-+

**Ash**

*Samples*. Jim needs to write how they were collected and analysed.

Data were pretreated application a Savitzky-Goulay algorithm (7 point smooth, quadratic, first derivative) to minimize the impact of the XRF baseline on subsequent analyses. A rough variable selection was performed by visually selecting ranges of energies spanning each of the 17 XRF peaks from the ensemble spectra. This reduced the length of each XRF spectrum from 2049 unique channels to 441 channels employed subsequent classification models. Each reduced spectrum was normalized to unit area to account for variability in ash loading and sample placement across all collected spectra.

For each classification model, all the replicate spectra of approximately 1/4th of the samples from each class were removed to form a test set. The 129 spectra in the training set were comprised of triplicate XRF spectra from 8 Kaibab limestone ash samples, 16 Moenkopi sandstone ash samples, and 19 Basalt/Andesite ash samples. The 39 test set spectra were collected in triplicate from 2 Kaibab limestone ash samples, 5 Moenkopi sandstone ash samples, and 6 Basalt/Andesite ash samples.

Figure 1

*Preliminary Analyses*. The classification challenge for determining soil type from tree ash lies in the large variability of the XRF signal within each class of soil. Comparing the mean ash spectrum from the 30 ‘Kaibab Limestone’, 63 ‘Moenkopi Sandstone’, and 75 ‘Basalt/Andesite’ samples indicates a unique XRF signature from each soil of origin. (**Figure 1a**) overlaying the 95% confidence interval for a sample from each class as determined by the standard deviation of all spectra in a class demonstrates how the natural spread of the data within a class overlaps the mean spectra of other classes. (**Figure 1b-d**) Concurrently, calculation of variance between the means of the three classes, mean variance within each of the three classes, and mean variance of the triplicate spectra from each ash pellet presents that the largest source of variance in affecting given spectrum is attributed to the natural spread of the spectra within a class. (**Table 1, column 2**) Superficially, application of EPO appears to greatly increase the capability to accurately classify the ash samples based on the three types of source soil. Without EPO, two-thirds of the variance of each individual sample is attributed to random sources, while with a full rank EPO, 58% of the sample variance comes from the class means. (**Table 1, row 2**) Pictorially, a scores plot of the first 2 PCs without EPO shows, at best, overlapping zones of the three classes while a scores plot of the first 2 PCs with full rank EPO presents nearly complete separation. (**Figure 2a and c**) The same degree of improvement is evident in the HCA dendrograms where application of full-rank EPO results in only 2 ‘Basalt/Andesite’ samples being associated with the ‘Moenkopi Sandstone’ sample.



Figure 2

**Table 1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Treatment** | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| Variance of class means | 2.48x10-4 (32.7%) | 2.11x10-4 (48.8%) | 9.66x10-6 (35.0%) | 1.33x10-6 (58.4%) |
| Mean variance within a class | 5.08x10-4 (67.3%) | 2.21x10-4 (51.1%) | 1.78x10-5 (64.5%) | 4.33x10-7 (19.0%) |
| Mean variance of replicates | 1.02x10-6 (0.013%) | 6.56x10-7 (0.15%) | 1.37x10-7 (0.50%) | 5.15x10-7 (22.6%) |
| PCA | # PC: 8  Cum Var: 95.2% | # PC: 7  Cum Var: 90.9% | # PC: 3  Cum Var: 44.4% | # PC: 3  Cum Var: 15.0% |
| KNN(1) Misclassified (129:39 split) | 0 Cal; 11 Pred | 0 Cal; 13 Pred | 13 Cal; 10 Pred | 62 Cal; 12 Pred |

However, the inherent concern of EPO eliminating useful variance not associated with in-class ‘clutter’ becomes apparent when more deeply considering the PCA and KNN results. In

**Table 2: PLS-DA Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** |  | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| ‘Kaibab’ | 90% +  80% - 90%  50% - 80%  < 50% | 6  0  0  0 | 6  0 (3 false pos.)  0  0 | 3 (2 false pos.)  0 (0 false pos.)  0 (1 false pos.)  3 | 2  0  0  4 |
| ‘Moen-Kopi’ | 90% +  80% - 90%  50% - 80%  < 50% | 9  3 (6 false pos.)  0  3 | 10  2 (6 false pos.)  0  3 | 9  0  0 (6 false pos.)  6 | 6 (2 false pos.)  1  0 (1 false pos.)  13 |
| ‘Basalt/Andesite’ | 90% +  80% - 90%  50% - 80%  < 50% | 5  4  9  0 | 18  0  0  0 | 18  0  0  0 | 14 (1 false pos.)  0  2  (5 not classified) |
| Heirarchical  Remove ‘B/A’  Classify ‘K’ v. ‘M’ | 90% +  80% - 90%  50% - 80%  < 50% | 11  1  3  6 | 12  0  3  6 | 14  2  0  5 | 11  2  0  2 (6 not classified) |

Performing SVM with full rank EPO as a 3-class problem returns all 39 samples in the test set as having between 0.663 and 0.668 probability of belonging to Moen-Kopi.

