Determining the Geographic Origin of Potatoes with Trace Metal Analysis Using Statistical and Neural Network Classifiers

Kim A. Anderson,*,†,‡ Bernadene A. Magnuson,† Matthew L. Tschirgi,‡ and Brian Smith§

Department of Food Science and Toxicology and Analytical Sciences Laboratory, Holm Research Center, University of Idaho, Moscow, Idaho 83844-2201, and Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164

The objective of this research was to develop a method to confirm the geographical authenticity of Idaho-labeled potatoes as Idaho-grown potatoes. Elemental analysis (K, Mg, Ca, Sr, Ba, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Mo, S, Cd, Pb, and P) of potato samples was performed using ICPAES. Six hundred eight potato samples were collected from known geographic growing sites in the U.S. and Canada. An exhaustive computational evaluation of the 608 × 18 data sets was carried out using statistical (PCA, CDA, discriminant function analysis, and *k*-nearest neighbors) and neural network techniques. The neural network classification of the samples into two geographic regions (defined as Idaho and non-Idaho) using a bagging technique had the highest percentage of correct classifications, with a nearly 100% degree of accuracy. We report the development of a method combining elemental analysis and neural network classification that may be widely applied to the determination of the geographical origin of unprocessed, fresh commodities.

Keywords: *Neural network; geographic authenticity; cononical discriminant analysis; discriminant function analysis; principal component analysis; elemental analysis; trace element analysis; potatoes*

Research on the determination of the geographic origin of commodities and food products is becoming an increasingly dynamic area. Financial incentives continue to drive retailers/resellers to misidentify the geographic origin of commodities and food products. The determination of the geographic origin is important for enforcement options for the food industry, protection of the consumer from overpayment and deception, and manufacturers using raw material that is variable (Desage et al., 1991; Flurur and Wolnik, 1994; Hernández and Rutledge, 1994a,b). The determination of the 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 and to a much lesser extent drugs of abuse, cocoa, coffee, and olive oil. Despite so many recent publications, as far as we know, geographic classification has never been applied directly to an unprocessed, fresh commodity using neural networks. Here we present a method for the determination of the geographic origin for potatoes from 2 different sites (identified as Idaho or non-Idaho) based on 2 growing seasons and representing a total of 608 × 18 data set.

Over two-thirds of all the research literature on the geographic origin of commodities involves the analysis of vitamins or other organic molecules (amino acids, triglycerides, etc.). Significant success (70−90% correct classification) has been reported using vitamin and/or amino acid assays to determine the geographic origin (Aires-De-Sousa, 1996; Ferland and Sadowski, 1992; Hulshof et al., 1997; Parcerisa et al., 1993, 1994, 1995); 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 (Lavedrine et al., 1997). 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 primarily because storage conditions either will not be known or will not be optimized for the analysis. Storage condition variability can and will compromise the analytical classifying technique. 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 since trace elements are significantly more stable in the commodity versus vitamins or some other types of organic compounds.

It is recognized that the 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. 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(s) must also pass through at least one cellular membrane in its movement from soil to plant. The selectivity of these processes of mineral accumulation within the

* Address correspondence to this author at the Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR 97331-7301.

† Department of Food Science and Toxicology.

‡ Analytical Sciences Laboratory.

§ Department of Pure and Applied Mathematics.

10.1021/jf980677u CCC: $18.00
Published on Web 03/20/1999

vegetable varies with different trace elements, with different plants, and the unique environment in which the commodity is grown.

The determination of the geographic origin of wines has been an active area of research for some years (Aires-De-Sousa, 1996; Armanino et al., 1990; Etievant et al., 1988; Latorre et al., 1994; Vanderschee et al., 1989). Most of 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 (Aires-De-Sousa, 1996; Armanino et al., 1990; Etievant et al., 1988; Latorre et al., 1994; Vanderschee et al., 1989). An excellent study by Day et al. (1995) 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. Many of the studies are a survey in nature (less than 30 samples), and therefore, general conclusions concerning the effectiveness of these techniques should be prudent.

