The Annals of Applied Statistics
2019, Vol. 13, No. 2, 1068–1102
https://doi.org/10.1214/18-AOAS1211

LEARNING ALGORITHMS TO EVALUATE FORENSIC GLASS EVIDENCE

BY SOYOUNG PARK AND ALICIA CARRIQUIRY[1]

Iowa State University

Glass fragments are often compared in the course of a forensic evaluation using their chemical composition determined with technologies such as LA-ICP-MS. At present forensic scientists advocate the use of two comparison criteria based on univariate intervals around all mean elemental concentrations for fragments originating from a known piece of broken glass. The main drawback of this approach is that it does not consider the dependencies between concentrations. Further, when the elemental concentrations are more variable within panes, it becomes harder to reject the null hypothesis of no difference between fragments. In the legal context higher variance would tend to incriminate the defendant because the intervals would tend to be wider. We demonstrate that a score-based approach to assess the probative value of evidence in glass comparisons outperforms the two standard interval methods and other methods proposed in the literature, at least in terms of minimizing classification error in the glass fragment sources we analyzed. We use machine learning algorithms to construct a *similarity score* between pairs of glass fragments. The learning algorithms exploit the dependencies among elemental concentrations and result in an empirical class probability; so, we can report the *degree of similarity* between two fragments. Our group is in the process of assembling the first glass composition database with enough information within and between glass samples to permit computing well-conditioned estimates of high-dimensional covariance matrices. These data will be available to anyone who wishes to carry out research in this area.

1. Introduction. In the United States' criminal justice system, jurors are typically tasked with deciding between the prosecutor's and the defense's propositions using summaries of the evidence presented by forensic scientists or other experts. In this paper, we focus on glass evidence that may arise when a glass object is broken during the commission of a crime. Small fragments from the broken object can transfer to clothing, hair or shoes of the perpetrator of the crime, or onto a victim in the crime scene. The question of interest then becomes whether the glass found on a suspect may have come from the broken glass object at the crime scene. From the juror's point of view, two important questions are whether the fragments

Received June 2018; revised September 2018.

[1]Supported in part by an endowment from the Iowa State University Foundation associated with the President's Chair in Statistics.

Key words and phrases. Multivariate measurements, random forest, out-of-bag errors, score likelihood ratio, forensic glass comparisons.

on the suspect and at the crime scene are indistinguishable (in some sense) and, if so, whether the observed degree of similarity is typical only when fragments have a common source. Fragments found at the crime scene are called control or known (K), and glass fragments recovered from the suspect are the questioned samples (Q). In the forensics literature, this is known as the *specific source* question: did the fragment on the suspect originate from the glass object broken in the commission of the crime? A related, but not identical question, is the *same source* question: could two fragments, one recovered from a victim and another from a suspect, have the same, but unknown source? Whether attempting to answer one or the other question, the statistic used to quantify the differences between fragments is the same. However, the approach for evaluating the *probative value of the evidence* is different depending on the question (e.g., Ommen and Saunders (2018)), and we revisit this issue later in the paper.

Glass objects broken during the commission of a crime include containers (bottles, jars, vials), architectural glass (windows, doors), car windshields and many more. In this paper, we focus on architectural float glass used in windows and doors. Glass is made by melting together sand, soda ash, dolomite, limestone and sodium sulfate at temperatures in excess of 1,500 C. Manufacturers also add *cullet* (recycled broken glass) to the mixture. In the 1950s Sir Alastair Pilkington invented the process to produce *float glass*; this process is used to this day. After raw materials are mixed in a batch plant, they are fed into a furnace where the batch is melted and mixed. The molten mixture is then extruded in the form of a wide ribbon onto a bath of molten tin that provides a flat, smooth surface for the glass. As the glass travels on the molten tin, it cools down gradually and, depending on settings, acquires the desired thickness. Once the glass has cooled down to about 1,000 C, it enters an annealing chamber, where controlled cooling is faster. The last steps in the manufacturing process consist in cutting the glass ribbon to specs, and the panes are then packaged for transportation. Figure 1 shows a schematic of the float glass manufacturing process.

A glass ribbon can have a thickness between 0.4 and 25 mm, and the length it travels between furnace and the end of the line is approximately 0.5 km, about the length of five American football fields. There are 370 float glass manufacturers worldwide who jointly produce almost a million tons of float glass per year.

Forensic scientists describe physical, optical and chemical properties of glass. Except in cases where the fragment on the suspect is large, it is often difficult to compare glass on the basis of physical properties. The refractive index (RI) of glass describes how light propagates through that glass fragment. In the past the RI varied between glass samples and was used as a discriminating feature (Curran (2003), Garvin and Koons (2011)), but improvements in the manufacturing processes have resulted in less variability in RI across glass samples (Koons and Buscaglia (2002)).

Today, it is widely agreed that the concentration of minor and trace elements in glass provides a more precise means to compare glass fragments. There are

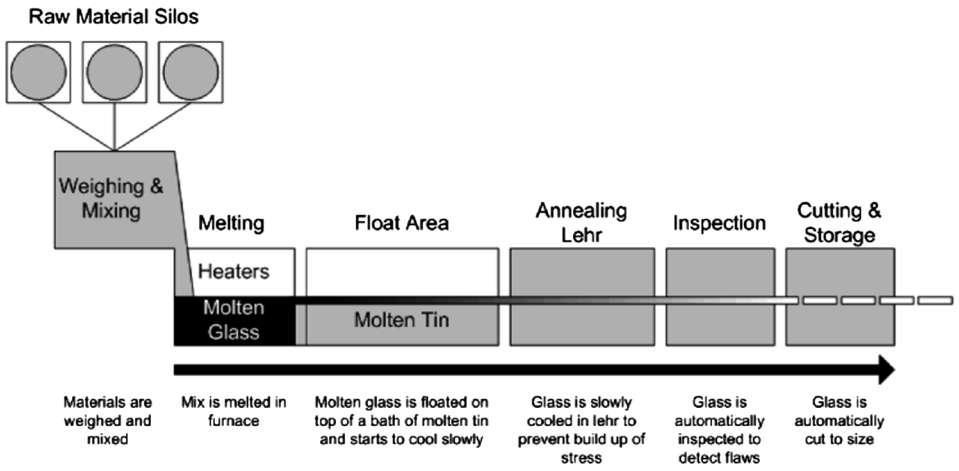

FIG. 1. *Production line to manufacture float glass* (*Tangram-Technology* (2004)).

several technologies that can be used to measure the concentration of elements in glass. For our analyses we used inductively coupled mass spectrometry with a laser add-on (LA-ICP-MS; e.g., Houk (1990), Curran et al. (1997a), Aeschliman et al. (2003), Trejos et al. (2013a), Weis et al. (2011), Dorn et al. (2015)). Briefly, ICP-MS works as follows: an inductively coupled plasma source ionizes the elements in the glass sample. The positively charged ions are then separated and routed to a mass spectrometer (MS) that can identify each ion by its atomic mass-to-charge ratio. Once the ions have been separated, they are detected (or counted) by a suitable detector that can estimate the concentration of each particular element in the sample given the number of the corresponding ions detected. The estimated concentrations are given in multiples of parts per million (ppm). Laser ablation consists in irradiating the surface of the glass sample with a high-pulse laser. When the laser beam hits the glass sample, it produces a plume (aerosol) of atoms that are then presented to the plasma for ionization. In forensic applications the number of elements used for characterizing a glass sample is typically 18 (Weis et al. (2011)).

The main objective of this work is to develop and evaluate a *data-driven score* that quantifies the similarity between two glass fragments using the concentrations of elements in the glass as discriminating features. To construct the score, we rely on supervised machine learning algorithms including random forests (RFs, Kam (1995), Breiman (2001)) and Bayesian Additive Regression Trees (BART, Chipman, George and McCulloch (2010)). The score itself, or a score-based likelihood ratio (SLR; Davis et al. (2012)), can then be used by the trier of fact (juror) to determine the probative value of the glass evidence presented in a specific case.

A second objective of this work is to outline a strategy for forensic glass examiners to compare glass fragments in real casework. Most forensic examiners in accredited laboratories follow the standards published by the American Society for Testing and Materials (ASTM). Glass examiners in particular rely on two

standards, ASTM-E2330-12 (2012) and ASTM-E2927-16 (2016). These standards provide guidance for sample preparation, analysis and interpretation. Both of these standards describe an approach to compare glass fragments using their elemental concentrations and to determine whether the fragments are chemically *indistinguishable*. We show later in this paper that the statistical methods presented in the ASTM standards for comparison of the elemental composition of glass do not perform well in terms of sensitivity and specificity when implemented on data other than the data that were used to develop those methods. In Section 2 we review those statistical approaches plus others that have been proposed in the literature and discuss their limitations. Most of the methods that have been proposed rely on a standard hypothesis testing set-up, where the null hypothesis is that there are no differences between mean concentrations in the fragments. By starting from a null hypothesis of no differences between the composition of the reference and question fragments, the methods appear to place the burden of proof on the defendant in the sense that the dissimilarity between the fragments has to be large enough to reject the null hypothesis in favor of the alternative.

Even though we focus on the classification performance of the various approaches for comparison of glass fragments, the overall goal is to develop a summary of the evidence to help jurors in their deliberations. A forensic scientist who follows the ASTM guidelines might report that two fragments of glass are chemically indistinguishable but this would be an incomplete summary of the evidence. To decide whether the suspect could have been at the crime scene, jurors would also need to know how rarely one would observe fragments with indistinguishable compositions if they originated from different sources. The data-driven score we propose is an ideal (from the jurors' perspective) summary with two elements—the similarity score and the range of values that the score can take on under the two different propositions. Suppose that the evidence is such that the forensic scientist concludes that it is five times more likely to observe a particular score if the fragment on the suspect originated from the crime scene window than if it had a different origin. The final decision is still the juror's who must determine whether 5 to 1 is high enough odds to place the suspect at the crime scene.

This paper focuses on forensic glass comparisons, but we note that the approach we discuss and that relies on the development of a similarity score is applicable in many other forensic problems. Song (2015) and Hare, Hofmann and Carriquiry (2017) proposed similar approaches to compare cartridge case and bullet striations, respectively. Trejos, Flores and Almirall (2010) used a multivariate score to compare the elemental composition of paper and ink and quantify similarity across and within samples. They focused on assessing the properties of the score in terms of classification accuracy but stopped short of discussing the question of weight of evidence. A comprehensive discussion of the analysis and interpretation of forensic glass evidence is presented in Curran, Champod and Buckleton (2000).

The rest of this paper is organized as follows. In Section 2 we review some of the statistical methods that have been proposed to compare the elemental composition of glass. Section 3 includes a more detailed discussion of the two prevailing

interval-based approaches–the standard $4-\sigma$ (Koons and Buscaglia (2002), Trejos et al. (2013b)) and the modified $4-\sigma$ criteria (Weis et al. (2011)) to compare two glass fragments. The section also includes a brief description of two parametric approaches included in the comparison (Hotelling's T^2 with a shrinkage estimator of the covariance matrix, Campbell and Curran (2009), and Parker's optimal H statistic, Parker and Holford (1968)) and the machine learning algorithms that we implemented in this paper. Section 4 describes the datasets that were used to carry out statistical analyses in this work. We present a brief exploratory analysis of the datasets in Section 5. Results are shown in Section 6 and Section 7. We illustrate the impact of the discriminating power of the classifier on the SLR in Section 8. Finally, we finish with a discussion and recommendations for practitioners in Section 9.

2. A brief history of the statistical analysis of glass. When a glass object is broken during the commission of a crime, glass fragments can transfer to the perpetrator or to others in contact with the crime scene. Forensic glass examiners are often tasked with answering the specific or the common source questions described in Section 1. The small size of questioned fragments on a suspect or on a victim almost always prevents comparison of the samples using physical characteristics such as color or thickness, so, in the past several years forensic examiners have relied on technologies such as LA-ICP-MS to accurately measure the concentration of a large number of elements in glass. Depending on the instrument used to obtain the measurements, the number of elements that can be detected can be as high as 40 (ASTM-E2330-12 (2012)) or as low as eight (Zadora (2009a)). In most applications no more than about 18 elements are used in forensic comparisons (e.g., Weis et al. (2011)).

