (*workflow.R* file) has been prepared, which complements the diagram and demonstrates how the actions presented on it can be performed in practice.

In the *first experiment*, we take into account the XPA datasets. It is of importance that in both XPA-SLAG and XPA-SI datasets we have a reliable assignment to individual groups (i.e., XP27, XP61, XP90). Therefore, we use XPA-SLAG assemblages as *training dataset* in order to present the methodology for creating classifiers, while XPA-SI assemblages are used as *test dataset* in order to demonstrate that the proposed approach is proper and works correctly. This experiment can be summarised as follows:

1. Both XPA-SLAG dataset and XPA-SI dataset were loaded (node A) and transformed, as it was shortly described in Electronic Supplements (see nodes B1 and B2).
2. Using the XPA-SLAG dataset a few classifiers were built (see nodes E and G). Six parametric (LDA, SVM, RF, NB, KNN, RPART) and one nonparametric (KDA) methods were used in this task (see nodes C1 and C2). The nonparametric method also requires dimensionality reduction. Some notes are given in Section 2.4 (see nodes D1 and D2).
3. The quality of individual classifiers was determined using four techniques, that is: a) *reclassification*, b) *holdout*, c) *k-fold cross validation* and d) *leave-one-out cross validation* (see nodes E and G). In this way, it was checked whether individual methods allowed to build classifiers with sufficiently good properties. Details are given in Section 2.5.
4. Finally, the XPA-SI dataset was classified (see nodes F and H) using the six aforementioned classifiers and the final result (that is, identification of SI provenance) was achieved using *majority voting technique*. Details on this technique are given in Section 2.6. As in this dataset we have a reliable assignment to individual groups, it has become possible to thoroughly check the practical usability of constructed classifiers (i.e., building confusion matrices and calculating percentages of well-classified cases, see node J1).

The results of the above steps are presented in Section 3.

Then, the quality of the classifiers which are applied in this paper was tested again with the use of the XPB assemblage. The XPB dataset has the same structure as the XPA dataset. Therefore, its analysis is identical to that for the XPA set. Detailed results are included in Electronic Supplements.

In the *second experiment*, we take into account the AD datasets. Note here that now only the SLAG dataset has a reliable assignment to seven individual regions (abbreviated as Baj, Bar, Bru, Fer, Min, SaH and SaM; see the caption of Table 1). As in the AD-SI dataset we do *not* have a reliable assignment to individual groups and it has *not* been possible to thoroughly check the practical usability of constructed classifiers (see node J2). The results of the above steps are presented in Section 4.

2.3. Classification methods

The main purpose of this research is to classify the AD-SI observations to a correct class (region in this case). The classifier is built based on one dataset (the AD-SLAG dataset, called a training one) and then the classifier constructed in this way is used to classify a completely different dataset (the AD-SI dataset, called a test one). It is obvious that the test set is independent from the training dataset.

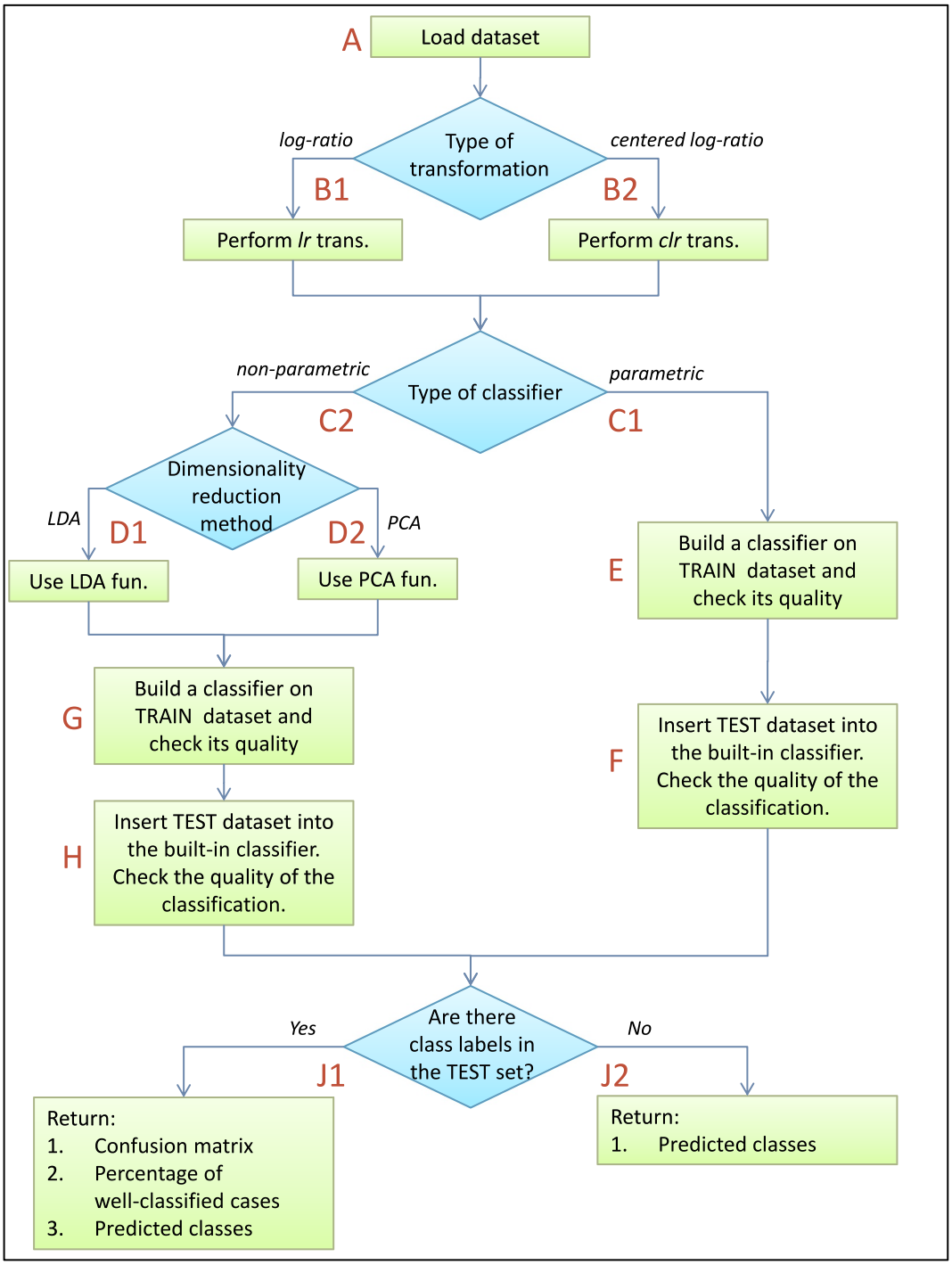

Fig. 1. Workflow diagram, from loading datasets to return the final results.

The main problem associated with the classification technique is that the target vector (region in our case) in the test dataset is completely unknown. Consequently, one must trust that the created model (classifier) correctly classifies the test dataset. In practical terms, there are no reliable methods for assessing the correctness of this process.

In order to improve the reliability of the final classification, it was decided to build several classifiers using different methods. The obtained partial results were then averaged in an appropriate manner and the final classification of the test dataset is a resultant of the scores of individual classifiers.

Six classical parametric classifiers and one nonparametric classifier were used. These are:

- LDA (*Linear Discriminat Analysis*),
- SVM (*Support Vector Machine*),
- RF (*Random Forests*),
- NB (*Naive Bayes*),
- KNN (*K-Nearest Neighbours*),
- RPART (*Recursive Partitioning and Regression Trees*),
- KDA (*Kernel Discriminant Analysis*).

These methods are widely known in the statistical world. On the other hand, we are fully aware that these may not necessarily be completely clear to every archaeologist. We do not discuss these methods in detail here, as all of them, are rather complex from a

mathematical point of view. We realise that anyone who wishes to know them better will anyway need to make a thorough study of relevant literature. Anyway, it seems to us that a detailed theoretical knowledge of each of these methods is not necessary for using them effectively in practice. For more information, the reader is sent to, e.g., Hastie et al. (2009); James et al. (2013). Different authors also propose new methods or modify existing ones, see for example Michalak and Kwaśnicka (2006); Kantavat et al. (2018); Taheri and Mammadov (2013).

A comment on the KDA method must be made here. While building the KDA-based classifier, it was necessary to reduce the dimensionality of the original *N*-dimensional data to its 2-dimensional equivalent. Such action is necessitated by the fact that the KDA method is basically unsuitable for data with dimensionality greater than 4–5, as it was reported by many authors. The reason for this is caused by phenomena known in literature as *curse of dimensionality* (Hastie et al., 2009). In practical terms, however, the resulting classifier is most conveniently presented as a 2D drawing and hence the need to reduce the data dimensionality to 2D.

