LLNL-CONF-414001

Nuclear Forensic Inferences Using Iterative Multidimensional Statistics

M. Robel, M. J. Kristo, and M. A. Heller

June 17, 2009

Institute of Nuclear Materials Management 50th Annual Meeting
Tucson, AZ
July 12 through July 16, 2009

Nuclear Forensic Inferences Using Iterative Multidimensional Statistics

Martin Robel*, Michael J. Kristo, and Martin A. Heller
Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94550

Abstract

Nuclear forensics involves the analysis of interdicted nuclear material for specific material characteristics (referred to as "signatures") that imply specific geographical locations, production processes, culprit intentions, etc. Predictive signatures rely on expert knowledge of physics, chemistry, and engineering to develop inferences from these material characteristics. Comparative signatures, on the other hand, rely on comparison of the material characteristics of the interdicted sample (the "questioned sample" in FBI parlance) with those of a set of known samples. In the ideal case, the set of known samples would be a comprehensive nuclear forensics database, a database which does not currently exist. In fact, our ability to analyze interdicted samples and produce an extensive list of precise materials characteristics far exceeds our ability to interpret the results. Therefore, as we seek to develop the extensive databases necessary for nuclear forensics, we must also develop the methods necessary to produce the necessary inferences from comparison of our analytical results with these large, multidimensional sets of data. In the work reported here, we used a large, multidimensional dataset of results from quality control analyses of uranium ore concentrate (UOC, sometimes called "yellowcake"). We have found that traditional multidimensional techniques, such as principal components analysis (PCA), are especially useful for understanding such datasets and drawing relevant conclusions. In particular, we have developed an iterative partial least squares-discriminant analysis (PLS-DA) procedure that has proven especially adept at identifying the production location of unknown UOC samples. By removing classes which fell far outside the initial decision boundary, and then rebuilding the PLS-DA model, we have consistently produced better and more definitive attributions than with a single pass classification approach. Performance of the iterative PLS-DA method compared favorably to that of classification and regression tree (CART) and k nearest neighbor (KNN) algorithms, with the best combination of accuracy and robustness, as tested by classifying samples measured independently in our laboratories against the vendor QC based reference set.

Introduction

Nuclear forensics is an important tool in the fight against illicit trafficking in nuclear material and in monitoring nuclear safeguards. The illicit trafficking of nuclear materials, including such low-threat materials as uranium ore concentrate (UOC) or "yellowcake," presents an ongoing threat to peace and security. One of the primary efforts to reduce or eliminate such trafficking is the identification and securing of sources of illicit material. In nuclear safeguards, undeclared or misrepresented transactions by states involving UOC are a potential proliferation concern. Inspection of facilities that produce or process UOC requires a means to subsequently verify the declared source, such as bulk or swipe sampling of materials. Both the illicit trafficking and nuclear safeguards scenarios require determination of the provenance of UOC samples based on measurable characteristics.

Characteristics that imply specific geographical locations, production processes, culprit intentions, etc. are referred to as "signatures." Signatures can be divided into two major categories: predictive

and comparative. Predictive signatures rely on expert knowledge of physics, chemistry, and engineering to develop inferences from these material characteristics. Comparative signatures, on the other hand, rely on comparison of the material characteristics of the questioned sample with those of a set of known samples. While the focus of this paper is the development of a comparative UOC signature capability, development of predictive signatures can ultimately benefit from insights gained by this effort. In the ideal case, the set of known comparison samples would be part of a comprehensive nuclear forensics database, a database which does not currently exist. Building such a database is a slow and expensive process, and might require data from UOC sources that no longer exist or are not readily accessible.

