Received: 16 December 2011, Revised: 18 January 2012, Accepted: 29 January 2012, Published online in Wiley Online Library: 10 April 2012

(wileyonlinelibrary.com) DOI: 10.1002/cem.2422

Support vector machine classification of suspect powders using laser-induced breakdown spectroscopy (LIBS) spectral data

Jessi Cisewski[a]*, Emily Snyder[b], Jan Hannig[a] and Lukas Oudejans[b]

Classification of suspect powders, by using laser-induced breakdown spectroscopy (LIBS) spectra, to determine if they could contain *Bacillus anthracis* spores is difficult because of the variability in their composition and the variability typically associated with LIBS analysis. A method that builds a support vector machine classification model for such spectra relying on the known elemental composition of the *Bacillus* spores was developed. A wavelet transformation was incorporated in this method to allow for possible thresholding or standardization, then a linear model technique using the known elemental structure of the spores was incorporated for dimension reduction, and a support vector machine approach was employed for the final classification of the substance. The method was applied to real data produced from an LIBS device. Several methods used to test the predictive performance of the classification model revealed promising results. Published 2012. This article is a US Government work and is in the public domain in the USA.

Keywords: classification; support vector machine; wavelet; dimension reduction; laser-induced breakdown spectroscopy

1. INTRODUCTION

When a large building, complex, or area has been contaminated with a powder substance that may contain *Bacillus* spores (causative agent for anthrax), it is crucial to determine if the substance is potentially harmful quickly and efficiently. These powder substances could be nonhazardous hoaxes (e.g., dust, chalk, or sugar), but they could also actually be or contain *Bacillus anthracis*. Laser-induced breakdown spectroscopy (LIBS) devices have the capability of generating characteristic spectra that can aid in determining if a substance is or contains a spore material like *B. anthracis*. In LIBS, a laser is focused onto a sample producing a plasma. This plasma atomizes, ionizes, and subsequently excites the interrogated sample. The light emitted from the plasma is collected, generating a characteristic wavelength spectrum. LIBS is an attractive technique for field analysis of suspect powders because it does not require preparation of samples, yields spectra in real time, and is easily made man-portable.

Differentiation of *B. anthracis* spore powder LIBS spectra from LIBS spectra of other innocuous powders via classification methods can be difficult because of the inhomogeneity of the spore powder itself and the variability typically associated with LIBS spectra. This variability is seen even when employing the same LIBS system and is because of a range of factors including the following: pulse-to-pulse variations in the laser energy and profile, sample topography (directly affects the distance of the plasma to the collection lens, which subsequently impacts the distance from the plasma to the collection fiber), creation of sampling craters (can be avoided by moving to a fresh spot for each laser shot), physical and chemical characteristics of the sample (surface adsorption, reflection, and thermal conductivity, which are determined by the composition, roughness, color, and moisture content of the sample), and matrix effects [1–5]. Normalization methods, such as the use of other emission lines from elements in the surrounding gas or reference elements in the matrix, and the use of excitation temperatures and/or electron temperatures are often applied to correct for these matrix effects [1]. However, these corrections are frequently not an option with heterogeneous samples and/or when ungated (nonintensified) charge-coupled device (CCD) detectors are used. These ungated CCDs are found on less expensive and portable LIBS systems.

Statistical methods have been implemented to overcome this issue, particularly in the area of analysis of biological agent powders. Using partial least squares discriminant analysis (PLS-DA), Gottfried and coworkers were able to differentiate stand-off LIBS spectra intensity ratios of pure spore powders and other powders such as talcum powder, sugar, dust, and flour on aluminum and glass substrates [6]. Previously, biological agent surrogate spectra were classified using linear and rank correlation [7]. These statistical techniques had difficulty distinguishing spore spectra, specifically *Bacillus atrophaeus* spectra, in mixtures of potentially interfering compounds such as urban particulate matter. Munson and coworkers explored the use of soft independent modeling of class analogies for classification of three *Bacillus* species, molds, Arizona road dust, and pollens as well as a mixture of Arizona road dust and *Bacillus globigii*. They found that soft independent modeling of class analogy models could be used to distinguish between spores

* Correspondence to: J. Cisewski, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 318 Hanes Hall, Chapel Hill, NC 27599-3260, USA
E-mail: cisewski@email.unc.edu

a J. Cisewski, J. Hannig
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260, USA

b E. Snyder, L. Oudejans
US Environmental Protection Agency, Office of Research and Development, National Homeland Security Research Center, Research Triangle Park, NC 27711, USA

in mixtures of the road dust and the road dust itself [8]. PLS-DA was also able to distinguish between spores in mixtures of the road dust and the road dust itself [6]. Employing these statistical methods still did not sufficiently resolve the issues of false positives for some materials (fertilizer and outdoor air particulate matter).