The KDA classifier was used with two nD→2D dimensionality reduction methods, i.e., LDA and PCA, see Section 2.4 (note that the LDA abbreviation is used in two senses: a) as an independent classification method, b) as a method of dimensionality reduction).

It must be noted that the QDA method (*Quadratic Discriminat Analysis*) was initially taken into account. However, due to the nature of our data (many variables and too few records; see Tables 1 and 2), it could not be applied. The QDA method requires that the number of records in each category cannot be less than the number of variables. For example, XP-SLAG dataset has as many as 26 variables and XP27 category has only 20 records. As $20 < 26$, the QDA method cannot be used in this case.

2.4. Dimensionality reduction

There are a couple of methods which can be used for dimensionality reduction. Many of them are mainly used as a convenient tool for representing (mainly as 2-dimensional plots) similarities and/or dissimilarites in data. In classification tasks the following two approaches are probably the most commonly used:

- Multidimensional scaling based on Principal Components Analysis (PCA),
- Multidimensional scaling based on Linear Discriminant Analysis (LDA).

More general dimensionality reduction methods can be most broadly divided into Metrical Multidimensional Scaling (mMDS) and Non-metrical Multidimensional Scaling (nMDS). The methods are based on calculating distances between individual samples using different distance measure methods. Different authors very often use here the default Euclidean measure, while we have also examined other available measures (for example, the *distance* function implemented in *philentropy* R package is able to compute 46 different distances/similarities measures). For details on the nMDS and mNDS methods see, e.g., Kruskal (1964); Kruskal and Wish (1978); Shepard (1962a, b). The literature on multidimensional scaling is very abundant, and two very often cited monographs on this subject are Borg and Groenen (2005) and Cox and Cox (2000).

It is worth noting that LDA was also used in Charlton et al. (2012) for 8D→2D dimensionality reduction.

2.5. Methods for assessing the quality of classifiers

A natural step after building a classifier is to evaluate its performance. A large number of measures have been developed and, typically, the training dataset is used for this task. Four approaches are the most common:

1. Reclassification method. After building a classifier using the training dataset the same dataset is used for evaluating its performance. In a sense, these results can be considered less binding, because it can be considered as controversial to use exactly the same full training dataset both for building and assessment of the resulted classified.
2. Holdout method. This is the most typical type of validation, in which the training dataset is divided randomly into independent sets: the training and the test one. Typically, the test set is less than 1/3 of the training set. Such procedure is carried out repeatedly hundreds of times and at the end the average rate of correct classifications is calculated.
3. K-fold cross validation method. The original dataset is randomly divided into K equal sized subsets. Out of these, a single subset is retained as the validation data for testing the model, and the remaining $K-1$ subsets are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsets being used exactly once as the validation data. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. A 10-fold cross-validation is commonly used but in general K remains an unfixed parameter.
4. Leave-one-out cross validation method. This is a variation of the K-fold approach when the N-element dataset is divided into N subsets, containing one element. The method involves using 1 observation as the test dataset and remaining $N-1$ observations as the training dataset. This method is often used for small datasets.

All the above mentioned approaches were used to assess the quality of constructed classifiers.

2.6. Classifier voting technique

As it was stated in Section 2.3 six different classifiers were used for the task of identifying the AD-SI provenance. Obviously, the returned six results are not identical. A natural approach here is to combine these results in order to obtain the final classification. It seems that this approach is the only one that can lead to an improvement of the final classification of the AD-SI dataset. We are dealing here with a kind of *Ensemble Vote Classifier*. This idea allows to immunise the result to various types of disorders, which is why it is a reliable solution when working with 'uncertain' data.

In this case, the group *G* of *N* classifiers is a set which usually consists of less complicated base classifiers $G = \{C_1, \ldots, C_k\};\ k = 1 \ldots N$ based on which the final decision is taken. Typical approaches are:

1. Majority voting (also known as hard voting),
2. Weighted majority voting,
3. More robust algorithms based on *bagging* and *boosting* ideas (see, e.g., Zho (2012); Kuncheva (2004)), such as, for example, a popular AdaBoost.M1 method, see Freund and Schapire (1997).

A remark is required in this place. Let the decision of the k^{th} which chooses the j^{th} class be denoted as

$$d_{k,j} \in \{0, 1\};\ k = 1, \ldots, N;\ j = 1, \ldots, M, \tag{1}$$

where *N* is the number of classifiers and *M* is the number of classes. If k^{th} classifier chooses class *J*, then $d_{k,J} = 1$, and $d_{k,J} = 0$, otherwise. Voting based methods operate on labels only, where $d_{k,J}$ is 1 or 0 depending on whether classifier k chooses *J*, or not, respectively. The ensemble then chooses class *J* that receives the largest total vote.

In the *majority voting* case, we predict the final class label as the class label that has been predicted most frequently by the individual classifiers. This is the simplest case of ensemble voting system. Here, the decision D_G is as follows:

$$D_G = \underset{j=1,\ldots,M}{\operatorname{argmax}} \sum_{k=1}^{N} d_{k,j}. \tag{2}$$

As an example, let us assume that we have a group of three classifiers $G = \{C_1, C_2, C_3\}$ which classify an observation to one of two classes 'A' or 'B':

- $C_1 \rightarrow$ class A,
- $C_2 \rightarrow$ class A,
- $C_3 \rightarrow$ class B.

Then, following (2) we obtain that

$$\begin{aligned} &\text{for class A} \rightarrow \sum_{k=1}^{3} d_{k,A} = 1 + 1 + 0 = 2, \\ &\text{for class B} \rightarrow \sum_{k=1}^{3} d_{k,B} = 0 + 0 + 1 = 1. \end{aligned} \tag{3}$$

It is therefore obvious that we would classify the sample as 'class A'.

The *weighted majority voting* case differs from the hard voting in that we currently define the factor ω_k which is the weight assigned to the k^{th} classifier C_k according to some measure of performance, e.g. measures described in Section 2.5. Equation (2) now has the following shape:

$$D_G = \underset{j=1,\ldots,M}{\operatorname{argmax}} \sum_{k=1}^{N} \omega_k d_{k,j}. \tag{4}$$

As another example, let us assume that we have a group of three classifiers $G = \{C_1, C_2, C_3\}$ and three weights $\omega_1, \omega_2, \omega_3$ which classify an observation as follows:

- $C_1 \rightarrow$ class A and $\omega_1 = 0.2$,
- $C_2 \rightarrow$ class A and $\omega_2 = 0.2$,
- $C_3 \rightarrow$ class B and $\omega_3 = 0.6$.

Then, following (4) we obtain that

$$\begin{aligned} &\text{for class A} \rightarrow \sum_{k=1}^{3} \omega_k d_{k,A} = 0.2 \times 1 + 0.2 \times 1 + 0.6 \times 0 = 0.4, \\ &\text{for class B} \rightarrow \sum_{k=1}^{3} \omega_k d_{k,B} = 0.2 \times 0 + 0.2 \times 0 + 0.6 \times 1 = 0.6. \end{aligned} \tag{5}$$

Therefore, it is obvious that we would classify the sample as 'class B'. It seems that the second approach is more appropriate for our purpose.

3. Experiments with the XPA datasets (XP27, XP61, XP90)

This section presents the classification results of the XPA datasets. In line with the workflow discussed in Section 2.2, in the first step XPA-SLAG and XPA-SI datasets were transformed accordingly. The raw and transformed datasets are included as Electronic Supplements. They are the basis for all further calculations. The two XPA sets are used to present the methodology for creating and testing classifiers.

The first task to be performed is to check the quality of constructed classifiers. This is based on the training dataset that was used to build them (see Section 3.1). If the result of this test is satisfactory, one can proceed to the classification of the XPA-SI test dataset. This issue is described separately for the classical parametric classifiers (see Section 3.2) and for the nonparametric KDA classifier (see Section 3.3).

Each of the classifiers mentioned in Section 2.3 returns slightly different results. In such a situation it is difficult to decide which result is the most reliable. Therefore, the classifier voting technique (see Section 2.6) was applied and the final classification of each specific record from the test dataset was made (see Section 3.4).

3.1. Quality of the applied classifiers

There is no question that it is indispensable to examine the quality of the constructed classifiers. Omitting this step may lead to a situation that we do not really know whether the classifiers have any practical value. Suggested methods for this task were briefly discussed in Section 2.5.

It is also worth remembering that the assessment of the classifier quality is made on the basis of the training dataset and not the test one. In consequence, we are not able to precisely determine how our classifier will work when we classify a completely different dataset than the one used to build the classifier.