As we seek to develop the extensive databases necessary for nuclear forensics, we must also develop the methods necessary to produce the necessary inferences from comparison of our analytical results with these large, multidimensional sets of data. Therefore, we have been investigating the use of various approaches to identifying unknown UOC samples using a large, vendor-supplied dataset of quality control (QC) analyses. Initially, we applied popular classification methods with the goal of clearly discriminating each source from the rest. However, we were thwarted by practical difficulties due to the similarity of many sources to each other, based on the available analytical results. Because high volume, production oriented QC analyses don't require very low detection limits for trace impurities, the specific concentrations of these trace impurities, information that might help to discriminate among different sources of high purity UOC, are not captured. Furthermore, the suite of trace elements measured for QC purposes are based on ASTM standards relating primarily to fuel performance criteria, and are not necessarily the same analyses one would choose for discriminating between sources. For example, of the 23 elements which were identified as most useful in characterizing UOC from three Australian mines [1], only 7 are in the QC dataset discussed here.

Further compromising efforts to distinguish samples from a specific mine from those from all other mines were the mathematical constraints imposed by building a classification model that encompasses all sources. While this set of constraints is not ideal, the challenges it presents have led to the development of a potentially more powerful and robust classification method with broad applicability to any number of multivariate classification problems. Early experiments, inspired by the data visualization of multidimensional data sets made possible with dimension reduction, led to the development of an adaptation of existing discriminant analysis algorithms that substantially improved the accuracy of our predictions. Validation of the methodology was performed both with data from the QC dataset as well as with data from physical samples, analyzed by both the vendor QC laboratory and our nuclear forensic laboratory. The terms 'source,' 'group,' and 'class' are used interchangeably throughout this discussion and all refer to a distinct production source for the UOC, a single mill, which may, or may not, have a single source of uranium ore. In many cases, there is a one-to-one correlation between processing facility and mine (and hence between UOC and ore source), but in some cases there may be material from multiple mines directed to a single processing facility. In this discussion, 'source' (or 'group' or 'class') can refer to either situation.

A discussion of multivariate statistics and visualization typically begins with Principal Components Analysis (PCA). PCA is technically defined as "the eigen-decomposition of the covariance matrix of the measured data matrix." The data matrix, in this case, consists of concentrations of various analytes, with rows representing different samples and columns representing different variables,

e.g., parts-per-million or ppm of Fe. The result of PCA is a transformation from a coordinate system with the materials characteristics as real variables into a new system whose new variables are linear combinations of the old variables (material characteristics). The effect is a rotation and scaling such that the maximum variance in the original data is represented with the minimum number of dimensions. The first dimension, or first Principal Component (1^{st} PC) is oriented along the direction of greatest variance in the data. The 2^{nd} PC is orthogonal to the first and captures the next greatest variance. For a dataset with n measured variables, the more the measured variables are correlated, the fewer PCs are required to capture the majority of the variance, and hence represent the significant information in the data. Because of the signal averaging nature of dimension reduction, noise is reduced as well.

When the classification of samples is not known *a priori*, PCA can reveal clustering in the transformed and reduced basis that may not be easily discernable in the original data. However, if classification *is* known *a priori*, as in the case of our vendor UOC dataset, i.e., our "training set," PCA doesn't use this important information. PCA will find the directions of greatest variance in the data, but these may or may not be the optimal directions for distinguishing one class from another. We would prefer a dimension reduction transformation that uses class identity information so that transformed coordinates are optimized for discriminating among classes. Fisher's LDA does just that, but has practical constraints which make it mathematically unstable if the measured data includes any redundancy (i.e., LDA requires matrices of full rank), making it unsuitable for a dataset which could include highly correlated variables (e.g., V and V_2O_5). The plsda function in PLS_Toolbox overcomes these practical mathematical constraints, while adding noise reduction (for the same reason PCA does), and returns a result which is effectively the same [2].

While PLSDA is more appropriate than PCA for classification, PLS-DA is most effective in separating two groups. A dataset representing UOC from 21 distinct sources is far from the two-group ideal, and discrimination performance is compromised by this broad, multigroup space. However, plotting the data with respect to the first two discriminant axes (figure 4) immediately reveals that some groups are quite distinct from the mass of overlapping groups near the origin (high purity sources). Simply removing those "outlier groups" from consideration and recalculating the model consistently gives better results for difficult unknowns (mostly those from high purity sources). A systematic method to determine which groups are outliers and which should be retained is described. This process of class elimination, and subsequent regeneration of the more focused PLSDA model, is what we have dubbed iterative-PLSDA or IPSLDA. The results from comparisons between IPLSDA and competing classification methods using the UOC data are also presented.

