Chemical Profiling To Differentiate Geographic Growing Origins of Coffee

KIM A. ANDERSON*[,†,‡] AND BRIAN W. SMITH[§]

Department of Environmental and Molecular Toxicology, Food Safety and Environmental Stewardship Program, and Department of Mathematics, Oregon State University, Corvallis, Oregon 97331

The objective of this research was to demonstrate the feasibility of this method to differentiate the geographical growing regions of coffee beans. Elemental analysis (K, Mg, Ca, Na, Al, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Mo, S, Cd, Pb, and P) of coffee bean samples was performed using ICPAES. There were 160 coffee samples analyzed from the three major coffee-growing regions: Indonesia, East Africa, and Central/South America. A computational evaluation of the data sets was carried out using statistical pattern recognition methods including principal component analysis, discriminant function analysis, and neural network modeling. This paper reports the development of a method combining elemental analysis and classification techniques that may be widely applied to the determination of the geographical origin of foods.

The determination of geographic origin of commodities and food products is becoming an increasingly active research area. The increasing demands on the agrifood industry from free-trade, globalization, and changing technology only further the drive to determine the authenticity of foods (*1*). This includes both geographic authenticity and adulteration of foods. Financial incentives continue to drive retailers/resellers to misidentify the geographic origins of commodities (*2*) and food products (*3*–*7*). The determination of geographic origin through chemical analysis coupled with sophisticated data classifying techniques is timely. Although recently publications in this area have begun to develop, geographic classification has focused on processed foods, most especially wines (*3*) and juices and to a much lesser extent drugs of abuse (*5*), cocoa (*6*, *7*), and olive oil. The objective of this study was to determine the feasibility of using a multielement analysis, by ICPAES, a single-analytical instrument, combined with statistical modeling methods (*2*) for differentiating (with the goal of ultimately determining) the geographic origin for coffee beans. Here we present data from eight different sites on three continents, representing a total 160 × 18 data set; nearly 3000 chemical parameters were measured.

Over two-thirds of all the research literature on geographic origin of commodities involves the analysis of vitamins or other organic molecules (amino acids, triglycerides, volatile aromatic compounds, etc.). Some success (60−90% correct classification) has been reported using vitamin and/or amino acid assays to determine geographic origin (*8*–*13*). However, a shortcoming of using vitamins (or other organic compounds) is their susceptibility to degradation (including enzymatic changes) from the time of harvest through storage to the time of analysis. Storage conditions may be especially important for some vitamin assays; for example, vitamin E is light sensitive, and changes in vitamin E content during storage have been reported (*14*). It is important, therefore, if one wants to develop a technique that will ultimately be used to determine the geographic origin of unknown samples, that effects from storage conditions be minimized. This is also important because coffee beans are processed and organic chemical profiles are likely to be especially susceptible. Therefore, a method that is robust and independent of variations from storage conditions is most desirable. The use of minerals and trace elements is therefore powerful because trace elements are significantly more stable in the commodity versus vitamins or some other types of organic compounds.

It is recognized that mineral and trace metal compositions of fruits and vegetables are a distorted reflection of the trace mineral composition of the soil and environment in which the plant grows (*15*). The soil−plant system is highly specific for different elements, plant species, and environmental conditions. Under most conditions, a trace element present in the vegetable/ fruit must have existed in the rooting zone of the plant, at least in a slightly soluble form. A trace element must also pass through at least one cellular membrane in its movement from soil to plant. The selectivity of these processes of mineral bioaccumulation within the vegetable varies with different trace elements, with different plants, and with the unique environment in which the commodity is grown.

The determination of geographic origin of wines has been an active area of research for some years (*8*, *16*–*19*). Most of

* Author to whom correspondence should be addressed [fax (541) 737-0497; e-mail kim.anderson@orst.edu].
† Department of Environmental and Molecular Toxicology.
‡ Food Safety and Environmental Stewardship Program.
§ Department of Mathematics.

Chemical Profiling To Differentiate Coffee Origins

J. Agric. Food Chem., Vol. 50, No. 7, 2002 **2069**

these studies involve the chemical determination of organic molecules, most commonly some combination of amino acids. Accuracy rates of prediction fall in the 73−91% range (*8, 16−19*). However, an excellent study by Day et al. (*20*) combined the analyses of ^{2}H NMR with multiple elemental and isotopic ratio determinations; here the technique classified wine samples with ≥99% accuracy. This approach, however, requires the use of several instruments including SNIF-NMR, elemental analyzer IRMS (isotope ratio mass spectrometry), FAAS (flame atomic absorption spectrometry), ETAAS (electrothermal atomic absorption spectrometry), and ICPAES (inductively coupled plasma atomic emission spectrometry). In addition, sophisticated techniques are necessary for the determination of the five isotopic ratios used. Overall, many authenticity studies are survey in nature (<30 samples), and therefore general conclusions concerning the effectiveness of these techniques should be prudent.

The purpose of this study is to determine the feasibility to differentiate coffees grown in the three different major regions of the world, Central and South America, East Africa, and Indonesia. In addition, the feasibility of further separation of these regions was investigated. The "coffee belt" is a globe-encircling geographic region between latitudes 30° N and 30° S, where coffee trees grow from sea level to 2000 m. Within the coffee-belt there are large variances in coffee quality. In-cup taste testers are used to recognize country of origin; however, testers cannot identify reliably a large number of coffee origins. Most importantly, the opinions of taste testers are subjective and often not always unanimous. In cases of arbitration (e.g., coffee producers, coffee traders, and coffee consumers) disagreements occur frequently. Import/export, legal implications, and financial concerns make determining country of origin for coffee important, with over $50 billion in U.S. dollars in coffee retail sales. Coffee is currently exported by more than 50 countries and on the international trade market ranks second only to petroleum, providing livelihood for over 100 million people worldwide (*21, 22*). Consequently, coffee producers and traders are motivated to discover more objective (chemical) techniques for determining the geographic origin of coffee.

