The Annals of Applied Statistics
2018, Vol. 12, No. 2, 788–814
https://doi.org/10.1214/18-AOAS1180

STATISTICAL MODELING AND ANALYSIS OF TRACE ELEMENT CONCENTRATIONS IN FORENSIC GLASS EVIDENCE

BY KAREN D. H. PAN[1] AND KAREN KAFADAR[2]

University of Virginia

The question of the validity of procedures used to analyze forensic evidence was raised many years ago by Stephen Fienberg, most notably when he chaired the National Academy of Sciences' Committee that issued the report *The Polygraph and Lie Detection* [National Research Council (2003) The National Academies Press]; his role in championing this cause and drawing other statisticians to these issues continued throughout his life. We investigate the validity of three standards related to different test methods for forensic comparison of glass (micro X-ray fluorescence (μ-XRF) spectrometry, ICP-MS, LA-ICP-MS], all of which include a series of recommended calculations from which "it may be concluded that [the samples] did not originate from the same source." Using publicly available data and data from other sources, we develop statistical models based on estimates of means and covariance matrices of the measured trace element concentrations recommended in these standards, leading to population-based estimates of error rates for the comparison procedures stated in the standards. Our results therefore do not depend on internal comparisons between pairs of glass samples, the representativeness of which cannot be guaranteed: our results apply to any collection of glass samples that have been or can be measured via these technologies. They suggest potentially higher false positive rates than have been reported, and we propose alternative methods that will ensure lower error rates.

1. Introduction. Trace element analysis has been used to evaluate the source of bullets, glass, paint, copper wire, and other types of physical evidence. The "working hypothesis" is that the concentrations of certain elements, those that presumably are highly specific to different pieces of evidence, provide a distinctive "signature" that allows one to connect evidence found at a crime scene with a specific source found in the possession of a suspect. Statistical issues surround this approach. For example, if the batch of material from which the evidence is manufactured is extremely homogeneous, then the measurements on *many* pieces from the same batch may be deemed "not distinguishable," depending on the level of error in the measurements themselves, thereby leading one to erroneously interpret

Received December 2017; revised May 2018.

[1]Supported in part by National Institute of Standards & Technology via subcontract from Iowa State University.

[2]Supported in part by Isaac Newton Institute for Mathematical Sciences, *Probability and Statistics in Forensic Science*, EPSRC Grant Number EP/K032208/1.

Key words and phrases. Robust methods, exploratory data analysis, multivariate lognormal distribution, covariance matrix, standard errors, error rates, ROC curve.

the "not distinguishable" conclusion as "came from the same source" and hence to potential false positives. Conversely, if the specific piece of evidence is itself rather inhomogeneous, then concentrations in pieces from two different parts of the *same* evidence may be different, leading to false negatives. Thus, trace element analysis of forensic evidence may be unsatisfactory for both inclusion and exclusion purposes. Although the terms "inclusion/exclusion" are never explicitly used, jurors in the United States may well understand "analytically indistinguishable" to mean "from the same source" [Gabel-Cino (2017)]. Wording in the standards, "[i]f the samples are distinguishable... in any of these observed and measured properties, it may be concluded that they did not originate from the same source of broken glass," also provides conditions under which "distinguishable" samples can be used for exclusion [ASTM E2330-12 Section 1.1, ASTM E2927-16 Introduction, ASTM 2926-13 Introduction (ASTM International (2012, 2013, 2016))]. In addition, ASTM E2927-16 asserts that following the technique it describes "yields high discrimination among sources of glass" and "provides high discriminating value in the forensic comparison of glass fragments." The problem is not just that jurors may misunderstand the meaning of "not distinguishable;" of much greater concern is that the standards themselves describe the probative value of this determination in a manner that may well be highly misleading.[3]

In this article, we describe a previous example of forensic trace evidence analysis, compositional analysis of bullet lead (CABL), which used similar inferential procedures for assessing two pieces of trace evidence material as "analytically [in]distinguishable." We then devote the rest of the paper to the description of procedures and data for the inference procedures that have been proposed for analyzing trace element concentrations in glass evidence. We provide alternative procedures which have lower false positive rates, and conclude with some cautions about using glass evidence as a definitive tool for suspect identification or exclusion.

2. Compositional analysis of bullet lead. Prior to the publication of the landmark report from the National Academy of Sciences (NAS), *Strengthening Forensic Science in the United States: A Path Forward* [National Research Council (2009), hereafter NRC (2009)], the NAS published its findings on a procedure used at the United States Federal Bureau of Investigation (FBI) known as Compositional Analysis of Bullet Lead [CABL; NRC (2004)]. The report provided an in-depth analysis of the statistical procedure used to compare the "signatures" between two samples of bullet lead [one from the crime scene (CS) and one from the potential suspect (PS), sometimes called in the forensics discipline "known" (K) and "questioned" (Q) or "recovered" (R)]. The "working hypothesis" was that a vector of seven measured trace elemental concentrations (silver Ag, antimony Sb, arsenic

[3]We thank Reviewer 2 for this remark.

As, bismuth Bi, cadmium Cd, copper Cu, lead Pb) would provide a unique "signature" that could be used to distinguish samples from different sources or confirm commonality of source. The FBI's "2-SD-overlap" procedure involved measuring, in triplicate, the concentrations of elements in both the K and Q bullets, calculating the sample mean and standard deviation on each element and for each bullet (or bullet fragment), and forming "mean $\pm$ 2 $\cdot$ SD" intervals. If the K and corresponding Q intervals overlapped for all seven elements, then K and Q were deemed "analytically indistinguishable." The FBI often went further in the courtroom by testifying that K and Q "likely originated from the same manufacturer's source of lead" or "must have come from the same box" [NRC (2004), pages 91–92]. The FBI calculated its "false positive error rate" (probability of claiming "same source" when the samples came from different sources) by counting the number of pairs between any two of 1837 samples in their "data base" that resulted in a "false match" by their procedure; i.e., they found 693 among their 1,686,366 pairs that resulted in a "false match." (In this paper, we refer to the rule used to determine "analytically indistinguishable" as the "match rule" and the proportion of times that the rule is satisfied as the "match rate.")