Nevertheless, the obtained results certainly tell us something about the quality of the classifiers we use. One has to be prepared for all kinds of surprises. Such surprise, for example, appears when we build classifiers based on the XPA-SLAG dataset. The results obtained (see Table 3) prove that this set is in a sense too good or too ideal. To put it simply, the separation between classes (regions marked as XP27, XP61, XP90) is in fact perfect.

3.2. Results for parametric classifiers

While interpreting results presented in Section 3.1, one must remember that the assessment of the classifiers quality was made on the basis of the training dataset (i.e., XPA-SLAG) and not the test one (i.e., XPA-SI). Having built our classifiers, we can use them to properly classify the test data (i.e., XPA-SI).

One of the most natural ways to demonstrate the quality of the classifier is to show it in the form of a so-called confusion matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). Table 4 shows the results for classification of the XPA-SI dataset with the LDA, SVM, NB, RF, KNN and RPART parameric classification methods. The best result was obtained for the LDA method and the worst one for the SVM method. The second result is rather surprising, because the SVM method is widely regarded as very robust and reliable one.

It is also worth noting that the classification of XP61 observations is the most peculiar. It is highly possible that this is to a great degree caused by the nature of XP61 SI data, which may have been considerably influenced by local conditions in the furnace. The samples from the XP61 bloom which were analysed (1A, 1B, 2A, 2B, 2D, and 7B) came from different locations within the bloom, corresponding to chemical variation observed in SI data (Birch (2014)). While Samples 2B and 2D came from the core, Sample 7B with a high CaO content was taken from a place near the blowing hole, and Samples 1A, 1B and 2A came from the top and edges of the bloom. Therefore, the composition of Samples 1A, 1B, 2A and 7B could be somewhat anomalous and thus not fully match the chemistry of slag (the symbols of individual observations, like 1A, 1B etc., can be found in the raw csv files attached in Electronic Supplements). This is an interesting example of the wide heterogeneity of bloom SIs (see Birch (2014)) and it certainly requires further studies. Having the results of not very high credibility, the classifier voting technique may already be very useful. It takes into account the 'knowledge' of several lower quality classifiers and on this basis it

Table 3
Results of assessing the quality of different classifiers created for the XPA-SLAG dataset. Four classical tests were used, as described in Section 2.5.

No.	Method	Reclassification	Holdout	K-fold CV	Leave-one-out CV
1	LDA	100.0	100.0	100.0	100.0
2	SVM	100.0	100.0	100.0	100.0
3	NB	100.0	100.0	100.0	100.0
4	RF	100.0	100.0	100.0	100.0
5	KNN	100.0	100.0	100.0	100.0
6	RPART	100.0	100.0	99.4	99.4
7	KDA-LDA	100.0	100.0	100.0	100.0
8	KDA-PCA	100.0	100.0	100.0	100.0

Table 4
Confusion matrices for classification of the XPA-SI dataset. The following methods are (from top to bottom): LDA, SVM, NB, RF, KNN and RPART.

	XP27	XP61	XP90	Total	% correct
LDA					
XP27	15	0	0	15	100%
XP61	0	23	17	40	57.5%
XP90	0	0	42	42	100%
Total	15	23	59	97	82.47%
SVM					
XP27	6	0	9	15	40%
XP61	0	4	36	40	10%
XP90	0	0	42	42	100%
Total	6	4	78	97	53.61%
NB					
XP27	15	0	0	15	100%
XP61	0	5	35	40	12.5%
XP90	0	5	37	42	88.1%
Total	15	10	72	97	58.76%
RF					
XP27	15	0	0	15	100%
XP61	0	14	26	40	35%
XP90	0	2	40	42	95.23%
Total	15	16	66	97	71.13%
KNN					
XP27	15	0	0	15	100%
XP61	0	16	24	40	40%
XP90	0	4	38	42	90.48%
Total	15	20	62	97	71.13%
RPART					
XP27	15	0	0	15	100%
XP61	1	23	16	40	57.5%
XP90	7	3	32	42	76.19%
Total	23	26	48	97	72.16%

prepares the final result with greater credibility.

3.3. Results for the KDA nonparametric classifier

In order to choose the best method to reduce the dimensionality of the XPA-SLAG and the XPA-SI datasets we have made a simple experiment. The XPA-SLAG and the XPA-SI datasets were reduced to 2D by two methods mentioned above and the results were depicted in Fig. 2. One can easily see that for this particular data the LDA method (upper plots) gives evidently better results then the PCA method. Regarding the XPA-SLAG set, the results are quite similar. In both cases, data from individual classes are clearly separated. However, as for the XPA-SI set, it can be seen that for the LDA method the separation of XP61 and XP90 classes is clearly better comparing to the PCA method. There is still an overlapping zone between them in the upper right plot. In the lower left plot such a zone does not occur. The XP27 class in both cases is 'safely' far from the two other classes. The advantage of the LDA method over the PCA method is not surprising here. PCA is an unsupervised learning technique (it does not use class information) while LDA is a supervised technique (it uses class information). Therefore, we would expect LDA to provide better data separation when compared to PCA, and this is exactly what we can see in Fig. 2. This kind of difference is to be expected since PCA tries to retain most of the variability in the data while LDA tries to retain most of the between-class variance in the data. In this example the first LD1 explains 83% of the between-group variance in the data while the first PC1 explains only 53% of the total variability in the data. These two values (i.e., LD and PC) are not directly comparable, but higher values always mean better performance and lower values mean worse performance.

It is also worth noting that in Fig. 2 (and also in Fig. 4) a *principal component loading vector* could also be included. This type of drawing is called *biplot*. The loading vector defines a direction in the feature space along which the data varies the most. The equivalents of Figs. 2 and 4 in the form of biplots were included in Electronic Supplements.

Among all classifiers used in our study, the KDA one is different than the others (i.e. LDA, SVM, RF, NB, KNN and RPART). A fundamental difference is that the KDA is based on probability density estimation of data under consideration. Although in theory the KDA method can be used for data of any dimensionality, practical considerations (the amount of available data, which rarely exceeds one thousand observations) suggest that 2 or possibly 3 dimensions are basically the upper limit. However, the data we work with have $8 + 18 = 26$ dimensions for XPA datasets. Therefore, dimensionality reduction is required before one can use the KDA approach. The final KDA-based classifiers are depicted in Fig. 3. Technical details of constructing such classifiers are beyond the scope of this paper. For additional information, readers may consult, e.g., Gramacki and Gramacki (2017); Gramacki (2018); Chaćon and Duong (2018) where simple demonstrative examples are presented and a basic mathematical background is given. Kernel Density Estimate (KDE) contours (one contour for data belonging to every given region XP27, XP61 and XP90) are marked with different colors. White color represents areas where the probability density is practically zero. For points located in this area it is not possible to assign them to a specific class.

In this place, it is worth noting that Charlton et al. (2012) proposed a way to verify whether all training sets could be rejected as possible provenance sources. For this purpose, they used 100% KDE contours in order to define boundaries of possible provenance fields and then they calculated mean SI percentile ranks within all provenance fields. Mean SI percentile ranks which did not plot within any field were considered a rejection of all possible provenance fields. In case the mean SI was within overlapping fields, it meant that one of two proposed provenance hypotheses could be considered more likely. This approach is of course viable, but an important reservation must be raised here.

In Fig. 3, graphs on the left demonstrate the case where the white areas are possibly the smallest (upper graph) or they do not appear at all (lower graph). This occurs if the "growth" of coloured areas is solely limited by a finite precision of calculations offered by present-day computers (double-precision floating-point format is used, which allows one to operate on numbers with really large but not infinite accuracy). Namely, the KDA is based on the use of the Gaussian function which has an infinite domain (from minus infinity to plus infinity). Therefore, from a strictly formal point of view, these white areas should not occur at all. However, every computer will sooner or later become unable to distinguish between extremely small numbers and zero. In our case, this results in the fact of appearance of areas where it is not possible to separate coloured zones.

On the other hand, it is possible to arbitrarily diminish the coloured areas, as demonstrated in the graphs on the right. In such a case, the white areas expand and some observations become non-classifiable. In our case, there are three such observations in the upper right graph and three in the lower right one. It is possible to control the contraction of the coloured areas with the use of the *supp* parameter in the *kde2D* R function. Interested readers can find it in the Electronic Supplements. A default value of this parameter is 100, and the result is that the white areas (if they occur at all) are as small as possible.

However, it must be very firmly underlined that a proper selection of the *supp* parameter is solely possible in a heuristic or experimental way. Such a procedure was applied when producing the graphs on the right. All in all, in the case of our XP data all the SI observation fell within the training set boundaries (Fig. 3, graphs on the left). The same was also the case for the AD dataset (see Fig. 5).