Other statistical methods have also been employed for classification of LIBS spectra of heterogeneous samples. Rehse *et al*. used discriminant function analysis to discriminate LIBS spectra of one genus of bacteria, applied as a thin smear on an agar plate, from another and obtained greater than 90% accuracy regardless of the nutrient medium in which the bacteria were cultured [9]. Hierarchical cluster analysis, artificial neural networks, and PLS-DA were used to classify LIBS spectra of chicken tissue samples (kidney, lung, liver, brain, muscle, and spleen) [10]. Artificial neural networks were also employed to classify rocks and soils, and average classification accuracy of 78% was observed when spectra that were not used to train the original model were classified (including some spectra of unknown rock and soil materials) [11].

In this paper, a new statistical technique is presented for distinguishing *B. anthracis* surrogate spore powders from other innocuous suspect powders by using LIBS spectra that can potentially provide better classification than the previously mentioned efforts. The proposed method exploits a known property of the *B. anthracis* and its surrogate spores to aid in classification. Specifically, there are eight elements typically detected in *B. anthracis* and their surrogate spores [7]. The principal idea of the proposed technique is, after preprocessing the data with the use of wavelets, to determine the combination pattern of the LIBS spectra for those eight elements that form the LIBS spectra of *B. anthracis* spore and other innocuous powders by using linear regression. The combination pattern is then used to build a classification model by using a support vector machine (SVM) approach to classify the substance as harmful, that is, *B. anthracis* spore, or not. The methodology was developed using pure spore specimens, pure confusant samples, and pure elements.

2. EXPERIMENTAL

This section describes the LIBS system and the materials used in the study. Spectra were collected for biological (*B. anthracis* spore surrogates) and nonbiological substances.

2.1. Laser-induced breakdown spectroscopy system

Data were collected on a bench-top LIBS system that consisted of a CFR400 Nd:YAG laser (Big Sky, Bozeman, Montana) operating at the fundamental wavelength of 1064 nm, a pulse duration of 8 ns, and maximum pulse energy of 400 mJ; a series of focusing and collection optics; and an LIBS 2000 broadband (200–980 nm with 0.1 nm resolution) spectrometer (Ocean Optics, Dunedin, Florida). During operation, a single laser pulse (≈65 mJ/pulse) from the laser is triggered by the LIBS software. This beam passes through a pierced parabolic mirror and is focused onto the sample surface with a 5-cm lens, producing the LIBS plasma. The resulting plasma emission is reflected by the pierced mirror to a 10-cm focal length lens that focuses the plasma emission onto a fiber optic bundle consisting of seven fibers. The fiber bundle delivers light to a broadband spectrometer that contains seven CCDs. Throughout operation of the system, the laser and the spectrometer are controlled by the Ocean Optics, Incorporated LIBS software. All spectra were taken at a delay time (time after plasma initiation) of 1.5 μs. Collection of plasma emission at this delay time optimizes the ratio of elemental emission lines to background plasma continuum emission.

2.2. Materials

Pellets (2.54 cm in diameter, 2–4 mm in thickness depending on substance) were made using a pellet press (XPRESS 3630, SPEX Sample Prep Metuchen, NJ) that applied 20 tons of pressure for 30 s. The nonbiological powder pellets analyzed via the bench-top LIBS system were as follows: Food Lion brand flour, Arm & Hammer detergent, Rumford baking powder, Arm & Hammer baking soda, BC powder, Crayola chalk, DiPel 150 dust, Equal artificial sweetener, Gain laundry detergent, Advil ibuprofen tablets, Johnsons baby powder, Food Lion brand powdered sugar, Food Lion brand sugar, Sweet'n Low artificial sweetener, Tide laundry detergent, and Tylenol acetaminophen capsules. The 5% elemental standard powder pellets were made from the following powders: magnesium sulfate (99%; Sigma Aldrich, St. Louis, MO; yields magnesium spectral lines), sodium chloride (99.999%, Sigma Aldrich, yields sodium spectral lines), potassium iodide (≥99.0%, Sigma Aldrich, yields potassium spectral lines) ferric sulfate hydrate (97%, Sigma Aldrich, yields iron spectral lines), manganese(II) sulfate monohydrate (≥98.0%, Sigma Aldrich, yields manganese spectral lines), sand (white quartz 50 + 70 mesh, Sigma Aldrich, yields silicon spectral lines), graphite (99.99% 100 mesh powder, Sigma Aldrich, yields carbon spectral lines), calcium chloride (≥99%, Sigma Aldrich, yields calcium spectral lines), and boron oxide (99.999%; Alfa Aesar, Ward Hill, MA; has minimal spectral features). Boron oxide was used as the diluent for the elemental standards because of its inertness and low spectral background. Spectral lines from the other component of the elemental standard (sulfur, iodine, chlorine) were not observed, but hydrogen, nitrogen, and oxygen spectral lines (because of the ambient air surrounding the sample) were observed in all samples except the graphite (which absorbs some emitted light from the plasma because of its color). The analyzed anthrax spore surrogate powders were *B. atrophaeus* (US Army Dugway Proving Ground, Dugway, Utah), *Bacillus cereus* (ATCC 14603), *Bacillus thuringiensis* (ATCC 51912), and *Bacillus stearothermophilus* (ATCC 12979). All ATCC spores were used as received from ATCC. The *B. atrophaeus* was prepared as an 80:20 mixture of dry spores to fumed silica particles by mass [12]. Spectra were also taken of a stainless steel coupon blank, used as the pellet backing during analysis. Note that DiPel 150 dust has traces (less than 0.065%) of *B. thuringiensis*; however, because the LIBS system would not detect this low concentration of *B. thuringiensis*, we include DiPel 150 dust as a confusant sample rather than a spore surrogate powder.