The NAS Committee concluded that (a) the "data base" of 1837 samples, which included "one specimen from each combination of bullet caliber, style, and nominal alloy class" [Koons (2003) and Koons and Buscalglia (2005)], could not be viewed as a representative sample of bullets; they were "selected" in hopes of spanning the space of possible bullet types, thereby resulting in pairs of bullets that could be expected to be more different than might be seen in a real case; (b) a procedure for estimating CABL's error rate is based more properly on statistical modeling of the covariance (or correlation) matrix among the seven elements in the proposed "signature," from which more valid estimates of sensitivity (given that the true concentrations differ by less than a prescribed "difference threshold," the probability that the FBI procedure properly concludes "same batch signatures") and specificity (given that the true concentrations differ by more than a prescribed "difference threshold," the probability that the FBI procedure properly concludes "different batch signatures") can be calculated. The Appendices in the NRC (2004) report concluded that the error rates of the FBI procedure will be much higher than the claimed 0.04%. The Committee found no fault with the actual measurement technique (inductively coupled plasma optical emission spectrometry, or ICP-OES), and had few recommendations on the laboratory procedures; rather, the concerns centered around the claimed error rates and the documented claims of "same source" identifications using the FBI's "match" procedure.

A major concern that was raised with CABL was the existence of thousands or even millions of bullets that may have similar chemical "signatures," simply due to the consistency in the lead manufacturing process: large homogenized batches of lead were likely to yield very similar concentrations for the thousands or millions of bullets that were created from the lead in a homogeneous batch. Further, once made into bullets and packaged into boxes of 25 or 50 bullets per box, no

box could be guaranteed to have bullets that came from only one batch of lead. Hence, a definitive statement such as "this bullet must have come from this box of 50 bullets" could not be supported, knowing that hundreds of other boxes likely contained bullets with the same signature. Moreover, bullets that did *not* satisfy the "match rule" did not guarantee that the two bullets came from different boxes, as bullets from different batches might have ended up in the same box. Thus, the procedure was useful for neither "inclusion" nor "exclusion" [for more about CABL, see Spiegelman and Kafadar (2006) and Giannelli (2010)].

3. Forensic comparison of glass: ASTM standards. Recently, an approach similar to CABL's "2-SD-overlap" procedure has been recommended for comparing glass samples found at a crime scene (CS) with those found on, or in connection with, a potential suspect (PS). Three standards from the American Society for Testing Materials (ASTM) have been published related to the use of measured trace element concentrations in glass, using three different techniques:

- ASTM E2330-12, *Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*.
- ASTM E2926-13, *Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*.
- ASTM E2927-16, *Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons*.

For simplicity, we will denote these three methods by ICP-MS, XRF, and LA-ICP-MS, respectively. Also, here we will refer to samples from completely different panes of glass as "samples," and to pieces from the same pane of glass as "fragments," because we expect more consistency in "fragments" from a single pane than between "samples" from different panes. (Indeed, this difference in consistency forms the basis of the ASTM standards on forensic glass evidence.)

Each standard includes a section entitled "Calculation and Interpretation of Results." The steps in this section are similar in each standard; below are those for E2330-12 (ICP-MS):

10.1.1 *For the Known source fragments, using a minimum of* 3 *measurements, calculate the mean for each element.*

10.1.2 *Calculate the standard deviation for each element. This is the Measured SD.*

10.1.3 *Calculate a value equal to* 3% *of the mean for each element. This is the Minimum SD.*

10.1.4 *Calculate a match interval for each element with a lower limit equal to the mean minus* 4 *times the SD (Measured or Minimum, whichever is greater) and an upper limit equal to the mean plus* 4 *times the SD (Measured or Minimum, whichever is greater).*

10.1.5 *For each Recovered fragment, using a minimum of* 3 *measurements, calculate the mean concentration for each element.*

10.1.6 *For each element, compare the mean concentration in the Recovered fragment to the match interval for the corresponding element from the Known fragments.*

10.1.7 *If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not "match" and the glass samples are considered distinguishable.*

For ASTM E2927-16 (LA-ICP-MS), "Calculation and Interpretation of Results" appears as Section 11, also with a "4-SD match interval"; for E2926-13, Section 10.7.3.2 uses a "3-SD match interval":

> "For each elemental ratio, compare the average ratio for the questioned specimen to the average ratio for the known specimens $\pm 3s$. This range corresponds to 99.7% of a normally distributed population. If, for one or more elements, the average ratio in the questioned specimen does not fall within the average ratio for the known specimens $\pm 3s$, it may be concluded that the samples are not from the same source."

[Note that the "99.7%" coverage applies only if the standard deviations were known, not estimated—as they are here, from possibly as few as three measurements—and only if the measurements come from a Gaussian (normal) distribution.] Whereas the FBI's CABL procedure involved calculating "mean $\pm$ 2 $\cdot$ SD" for the K and Q specimens, the glass standards calculate instead intervals of the form "mean $\pm$ 4 $\cdot$ SD" for only the K fragment and check to see if the means for the Q fragment fall in the corresponding intervals (i.e., only one set of SDs is calculated, for the concentrations in the K fragments). The glass standards also recommend the use of 8–17 elements, not just seven.