The purpose of this study is to differentiate between potatoes grown in Idaho from the other potato-growing regions in North America. In the 1990s, Idaho became the number one potato producer for the U.S., growing nearly 30% of the nation's crop (Idaho Agricultural Statistics Service, 1992). Idaho produces approximately 13.8 billion pounds (U.S. units) of potatoes annually (Idaho Agricultural Statistics Service, 1992). In addition to being the production leader, Idaho is also the price leader. On average, from 1981 to 1990, retail prices for Idaho Russet potatoes were 2.3 times higher per hundred weight (cwt) versus Northeastern white potatoes (Idaho Agricultural Statistics Service, 1992). This consistent trend of price differential is largely the result of 60 years of successful advertising, promotion, and quality control. Quality control of potatoes is based on taste, solids, texture, and consistency. The Idaho potato is especially high in solids (21%), as compared to other growing regions, creating the "fluffy"/dry consistency recognized by consumers. The potato industry in Idaho represented over 2.5 billion dollars to the state's economy, representing over 15% of Idaho's gross income (U.S. Department of Commerce, 1992). Protecting Idaho's market share, reputation, and consumer confidence to pay a premium for Idaho potatoes is meaningful to the industry and the state's economy. Unscrupulous resellers/retailers misidentifying potatoes from Idaho cost Idaho growers both in the short and long term as well as deceiving the consumer. Further lasting effects include jeopardizing consumer confidence in the quality of Idaho potatoes (if they have been unknowingly switched with lower quality potatoes) and affecting the consumers' willingness to pay premium prices for Idaho potatoes. Therefore, developing a method that can identify the origin of potatoes is important to protect the potato industry as well as the economy of Idaho. We report the development of a method capable of determining the geographical origin of fresh potatoes with nearly 100% degree of accuracy, using neural network classification of the concentrations of 14 elements in potatoes sampled from 2 seasons.

METHOD

Reagents. The source of chemicals and reference materials was as follows: concentrated, nitric acid trace metal analysis grade (J. T. Baker, St. Louis, MO); elemental stock standard solutions (J. T. Baker, St. Louis, MO); 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, Ontario, 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 LPM (liters/min), nebulizer 41 psi, auxiliary flow 0.20, pump rate 1.0 mL/min, scan intergration time 0.25 s, Mn1 peaking wavelength, acid flexible tubing 0.030 mm ID (internal diameter), wavelengths and background corrections previously reported (Anderson, 1996). The temperature controller/digester used was a digestion system 40, 1016 digester, and Autostep 1012 controller (Tecator, Sweden), fitted with an aluminum adapter plate 3 mm thick with 40−17-mm holes on top overlaid on the heater block.

Sampling, Preparation, and Analysis. To ensure that we used only authentic samples with precisely known origin, samples where gathered by the Idaho Potato Commission (IPC) or one of their delegates directly from farms or producers' storage units. Samples were shipped within days of collection with chain-of-custody documentation to the University of Idaho Analytical Sciences Laboratory. Samples were stored under controlled access at 4 °C until analysis, typically within 2 weeks.

Potatoes were collected from most major fresh-market geographic locations in North America (U.S. and Canada), based primarily on the number of acres in fresh potato production. Idaho potatoes are grown primarily in the Snake River Plain, representing about 400 000 acres in current potato production. Within this region, 342 samples were collected from widely distributed locations within Idaho. Non-Idaho potatoes include samples taken from the following geographic locations: Colorado, Washington, Wisconsin, Maine, Michigan, and Canada (Prince Edward Island and New Brunswick). Two hundred sixty-six non-Idaho samples were collected. For the purposes of this paper, all of these regions are combined into one category called non-Idaho.

Each potato was hand-rinsed under a stream of tap water for 20−30 s. Dirt was removed by gently rubbing by hand under the water stream. After rinsing, the potatoes were shaken to remove any excess water, gently blotted with a paper towel, and placed in a lab-mat covered tub to air-dry prior to processing (1−2 h). A ca. 1.0-g cross-sectional slice of whole tuber was taken, see Figure 1, 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. Similarly, a ca. 1-g sample of pulp only was taken from each potato as a cross-sectional slice, and a 1−3-mm-thick slice of peel was taken. These three samples represented the whole tuber, pulp only, and peel only subsamples, respectively.

The samples were allowed to react for ca. 4−8 h in a hood at ambient temperature. Then the samples were digested using a heating block (or a programmable digester may be used). The samples were heated to 180 °C for 3−4 h. Digestion was confirmed complete when no nitrous oxide gases were evolved (i.e., orange gas production). Samples were diluted with type 1 water (18 Ω·cm) and mixed thoroughly using a vortexer. Analysis was by ICPAES.

The percent moisture for each individual sample was determined in duplicate. The percent moisture method used was a modification of AOAC Method 984.25 (Association of Official Analytical Chemists, 1990). The samples are placed

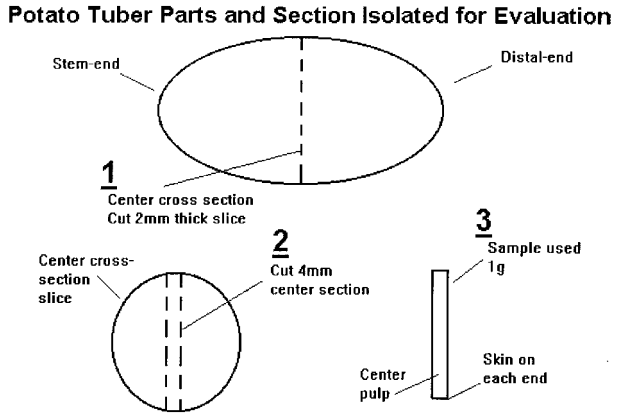

Figure 1. Graphical representation of the sampling technique developed for subsampling the potato.

