Talanta 83 (2011) 1166–1172

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Anionic forensic signatures for sample matching of potassium cyanide using high performance ion chromatography and chemometrics

Carlos G. Fraga*, Orville T. Farmer, April J. Carman

Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA 99352, USA

ARTICLE INFO

Article history:
Available online 21 August 2010

Keywords:
Chemical forensics
Feature selection
Supervised classification
Ion chromatography

ABSTRACT

Potassium cyanide was used as a model toxicant to determine the feasibility of using anionic impurities as a forensic signature for matching cyanide salts back to their source. In this study, portions of eight KCN stocks originating from four countries were separately dissolved in water and analyzed by high performance ion chromatography (HPIC) using an anion exchange column and conductivity detection. Sixty KCN aqueous samples were produced from the eight stocks and analyzed for 11 anionic impurities. Hierarchal cluster analysis and principal component analysis were used to demonstrate that KCN samples cluster according to source based on the concentrations of their anionic impurities. The Fisher-ratio method and degree-of-class separation (DCS) were used for feature selection on a training set of KCN samples in order to optimize sample clustering. The optimal subset of anions needed for sample classification was determined to be sulfate, oxalate, phosphate, and an unknown anion named unk5. Using *K*-nearest neighbors (KNN) and the optimal subset of anions, KCN test samples from different KCN stocks were correctly determined to be manufactured in the United States. In addition, KCN samples from stocks manufactured in Belgium, Germany, and the Czech Republic were all correctly matched back to their original stocks because each stock had a unique anionic impurity profile. The application of the Fisher-ratio method and DCS for feature selection improved the accuracy and confidence of sample classification by KNN.

1. Introduction

Cyanide salts are of particular interest to criminals or potential terrorists. They are highly toxic as an inhalation hazard through the release of toxic hydrogen cyanide gas when mixed with dilute acid and as a poison hazard through ingestion. Cyanide salts have been used in a number of high-profile poisonings such as the product-tampering crimes involving Tylenol® that killed 7 people in 1982 and Excedrin® that killed 2 in 1986 [1]. Cyanide-laced Kool-Aid® was also used by the Reverend Jim Jones to kill more than 900 of his followers in 1978 [1]. Currently, the Centers for Disease Control and Prevention includes cyanide as a probable agent for chemical terrorism [2]. Given these facts, commercial potassium cyanide was studied here as a model cyanide salt to determine the feasibility of using impurities in cyanide salts as a potential forensic signature.

KCN is a white, granular or crystalline solid that is manufactured commercially by the neutralization of aqueous potassium hydroxide with hydrocyanic acid [3]. After neutralization, the aqueous KCN is concentrated through the evaporation of water typically under reduced pressure and elevated temperatures. The KCN eventually precipitates and is separated by filtration followed by drying [3]. The primary impurities in commercial KCN (>99.0% purity) are potassium carbonate, potassium hydroxide, and potassium formate. These salts can be present in several tenths of a percent (w/w) as illustrated by a KCN lot in this study which was assayed to be 99.3% KCN, 0.37% K_2CO_3, 0.36% KOH, and 0.15% $KHCO_2$ according to its certificate of analysis. Given the known presence of these salt impurities and other lesser ones, we hypothesized that anionic impurities in commercial KCN may be of forensic value for sample matching. Our previous work demonstrated the forensic value of organic impurities for sample matching of commercial nerve-agent precursors based on stock [4,5]. Herein, we extend this work to the anionic impurities found in eight KCN stocks while also illustrating, for the first time, the value of feature selection for chemical forensics.

Each of the eight KCN stocks had a unique lot number listed on its original storage container. Each stock also had a certificate of origin listing its country of origin. The country of origin for any good is the country in which that good is first manufactured, extracted, produced or grown [6]. Further work or material added to the good in another country must result in a new and different good, which then has a name, character, use, and tariff code different from those of its constituent materials in order to change the good's country of origin [6]. In the case of KCN, simply repackaging and reanalyz-

---

* Corresponding author. Tel.: +1 509 371 7589.
E-mail addresses: carlos.fraga@pnl.gov, carlos.fraga@edwards.af.mil (C.G. Fraga).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.08.017

ing the purity of a KCN shipment by a supplier does not change its country of origin. The above definition of country of origin was confirmed by a customer representative from a major chemical supplier.

In this study, portions of eight KCN stocks originating from four countries were separately dissolved in water and analyzed by high performance ion chromatography (HPIC) using an anion exchange column and conductivity detection. Sixty KCN aqueous samples were produced from the eight stocks and analyzed for 11 anionic impurities. Hierarchal cluster analysis (HCA) and principal component analysis (PCA) were used to demonstrate that KCN samples cluster according to source based on the concentrations of their anionic impurities. The Fisher-ratio method and degree-of-class separation (DCS) were used for feature selection on a training set of KCN samples in order to optimize sample clustering. Finally, *K*-nearest neighbors (KNN) was used on a test set to demonstrate the forensic value of anionic impurities as a means of matching an unknown KCN sample to a specific source.