**Table 2b: PLS-DA Classification w/ Discretization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** |  | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| ‘Kaibab’ | 90% +  80% - 90%  50% - 80%  < 50% | 6 (1FP)  0  0  0 | 6 (1FP)  0  0  0 | 6 (1FP)  0  0  0 | 6 (6FP)  0  0  0 |
| ‘Moen-Kopi’ | 90% +  80% - 90%  50% - 80%  < 50% | 12 (2 FP)  0 (1 FP)  0  3 | 12 (2 FP)  0 (1 FP)  0  3 | 12 (2 FP)  0 (1 FP)  0  3 | 9 (1 FP)  0  0  6 |
| ‘Basalt/Andesite’ | 90% +  80% - 90%  50% - 80%  < 50% | 18  0  0  0 | 18  0  0  0 | 18  0  0  0 | 18  0 (1 FP)  0  0 |
| Heirarchical  Remove ‘B/A’  Classify ‘K’ v. ‘M’ | 90% +  80% - 90%  50% - 80%  < 50% |  |  |  |  |

**Table 3: SVM Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** |  | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| ‘Kaibab’ | 90% +  80% - 90%  50% - 80%  < 50% | 0  3  2  1 | 0  0  5  1 | 0  0  0 (4 false pos.)  6 | 0  0  0  6 |
| ‘Moen-Kopi’ | 90% +  80% - 90%  50% - 80%  < 50% | 9  3 (3 false pos.)  0 (3 false pos.)  3 | 9 (1 false pos.)  3 (2 false pos.)  0 (3 false pos.)  3 | 9  0  1 (6 false pos.)  5 | 0  0  1 (2 false pos.)  14 |
| ‘Basalt/Andesite’ | 90% +  80% - 90%  50% - 80%  < 50% | 14  4  0  0 | 13  5  0  0 | 18  0  0  0 | 0  0  13 (1 false pos.)  5 |
| Heirarchical  Remove ‘B/A’  Classify ‘K’ v. ‘M’ | 90% +  80% - 90%  50% - 80%  < 50% | 9  3  6  3 | 10  2  2  7 | 9  0  3  9 | 0  0  15  6 |

**Red Glasses**

*Samples.* Soft glass (coefficient of expansion 104) samples were purchased from Devardi Glass of the type appropriate for lampworking projects. The glasses were a “set of mixed reds” of various hues, both opaque and transparent, each approximately 25 cm long and 6 – 10 mm in diameter. Inspection of the 21 rods received and comparison to the Devardi catalog color chart indicates that the set contains 2 duplicate and 1 triplicate colored rods. Other rods might be duplicate colors, but these were not readily identified by visual inspection. Prior to analyses, the rods were designated ‘A’ through ‘U’. It was determined that ‘I’/’G’ and ‘K’/’S’ were duplicate pairs and ‘E’/’J’/’R’ was a triplicate pair. Consequently, the set of 21 rods spans at most 17 unique colors.

LIBS spectra were collected with a SciAps Z300 hand-held LIBS analyzer. Samples were aligned by manually holding each glass rod in the v-shaped alignment groove on the Z300 faceplate. Twelve spectra were collected were collected from random locations on each glass rod. Three spectra of each rod was collected in one sitting on 10/10. Nine spectra on rods ‘A’ through ‘R’ were collected on 10/29. The remaining 9 spectra on rods ‘S’ through ‘U’ were collected on 11/18. Consequently, each set of 12 spectra spans at least 2 collection periods. In total, 252 spectra were collected for this data set.

As with the Dalbergia samples, the LIBS baseline contribution was minimized by a Savitzky-Goulay algorithm (7 point smooth, quadratic fit, first derivative). The derivatized signal at every wavelength was transformed by applying the square root of its absolute value; this normalized the error distribution across peaks of vastly different scale. Each spectrum was then normalized to unit area to account for efficiencies in placing the sample on the LIBS analyzer. Based on the mean spectra of the entire data collection, a threshold value of 0.004 units was determined to separate ‘baseline’ from ‘LIBS’ channels. In this manner the number of channels employed in each spectrum was reduced from 23,431 to 8,169. This is referred to as the ‘basic’ preprocessing treatment in the following LIBS analyses.