3. STATISTICAL ANALYSIS

The proposed methodology provides a means for classification of LIBS data as *Bacillus* spores or not. The idea after preprocessing the data was to reduce the dimension of the data by using known structural information about *B. anthracis* spores and linear models and build an SVM classification model by using the dimension-reduced data. The statistical analysis begins with preprocessing of the LIBS spectra produced for the substances described in Section 2.2. The first step in preprocessing was to remove outlying spectra followed by logarithmic and wavelet transformations. Wavelet transformations are described in the succeeding paragraphs along with an introduction to SVMs, which were used for the classification model. A description of the proposed method follows these introductions, and the section is concluded with an analysis and discussion of the performance of the model. The statistical analysis was completed using the statistical software R [13] and packages kernlab [14] and e1071 [15]. A comparison of these packages can be found in Karazoglou *et al* [16].

3.1. Wavelets

The wavelet transformation is a methodology useful in modeling data characterized by sharp peaks, or spikes, and other local features by using a set of wavelet basis functions [17]. There are several wavelet families, and among the most popular are the Daubechies wavelets, which form an orthonormal basis in the space of square-integrable functions. In modeling, the mother wavelet ψ is dilated and translated, that is, stretched, squeezed, and shifted, to represent some function f where, in this analysis, f is an LIBS spectrum. The general form of the representation of f by using a wavelet basis is

$$f(x) = c_{00}\varphi_0(x) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}(x) \qquad (1)$$

where φ_0 is referred to as the father wavelet, or the scaling function, with coefficients c_{00}, and $\psi_{jk}(x)=2^{0.5j}\psi(2^jx-k)$ with integers j and k indexing the dilations and translations, respectively, of the mother wavelet. The dilation can be thought of as the window width of the wavelet. The d_{jk} is the wavelet coefficient at level j and location k, defined as

$$d_{jk} = \int f(x)\psi_{jk}(x)dx \qquad (2)$$

In practice, f is observed at discretized points so Equation (2) is replaced by an approximation, and the range of the indices of the summations in Equation (1) are truncated on the basis of the values computed by the available data. Specifically, the sum indexed by j is truncated to $\log_2(n)-1$, where n is the number of observations of the function f, and $k=0,1,\ldots,2^j-1$.

In the proposed methodology, a wavelet transformation was taken of each spectrum, and only the estimated wavelet coefficients, called the discrete wavelet coefficients (DWC), were retained for the classification model. Including a wavelet transformation in the framework provides flexibility when more preprocessing of the data is required, for example, for noisy data, it can be appealing to threshold the DWC. More details about the wavelet transformation used in the proposed methodology can be found in Section 3.3.

3.2. Support vector machines

Support vector machines have many uses in statistics, in particular for classification. An overview of the methods can be found in Shawe-Taylor and Cristianini [18] or Hastie *et al.* [19]. The main idea for SVM classification is to find the hyperplane that best separates the data into two classes by maximizing the margin between the closest points in each class. These closest data points are known as the *support vectors*.

Consider a data set $\{\mathbf{x}_i,\mathbf{y}_i\}$ where $i=1,\ldots n$, $\mathbf{y}_i=\{-1,\ 1\}$, and $\mathbf{x}_i \in \Re^d$ where d is the dimension of the data set. For simplicity, suppose $d=2$. Then, there is a two-dimensional vector of n data points, and each point is assigned into classes -1 or 1. The goal is to find the hyperplane, which for $d=2$ is a line that best separates the two classes. Specifically, one needs to find variable b and vector $\mathbf{w}$ that defines the hyperplane

$$(\mathbf{x}_i\cdot\mathbf{w}+b)y_i \geq 1 \text{ for } i=1,\ldots,n, \qquad (3)$$

where the hyperplane is such that it is as far as possible from the closest data points of each class. That is, the goal is to maximize the margin between the two classes. The distance between the hyperplane of Equation (3) and the support vectors of each class is equal to $||\mathbf{w}||^{-1}$, where $\mathbf{w}$ is orthogonal to the hyperplane. Hence, the distance between the two classes (i.e., the margin) is equal to $2||\mathbf{w}||^{-1}$. To maximize the margin, one can minimize $||\mathbf{w}||$ subject to the constraints defined in Equation (3). To make this computationally easier, $0.5||\mathbf{w}||^2$ is minimized with the same constraints. The variable b is the offset from the origin of the hyperplane.