A question may be asked whether the KDA method has any advantage over classical parametric methods (see Section 3.2). While comparing the results in Tables 4 and 5, it can be seen that the LDA and KDA-LDA methods produce the best results and these results are

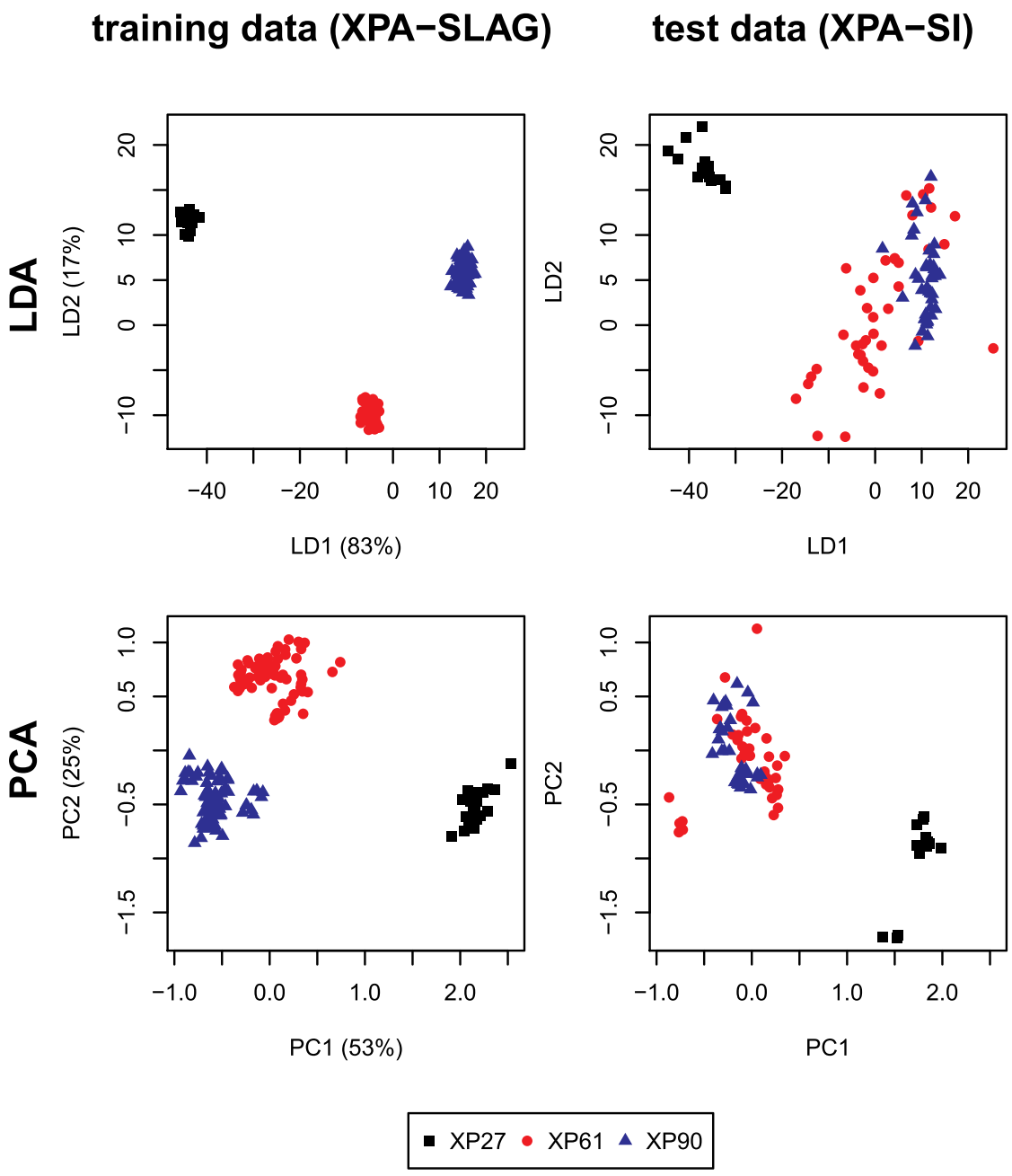

Fig. 2. Visualization of the XPA-SLAG and the XPA-SI datasets' transformations according to the LDA and PCA dimensionality reduction methods. Next, the XPA-SI testing set was projected into the space created by the XPA-SLAG training set (plots in right column). Such 2D projections show that they are much more dispersed (as should generally be expected). In this example the first LD explains 83% of the between-group variance in the data while the first PC explains only 53% of the total variability in the data.

identical for data that is analysed. There is no question that an advantage of the KDA method is an opportunity for a rapid visual assessment of the analysed data. In particular, one can quickly assess how data from individual classes are arranged relative to each other. On the other hand, one major disadvantage of the KDA method is its relatively slow computation time. Parametric methods usually do not offer such a possibility, especially if the data is strongly multidimensional. One must of course bear in mind that while using the KDA method a certain loss of information is inevitable. It results from the need of reduction of multidimensional data into 2D or possibly 3D. In the case of the XPA-SLAG data this loss is anyway not very significant, especially for the LDA method, where the first LD explains 83% of the between-group variance in the data. A PCA-based dimensionality reduction method should theoretically produce somewhat worse results than the LDA-based reduction. However, in this paper results produced by both methods are given. It seems therefore that the use of the KDA method is always worth considering, even if it does not offer a decisive advantage over more classical approaches. A sort of shortcoming of the KDA method is its computational complexity. On the other hand, for data assemblages which are not more numerous than a few thousands it is not a too great difficulty.

3.4. Final results of SI provenance for the XPA dataset

The final results are summarised in Table S1 in *Supplementary_Materials.html* file in Electronic Supplements. Numbers in parentheses indicate a posteriori probabilities of a choice of a given class (XP27, XP61, XP90). The *XP* column is the number of experiment in which a