2. Theory

The following subsections describe the algorithms utilized in this report.

2.1. HCA and PCA

HCA and PCA are two unsupervised pattern recognition techniques widely used in chemometrics [7,8]. HCA involves measuring the distance between samples in a multivariate space and then progressively linking samples that are closest together. Eventually, all samples are linked creating a dendrogram that illustrates sample relationships based on their similarity among the different measured variables, e.g., anion concentrations. For instance, samples that are similar will have measured values that are similar and will therefore be the first linked together by HCA. On the other hand, PCA converts the original measured variables to a set of relevant principal components that are much smaller in number than the original measured variables. Each principal component consists of a scores vector and loadings vector. Each scores vector contains the new measured value for each sample for a given principal component. The loadings vector contains the weights or loadings given to each original variable to create the given principal component. In summary, PCA compresses the original sample data such that relationships among samples can be typically viewed with a two or three dimensional scores plot, with each dimension representing a PC. For both PCA and HCA, the sample data is typically preprocessed by mean centering. In addition, normalizing the measurements for each variable to unit variance is also accomplished when the units among variables are different.

2.2. Fisher-ratio method and DCS

The Fisher-ratio method [9] is used to discard variables (e.g., anions) that are not useful for classification while retaining variables that have chemical information correlating with known sample classes. Variables or features that have a Fisher ratio that exceed a set threshold are retained. The Fisher ratio is the among-class variance in measured variables divided by the summed within-class variance in the measured variables for a training set of known sample classes [9]. The among-class variance (s_{ac}^2) is calculated for a variable *x* as:

$$s_{ac}^2 = \sum_{i=1}^{k} \frac{(\bar{x}_i - \bar{x})^2 n_i}{k-1}$$

where *k* is the number of classes, n_i is the number of class members in the *i*th class, $\bar{x}_i$ is the mean value for variable *x* in the *i*th class, and $\bar{x}$ is the overall mean of *x* across all classes. In this report, the summed within-class variance (s_{wc}^2) is calculated as:

$$s_{wc}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{N-k}$$

where $\bar{x}_{ij}$ is the value for variable *x* in the *j*th member of the *i*th class and *N* is the number of all members in all classes. The Fisher ratio is then calculated as the ratio of the two variances:

$$f\,\text{ratio} = \frac{s_{ac}^2}{s_{wc}^2}$$

Once the Fisher ratio is calculated for each variable, then the DCS is calculated using the variables that exceed a Fisher-ratio threshold that is incrementally varied. The DCS [10] is a useful metric to help determine the optimum threshold and therefore the subset of variables needed for optimal sample classification. In this report, the DCS is defined as the Euclidean distance between the means of two classes divided by the square root of the summed variances in the Euclidean distance of each member relative to the mean of its class. The Euclidean distance between the means of two classes *A* and *B* ($d_{\bar{A}\bar{B}}$) is calculated as:

$$d_{\bar{A}\bar{B}} = \sqrt{\sum_{x=1}^{z} (\bar{A}_x - \bar{B}_x)^2}$$

where *z* is the number of variables (e.g., anions) and $\bar{A}_x$ and $\bar{B}_x$ are the means for variable *x* in classes *A* and *B*, respectively. The DCS is then calculated as:

$$\text{DCS} = \frac{d_{\bar{A}\bar{B}}}{\sqrt{s_A^2 - s_B^2}}$$

where s_A^2 and s_B^2 are the variances in the Euclidean distance of each member relative to its class mean for classes *A* and *B*, respectively [10]. After calculating the DCS for multiple variable subsets (each subset determined by a different Fisher-ratio threshold), then the variable subset giving the largest DCS for a class pair is considered optimal.

2.3. KNN

KNN is a simple supervised classification technique widely used in chemometrics to determine the class of unknown samples [7]. It classifies unknown samples based on the nearness of the unknown sample to a set of known training samples in multivariate space. Each unknown sample is classified by the number of nearest neighbors or *K*. Those known samples that make up the majority of the *K* neighbors decide the class membership of the unknown sample.

3. Experimental

3.1. KCN

Eight KCN stocks (A–H) with different lot numbers and a listed purity of at least 98% were obtained from five chemical suppliers. The word stock refers to solid KCN from a specific lot-labeled bottle. Table 1 lists the country of origin for each of the KCN stocks as stated by each stock's certificate of origin provided by the supplier. In addition, some suppliers identified the specific manufacturer such that some of the stocks are known to be from the same manufacturer (see Table 1).

Table 1
KCN stocks.

Stocks[a]	Country of origin[b]
A[c]	United States
B	Belgium
C[c]	United States
D	Czech Republic
E	Germany
F[c]	United States
G[c]	Untied States
H	United States

[a] Letter represents a specific KCN stock with a unique lot number. The letter is the first term in the sample label scheme: *stock-solution-duplicate*.
[b] As stated on the certificate of origin.
[c] Stocks known to be from the same manufacturer based on supplier information.