Whether the question is one of specific or of common source, the forensic examiner must quantify the difference between two fragments and decide whether fragments are similar enough that the possibility of common source cannot be excluded. To aid in this decision, we propose a classification method that has high sensitivity meaning that it correctly detects fragments that have a common source (whether specific or not) and high specificity or a high rate of correctly identifying pairs of fragments that have a different source. In Courts in *Daubert* jurisdictions, other relevant performance criteria to evaluate the classifier might include the *positive predictive value* (PPV) and the *negative predictive value* (NPV) of the classifier. The PPV in the glass context is the proportion of same source pairs of fragments among those classified as such by the algorithm. Similarly, the NPV is the proportion of different source fragments among pairs of fragments classified as having a different origin by the algorithm.

Hickman (1987) and Koons, Fiedler and Rawalt (1988) were among the first to use ICP-MS to discriminate between glass fragments from sheet glass and from glass containers using simple clustering methods, so, ours is not a new problem.

Among forensic practitioners comparison criteria based on range overlap or other interval-based approaches are the tools of choice. In 1991, Koons, Peters and Rebbert (1991) proposed a comparison criterion called *range overlap*, where examiners first obtain elemental concentrations from several known fragments and compute the range of values, element-wise. The same elemental concentrations are then obtained in the Q fragment(s). Suppose that there are p trace elements measured; if all p concentrations in the Q fragment fall within the corresponding range obtained from the K samples, then the two fragments are declared to be chemically indistinguishable. If one or more of the p elemental concentrations in Q fall outside of the corresponding range, K and Q are determined to have a different source. A variation of this approach consists in computing the standard deviation of the measurements for each element in the K samples and then construct a $s-\sigma$ interval around the mean, for some s. As before, the elemental concentrations in Q would then be declared to be chemically indistinguishable from K if all p values fall within the corresponding $s-\sigma$ interval.

Comparison criteria based on univariate intervals or ranges have obvious drawbacks. First, when measurement uncertainty or the variability of the elemental concentrations increases, the width of the intervals increases as well which has the unintuitive effect of making it harder to reject the hypothesis of same source. Second, the fact that intervals are constructed elementwise, ignores the presence of correlations (sometimes very high) among elemental concentrations, and consequently the approach is inefficient. Finally, this interval based approach does not consider the probability of a *coincidental match* defined as the chance that two fragments are indistinguishable even if they have a different source. Nonetheless, the recently revised standard for the analysis and interpretation of elemental concentrations obtained by LA-ICP-MS (ASTM-E2927-16 (2016)) recommends that p univariate $s-\sigma$ intervals (slightly modified as described in Section 3) be used as a comparison criterion (Trejos et al. (2013a, 2013b), Dorn et al. (2015)), for $s=4$. Weis et al. (2011) proposed a modified $s-\sigma$ criterion and again suggested that $s=4$. In the remainder we fix the value of s at 4, as recommended by Trejos et al. (2013a) and by Weis et al. (2011).

In parallel to these developments, statisticians over the years have proposed approaches that account for the multivariate nature of the measurements. Parker (1966) introduced the concept of an index C to assess the similarity between two items, when the measured attributes are uncorrelated normal variates with known standard deviations. Curran et al. (1997b) and Campbell and Curran (2009) proposed the use of the Hotelling T^2 statistic with a shrinkage estimator of the covariance matrix to compare two multivariate mean vectors. By using a shrinkage estimator of the covariance matrix, the Hotelling T^2 test is still valid when the number of features p exceeds the number of observations used for estimation. To avoid the need to rely on strong distributional assumptions, Campbell and Curran (2009) proposes a permutation test to derive the null distribution of the statistic.

Lindley (1977) was the first to move beyond the binary same source/different source framework and to propose the use of likelihood ratios to compute the odds of observing a match between two fragments under the competing hypothesis of same and different source. If H_s and H_d denote the competing propositions of same and different source, respectively, and $f_s(y|\theta_s)$ and $f_d(y|\theta_d)$ are the corresponding multivariate densities of the vector of measurements y where f_s and f_d are the probability model for y under the same source and different source hypotheses, then the likelihood ratio is computed as

$$LR = \frac{f_s(y|\theta_s)}{f_d(y|\theta_d)}. \tag{2.1}$$

High values of the LR support the same source hypothesis. Curran et al. (1997a) revisited the likelihood ratio framework and illustrated its use in a small dataset. Aitken and Lucy (2004) compared interval-based methods to likelihood-ratio based methods using a dataset consisting of one fragment from each of 62 windows, with five replicate measurements obtained on each fragment. They carried out a three-dimensional analysis by considering ratios of elemental concentrations and focusing on those which they believed to be most discriminating. In a simulation study they found that a likelihood ratio with a kernel density estimator of the variability across sources outperformed the other methods at least in terms of minimizing the false match and the false nonmatch rates. Scheer (2006) compared the performance of the likelihood ratio when different methods are used to approximate the density of the measurements under both hypotheses and when varying the number of elements used in the comparison. It is fair to say that none of these methods has been adopted for use in actual casework.

Except for the work by Zadora (2009a, 2009b), there have been no other attempts to use machine learning algorithms to either classify glass fragments into various categories or to compare the elemental composition and the refractive index of glass fragments. Zadora addressed the question of classification of glass fragments into a small number (five or six) of different categories–containers, bulbs, windows, car windows and headlamps. Working with a dataset from Poland that included one glass sample from each of 23 windows, 25 bulbs, 32 car windows, 57 containers and 16 headlamps, he found that a Support Vector Machine (SVM) and a naive Bayesian classifier (NBC) did reasonably well when comparing fragments that had different composition and refractive index but resulted in non-negligible false same class and false different class decisions when fragments had similar characteristics (e.g., when classifying fragments of float glass from architectural and automobile windows). Zadora (2009b) had limited samples of glass of different types on which to train the algorithms. Furthermore, measurements were obtained from only four fragments per sample of glass, and each fragment was replicated three times. Measurements were obtained using scanning electron microscopy (SEM-EDX) which limited the number of elements with detectable

concentrations to just eight. In this light we propose the first, to our knowledge, use of a score-based rule to address the question of source.

Research to develop statistical methods to compare elemental composition of glass that minimize classification error and that are robust in the sense of performing well across a variety of datasets has been limited because of the dearth of adequate data. We describe the data that are available to researchers in Section 4 but note here that there are no datasets in the United States with more than just a few fragments (three or four) from each pane of glass. As a consequence there are no datasets that permit estimating consistently a $p \times p$ covariance matrix when p exceeds two or three. As mentioned above measurements obtained via LA-ICP-MS include values for about 18 elements, so to estimate the within-sample covariance matrix we would need no fewer than 20 fragments on at least on a subsample of the glass panes. A dataset collected by the Bundeskriminalamt (BKA, German Federal Criminal Police) includes one glass pane from which 34 different fragments were measured. To help alleviate this problem, we are constructing a dataset with elemental composition of glass fragments using LA-ICP-MS. The dataset includes measurements of 18 elements on each of 24 fragments per pane (five replicates per fragment) on as many panes as our budget allows. We describe the protocol for the collection of these data in Section 4. In this work, and as described in Section 4, we use two sources of data.

We note that there have been no large, well-designed studies that explore whether the elemental composition of float glass is stable over time, even within a single manufacturer. Koons and Buscaglia (2002) mentions that it is possible to detect differences in the elemental composition of glass within a manufacturer between production runs. The likelihood-based approach proposed by Aitken and Lucy (2004) was criticized because of its reliance on a reference population from which the covariance matrices of elemental compositions across and within glass panes are obtained. We have observed a time trend in the concentration of some elements (see Figure 3) suggesting that some drift in the chemical composition of float glass can be expected. The fact that it is likely that the background population of float glass has a variable elemental composition is one of the motivations for the development of comparison approaches that only rely on glass fragments collected from the crime scene and from the defendant or the victim. With regard to the learning approach, we propose in this paper that it reinforces the notion that algorithms may need to be periodically retrained using updated databases.

3. Methods to compare the elemental concentrations in glass fragments.

3.1. *Interval-based match criteria.* Suppose that the concentration of p elements is measured on J fragments from a sample of glass. Each measurement is replicated L times.

Let y_{ijl} be the concentration of the ith element in the lth measurement of the jth fragment; $i = 1, \ldots, p$, $j = 1, \ldots, J$, $l = 1, \ldots, L$. The ith mean concentration in

the jth fragment is denoted by $\bar{y}_{ij,}$ and the standard deviation of concentrations is denoted by SD_{ij} which reflects the measurement error variability. The relative standard deviation (RSD) is the name given by forensic scientists to the coefficient of variation calculated as the ratio $SD_{ij}/\bar{y}_{ij}$.

As mentioned in Section 2, forensic scientists have proposed and evaluated several interval-based comparison criteria. We focus on the two criteria that have been recommended–the standard $4-\sigma$ criterion and the modified $4-\sigma$ criterion (Weis et al. (2011), ASTM-E2330-12 (2012), Trejos et al. (2013a), Dorn et al. (2015), ASTM-E2927-16 (2016)). Both of these criteria are implemented by carrying out p element-wise comparisons.

Standard $4-\sigma$ *interval criterion*. The method described in Sections 10 of ASTM-E2927-16 (2016) and ASTM-E2330-12 (2012) consists of the following. Suppose that we have two glass samples, Q and K, for K the known or reference sample. Using K and for $J, L \geq 3$, compute the p concentration means $\bar{y}_{Ki}$ and the p standard deviations, SD_{Ki}, $i = 1, \ldots, p$, over the $L \times J$ measurements. Neither standard explains precisely whether the SD_{Ki} is computed using observations or fragment means; here, we interpret SD_{Ki} as the standard deviation of the observations. ASTM-E2927-16 (2016) recommends that a minimum of nine measurements of elemental concentrations be obtained from the K sample (three fragments, three replicate measurements from each) and "as many measurements as are practical" be obtained from the Q sample. ASTM-E2330-12 (2012) mentions "a minimum of three measurements" (see Section 10.1.1) from the K sample but does not specify the number of fragments. A minimum SD (MSD_{Ki}) is fixed to be 3% of the mean for the ith element in the K sample. Note that regardless of the number of fragments obtained from K, the standard deviation used to construct the intervals cannot fall below 3% of the corresponding mean concentration. Further, intervals are constructed using the SD of the measurements and not of the mean of measurements, so increasing the number of fragments from K does not necessarily result in narrower intervals. The ith comparison interval for sample K is then computed as

$$\bar{y}_{Ki} \pm 4 \times \max(SD_{Ki}, MSD_{Ki}). \tag{3.1}$$

For $J = 3$ fragments from K, the 0.975 tail quantile of a t distribution with two degrees of freedom is 4.3. Therefore, the interval in equation (3.1) is reminiscent of the standard two-tailed t interval with type I error fixed at 0.05 that would be used to test the null hypothesis of equal means against the alternative of different means. Next, elemental concentrations in sample Q are compared to the p intervals computed as in equation (3.1), element by element. If all elemental concentrations in sample Q are contained in the corresponding intervals, then the two samples are said to be *chemically indistinguishable*. This decision is equivalent to failing to reject the univariate null hypothesis of equal mean concentrations for all p elements. If one or more elemental concentration in sample Q is outside the corresponding

interval obtained from the $J \times L$ measurements, then the two samples are declared to be *distinguishable*. To quote from ASTM-E2330-12 (2012),

> If the samples are indistinguishable in all of these observed and measured properties, the possibility that they originated from the same source of glass cannot be eliminated.