The justification for this procedure [Dorn et al. (2015), Trejos et al. (2013), Weis et al. (2011), Koons and Buscaglia (2001)] appears to be based on empirically observed "error rates" calculated among all pairs of glass samples from different sources. For example, Weis et al. (2011) measured 62 different samples, mostly from different manufacturers, but some from the same manufacturer produced from different batches at different time periods. The "error rate" was then calculated as the proportion of all pairs that satisfied the "match" criterion, even though the two samples in the pair came from different sources. Comparing each one of the 62 samples as the K with any one of the other 61 samples as the Q, they found two of the 1891 pairs satisfied their "modified n-sigma criterion with fixed relative standard deviations (FRSDs)" (Type II error rate 0.11%), where the FRSDs varied between 3.0% and 8.9% (see Table 7 on page 1281). Dorn et al. (2015) used a similar "4-SD match criterion," but with an RSD_{min} set to 3% for the concentrations of the 10 elements in their study (page 89). They found similarly small "error

rates": 0.27% (6/2256, 48 same-source samples)[4] for "Type I error rate" (two samples from same source failed to satisfy the "match" criterion), and 0.11% (7/6642, 82 different-source samples) for "Type II error rate" (two different samples satisfied the "match" criterion).

The reported error rates from four commonly referenced papers [Dorn et al. (2015), Koons and Buscaglia (2001), Trejos et al. (2013), Weis et al. (2011)] are very low, typically $<1\%$. They calculate these error rates by comparing two different-source samples from among all possible pairs in the data set (so the same sample is used for multiple different-source comparisons): if the sample means for the elemental concentrations from the assigned "R" ("recovered") fragment fall within all corresponding "mean $\pm$ 4 $\cdot$ SD" intervals calculated from the measured elemental concentrations from the assigned "K" ("known") fragment, then a "false positive" is recorded. (Note that the conclusions may be different if samples i and j are "R" and "K", versus if they are "K" and "R", respectively.) Unfortunately, with this method, the false positive rate (FPR) will depend on the nature of the samples in the data set. If the glass samples in the data set all are of the same type (e.g., Honda windshields) that were all manufactured at nearly the same times, then one may well expect that the mean differences among these samples will be closer than if the data set contains samples of very different types (e.g., car windshields and baby food jars). Consequently, the estimated false positive rate in the first data set (highly similar samples) may well be higher than that from the second data set (very different samples). To eliminate this dependence on data set (whose true concentrations can be only estimated anyway via measurements), the only way to understand the probability of a false positive, when two samples' mean concentrations are close versus far, is to calculate the rate of false positives among hundreds of comparisons whose mean concentrations differ by a *known pre-specified amount*, say δ_0. If δ_0 is huge (as it may be for some data collections), we'd expect a low false positive probability. But as δ_0 gets smaller (as may occur for different glass panes manufactured at nearly the same time), the FPP may be much higher. For these reasons, we chose a more formal statistical modeling approach to estimating error rates.

As was done in the NAS report for CABL, our statistical modeling approach provides more accurate estimates of error rates in the proposed procedure for assessing "distinguishability" between two glass samples. The need for modeling is even more critical here because (i) the data bases are necessarily much smaller than 1837 (the size of the data set shared by the FBI to the NAS Bullet Lead Committee in 2003); typically a lab has the facilities to measure and store at most only a few hundred samples; and (ii) the procedure uses a "minimum SD" method—the maximum of the calculated SD from all K fragments ("at least three" replicates)

[4]Note that Dorn et al. (2015) actually measured 24 fragments nine times each, and a 25th fragment 24 times, which is quite different from 48 fragments. See page 87, "*Group 1*", for details.

and 3% of the mean (from ideally three fragments, but not required). Such a procedure is not easily analyzed theoretically[5] so we resort to statistical modeling of the available data, validation of this model, and then simulation of samples according to this model, from which error rates can be calculated. Two critical advantages of this approach are (a) we are able to simulate "measurements" that may come from an idealistic Gaussian distribution *as well as* from more realistic distributions (e.g., from distributions that have heavier tails or outliers more often than the presumed idealistic Gaussian does); and (b) we can quantify more precisely the error rate *when the true difference in elemental concentration is a known stated percentage* (e.g., if the true difference is 3%), because we can simulate samples that have concentrations at specific levels. By doing so, we are able to quantify more precisely the level of the difference in concentrations at which the error rate of a "4-SD match interval" procedure will fall below a specific target.

In the next section, we describe the basic statistical modeling procedure and the features of the data sets that we used to develop appropriate models for this purpose. Section 5 describes our approach to estimating the means and covariance matrices from these data sets: one that uses ICP-MS (from FIU, Florida International University) and two using LA-ICP-MS [one from Germany (GER) courtesy of Peter Weis as reported in Weis et al. (2011), and one from Canada (CAN) courtesy of David Ruddell as reported in Dorn et al. (2015)], leading to our estimates of error rates in Section 6. Unfortunately, no data sets measured by XRF seem to be available; we describe in Section 7 the challenges of deriving similarly appropriate error rates for the ASTM E2926-13 standard without data. We conclude in Section 8 with some recommendations on this approach to forensic glass evidence. All analyses were conducted in R (Version 3.3.1, 2016-06-21) and programming details can be obtained from the authors.

4. Statistical modeling. Chemists refer to "relative standard deviation" (RSD) rather than the raw SD, because the SD of elemental concentrations tends to be related to the mean. Hence, six measurements of ^{7}Li on a glass fragment might be (4.56, 4.68, 4.79, 4.25, 4.33, 4.49) whose mean is 4.517 and SD is 0.205 (RSD = 0.205/4.517 = 4.5%); but six measurements of ^{90}Zr might be (54.16, 55.25, 51.93, 50.13, 49.97, 49.44) whose mean and SD are 11.5 and 11.8 times larger (51.813 and 2.416, respectively) but whose RSD is nearly the same (2.416/51.813 = 4.7%). It is well known that, when the RSD is small (<5%), the standard deviation of the *logarithms* of the observations is very close to the RSD; using the same ^{7}Li data, their logs are (1.517, 1.543, 1.567, 1.447, 1.466, 1.502) for which the mean is 1.507 [close to log(4.517) = 1.508] and the SD is 0.045 (virtually identical to the RSD of the original measurements). Moreover,

[5]If the data are guaranteed to come from a lognormal distribution (which is rarely true due to outliers and other causes for unusual departures), theory would require the distribution of $\max\{s, 3\%\}$, which is the square root of a truncated chi-squared distribution.

while the original measurements may tend to have skewed distributions, the distribution of their logarithms tends to be more symmetric. For both these reasons—interest in RSD, not the raw SD, and greater symmetry in the distribution of the measurements—we will model the *logarithms* of the measured concentrations on all elements, so the estimated SDs are approximately the RSDs.