Experimental

A vendor provided us with their database of UOC quality control measurements. The database included 30 discrete measured variables (Table 1). There are 21 mines/processing plants from 7 different countries represented (Table 2). The number of samples for a given source varies from 1 to 397. Values reported as either 0 or < detection limit (non-detects, or ND) were treated in two different ways, either by replacing those values with 0 or by replacing them with ½ the reported, or inferred, detection limit. Both approaches yielded the same result within less than 1%.

Table 1. Summary of analyses performed for all samples in the QC dataset. 'ND' indicates a value below the detection limit, but for which the detection limit was not provided. Elements/isotopes in bold are common to both the QC dataset and LLNL analyses. Those elements marked with an * were identified as most descriptive of variance in Australian UOC [1].

Units	Element/Isotope /Compound	MAX	MIN	AVG	STDEV	Instrument
percent	B	8.00E-03	<3.66E-05	4.07E-04	6.24E-04	**Thermo X-7**
percent	**Na***	2.60E+00	<1.00E-02	6.33E-02	1.32E-01	**PE AA400**
percent	**Mg***	1.79E-01	<2.46E-05	1.09E-02	1.81E-02	**Thermo X-7**
percent	Si	4.52E-01	<5.66E-03	3.25E-02	5.01E-02	**Thermo X-7**
percent	P	2.10E-01	<3.34E-03	1.39E-02	1.88E-02	**Thermo X-7**
percent	**K**	2.14E+00	<1.00E-02	1.70E-02	8.90E-02	**PE AA400**
percent	Ca	1.33E+00	<1.00E-02	8.15E-02	9.36E-02	**PE AA400**
percent	**Ti**	6.40E-02	<6.21E-05	8.53E-04	2.74E-03	**Thermo X-7**
percent	**V**	7.87E-01	<3.33E-05	6.26E-03	3.26E-02	**Thermo X-7**
percent	**Cr***	5.40E-02	<3.21E-05	3.77E-04	1.96E-03	**Thermo X-7**
percent	**Fe***	6.97E+00	<1.00E-02	1.20E-01	4.66E-01	**PE AA400**
percent	As	6.50E-02	<1.05E-05	2.57E-03	5.31E-03	**Thermo X-7**
percent	Se	1.72E-01	<9.69E-05	1.97E-03	1.09E-02	**Thermo X-7**
percent	**Zr***	3.11E+00	<2.85E-05	1.87E-02	9.25E-02	**Thermo X-7**
percent	**Mo***	2.50E-01	<9.60E-06	3.20E-02	4.05E-02	**Thermo X-7**
ppb	**Ag**	2.73E+05	<1.50E+01	2.58E+03	1.25E+04	**Thermo X-7**
ppm	**Cd**	5.24E+01	<3.60E-02	3.48E+00	7.66E+00	**Thermo X-7**
ppm	**Ba**	1.49E+02	<1.62E-01	2.02E+00	5.57E+00	**Thermo X-7**
ppm	**Pb**	1.30E+02	<6.60E-02	2.10E+00	6.53E+00	**Thermo X-7**
ppm	**Th***	3.45E+03	<1.41E-01	4.73E+01	2.20E+02	**Thermo X-7**
ppm	**^{234}U**	8.43E+01	4.70E+01	5.37E+01	3.04E+00	Not Reported
percent	SiO2	9.67E-01	ND	6.96E-02	1.07E-01	Not Reported
percent	PO4	6.45E-01	1.03E-02	4.28E-02	5.78E-02	Not Reported
percent	V2O5	1.41E+00	ND	1.12E-02	5.84E-02	Not Reported
percent	235U	7.14E-01	7.08E-01	7.11E-01	1.23E-03	Not Reported
percent	**^{238}U**	8.76E+01	6.12E+01	8.24E+01	4.45E+00	Not Reported
percent	SO4--	8.71E+00	ND	9.40E-01	1.19E+00	Not Reported
percent	halides	2.34E+00	ND	1.13E-02	7.99E-02	Not Reported
percent	F-	1.00E+00	ND	3.41E-03	3.90E-02	Not Reported
percent	H2O	8.00E+00	ND	2.89E-01	9.26E-01	Not Reported

Table 2. Uranium ore concentrate vendor QC data sources and numbers of samples.