Table 1. Recovery and Standard Deviations of Seven Different Standard Reference Materials Used during the Study[a]

	% recovery						
element	oyster tissue[b]	pine needle[c]	rice flour[d]	lobster[e]	corn bran[f]	alfalfa[g]	bovine liver[h]
Ba	NA	NA	NA	NA	96.1	NA	NA
Ca	101.4	105.3	112.1	NA	106.1	99.2	113.2
Cd	115	NA	NA	116.1	NA	NA	99.3
Co	BDL	NA	NA	112.3	NA	NA	NA
Cr	NA	102.8	NA	NA	NA	113.5	NA
Cu	102.8	115.0	112.8	95.0	115.0	104.4	106.9
Fe	95.1	86.2	89.5	94.9	92.1	97.5	108.9
K	102.7	105.1	104.7	NA	107.6	109.4	105.7
Mg	103.6	NA	97.1	NA	107.8	96.4	116.4
Mn	BDL	102.3	99.7	84.7	85.0	NA	85.7
Mo	NA	Na	136.4	114.4	NA	NA	116.2
Ni	123.9	NA	NA	101.6	NA	NA	NA
P	103.4	108.9	112.3	NA	103.9	104.5	120.5
Pb	NA	116.4	NA	NA	NA	NA	NA
S	104.0	NA	104.0	NA	102.3	108.2	106.6
V	123.5	NA	NA	103.7	NA	NA	NA
Zn	105.4	NA	108.3	105.3	104.0	104.5	107.6

[a] NA = certified value not available, BDL = below detection limit. % SD = percent standard deviation; reference values not available for strontium. [b] $n = 72$, NIST SRM 1566a oyster tissue, % SD ranged from 7.9 to 9.8%, Ni is near MDL, and % SD was >10% [c] $n = 95$, NIST SRM 1575 pine needles, % SD ranged from 8 to 22%. [d] $n = 90$, NIST SRM 1568a rice flour, % SD ranged from 7 to 22%. [e] $n = 8$, CRC TORT2 lobster hepatopancreas, % SD ranged from 2 to 22%, Mo is near MDL, and % SD was >10%.[f] $n = 84$, NIST SRM 8433 corn bran, % SD ranged from 10.3 to 23%. [g] $n = 12$, house SRM alfalfa, % SD ranged from 3.5 to 12%. [h] $n = 3$, NIST SRM 1577a bovine liver, % SD ranged from 1 to 5%.

in a convection oven at 105 °C for 5 days. The mineral and trace element concentrations were standardized to a dry weight, based on the moisture content.

Quality Control. Each analytical batch contained a minimum of 25% quality control samples, including check standards, duplicates, spikes, and standard reference materials (SRMs). The percent recovery and percent standard deviation for SRM are given in Table 1. During the course of the study, over 360 SRM samples were analyzed; SRMs were dominantly plant matrixes where available; in all cases, the SRMs represented analyte concentration ranges typically found in plant tissues. The percent recovery ranged from 86 to 136%. The percent standard deviation ranged from 2 to 39%. Typical percent standard deviation (% SD) was <10%, although analytes close to method detection limits (MDLs) had higher % SDs. Spike recoveries and check standards where typically within ±10% of their true value.

Computational Analysis. The data were analyzed in an effort to classify potato samples as having originated from Idaho or from outside Idaho based on the trace element profile of each sample. Basic statistical analyses and several pattern recognition methods were applied to the data. Neural network methods were applied utilizing software (NeroShell for Microsoft Windows, Release 4.6) supplied by Ward Systems Group Inc. (Frederick, MD). Neural network analysis included variants of feed-forward back-propagation architectures and included combining classifiers in a "bagging" strategy. Basic statistical analyses and pattern recognition techniques, excluding neural network, were performed utilizing the SAS System for Windows analysis package (Release 6.11, SAS Institute Inc., Cary, NC). Descriptive statistics included the following: Student *t* test, assessment of normality of the data distribution, principal component analyses, canonical discriminant analysis, discriminate function analyses, nonparametric *k*-nearest-neighbor analyses. The data set was standardized to account for differing variable scales by subtracting from each entry its associated variable mean and then dividing by the variable standard deviation. The standardized data corresponding to each variable thus has a mean equal to zero and standard deviation equal to one.

Descriptive Statistics. Descriptive statistics (the mean, standard deviation, minimum and maximum values) for each element in each group were determined (SAS). The TTEST procedure was used to compute a *t* statistic for testing the hypothesis that the means of the elemental concentrations of the two groups of potatoes are equal. The UNIVARIATE procedure was used to test for normality using the Shapiro−Wilk statistic and data distribution plots. Small values of *W* lead to the rejection of the null hypothesis.