For instance, there is a consistent trend of price differential for coffee based on quality, taste, and body (*23*). The price differential is dramatic; some of the most expensive coffees sell for $35/pound (U.S. dollars), whereas some of the least expensive coffees sell for $1/pound. This has led some people to pass off cheaper coffee, or a mix of coffees, as pure expensive types. Recently, Hawaiian coffee growers settled a lawsuit alleging distributors sold cheaper South American beans as genuine Kona coffee from Hawaii (*24*). For example, it is estimated that 20 million pounds of Kona coffee is sold annually, whereas only 2 million pounds are produced (*24*). Protecting market share, reputation, and consumer confidence to pay a premium for specific growing regions of coffee is meaningful to the industry. The misidentification of coffee by unscrupulous resellers/retailers can affect future consumer choices in both the short and long term as well as deceive the consumer. Further lasting effects include jeopardizing consumer confidence in the quality of coffee (*25*), if coffees have been unknowingly switched with lower quality beans, and affecting the consumers' willingness to pay premium prices for coffee. Therefore, developing a method that can identify the origin of coffee is important to protect the coffee industry.

Surprisingly, there are very few studies reported for the determination of origin of coffee (*26*). Much of the existing literature on the chemical composition of coffee (green beans, roasted beans, or prepared) has focused on food quality aspects or adulteration with noncoffee products (*26*). Food quality chemical analyses are subsequently dominated by volatile headspace techniques for organic compounds. This type of analysis is inappropriate for geographic authenticity as these techniques are subjective to the storage conditions and brew techniques, and the analysis technique itself influences the "aroma" profile of volatile headspace analyses (*26*). In addition, these types of techniques may tend to be too subjective or too laboratory specific to provide the industry with the necessary definitive, robust, reproducible method needed for cases in dispute. A thermogravimetric chemical ionization mass spectrometry technique was not successful in differentiating 13 green coffees (*27*). A study using a suite of metals (and four instruments) on 10 coffee samples found some differences, but the study was very sample limited so conclusions should be prudent (*26, 28*). Also, one of the four methods used was neutron activation analysis, which requires an instrument not routinely available to a wide research audience. We report on the feasibility of using an elemental analysis method to determine the geographic growing origin of coffee from eight different sites, including all three major coffee-growing regions in the world. The method presented requires only one instrument, is flexible, and demonstrates the feasibility to differentiate coffee beans from different geographic origins with samples that are likely to be in dispute. The method presented here is not susceptible to storage conditions (*2*), requires only small samples, and uses common automated equipment.

MATERIALS AND METHODS

Reagents. The sources of chemicals and reference materials were as follows: concentrated, nitric acid trace metal analysis grade (J. T. Baker, St. Louis, MO); elemental stock standard solutions (J. T. Baker); reference materials, NIST 1575 pine needles, NIST oyster tissue 1566a, NIST rice flour 1568a, NIST 1577b bovine liver, NIST 8433 corn bran (National Institute of Standards and Technology, Gaithersburg, MD); NRC TORT-2 lobster hepatopancreas (National Research Council Canada, Institute National Measurements Standards, Ottawa, ON, Canada).

Apparatus. The inductively coupled argon plasma atomic emission spectrometer (ICPAES) was equipped and set up as follows: model, Leeman 1000 ICPAES; power, 1.1 kW; coolant, 16 L/min (Lpm); nebulizer, 41 psi; auxiliary flow, 0.20 L/min; pump rate, 1.0 mL/min; scan intergration time, 0.25 s; Mn1 peaking wavelength; acid flexible tubing, 0.030 i.d. mm, wavelengths and background corrections are given in **Table 1** (*29, 30*). Temperature controller/digester used was a digestion system 40, 1016 Digester, and Autostep 1012 controller (Tecator). It was fitted with an aluminum adapter plate 3 mm thick with 40 17-mm holes on top overlaid on the heater block. Measurements were made at the wavelengths listed below for the macro- and microelements; upper and lower background corrections are used as listed in **Table 1**.

Sampling, Preparation, and Analysis. Samples were purchased at a local coffee outlet. All samples were roasted coffee beans (as purchased by the coffee outlet). Four samples from Central and South America were purchased and identified as being from Costa Rica, Colombia, Panama, and Guatemala. Two samples from Indonesia were identified as being from Sulawesi and Sumatra. Two samples from East Africa were identified as being from Kenya and Ethiopia.

Each coffee was analyzed as the roasted bean. No further preparation of the sample was required (i.e., no further drying). A ∼1.0 g sample was taken, representing four to six beans, and the sample was digested with 3.0 mL of nitric acid (trace metal grade) in a 10 mL graduated Kimax culture tube on a programmed heating block. The samples were allowed to react for ∼4−8 h in a hood at ambient temperature. Then the samples were digested using a heating block (programmable digester

2070 *J. Agric. Food Chem.,* Vol. 50, No. 7, 2002

Anderson and Smith

Table 1. Wavelengths and Background Corrections for Macro- and Microelements in Coffee Samples, Determined by ICPAES, and Method Detection Limits

element	emission λ (nm)	upper and lower background correction from emission λ	method detection limit[a] (μg/g)
aluminum	308.215	none used	0.3
cadmium	214.438	0.047/0.036	0.04
calcium	317.933	0.053/none	0.17
chromium	206.149	0.026/0.049	0.12
cobalt	228.616	0.033/0.038	0.12
copper	324.754	0.041/0/038	0.07
iron	259.940	0.074/0.056	0.06
lead	220.353	0.034/0.025	0.5
magnesium	279.553	0.032/0.035	0.04
manganese	257.610	0.052/0/042	0.02
molybdenum	202.030	0.037/0.037	0.34
nickel	231.604	0.029/0.029	0.13
phosphorus	214.910	0.034/none used	0.5
potassium	766.490	none used	10
sodium	589.592	none used	1
vanadium	310.230	0.025/0.034	0.1
zinc	213.856	0.060/0.032	0.05