Source ID (Class)	Source Name	Samples	Class	Source Name	Samples
1	USA 1	6	12	Australia 3	233
2	USA 2	3	13	Kazakhstan 1	37
3	USA 3	8	14	Kazakhstan 2	7
4	USA 4	85	15	Kazakhstan 3	4
5	USA 5	21	16	Kazakhstan 4	1
6	Canada 1	301	17	Kazakhstan 5	17
7	Canada 2	94	18	Uzbekistan 1	177
8	Canada 3	5	19	South Africa 1	56
9	Canada 4	11	20	Namibia 1	169
10	Australia 1	397	21	Czech Republic 1	27
11	Australia 2	59			

Software

PLS_Toolbox is a suite of chemometric applications available through Eigenvector Research, Inc. which includes many popular methods, including PCA and PLSDA [3]. The toolbox is available both as a standalone application and as a MATLAB™ add-on, the latter of which was used for this study. The command line version of the 'plsda' function supplied with PLS_Toolbox was used to perform the discriminant analysis at each iteration of the IPLSDA process. MATLAB R2008b served as the development environment for the IPLSDA implementation function. Additional analyses using classification tree (CT) and k-nearest neighbor (KNN) methods were performed in R, a free statistical computing and graphics language and environment available for download from http://www.r-project.org.