4.1. *Sources of variability.* The measurements of trace elements in a medium such as glass or bullet lead involve four sources of variation, listed in (generally) increasing order of magnitude:

1. Measurement variation: variability among measurements taken on a single fragment at nearly the same time (i.e., only at most a few minutes apart), denoted by σ_e;
2. Time variation: variability among measurements taken on a single fragment at different times (e.g., on different days, perhaps as much as weeks apart), denoted by σ_t;
3. Fragment variability: variability in measurements taken on *different* fragments from the *same* pane of glass, denoted by σ_f;
4. Source variability: variability in measurements taken on samples from different panes of glass, denoted by σ_B.

Using data from Weis et al. (2011) (see Section 5), σ_e tends to be quite small, usually 1–4% (RSD) for most elements. Measuring the same fragment on different days suggests that $\sigma_t \approx$ 3–8% for all elements except ^{90}Zr (12%) and ^{178}Hf (13.6%). Because a measurement on a single fragment involves both sources of variation (to properly characterize the range of variability if it had been measured again and/or on another day), the root mean square of these two sources (i.e., $\sqrt{\sigma_e^2 + \sigma_t^2}$) is approximately 3–9%.

The underlying justification for using trace element concentrations in glass as forensic evidence rests with the idea that σ_B far exceeds σ_e, σ_t, and σ_f combined; that is, the procedure can correctly distinguish fragments that came from the *same* source from those that came from *different* sources. Error rates depend crucially on good estimates of the magnitudes of these sources of variation, along with the correlations on the measurements in *pairs* of elements. We discuss in Section 5 the data that we used for estimating these standard deviations and the pairwise correlations between them.

4.2. *Lognormal distributions for measured concentrations.* We assume initially that the log concentrations are normally distributed with a mean of μ and a variance of σ^2, where both μ and σ are estimated from available glass data sets. (Later, we will assume that the log concentrations have a heavier-tailed distribution, such as Student's t.) On this log scale, the SD (σ) for most elements is in the range of 0.02–0.06 (RSD of 2–6%), in accordance with Weis (2011), page 1281.

Because the ASTM standards recommend measuring 8–17 elements, and none of the three data sets that we analyze here has more than nine replicates per sample, Weis et al. (2011) dismiss the use of Hotelling's T^2 statistic, a multivariate version of Student's t, for assessing the significance of the measurement difference in elemental concentrations in two samples:

> "... at least 10 replicate measurements of both samples to be compared must be conducted for the Hotelling's T^2-test to be applicable. If only six replicate measurements are carried out for each of the two samples to be compared, the number of elements used for the comparisons has to be reduced to 10, which leads to a loss of evidential value. Hence, Hotelling's T^2-test calculations will not be addressed in this paper."

In fact, having fewer replicates than elements does *not* relieve us of the need for more replicates, because we still need to estimate the *correlations* in the measurements among the different elements. Forensic glass experts are well aware of the correlations among certain elements, based on their chemical properties.[6] The correlation (or covariance) matrix is used explicitly in Hotelling's T^2 statistic, but even if not used explicitly, knowing the correlations between each pair of elements removes the temptation to treat individual "match intervals" as independent (which they surely are not; see Section 5). We therefore resort to estimating variances and covariances among all pairs of elements in a given sample and pooling these estimates across all samples, for both measurement variation (1) and time variation (2). We denote these pooled covariance matrices as V_e and V_t, respectively. Furthermore, some of our data sets also allow for estimation of fragment variability (3), V_f. Hence, the difference in the concentrations between two fragments from the same glass pane can be expected to vary due to V_e, V_t, and V_f.

We denote by p the number of elements in each standard,[7] and model the logarithms of the p measured concentrations initially as Gaussian with mean μ and covariance matrix Σ. Thus, a vector X of p concentrations has a distribution that we will denote as $N_p(\mu, V)$, where $V = V_e + V_t + V_f$. We will see in Section 5 that the standard deviation measurements made on different fragments tend to be about 1.0–1.6 times larger than those on the same fragment on different days (i.e., on a per-element basis, $\sigma_f/\sigma_t \approx 1.0$–$1.6$), and about 1.5–4.5 times larger than those on the same fragment at the same time (i.e., on a per-element basis, $\sigma_f/\sigma_e \approx 1.5$–$4.5$). Because $\sigma_t \approx \sigma_f$ for many elements based on limited data, and because most trace element concentrations in glass evidence are measured on the same day, we will

[6]For example, the very high correlation between hafnium and zirconium is well known.

[7]The elements in the "signature" differ for each standard. Standard ASTM E2330-12 for ICP-MS recommends 14 elements: magnesium (Mg), aluminum (Al), iron (Fe), titanium (Ti), manganese (Mn), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba), lanthanum (La), cerium (Ce), neodymium (Nd), samarium (Sm), and lead (Pb). ASTM E2927-16 for LA-ICP-MS recommends all of the same except Sm, plus lithium (Li), potassium (K), calcium (Ca), and cerium (Ce) (17 elements). The standard for XRF is less specific; see ASTM E2926-13 Section 10.6.2.1 and discussion in Section 7.

consider only the effects of variation due to measurements and fragments [sources (1) and (3), respectively] in our modeling approach.