We can represent this comparison criterion in the form of a *score*, computed as the absolute value of the difference between the two elemental concentrations in samples K and Q. Let $S_{\text{ASTM},i}$ denote the score for sample K and Q computed for the ith element. Then

$$S_{\text{ASTM},i} = \left| \frac{\bar{y}_{Ki} - \bar{y}_{Qi}}{\max(0.03 \times \bar{y}_{Ki}, SD_{Ki})} \right|, \tag{3.2}$$
$$S_{\text{ASTM}} = \max(S_{\text{ASTM},i}), \qquad i = 1, \ldots, p,$$

where $\bar{y}_{Ki}$, $\bar{y}_{Qi}$ are the mean concentration of the ith element in the sample K and Q, respectively, and SD_{Ki} is the standard deviation of the ith element measurements on the sample K. As stated in ASTM-E2927-16 (2016), ASTM-E2330-12 (2012), if any of the p $S_{\text{ASTM},i}$ larger than 4, then two fragments are declared to be *distinguishable*.

Modified $4 - \sigma$ criterion with fixed relative SD (FRSD). Weis et al. (2011) proposed an interval-based criterion called modified $s - \sigma$ criterion. They found that $s = 4$ leads to the best compromise between sensitivity and specificity. Weis et al. (2011) obtained 90 measurements (mean of three replicates each) from the German glass standard DGG 1 (Deutsche Glastechnische Gesellschaft, Germany), and from these, computed a fixed relative standard deviation (FRSD), (expressed as percent of the mean) for each of 18 elements. When a value was below 3%, the FRSD was set to 3%. The values of the FRSD are shown in Weis et al. (2011), Table 7. As in the ASTM standards, analyzing additional fragments from K contributes to more reliable estimation of the mean, but not to shorter comparison intervals, since their width is fixed by the FRSD. Using these FRSD, Weis et al. (2011) propose constructing intervals for each element as shown in equation (3.3):

$$\left(\frac{\bar{y}_{Ki}}{(1 + 4 \times FRSD_i)}, \bar{y}_{Ki} \times (1 + 4 \times FRSD_i) \right), \qquad i = 1, \ldots, p, \tag{3.3}$$

where $\bar{y}_{Ki}$ is the mean concentration of the ith element in sample K. As before, if the mean concentrations of all 18 elements in Q fall within the corresponding interval, then the two samples are declared to be chemically *indistinguishable*. If one or more mean concentrations in Q is not contained in its interval, then the two samples are declared to be nonmatches.

The modified $4-\sigma$ criterion in equation (3.3) can also be transformed into a score as in equation (3.4):

$$S_{\mathrm{BKA},i} = \left| \frac{\exp(|\log \bar{y}_{Ki} - \log \bar{y}_{Qi}|) - 1}{FRSD_i} \right|, \tag{3.4}$$
$$S_{\mathrm{BKA}} = \max(S_{\mathrm{BKA},i}), \qquad i = 1, \ldots, p.$$

3.2. *Parametric approaches.* Campbell and Curran (2009) suggested that a better alternative to the range overlap or the $4-\sigma$ methods was to use a Hotelling T^2 test for the comparison of two or more multivariate mean vectors. To overcome the challenges imposed by the limited information available for estimation of well-conditioned p-dimensional covariance matrices, they recommended a shrinkage estimator of the covariance matrix. The form of the Hotelling T^2 statistic is the usual, but the sample covariance, S, is replaced by a shrunken estimate, $\hat{\Sigma}_s$. The statistic is

$$T^2 = \left(\frac{M_K \times M_Q}{M_K + M_Q} \right) (\bar{\mathbf{y}}_K - \bar{\mathbf{y}}_Q) \hat{\Sigma}_s^{-1} (\bar{\mathbf{y}}_K - \bar{\mathbf{y}}_Q), \tag{3.5}$$

where M_K and M_Q are the number of observations in samples K, Q and $(\bar{\mathbf{y}}_K - \bar{\mathbf{y}}_Q)$ is the difference in mean vectors in samples K, Q.

When the number of measurements M_K is smaller than the number of elements p, the shrinkage estimator of the covariance matrix is more efficient, always positive definite and does not rely on assumptions about the underlying distribution of the measurements. Following Schäfer and Strimmer (2005), Campbell and Curran (2009) estimated the covariance matrix by shrinking (James-Stein shrinkage estimator) the sample covariance matrix S toward a target structured matrix F, so that

$$\hat{\Sigma}_s = \hat{\delta}^* F + (1 - \hat{\delta}^*) S,$$

where $\hat{\delta}^*$ is an optimized shrinkage constant and the target F is the p-dimensional matrix with identical pairwise correlations (Ledoit and Wolf (2003)). We use T^2 as a test statistic (score) to quantify the similarity of float glass fragments.

Parker (1966, 1967) proposed an index, C, to quantify the similarity between two items when the features are uncorrelated and have known standard deviations. Parker (1967), Parker and Holford (1968) discussed the effects of correlation among attributes on the discrimination question and suggested a test statistic H for which they derived a sampling distribution. They showed that H is the optimal test in the sense of Birnbaum (1954) when attributes are correlated normal variates with unknown standard deviations. Because elemental concentrations in glass fragments tend to exhibit dependencies, we used the optimum test statistic H as the additional test statistic (score) to assess the similarity of float glass fragments.

3.3. *Supervised learning approaches*. In practice, forensic scientists do not have much information with which to compare glass from a crime scene and glass recovered from the suspect. This poses a challenge. If the comparison is based exclusively on the data at hand (as is proposed in ASTM-E2927-16 (2016)), the standard deviation *SD* of the p elements in the K sample will be poorly estimated, unless a very large number J of fragments from K is included in the comparison. On the other hand, if the comparison relies on estimates of standard deviation such as the FRSD in Weis et al. (2011), then the forensic practitioner needs to justify that those values are plausible when comparing fragments obtained in the specific case under investigation.

We propose a different approach. Suppose that we have a large number of fragments from samples of float glass from a wide variety of manufacturers in the United States (and perhaps other countries as well) for which we know *ground truth*. That is, we know which fragments in the dataset came from which pane of glass. Just 2000 glass fragments allows for almost 2,000,000 different pairwise comparisons, some of which will be between known mated (KM) and some between known nonmated fragments (KNM). We propose to quantify the similarity between the two fragments in each pairwise comparison via a *data-driven score*. If the distributions of values of the score among KM and among KNM pairs do not overlap or have minimum overlap, then the score can be used to classify a pair of fragments as mated or as nonmated. Figure 2 illustrates the idea.

In practice, a forensic examiner would compute the score for the pairwise comparisons in the case on which she is working. Suppose that, following ASTM-E2927-16 (2016), the examiner obtains $L = 3$ replicate measurements on $J = 3$ fragments from K. As, in current practice, the practitioner would obtain an average elemental concentration to represent K and would then calculate the similarity

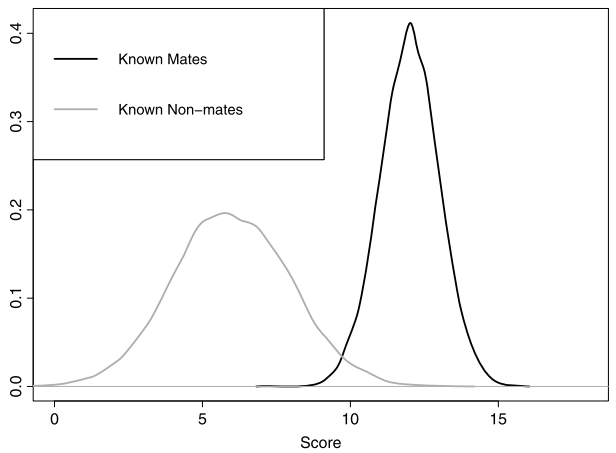

FIG. 2. *Hypothetical distributions of a score among mated and nonmated pairs of fragments.*

between the "average" K fragment and the Q fragment(s). Alternatively, the practitioner might compute three similarity scores, one between each fragment from K and the Q fragment(s), and use the smallest (most favorable to the suspect) to address the question of source. We discuss these alternatives later in this paper. If the sample of fragments used to construct the distributions in Figure 2 is drawn from a relevant population of glass fragment sources, then the examiner can evaluate the *probative value of the evidence* by comparing her score(s) to the reference distributions of scores. Under the hypothetical distributions shown above, a score below 6 or 7 suggests that the fragments are distinguishable, while a score of 12 or above would suggest that the two fragments have very likely originated from the same source of broken glass. Scores between 9 and 11 in this hypothetical example are equally likely under both distributions and therefore do not support any of the two decisions. Clearly, given one or more databases of glass fragments as the one we described above, it would also be possible to proceed as in Aitken and Lucy (2004) or as in Campbell and Curran (2009) and implement a parametric approach to compare compositional means. In this case, it would be necessary to rely on additional assumptions.

In the forensic context, learning algorithms present several advantages: (1) They account for the multivariate nature of elemental compositions. We know that elemental concentrations tend to be highly correlated (see Figure 4) and, potentially, also associated in nonlinear ways. (2) Learning algorithms provide a ranking of the variables that are most discriminating. (3) Most algorithms compute an *empirical class probability* (*score*) or the empirical membership probability that the pair of fragments have the same source or a different source. We can use the estimated empirical probability of common source as the score. High scores are then suggestive of a common source for the fragments, whereas low scores would be associated with pairs of fragments known to originate from different pieces of broken glass. The rate of correctly determining whether two fragments have a common or a different source depends on the threshold we select for the score. ROC curves can help select a threshold that minimizes the false match and the false nonmatch decisions. (4) Once the algorithm has been trained it can be used to compute the similarity between a single pair of fragments or to compare multiple fragments from two panes of glass.

A drawback associated with supervised learning methods is that they depend critically on the data used to train the algorithms and often suffer from overfitting. Overfitting occurs when the classifier mistakenly includes noise in the training data as part of the information contained in the features. In this light, it is important to note that the similarity scores produced by these methods are strongly data dependent. Two approaches to minimize overfitting include resampling or k-fold validation. In addition, setting aside a portion of the data for testing purposes only is also recommended. We implemented both of these measures in our analysis. Also, except in simple cases, algorithms are "black boxes" in that the relationships between the predictors and the response are not explicitly estimated. In the forensic

context, the items that are used to estimate the distribution of scores among non-mated pairs depends on whether the examiner is attempting to answer a specific source or a common source question. This topic is the source of much discussion in the forensic literature (e.g., Hepler et al. (2012), Morrison and Enzinger (2016), Lund and Iyer (2017), Ommen and Saunders (2018)). Because the classification rule obtained from a learning algorithm may change if more pairs of fragments are included in the training dataset, it is important to clearly define the concept of a *relevant background population* and to explore the variation in the chemical composition of glass over time and over manufacturers.

There are several algorithms in the general class of supervised learning methods. Here, we focus on two classifiers: random forests (Breiman (2001)) and Bayesian Additive Regression Trees (BART; Chipman, George and McCulloch (2010)). We compare their performance to that of the two interval based classifiers in ASTM-E2330-12 (2012) and in Weis et al. (2011) and to the two parametric methods proposed by Campbell and Curran (2009) and Parker and Holford (1968). The comparison criterion is the classification error (false positives and false negatives) that results when applying the algorithm to a set of pairs of fragments that were not included in the dataset used to train the algorithms (in the case of RFs and BART).

The random forest and the BART methods produce estimated empirical class probabilities for each sample comparison. The similarity score of the random forest is computed as the average of the empirical class probabilities predicted by each tree from a set of bootstrap samples. The empirical class probability of the BART is the conditional probit (CDF of the standard normal distribution) evaluated at the sum of tree predictions given a specific set of features. We use a threshold for the score equal to 0.5. If the empirical class probability for "same source" exceeds 0.5, then we say that the evidence supports the common source proposition. If not, we conclude that the fragments originate from different pieces of broken glass. Scores that are close to the threshold suggest more uncertainty about the decision than scores that are close to 0 or to 1.