Principal Component Analysis. The principal component analysis (PCA) generates principal components that are linear combinations of the original variables. The first principal component (PC) describes 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 original variables. Analyzing the data with respect to principal components can thus sometimes effectively reduce the number of variables, especially if a large percentage of the total variation is described by a few principal components. One- or two-dimensional plots of data with respect to selected principal components can sometimes provide visual insight into the data, offering a visual description of group differences or clustering, and outliers. PCA has been applied to geographical classification applications of various foods including processed orange juice (Nikdel et al., 1988), wine (Latorre et al., 1994; Day et al., 1995), honey (Sanz et al., 1995), and cocoa (Hernández and Rutledge, 1994a,b). PCA was applied to our data using the PRINCOMP procedure, and the details of the results appear under Results and Discussion.

Canonical Discriminant Analysis. 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 total variation of the data. Like PCA, CDA can be used to effectively reduce the number of variables and is particularly useful for producing one- or two-dimensional visualizations of the data since the "views" optimize the between-class differences. The default number of canonical variables generated is the minimum of the number of classes minus one and the number of original variables. Different views of our data were obtained by defining the number of classes to be two (Idaho versus non-Idaho). CDA has been applied to data for the purpose of geographical classification of wine (Day et al., 1995). The SAS procedure used for our analysis was the CANDISC procedure, and the results are discussed under Results and Discussion.

Discriminant Function Analysis. The DISCRIM procedure was used for both parametric and nonparametric discriminant function analyses. The parametric procedure determines a discriminant function of classification criterion by a measure of the generalized squared distance (Rao, 1973). This procedure assumes a multivariate normal distribution. Selection of variables to be included is discussed under Results and Discussion. In this case, the classification criterion was

based on an individual within-group covariance matrix, yielding a quadratic function. There was no difference in the classification of samples when either equal or prior probabilities of the groups were used (data not shown). Two error rates are computed. The first is an estimate of the probability of misclassification of future samples using the discriminant function created by the entire training set ($n = 608$). The second is the error rate incurred during a cross-validation step, in which each sample is removed from the training set and tested against the resultant discriminant function created by the remaining samples ($n = 607$). In all cases, error rates given under Results and Discussion are those from the cross-validation test. Validation of the discriminant function was also conducted by withholding one-half of the samples from the training set and using them as a test set against the discriminant function created by the remaining 304 samples. This was then repeated in reverse. The nonparametric procedure used was the *k*-nearest-neighbor method, where $k = 10$. As no assumption is made in this procedure regarding the nature of the data set, all variables were included.

Neural Network Analysis. Feed-forward back-propagation neural network methods were also applied to the data in an effort to classify the samples by geographic origin as Idaho or non-Idaho samples. To prevent over*fitting* or over*training*, an early stopping strategy was employed to enhance the ability of the networks to generalize well (perform well on new data). The data were divided into two disjoint subsets: a training set and a test set. Networks were trained using half of the data (training set). During the training process, the remaining half of the data (test set) was periodically presented to the networks for classification. The final values of the network parameters were those corresponding to optimum test set performance. Further generalization enhancements are possible by employing a bootstrap aggregation (bagging) strategy (Breiman, 1996). Here multiple networks are trained using randomly selected (sampling with replacement) training sets corresponding to half the data. Final classification is then determined by voting. This has the effect of reducing the high variance inherent in neural networks, resulting in improved generalization. The results of these strategies are discussed under Results and Discussion.

RESULTS AND DISCUSSION

Chemical Analysis. There are several unique aspects to optimizing a set of chemical measurements that can be used to determine geographic origin of fresh commodities. This includes the determination of the most appropriate portion of the commodity to test, determination of factors that might mask or dominate over subtle trends, and determination of the most applicable set of chemical measurements to be made on the sample of choice.

Fresh commodities may be stored for long periods (1−9 months); during storage, fresh produce may lose moisture. For example, in a study on walnuts and storage influence, the authors proposed that even at 4 °C (3 months), desiccation of the walnuts occurred (Lavedrine et al., 1997). In the case of potato tubers, the percentage of water may vary 5−20% from the time of harvest to the time of use (1−9 months later). The percent moisture content will affect the relative concentration of trace elements (e.g., weight/weight). Therefore, the percentage moisture must be equalized such that it does not dominate or mask the variations of the elemental concentrations, which are due to geographic growing conditions of fresh commodities versus the effects of dehydration during storage. The potato tuber was not dried prior to subsampling due to the difficulty in subsampling a portion that had a consistent pulp/skin ratio (see below). Desiccation, by freeze-drying, would be a viable option; however, this equipment was not available to us. Therefore, the percent moisture was determined (in duplicate) for each individual tuber. The percent moisture was then used to determine the elemental concentrations on a dry weight basis for each individual tuber. In this way, the loss or variation of water would not mask the variations that are due to geographic growing conditions. The procedure developed here was tested with samples over 4 months and found that when the % moisture was compensated for, the elemental concentrations were consistent regardless of storage time. This method therefore is robust in its applicability independent of storage time.