[a] Detection limits are from our typical plant analysis screen; subsequent optimization for dry samples such as coffee beans would decrease detection limits by a minimum of a factor of 4–6.

may be used). The samples were heated to 180 °C for 3–4 h. Digestion is confirmed to be complete when no nitrous oxide gases are evolved (i.e., orange gas production). Samples are diluted with type 1 water (18 Mohm·cm) and mixed thoroughly using a vortexer. Analysis is by ICPAES.

Quality Control. Each analytical batch contained a minimum of 25% quality control samples, including check standards, duplicates, spikes, and standard reference materials (SRM). Calibration curves consisted of three to four standards each, with correlation coefficients of >0.98. During the course of the study >50 SRM samples were analyzed; SRM were dominantly plant matrices when available. Typical percent standard deviation (% SD) was <10%, although analytes close to method detection limits (MDL) had higher % SD. Spike recoveries and check standards were typically within ±10% of their true value.

COMPUTATIONAL ANALYSIS

The data were analyzed to explore the feasibility of classifying coffee samples according to geographic origin. Multivariate pattern recognition methods were used to analyze the data to determine if the geographic origin of a sample could be characterized by the proportions of its trace metal constituents. The methods employed include principal component analysis, canonical discriminant function analysis, quadratic discriminate function analysis, linear discriminate function analysis, and neural network modeling. The computational analysis was similar to that done by the authors to determine the feasibility of geographic classification of potatoes (2). Principal component analysis (PCA) and canonical discriminant analysis (CDA) are variable reduction methods and were used for exploratory data visualization to determine to what extent we could discern differentiation of the samples according to geographic origin by looking at appropriate one- and two-dimensional displays of the data. The other methods are model classifier methods: known samples are used to build or "train" a classification model. The trained model will then make a prediction about the geographic origin of an unknown sample based on the proportions of trace metals found in the sample.

Visualization. *(a) Principal Component Analysis.* PCA generates principal components that are linear combinations of the original variables. The first principal component (PC)

summarizes the maximum possible variation that can be projected onto one dimension, the second PC captures the second most, and so on. The principal components are orthogonal in the original space of variables, and the number of principal components can equal the number of the original variables. However, it is sometimes the case that a large percentage of the total variation can be explained by the first few principal components, effectively reducing the number of variables needed to describe variation among individual samples. In this case, plotting the samples with respect to one or two principal components facilitates one- or two-dimensional views of how individual samples differ from one another (in the variation sense). For a geographic classification task it is desirable to have group differences explicitly manifest with a low-dimensional view. However, this is not always the case because this method measures variation in the elemental concentrations in the samples but does not take into account group (geographic origin) membership. To get the best possible view of group clustering, we use CDA.

(b) Canonical Discriminant Analysis. CDA generates canonical variables, which are linear combinations of the original variables, that describe the variation between prespecified classes in a manner analogous to the way in which PCA summarizes the variation among individual samples. CDA can effectively reduce the number of variables and provide optimum low-dimensional "views" of the data, which display the maximum possible variation *among different groups* and the minimum possible variation *within the same group.* CDA has been applied to data for the purpose of geographical classification of potatoes (2) and wine (20).

Classification Models. *(a) Discriminant Function Analysis.* Discriminant function analysis here refers to a group of pattern recognition classification methods that use known data to determine a discriminant function, which can then be used to classify unknown samples into predetermined classes. Two types of discriminant functions were used for this study: a linear discriminant function and a quadratic discriminant function. Details of how each of these work can be found in the description of the DISCRIM procedure in the SAS technical manual (31).

(b) Neural Networks. Feed-forward back-propagation neural network methods were also applied to the data in an effort to classify samples according to geographic origin. A neural network is "trained" using known samples by adjusting internal parameters called weights so that an error measure of actual versus predicted results is minimized. The trained model is then used to classify unknown samples. Neural network models are inherently very flexible and are able to model complex boundaries between groups.

Usually, some measures must be taken to prevent the tendency of neural networks to "overfit" the known training data. An overfitted model will predict the training data well but may perform poorly when new samples not used in the training set are classified. One method is to use early stopping. The known data is split into training and test sets. As the network is trained using the training set it is periodically presented with the task of classifying the test set. The network is saved when the test set error is minimized (as opposed to when the training set error is minimized). The idea is to ensure that the neural network model will perform well when it is used to classify new samples that it has never seen. This property is called generalization. Further generalization improvements can usually be achieved by employing a bootstrap aggregation ("bagging") strategy (32). Multiple networks are trained using randomly selected (sampling