Because most data will not be perfectly linearly separable, Equation (3) is modified to allow for misclassifications (data points on the wrong side of the separating margin) as follows:

$$(\mathbf{x}_i\cdot\mathbf{w}+b)y_i \geq 1-\xi_i,\ \xi_i \geq 0 \text{ for } i=1,\ldots,n, \qquad (4)$$

This now allows for some values to be misclassified, and the objective function is modified to

$$\min\left(0.5||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i\right) \qquad (5)$$

subject to the constraints of Equation (4). The parameter C is chosen to reflect the degree to which misclassifications are penalized and is referred to as the cost parameter, and the variables ξ_i are measures of the degree of the misclassification of $\mathbf{x}_i$. A nonzero ξ_i suggests that the datum $\mathbf{x}_i$ is on the "wrong" side of the hyperplane resulting in the application of a penalty.

The SVM classification model is completely defined by the vector $\mathbf{w}$ and b where a new point $\mathbf{x}^*$ is classified by the sign of $\mathbf{w}\cdot\mathbf{x}^*+b$. If there is asymmetry in the number of observations falling into each class, there are procedures for reducing the impact of the imbalance [20].

The derivation assumes that the data are linearly separable, but this is not always the case; for example, one of the classes could be circumscribed by the other class. However, the data can be transformed using a kernel function (determined by the data) onto feature space to improve linear separability [18].

For the proposed methodology, $\mathbf{x}$ are a measure of the presence of the eight elements typically detected in anthrax surrogates (described in detail hereafter), and the $\mathbf{y}$ indicate if the substance is a spore powder or not.

3.3. Data preprocessing

The data used in the analysis were the spectra (intensity at different wavelengths) collected from the LIBS system described in Section 2.1 and the substances named in Section 2.2. The spectra used to build the model had 13 701 data points (one for each wavelength at which the system recorded a measurement). A sample spectrum of baking powder and *B. stearothermophilus* (ATCC 12979) are displayed in Figure 1a and b, respectively.

The overall idea of the model is to first regularize the data via wavelets, then reduce the dimension of the samples by using a linear model, and finally build a classification model that will categorize an unknown substance as a *Bacillus* spore, or not,

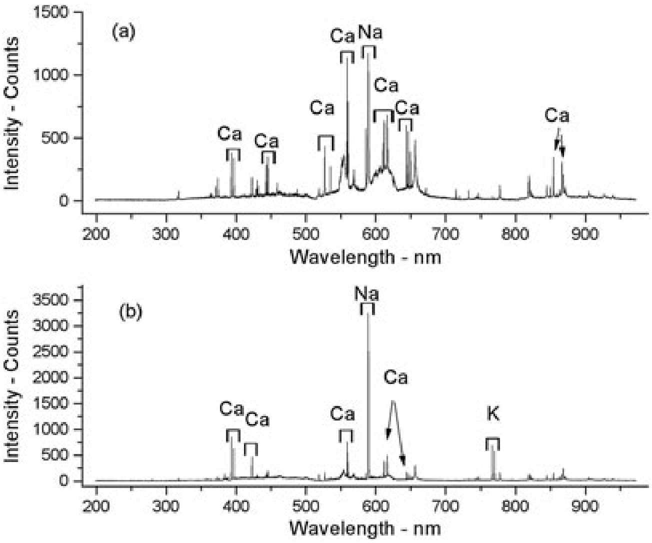

Figure 1. A spectrum generated by a laser-induced breakdown spectroscopy system of samples of (a) baking powder and (b) *B. stearothermophilus* (ATCC 12979).

using SVMs. This section addresses all the steps preceding the SVM classification model.

In addition to the LIBS spectra for *B. anthracis* surrogate spore powders and other innocuous powders described in Section 2.2, LIBS spectra for the eight elements typically detected in *B. anthracis* spores—sodium, potassium, magnesium, manganese, silicon, carbon, calcium, and iron—were considered. Each element has a characteristic spectrum with peaks at known wavelengths, and the characteristic elemental spectra were used in the analysis to aid in dimension reduction of the other LIBS spectra and provided a way to focus on the wavelengths of the spectra where peaks are expected when *B. anthracis* surrogate spores are present.

The first step in the analysis was to remove obvious outliers in the data. To locate suspected outliers, each set of spectra for each substance was analyzed in the following manner. For each substance (e.g., Sweet 'n Low spectra), mean and median spectra were defined. The sum of the absolute difference between each sample and the mean and median spectra were calculated along with noting the sample's minimum and maximum intensities. Four plots were generated to compare these four values between the samples. If a sample's point for any of the plots did not follow a pattern similar to the majority of the other points, its spectrum was visually compared with the mean and median spectra. In the few cases where the spectra were clear outliers (e.g., no peaks were appearing at any wavelength or for large portions of wavelength ranges), the associated samples were removed from the analysis (less than 0.6% of the sample spectra were removed as outliers).