3.2. Sample preparation

A 1-g portion of a KCN stock was dissolved in deionized (DI) water from a Barnstead Nanopure Diamond water purifier (Thermo Fisher Scientific Inc., San Jose, CA) to produce approximately 30 g of an aqueous KCN solution. The actual masses of the KCN and KCN solution were measured using a Sartorius R2000 balance (Sartorius Corp., Edgewood, NY) capable of measuring to ten thousandths of a gram. The actual mass of the KCN solution was measured after sealing the solution's container. Additional 1-g portions from the same KCN stock were dissolved on different dates to produce up to a total of three 30-g aqueous KCN solutions (see Table 2). This was accomplished in order to account for compositional differences within a given KCN stock. A 0.15-mL aliquot from each KCN solution was then diluted in DI water to produce an approximate 5-g aqueous KCN sample for HPIC analysis. The actual masses of the aliquot and diluted aqueous KCN sample were measured and recorded. Two more 0.15-mL aliquots from each KCN solution were diluted in DI water to produce a total of three duplicate aqueous KCN samples. This was accomplished in order to account for variability associated with sample preparation and analysis. In summary, six or nine KCN samples were made from each KCN stock for a total of 60 KCN samples as listed in Table 3. All KCN samples were labeled using a label scheme (see Tables 1–3) such that each sample had a unique label that also identified its source. For example, C-2-3 was the diluted third aliquot of a KCN solution made on 01 February 2010 from a United States stock known as stock C.

3.3. HPIC analysis

A Dionex ICS-3000 HPIC system (Dionex Corp., Sunnyvale, CA, USA) was fitted with an IonPac AS-18 4 mm × 250 mm anion exchange column. A 25 μL sample loop was used to inject each KCN sample onto the column followed by an isocratic separation using 33 mM KOH (aq.) at a flow rate of 1 mL/min and a column temperature of 30 °C. Each KCN sample listed in Table 3 was analyzed starting the day after solution preparation (see Table 2) for a total of three separate analysis periods. The KCN samples for a given analysis period were analyzed in groups of 12. Each group of 12 KCN samples was preceded by the analysis of a DI water blank and three calibration standard solutions ran in triplicate. In addition, a calibration standard was analyzed in triplicate after every six sample analyses as a check standard. Each calibration standard consisted of eight anions at known concentrations. The anions were fluoride, chloride, nitrite, sulfate, bromide, nitrate, oxalate, and phosphate. Anions in each KCN sample that were in the calibration standards were identified by their retention times and quantified by a three-point calibration curve. The final concentrations of these identified anions were reported as μg/g of KCN stock. Anions present in the KCN samples but not in the calibration standards had their concentrations reported as total peak signal per gram of KCN or μS/g of KCN stock.

Table 2
KCN solutions.

Solution[a]	Date of preparation
1	19 January 2010
2	01 February 2010
3	21 April 2010

[a] Number is the second term in the sample label scheme: *stock-solution-duplicate*. It represents the specific date when a 1 g portion of KCN was taken from a stock bottle and dissolved to create an aqueous stock solution. Three duplicate aliquots from each stock solution were prepped and analyzed starting the following day.

Table 3
KCN samples (label scheme: *stock-solution-duplicate*).

A-1-1	B-1-1	C-1-1	D-1-1	E-1-1
A-1-2	B-1-2	C-1-2	D-1-2	E-1-2
A-1-3	B-1-3	C-1-3	D-1-3	E-1-3
A-2-1	B-2-1	C-2-1	D-2-1	E-2-1
A-2-2	B-2-2	C-2-2	D-2-2	E-2-2
A-2-3	B-2-3	C-2-3	D-2-3	E-2-3
F-1-1	G-1-1	H-1-1	B-3-1	E-3-1
F-1-2	G-1-2	H-1-2	B-3-2	E-3-2
F-1-3	G-1-3	H-1-3	B-3-3	E-3-3
F-2-1	G-2-1	H-2-1	D-3-1	F-3-1
F-2-2	G-2-2	H-2-2	D-3-2	F-3-2
F-2-3	G-2-3	H-2-3	D-3-3	F-3-3

A sample's label is deciphered using Tables 1 and 2 and knowing that *duplicate* (third term in label scheme) represents one of the three possible aliquots from a stock solution.

3.4. Chemometric analysis

HCA and PCA from PLS Toolbox 5.0 (Eigenvector Research, Inc., Manson, WA) were implemented with a personal computer running Matlab 7.0 (The MathWorks, Inc., Natick, MA). In addition, KNN and feature selection using the Fisher-ratio method and DCS were implemented in Matlab with code written in house.