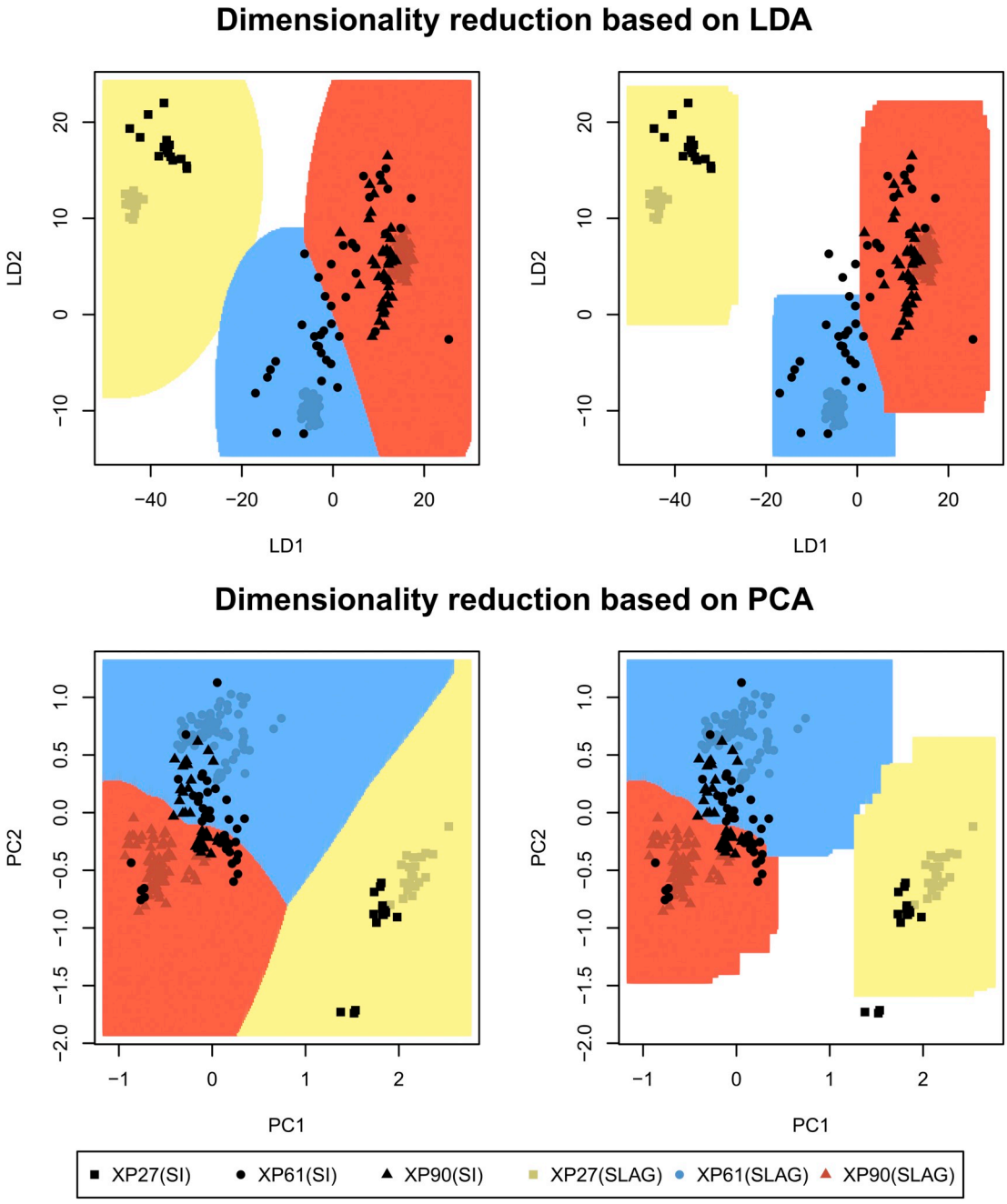

Fig. 3. KDA-based classifiers. 8D→2D reduction was made using LDA and PCA methods The locations of individual points are exactly the same as depicted in Fig. 2.

given artefact was manufactured (see the raw scv files attached in Electronic Supplements). The *Final* column gives the result of the classification after employing the voting technique, taking into account the six methods (LDA, SVM, NB, RF, KNN and RPART) for which it is possible to assess the a posteriori probabilities. The *KDA-LDA* and *KDA-PCA* columns give the result for the KDA method (dimensionality reduction was made by the LDA and PCA approaches). Note also that for the KDA method the voting technique cannot be used as the posteriori probabilities are undefined here.

In order to help the reader understand the results given in the *Final* column in Table S1, details of how to obtain them are offered below. As an example, the results for observation #20 are calculated.

1. Six classifiers (LDA, SVM, NB, RF, KNN, RPART) decided to classify observation #20 in the following manner: (numbers in parentheses indicate a posteriori probabilities of selection of a given class – XP27, XP61 or XP90)
   LDA SVM NB RF KNN RPART XP61(1) XP90(0.61) XP90(1.00) XP61 (0.58) XP61(1.00) XP61(1)
2. Then, a posteriori probabilities have been normalised to sum up to 1, with the following results:
   LDA SVM NB RF KNN RPART XP61(0.193) XP90(0.117) XP90 (0.193) XP61(0.112) XP61(0.193) XP61(0.193)
3. Next, Equation (4) produced the following result:

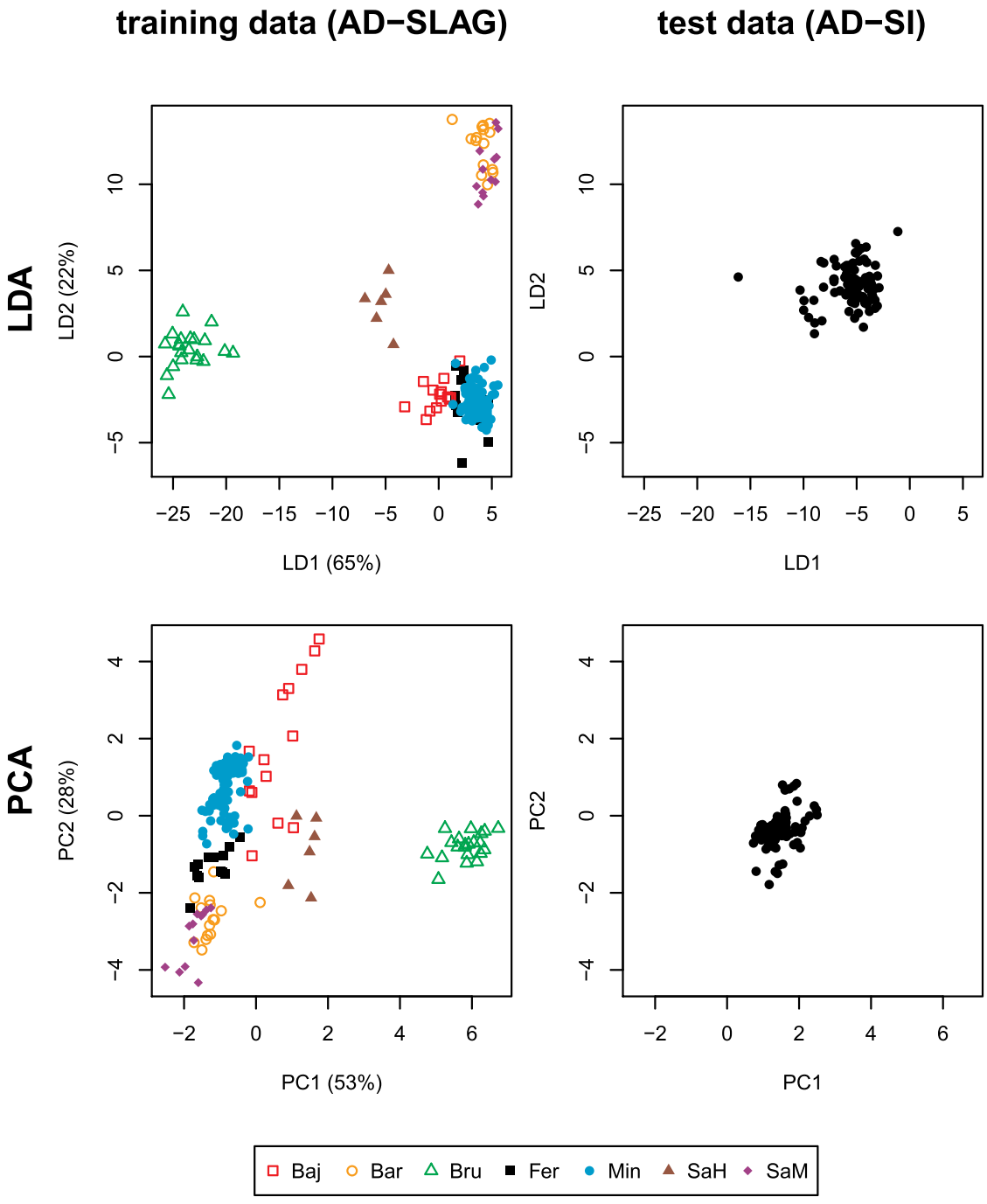

Fig. 4. Visualization of the AD-SLAG and the AD-SI datasets' transformations according to the LDA and PCA dimensionality reduction methods. In this example the first LD explains 65% of the between-group variance in the data while the first PC explains only 53% of the total variability in the data.

$$\text{class XP27} \rightarrow \sum_{k=1}^{6} \omega_k d_{k,j} =$$
$$0.193 \times 0 + 0.117 \times 0 + 0.193 \times 0 + 0.112 \times 0 + 0.193 \times 0 + 0.193 \times 0 = 0.000$$
$$\text{class XP61} \rightarrow \sum_{k=1}^{6} \omega_k d_{k,j} =$$
$$0.193 \times 1 + 0.117 \times 0 + 0.193 \times 0 + 0.112 \times 1 + 0.193 \times 1 + 0.193 \times 1 = 0.691$$
$$\text{class XP90} \rightarrow \sum_{k=1}^{6} \omega_k d_{k,j} =$$
$$0.193 \times 0 + 0.117 \times 1 + 0.193 \times 1 + 0.112 \times 0 + 0.193 \times 0 + 0.193 \times 0 = 0.422$$