4. Data sources. We use three datasets to train, test and compare algorithms. In all three datasets the concentrations (in ppm) of 18 elements were measured using LA-ICP-MS. Following ASTM guidelines, we used the NIST 1831 standard and two German standards FGS-2 and DGG 1 to calibrate the instruments and monitor drift.

Datasets 1 *and* 2. The first two datasets used in this paper are described in Weis et al. (2011). Dataset 1 includes one fragment from each of 62 different float glass samples and obtained from different countries and manufacturers. Dataset 2 consists of multiple fragments from a single glass pane purchased in Virginia and analyzed by the FBI. A total of 34 different fragments from the Virginia pane were analyzed. In addition, one of the fragments (fragment 104G) was reanalyzed on 11 consecutive days. Therefore, there are 44 18-dimensional measurement vectors

from the Virginia pane. In the remainder we use X to denote the Virginia pane. In both datasets measurements on each fragment were replicated six times.

Dataset 3. These data were collected by Iowa State University, in collaboration with University of Iowa, as part of an effort to construct a dataset to be put in the public domain. At present the dataset includes 31 panes manufactured by Company A and 17 panes manufactured by Company B. The Company A panes are labeled AA, AB, ..., AAR, and the Company B panes are labeled BA, BB, ..., BR. Because the panes from Company A were produced within three weeks (January 3 to 24, 2017) and the panes from Company B were produced within two weeks (December 5 to 16, 2016), we expect them to be more similar to each other than to other panes produced by different manufacturers or by the same manufacturer but at a different time. To understand variability within a ribbon of glass, two glass panes were collected on almost all days–one from the left side and one from the right side of the ribbon. Twenty four fragments were randomly sampled from each glass pane. Five replicate measurements were obtained for 21 of the 24 fragments in each pane; for the remaining three fragments in each pane, we obtained 20 replicate measurements. Dataset 3 contains almost 8000 18-dimensional measurement vectors from over 1150 fragments. Resources including the manuscript, glass measurements and R code to reproduce the analysis can be found in https://github.com/CSAFE-ISU/AOAS-2018-glass-manuscript. We do not post the German Datasets 1 and 2, because they do not belong to us. We also include in the repository a detailed explanation of the analytical methods used to obtain the elemental concentrations.

5. Exploratory data analysis. As a first step we log transformed the elemental concentrations for each element, so that their distributions were less skewed. The 18 elements for which we obtained a concentration were Ca, Na, Mg, Al, K, Fe, Li, Ti, Mn, Rb, Sr, Zr, Ba, La, Ce, Nd, Hf and Pb. Figure 3 shows the distribution of log values of Na, Ti, Zr and Hf by manufacturer. In the figures, the 31 panes obtained from Company A and the 17 panes from Company B are shown in order of production date. The last box corresponds to pane X from Dataset 2. For Ti there is a large difference in concentration between samples from Company A and from Company B. Pane X from Virginia differs from the samples from Company A and Company B with respect to almost all elements. Over the three weeks of sampling, most elemental concentrations in Company A and Company B panes stay approximately constant; the exceptions are Zr and Hf in panes from Company A, where a decreasing trend in time is apparent. We drew the same boxplots using the elemental concentrations of the 62 fragments from different sources in Dataset 1 (figure not shown). As expected, we observed larger variability in elemental concentrations among these fragments.

In contrast to the statement in Curran et al. (1997b), we find that elemental concentrations tend to be highly correlated within pane. In Figure 4 we show the

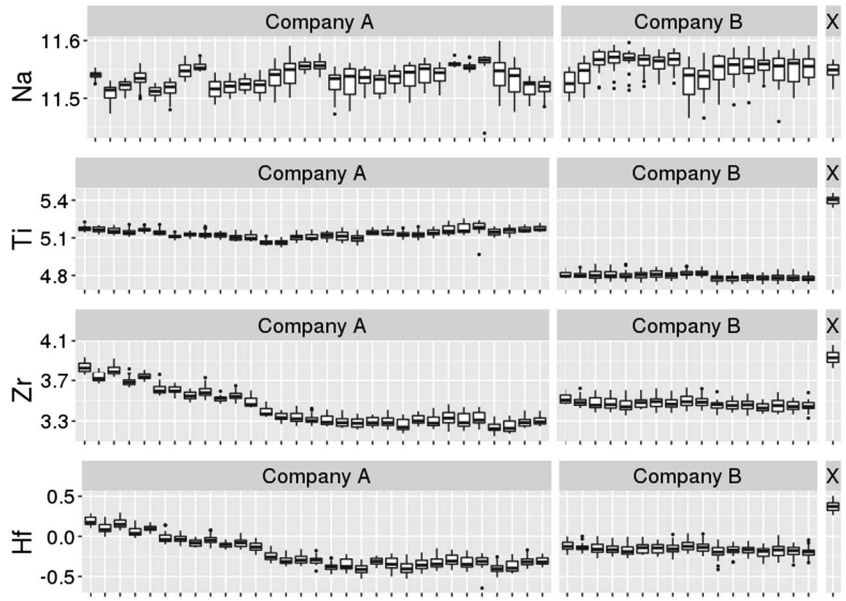

FIG. 3. *Box plot of four elemental compositions in* 49 *panes from Company A and Company B by date of production, and pane X.*

correlations among 18 elements for panes AAR and X—just as illustration. The shaded entries correspond to absolute correlations above 0.5. Note that at least for these two panes, the estimated correlation matrices appear to be different.

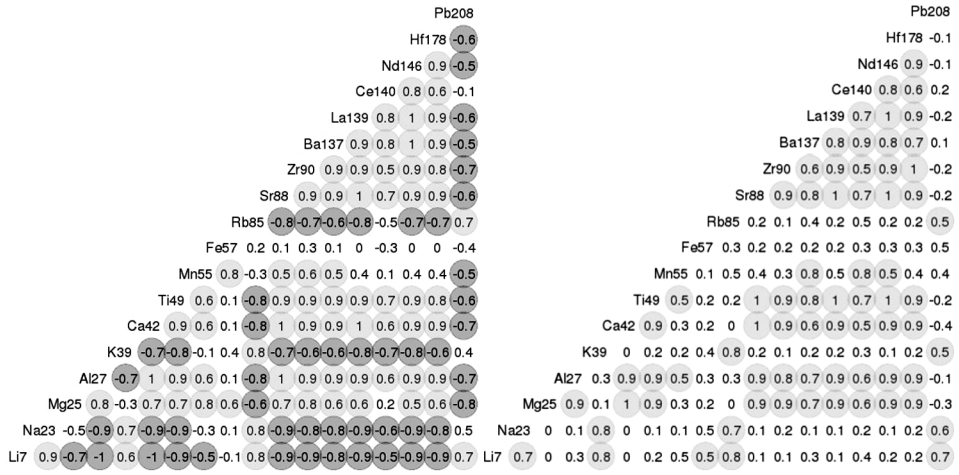

FIG. 4. *Correlations among elemental concentrations in panes AAR and X.*

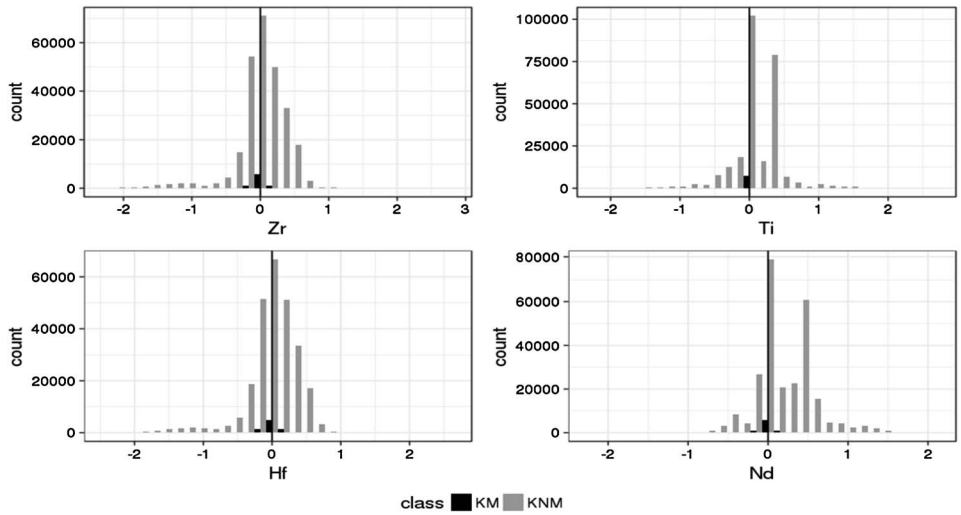

FIG. 5. *Histogram of four feature values among mated* (*black*) *and nonmated* (*gray*) *pairs in the training set.*

Figure 5 shows the distribution of differences for four of the elements (Zr, Ti, Hf, Nd) when pairs of fragments are mated (KM, shown in black) or nonmated (KNM, shown in gray). The distributions shown in black correspond to differences in concentrations among KM pairs of fragments and, as expected, are concentrated around 0 (black vertical line). Differences in concentrations among KNM pairs shown in gray; however, they are more spread out and not centered at 0.

6. Analyses and results. We carried out several separate analyses using different combinations of the datasets to train the supervised learning algorithms and obtained comparable results in all cases. We report on one of them only. We found that, at least in this particular application and with these particular datasets, learning models outperform both of the interval-based criteria that are currently in use, as well as the two parametric approaches we considered. In the remainder of this paper, we analyze Datasets 1, 2 and 3. Let y_{hijl} denote the log of the concentration of the ith element in the jth fragment of the hth pane, for the lth replicate, for $i = 1, \ldots, 18$, $j = 1, \ldots, 24$, $h = 1, \ldots, 49$ (except for pane X for which $j = 1, \ldots, 34$), and $l = 1, \ldots, 5$ (except for pane X, where $l = 1, \ldots, 6$, and for three fragments in each pane for which we obtained 20 replicate measurements). In Dataset 1, $h = 1, \ldots, 62$ and $j = 1$. The average measurements over the five (or six) replicates are denoted $\bar{y}_{hij}$ for each element in each fragment and pane. The *vectors of features* are the 18 differences in concentrations $\bar{y}_{hij} - \bar{y}_{h'ij'}$. When $h = h'$, $j \neq j'$, the comparison is among mated pairs of fragments, and when $h \neq h'$, $j \neq j'$, the comparison involves nonmated pairs. These feature vectors

are computed for all possible pairs of fragments. We center and scale the measurements in the subsets of the data that we use to train the supervised learning algorithms. The models are fitted using a 10-fold cross-validation, as explained below. In each model the tuning parameters were optimized to improve model fitting. To do so, we used the package `caret` for the construction of random forest model and `bartMachine` (Kapelner and Bleich (2013)) for implementing BART in `R`, version 3.3.3. In the tuning step, which was repeated three times, performance of an algorithm with a specific set of parameters is assessed using the area under the ROC curve (AUC), the sensitivity and specificity of the classifier. For the BART model we used the default values in `bartMachine` (Kapelner and Bleich (2013), Chipman, George and McCulloch (2010)) for the hyperparameters for the underlying Bayesian probability model with the number of trees (m) fixed at 100.

To implement the supervised machine learning methods, we divided our dataset into two portions, one that we used for training and validating, and another one that we used to estimate an honest out of bag (OOB) error rate. The training data consisted of 19 panes produced by Company A, and nine panes produced by Company B, for a total of 7705 pairs of fragments known to come from the same pane (KM). For creating the known nonmated pairs we included the 62 float glass samples from Dataset 1, in addition to fragments from 28 panes in Dataset 3. We had a total of 260,573 pairs known to come from different panes (KNM). Several of those panes were manufactured on consecutive days; so, we expected that it would be difficult to correctly allocate fragments to panes when two panes from the same manufacturer were produced one day apart. To carry out the 10-fold validation, the training data were divided into 10 equally sized partitions; nine of the partitions were used to build a random forest that was then tested on the 10th partition. The final forest is obtained as an average over the 10 validation replicates. We do not report the classification error obtained from the internal validation samples.