4. Results and discussions

Two representative chromatograms obtained from the HPIC analysis of KCN samples originating from two different stocks (B and D) are depicted in Fig. 1. Most of the peaks depicted in the chromatograms were identified using the known retention times of the anions found in the calibration standards. The carbonate anion, while not included in the standards, was identified in the KCN chromatograms based on its retention time relative to the other identified peaks. Interestingly, a large cyanide peak is absent in the KCN chromatograms because most free cyanide ions form HCN in water due to HCN's low dissociation constant ($pK = 9.2$) [11]. This is why the cyanide ion is not detected directly by conductivity detection in HPIC analyses [11,12]. It was not feasible to confidently identify the cyanide peak or any other peaks labeled as unknowns in Fig. 1 without further analytical work. Regardless, the true identities of all anionic peaks in this study were not critical in terms of investigating the feasibility of using anions for forensic sample matching.

By comparing the two overlaid chromatograms in Fig. 1, one can see the potential of using anions to "fingerprint" different KCN stocks based on the obvious difference in anionic peak intensities between the two chromatograms. Assuming that at least some of the differences in peak intensities are reproducible among samples from the same stock or from the same manufacturer, then it may be possible to fingerprint a specific stock or country of origin. In order to test this hypothesis, a training set of KCN samples was

Table 4
Training matrix (μg/g KCN for identified anions except CO_3; μS/g KCN for unknowns and CO_3).

Sample	unk1	unk2	Cl	unk3	NO_2	unk4	CO_3	unk5	SO_4	OX	PO_4
B-1-1	7027.81	19,485.71	18.76	13,024.25	30.78	6508.19	4068.40	915.22	183.46	114.89	310.08
B-1-2	6449.22	16,487.31	14.08	12,821.22	24.46	6496.50	4124.36	908.65	176.96	107.79	314.47
B-1-3	6353.03	15,912.65	13.85	12,748.74	22.74	6557.82	4214.34	905.91	175.83	109.29	314.09
B-2-1	6126.91	15,436.43	10.43	12,142.60	24.30	6600.58	4150.60	754.66	182.21	115.74	316.11
B-2-2	6019.10	14,256.95	8.95	12,198.00	21.68	6558.43	4338.93	782.03	175.95	113.71	317.53
B-2-3	6022.42	14,434.32	7.77	12,118.94	21.51	6504.01	4508.93	777.80	174.72	114.43	316.13
C-1-1	7027.81	16,003.74	27.83	12,980.61	7.74	10,718.89	9340.56	2677.58	88.88	22.62	0.00
C-1-2	6449.22	15,875.69	27.36	12,921.75	7.77	10,621.51	9377.63	2674.52	87.97	22.44	0.00
C-1-3	6353.03	15,929.75	27.52	12,822.41	7.57	10,579.01	9563.30	2675.73	87.73	22.09	0.00
C-2-1	6126.91	15,647.29	18.92	13,215.68	8.01	11,809.05	6960.84	2551.08	91.48	22.62	0.00
C-2-2	6019.10	15,550.16	18.45	13,153.06	7.95	11,908.26	6997.43	2617.43	89.50	22.25	0.00
C-2-3	6022.42	15,616.00	18.99	13,054.91	7.80	11,929.26	6953.46	2661.73	87.91	22.43	0.00
D-1-1	1293.05	5662.88	58.98	12,811.67	2.23	543.69	7029.89	0.00	4.02	3.52	0.00
D-1-2	1313.61	5793.82	59.39	12,773.76	2.70	526.00	7130.26	0.00	3.71	2.94	0.00
D-1-3	1312.89	5791.18	58.03	12,769.52	2.37	531.41	7222.14	0.00	2.61	2.97	0.00
D-2-1	1028.82	5273.42	50.72	13,093.06	1.89	950.15	10,603.13	0.00	2.86	2.94	0.00
D-2-2	1014.20	5324.08	50.98	13,085.61	2.08	920.97	10,717.52	0.00	2.69	1.85	0.00
D-2-3	997.18	5339.64	50.79	12,964.05	2.01	910.68	11,027.85	0.00	2.56	1.57	0.00
E-1-1	8330.43	26,175.67	9.05	13,286.69	10.26	3124.66	8807.49	1155.66	210.81	63.58	371.40
E-1-2	8280.32	25,851.47	8.79	13,023.15	9.02	3134.06	8887.35	1163.75	207.21	63.11	365.98
E-1-3	8251.93	25,863.75	9.31	12,978.89	8.97	3123.36	8869.77	1165.66	207.93	62.56	365.45
E-2-1	7981.39	13,633.55	2.21	13,043.42	8.33	3660.34	5988.51	1266.02	211.69	60.96	358.55
E-2-2	7964.25	13,400.09	2.48	12,937.81	7.61	3664.66	6104.78	1278.10	211.20	62.10	357.73
E-2-3	7985.91	13,499.22	2.32	12,927.57	7.92	3656.56	6366.81	1267.06	211.32	60.96	365.19
F-1-1	425.00	38,768.03	93.34	12,307.60	6.95	8692.63	12,756.64	2402.97	91.22	36.91	0.00
F-1-2	418.89	38,824.73	93.07	12,351.27	6.97	8702.62	12,648.71	2422.45	90.97	36.45	0.00
F-1-3	451.84	39,260.50	94.21	12,300.50	7.74	8693.42	12,844.46	2405.37	92.53	36.71	0.00
F-2-1	360.00	15,503.36	30.61	12,022.88	7.22	10,203.47	5819.65	2656.46	93.17	30.15	0.00
F-2-2	352.49	15,662.12	30.50	12,133.14	7.60	10,158.58	5895.10	2660.55	92.85	29.82	0.00
F-2-3	371.13	15,828.06	31.42	12,048.75	8.27	10,118.28	5994.93	2685.27	93.09	30.06	0.00