4.3. *Simulation strategy.* All three standards begin with two samples. For the R (or Q) sample, we simulate three measuements (the minimum number of replicates required by the ASTM standards) from $N_p(\mu, V^* = V_f^* + V_e^*)$, where μ is a vector of length p of all zeroes, and V_f^* and V_e^* are estimates of the between-fragment and within-fragment covariance matrices, respectively. (See Section 5.4 for a description of our estimates of V_f and V_e from the available data sets.) For the K sample, we simulate another three measurements, this time from $N_p(\mu+\delta, V^*)$, where δ is a vector of length p of differences in the means between the logarithms of the measurements. [A change of δ on the log scale corresponds to a relative difference in the means between the two fragments on the original scale of $\exp(\delta) - 1$. For example, if $\delta = 1.5$, then the means of an element of the K and R samples may be $\exp(3) = 20.1$ and $\exp(4.5) = 90.0$, for a relative difference of $(90.0 - 20.1)/20.1 = \exp(1.5) - 1 = 3.48$.] For our simulations, we will set values of δ, the absolute difference in the two means on the log scale or the relative difference on the original scale, and then count the proportion of our simulated samples that meet the "match" criterion. In this way, we know the *true* difference in the means, and calculate the "match rate" for a theoretical set of thousands of samples, not dependent on a particular set of collected samples.

Formally, the simulation proceeds as follows, for p elements (p will be 10–17, depending on the data set we use):

1. Set `matchcount` to 0.
2. Simulate two covariance matrices $\hat{V}_1$, $\hat{V}_2$ from a Wishart distribution, assuming V^* is the "true" covariance matrix. This takes into account the variability in estimating V^* from data.
3. Generate a sample of 3 (or 6, or 9) measurements from $N_p(0, \hat{V}_1)$, representing 3 (or 6, or 9) measurements of concentrations on p elements for the K fragment. Let $\bar{X} = (\bar{x}_1, \ldots, \bar{x}_p)$ and $S_x = (s_1, \ldots, s_p)$ represent the vector of means and standard deviations, respectively, for each of these p elements, and let $S_x^* = (s_1^*, \ldots, s_p^*)$ where each $s_i^* = \max(0.03, s_i)$.
4. Calculate the "match interval" for the ith element as $(\bar{x}_i - 4s_i^*, \bar{x}_i + 4s_i^*)$.
5. Generate another sample of 3 (or 6, or 9) measurements from $N_p(\delta, \hat{V}_2)$, representing 3 (or 6, or 9) measurements of concentrations on p elements for the R fragment. Let $\bar{Y} = (\bar{y}_1, \ldots, \bar{y}_p)$ represent the vector of means of for each of these p elements.
6. If $\bar{y}_i \geq \bar{x}_i - 4s_i^*$ and $\bar{y}_i \leq \bar{x}_i + 4s_i^*$ for each element $i = 1, \ldots, p$, then increase `matchcount` by 1.

We repeat steps 1–5 100,000 times for various values of δ between 0.00 (true matches) and 6.00 [relative change in means on raw scale is $\exp(6) - 1$]. Note

that $\delta = 0.3$ corresponds to a relative change in raw means of 35%, and $\delta = 0.5$ is a relative change in raw means of 65%. Note also that the "match rate" at $\delta = 0$ provides the probability of false exclusions; in fact if measurements typically vary 10–15% anyway, one may wish to consider samples whose means differ by no more than $\delta = 0.15$ (16% relative difference) as "indistinguishable," and consider the "match rate" at $\delta \leq 0.15$ the "false exclusion rate." We expect that the probability of a "match" as δ increases should fall to zero, because large differences in means should be increasingly easy to detect.

We repeat the steps above, but where the log(concentrations) come from a heavier-tailed distribution than the Gaussian. The family of t-distributions satisfies this purpose: t_{30} (30 degrees of freedom) is rather close to Gaussian, while t_3 (3 degrees of freedom) is considerably heavier-tailed. While these distributions look quite similar except for the tails, we will see in Section 6 that the effect of heavy-tailed distributions on the error rates is quite substantial.

5. Preliminary analyses: ICP-MS and LA-ICP-MS data. In this section we describe three data sets that formed the basis of our statistical modeling approach, along with some preliminary exploratory analyses.

5.1. *ICP-MS*. We obtained data from Florida International University (FIU) in which concentrations of 16 elements were measured on multiple glass samples via ICP-MS. The elements include 13 of the 14 cited in E2330-12 (with two isotopes of strontium, ^{86}Sr and ^{88}Sr) minus neodymium (Nd), plus antimony (^{121}Sb and ^{123}Sb), gallium (^{71}Ga), and hafnium (^{178}Hf). Each sample had three measurements, and the collection of 590 samples included seven types of glass: 160 Container glass samples, 189 Float Architecture, 46 Float Autowindow (CFS), 97 Float Autowindow (non-CFS), 45 Headlamp, 10 Laboratory, and 43 "Rare." Not all types of glass had elemental concentration measurements for all 16 elements.

Because the types of glass are so different, from container to decorative architectural to automotive, we chose to estimate covariances and correlations between elements separately for different glass types. Table 1 shows several pairs of elements with consistently high correlations across all types.