The internal classification error computed from the 10-fold validation subsamples is likely to underestimate the true classification error because the training and validation subsamples inevitably include fragments from the same pane. Instead, we computed an honest OOB error rate as follows. We set aside a portion of the measurements consisting of 12 Company A and eight Company B panes, plus pane X, for the purpose of testing the performance of the machine learning algorithms. Neither the fragments nor the panes in the test dataset were included in the data set used for training and validating the RF and BART. Of a total of 111 panes of glass in our combined dataset (see Table 1), 90 panes were used to train the RF and the BART, and 21 panes were used to compare the classification performance of the six algorithms we consider here.

Sampling pairs of fragments to train the RFs and BART for classification. The subset of the data we use to train the algorithms is unbalanced, in that there are almost 30 times more pairs of fragments from different panes than from the same

TABLE 1
Panes included in the training/validation and in the testing data subsets

Training and validation set				Test set			
Pane	**Date**	**Pane**	**Date**	**Pane**	**Date**	**Pane**	**Date**
AA	1/3	AAH	1/19	AB	1/3	BF	12/7
AC	1/4	AAI	1/20	AD	1/4	BH	12/9
AE	1/5	AAK	1/21	AF	1/5	BJ	12/9
AG	1/6	AAM	1/22	AH	1/6	BL	12/12
AI	1/7	AAQ	1/24	AJ	1/7	BO	12/15
AK	1/8	BA	12/5	AL	1/8	BR	12/16
AM	1/9	BC	12/7	AX	1/14	X	NA
AO	1/10	BE	12/7	AAB	1/16		
AV	1/13	BG	12/9	AAD	1/17		
AW	1/14	BI	12/9	AAJ	1/20		
AY	1/15	BK	12/12	AAL	1/21		
AAA	1/16	BM	12/14	AAR	1/24		
AAC	1/17	BN	12/15	BB	12/5		
AAF	1/18	BP	12/16	BD	12/7		
	Data 1 : 62 panes						
	28 panes + 62 panes				21 panes		

pane. As a consequence the information contributed by the KNM pairs can dominate the learning process. Several approaches have been proposed in the literature that address the question of imbalance for random forests and other classifiers. Those approaches are based on the idea of *differential weighting*, which has the effect of increasing the cost of misclassification in the minority class, or, on the idea of *sampling*, to even out the number of observations in each of the classes. Sampling can consist in downsampling the majority class, upsampling the minority class or a combination of both (e.g., Random Over-Sampling Examples, ROSE (Lunardon, Menardi and Torelli (2014)); Synthetic Minority Oversampling Technique, or SMOTE (Chawla et al. (2002))). We implemented five different approaches to address the imbalance in our sample; a comparison of the performance of the different approaches is shown in Figure 6. Figure 6 shows the range of values of AUC, sensitivity and specificity from the 30 resampling process (10-fold validation and tuning with three replicates). From the figure it seems that for the internal validation set, SMOTE and downsampling outperform the other approaches.

Figure 7 confirms those results. In the figure we show the ROC curves for the RF fitted to the imbalanced test data and for the five sampling or weighting schemes we considered. The bottom panel of the figure zooms into the upper left-hand corner of the ROC in the top panel and shows that downweighting the majority class or using a combination of down and upweighting (SMOTE) results in the classifiers with best performance in the sense of maximizing the AUC, sensitiv-

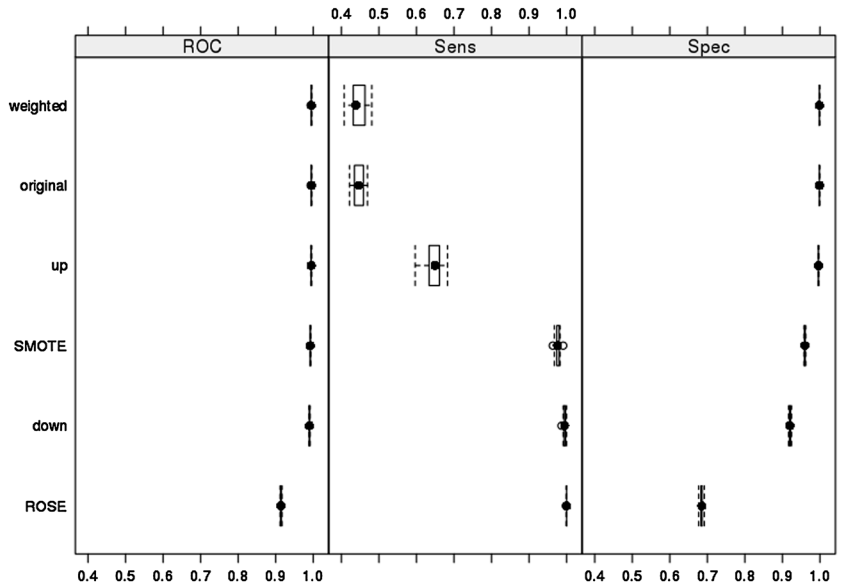

FIG. 6. *ROC (AUC), sensitivity and specificity of RF with optimized parameters in the training set by sampling technique.*

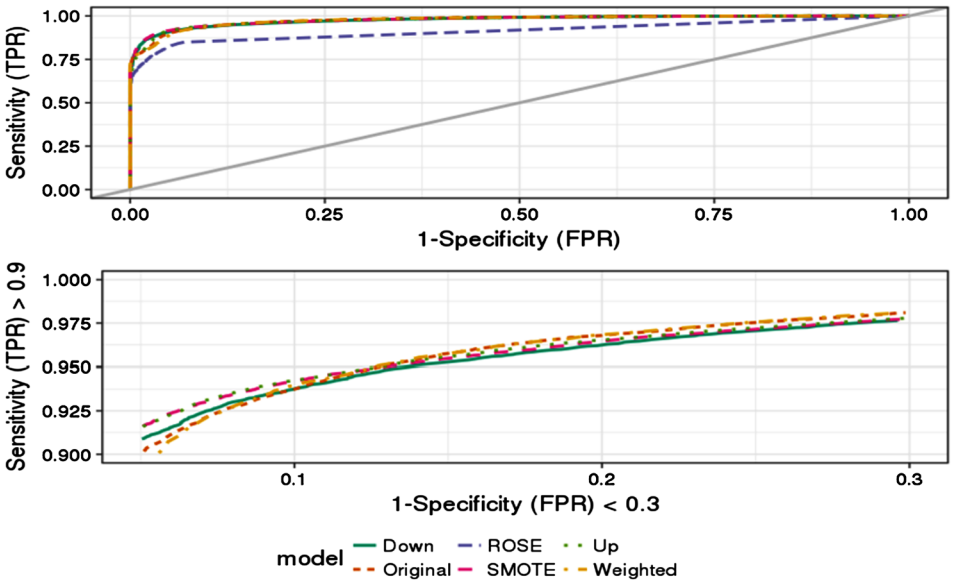

FIG. 7. *ROC curves for random forests by weighting and sampling techniques.*

TABLE 2
TPR at fixed FPR at 5%, 10% and 15% for random forests by weighting and sampling techniques

Sampling	TPR (5%-FPR)	TPR (10%-FPR)	TPR (15%-FPR)
SMOTE	0.863	0.963	0.984
Down	0.853	0.958	0.983
UP	0.868	0.958	0.974
Weighted	0.869	0.943	0.963
Original	0.868	0.940	0.959

ity and specificity. The estimated AUC for downsampling and SMOTE are 0.977 and 0.978, respectively, with approximate 95% confidence intervals for the true AUC equal to $(0.938, 1.000)$ for downsampling and $(0.941, 1.000)$ for SMOTE (DeLong, DeLong and Clarke-Pearson (1988)), suggesting that there are no significant differences in AUC between the two approaches. Here, we used a conservative estimate of the variance of AUC where instead of the number of pairs of fragments in the comparisons we used the number of independent panels of glass in the denominator. The rest of the sampling methods resulted in estimates of AUC that were significantly lower. Table 2 shows the TPR values when FPR is fixed at 5%, 10% and 15%. The random forest trained with SMOTE or downsampling outperform the alternatives, confirming the findings discussed above. In the remainder we strike a compromise between performance and computational efficiency and use downsampling of the majority class to ameliorate the effect of imbalance.

7. Comparisons among methods. We compare the performance of the two learning algorithms we consider to the performance of the two prevailing interval-based methods: the standard $4-\sigma$ criterion (ASTM-E2927-16 (2016), Trejos et al. (2013b)), and the modified $4-\sigma$ criterion (Weis et al. (2011)). We also include two parametric tests in the comparison: the optimum test statistic H (Parker and Holford (1968)), and the Hotelling T^2 statistic with a shrinkage covariance estimator (Campbell and Curran (2009)) which can be implemented using the R package `Hotelling`. As recommended by Parker and Holford (1968), the statistics H is tested on the log transformed values. We use the same test dataset in the comparison, but note that for the two interval-based and the two parametric classifiers, we do not need to train the algorithms.

The number of mated and nonmated pairs on the set-aside data that we use to compare the classification performance of the different methods depends on the number J of fragments from K obtained by the forensic scientist. If $J = 1$, there are 5590 known mated pairs of fragments and 123,805 known nonmated pairs of fragments in the 21 panes in the test dataset (see Table 1). ASTM-E2927-16 (2016)

recommends that at least three fragments with at least three replicated measurements be used to compute the mean and SD of each of the p elemental concentrations. Here, we followed those recommendations, and, in what follows, both interval-based methods and both parametric methods were implemented using the mean of 15 measurements (three fragments, five replicated measurements). For the two learning algorithms we report two sets of results. The first set is obtained by computing the similarity score between the mean of the 15 measurements from K and the average of five replicated measurements from Q. The second set is obtained by computing three similarity scores, one for each fragment mean from K and the fragment mean from Q, and then using the smallest score for classification. This approach is favorable to the defense, in that it results in lower probability of declaring that two fragments are chemically indistinguishable. In the second case we randomly select three fragments from K for each questioned fragment. For the purposes of these comparisons, we constructed 30 comparisons for each questioned fragment Q, each including three fragments in K. This resulted in 15,300 pairs of fragments known to have originated from the same piece of broken glass and 150,060 pairs of fragments known to have originated from different panes.

Table 3 shows the honest OOB classification errors when each of the methods was used to classify the pairs of fragments in the test set. We used a threshold equal to 0.5 in the RF and in BART, so that two fragments were declared to originate from the same pane of glass when the score exceeded 0.5. We only show the results obtained when the majority class was downsampled; results obtained using SMOTE were almost identical. We implemented the Hotelling T^2 approach using the `R` package `Hotelling` Campbell and Curran (2009) with a shrinkage covariance estimator and a randomization method to test the null hypothesis of no difference between fragments at the 5% significance level. To compute the p-value for Parker's optimum test statistic, we used the table on page 244 in Parker and Holford (1968). Since we have 18 features, we approximated the upper 5%

TABLE 3

Out-of-bag classification errors on test set. RF-Mean and RF-Min are labels for the RF classifier based on the average of 15 *measurements on the* K *sample or on the minimum of three similarity scores obtained from three* K *fragments, respectively. The same is true for the BART-Mean and BART-Min labels. FNR is False Negative Rate and FPR is False Positive Rate*

Error	**RF-Mean**	**RF-Min**	**BART-Mean**	**BART-Min**
FNR	0.0235	0.1290	0.0220	0.1341
FPR	0.0964	0.0564	0.0954	0.0490
Error	**Standard** $4-\sigma$	**Modified** $4-\sigma$	**Hot-**T^2**-shrinkage**	**Opt. test**
FNR	0.0559	0.4482	0.7798	0.7657
FPR	0.1866	0.0628	0.0006	0.0042

tail of the distribution of the statistic by simulation, as suggested by the author, and found that the threshold, for the test statistic when the number of features is 18 and the degrees of freedom is also 18, was 30.03.