created. This training set consisted of 30 samples that were all the B, C, D, E, and F samples from their respective 1 and 2 KCN solutions (see Table 2). The 11 anion concentrations from each of the 30 training samples were assembled into a data matrix shown as Table 4.

Table 4 shows some obvious differences in the anionic impurity profiles among KCN samples from different sources. For instance, only the KCN samples from the Belgium and German stocks (B and E) had detectable levels of the phosphate ion while the KCN samples from the Czech stock (D) had much lower sulfate concentrations and no detectable levels of unk5. There were other differences among the KCN samples that related back to their stock source. Those differences and the reproducibility of the differences can be effectively illustrated by HCA.

The dendrogram from HCA of the training set is shown in Fig. 2. The training set was autoscaled prior to HCA primarily to compensate for having different concentration units between identified and unknown anions. All KCN samples in the dendrogram, except for

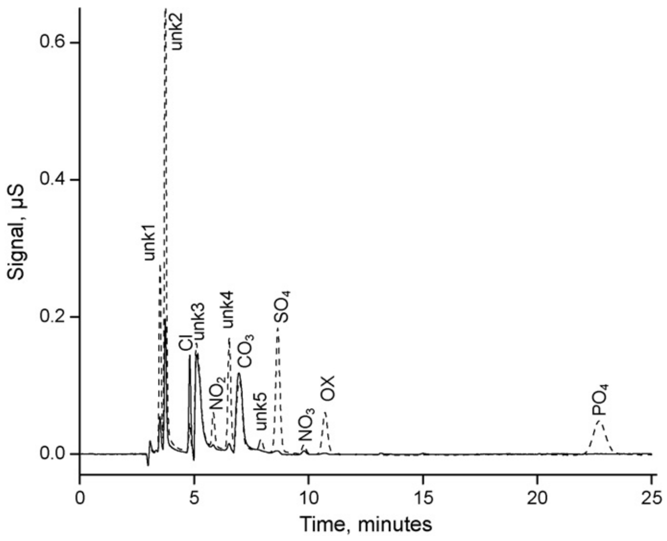

Fig. 1. Overlaid HPIC chromatograms from two different KCN stock samples known as B-1-1 (dashed) and D-1-1 (solid). The identified anions are chloride (Cl), nitrite (NO_2), carbonate (CO_3), sulfate (SO_4), nitrate (NO_3), oxalate (OX), and phosphate (PO_4). The unknown anions are labeled as unk1, unk2, unk3, unk4, and unk5.

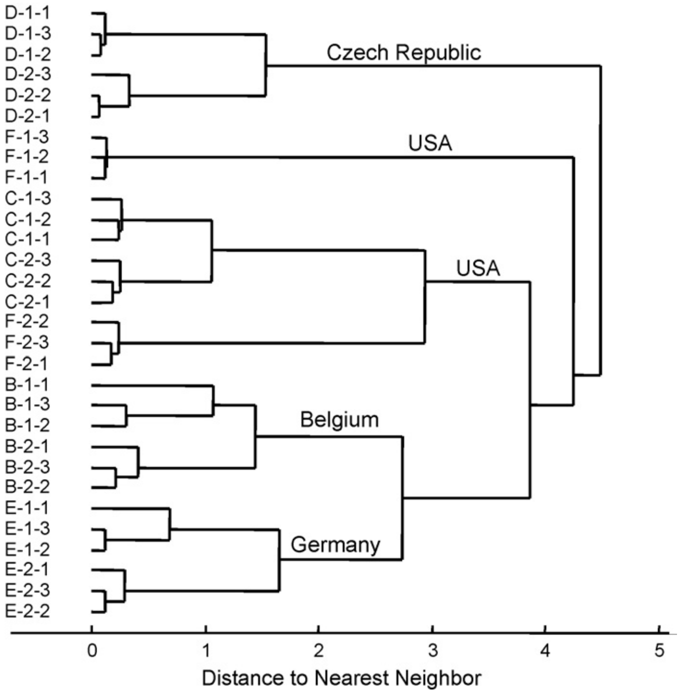

Fig. 2. Dendrogram from HCA of training set using all 11-anion peaks. The training set was autoscaled prior to HCA.