Chemical Profiling To Differentiate Coffee Origins

J. Agric. Food Chem., Vol. 50, No. 7, 2002 **2071**

Table 2. Mean Concentrations and Standard Deviations of Dry Weight for 11 of the 17 Elements Determined in Roasted Coffee Bean[a]

growing region		statistics, μg/g (ppm)										
		P	Zn	Mn	Fe	Mg	Al	Cu	Ca	K	S	Na
Costa Rica	av	1920	7.97	23	15	2203	13	18.1	1079	18570	1640	41.5
($n = 20$)	SD	140	1.4	4.0	3	100	2	2.6	150	1190	55	7.0
Colombia	av	1980	8.01	38	17	2268	19	17.2	1129	19170	1480	40.1
($n = 20$)	SD	200	1.8	10	3	140	4	1.7	150	1380	130	11
Guatemala	av	1960	8.03	25	13	2410	8	14.0	1234	19010	1640	28.9
($n = 20$)	SD	180	1.4	5	2	140	1	2.3	290	8840	74	5.9
Panama	av	1740	7.04	26	20	2174	3	16.8	997	18680	1500	9.6
($n = 20$)	SD	90	1.8	4	5	110	2	1.8	170	1260	92	2.3
Ethiopia	av	1860	7.82	21	12	2058	7	13.8	1013	19280	1450	20.1
($n = 20$)	SD	150	1.9	4	3	130	2	0.9	140	1380	83	4.7
Kenya	av	1710	7.15	39	15	2150	4.4	17.8	976	17500	1420	37.3
($n = 20$)	SD	140	1.9	11	5	160	3	2.2	170	1030	97	10.0
Sulawesi	av	2110	7.87	29	21	2347	13	12.5	934	19160	1431	1467
($n = 20$)	SD	160	1.7	7	6	160	6.8	3.4	120	1090	124	5.3
Sumatra	av	1940	6.51	19	31	2098	36	13.2	1141	19600	1490	10.60
($n = 20$)	SD	210	1.1	5	7	110	13	2.0	190	1400	85	9.7

[a] Each mean represented 20 individual samples, a total of 160 samples. Data are not shown for molybdenum, cadmium, lead, cobalt, nickel, and chromium; these elements had means that were either below method detection limits or near method detection limits that resulted in large standard deviations (>30%).
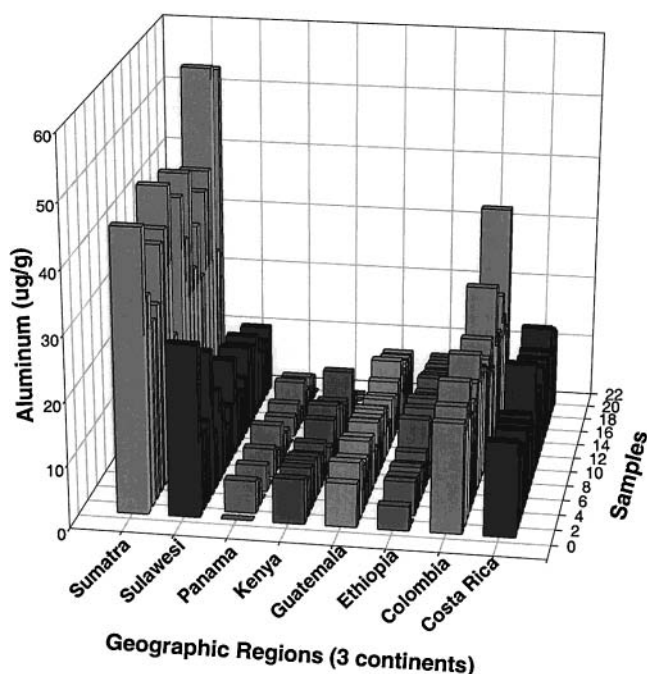

Figure 1. Concentration of aluminum in roasted coffee beans versus geographic growing origin.

Figure 2. Concentration of iron in roasted coffee beans versus geographic growing origin.

with replacement) training sets, and final classification is obtained by voting. This has the effect of reducing the high variance inherent in neural networks, resulting in improved generalization. Typically, if safeguarded against overfitting, neural network models perform as well as or better than other classification methods.

RESULTS AND DISCUSSION

Chemical Analysis. An important attribute of this approach is that all of the chemical data can be determined with the use of a single analytical instrument, ICPAES. Whereas other geographic authenticity approaches require the use of several instruments, this technique requires only a single, commonly available automated instrument. In this study 18 elements were determined, additional elements could be added, and unlike chromatography techniques, spectroscopy data analysis requires little analyst time or special expertise. In this approach, the data

are used directly from the ICPAES into the computational models requiring no prior mathematical or interpretive data analyses, as is often the case with other geographic authenticity approaches.

The analytical technique is well suited to analysis of small samples. A minimum of 500 mg can be used; 1 g was used in this study. Dilution factors are minimized here—only a factor of 10 as compared to typical digestions that involve dilution factors of ≥50. This small dilution factor permits determination of additional elements that would otherwise be below instrument detection limits. In addition, as a pollution prevention mechanism, this technique uses fewer reagents and thus reduces waste.

Of the 18 elements tested, 11 were routinely above the detection limit (see **Table 2**). Phosphorus, zinc, magnesium, calcium, potassium, and sulfur had concentrations within 10% of the overall mean. Therefore, individually none of these elements alone appears to have discriminating power with the geographic regions tested. Copper, sodium, manganese, and iron have some discriminating power with the geographic regions

Figure 3. Chemical profile of three elements (Al, Na, and Mn).

tested, but one cannot determine origin with these elements alone. However, more sophisticated computational analysis indicates these data have value increasing modeling success (see below). Aluminum ranged from 3 to 36 μg/g, a factor of 12 difference between geographic regions (see **Figure 1**). Iron ranged from 12 to 31 μg/g, a factor of 2.5 difference between geographic regions (see **Figure 2**). Manganese for all tested

