Making Your BEST Case: Near-IR Spectral Identification of Soil

Yi Zou, Yu Xia, Angela R. Jones, and Robert A. Lodder
Department of Chemistry and College of Pharmacy
University of Kentucky
Lexington, KY 40536-0082

Soil analysis has played an important role in criminal investigations. It has been used successfully for purposes ranging from the recovery of stolen artifacts to definitive conclusions in murder cases.

Federal land in the United States often includes ancient sites formerly inhabited by Native Americans. Artifacts from these sites are valuable and are sometimes removed by looters who, when caught, claim that the items were obtained legally on private land. Elemental "fingerprinting" of public and private land, and comparison with soil samples obtained from the artifacts, has led to successful prosecution of artifact thieves (*1*).

The analysis of soil has frequently been used to assist in other criminal cases. One incident involved an odd set of circumstances in two different states separated by thousands of miles (*2*). Officials were left with few clues involving a homicide in Morrison, CO, and had to piece together a series of unconnected and unlikely events.

It began much like a mystery novel. A police officer driving along the highway spotted a car beside the road with no driver. Walking toward the car, the officer realized that the engine was running. Peering inside

REPORT

the car, he found bloodstains and a pair of glasses. The owner's fate and whereabouts were unknown.

Later, across the country in Atlantic City, NJ, a seemingly unrelated event occurred. A car was found burning at a salvage dump. The vehicle was identified as the same suspect car in the Colorado murder mystery, which was no longer in police custody. How the vehicle got from Colorado to New Jersey is unknown.

Analysis of soil from the car revealed four distinct soil layers. Three layers contained mineral grains indicative of the Rocky Mountain area near Denver. More than 360 soil samples from Morrison were gathered and compared with the soil found on the burnt automobile.

Authorities were able to make a connection between the murder scene and the assailant's attempt to destroy evidence. Soil samples were analyzed and used to find the body of the victim, buried 27 miles south of Denver. Further analysis revealed that the soil type was the same as that found on the victim's ranch. The assailant was convicted of murder and kidnapping.

Until recently, the cost of sophisticated soil tests—which include X-ray diffraction and electron microscopic techniques—has been prohibitive. The equipment has been too cumbersome to permit extensive on-site soil sampling. These "traditional" methods concentrate largely on inorganic constituents of the soil as a way to make a positive identification. However, inorganic salts sometimes dissolve in water and travel significant distances from their points of origin.

Near-IR spectrometry of soil samples makes it possible to include organic constituents in the "fingerprint" of the sample. In this REPORT, we describe the extended quantile BEST (bootstrap error-adjusted single-sample technique), which was used for the first time on an IBM PC-compatible computer to identify organic and inorganic constituents in soil samples from different sources. Until now, this algorithm has been available for use only on mainframes and supercomputers.

Theory

A population **P** in a hyperspace **R** represents the universe of possible spectrometric samples (the rows of **P** are the individual samples, and the

0003-2700/93/0365-434A/$04.00/0

columns are the independent information vectors such as wavelengths or energies). $\mathbf{P}^*$ is a discrete realization of $\mathbf{P}$ based on a calibration set $\mathbf{T}$, chosen only once from $\mathbf{P}$ to represent as nearly as possible all the variations in $\mathbf{P}$.

$\mathbf{P}^*$ is calculated using a bootstrap process by an operation $\kappa(\mathbf{T})$. $\mathbf{P}^*$ has parameters $\mathbf{B}$ and $\mathbf{C}$, where $\mathbf{C} = E(\mathbf{P})$ (the group mean of $\mathbf{P}$) and $\mathbf{B}$ is the Monte Carlo approximation to the bootstrap distribution (*3*).

Given two calibration sets $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ with an equal number of elements κ, it is possible to determine whether $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ are drawn from the same population even if the distance between them is < 3 SDs (standard deviations). Quantile–quantile (QQ) plots and a simple correlation test statistic are used (*3*).

$$\rho\left(\left\{\int_R \mathbf{P}_1^*\right\}, \left\{\int_R \mathbf{P}_1^*\right\} \bigcup \left\{\int_R \mathbf{P}_2^*\right\}\right) \quad (1)$$

A bootstrap method is employed to set confidence limits on ρ, the correlation coefficient. The central 68% confidence interval on ρ is also used to calculate σ_ρ, a distance in SDs that is sensitive to small differences in location and scale between $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$.