After removing outlying spectra, a logarithmic transformation was taken of all the remaining data. There are a few points to note about the sample spectra displayed in Figure 1a and b. First, the spectra have slightly irregular patterns tracing their bases (the irregularities are more pronounced after the logarithmic transformation); second, the spectra have a number of sharp peaks; and lastly, the spectrum in Figure 1a has peaks in locations and at heights different from the spectrum in Figure 1b. For these reasons, along with potential for thresholding and other forms of regularization, a wavelet transformation was employed rather than another functional representation of the data. Although a wavelet transformation of the LIBS spectra was not required for the use of the proposed classification model, the irregular base pattern was removed by only retaining the DWCs, that is, the estimated d_{jk} from Equation (1), and discarding the information related to the scaling function while preserving information about the peaks. The type of wavelet filter selected for the analysis was the Daubechies 4 filter, that is, with two vanishing moments. Symmetric boundary conditions were selected because it was reasonable to assume the unobserved signal to the right of the domain would be better represented as a continuation of the right part of the spectrum (rather than the left part of the spectrum), and vice versa for the region to the left of the domain. Only 9 of the 13 possible levels, that is, dilations of the mother wavelet, of coefficients were needed to capture the required detail of the spectra. The DWCs for each spectrum were vectorized and replaced the raw spectra as the data.

Next, a linear regression model was used to determine the combination pattern of the characteristic elemental spectra for the *Bacillus* spores LIBS spectra and the other substances. The components of the linear model are described in Equation (6).

For $e = 1, \ldots, 9$,

$$E^e = \text{ vector of DWCs for chemical element } e \qquad (6)$$

where the ninth element included is boron because of its role in obtaining the elemental LIBS spectra. Including boron in the determination of the combination pattern, but not including it in the classification model, kept the presence of the boron in the elemental spectra from influencing classification. In addition to the E^e's, indicators for the seven CCDs (see Section 2.1), defined by wavelength ranges, were carried through the wavelet transformation. The seven indicator vectors initially contained 1 for wavelengths within the range of the corresponding CCD and 0 everywhere else. The wavelength ranges are disjoint between the seven CCDs, and therefore for every wavelength, there is only one column with the corresponding entry 1. These indicators were then subjected to the same wavelet transformation as the sample spectra, vectorized as previously, and then defined for measuring device $i = 1, \ldots, 7$ as

$$I^i = \text{ vector of DWCs for measuring device indicator } i \qquad (7)$$

The I^i's account for any overall inconsistencies between the wavelength ranges of each measuring device. Note that an LIBS spectrum was not used to produce the indicators of Equation (7), but only known information (i.e., the wavelength ranges of the seven CCDs) about the LIBS device.

Every sample spectrum was transformed onto the wavelet domain by using the same wavelet transformation with the DWCs retained. Each collection of DWCs was vectorized and defined as $Y_{sj} = \text{ vector of DWCs},$ for substance $s = 1, \ldots, S$ sample $j = 1, \ldots, J_s$ (each substance had 20–25 sample spectra).

The combination pattern of the elemental spectra for each substance is called its loadings. To obtain the loadings for every sample of each substance, the following linear model was fit:

$$Y_{sj} = \mathbf{X}\beta_{sj} + \varepsilon \qquad (8)$$

where **X** is a matrix composed of a column of ones, followed by I^i, $i = 1, \ldots 6$ defined in Equation (7) (only six of the seven CCDs were used because of the multicollinearity that would result, and thus nonidentifiability of parameters; this is a mathematical issue and does not impact the results), and E_e, $e = 1, \ldots, 9$ as defined in Equation (6). Furthermore, β_{sj} is the unknown 16-dimensional parameter vector for sample j of substance s, and ε is the unknown error in the model (note that normality of the data is not assumed). The loading vector is defined as part of the least-squares estimates, $\widehat{\beta}_{sj}$, where $\widehat{\beta}_{sj} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y_{sj}$. The loading vector is denoted as $\breve{\beta}_{sj}$ for sample j of substance s, where this only includes the values from $\widehat{\beta}_{sj}$, corresponding to the eight elements, that is, excluding boron.

In summary, each sample of each substance had an eight-dimensional loading vector associated with it. The dimension of data was reduced from 13 701 to 8 by focusing on the locations of the spectra where the signature elements of *Bacillus* spores were found to have peaks. The loadings for the *Bacillus* spores (the particular combination of the elemental spectra for each sample) are displayed in Figure 2a–d. The behavior of the loadings for each of the surrogates appears to be similar, suggesting that these elemental spectra do combine in a similar way to form the *Bacillus* spore spectra, whereas the loadings for two examples of innocuous substances, baking soda, and baking powder (see Figure 2e and f) clearly follow a different pattern. These eight-dimensional loading vectors were used in the classification model described in the next section.