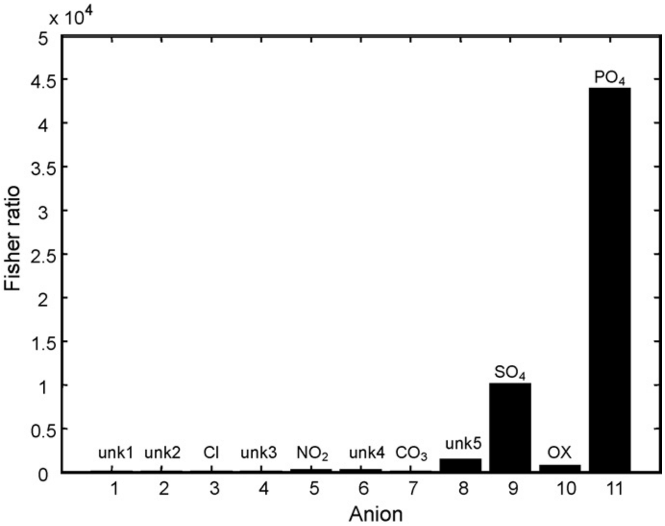

Fig. 3. Fisher ratios for 11 anions from the training set.

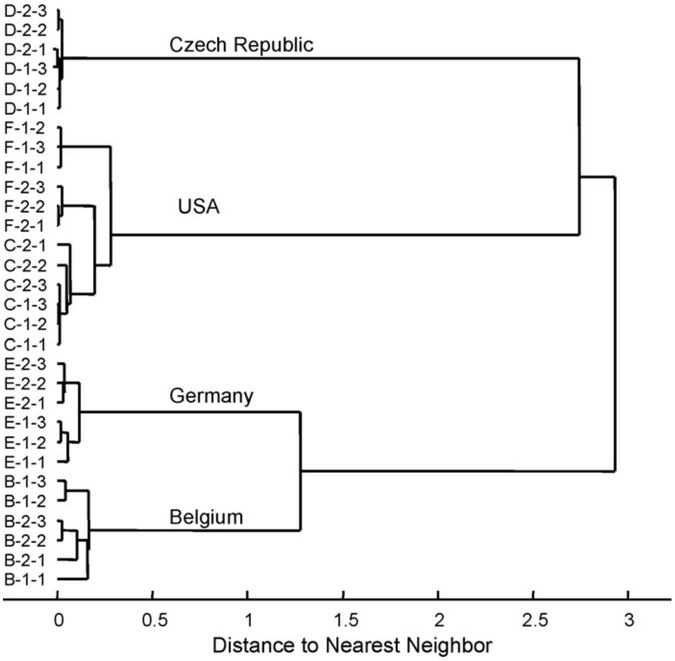

Fig. 4. Dendrogram from HCA of training set using peaks for unk5, SO_4, OX, and PO_4. The training set was autoscaled prior to HCA.

the three duplicate F-1 samples, formed a subgroup with samples from the same stock or country. For example, all samples from the Belgium stock (B) formed a subgroup that was linked to the subgroup of samples from the German stock (E). The linkage between the Belgium and German subgroups was excepted because the German and Belgium samples were the only ones having detectable phosphate levels. As previously mentioned, the F-1 samples did not cluster themselves with the other F-2 samples nor with the other United States samples (C1 and C2 samples). Hence, feature selection was a logical next step to determine if a subset of anions would provide better clustering particularly for the F samples.

Feature selection using the Fisher-ratio method was performed by first calculating the Fisher ratio for all 11 anions in the training set. Fig. 3 depicts the *F*-ratios for all 11 anions using the data from the training set. The *F*-ratios were based on four known sample classes: (1) United States C and F samples, (2) Belgium B samples, (3) Czech Republic D samples, and (4) Germany E samples. As seen in Fig. 3, not all 11 anions were relevant in terms of distinguishing classes as illustrated by the few peaks that have relatively larger *F*-ratios. In order to select the optimum anion subset, the DCS was calculated for all four classes as a function of Fisher-ratio threshold starting with a value of zero. All subsequent Fisher-ratio thresholds were increased such that each increment eliminated one anion from the previous anion subset. In this approach, as the Fisher-ratio threshold increased, the anions with the smallest *F*-ratios were progressively eliminated. Table 5 depicts the median DCS for the six possible class pairs in the training set as a function of anion subsets. For brevity, only the subsets containing eleven, six, four, three, two, and one anions are listed in Table 5. Sulfate rather than phosphate is listed as the one-anion subset because it was the anion with the largest *F*-ratio for which a DCS value could be calculated for all pairs. Table 5 also lists the DCS for the B and E pair because these two classes were the most similar and therefore the most difficult to separate. As shown in Table 5, the original 11 anions do not provide the best DCS values, i.e., the largest values. Indeed, the median DCS continued to increase as the number of anions was reduced with sulfate and phosphate providing the largest median DCS. However, only focusing on the median DCS was misleading because it ignored the DCS between the critical class pair, i.e., B and E. The optimal anion subset for this critical pair was four anions: unk5, sulfate, oxalate, and phosphate. The critical class pair was the deciding factor in finding the optimal anions or features for overall class separations. The effectiveness of these four anions for sample clustering based on country of origin is illustrated well by Fig. 4. Fig. 4 is the dendrogram obtained by HCA of the autoscaled training set using concentrations from only unk5, sulfate, oxalate, and phosphate. One can clearly see much tighter clusters within a class and greater distance among different classes when Fig. 4 is compared with the dendrogram obtained using all 11 anions (see Fig. 2). This is also illustrated by the PCA scores plot