5.2. *LA-ICP-MS*: *Data Set* 1. Dr. Peter Weis (Bundeskriminalamt/Federal Criminal Police Office, Forensic Science Institute, KT 42—Inorganic Materials and Microtraces, Coatings, Wiesbaden, Germany) kindly shared the data that were published in Weis et al. (2011). Each fragment was measured six times, which allows for reliable estimates of within-fragment variability for all 20 elements that were measured. The elements include all 17 of the elements cited in E2927-16, plus sodium (^{23}Na), tin (^{118}Sn), and silicon (^{29}Si, as the constant standard). In this collection, data set (A) "Same" consisted of 33 fragments from the same pane of glass, plus a 34th fragment that was measured six times on each of 11 consecutive days, permitting rough estimates of between-fragment variability (among the 33

TABLE 1

*Robust correlations for FIU data (Italic: $0.7 \le |x| < 0.8$, **Bold:** $0.8 \le |x| \le 1$). *Float Auto (CFS) does not have measurements for La; all three Float glass types do not have measurements for Sb. †Lab has only 10 samples (and some missing values), not enough to calculate robust correlations. Classical correlation values are shown*

	Ce–La	**Ce–Sm**	**La–Sm**	**Mn–Sm**	**Ba–Mn**	**Ba–Sm**	**Mn–Ti**	**La–Mn**	**Sm–Ti**
Container	**0.98**	**0.92**	**0.94**	**0.83**	0.47	0.65	0.63	**0.83**	*0.74*
Float Arch*	**0.96**	**0.92**	**0.95**	*0.76*	*0.70*	**0.82**	*0.77*	*0.70*	0.43
Float Auto (CFS)*	—	0.37	—	**0.87**	**−0.83**	*−0.75*	*−0.77*	—	*−0.73*
Float Auto (non-CFS)*	**0.89**	**0.92**	**0.95**	**0.83**	**0.83**	**0.92**	*0.79*	**0.81**	**0.87**
Headlamp	**0.98**	**0.96**	**0.92**	−0.32	0.17	0.37	−0.23	−0.29	0.48
Lab†	**0.98**	**1.00**	**0.98**	*0.71*	**0.97**	**0.86**	**0.88**	**0.82**	**0.96**
Rare	**0.99**	**0.92**	**0.95**	0.42	**0.90**	*0.72*	0.54	0.41	*0.80*

fragments) and between-day variability (among the 11 days). We also can verify that the within-fragment variability (measurement variation among the six replicates) is consistent with the within-day variability (also six replicates each on 11 days). Set (B) "Different" consisted of 62 samples mostly from different sources, but included some from the same manufacturer and even the same production year but different batches; for example, Samples 10_01 and 11_01 were both produced in Flachglas' Gladbeck 1 plant and were clear and 3.8mm thick, but one was produced on 24 July 1994 and the other was produced on 11 Oct 1994. Thus, we are able to assess some degree of consistency in trace element concentrations for glass from the same, versus from different, manufacturers.

5.3. *LA-ICP-MS: Data Set 2*. Dr. David Ruddell (Centre of Forensic Sciences, Toronto, Canada) kindly shared the data from Dorn et al. (2015). The data from the "pane study" consisted of 48 "samples" taken from a single $4' \times 6'$ pane of glass: 24 fragments were cut from the glass pane and measured 9 times each; a 25th fragment was measured 24 times (see page 87, "*Group 1*," for details). The 23 elements measured include all 17 elements cited in E2927-16, plus silicon (^{29}Si), cobalt (^{59}Co), tin (^{118}Sn), antimony (^{121}Sb), thorium (^{232}Th), and uranium (^{238}U). These data can be used to corroborate the estimates of measurement variability found from Weis' 33 same-source samples as well as within-fragment variability (from the 25th fragment measured 24 times). Table 2 shows very good agreement (except for ^{7}Li) in the estimated standard deviations from these two data sets.

5.4. *Estimating parameters from data sets*. For all three data sets, Gaussian quantile-quantile plots suggest that most of the means for the different elemental log concentrations on different glass samples tend to be Gaussian, but high outliers are common.

TABLE 2
Within-fragment measurement variability in LA-ICP-MS measurements on 17 *elements* (*"elt"*) *using data from Canada* (*CAN*) *and Germany* (*GER*). *Standard deviations on log measurements* (*approximately relative standard deviations on raw scale*)

Elt	CAN	GER	Elt	CAN	GER	Elt	CAN	GER
^{7}Li	10.48	2.41	^{55}Mn	2.22	1.90	^{139}La	2.01	2.57
^{25}Mg	1.59	0.88	^{57}Fe	2.56	0.96	^{140}Ce	2.09	1.66
^{27}Al	1.60	2.54	^{85}Rb	2.61	2.10	^{146}Nd	2.65	3.39
^{39}K	1.57	1.32	^{88}Sr	1.76	1.65	^{178}Hf	3.27	4.06
^{42}Ca	1.37	1.35	^{90}Zr	2.24	2.84	^{208}Pb	4.57	2.55
^{49}Ti	1.77	1.74	^{137}Ba	2.77	2.39			

Ideally, we seek to estimate the covariance (or correlation) matrix among the p elements measured in the data set. Because no data set has at least (and preferably much more than) p replications, and because outliers in the data are typical, the usual Pearson correlation matrix cannot be calculated. We address this problem by using the two LA-ICP-MS data sets on which multiple measurements were taken *on the same piece of glass* on multiple occasions:

1. Weis et al. (2011): The authors of this paper measured sample 104G six times *on 11 separate days*, to assess day-to-day variability. If the day effect is absent, then one would have 66 measurements to estimate the 17×17 covariance matrix, from which one can develop realistic simulations (see Section 4 above).
2. Dorn et al. (2015): In this paper, the authors measured the same 3 cm $\times$ 3 cm piece of glass nine times *on* 24 *separate occasions*. Generally, three sets of nine measurements each were run in a single day. The 24 sets were denoted 1A, ..., 1D, 2A, ..., 2D, ..., 6D. If the occasion effect is absent, then we would have 210 measurements from which to estimate pairwise covariances and correlations. (Set 1C had only three measurements and was removed to keep the data set nicely balanced, so we have 207 measurements.) To our knowledge, the names of the data sets did not carry any information beyond the fact that measurements were taken in the order 1A, 1B, ..., 6D.

These two data sets are sufficiently large enough that we can estimate the pairwise correlations among the 17 elements listed in the LA-ICP-MS standard *if the effects of day (German) and occasion (Canadian) are absent*. In addition, because outliers in such numerous data sets are likely to occur (few data sets are completely error- or outlier-free), we compute a *robust* pooled estimate of the covariance matrix that downweights obviously discrepant observations. Fast MCD [minumum covariance determinent; see Rousseeuw and Van Dreissen (1999)] was computed for each data set using package MASS in R [Ripley (2015)].