*Preliminary Analyses.* 1

**Table 4:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Treatment** | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| Variance of class means | 0.0775 (79.3%) | 0.0670 (95.7%) | 0.0544 (98.7%) | 0.0127 (99.9%) |
| Mean variance of replicates | 0.0161 (20.7%) | 0.00285 (4.3%) | 0.000724 (1.3%) | 0.0000229 (0.2%) |
| PCA | # PC: 7  Cum Var: 82.2% | # PC: 6  Cum Var: 80.6% | # PC: 5  Cum Var: 81.6% | # PC: 5  Cum Var: 79.8% |
| KNN Misclassified (9:3 split) | 1NN: 1 Cal; 1 Pred  3NN: 5 Cal; 1 Pred | 1NN: 2 Cal; 1 Pred  3NN: 7 Cal; 1 Pred | 1NN: 1 Cal; 1 Pred  3NN: 5 Cal; 1 Pred | 1NN: 0 Cal; 0 Pred  3NN: 0 Cal; 0 Pred |

Like discretization, External Parameter Orthogonalization (EPO) seeks to highlight the between-class variance of a dataset by reducing the within-class variance. Applying an increasingly aggressive EPO algorithm to remove the within class variance greatly decreases the within-class variance (So called ‘clutter’.) at the expense of slightly decreasing the between-class variance; overall the percent variance attributed to within-class clutter decreases from 20.7% without EPO to 0.02% with full rank EPO. (**Table 4, row 3**) Fewer PCs are needed to describe the data. (**Table 4, row 4**) The PC plots show increasingly sharper clusters with clean separation among classes. (**Figure 3c for example**). With full rank EPO, the HCA dendrogram shows clear delineation among the 21 putative classes. However, there is no real improvement in the KNN classification model short of full rank EPO application (**Table 4, row 5**)

Classification of the red glass rods by PLS-DA highlight the potential for EPO to create misclassifications by over-fitting the within-class variance. Consequently, EPO suggests class boundaries that are much tighter than are warranted by the natural variability within each class. Training a PLS-DA model on the 12 replicate spectra from one glass rod in each of the 17 unique red hues should construct a model that places the 12 replicate spectra from the duplicate/triplicate hues into the appropriate class. That is, rods ‘E’, ‘I’ and ‘K’ are included in the training set with the 14 unique colors and rods ‘G’, ‘J’, ‘R’, and ‘S’ form the training set. Prior analyses shows that the duplicate rods project near to each other in a both the PCA and HCA plots with distinction among the duplicates evident in higher dimension PCs and in the dendrogram following application of EPO filtering. (**Figure 3**)



Figure 3

**Table 5: PLS-DA Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** |  | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| ‘G’ | Factors:  Correct:  Indeterminate:  Wrong: | 1  12/12  0/12  0/12 | 1  12/12  0/12  0/12 | 1  12/12  0/12  0/12 | 1  0/12  12/12  0/12 |
| ‘J’ & ‘R’ | Factors:  Correct:  Indeterminate:  Wrong: | 5  24/24  0/24  0/24 | 4  24/24  0/24  0/24 | 3  24/24  0/24  0/24 | 3  1/24  23/24  0/24 |
| ‘S’ | Factors:  Correct:  Indeterminate:  Wrong: | 1  12/12  0/12  0/12 | 1  12/12  0/12  0/12 | 1  11/12  0/12  1/12 | 1  2/12 : 7/12\*  10/12 : 1/12  0/12 : 4/12 |
|  |  |  |  |  |  |

**Table 6: SVM Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** |  | **Basic** | **Basic + EPO(1)** | **Basic + EPO(6)** | **Basic + EPO(Full Rank)** |
| ‘G’ | 90% +  80% - 90%  50% - 80%  < 50% |  |  | 12  0  0  0 | 1  7  4  0 |
| ‘J’ & ‘R’ | 90% +  80% - 90%  50% - 80%  < 50% | 13  6  5  0 | 16  5  3 | 14  5  5 | 15  10  2  0 |
| ‘S’ | 90% +  80% - 90%  50% - 80%  < 50% | 2  5  5  0 | 2  7  3  0 | 3  6  3  0 | 3  4  1  4 |
|  |  |  |  |  |  |