It has been reported that the elemental distribution in a fresh commodity will be different for different parts of the commodity (Esechie, 1992). For example, the concentration of various elements within a potato will be different in the skin versus the pulp. There is evidence (Anderson, unpublished results) that some elements may be concentrated in the potato skin relative to the potato pulp. In addition, some elements in the skin may be an enhanced (or distorted) reflection of geographic conditions. However, the pulp, which represents the largest portion by weight of the commodity, may have unique elemental distribution tendencies relative to other portions of the commodity. Therefore, the challenge is to analyze sample components that maximize the effects of geographical conditions and yet are reasonable to prepare for analytical determination.

Three sample component parts for the potato commodity were analyzed; skin only, pulp only, and whole tuber. A preliminary data analysis using 70 samples (computational modeling) was used to screen the viability of each sample component part. In addition, practical aspects such as the reliability and consistency that could be achieved at the bench level during sample preparation for the chemical analysis were evaluated. The most optimal sample component type was determined to be whole tuber. However, an important caveat of this sample type was the importance of the ratio of skin to pulp. It was determined that the skin-to-pulp ratio (by weight) should be consistent between all samples. A protocol was developed which provided a method to subsample from the tuber that could consistently represent the same pulp/skin ratio (see Figure 1).

Elemental distribution within a single commodity component (e.g., pulp only) may vary within the commodity itself. For example, there is evidence that some chemicals within a potato tuber are not evenly distributed in the given potato component (i.e., the pulp) from the stem end to the distal end (Al-Saikhan et al., 1995). Here we developed a protocol that isolated a consistent potato tuber section. The center section was determined to be the least affected by any variations that might exist between the stem end and the distal end.

The drying of a plant tissue sample is a balancing act between too low of a temperature over a prolonged period that will encourage and promote biological activity and too high of a temperature over a short period that may result in the loss of volatile analytes. We performed a 10-day study ($n = 3$) of drying times versus temperatures. After 5 days, the percentage of moisture at 105 °C changed by less than 0.2% on average. Lower temperatures (<85 °C) required longer drying times (>7−8 days), which risked biological growth, and temperatures > 105 °C were determined to increase the risk of other volatile analytes losses. The above-described

Table 2. Dry Weight Elemental Analysis (mg/kg) of Idaho and Non-Idaho Potatoes

element	location	N	mean	SD	min[a]	max	P[b]
Ba	Idaho	342	1.43	0.63	0	3.66	0.000
	non-Idaho	266	1.78	1.72	0	7.61	
Ca	Idaho	342	532.7	164.0	197.1	1172.2	0.000
	non-Idaho	266	357.5	184.3	100.1	1163.7	
Cd	Idaho	342	0.33	0.33	0	1.52	0.004
	non-Idaho	263[c]	0.25	0.35	0	1.48	
Cr	Idaho	342	0.38	0.48	0	2.7	0.000
	non-Idaho	266	0.73	0.63	0	2.12	
Co	Idaho	342	0.44	0.68	0	2.96	0.933
	non-Idaho	266	0.43	0.78	0	3.28	
Cu	Idaho	342	4.26	1.47	0	8.54	0.000
	non-Idaho	266	5.60	2.76	0	18.15	
Fe	Idaho	342	34.95	13.7	12.93	90.83	0.000
	non-Idaho	266	40.58	17.70	11.71	131.05	
K	Idaho	342	20902.7	3370.7	10281.9	32770.4	0.167
	non-Idaho	266	21259.6	2866.0	13587.7	31277.1	
Mg	Idaho	342	1204.6	190.0	766.2	2015.2	0.013
	non-Idaho	266	1166.0	191.9	732.5	1858.5	
Mn	Idaho	342	6.85	1.74	1.61	18.69	0.000
	non-Idaho	266	10.43	5.03	1.46	35.25	
Mo	Idaho	342	0.38	0.82	0	3.47	0.217
	non-Idaho	266	0.47	0.99	0	5.07	
Ni	Idaho	342	1.03	1.01	0	4.35	0.011
	non-Idaho	266	0.8	1.33	0	4.98	
P	Idaho	342	2506.4	686.8	1173.0	5135.0	0.144
	non-Idaho	266	2585.0	681.7	1252.0	4424.4	
Pb	Idaho	342	2.09	2.68	0	10.12	0.007
	non-Idaho	266	1.49	2.87	0	12.83	
S	Idaho	342	1675.8	322.2	1049.6	4020.5	0.000
	non-Idaho	266	1562.4	310.1	919.5	2530.8	
Sr	Idaho	342	2.38	1.26	0	6.08	0.000
	non-Idaho	266	1.59	1.81	0	9.89	
V	Idaho	342	1.29	1.12	0	4.32	0.106
	non-Idaho	266	1.12	1.57	0	7.1	
Zn	Idaho	342	12.58	3.4	3.95	23.48	0.000
	non-Idaho	266	17.79	4.83	6.54	58.18	

[a] For statistical analysis, below detection limit samples were set at a value of 0. [b] *t* statistic. [c] Three cadmium data results accidentally not collected.

procedure therefore was determined to be optimal for fresh commodities by minimizing any volatilization and producing a consistent dried weight while avoiding biological growth.