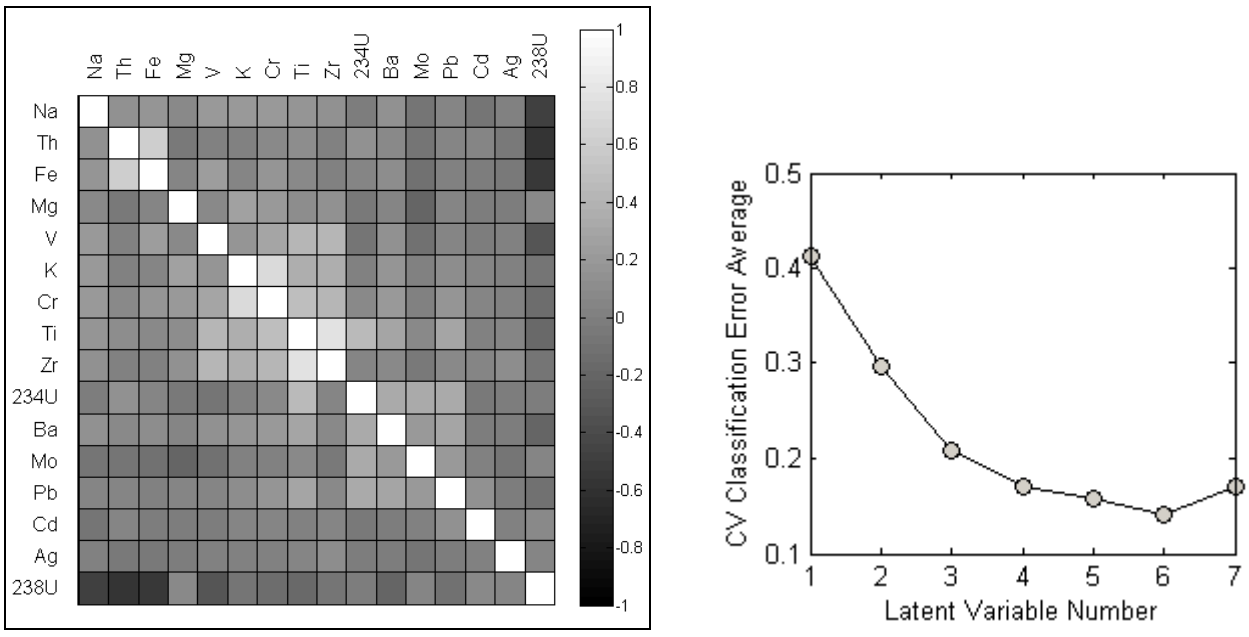
Implementation of IPLSDA

To determine which classes (sources) to remove from consideration (exclude from the model) and which to retain, the error minimizing decision boundary feature of the plsda function was used. The decision boundary is calculated separately for each class. If the unknown maps to the same side of the decision boundary as a class, that class is retained for the next iteration. This is essentially a 'survival of the fittest' approach to narrowing down the possible sources until a final single answer is reached. Figure 5 shows what happens if this approach is not used with the UOC dataset, and instead the class with the highest probability is chosen from the global model.

Results and Discussion

Prior to performing PLS-DA, the measured data were autoscaled, i.e., mean-centered and scaled to unit variance by dividing each column by its standard deviation [3]. The correlation matrix (Figure 1a) shows that there is little strong correlation amongst the 16 variables used for the model based upon the results from LLNL analyses of physical samples. This explains why the 3 LV model captures less than 50% of the total variance in the measured data (Figure 4). The number of latent variables to use in the model is best determined from a plot of classification error vs. latent variable (Figure 1b). Choosing up to 6 latent variables yields improved results with a training set, but risks overfitting the data, causing poor performance of the model when applied to new samples. Choosing 4, or even 3 LVs, produces a model with more robust behavior. Four LVs were used for the results presented here.

To fully characterize the performance of a classification algorithm, both false positives and false negatives must be considered. For classes with small numbers of samples, leave-one-out (LOO) validation is appropriate. For larger classes, this method is prone to over-fitting. For the purposes of comparing methods applied to the UOC dataset, we consider LOO validation to be a useful tool. It is also reasonable for many of the classes in the dataset, due to small numbers of samples. The validation tabulates false positives and false negatives. False positives are defined as the number of samples from outside a given class which are incorrectly attributed to it. False negatives are defined as the number of samples from a given class which fail to classify as in that class. The fraction of false positives is the number of false positives divided by the total number of samples in the dataset from other classes. The fraction of false negatives is the number of false negatives in a class divided by the total number of samples in that class. Classification error is defined as the average of the fractions of false positives and false negatives, the intent being to produce a model with a good balance between sensitivity and specificity [3].



(a) (b)

Figure 1. Pre-processed data represented with a correlation matrix (a) for those variables which are common to both the QC dataset and LLNL analyses. Leave-one-out validation is used to derive classification error for the QC dataset (b).

Loadings plots (Figures 2 and 3) show how much influence each real, measured variable has on each LV. Trends, such as the anti-correlation between 238U (uranium assay) and all other variables in LV 1, are a strong function of the characteristics of outlier classes in the global model. Removing a single source with unusually high isotopic concentration of ^{234}U (234U) and high

impurities from the model causes 234U to group with 238U and away from most of the trace impurities. This sensitivity to the characteristics of sources present in the model helps explain why the iterative method is effective. Variance that is descriptive of the differences between high purity sources has less influence on the model when more extreme variance in the measured data is "using up" a significant fraction of the variance captured in the model. This example also illustrates the arbitrariness of the signs of the loadings, which are only significant relative to each other.