From the results in Table 3, we see that the two learning algorithms, when implemented with downsampling with all 18 features, strike a good compromise between minimizing the false positive and the false negative rates (FPR and FNR, respectively). The learning classifiers that compare the average fragment in K with a fragment from Q exhibit FPR between 9.5% and 9.6% and FNR between 2.2% and 2.3% approximately. As expected, the classifier that is most favorable to the defendant has smaller FPR of 4.9% and 5.6% and FNP about 13%. The Hotelling T^2 shows the smallest FPR among all methods, but at the expense of incorrectly misclassifying 78% of all pairs of fragments known to have a common source. Parker's optimum test statistic also exhibits a low (0.5%) false positive rate but a high false negative rate. The standard interval approach proposed in ASTM-E2927-16 (2016) is dominated by all of the learning algorithms, including the RF that relies on the minimum of three scores. The poor performance of the two interval-based criteria is most likely due to the fact that both approaches ignore the dependence structure among elemental concentrations and rely on poorly estimated standard deviations. We discuss this further in Section 9.

One useful attribute of BART over other classifiers is that it provides a measure of the uncertainty associated with estimated empirical class probabilities. Figure 8 shows the estimated empirical class memberships and corresponding 95% credible sets for pairs of known mates (left panel) and known nonmates (right panel) for 100 randomly selected pairs in each class. Notice that low (high) estimated empirical probabilities of membership in the mated (nonmated) class have wide credible

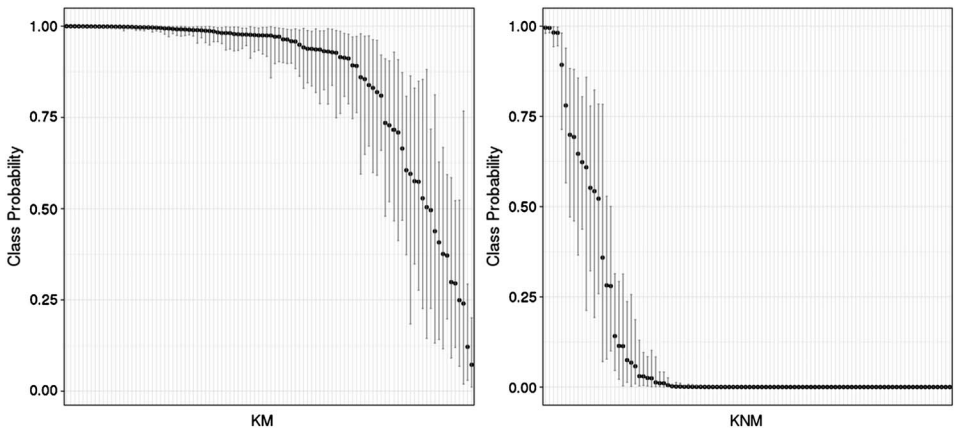

FIG. 8. *Class probability by BART down sampling and its credible interval on* 100 *random KM and KNM in the test set.*

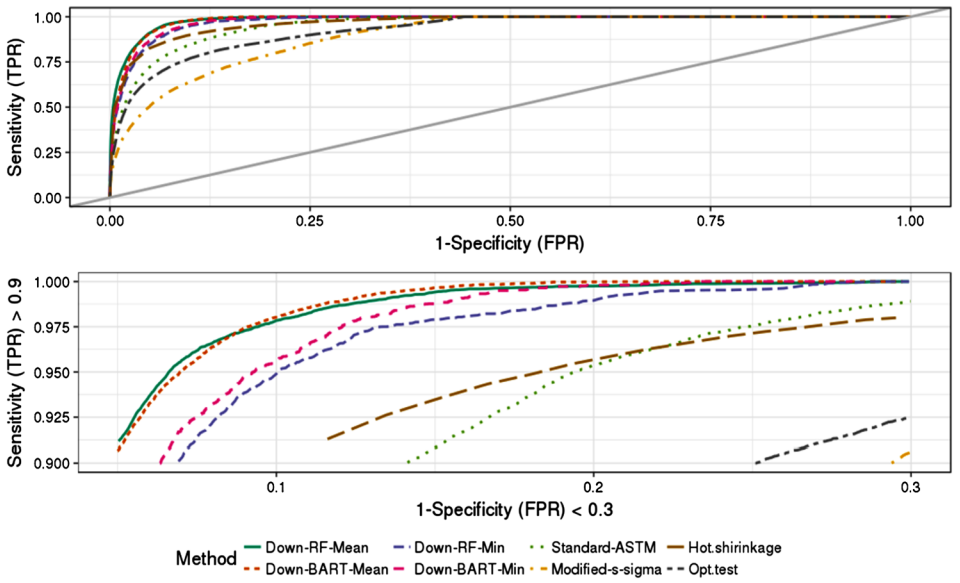

FIG. 9. *ROC curve of eight classifiers on the same out-of-bag test set.*

sets. This suggests that false positive and false negative classifications are subject to more uncertainty than true positive or negative classifications.

As an additional diagnostic we inspect the ROC curves for the honest out-of-bag test set for the eight classifiers in Table 3. To draw the curves, we considered 10,000 thresholds for each classifier. In the case of RF and BART, the 10,000 thresholds took on values between 0 and 1. For all other classifiers we considered 10,000 threshold values between the minimum and the maximum values of the corresponding scores. For the four machine learning classifiers we use majority voting rate as the score, as discussed earlier. The form of the scores that correspond to the two interval-based criteria are given in equation (3.2) and equation (3.4). In the case of the two parametric tests, the test statistics themselves constitute the score. Figure 9 shows the ROC curves corresponding to the eight classifiers (top panel), and a zoom-in to the upper left corner of the figure in the bottom panel. Results are consistent with those shown in Table 3.

An alternative way to evaluate the different approaches is to estimate the threshold value at which classification performance is optimized. Table 4 shows the estimated AUC and equal error rate for all classifiers, as well as the lowest achievable (for these particular data) false positive and false negative rates given an optimal threshold. Note that in the case of the standard $4 - \sigma$ criterion, the optimal threshold is 3.3, relatively close to the stated threshold equal to 4. For the random forest and BART classifiers that use the mean measurement vector of the K fragments, the optimal thresholds are close to 0.5 which is sometimes adopted as the default threshold when scores take on values between 0 and 1. Results shown in Table 4

TABLE 4
Area Under the ROC Curve (AUC) and equal error rate (EER) of existing classifiers

Model	AUC	EER	Opt. threshold	FPR	FNR
RF-Mean	0.984	0.061	0.590	0.076	0.037
BART-Mean	0.982	0.062	0.537	0.090	0.026
BART-Min	0.978	0.075	0.228	0.095	0.047
RF-Min	0.975	0.080	0.330	0.101	0.049
Hotelling T^2 shrinkage	0.966	0.100	244.208	0.096	0.104
Standard-ASTM	0.954	0.122	3.300	0.142	0.0984
Optimum test statistic	0.926	0.162	125.956	0.136	0.184
Modified $s - \sigma$ Criterion	0.899	0.204	12.961	0.298	0.096

suggest that–at least for the glass samples we analyzed–the RF and BART classifiers with the mean K measurement vector perform best in the sense of jointly minimizing the false positive and false negative rates.

8. The impact of the classifier on the score-based likelihood ratio statistic (SLR). The use of a likelihood ratio framework to quantify the strength of the evidence has been proposed by, for example, Lindley (1977), Aitken and Lucy (2004), and Hepler et al. (2012). A likelihood ratio (equation (2.1)) represents the odds of observing a match between two fragments under the competing hypothesis of same or different source. A high value of the LR supports the same source hypothesis, whereas low LR values close to 0 tend to support the hypothesis of different source. Some authors however have urged caution, in that producing an LR statistic in any specific situation typically relies on assumptions that can have impact on the resulting statistic (Lund and Iyer (2017)). In this section, we illustrate that even when using the same set of data, the LR based on a classification score can vary dramatically, depending on the properties of the score.

For illustration we compute the distribution of scores obtained from a random forest with downsampling and the standard $4 - \sigma$ interval approach proposed by Trejos et al. (2013b) using the 6710 pairs of mated fragments and 54,020 pairs of nonmated fragments with panes used to train the algorithms—described in Section 6. In both cases we used the mean of 15 measurements obtained from the K sample. We used the training dataset to mimic the scenario where a forensic scientist has a reference set of scores that she can use to determine the significance of a similarity computed from a casework sample. We used a nonparametric density estimator to estimate the densities f_s, f_d shown in Figure 10 and checked their reasonableness using a goodness-of-fit test. Here, f_s and f_d denote the densities of the scores of mated and nonmated pairs of fragments.

Figure 10 shows the two estimated densities for scores produced by a random forest with downsampling and by the standard $4 - \sigma$ interval method (ASTM-

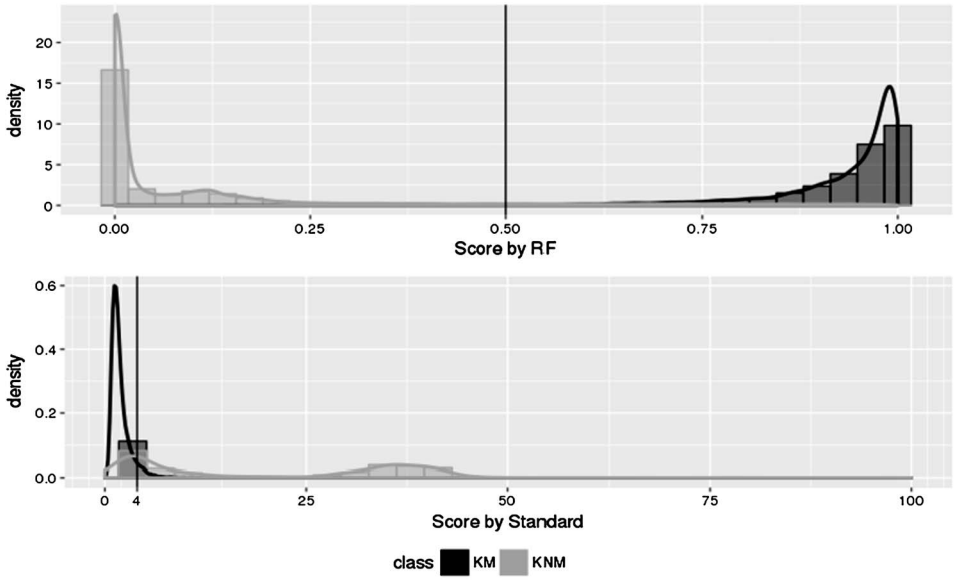

FIG. 10. *Density plot of scores computed by random forest and using the standard* $4-\sigma$ *approach.*

E2927-16 (2016)). In both panels the dark gray density corresponds to scores obtained for KM pairs, and the light gray density corresponds to the scores computed from the KNM pairs. On the top panel the black vertical line represents the threshold equal to 0.5, which we used to classify pairs into one of the two classes. In the bottom panel, the tail of the gray distribution extended to values over 4,000; we only show the estimated density for values of the score less than 100. In the bottom panel, the vertical line is drawn at the value 4, the threshold implied in the Trejos et al. (2013b) method. Both classifiers made the most mistakes when comparing two fragments from different panes that were produced by the same manufacturer on consecutive days.

The mated and nonmated score distributions obtained from the random forest scores have a small overlap that is due to the long right tail of the distribution of nonmated scores. About 4% of the mass of the nonmated score distribution is on scores over 0.5. For the densities computed using the standard interval scores, the overlap is higher, in particular for values of the score below 10. The estimated distribution of nonmated score by standard interval is also very long tail with 1% of them larger than 500 with maximum of 4879. It is also bimodal, and about 46% of its mass is on scores below 25. These scores arose from the comparison between fragments from different panes but produced by the same manufacturer and within a day or two.