An important attribute of this approach is that all of the chemical data can be determined with the use of only a single analytical instrument, ICPAES. Whereas other geographic authenticity approaches require the use of several instruments and sophisticated approaches for data analysis, this technique requires only a single, commonly, available instrument. In this approach, the data are used directly from the ICPAES into the computational models requiring no prior mathematical or interpretive analyses as is often the case with other geographic authenticity approaches.

The Idaho Snake River Plain is a unique area composed of rich volcanic soil in an arid to semiarid (irrigated) environment. The soils in this region are xerolls, which are unique as compared to other potato-producing geographic regions. The soil and environmental growing conditions clearly provide mineral and trace element tuber uptake that is unique and provide the necessary chemical profile difference to differentiate between potatoes grown in Idaho versus outside Idaho.

Computational Analysis. *Descriptive Univariate Statistics.* The means, standard deviations, and minimum and maximum values for the elemental content of potatoes from Idaho and non-Idaho locations are shown in Table 2. Idaho potatoes had higher concentrations of Ca, Cd, Mg, Ni, Pb, S, and Sr compared to non-Idaho potatoes, whereas the concentrations of Ba, Cr, Cu, Fe, Mn, and Zn were lower in Idaho compared to non-Idaho potatoes. The concentrations of Co, K, Mo, P, and V in the two groups were not significantly different. Despite these differences, examination of the minimum and maximum values illustrated that there was not a single element that could correctly classify the potato samples as to location, as the ranges of concentration for each group overlapped for every element. Therefore, multivariate classification techniques were examined.

Table 3. Principal Component (PC) Analysis of the Elemental Analysis of Potatoes

	eigenvalue	proportion	cumulative
PC 1	2.701	0.1931	0.1931
PC 2	2.378	0.1698	0.3630
PC 3	1.779	0.1271	0.4901
PC 4	1.314	0.0938	0.5839
PC 5	1.043	0.0744	0.6584

Tests for Normality. The concentrations of several elements in the potato samples were very close to the detection limit of the chemical analysis method. For the purposes of statistical analyses, any value that was below the detection limit was set to a value of zero. This resulted in highly nonnormal distributions (*W* less than 0.8) for Co, Mo, and Pb. These variables were subsequently eliminated from parametric analyses (PCA and discriminant function). Cr, Ni, and V were also somewhat nonnormal with *W* less than 0.9. Each of these variables was systematically tested for contribution to the parametric discriminant function analysis as described below.

Principal Component Analysis. PCA demonstrates that a small number of variables were not dominating the total variability, as the first three principal components accounted for only 49% of the total variability (Table 3). Only modest visual clustering was apparent when the data were displayed with respect to the first two principal components. This was not surprising since the first principal component accounts for the maximum possible one-dimensional projection of the total variation of the individual data points, which does not necessarily correspond to the maximum variation between defined classes. Better visual results were obtained using CDA (discussed below).

Canonical Discriminant Analysis. CDA was applied to the data using two defined classes, Idaho and non-Idaho. Figure 2 shows a frequency chart of the data using the first canonical variable and depicts a reasonably good separation of classes using only one variable.

Discriminant Function Analyses. The addition of V (vanadium) values to the parametric discriminant function, generated with the 15 remaining elements, increased the number of misclassified samples, and they were therefore removed from the analyses. Elimination of either Cr or Ni values reduced the number of misclassified samples, and therefore, these variables were included in the model. The final model included 14 elements (Ba, Ca, Cd, Cr, Cu, Fe, K, Mg, Mn, Ni, P, S, Sr, and Zn). The error rates of the quadratic discriminant function calculated using 14 element concentrations of 342 known Idaho potato samples and 266 known non-Idaho potato samples were 3.5 and 5.6% respectively, resulting in 330 Idaho (97%) and 251 non-Idaho (95%) correctly classified samples in cross-validation testing (Table 4). The data set was randomly divided into 2 halves of 304 samples. Cross-validation testing using only 304 samples as the calibration or