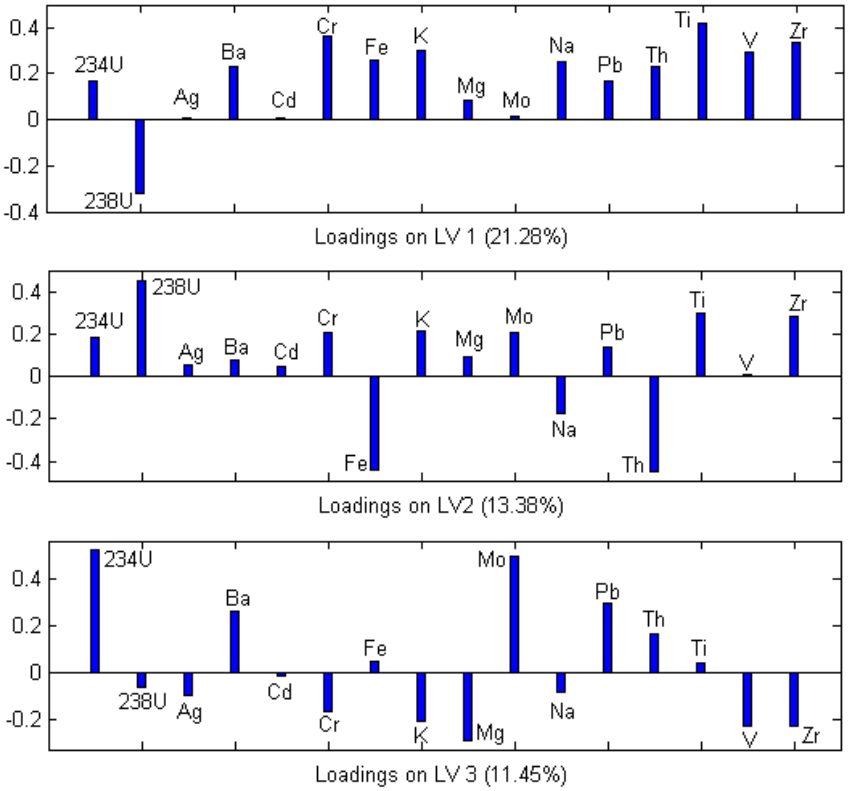


Figure 2. Loadings plots showing the influence of measured variables on the first 3 latent variables when all sources are included in the model.

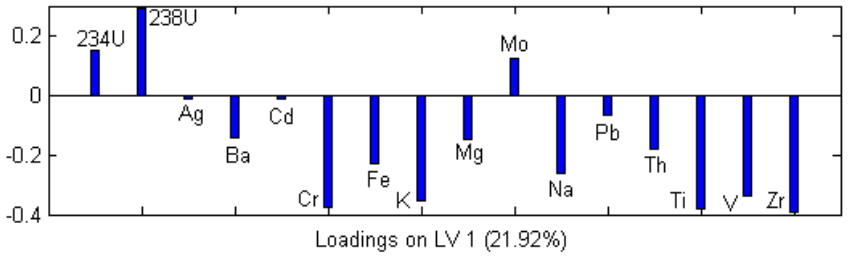
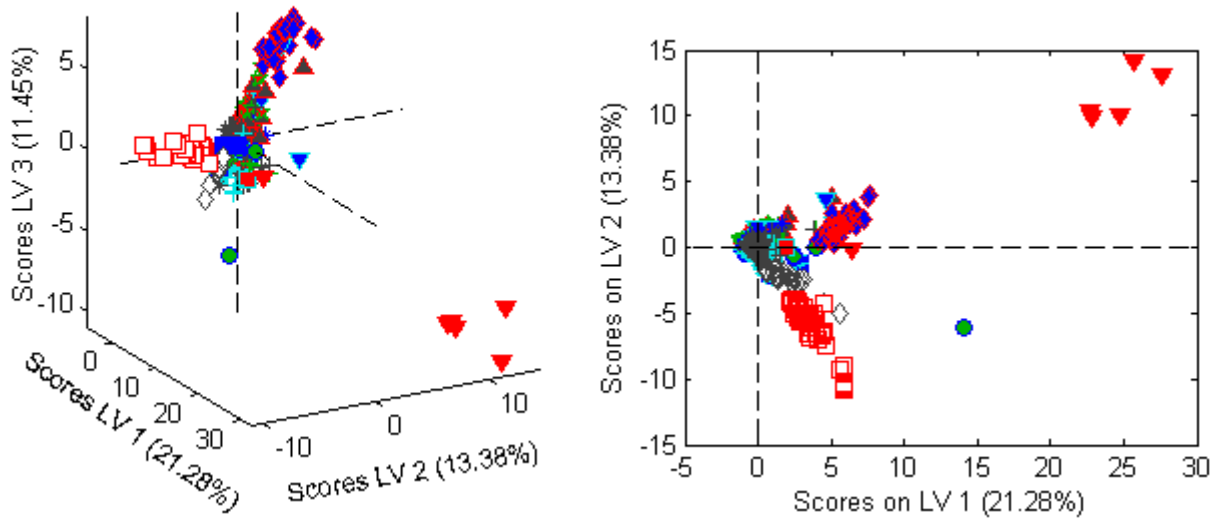


Figure 3. Loadings on LV1 after removal from the model of a single source with unusually high 234U concentration.