This approach to spectral analysis has significant advantages. More wavelengths can be used in the calibration than there are samples in the calibration set, without degrading the results. Full spectra can be used without down-weighting some of the information at certain wavelengths, reducing the possibility of missing something new that may appear in future samples. Also, the method is completely nonparametric, and the shape, scale, and skew of the soil sample distributions do not affect the quality of the results.

The analysis

Samples. In one experiment, 50 1-g soil samples were collected from five different locations (10 samples at each location). All samples were collected within Fayette County, KY. Group 1 samples were collected within 3 m of highway 68. Group 2 samples were collected within 3 m of a creek. Group 3 samples were collected within a 20-m circle on a hilltop. Group 4 samples were collected within a 20-m circle on a suburban lawn around a house, and group 5 samples were collected within a 3-m circle in a cultivated garden. The samples were stored in capped glass scintillation vials.

The samples were then uncapped and dried in an oven at 40 °C until their evaporative mass loss in successive weighings was < 1% of the sample mass, a process that required 12 h. Drying the samples prevented differences in water content alone from separating the different soil types. The samples were crushed into small particles in the vials with a glass rod and mixed by shaking, then scanned directly in the glass vials using the near-IR spectrometer.

Instrumentation. The samples were scanned on a Bran+Luebbe near-IR spectrometer. Sample spec-

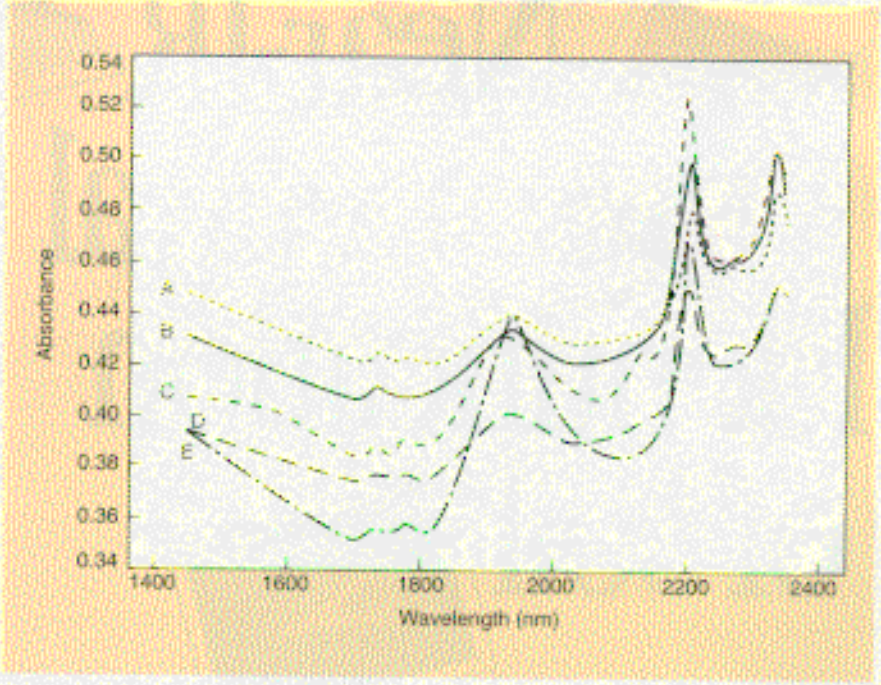

Figure 1. Five averaged spectra of 10 soil samples from each of five locations. A—creek, B—road, C—yard, D—garden, and E—hilltop.

Table I. Distance in SDs of ρ (from Equation 1) between soil samples obtained from different locations[a]

Test set	Calibration set				
	Road	Creek	Hilltop	Yard	Garden
Road		-0.074	4.515	6.665	**11.573**
Creek	**16.652**		13.257	25.885	**30.207**
Hilltop	**5.919**	**22.068**		12.586	14.961
Yard	5.309	19.746	7.644		18.588
Garden	10.729	26.406	**24.631**	**22.199**	

[a]Actual distance reported by the algorithm is shown in bold.