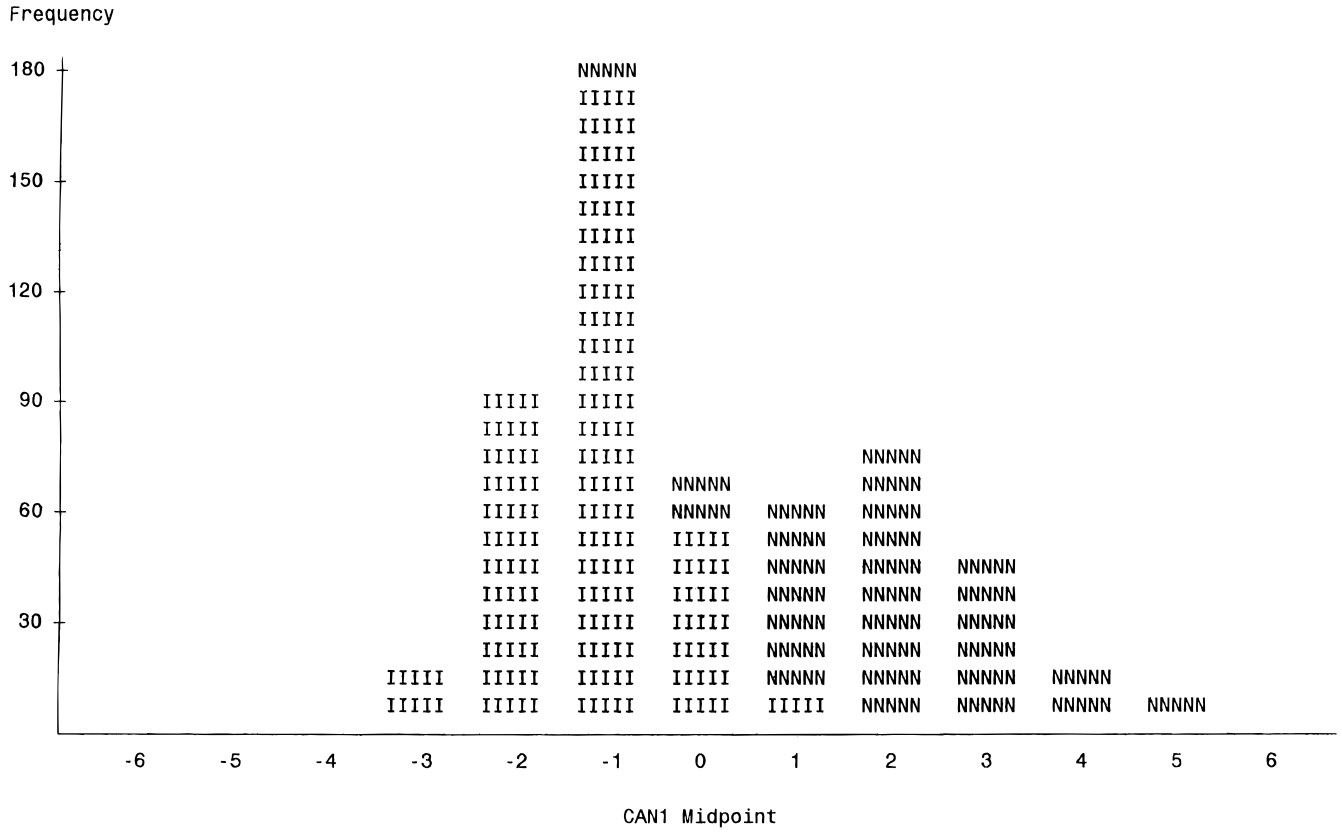

Figure 2. CDA frequency chart using the first canonical variable. symbols: I = Idaho, N = all non-Idaho locations. This simplified visual representation demonstrates the separation of Idaho versus non-Idaho; any overlap viewed in this one-dimensional representation is not indicative of any intractable classification task; all 14 available dimensions are utilized; see text.

Table 4. Parametric Discriminant Function Analysis of the Elemental Concentrations of Potatoes

		error rates for	
training set	test set	Idaho	non-Idaho
entire database, $n = 608$	cross-validation	3.5% (12/342)	5.6% (15/256)
half (no. 1) of database, $n = 304$	cross-validation	4.7% (8/171)	4.5% (6/133)
	remaining database ($n = 304$)	4.7 (8/171)	6.0% (8/133)
half (no. 2) of database, $n = 304$	cross-validation	4.1% (7/171)	9.3% (11/133)
	remaining database ($n = 304$)	4.7% (8/171)	3.8% (5/133)

training set had error rates of 4.1 and 4.7% for known Idaho potatoes and 4.5 and 9.3% for non-Idaho potatoes (Table 4). When the remaining database of known samples was used as a testing set against the 304 potato training set, error rates were 4.7% for the Idaho potatoes and 3.8−6.0% for the non-Idaho potatoes (Table 4). The nonparametric *k*-nearest-neighbor analysis using all 18 variables and all 608 samples gave low error rates for Idaho potatoes (1.2%) but had higher error rates for non-Idaho potatoes (8.4%).