Suppose that a forensic examiner in the course of evaluating some crime scene evidence, has to compare the pairs of fragments shown in the first two columns of Table 5. Suppose further that the examiner computes the corresponding scores

TABLE 5
The impact of two classifiers on the value of the SLR

Comparison			Random forest		Standard $4-\sigma$	
Pane-Frag.	**Pane-Frag.**	**Truth**	**Score**	**SLR**	**Score**	**SLR**
AB-2	AB-24	SP	0.986	175.81	1.59	14.40
BB-1	BB-5	SP	0.946	32.33	3.20	2.36
AB-2	BB-2	DP	0.000	4.30×10^{-14}	35.9	3.59×10^{-11}
AB-2	AF-2	DP	0.160	9.14×10^{-3}	5.23	0.471
BB-14	BD-2	DP	0.556	0.418	3.78	1.42

using the RF trained on the background data and the interval score described in Trejos et al. (2013b) and in ASTM-E2927-16 (2016). Using the estimated densities from Figure 10, we computed the SLR for each of the five comparisons and the two different scores. With the exception of the last pair of fragments, an examiner who uses the RF scores, will classify all pairs correctly and will obtain LR values that are clearly in support of the same or of a different source decision. The last pair of fragments was produced on consecutive days; fragments are similar enough to result in a score just barely above the 0.5 threshold. The corresponding SLR, however, would lead the examiner to correctly conclude that fragments come from different panes. If instead she uses the interval-based score, she will also incorrectly classify the last pair of fragments in the table and, in addition, will obtain SLR values that are ambiguous for most of the comparisons. The conclusion is that the range of values that the SLR can take on, and, therefore, the assessments of the score-based weight of evidence, is strongly dependent on the discriminating power of the scores on which classification decisions are based. Several authors (e.g., Morrison and Enzinger (2016), Hepler et al. (2012)) have observed that LR based on scores can exhibit unexpected behavior.

9. Discussion. In the United States' criminal justice system, *triers of fact* or jurors are typically expected to decide whether the evidence supports the prosecutor's or the defense's propositions. To do so, jurors rely on summaries of the evidence presented by experts during trial. In this paper, we propose a learning approach to summarize the evidence and compare its performance to the performance of other methods in the literature.

The community of forensic glass examiners has for years advocated the use of interval-based match criteria to decide whether two glass fragments are chemically indistinguishable. The two approaches that are currently considered to be state of the art were developed using limited data. We do not know of any dataset with elemental concentrations in glass that would permit obtaining well-conditioned estimates of covariance matrices of 12 or more elemental concentrations. This is a serious handicap given that dependencies among elemental concentrations are

present, and estimated pairwise correlations tend to be large in absolute value. We have begun collecting elemental concentration data using LA-ICP-MS in collaboration with colleagues in the University of Iowa, and we plan on putting the data in the public domain for the benefit of the general scientific community. Forensic databases tend to be proprietary, unfortunately. For example we requested the data that were analyzed by Dorn et al. (2015) but were denied access to them.

Our results suggest that supervised machine learning algorithms may provide a better summary of the evidence than interval-based methods. The two algorithms on which we focused here—random forests and BART–exhibited good classification performance when tested on a dataset that was not used in the training or validation of the algorithms. Furthermore, the learning algorithms do not rely on the standard hypothesis testing framework which, as discussed in Section 1, appears to violate the principle of "innocent until proven guilty." To implement BART, a probability model is implied; this is not true for RFs which are fully nonparametric methods. Both algorithms make no assumptions about the structure of the relationships among features. Two major challenges with supervised learning methods are their reliance on the training dataset and their potential for overfitting; both of these shortcomings affect the predictive ability of the algorithms. In the forensic context, the data on which the algorithms are trained can have an enormous effect on the answers to specific or common source questions of interest to forensic scientists. In the case of the forensic analysis of glass, it might be necessary to assemble different training datasets for different types of glass, for example, from automobile windows, from containers and bottles, from headlamps, etc., if it is found that these types of glass differ significantly in terms of their chemical composition. This is a matter that requires further research and much larger datasets. In this work, we did not include RI as an additional classification feature because we did not measure it for Dataset 3 fragments. However, a more complete reference dataset might include RI as well as other potentially discriminating features in glass.

One additional drawback of machine learning algorithms is that they tend to behave like black boxes. An attractive property of RFs, however, is that they permit ranking the features in terms of their importance for classification purposes. There are several reasons to carry out variable selection when growing a random forest: (1) To decrease training time, (2) to avoid the curse of dimensionality and overfitting and (3) to simplify the model and improve interpretability. In addition, we expect that in a highly multivariate data setting, not all features (elemental concentrations) will be equally discriminating. In the case of comparisons among glass fragments, it may be possible to work with fewer than 18 elements without losing discrimination power.

Figure 11 shows the variable importance estimated by a random forest trained with downsampling (left panel) and SMOTE (right panel). Importance is scaled to take on values between 0 and 100, and in Figure 11 elements are shown in decreasing order of importance. To select the most discriminating subset of features, we monitor the increase in error (or decrease in importance) as we move down the list.

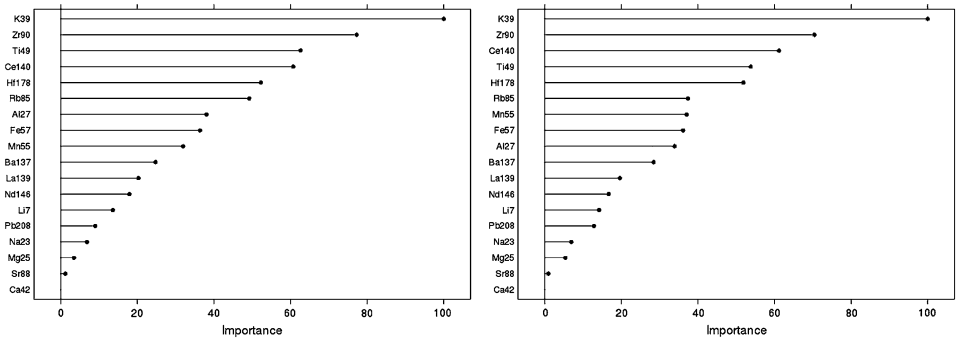

FIG. 11. *Variable importance from the random forest with down sampling or SMOTE.*

The set of nine most discriminating elements includes K, Ce, Zr, Ti, Hf, Rb, Al, Fe and Mn, regardless of the sampling approach. We compared the performance of the classifiers when using the full set of 18 features, the nine most discriminating features and the subset of 12 elements that have been identified as important and that are known to be good discriminators (e.g., Dorn et al. (2015)). In terms of AUC, results were similar whether based on the full set of features or on the subsets with nine or 12 features, but there was some loss in sensitivity and specificity when subsets of the features were used to train the classifiers.

Oftentimes in casework, forensic scientists are able to obtain multiple fragments from the broken glass at the crime scene. The ASTM standards recommend that at least three fragments be obtained from the reference sample at the crime scene, and that on each fragment, measurements be replicated three times. To explore whether increasing the number of reference fragments would have an impact on the classification ability of the standard $4 - \sigma$ and the modified $4 - \sigma$ approaches, we repeated the comparison described in Section 6 using three, six, nine and 12 fragments from K with measurements on each fragment replicated five times. Table 6

TABLE 6
Out-of-bag classification errors on test set using 3, 6, 9 *or* 12 *control samples*

	Standard $4 - \sigma$			
Error	**3 controls**	**6 controls**	**9 controls**	**12 controls**
FNR	0.0559	0.0176	0.0067	0.0042
FPR	0.1866	0.1948	0.2017	0.2043
	Modified $4 - \sigma$			
Error	**3 controls**	**6 controls**	**9 controls**	**12 controls**
FNR	0.4482	0.4303	0.4184	0.4203
FPR	0.0628	0.0646	0.0662	0.0674

shows the results we obtained. The performance of the modified $4-\sigma$ algorithm did not change. This was expected, because the algorithm relies on a fixed standard deviation estimate that is independent of the observed measurements. More interesting was the behavior of the standard $4-\sigma$ classifier as a function of the number of reference fragments. While the FNR decreased by over five percentage points, the FPR increased slightly, by a bit over one percentage point, as the number of reference fragments increased from three to 12. An explanation for these modest changes in classification performance, even when the reference sample size quadrupled, is the fact that the estimate of σ is bounded below by a value equal to 3% of the corresponding mean elemental concentration. When the number of reference fragments J was equal to three, the SD of the measurements exceeded the 3% of the mean floor for 12 elements, and, therefore, the SD was set to 3% of the mean elemental concentration for six out of 18 elements. When $J=12$, however, the actual SD was used to construct the $4-\sigma$ intervals for 15 out of 18 elements.

We envision that at some point there will be relevant and stable enough training databases, so that the distributions of scores under the two scenarios (same or different source) can be well estimated. If so, then a practicing forensic scientist who must compare a single fragment recovered from a suspect to a few fragments known to originate from the crime scene would be able to compute the comparison score on the pairs of evidence fragments and then decide whether the score is high enough to suggest same source. To determine what is "high enough," the forensic practitioner would refer to the relevant distributions of the score under the two competing hypotheses of same or different source calculated from appropriate reference datasets. We insist on the importance of developing those reference score distributions using the appropriate datasets. In the case of glass, these reference datasets might need to be updated on a regular basis. Alternatively, the forensic scientist could also compute a score-based likelihood ratio statistic as in Section 8 but at the cost of making additional assumptions about the score distributions that may or may not be plausible.

When applied to the dataset that combines the BKA measurements and the measurements obtained by Iowa State, the two interval-based criteria exhibited higher miss-classification errors than the learning-based methods. Both the random forest and standard $4-\sigma$ criterion tend to misclassify pairs of fragments that originate from panes produced within one to two days in the same manufacturing facility.

To investigate why these interval based algorithms exhibit larger FNR, we carried out a small simulation study as follows. We considered three scenarios: a fragment compared to itself, a fragment compared to another fragment from the same pane, and a fragment compared to a fragment from a different pane. We estimated an 18-dimensional mean vector and covariance matrix for each fragment included in the simulation, by using a subset of fragments in Dataset 3 for which we had 20 replicate measurements on each. The fragments we used in this simulation were AB-14, AB-24, AAR-14, AAR-24, AL-14, AL-24, BJ-14 and BJ-24. We also included fragment 104G from pane X. For this fragment we had 11 measurements

TABLE 7
Simulation results: *estimated classification error rate*

Same Pane & Same Fragment					
Pane	**Fragment**	**Pane**	**Fragment**	**Error (Modified $4-\sigma$)**	**Error (RF)**
AB	14	AB	14	0.214	0
AB	24	AB	24	0.208	0
AAR	24	AAR	24	0.005	0
AAR	14	AAR	14	0.094	0
X	104G	X	104G	0	0
Same Pane & Different Fragment					
Pane	**Fragment**	**Pane**	**Fragment**	**Error (Modified $4-\sigma$)**	**Error (RF)**
AAR	14	AAR	24	1	0
AB	14	AB	24	0.945	0
AL	14	AL	24	0.692	0
BJ	14	BJ	24	1	0
Different Pane					
Pane	**Fragment**	**Pane**	**Fragment**	**Error (Modified $4-\sigma$)**	**Error (RF)**
AB	24	AAR	14	0	0
AB	14	BJ	14	0	0
X	104G	BJ	14	0	0

made on consecutive days, each replicated six times, for a total of 66 observation vectors. Next, using the estimated mean vector and covariance matrix from each fragment, we generated five random draws from a multivariate log-normal distribution. We used these five replicates to compare pairs of fragments using the modified $4-\sigma$ criterion and the random forest with down sampling on 18 variables built in Section 6. We generated 1,000 sets of five replicates for each fragment and used those to carry out the comparisons shown in Table 7. Results are shown in the last two columns of the table.

In these particular simulation scenarios, when comparing a fragment to itself or to a fragment from the same source, the modified $4-\sigma$ criterion results in false negative errors in a large proportion of cases. For example, the modified $4-\sigma$ criterion incorrectly classified fragments 14 and 24 from pane AB as originating from a different source almost 95% of the time. The modified $4-\sigma$ criterion did well, however, when comparing fragments from different panes. This suggests that the modified $4-\sigma$ criterion performs well when fragments are sufficiently different, but results in a large number of false exclusions (different source conclusions) when fragments are similar. The random forest, on the other hand, performed well in all cases.