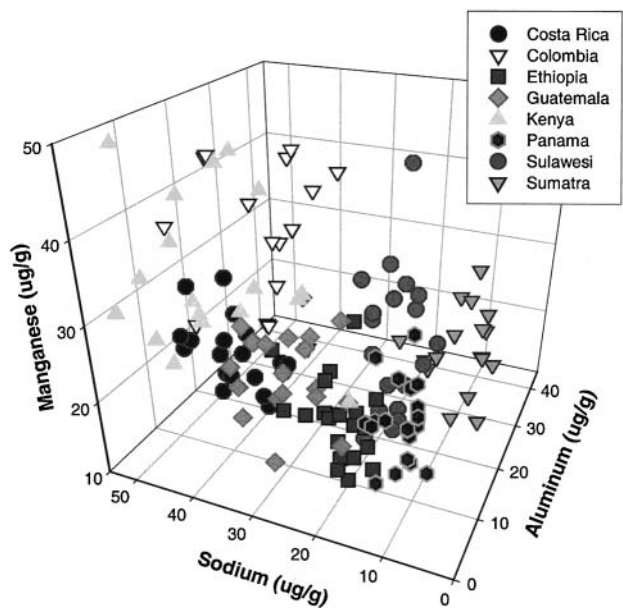samples ranged from 19 to 39 μg/g, a factor of 2 difference between geographic regions (see **Table 2**). Copper for all tested samples ranged from 12 to 18 μg/g, which is a factor of 1.5 difference between geographic regions. Sodium for all tested samples ranged from 10 to 41 μg/g, a factor of 4 difference between geographic regions. Although individually no element is diagnostic of origin, by combining elements there is better discrimination. For example, from a three-dimensional plot of aluminum, sodium, and manganese, one can see that origins are beginning to separate (see **Figure 3**). With more dimensions and modeling (see below), better separations are possible.

The ranges of all elements tested were in the same range as for green coffee beans, indicating that the roasting process probably has little effect on element concentrations (*26, 28*) unlike organic compounds.

Aluminum, cadmium, phosphorus, and sulfur have not been previously determined for origin determination in coffee beans. Some elements determined here, such as Fe, did not show as much discriminating power in green coffee beans as discovered here (*28*). However, the error limits in the green coffee bean study were typically a bit larger (e.g., 25% for Fe), such that subtle differences may not have been revealed. In this same study manganese was the most discriminating element, with a difference of a factor of 2.5 seen between regions.

Another important result of the element concentration distribution is that no one region is responsible for all of the high or low concentrations. For example, Costa Rica has the highest average copper and sodium concentrations and the highest sulfur concentration. Colombia has the highest zinc, and Guatemala has the highest calcium and sulfur concentrations. Panama has
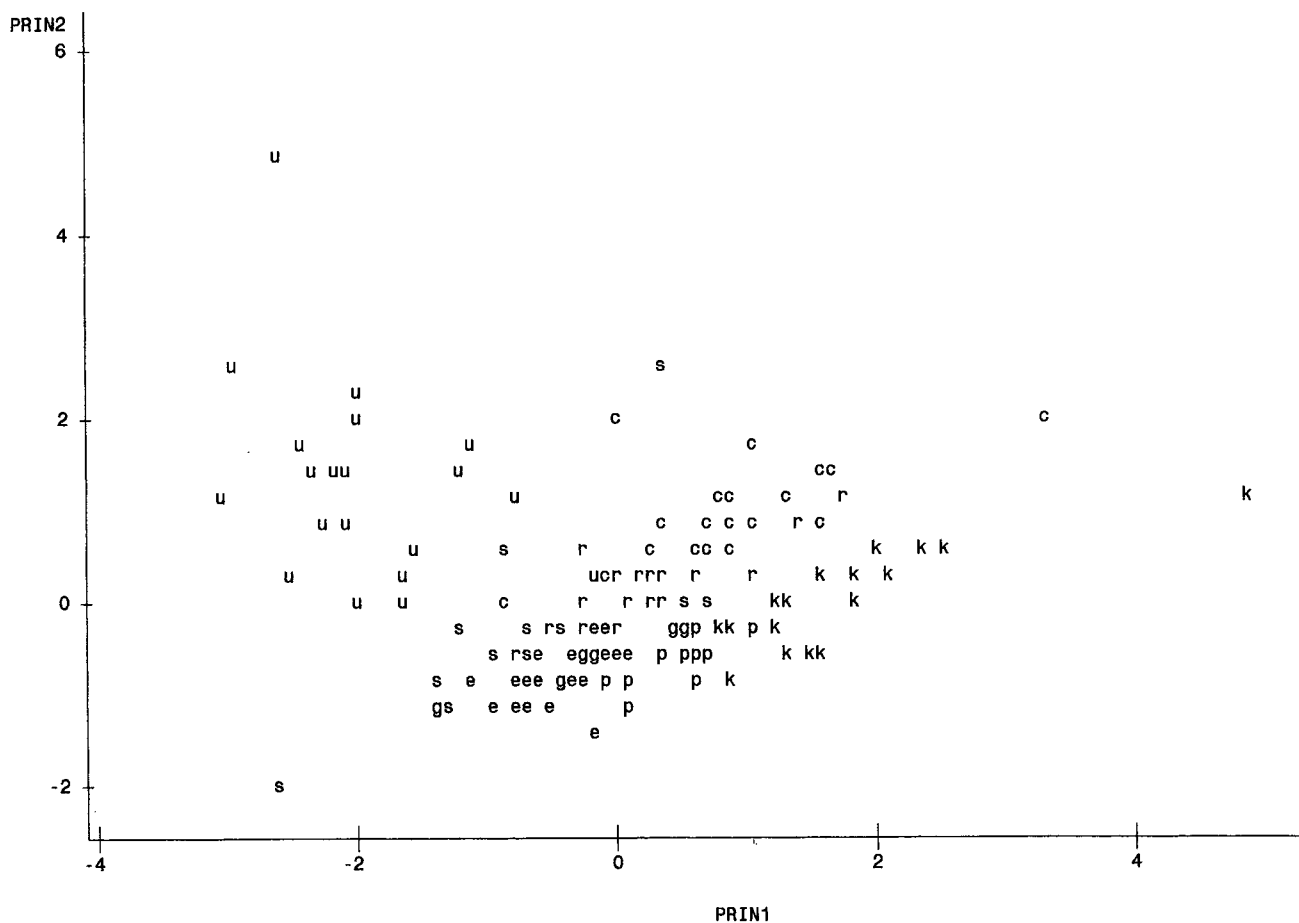