Table 5
Anion influence on degree-of-class separation (DCS).

Anions	Median DCS	DCS for B and E
All 11	7.8	9.1
NO_2, unk4, unk5, SO_4, OX, PO_4	19.9	9.5
unk5, SO_4, OX, PO_4	49.6	53.5
unk5, SO_4, PO_4	56.4	33.7
SO_4, PO_4	118.8	22.7
SO_4	99.1	19.6

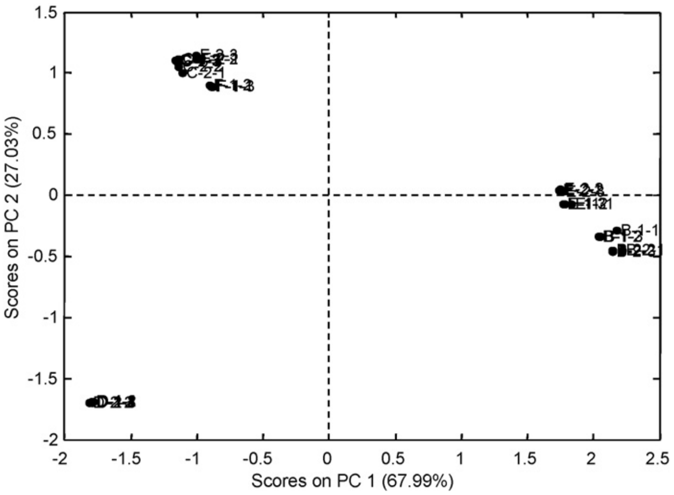

Fig. 5. PCA scores plot for the training set using anions unk5, SO_4, OX, and PO_4. The training set was autoscaled prior to PCA. Each training sample is labeled.

Table 6
KNN classification results for the test set.

Sample[a]	True class	Predicted class (11 anions)	Nearest neighbors[b] (11 anions)	Predicted class (four anions)	Nearest neighbors[b] (four anions)
A-1	United States	United States	CCCCE	United States	FFFFF
A-2	United States	United States	FFFDD	United States	FFFFC
H-1	United States	United States	FFFEF	United States	FFFFF
H-2	United States	United States	CCCCC	United States	CCCCF
G-1	United States	United States	FCCFC	United States	CFFFC
G-2	United States	United States	FFFCC	United States	CFFFC
F-3	United States	United States	FFFCC	United States	FFFFF
B-3	Belgium	Germany	EEEBB	Belgium	BBBBB
E-3	Germany	Germany	EEEEE	Germany	EEEEE
D-3	Czech Republic	Czech Republic	DDDDD	Czech Republic	DDDDD

[a] Analyte concentrations from three duplicates of a given sample were averaged.

[b] The stock designation for the nearest KCN samples listed left to right starting with the closest sample.

for the training set depicted in Fig. 5. The class separation depicted in Fig. 5 between E and B is optimal and the separations among the other class pairs, while not optimal for all class pairs, appears very good.

In order to objectively evaluate the effectiveness of the four selected anions for sample matching, a test set was projected into the scores plot to determine if test samples fall near members of their known respective classes. The test set consisted of concentrations for unk5, sulfate, oxalate, and phosphate for 10-test samples: A-1, A-2, B-3, D-3, E-3, F-3, G-1, G-2, H-1, and H-2. The anion concentrations of three duplicate KCN samples from a specific KCN solution were averaged to produce a test sample. This was accomplished for brevity and because sample variability within a stock is much larger than variability due to sample preparation and analysis. Fig. 6 depicts the plotted score values for both the training samples and the projected test samples. As seen in this scores plot, the test samples clearly fall closest to training samples from the same class. It is worth mentioning that while there appears to be significant distance between the United States test samples (C and F) and most of the United States training samples (A, G, and H), all training samples from the United States fell within the 99% confidence-interval ellipsoid of the test set generated by PCA of the United States test samples. All other non-United States samples were outside the ellipsoid.