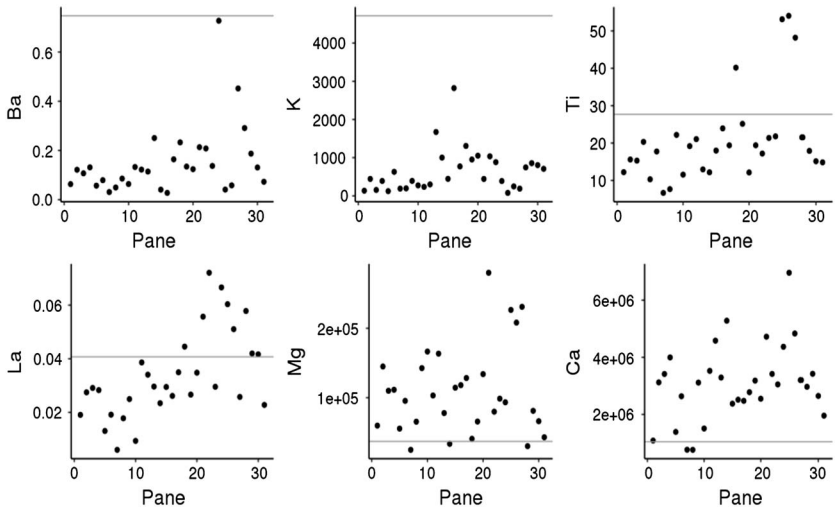

FIG. 12. *Within pane variances* (*dots*) *from* 31 *Company A panes and between pane variance* (*line*).

Why would the modified $4 - \sigma$ criterion fail when fragments have similar (but not identical) elemental composition? The problem arises when elements are more variable within source than between sources. Figure 12 shows the within-pane variances for six elements (diagonal elements of the covariance matrices) as dots. The between-source variances are shown as a constant line. Intuitively, we would expect to see that the within-pane variances are smaller than the between-pane variances, but that is not the case for several elements in our data. This constitutes a challenge for both interval-based methods because they rely exclusively on within-pane variances. The random forest, on the other hand, "learns" which features are more discriminating by looking at the within and between pane variance of elemental concentrations. From Figure 12 we see that elements such as Ca and Mg (also Sr, Na, Pb, not shown) have larger within pane variance than between pane variance (line). Those elements are ranked as less important for classification, as shown in Figure 11.

Finally, we note that while this paper has focused on forensic glass comparisons, the protocols outlined here are broadly applicable to many other forensic disciplines, including those that rely on pattern recognition; see, for example, Song (2015), Hare, Hofmann and Carriquiry (2017), Swofford et al. (2018). The basic two-step approach, appropriately tailored to the measurements that can be made in the various contexts, can be an appealing alternative to a probability model-based likelihood ratio approach. At the very least it can provide a valid means to compare two or more items and to assess the probative value of the evidence while research on likelihood-based or Bayes factor-based approaches is ongoing.

Acknowledgments. We wish to express our gratitude to Dr. Peter Weis from the German BKA (Wiesbaden) who has so generously shared his knowledge and

his data with us. Dr. Weis is a true scientist, interested in seeking the best methods to use in forensic casework. We have also benefited from discussions with Dr. Tatiana Trejos from the University of West Virginia and with Dr. Joann Buscaglia (FBI) about specific attributes of float glass. Dr. Karen Kafadar, who is also working with elemental concentrations in float glass, was a good referent for us. Dr. David Peate from the University of Iowa has been a valuable collaborator, and we thank him for the care with which he has produced the analytical measurements we use in this study. The glass samples we analyzed were provided to the Ames National Laboratory by two glass manufacturers in the United States whom we wish to acknowledge; we promised not to reveal their names, so we have referred to them as Company A and Company B. Finally, we also wish to thank two anonymous referees and an Editor for their insights and for their many constructive comments.

REFERENCES

AESCHLIMAN, D. B., BAJIC, S. J., BALDWIN, D. P. and HOUK, R. (2003). Spatially-resolved analysis of solids by laser ablation-inductively coupled plasma-mass spectrometry: Trace elemental quantification without matrix-matched solid standards. *J. Anal. At. Spectrom.* **18** 872–877.

AITKEN, C. G. G. and LUCY, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 109–122. MR2037883

ASTM-E2330-12 (2012). Standard test method for determination of concentrations of elements in glass samples using inductively coupled plasma mass spectrometry (ICP-MS) for forensic comparisons. Technical report, ASTM International. Available at https://doi.org/10.1520/E2330-12.

ASTM-E2927-16 (2016). Standard test method for determination of trace elements in soda-lime glass samples using laser ablation inductively coupled plasma mass spectrometry for forensic comparisons. Technical report, ASTM International. Available at https://doi.org/10.1520/E2927-16.

BIRNBAUM, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49** 559–574. MR0065101

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CAMPBELL, G. P. and CURRAN, J. M. (2009). The interpretation of elemental composition measurements from forensic glass evidence III. *Sci. Justice* **49** 2–7.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. and KEGELMEYER, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Res.* **16** 321–357.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

CURRAN, J. M. (2003). The statistical interpretation of forensic glass evidence. *Int. Stat. Rev.* **71** 497–520.

CURRAN, J. M., CHAMPOD, T. N. H. and BUCKLETON, J. S. (2000). *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton, FL.

CURRAN, J., TRIGGS, C., ALMIRALL, J., BUCKLETON, J. and WALSH, K. (1997a). The interpretation of elemental composition measurements from forensic glass evidence: II. *Sci. Justice* **37** 245–249.

CURRAN, J., TRIGGS, C., ALMIRALL, J., BUCKLETON, J. and WALSH, K. (1997b). The interpretation of elemental composition measurements from forensic glass evidence: I. *Sci. Justice* **37** 241–244.

DAVIS, L. J., SAUNDERS, C. P., HEPLER, A. and BUSCAGLIA, J. (2012). Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic Sci. Int.* **216** 146–157.

DELONG, E. R., DELONG, D. M. and CLARKE-PEARSON, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44** 837–845.

DORN, H., RUDDELL, D. E., HEYDON, A. and BURTON, B. D. (2015). Discrimination of float glass by LA-ICP-MS: Assessment of exclusion criteria using casework samples. *Can. Soc. Forensic Sci. J.* **48** 85–96.

GARVIN, E. J. and KOONS, R. D. (2011). Evaluation of match criteria used for the comparison of refractive index of glass fragments. *J. Forensic Sci.* **56** 491–500.

HARE, E., HOFMANN, H. and CARRIQUIRY, A. (2017). Automatic matching of bullet land impressions. *Ann. Appl. Stat.* **11** 2332–2356. MR3743299

HEPLER, A. B., SAUNDERS, C. P., DAVIS, L. J. and BUSCAGLIA, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* **219** 129–140.

HICKMAN, D. (1987). Glass types identified by chemical analysis. *Forensic Sci. Int.* **33** 23–46.

HOUK, R. S. (1990). Elemental analysis by atomic emission and mass spectrometry with inductively coupled plasmas. In *Handbook on the Physics and Chemistry of Rare Earths* **13** 385–421.

KAM, H. T. (1995). Random decision forest. In *Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition*, *Montreal*, *Canada*, *August* 14–18.

KAPELNER, A. and BLEICH, J. (2013). bartMachine: Machine learning with Bayesian additive regression trees. Preprint. Available at arXiv:1312.2171.

KOONS, R. D. and BUSCAGLIA, J. (2002). Interpretation of glass composition measurements: The effects of match criteria on discrimination capability. *J. Forensic Sci.* **47** 505–512.

KOONS, R. D., FIEDLER, C. and RAWALT, R. (1988). Classification and discrimination of sheet and container glasses by inductively coupled plasma-atomic emission spectrometry and pattern recognition. *J. Forensic Sci.* **33** 49–67.

KOONS, R. D., PETERS, C. A. and REBBERT, P. S. (1991). Comparison of refractive index, energy dispersive X-ray fluorescence and inductively coupled plasma atomic emission spectrometry for forensic characterization of sheet glass fragments. *J. Anal. At. Spectrom.* **6** 451–456.

LEDOIT, O. and WOLF, M. (2003). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management* **30**.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213. MR0518935

LUNARDON, N., MENARDI, G. and TORELLI, N. (2014). ROSE: A package for binary imbalanced learning. *R J.* **6** 79–89.

LUND, S. P. and IYER, H. K. (2017). Likelihood ratio as weight of forensic evidence: A closer look. Preprint. Available at arXiv:1704.08275.

MORRISON, G. S. and ENZINGER, E. (2016). What should a forensic practitioner's likelihood ratio be? *Sci. Justice* **56** 374–379.

OMMEN, D. M. and SAUNDERS, C. P. (2018). Building a unified statistical framework for the forensic identification of source problems. *Law Probab. Risk* **17** 179–197.

PARKER, J. B. (1966). A statistical treatment of identification problems. *Sci. Justice* **6** 33–39.

PARKER, J. B. (1967). The mathematical evaluation of numerical evidence. *Sci. Justice* **7** 134–144.

PARKER, J. B. and HOLFORD, A. (1968). Optimum test statistics with particular reference to a forensic science problem. *Appl. Stat.* **17** 237–251.

SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. MR2183942

SCHEER, S. (2006). The evidential value of elemental composition in forensic glass examination: The use of multivariate likelihood ratio methods. Available at https://www.forensic.to/webhome/statistiek/AfstudeerscriptieSonja.pdf.

SONG, J. (2015). Proposed congruent matching cells (CMC) method for ballistic identification and error rate estimation. *AFTE J.* **3** 177–185.

SWOFFORD, H., KOERTNER, A., ZEMP, F., AUSDEMORE, M., LIU, A. and SALYARDS, M. (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Sci. Int.* **287** 113–126.

TANGRAM-TECHNOLOGY (2004). Float glass production. Available at http://www.tangram.co.uk/TI-Glazing-Float%20Glass.html.

TREJOS, T., FLORES, A. and ALMIRALL, J. R. (2010). Micro-spectrochemical analysis of document paper and gel inks by laser ablation inductively coupled plasma mass spectrometry and laser induced breakdown spectroscopy. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **65** 884–895.

TREJOS, T., KOONS, R., BECKER, S., BERMAN, T., BUSCAGLIA, J., DUECKING, M., ECKERT-LUMSDON, T., ERNST, T., HANLON, C., HEYDON, A. et al. (2013a). Cross-validation and evaluation of the performance of methods for the elemental analysis of forensic glass by μ-XRF, ICP-MS, and LA-ICP-MS. *Anal. Bioanal. Chem.* **405** 5393–5409.

TREJOS, T., KOONS, R., WEIS, P., BECKER, S., BERMAN, T., DALPE, C., DUECKING, M., BUSCAGLIA, J., ECKERT-LUMSDON, T., ERNST, T. et al. (2013b). Forensic analysis of glass by μ-XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: Evaluation of the performance of different criteria for comparing elemental composition. *J. Anal. At. Spectrom.* **28** 1270–1282.

WEIS, P., DÜCKING, M., WATZKE, P., MENGES, S. and BECKER, S. (2011). Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *J. Anal. At. Spectrom.* **26** 1273–1284.

ZADORA, G. (2009a). Classification of glass fragments based on elemental composition and refractive index. *J. Forensic Sci.* **54** 49–59.

ZADORA, G. (2009b). Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian network approaches. *Anal. Chim. Acta* **642** 279–290.

DEPARTMENT OF STATISTICS AND
STATISTICAL LABORATORY
IOWA STATE UNIVERSITY
195 DURHAM
AMES, IOWA 50011-1210
USA
E-MAIL: sypark@iastate.edu
alicia@iastate.edu