Table II. Confidence limits[a] on correlation for each soil set

Group no.	Sample origin	Limit	Mean ρ	SD
1	Road	0.96529	0.98481	0.0097621
2	Creek	0.97466	0.98731	0.0063259
3	Hilltop	0.96408	0.98725	0.011584
4	Yard	0.97559	0.98817	0.0062872
5	Garden	0.97166	0.98506	0.0067034

[a]98% level calculated from the mean and SD of 100 bootstrap samples of each training set.

tra were analyzed on a battery-powered IBM PC-compatible notebook computer with an 80486DX CPU and 12 Mb RAM. All software was written in an alpha-test version of Speakeasy IV Eta for the PC (Speakeasy Computing Corp., Chicago, IL).

Figure 1 shows the mean near-IR spectra of the five types of soil samples. These spectra indicate chemical differences among the soils from different sources. Soil samples from the hilltop and the suburban lawn produced diffuse-reflectance spectra with the lowest number of scattering events per unit volume and the largest peaks in the raw data.

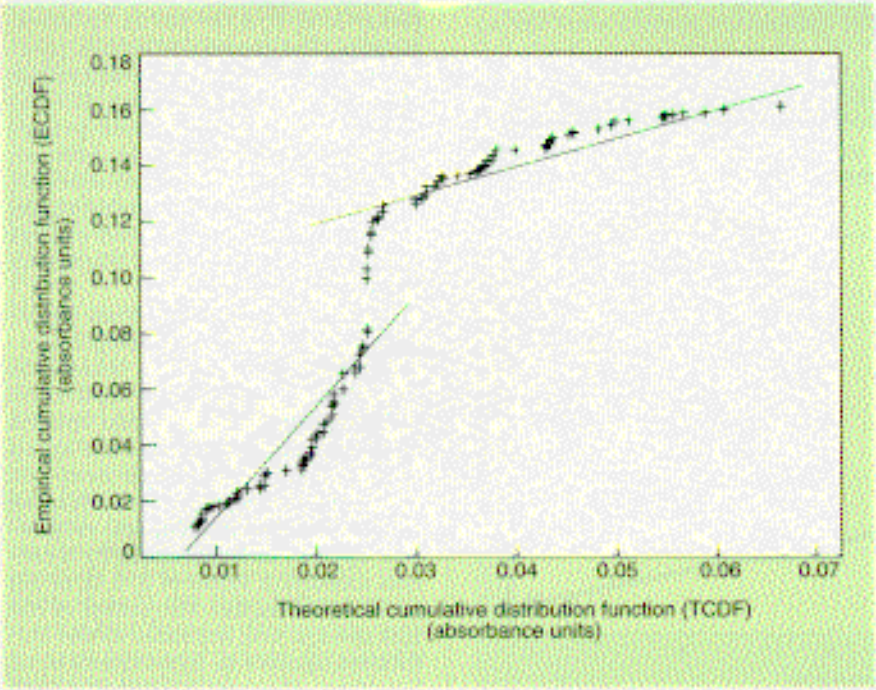

Figure 2. QQ plot showing that spectra of soil samples obtained near the road have greater variability than those obtained from the garden.

The line segment with the larger slope is generated by the spectra of samples from near the road. The TCDF is obtained from soil samples collected near the road, and the ECDF is obtained from the road and garden samples.