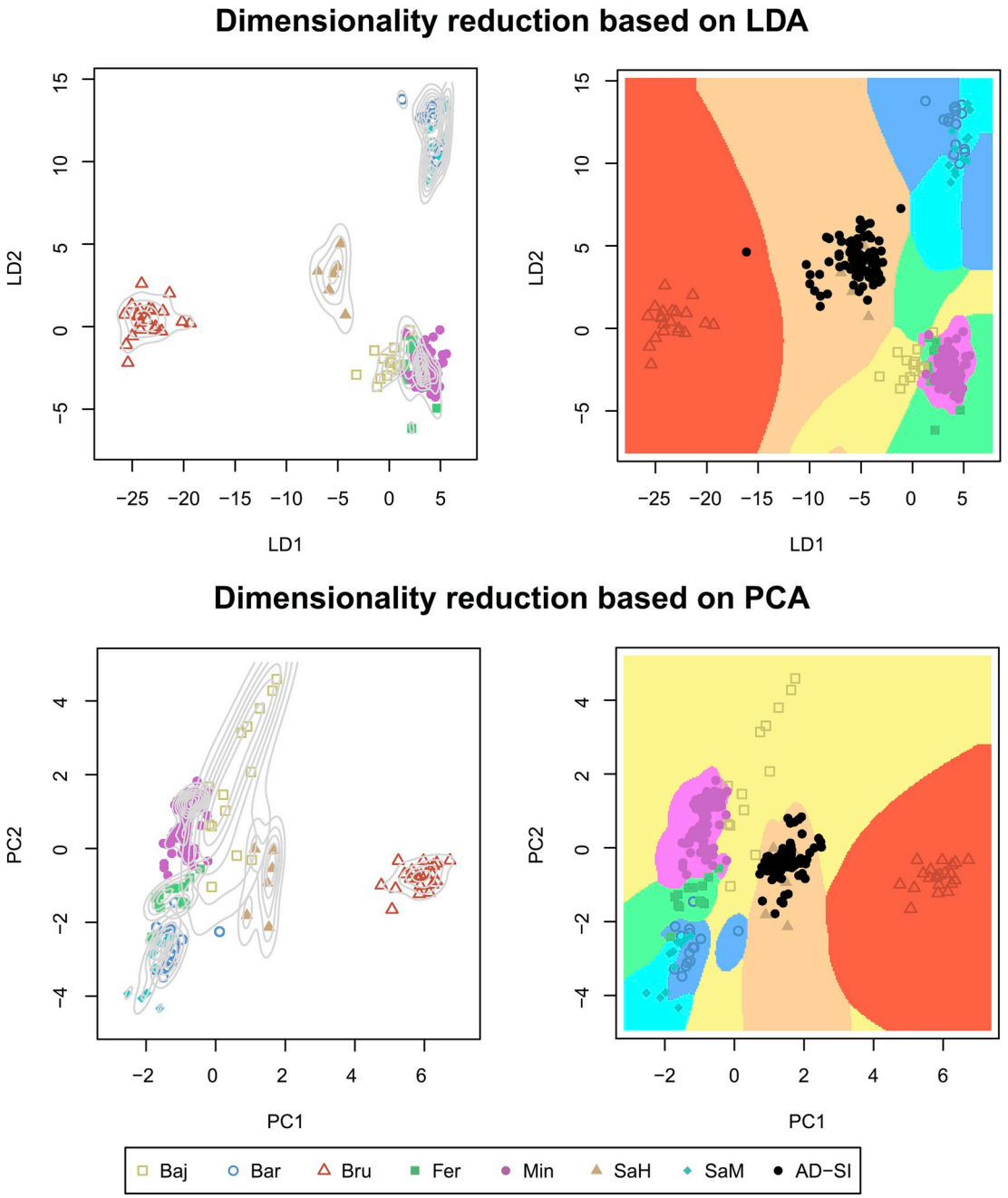

Fig. 5. KDA-based classifier. 8D→2D reduction was made using LDA (upper line) and PCA (lower line) methods. The left plots show the Kernel Density Estimate (KDE) contours (one contour for data belonging to a given region). The right plots show graphically the obtained classifier. The locations of individual points are exactly the same as depicted in Fig. 4.

Eventually, we find out that observation #20 is assigned to the class 'XP61' with a confidence level of 69%. As it could be seen, in the case of observation #20 two classifiers (SVM, NB) classified it to XP90, four classifiers (LDA, RF, KNN and RPART) assigned it to XP61, while none classifier voted for XP27. Although individual a posteriori probabilities are sometimes high (1.00), it is difficult to clearly determine which result should be accepted. Therefore, the only logical conclusion is that we tend to assign observation #20 to class XP61 but with a low confidence level (here only 69%).

While comparing the results obtained with these six classifiers they are quite in line with each other. Of course, there is no full agreement between all the classifiers used in our research. Therefore, this is a strong argument for using the voting techniques (see Section 2.6) in order to obtain sufficiently reliable results.

3.5. *Experiments with the XPB datasets (XP17, XP23, XP26)*

As mentioned above, due to considerable differences between the ore chemistries that were used in XP26, XP61 and XP90, it was decided to test the proposed approach with the use of data from another three experiments. For detailed results, the reader is sent to the *Supplementary_Materials_XPB.html* file in the Electronic Supplement and in this section only the main points are discussed. The LDA demonstrated a reasonable separation between both the test set and the training set (XP17; XP23; XP26), although some overlapping between XP23 and XP26 can be seen. The result is worse for the PCA, due to reasons stated in Section 3.3. Concerning the validation of the classification, the best results were produced by the RF and the SVM methods, while the performance of the KN approach seems to be the worst. As regards the

Table 5
Confusion matrices for classification of the XPA-SI dataset using the KDA-LDA and KDA-PCA methods.

	XP27	XP61	XP90	Total	% correct
KDA-LDA					
XP27	15	0	0	15	100%
XP61	0	23	17	40	57.5%
XP90	0	0	42	42	100%
Total	15	23	59	97	82.47%
KDA-PCA					
XP27	15	0	0	15	100%
XP61	0	21	19	40	52.5%
XP90	0	15	27	42	64,29%
Total	15	21	61	97	64.95%

confusion matrices of the SI data, it was the NB and the SVM methods that performed the best, while the results yielded by the LDA were the worst. As regard the final results of the provenance assessment, all SI observations from the XP17 dataset and most SI observations from the XP26 dataset were correctly classified, while there was some confusion concerning the XP23 dataset. All in all, the obtained results demonstrate that the performance of the proposed method is sound and reasonable.

4. Experiments with the AD datasets

Our main goal is to build a classifier using the AD-SLAG data and then to classify the AD-SI data using this classifier. We use exactly the same methodology as it was discussed in Section 3. Therefore, in this section, we do not repeat the information given therein and only discuss the final results.

Table 6 offers results of assessing the quality of classifiers. The results are generally very good for parametric methods; however, the results for the two variants of the KDA method are slightly worse. In the case of the AD assemblage we have as many as 7 classes, while there were only 3 classes in the XP data. In such circumstances, the KDA approach becomes slightly less efficient. What is more, the numerical strength of individual classes is sometimes not very high, e.g., for 'SaH' there are only 6 observations (see Table 1). This is perhaps the main reason why the results produced by the Holdout, K-kold CV oraz Leave-one-out CV methods are a little worse (but these are still not bad).

As mentioned before, in order to use the nonparametric KDA classifier the nD→2D dimensionality reduction is required. In order to choose the best 2D conversion method, an experiment which is similar to the aforementioned one was carried out, (see Fig. 4).

As it was the case with the XP data, the LDA-based dimensionality reduction method produces better results and individual classes are better separated. This visual assessment is also confirmed by the numerical results. In this example the first LD explains 65% of the between-group variance in the data while the first PC explains only 53% of the total variability in the data. It is also worth stressing that data from 'Bar' and 'SaM' classes, as well as 'Baj', 'Fer', and 'Min' classes are practically non-separable. Thus, it cannot be expected that the KDA method will perform well. As regards the AD-SI data, after the reduction to 2D it can be seen that both for the LDA and PCA methods individual observations are allocated to one pretty compact region. On this basis it could be tentatively expected that they come from one region or several regions which are similar to one another.

Table 6
Results of assessing the quality of different classifiers created by the authors for the AD dataset. Four classical tests were used, as described in Section 2.5.

No.	Method	Reclassification	Holdout	K-fold CV	Leave-one-out CV
1	LDA	100.0	100.0	99.5	99.5
2	SVM	98.9	97.0	97.3	97.3
3	NB	96.8	95.0	95.7	95.2
4	RF	96.8	98.0	96.8	96.8
5	KNN	98.9	97.0	96.8	97.6
6	RPART	98.9	90.0	86.2	93.1
7	KDA-LDA	91.5	88.0	88.3	87.8
7	KDA-PCA	95.7	90.0	88.8	89.9

Final results of SI provenance classification for the AD dataset are given in Table S2 in *Supplementary_Materials.html* file in Electronic Supplements. In addition, the classification results for two variants of the KDA method are also included in this table. The KDA-LDA and KDA-PCA columns present results for the case where 2D reduction was done using the LDA and PCA methods respectively (see Section 2.4). The two methods give in fact different results and unfortunately it is not possible to reliably indicate which one should be used. The graphical presentations of the KDA-based classifiers are depicted in Fig. 5. It can be seen that both classifiers give essentially different results. In principle, it is not possible to indicate which results are better and which ones are worse. The final results are to a great degree convergent with those achieved by Disser and his team (Disser et al., 2017). After the application of the 'classifier voting technique' (see Section 2.6) a vast majority of the AD-SI data is classified into the 'Min' class. Concerning these observations which were assigned to different classes, it can be assumed that the final classification produced by a set of independent methods can perhaps be considered more reliable.