(a) (b)

Figure 4. Global UOC dataset score plot in PLSDA model space. Each point represents a single sample, with a total of 1718 in these plots. Each latent variable (LV) captures a percentage of the variance in all measured variables.

IPLS-DA was compared to the alternative classification techniques CT (a subset of the Classification And Regression Tree (CART) method) and KNN [4, 5]. KNN, with K=5, was chosen for the comparison using a subset of QC data only (samples >5), but was not practical for the purpose of testing with LLNL analyses of physical samples because some sources have 5 or fewer samples. Similarly, the performance of CT varies depending in part on the tuning parameter CP, or "cost complexity factor," which uses feedback from cross-validation errors [6]. The value of CP was set to 0.005 for this study to avoid overfitting, but this produced default false negatives for classes 1-3, 8, 9, 14, and 15. More important than a single performance metric, such as Classification Error (Figure 6), is the fundamental difference between CT and PLS-DA with regard to how predictor variables are used. CT considers one variable at a time, whereas PLS-DA uses a noise-reduced and discriminant optimized weighted linear combination of all the variables simultaneously [6], making PLS-DA a more inherently robust tool.

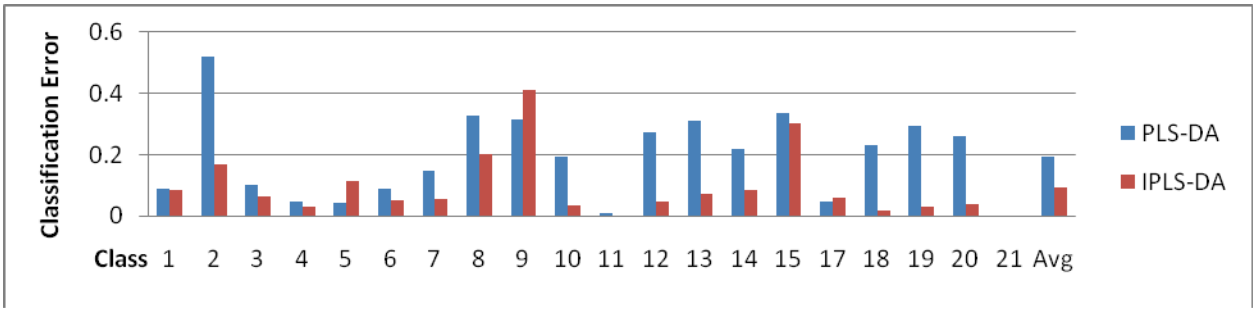


Figure 5. Comparison between PLS-DA classification error using the global dataset and using the iterative PLS-DA method. Only the 16 variables common to both LLNL and the QC datasets were used. Class 16 was excluded because it contains only 1 sample, making validation impossible.