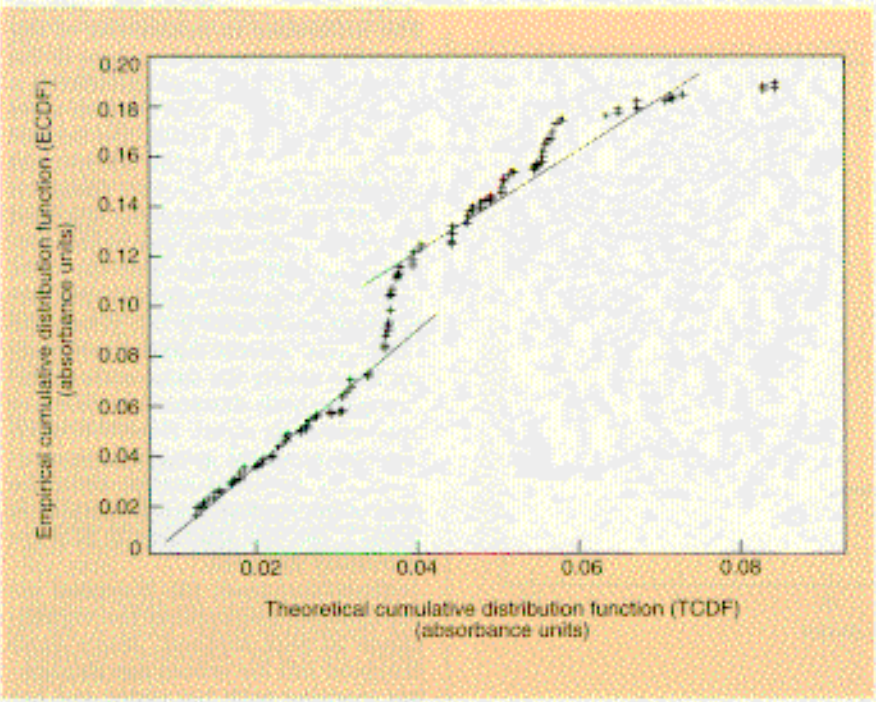

Figure 3. QQ plot showing that the variability in spectra of soil samples from near the road and from the hilltop is very similar.

The TCDF represents samples obtained from the hilltop, and the ECDF is obtained from hilltop and road samples.

Table I shows the distance in SDs between each soil group and every other soil group. The calculation of each distance (in SDs of the correlation coefficient) required 1.3 s on the PC. Distances in SDs are always measured with reference to a calibration set of spectra. When two sets have different locations, scales, and skews in spectral hyperspace, the distance between the two spectral sets will differ—depending on which spectral set is selected as the calibration set.

In Table I, each set of soil spectra serves as both a calibration set and a test set. In routine use, however, the algorithm reports only the largest of the two values (shown in bold type in the table) for the comparison of two sample groups—in effect producing only one triangle of the rectangular table as output. Both the upper and the lower triangles must be taken into account, because two soil groups are considered to be drawn from the same population only when their mutual correlation coefficient is greater than *both* bootstrap confidence limits on the individual calibration groups.

The confidence limit on ρ for each soil group at the 98% level is shown in Table II. The calculation of each confidence limit required 30 s on the PC. This time frame is not a large burden, because a confidence limit is calculated only once for each training set and can be used repeatedly to analyze subsequent new samples. The biggest difference in SDs exists between group 2 (from soils near the creek) and group 5 (from a garden), which has a small correlation coefficient (0.785). The smallest difference in SDs exists between group 1 (from a road) and group 3 (from a hilltop), which has a large correlation coefficient (0.933). On inspection, the hilltop soil and the soil from beside the highway contained the largest amounts of inorganic matter in the form of gravel and sand.

The differences among soil types are visualized in QQ plots (see Figures 2 and 3). The intercepts of the line segments in QQ plots provide information on the location of spectral sets in hyperspace, and the slope of the line segments provides information about the scale of the groups.

The QQ plot in Figure 2 shows that the spectra of soil samples obtained near the road have greater variability than those of samples obtained from the garden. The QQ plot in Figure 3 shows that the variability in spectra of soil samples from near the road and from the hilltop are very similar.

Figure 4 shows how ρ from Equation 1 drops as the test sample spectral cluster increases in scale with respect to the calibration cluster. The calibration spectra were the actual spectra of the first group of soil samples—those obtained near the road. The test spectra were formed simply by increasing the scale of the calibration spectra in hyperspace by a certain amount (a multiplicative factor of 2, 3, 5, or 10) at 19 wavelengths. The center of the test spectra was the same as the calibration spectra in hyperspace; thus, no location change occurred (both groups had the same average spectrum). The algorithm is thus quite sensitive to changes in spectral group scale, independent of any location change (which would indicate a change in the average spectrum).