For this specific data one should rather prefer the KDA-LDA method as the LDA produces generally better results for the dimensionality reduction task. This is demonstrated both by the visual assessment of the 2D models as shown in Fig. 4 and by numerical indicators (between-group variance and total variability in the data). The latter, however, are slightly better for the LDA than for the PCA method.

It can be also asked whether the KDA method is of any practical use here if the results it produces are so different than those in the 'Final' column in Table S2. However, it can be seen in Fig. 4 that the AD-SI data anyway falls within the 'Min' region. Although the results from Fig. 5 classify some observations from the AD-SI data to other classes than 'Min', their proximity to the 'Min' class can be perceived as a sort of positive verification of the results which were produced by the parametric methods and were then aggregated into the final outcome with the use of the classifier voting technique.

5. Conclusions

This paper proposes a new approach to the issue of provenance of archaeological iron artefacts making use of major oxides and trace elements. This approach significantly improves the credibility of final results. What is original in the proposed method is the fact that in the first step we classify data using a few very different (concerning their principles of operation) classification algorithms. In the next step, we aggregate such partial results in order to obtain a final classification of the studied data assemblage. To our knowledge, such an approach has not been applied so far in archaeometry.

As stated in the Introduction, the use of multivariate methods is already well-established in provenance studies. Among the methods which were used in previous works there were LDA and AHC (e.g., Leroy et al. (2012)), sometimes supported with MCA (Leroy et al. (2017)), a two-step PCA and AHC analysis (Disser et al. (2016), Disser et al. (2017); see also Bauvais et al. (2017), or a multi-stage PCA-AHC approach (Dillmann et al. (2017)). However, a common trait of many studies on iron provenance that one method is preferred and the reader has no chance to see how the final results could be compared with those achieved with different approaches. In our opinion, a provenance assessment which is produced by several different methods where a final result is a majority-voted outcome may be considered more reliable.

Concerning the variable selection for our provenance study, the major oxides or major elements (the experimental data - MgO, Al_2O_3, SiO_2, K_2O, CaO, TiO_2, MnO, BaO; the AD data - Mg, Al, Si, K, Ca, Mn) are commonly used for such purposes and as such do not necessitate a

further discussion. The matter is more complex with trace elements (in our case, V, Cr, Rb, Sr, Y, Zr, Nb, Ce, Nd, Sm, Eu, Tb, Tm, Yb, Hf, Ta, Th, U for the experimental data, and Ce, Eu, Gd, Hf, La, Nb, Nd, Pr, Sm, Tb, Th, U, Y, Yb for the AD data). As it can be seen, the variables overlap to a considerable extent (Ce, Eu, Hf, Nb, Nd, Sm, Tb, Th, U, Y, Yb), but in each case a decision must be made individually, as stated in Section 2. A case-specific selection of trace elements was also applied in many previous studies (e.g., Coustures et al. (2003); Coustures et al. (2006); Desaulty et al. (2008); Leroy et al. (2012); L'Héritier et al. (2016); Disser et al. (2016); Bauvais et al. (2017); Dillmann et al. (2017).

In our paper we used three assemblages of data. The first two is experimental data, marked as XPA and XPB (see Section 2, Tables 1 and 2). It consists of two sub-assemblages: smelting slag (SLAG) and slag inclusions (SI) and it contains both major oxides (M) and trace elements (T). The experimental data assemblages were used to test our new approach to the issue of classification. On the basis of these assemblages we demonstrate technical details of the method (Section 2) and then we prove its considerable practical value (Section 3). A significant new element is the use of the so-called Classifier Voting Technique. Thanks to it, we are able to strongly enhance the credibility of final results of classifying of archaeological artefacts of unknown provenance to individual regions. What is more, we can express this trustworthiness in a quantitative manner as a probability of classification of a given observation to a given group. The experimental assemblages also underwent classification with the use of the non-parametric method, that is, Kernel Discriminant Analysis (KDA). It significantly differs from other parametric methods which were used in our study, i.e., Linear Disriminat Analysis (LDA), Support Vector Machine (SVM), Random Forests (RF), Naive Bayes (NB), K-Nearest Neighbours (KNN), as well as Recursive Partitioning and Regression Trees (RPART). The obtained results are of reasonable quality (see Fig. 3). However, it is perhaps the experimental nature of the data that produced very good classification results (accuracy of about 85% for the KDA-LDA method). This may not necessarily be the case for actual archaeological data assemblages.

The third analysed assemblage is actual archaeological data provided by Alexandre Disser (marked as AD, see Section 2, Tables 1 and 2). As in the previous case, this assemblage is also divided into SLAG and SI groups and it consists of M and T data (see Tables 1 and 2). The classification results for this assemblage are offered in Section 4. On the basis of the AD assemblage 6 different parametric classifiers (using the aforementioned LDA, SVM, NB, RF, KNN, and RPART methods) and 2 non-parametric KDA classifiers were built. In the latter case, two different methods of data multidimensionality reduction were applied, that is, Linear Discriminat Analysis (LDA) and Principal Component Analysis (PCA). The quality of these classifiers was verified with a series of methods which are typical for such cases (see Table 6). Then, the final (i.e., test set) data was classified. As in the case of the experimental assemblages, the final classification of the AD test data (SI group) was obtained with the use of the classifier voting technique. The final result is in most cases convergent with that obtained by Disser et al. (2017). As said above, the classification of such observations that were assigned to different classes than in Disser et al. (2017) is perhaps more credible, as it has been produced by the voting technique on the basis of a few independent methods. It is worth stressing that the results of classification of the AD assemblage with the use of the KDA method are pretty debatable and thus a practical value of this approach is rather dubious. Although very good results were obtained for the experimental assemblages, the classification obtained for the AD data demonstrates that this is not necessary a general rule. What is a significant problem here is a great discrepancy of results (see Fig. 5). This is caused by the fact that a key operation in the KDA method is multidimensionality reduction of input data. Depending on the approach we apply, we usually receive results which strongly differ from each other in the 2D space (compare Figs. 2 and 4). What is more, there is virtually no credible method of assessing which result is the best. Therefore, these observations are somewhat contradictory to the results obtained with the KDA method, as discussed by Charlton et al. (2012). In our opinion, cross-validation is a very important step to assess the quality of the classification. This research has developed an example of a convenient framework to achieve this objective.

Eventually, it is worth stressing that an indisputable advantage of the KDA method is the fact that it clearly visualises data. This is especially significant if one works with multidimensional data, where a visualization of raw data is not very convenient. The analysed assemblages have 26, 7 and 20 dimensions respectively (see Table 2). Figs. 2 and 4 clearly demonstrate observations in individual classes. The non-parametric classifiers which were constructed on the basis of the data which was rescaled to 2D (see Figs. 3 and 5) can also be a very useful tool which increases the credibility of results produced by the parametric classifiers, as mentioned in the end of Section 4.

Author contributions

Żabiński and Mišta-Jakubowska proposed the concept of the paper and carried out an initial processing of data. Gramacki and Gramacki developed the analytical protocol for statistical examinations and produced the R codes. Birch and Disser provided the experimental and archaeological datasets for testing the statistical approach in a revised version of the manuscript, to which all authors contributed and approved.

Declaration of competing interest

None.

Acknowledgement

Thanks must go to Peter Crew for supplying the experimental assemblage that Thomas Birch analysed and for his helpful comments on the nature of this assemblage. We also wish to extend our thanks to Prof. Michael Charlton and reviewers for their feedback and comments that went into improving the manuscript and data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jas.2019.105055.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London. https://doi.org/10.1002/bimj.4710300705.

Bauvais, S., Berranger, M., Boukezzoula, M., Leroy, S., Disser, A., Vega, E., Aubert, M., Dillmann, Ph, Fluzin, P., 2017. Guard the good deposit: technology, provenance and dating of iron bipyramidal semi-products of the Durrenentzen deposit (Haut-Rhin, France). Archaeometry 41 (1), 1–18. https://www.cairn.info/revue-archeosciences-2017-1-page-45.htm.

Birch, T., 2014. The Provenance and Technology of Iron Age War Booty from Southern Scandinavia. PhD thesis. University of Aberdeen, Department of Archaeology.

Birch, T., Martinón-Torres, M., 2015. The iron bars from the 'Gresham Ship': employing multivariate statistics to further slag inclusion analysis of ferrous objects. Hist. Metall. 48, 69–78.

Blakelock, E., Martinón-Torres, M., Veldhuijzen, H., Young, T., 2009. Slag inclusions in iron objects and the quest for provenance: an experiment and a case study. J. Archaeol. Sci. 36, 1745–1757. https://doi.org/10.1016/j.jas.2009.03.032.