3.4. Classification model

Support vector machines were used to develop a classification model using the loadings associated with the data described in Section 2.2. The two classes were *Bacillus* spores and other nonbiological confusant substances. As noted in Section 3.2, the goal of SVM was to minimize Equation (5) subject to the constraints of Equation (4). In this analysis, $y_i = 1$ for spores, $y_i = -1$ for other substances, and the corresponding vectors $\mathbf{x}_i$ were the eight-dimensional loading vectors defined previously. In addition to our confusant powders, we included loadings for the spectra of the eight elements in the non-*Bacillus* group ($y_i = -1$). Including these elemental loadings helped to guard against an unknown confusant substance being misclassified as *Bacillus* spores simply because it was made up of only one or several of these elements. For SVMs, disproportionate class sizes, if not accounted for, can lead to incorrect classification [21,22]. Because of the imbalance between the number of spore samples and other samples (94 sample spore spectra were used compared with 419 confusant substances), the substances were adjusted to alleviate the disproportion (to diminish the effect of asymmetric class sizes) by using the *class.weights* option in the R packages mentioned in the introduction of Section 3. This option allows the user to mitigate the class size imbalance between the two classes by assigning a higher weight to the class with fewer samples and a lower weight to the class with more samples.

The model was built using 70% of the data (randomly selected for each substance) and then verified using the remaining 30%. Because the data appeared to be linearly separable, the linear kernel was sufficient for this model, that is, a hyperplane, or flat surface, separated the two classes of data, and no additional transformation was needed.

During the construction of the classification model, the cost parameter C was set using a grid search over a specified range of values. The final C was chosen on the basis of 10-fold cross validation error over the grid. The performance based on model and 10-fold cross validation errors was low for all C's presented in the grid. Although all of the values considered for C perform well, the grid-search algorithm used for tuning parameters of SVM selected $C = 2$ as the best value on the basis of 10-fold cross

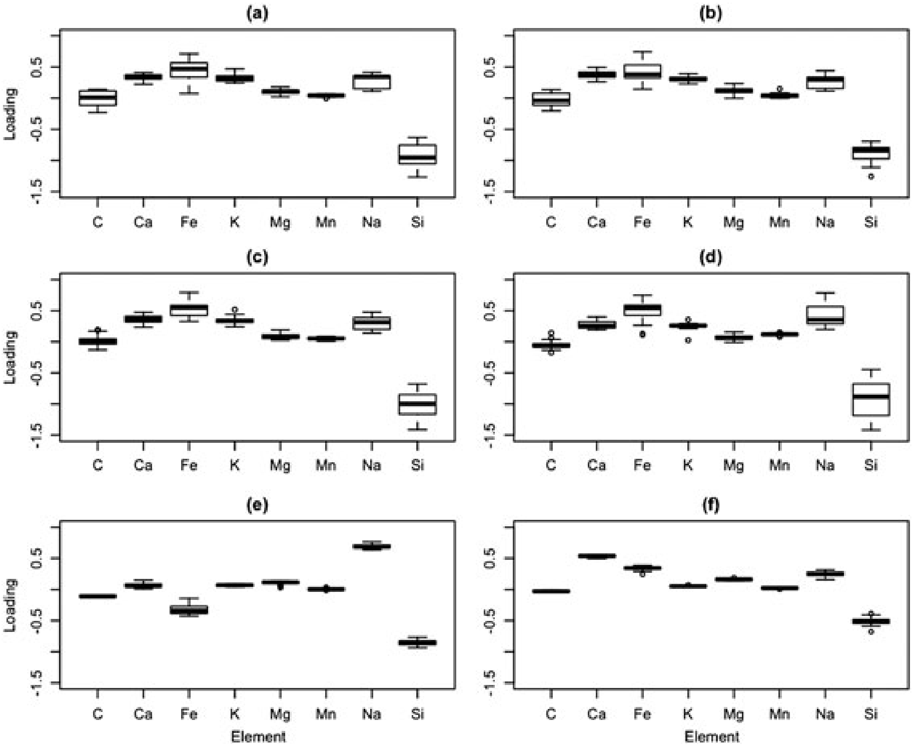

Figure 2. Box plots of the loadings for *Bacillus* spores (a) *Bacillus stearothermophilus* (ATCC 12979), (b) *Bacillus thuringiensis* (ATCC 51912), (c) *Bacillus atrophaeus* (ECBC), and (d) *Bacillus cereus* (ATCC 14603), and confusant substances (e) baking soda and (f) baking powder. The horizontal axis displays the eight elements typically detected in *Bacillus* spores, and the values along the vertical axis are the corresponding loading values for each sample summarized as a box plot. There were 21–29 samples used for each plot.

validation. Once the model was built, the remaining 30% of the data was put into the classification model. Misclassifications, or prediction error, occurred if the outcome of the model did not accurately predict the class of the substance, that is, if an anthrax surrogate powder was classified as a confusant powder, or vice versa. Table I displays the results of the prediction error of the fitted model by using the 30% of the data not included in the model selection and parameter tuning.