Figure 4. Principal component one versus principal component two, for chemical profile of elements in roasted coffee beans from eight different growing regions: r, Costa Rica; c, Colombia; p, Panama; s, Sulawesi; u, Sumatra; k, Kenya; e, Ethiopia. There are 41 observations that are hidden.
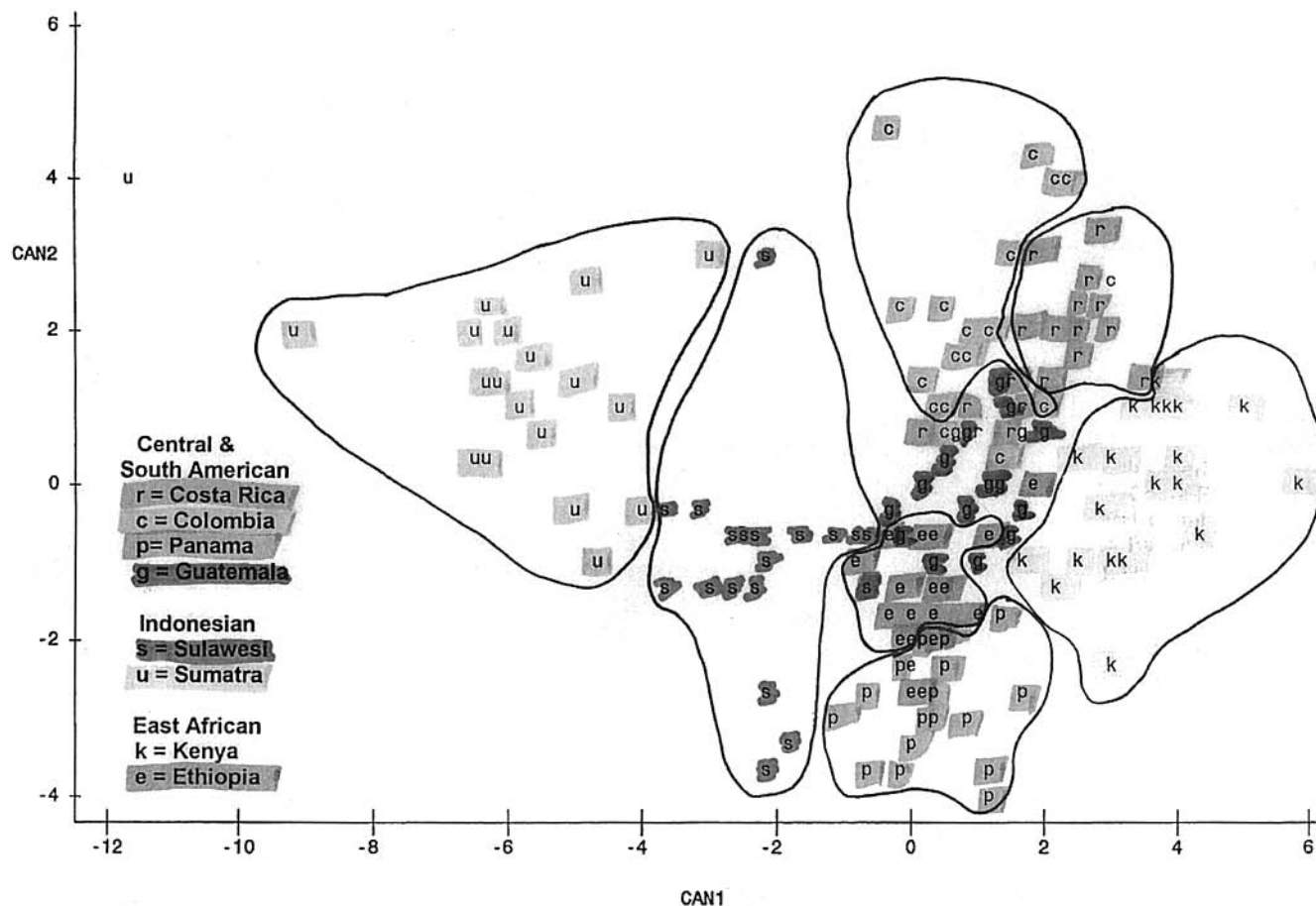
Chemical Profiling To Differentiate Coffee Origins

J. Agric. Food Chem., Vol. 50, No. 7, 2002   **2073**

Figure 5. Plot of the first canonical functions (country codes are given in the figure). This simplified visual representation demonstrates the separation of regions; any overlap viewed in this one-dimensional representation is not indicative of any intractable classification task. All 16 available dimensions are utilized (see text).

the lowest average aluminum concentrations of the regions tested. Ethiopia has the lowest iron and magnesium. Kenya has the highest manganese and lowest potassium and sulfur concentrations. Sulawesi has the highest magnesium and the lowest average copper and calcium concentrations. Finally, Sumatra, of all the regions tested, is probably the most different, with the lowest zinc, manganese, potassium, and sodium and the highest iron and aluminum. Overall, with so many differences, on the whole, the use of computational modeling with elemental concentrations is powerful.