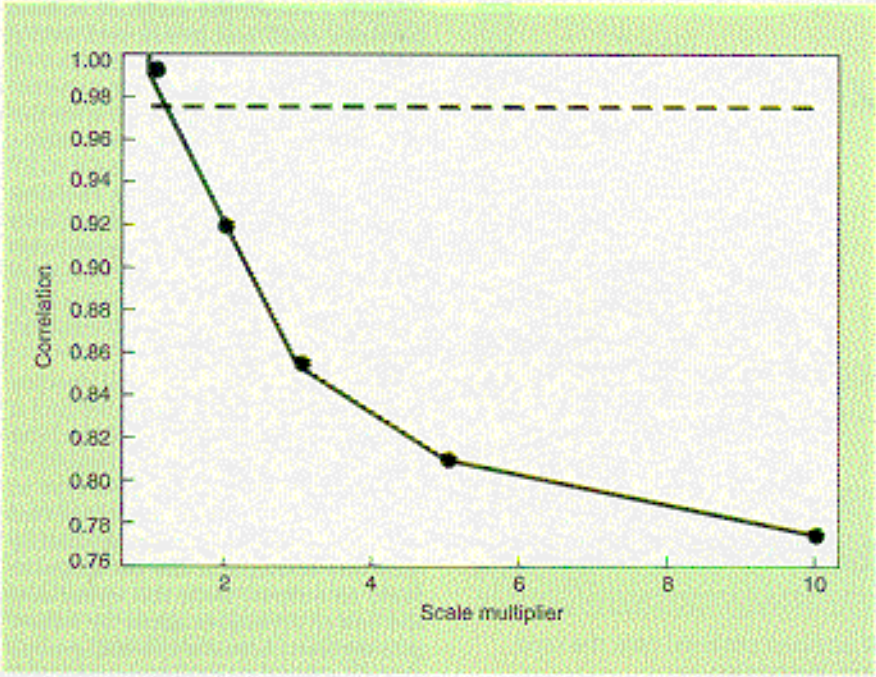

Figure 4. Plot showing how ρ from Equation 1 drops as the test sample spectral cluster increases in scale with respect to the calibration cluster.

The dashed line is the 98% confidence limit on ρ.

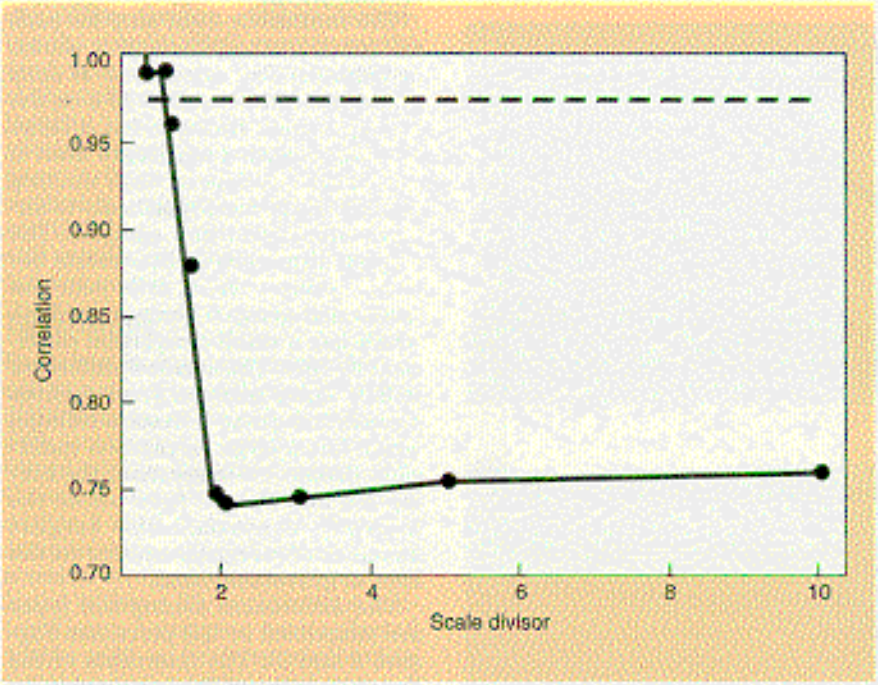

Figure 5. Plot showing that ρ from Equation 1 drops much more rapidly as the test sample spectral cluster decreases in scale with respect to the calibration cluster.

The dashed line is the 98% confidence limit on ρ.