KNN was also applied in order to objectively match the test samples to the training samples according to their distances from one another. This was accomplished by calculating the Euclidean distance between each test sample and all training samples using the autoscaled concentrations of unk5, sulfate, oxalate, and phosphate. The top five nearest neighbors ($K = 5$) were used to determine the class, i.e., country, for each test sample. A test sample was matched to the class with the greatest number of known class members that made up the five nearest neighbors to the test sample. $K = 5$ was selected because it was the largest odd number that did not exceed the number of known samples in each stock class ($n = 6$). A large K provided more confidence in classification, while an odd number K ensured that a test sample was only matched to one class in potential cases where an even number of nearest neighbors were equally split among two or more classes. The classification results for KNN are provided in Table 6 including those results obtained using all autoscaled 11 anions. It is quite clear by comparing the results of four versus 11 anions that feature selection was necessary to correctly classify all 10-test samples. In addition, greater confidence in classification was obtained using the four feature-selected anions. For example, both approaches correctly classified samples A-2 and F-3; however, greater confidence in classification was obtained using the four anions which had all five nearest neighbors for A-2 and F-3 belonging to the correct class versus three-out-five neighbors for the 11-anion approach.

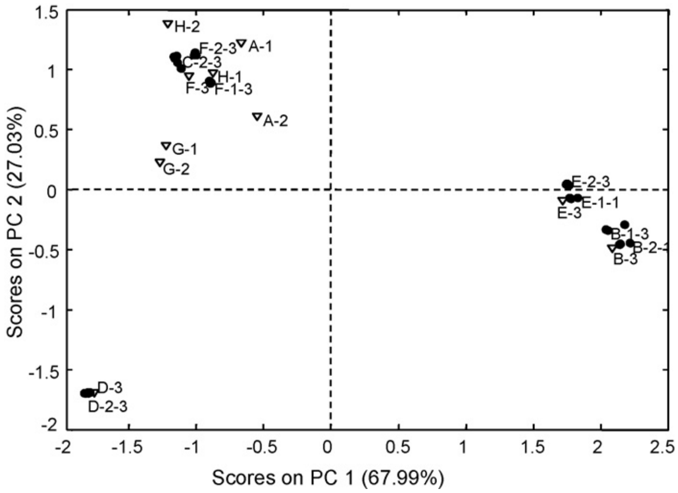

Fig. 6. PCA scores plot for the training set and projected test set using anions unk5, SO_4, OX, and PO_4. All tests samples (▽) and some of the training samples (●) are labeled for clarity.

5. Conclusion

This report demonstrated the value of anionic impurity profiles obtained by HPIC for sample matching KCN samples according to their source and the value of feature selection for obtaining reliable forensic signatures. The concentrations of sulfate, oxalate, phosphate, and an unknown anion named unk5 were used to successfully match a test set of KCN samples to their country of origin. While test samples from three different KCN stocks were correctly determined to be manufactured in the United States, this was not fully demonstrated for KCN samples from the other three countries because only one stock per country was obtained. However, the KCN samples from these countries were all correctly matched back to their original stocks because each stock had a unique anionic impurity profile. Future studies will involve obtaining more stocks from these countries in order to determine if the unique impurity profiles are dependent on the country of origin as shown for the United States stocks. Our hypothesis is that these anionic impurities are dependent on the geographic location of manufacture given that the potassium hydroxide, water, air and dust present during KCN syntheses have anions that are characteristic of the local environment. Regardless of what is determined in the future, this study also demonstrated the application of Fisher ratios and DCS for feature selection in order to improve the accuracy and confidence of sample classification.

Acknowledgement

Funding for this work provided by the Science and Technology Directorate, U.S. Department of Homeland Security.

References

[1] M.E. Keim, Prehosp. Disaster Med. 21 (2006) s56–s60, Available at http://pdm.medicine.wisc.edu (serial on the Internet).
[2] A.S. Khan, A.M. Levitt, M.J. Sage, Morb. Mortal. Wkly. Rep. 49 (RR04) (April 21, 2000) 1–14.
[3] A. Rubo, R. Kellens, J. Reddy, N. Steier, W. Hasenpusch, Ullmann's Encyclopedia of Industrial Chemistry: Alkali Metal Cyanides, Wiley-VCH, Weinheim, Germany, 2005.
[4] J.C. Hoggard, J.H. Wahl, R.E. Synovec, G.M. Mong, C.G. Fraga, Anal. Chem. 82 (2010) 689.
[5] C.G. Fraga, B.H. Clowers, R.J. Moore, E.M. Zink, Anal. Chem. 82 (2010) 4165.
[6] Rules of Origin, Title 19, Part 102, Code of Federal Regulations. Available at http://www.gpoaccess.gov/cfr/index.html.
[7] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, 1998.
[8] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, John Wiley & Sons, Inc., New York, 1986.
[9] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225.
[10] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101.
[11] M. Nonomura, Anal. Chem. 59 (1987) 2073.
[12] S.V. Karmarkar, J. Chromatogr. A 956 (2002) 229.