Seasonal variability was not investigated here. In another study, with green coffee beans, although data were limited, over three growing seasons, little variation was seen (*28*). We have analyzed >2000 potatoes over several seasons and have also seen little variation (*2*), although generation of databases for each season is likely (and necessary) to ensure good predictability of any model used routinely.

Computational Analysis. *Introduction.* A total of 160 coffee samples representing eight geographic regions with 20 samples from each were analyzed. Each sample was measured for the proportions of 16 (Pb and Mo not used) trace elements present. For initial data exploration PCA was applied to the data. To adjust for different scales of measurement between trace elements, the data were normalized by subtracting the elemental means from each entry, and then each resulting difference was divided by the corresponding standard deviation. Thus, each trace element had an adjusted mean of zero and an adjusted standard deviation of one.

Visualization. (a) Principal Component Analysis. PCA has been applied to geographical classification applications of various foods including processed orange juice (*33*), wine (*18, 20*), honey (*34*), and cocoa (*6, 7*). PCA demonstrates that a small

number of variables were not dominating total variability. Only modest visual clustering was apparent when the data were displayed with respect to the first two principal components (see **Figure 4**). Similar results were seen from principal component one versus three or principal component two versus three (data not shown). This was not surprising because the first principal component accounts for the maximum possible one-dimensional projection of total variation of the individual points, which does not necessarily correspond to the maximum variation between classes. Better visual results were obtained with CDA (discussed below).

(b) Canonical Discriminant Analysis. CDA was applied to our data using eight defined classes. The data were plotted using the first canonical variable alone, as shown in **Figure 5**. **Figure 6** shows samples plotted with respect to the first and second canonical variables. Both figures display reasonably good visual separation of the samples based on geographic origin.

Classification Models. From each of the eight geographic regions, 4 samples (from 20) were randomly selected to form a test set of 32 samples. The remaining 128 samples were used as the training set for the classification models. Once trained, each model was then used to classify the 32 "unknown" samples in the test set. Also, each model was presented with the training set for classification. Classification performances for each classification model are discussed below.

(a) Discriminant Function Analyses. The linear discriminant function model classified 23 of 32 samples (~72%) in the test set correctly and correctly classified 90 of 128 samples (~71%) in the training set. The quadratic discrimination function model classified 23 of 32 (~72%) samples in the test set correctly and correctly classified 102 of 128 samples (~80%) in the training set.

Figure 6. Plot of scores of first and second canonical functions for 160 coffee samples (country codes are given in the figure). For clarity the different continent areas of origin have been outlined by hand.

(b) Neural Network Analysis. Fifty neural network models were trained using the training set. Each of these models was safeguarded to some extent using an early stopping strategy. That is, to construct each of the 50 models, the training set of 128 samples was randomly split into two parts: 77 samples in a "local training" set and 51 samples in a validation set. (Each model had its own random partitioning of the training set.) The weight parameters that determine a given model were saved at a point where an error measure on the validation set was minimized. Once trained, the 50 models were aggregated to form a composite model. The composite model prediction for a given sample is determined by voting with the 50 individual models. The composite model classified 26 of 32 samples (~81%) in the training set correctly and correctly classified 110 of the 128 samples (~86%) in the training set.

Within the framework of this study it appears that the geographic origin of coffees may be determined by their chemical profile. Statistical analysis revealed groupings between the three major geographic regions of coffee production in the world, Indonesia, East Africa, and Central/South American. Although differences in elemental concentrations were determined, simple inspection of elemental concentrations cannot be used to differentiate growing origin. Use of neural network models and discriminate function analysis both successfully differentiated coffees relative to their subregional growing origin (70−86% successful classification). Work is in progress to further substantiate the effects of seasonal and coffee variety influences. As well, further classification breakdown of subregions is currently underway in our laboratory.

LITERATURE CITED

(1) U.S. Department of Commerce, Bureau of Economic Analysis, Regional Division, 1992.
(2) Anderson, K. A.; Magnuson, B. A.; Tschirgi, M. L.; Smith, B. Determining the geographic origin of potatoes with trace metal analysis using statistical and neural network classifiers. *J. Agric. Food Chem.* **1999**, *47*, 1568−1575.
(3) Danzer, K.; Carcia, D. D. C.; Thiel, G.; Reichembacher, M. Classification of wine samples according to origin and grape varieties on the basis of inorganic and organic trace analyses. *Am. Lab.* **1999**, *31* (20), 29−34.
(4) Desage, M.; Guilluy, R.; Brazier, J.; Chaudron, H.; Girard, J.; Cherpin, H.; Jumeau, J. Gas chromatography with mass spectrometry or isotope-ratio mass spectrometry in studying the geographical origin of heroin. *Anal. Chem. Acta* **1991**, *247*, 249−254.
(5) Flurur, C.; Wolnik, K. Chemical profiling of pharmaceuticals by capillary electrophoresis in the determination of drug origin. *J. Chromatogr. A* **1994**, *674*, 153−163.
(6) Hernández, C.; Rutledge, D. Characterization of coca masses: low resolution pulse NMR study of the effect of geographical origin and roasting on fluidification. *Food Chem.* **1994**, *49*, 83−93.

Chemical Profiling To Differentiate Coffee Origins

J. Agric. Food Chem., Vol. 50, No. 7, 2002 **2075**

(7) Hernández, C. V.; Rutledge, D. N. Multivariate statistical analysis of gas chromatograms to differentiate cocoa masses by geographical origin and roasting conditions. *Analyst* **1994**, *119*, 1171−1176.

(8) Aires-De-Sousa, J. Verifying wine origin: a neural network approach. *Am. J. Enol. Vitic.* **1996**, *47*, 410−414.

(9) Ferland, G.; Sadowski, K. Vitamin K$_1$ (phylloquinone) content of green vegetables: effects of plant maturation and geographical growth location. *J. Agric. Food Chem.* **1992**, *40*, 1874−1877.