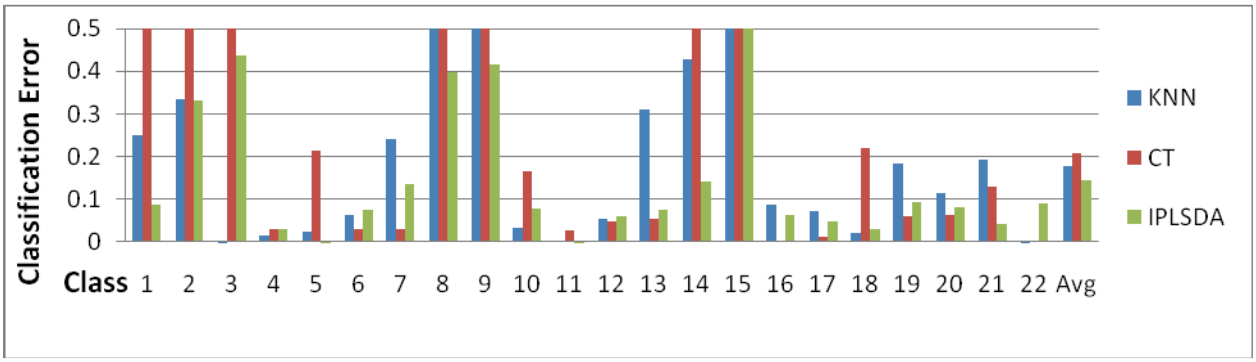


Figure 6. Comparison between three classification methods using cross-validation of the QC dataset. All 30 variables in the QC dataset were used. Class numbering is changed relative to figure 5 and two classes have been split into sub-classes.

Table 3. Comparison of classification performance for different classification methods using a subset of the QC dataset (sources with > 5 samples).

Method	Average Classification Error
IPLSDA	0.15
CT	0.21
KNN (K=5)	0.18

Table 4. Comparison between IPLSDA and CT predictions using LLNL analyses of physical samples from the vendor. All 20 modeled sources were represented.

Method	Correct Country (%)	Correct Source (%)
IPLSDA	85	60
CT	75	45

Correct predictions for IPLSDA on analyses of physical samples were later increased to 90% correct country and 65% correct source after applying a regression coefficient vector derived from observed bias between vendor data and LLNL measurements for certain trace elements (Table 4). This highlights the importance of a good calibration between the vendor provided QC data and the highly precise and accurate analyses performed either for the purpose of forensic analysis, or to expand the database with new samples and locations. Assuring such inter-laboratory consistency is vital to exploit the full potential of the comparative approach to sample attribution. Unfortunately, elimination of any such systematic error in the combination of data from multiple labs does not address the potential for significant inhomogeneity for material from a given location. This is a fundamental limitation of the comparative signature approach.

Table 4. Significance test for systematic bias between QC data and LLNL measured data for elements common to both datasets.

Element	**Two-sided p-Value**	**Element**	**Two-sided p-Value**
Ag	0.28	Mo	0.01
Ba	0.11	Na	0.69
Ca	0.24	Pb	0.30

Cd	0.16	Th	0.28
Cr	0.64	Ti	0.04
Fe	0.39	V	0.04
K	0.31	Zr	0.31
Mg	0.10		

Conclusion

We have found that traditional multidimensional techniques, such as principal components analysis (PCA), are especially useful for understanding such datasets and drawing relevant conclusions. In particular, we have developed an iterative partial least squares-discriminant analysis (PLS-DA) procedure that has proven especially adept at identifying the production location of unknown UOC samples. By removing classes which fell far outside the initial decision boundary, and then rebuilding the PLS-DA model, we have consistently produced better and more definitive attributions than with a single pass classification approach. Performance of the iterative PLS-DA method compared favorably to that of classification and regression tree (CART) and k nearest neighbor (KNN) algorithms, with the best combination of accuracy and robustness, as tested by classifying samples measured independently in our laboratories against the vendor QC based reference set.

References

1. E. Keegan, S. Richter, I. Kelly, H. Wong, P. Gadd, H. Kuehn, and A. Alonso-Munoz, "The provenance of Australian uranium ore concentrates by elemental and isotopic analysis," Applied Geochemistry, 23 pp. 765–777, 2008.

2. M. Barker and W. Rayens, "Partial least squares for discrimination," J. Chemometrics, 17, pp. 166-173, 2003.

3. B. M. Wise, J. M. Shaver, N. B. Gallagher, W. Windig, R. Bro, and R. S. Koch, "PLS_Toolbox 4.0 for use with MATLAB™," Eigenvector Research, Inc., 2006.

4. L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," Chapman and Hall/CRC, 1998.

5. T. M. Cover and P.E. Hart, "Nearest neighbor pattern classification", IEE transactions on Information Theory, IT 13(1), 21–27, 1967.

6. StatSoft Electronic Statistics Textbook, www.statsoft.com/textbook/stathome.html, StatSoft, Inc. (1999).

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.