Figure 5 shows how ρ from Equation 1 drops as the test sample spectral cluster decreases in scale with respect to the calibration cluster. As in Figure 4, the calibration spectra were the actual spectra of the group of soil samples obtained near the road. The test spectra were formed by dividing the scale of the calibration spectra in hyperspace by a certain amount (a factor of 2, 3, 5, or 10) at 19 wavelengths. The center of the test spectra was the same as the calibration spectra in hyperspace, so no location change occurred (both groups had the same average spectrum).

The algorithm is therefore extremely sensitive to a reduction in the scale of the test spectra—even more sensitive than it is to an increase in the scale of the test spectra. The difference in sensitivity of the algorithm to size relationships is the reason the algorithm calculates a distance in SDs of ρ in the two ways shown in Table I. Using the largest group of spectra as the theoretical cumulative distribution function (TCDF) in the empirical QQ plot puts the bends in the line in the center of the plot instead of at the ends, producing a greater effect on the correlation coefficient ρ (*3*). When the calibration spectra and the test spectra have nearly the same scale in hyperspace, both distances in SDs are approximately equal. However, when the two spectral groups have dramatically different scales, as in the case of the road and creek samples in Table I, the two distances in SDs are extremely different.

One final case, the Haygood investigation, exemplifies the significance of this method. Mary Jane Haygood left for school one September morning with her sister and her boyfriend, setting in motion events that would lead to the discovery of her death only days later (*4*).

Her mother had forbidden the re-

lationship between Haygood, 15, and Gerald Pate, 22. Pate decided to get even with her mother for her interference. He took Haygood to his secluded barn, isolated from the farm community's only road, and talked of marriage. When the two had a disagreement, she threatened to tell authorities about the numerous robberies Pate had committed. Pate choked her to death; he later bragged about using a hold he had learned in the Marines. He buried her naked body with her school notebook. Later, when Haygood was reported missing, the police went to Pate's farm, where they found him hiding in a closet.

Although Pate confessed to the crime, his mental state prevented him from remembering where he had disposed of the body. During the investigation, a shovel was discovered in the barn among some blankets and Pate's Marine pack. Upon examination, authorities found fresh soil on the shovel; results from the soil analysis led searchers directly to a ditch close to the river bottom where a volunteer found a depression near a large tree and where Haygood's body was uncovered.

Despite his mental state, Pate eventually may have been able to lead authorities to Mary's body. With the soil analysis, however, authorities were able to find the body in time to perform an autopsy and corroborate his confession. Pate was later convicted in Oklahoma State Court.

As demonstrated by these examples, near-IR spectrometers have become sufficiently rugged, portable, and inexpensive to be used in field investigations. The newest notebook-sized personal computers with 80386 and 80486 CPUs provide an abundance of computing power to perform sophisticated spectral identification tasks. More widespread application of near-IR spectrometry will likely follow.

References

(1) Monastersky, R. *Science News* **1990**, *138*, 392–94.
(2) Murray, R. C. *Forensic Geology: Earth Sciences and Criminal Investigation*; Rutgers University Press: New Brunswick, NJ, 1975.
(3) Lodder, R. A.; Hieftje, G. M. *Appl. Spectrosc.* **1988**, *42*, 1500–12.
(4) No. A-12901, Court of Criminal Appeals of Oklahoma, 361 P.2d. 1086, April 19, 1961.

Robert A. Lodder is an assistant professor of pharmaceutical and analytical chemistry. He received a Ph.D. in analytical chemistry (1988) from Indiana University. He is the recipient of several awards, including the 1992 NYI award from the National Science Foundation, first prize in the 1990 IBM supercomputing contest, and the 1988 Tomas Hirschfeld Award.

Angela R. Jones is working toward a B.A. degree in journalism at the University of Kentucky. A technical writer for Lodder, she is the only four-year recipient of the national Freedom Forum Journalism Scholarship. She has interned as a general assignment reporter for several daily newspapers.

Yi Zou is completing his Ph.D. in analytical chemistry at the University of Kentucky. He received his B.S. and M.S. degrees from Xiamen University in China. His research includes the use of pattern recognition and multivariate data analysis with near-IR spectroscopy.

Yu Xia was a visiting scholar in Lodder's laboratory at the University of Kentucky, and is now a lecturer in the pharmacy department of Shandong Medical University in China. Her primary interests are the use of HPLC and UV spectrometry for pharmaceutical analysis.

Reprinted from Analytical Chemistry, 1993, 65.