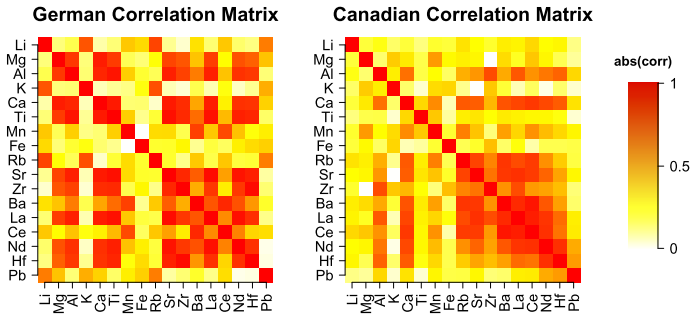

FIG. 1. *Classical correlation matrices for German and Canadian data sets. Robust versions using fast MCD are similar to those shown here.*

Figure 1 shows classical correlation matrices estimated from the German and Canadian data sets to be rather different.[8] Note that the German data set measured only the upper triangular half of a pane of glass, while the Canadian data measured an entire pane. This might account for some of the differences between correlation matrices and match rates (Figure 2).

6. Estimating error rates from populations.

6.1. *Canadian and German data simulations.* The following simulations use the covariance matrices estimated from the Canadian and German data sets as the "true" covariance V, from which we estimate the error rates as the proportion of times that the 4-SD approach (in ASTM E2927-16) fails to identify samples as "distinguishable" when in fact the true difference between the mean concentrations is specified as $\delta = 0.1, \ldots, 0.6$ on the log scale [$\exp(\delta) - 1$ on the relative means scale]. In each simulation run, we simulated not only the two sets of r (number of replicates) p-dimensional vectors (representing the lognormally distributed concentrations from the K and R fragments), but also generated a covariance matrix from a Wishart distribution with mean $(\mathrm{df}) \cdot V$, where df = degrees of freedom on which V is based.[9]

Because "REAL DATA OFTEN FAIL to be Gaussian IN MANY WAYS" [Brillinger and Tukey (1985), page 1020], our simulations generate samples from both the (optimistic) multivariate Gaussian distribution as well as from (heavier-tailed) multivariate t distributions with 3, 6, and 9 degrees of freedom.[10] Figure 2

[8]These are extremely similar to the robust correlation matrices, thus the classical correlations are shown and used in analyses. The discrepancy is slightly larger for the German data set, possibly due to different day/occasion effect or the difference in sample size.

[9]The simulated Wishart had df $= 40$ and 100 degrees of freedom for the covariance matrices estimated from the German and Canadian data sets, respectively, less than the nominal $66 - 17 = 49$ and $207 - 17 = 190$ degrees of freedom, to account for other (unknown) sources of variability.

[10]R package mvtnorm was used to sample from multivariate t distributions [Genz et al. (2016)].

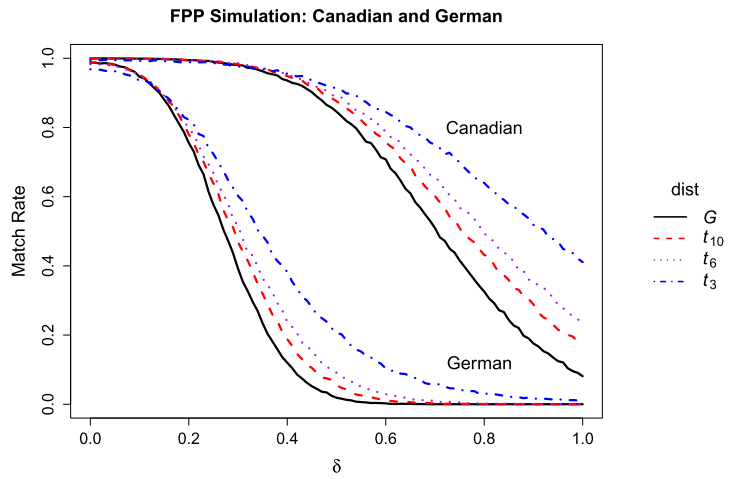

FIG. 2. *Match rates from Canadian and German simulations for data from four different distributions. δ gives the approximate relative change in means.*

shows two sets of curves for the match rates, one for each covariance matrix (top: Canadian; bottom: German). The lowest line in each set corresponds to match rates from samples generated from a multivariate Gaussian distribution, and the other three lines are analogous for three multivariate t distributions. To remain consistent with the actual data sets, the Canadian simulations used nine measurements while the German simulations used six measurements from each distribution.

The horizontal axis is δ, the "known" difference (set in the simulation) in elemental concentration for all elements (log scale), and the vertical axis shows the match rate. Ideally, we want to see high match rates at low values around $\delta = 0$, and decreasing match rates as δ increases. However as the plot indicates, the match rates do not decrease quickly. For example, two samples that come from batches whose mean log concentrations differ by $\delta = 0.5$ (i.e., a rather large ratio of mean raw concentrations of 1.65) in all 17 elements would not be "considered distinguishable" (ASTM E2927-16, Section 11.1.7) 84.6%–91.3% of the time (lognormal-t_3) using the estimate of the covariance matrix from the Canadian data. The rates are much lower using the estimate from the German data (1.9%–20.6%), but they are still rather high (39.2%–60.2%) when $\delta = 0.3$ (ratio of mean raw concentrations is 1.35). The match rates are very different depending on the data set used to estimate the covariance matrix, demonstrating the importance of collecting much more "same-fragment" data from different laboratories and on different types of glass. Table 3 provides these simulated match rates for specific values of δ between 0 ("same," where we expect very high match rates) to 0.8 (mean raw concentrations more than double).