To further test the model, sample spectra of two different *Bacillus* spores not used to build the model were classified using the model: *B. thuringiensis* (ATCC51912; eight spectra) and *B. stearothermophilus* (ATCC12979; five spectra). All 13 spectra were correctly classified.

To further verify the performance, and because of the limited data available, a "leave-one-out" method was employed. The model was built using spectra from only three of the four *Bacillus* spore samples (i.e., one *Bacillus* spore sample was left out of the model building stage). Seventy per cent of the three remaining *Bacillus* spore samples and confusant powder samples were randomly selected to build the model. The model's predictive power was tested on the *Bacillus* spore sample left out and the remaining 30% of the data. This was repeated for each *Bacillus* spore sample, and the results of this test are displayed in Table II. The prediction error was between 0.0% and 3.4%.

A similar procedure was performed by randomly selecting five nonspore samples to be left out of building the model: four confusant substances (baking soda, Gain laundry detergent, sugar, and Tide laundry detergent) and the blank stainless steel spectra. These five substances were grouped together in the determination of the prediction error. The results are also listed as the last row in Table II. The prediction error was 0.9% and 3.3% for confusant powders and *Bacillus* spores, respectively. The higher prediction error for the *B. atrophaeus* could potentially be attributed to its distinct spore preparation.

Table I. Prediction error using cost parameter $C=2$. The prediction error is determined by the frequency of misclassifications of the 30% of the data values not used in model selection. Note that less than 0.6% of the sample spectra was removed as outliers

Confusant powders	*Bacillus* spore powders
0.6%	3.3%

Table II. Prediction error of model built with all samples except the samples listed under "testing substance left out". *Bacillus stearothermophilus* (ATCC 12979), *Bacillus thuringiensis* (ATCC 51912), *Bacillus cereus* (ATCC 14603), and *Bacillus atrophaeus* (ECBC) are spores. The category "confusant" includes stainless steel, baking soda, Gain laundry detergent, sugar, and Tide laundry detergent. Note that less than 0.6% of the sample spectra was removed as outliers

Testing substance left out	Confusant powders (%)	*Bacillus* spore powders (%)
B. stearothermophilus (ATCC 12979)	0.0	0.0
B. thuringiensis (ATCC 51912)	1.3	0.0
B. cereus (ATCC 14603)	0.6	0.0
B. atrophaeus (ECBC)	0.0	3.4
Confusant	0.9	3.3

4. CONCLUSION

The proposed methodology provides a way to classify suspect powders, like *Bacillus* spores, from other substances by using LIBS spectra generated using the same LIBS system. Several statistical techniques were brought together to produce the classification model. A wavelet transformation was used to reduce irregularities in the LIBS spectra and focus the classification analysis on the peaks. Regressing the DWCs of the spores and other substances on the DWC of the eight elements helped to both reduce the dimension of the data and focus on the regions of a spectrum where peaks were expected if spores are present in the substance. Finally, the output loading vectors were then used to build the classification model by using the SVM approach.

The overall classification model performed well for the data and setting presented and could be used in other cases where one of the classes has some known elemental structure. The methodology was developed using pure substances, but an interesting and important extension would be to consider spore specimens mixed with various confusant substances as well as various spore preparations for the same spore species. More complex classification goals would require an increased number of samples of *Bacillus* spores and investigation of a nonlinear relationship between spores' spectra and the elemental spectra. To generalize this method to other LIBS systems, random effects could be incorporated to account for the expected variation differences between LIBS systems. After obtaining samples from several LIBS devices (of the same design), a random effect would be incorporated into the model Equation (8) to capture the population-level variation of the LIBS devices to generalize the methodology to all LIBS devices of this type. The usefulness of incorporating random effects would be assessed by comparing the results of the method outlined in this paper with the results of the same method proposed previously, except using a mixed linear model to determine the loading vectors rather than the linear model used in Equation (8).

Acknowledgements

The research of Jan Hannig and Jessi Cisewski was supported in part by the National Science Foundation under Grant nos. 1007543 and 1016441. The authors wish to thank Professor M. R. Leadbetter (Department of Statistics and Operations Research, University of North Carolina at Chapel Hill) for his contribution during the initial stages of this project. Emily Snyder and Lukas Oudejans wish to thank the assistance of Steve Terll, Matt Clayton, and Craig Williams of ARCADIS US, Inc. in creating pellets and recording of LIBS spectra.