Neural Network Analysis. Originally all 18 candidate trace metals were considered. It was found that superior classification results were obtained by considering only the 14 trace metals used in the parametric discriminant function analysis. This is most likely attributable to the fact that for a large number of samples, the measured quantities of the four unused trace metals were below detection limits, resulting in artificially truncated frequency distributions for these metals.

An early stopping strategy was first examined. Fifty neural network models were generated. Each model used 50% (304 samples) of the data for the training set and 50% for the test set. The model architecture was the same for each model; the difference in the models was due to the difference in training and test sets, which were selected randomly (as disjoint complements) for each model. Individual model classification performance on the known data (training and validation sets together) ranged from 92 to 98%.

To investigate a bagging strategy, a universal test set of 46 samples was selected from the original data and set aside. This universal test set was selected so as to represent a typical cross section of the original data. Fifty neural network models were then generated using the remaining data (562 samples), which was now considered as the "known" data set. As before, each individual model was generated using 50% (281 randomly selected samples) of the known data for training, and the remaining complementary set was used as a test set. Individual model performance ranged from 92 to 98% correctly classified on the known data (training and test sets together) and 89−98% on the universal test set. Generally the relative performances of individual models on the known data and the universal test set were not strongly correlated. When the 70 independent classifiers were combined (bagged), the resulting aggregate model correctly classified 98% of the universal test set samples, missing only 1 out of the 50 samples.

Also, when the aggregate model was applied to the known data set, over 99% of the samples were correctly classified.

To infer a comparison between how well our best neural network strategy (bagging) and the optimized parametric discriminant function analysis would perform, the universal test set (46 samples) was removed and parametric discriminate functions were generated using the remaining data. The discriminant function analysis correctly classified 89% of the universal test set (41 out of 46 samples) and 95−96% of the known data set (562 samples) in cross-validation testing. Therefore, neural network bagging does appear to be a worthwhile strategy, producing superior results over single-model discriminant analysis.

Conclusion. The concentration of selected mineral and trace elements in potatoes was used to differentiate between potatoes grown in Idaho or outside of Idaho (non-Idaho). The elemental determinations can be done precisely and accurately using commonly available automated equipment and enable the geographical origin authentication of potatoes to be considered as a standard procedure. The content of selected minerals and trace elements is a reflection of the soil type and, importantly, the environmental growing conditions. The geographic origin of potatoes can be determined by their chemical profile. Statistical analysis revealed groupings between the Idaho and non-Idaho potatoes; however, simple inspection of elemental concentrations cannot be used to differentiate the growing origin. Use of neural network models and discriminate function analysis both successfully classified potatoes relative to their origin; however, higher rates of correct classifications (98−99%) were obtained with neural networks using a bagging strategy. The nature of some potato varieties modifies slightly the mineral and trace element profiles. The nature of seasonal and environmental conditions may also slightly modify the mineral and trace element profiles. Here we included two seasons of potatoes and found that seasonal variation did not compromise the model's classification success. Work is in progress to further substantiate the effects of seasonal and potato variety influences. As well, further classification breakdown of subregions is currently underway in our laboratory.

ACKNOWLEDGMENT

We thank Dr. Seifollah Nikdel, FDOC, and Bill Price, University of Idaho, for helpful discussions.

LITERATURE CITED

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

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

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

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.

Association of Official Analytical Chemists. *AOAC Official Methods of Analysis of AOAC International*; Association of Official Analytical Chemists: Washington, DC, 1990.

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

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.

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.

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.

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.

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.

Flurur, C.; Wolnik, K. Chemical profiling of pharmaceuticals by capillary electrophoresis in the determination of drug origin. *J. Chromatogr. A* **1994**, *674*, 153−163.

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.* **1994a**, *49*, 83−93.

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

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.

Idaho Agricultural Statistics Service. Idaho Agricultural Statistics. 1992.

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.

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.

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

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.

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.

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.

Rao, C. R. *Linear Statistical Inference and Its Applications*, 2nd ed.; Wiley: New York, 1973.

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.

Smith, B.; Anderson, K. A. Defining Sub-Regional Geographic Classification of Potatoes. Manuscript in preparation.

U.S. Department of Commerce Bureau of Economic Analysis, Regional Division, 1992.

Vanderschee, H. A.; Bouwknegt, J. P.; Tas, A. C.; Maarse, H.; Sarneel, M. M. The authentication of sherry wines using pattern-recognition − an inter laboratory study. *Z. Lebensm. Unters. Forschung.* **1989**, *188*, 324−329.

Received for review June 19, 1998. Revised manuscript received January 14, 1999. Accepted January 19, 1999. We thank the Idaho Potato Commission for partial funding of the project.

JF980677U