(10) Hulshof, P.; Xu, C.; van de Bovenkamp, P.; Muhilal; West, C. Application of a validated method for determination of provitamin A carotenoids in Indonesian foods of different maturity and origin. *J. Agric. Food Chem.* **1997**, *45*, 1174−1179.

(11) Parcerisa, J.; Boatella, J.; Codony, R.; Farràn, A.; Garcia, J.; López, A.; Rafecas, M.; Romero, A. Influence of variety and geographical origin on the lipid fraction of hazelnuts (*Corylus avellana* L.) from Spain: I. Fatty acid composition. *Food Chem.* **1993**, *48*, 411−414.

(12) Parcerisa, J.; Rafecas, M.; Castellote, A.; Condony, R.; Farràn, A.; Garcia, J.; López, A.; Romero, A.; Boatella, J. Influence of variety and geographical origin on the lipid fraction of hazelnuts (*Corylus avellana* L.) from Spain: II. Triglyceride composition. *Food Chem.* **1994**, *50*, 245−249.

(13) Parcerisa, J.; Rafecas, M.; Castellote, A.; Condony, R.; Fàrran, A.; Garcia, J.; Gonzalez, C.; López, A.; Romero, A.; Boatella, J. Influence of variety and geographical origin on the lipid fraction of hazelnuts (*Corylus avellana* L.) from Spain: III. Oil stability, tocopherol content and some mineral contents (Mn, Fe, Cu). *Food Chem.* **1995**, *53*, 71−74.

(14) Lavedrine, F.; Ravel, A.; Poupard, A.; Alary, J. Effect of geographic origin, variety and storage on tocopherol concentrations in walnuts by HPLC. *Food Chem.* **1997**, *58*, 135−140.

(15) Esechie, H. Distribution of chemical constituents in the plant parts of six tropical-origin forage grasses at early anthesis. *J. Sci. Food Agric.* **1992**, *58*, 435−438.

(16) Armanino, C.; Fornia, M.; Castino, M.; Piracci, A.; Ubigli, M. Chemical investigation of four red wines from a single cultivar grown in the Piedmont region. *Analyst* **1990**, *115*, 907−910.

(17) Etievant, P.; Schlich, P.; Bouvier, J. C.; Symonds, P.; Betrand, A. Varietal and geographic classification of French red wines in terms of elements, amino acids and aromatic alcohols. *J. Sci. Food Agric.* **1988**, *45*, 25−51.

(18) Latorre, M. J.; García-Jares, C.; Mèdina, B.; Herrero, C. Pattern recognition analysis applied to classification of wines from Galicia (Northwest Spain) with certified brand of origin. *J. Agric. Food Chem.* **1994**, *42*, 1451−1455.

(19) Vanderschee, H. A.; Bouwknegt, J. P.; Tas, A. C.; Maarse, H.; Sarneel, M. M. The authentication of sherry wines using pattern-recognition—an interlaboratory study. *Z. Lebensm. Unters. Forsch.* **1989**, *188*, 324−329.

(20) Day, M.; Zhang, B.; Martin, G. Determination of the geographical origin of wine using joint analysis or elemental and isotopic composition. II. Differentiation of the principal production zones in France for the 1990 vintage. *J. Sci. Food Agric.* **1995**, *67*, 113−123.

(21) International Coffee Organization. *The Coffee Story*; ICO Information Centre: London, U.K., 1989.

(22) International Coffee Organization, 2001; http://www.imf.org/external/np/sec/decdo/ico.htm.

(23) Bicchi, C. P.; Binello, A. E.; Legovich, M. M.; Pellegrino, G. M.; Vanni, A. C. Characterization of roasted coffee by S-HSGC and HPLC-Uv and Principal Component Analysis. *J. Agric. Food Chem.* **1993**, *41*, 2324−2328.

(24) Dworkin, A. Coffee growers settle suit about Kona name. *The Oregonian* **1999** (Oct 1).

(25) Al-Saikhan, J.; Howard, L. R.; Miller, J., J. C. Antioxidant activity and total phenolics in different genotypes of coffee. *J. Food Sci.* **1995**, *60*, 341−346.

(26) Prodolliet, J. Authenticity of Coffee. In *Food Authentication*; Ashurst, P. R., Dennis, M. J., Eds.; Blackie Academic and Professional: London, U.K., 1996.

(27) Dyszel, S. M. Characterization of green coffee beans by combined thermogravimetric analysis atmospheric press chemical ionization mass spectrometry. *Thermochim. Acta* **1985**, *87*, 89−98.

(28) Krivan, V.; Barth, P.; Morales, A. F. Multielement analysis of green coffee and its possible use for the determination of origin. *Mikrochim. Acta* **1993**, *110*, 217−236.

(29) Anderson, K. A. Micro-digestion and ICPAES analysis for the determination of macro and micro elements in plant tissues. *At. Spectrosc.* **1996**, *1/2*, 30−33.

(30) Anderson, K. A. *Analytical Techniques for Inorganic Contaminants*; AOAC International: Gaithersburg, MD, 1999.

(31) SAS Systems for Windows, release 6.11; SAS Institute, Inc., Cary, NC.

(32) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123−140.

(33) Nikdel, S.; Nagy, S.; Attaway, J. Trace metals: Defining geographical origin and detecting adulteration of orange juice. *Sci. Toxicol.* **1988**, *596*, 81−105.

(34) Sanz, S.; Perez, C.; Herrera, A.; Sanz, M.; Juan, T. Application of a statistical approach to the classification of honey by geographic origin. *J. Sci. Food Agric.* **1995**, *69*, 135−140.