Borg, I., Groenen, Patrick, 2005. Modern Multidimensional Scaling Theory and Applications, second ed. Springer. https://doi.org/10.1007/0-387-28981-X.

Brauns, M., Schwab, R., Gassmann, G., Wieland, G., Pernicka, E., 2013. Provenance of Iron Age iron in southern Germany: a new approach. J. Archaeol. Sci. 40, 841–849. https://doi.org/10.1016/j.jas.2012.08.044.

Buchwald, V.F., 2005. Iron and Steel in Ancient Times, vol. 29. Historisk-filosofiske Skrifter, Copenhagen.

Buchwald, V.F., Wivel, H., 1998. Slag analysis as a method for the characterization and provenancing of ancient iron objects. Mater. Char. 40, 73–96. https://doi.org/10.1016/S1044-5803(97)00105-8.

Chaćon, J.E., Duong, T., 2018. *Multivariate Kernel Smoothing and its Applications*, Volume 160 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC. http://mvstat.net/mvksa/mvksa.pdf.

Charlton, M., Blakelock, E., Martinón-Torres, M., Young, T., 2012. Investigating the production provenance of iron artifacts with multivariate methods. J. Archaeol. Sci. 39 (7), 2280–2293. https://doi.org/10.1016/j.jas.2012.02.037.
Charlton, M.F., Crew, P., Rehren, T., Shennan, S.J., 2013. Measuring variation in iron smelting slags: an empirical evaluation of group-identification procedures. In: The World of Iron, pp. 421–430. London.
Coustures, M.-P., Béziat, D., Tollon, F., 2003. The use of trace element analysis of entrapped slag inclusions to establish ore-bar iron links: examples from two Galloroman iron-making sites in France (Les Martys, Montagne Noire, and Les Ferrys, Loiret). Archaeometry 45, 599–613. https://doi.org/10.1046/j.1475-4754.2003.00131.x.
Coustures, M.-P., Rico, C., Béziat, D., Djaoui, D., Long, L., Domergue, C., Tollon, F., 2006. La provenance des barres de fer romaines des Saintes-Maries de-la-Mer (Bouches-du-Rhône). Etude archéologique et archéométrique. Gallia 63, 243–261. https://doi.org/10.3406/galia.2006.3297.
Cox, T., Cox, M., 2000. Multidimensional Scaling, second ed. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.
Crew, P., 2013. Twenty five years of bloomery experiments: perspectives and prospects. In: Dungworth, D., Doonan, R. (Eds.), Accidental and Experimental Archaeometallurgy, vol. 7. HMS Occasional Publication, pp. 25–50. Historical Metallurgical Society (London).
Desaulty, A., Mariet, C., Dillmann, P., Joron, J., Fluzin, P., 2008. The study of provenance of iron objects by ICP-MS multi-elemental analysis. Spectrochim. Acta B 63, 1253–1262. https://doi.org/10.1016/j.sab.2008.08.017.
Dillmann, Ph, Schwab, R., Bauvais, S., Brauns, M., Disser, A., Leroy, S., Gassmann, G., Fluzin, P., 2017. Circulation of iron products in the North-Alpine area during the end of the First Iron Age (6th-5th c. BC): a combination of chemical and isotopic approaches. J. Archaeol. Sci. 87, 108–124. https://doi.org/10.1016/j.jas.2017.10.002.
Disser, A., Dillman, P., Leroy, M., L'Héritier, M., Bauvais, S., Fluzin, P., 2017. Iron supply for the building of Metz Cathedral: new methodological development for provenance studies. Archaeometry 59, 493–510. https://doi.org/10.1111/arcm.12265.
Disser, A., Dillman, Ph, Bourgain, C., L'Héritier, M., Vega, E., Bauvais, S., Leroy, S., 2014. Iron reinforcements in Beauvais and Metz Cathedrals: from bloomery or finery? The use of logistic regression for differentiating smelting processes. J. Archaeol. Sci. 42, 315–333. https://doi.org/10.1016/j.jas.2013.10.034.
Disser, A., Dillmann, Ph, Leroy, M., Merluzzo, P., Leroy, S., 2016. The bridge of Dieulouard (Meurthe-et-Moselle, France): a fresh perspective on metal supply strategies in Carolingian economy. ArchéoSciences 40, 149–161. https://doi.org/10.4000/archeosciences.4830.
Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139. https://doi.org/10.1006/jcss.1997.1504.
Gramacki, A., 2018. Nonparametric Kernel Density Estimation and its Computational Aspects, vol. 37. Springer. https://doi.org/10.1007/978-3-319-71688-6 of Studies in Big Data.
Gramacki, A., Gramacki, J., 2017. FFT-based fast bandwidth selector for multivariate kernel density estimation. Comput. Stat. Data Anal. 106, 27–45. https://doi.org/10.1016/j.csda.2016.09.001.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, second ed. Springer. Springer Series in Statistics.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer Series in Statistics. Springer.
Kantavat, P., Kijsirikul, B., Songsiri, P., Fukui, K.-I., Numao, M., 2018. Efficient decision trees for multi-class support vector machines using entropy and generalization error estimation. Int. J. Appl. Math. Comput. Sci. 28 (4), 705–717.
Kruskal, J., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29 (1), 1–27.
Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. Sage Publications.
Kuncheva, L., 2004. Combining Pattern Classifiers: Methods and Algorithms. CRC Wiley.
Leroy, S., Cohen, S., Verna, C., Gratuze, B., Téreygeol, F., Fluzin, P., Bertrand, L., Dillmann, Ph, 2012. The medieval iron market in Ariège (France). Multidisciplinary analytical approach and multivariate analyses. J. Archaeol. Sci. 39, 1080–1093. https://doi.org/10.1016/j.jas.2011.11.025.
Leroy, S., Hendrickson, M., Bauvais, S., Vega, E., Blanchet, T., Disser, A., Delque-Kolic, E., 2017. The ties that bind: archaeometallurgical typology of architectural crampons as a method for reconstructing the iron economy of Angkor, Cambodia (tenth to thirteenth c.). Archaeol. Anthropol. Sci. 1–21. https://doi.org/10.1007/s12520-017-0524-3.
L'Héritier, M., Leroy, S., Dillmann, Ph, Gratuze, B., 2016. 14. characterization of slag inclusions in iron objects. In: Recent Advances in Laser Ablation ICP-MS for Archaeology. Natural Science in Archaeology. Berlin-Heidelberg, pp. 213–230. https://doi.org/10.1007/978-3-662-49894-1.
Michalak, K., Kwaśnicka, H., 2006. Correlation-based feature selection strategy in classification problems. Int. J. Appl. Math. Comput. Sci. 16 (4), 503–511.
Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Wiley.
Pryce, T.O., Hendrickson, M., Phon, K., Chan, S., Charlton, M.F., Leroy, S., Dillmann, P., Hua, Q., 2014. The Iron Kuay of Cambodia: tracing the role of peripheral populations in Angkorian to colonial Cambodia via a 1200 year old industrial landscape. J. Archaeol. Sci. 47, 142–163. https://doi.org/10.1016/j.jas.2014.04.009.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Schwab, R., Heger, D., Höppner, B., Pernicka, E., 2006. The provenance of iron artefacts from Manching: a multitechnique approach. Archaeometry 48, 433–452. https://doi.org/10.1111/j.1475-4754.2006.00265.x.
Serneels, V., 1995. Du minerai à l'objet: un village de sidérurgistes du VIIIe au XIIe siècle à Liestal-Röserntal (BL/Switzerland). In: Magnusson, G. (Ed.), The Importance of Ironmaking, Technical Innovation and Social Change. Vol.1, Norberg Conference, 8-13.05.95. Jernkontorets Berghistoriska Utskott 58. Jernkontoret, Stockholm, pp. 124–131.
Shepard, R., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. I. Psychometrika 27, 125–140.
Shepard, R., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. II. Psychometrika 27, 219–246.
Taheri, S., Mammadov, M., 2013. Learning the naive Bayes classifier with optimization models. Int. J. Appl. Math. Comput. Sci. 23 (4), 787–795.
van den Boogaart, K., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer. https://doi.org/10.1007/978-3-642-36809-7.
Żabiński, G., Biborski, M., Mińta-Jakubowska, E., 2018. A late medieval or early modern light gun barrel from the Castle Museum in Malbork–typology, technology of manufacture and identification of the smelting process. Archaeol. Anthropol. Sci. 1–21. https://doi.org/10.1007/s12520-018-0653-3.
Zho, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC Press.