The US Environmental Protection Agency through its Office of Research and Development completed the research described here under ARCADIS G&M contract EP-C-04-023. It has been reviewed by the US Environmental Protection Agency's National Homeland Security Research Center and approved for publication.

Approval does not signify that the contents reflect the views of the agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

REFERENCES

1. Tognoni E, Palleschi V, Corsi M, Cristoforetti G, Omenetto N, Gornushkin I, Smith BW, Winefordner JD. From sample to signal in LIBS: a complex route to quantitative analysis. In Laser-induced Breakdown Spectroscopy (LIBS): Fundamentals and Applications, Miziolek AW, Palleschi V, Schechter I (eds.). Cambridge University Press: Cambridge, UK, 2006.
2. Rat VN, Thakur SN. Physics of plasma in LIBS. In Laser-induced Breakdown Spectroscopy, Singh JP. Thakur SN (eds.). Elsevier: New York, NY, 2007.
3. Wisbrun R, Schechter I, Niessner R, Schroder H, Kompa KL. Detector for trace elemental analysis of solid environmental-samples by laser-plasma spectroscopy. *Anal. Chem.* 1994; **66**(18): 2964–2975.
4. Tognoni E, Cristoforetti G, Legnaloli S, Palleschi V, Salvetti A, Mueller M, Panne U, Gomushkin I. A numerical study of expected accuracy and precision in calibration-free laser-induced breakdown spectroscopy in the assumption of ideal analytical plasma. *Spectrochim Acta B* 2007; **62**(12): 1287–1302.
5. Lednev V, Pershin SM, Bunkin AF. Laser beam profile influence on LIBS analytical capabilities: single vs. multimode beam. *J Anal Atom Spectrom* 2010; **25**(11): 1745–1757.
6. Gottfried JL, De Lucia FC, Munson CA, Miziolek AW. Standoff detection of chemical and biological threats using laser-induced breakdown spectroscopy. *Appl. Spectrosc.* 2008; **62**(4): 353–363.
7. Gibb-Snyder E, Gullett B, Ryan S, Oudejans L, Touati A. Development of size-selective sampling of Bacillus anthracis surrogate spores from simulated building air intake mixtures for analysis via laser-induced breakdown spectroscopy. *Appl. Spectrosc.* 2006; **60**(8): 860–870.
8. Munson CA, De Lucia FC, Piehler T, McNesby KL, Miziolek AW. Investigation of statistics strategies for improving the discriminating power of laser-induced breakdown spectroscopy for chemical and biological warfare agent simulants. *Spectrochim Acta B* 2005; **60**(7–8): 1217–1224.
9. Rehse SJ, Jeyasingham N, Diedrich J, Palchaudhuri S. A membrane basis for bacterial identification and discrimination using laser-induced breakdown spectroscopy. *J. Appl. Phys.* 2009; **105**(10): 102034–102047.
10. Singh JP, Yueh FY, Zheng HB, Burgess S. Preliminary evaluation of laser-induced breakdown spectroscopy for tissue classification. *Spectrochim Acta B* 2009; **64**(10): 1059–1067.
11. Koujelev A, Sabsabi M, Motto-Ros V, Laville S, Lui SL. Laser-induced breakdown spectroscopy with artificial neural network processing for material identification. *Planet Space Sci* 2010; **58**(4): 682–690.
12. Brown GS, Betty RG, Brockmann JE, Lucero DA, Souza CA, Walsh KS, Boucher RM, Tezak M, Wilson MC, Rudolph T. Evaluation of a wipe surface sample method for collection of Bacillus spores from non-porous surfaces. *Appl. Environ. Microbiol.* 2007; **73**(3): 706–10.
13. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2011.
14. Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab - an S4 package for kernel methods. *R. J. Stat. Softw.* 2004; **11**(9): 1–20.
15. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. E1071 – misc functions of the Department of Statistics. R package version 1.6. http://www.r-project.org/ (accessed September 20, 2011).
16. Karatzoglou A, Meyer D, Hornik K. Support vector machines in R. *J. Stat. Softw.* 2006; **15**(9): 1–28.
17. Nason CP. Wavelet Methods in Statistics with R. Springer Publishing Company: New York, NY, 2008.
18. Shawe-Taylor JCN. An Introduction to Support Vector Machine and Other Kernel-based Learning Methods. Cambridge University Press: New York, NY, 2000.
19. Hastie TTR, Friedman J. 2001. The Element of Statistical Learning: Data Mining, Inference, and Prediction. Springer Publishing: New York, NY, 2001.
20. Chawla NV, Japkowicz N, Ko lcz A. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations* 2004; **6**(1): 1–6.
21. Shin H, Cho S. How to deal with large dataset and binary output in SVM based response model. *Proceedings of the Korean Data Mining Society Conference* 2003, 93–107.
22. Wu G, Chang EY. KBA: kernel boundary alignment considering imbalanced data distribution knowledge and data engineering. *IEEE Transactions on* 2005; **17**(6): 786–795.