TABLE 3
Canadian and German data simulation match rates at various δ

(δ)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
					(a) Canadian data match rates				
t_3	0.994	0.992	0.989	0.976	0.956	0.913	0.845	0.750	0.640
t_6	0.999	0.999	0.996	0.986	0.952	0.888	0.790	0.657	0.493
t_{10}	1.000	0.999	0.995	0.983	0.948	0.876	0.756	0.602	0.432
G	0.999	0.998	0.994	0.981	0.936	0.846	0.708	0.509	0.326
					(b) German data match rates				
t_3	0.968	0.934	0.822	0.602	0.383	0.206	0.104	0.059	0.030
t_6	0.983	0.950	0.797	0.516	0.238	0.091	0.029	0.011	0.002
t_{10}	0.986	0.953	0.780	0.468	0.188	0.055	0.012	0.002	0.000
G	0.988	0.946	0.755	0.392	0.120	0.019	0.002	0.000	0.000

6.2. *Simulations using Canadian covariance matrix.* Further simulations were conducted with the Canadian covariance matrix, which not only seemed more stable but also had more observations for estimating the covariance matrix V. The figures and tables below will use the covariance matrix containing all 17 elements. However, we noticed tremendous set-to-set variation in the 23 sets of nine replicate measurements of ^{39}K (potassium) and ^{57}Fe (iron). We understand that the huge variation may be due to (i) the ubiquity of ^{39}K and ^{57}Fe in the surrounding environment which may be present at different levels on different days resulting in varying levels of contamination; and (ii) interference from the plasma. Consequently, many forensic scientists ignore them for casework. However, both remain listed in the standard, so we simulated error rates for both $p = 17$ elements and $p = 15$ elements (all but ^{39}K and ^{57}Fe).[11]

Figure 3 analyzes the effect of sample size on the match rates at various levels of n, where n represents the multiplier in the "n-SD" approach. The ASTM standards require a minimum of three measurements, the Canadian data set had nine, and 12 would be slightly more than the minimum required to perform Hotelling's T^2 test. Figure 3 shows match rates using a sample size of three to be considerably different than nine and 12. The points at $n = 4$ correspond to the 4-SD approach, and are shown for Gaussian data (top row) and t_3 distributed data (bottom row). Table 4 shows the match rates at $n = 4$ (current choice for ASTM standard E2927-16) for various δ.

To compare performances of Hotelling's T^2 and the ASTM 4-SD approach, data were generated with sample sizes of 12 and 20. The bottom pairs of lines in Figure 4 are match rates using Hotelling's T^2, and the top two pairs correspond to the 4-SD approach. Albeit more complicated than the 4-SD approach, Hotelling's

[11] Figures for $p = 15$ are available in a supplementary file [Pan and Kafadar (2018)].

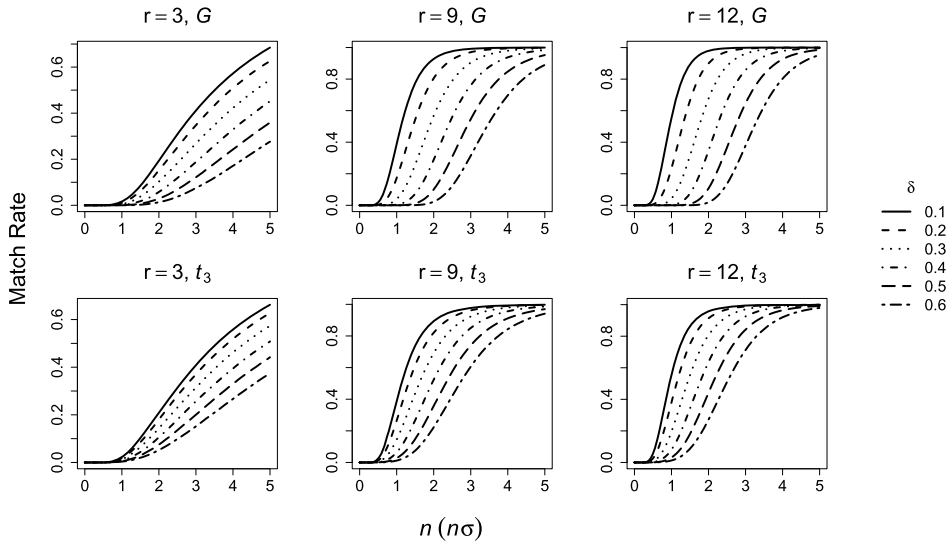

FIG. 3. *Simulation match rates using the Canadian covariance matrix by sample size and distribution for six δ values with $n = 0, \ldots, 5$.*

T^2 provides considerably lower match rates at larger values of δ; that is, much higher chances of claiming "distinguishable" samples that truly differ in their mean concentrations.

Figure 5 overlays 95%, 99.8%, and 99.9% t confidence intervals onto Figure 4. The latter two ($r = 12$) overlap slightly with Hotelling's T^2 ($r = 20$), but are overall more conservative. These differ from the 4-SD method in that, as the name implies, the 4-SD method uses the standard deviation and not the standard error in calculation, resulting in higher match rates as n increases (see Figure 4) because the confidence level also increases when not normalizing by $\sqrt{n}$.

Lastly, Figure 6 shows receiver operating characteristic (ROC) curves for the n-SD simulation results, which can be viewed as plotting power versus Type I er-

TABLE 4
Match rates by sample size at $n = 4$ (G = Gaussian, $t_3 = t$ with $\mathrm{df} = 3$) for various δ

(δ)	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
3 G	0.572	0.509	0.422	0.328	0.242	0.171
3 t_3	0.558	0.519	0.462	0.394	0.330	0.267
9 G	0.999	0.994	0.980	0.938	0.848	0.697
9 t_3	0.993	0.990	0.979	0.956	0.911	0.843
12 G	1.000	0.999	0.996	0.978	0.926	0.803
12 t_3	0.998	0.997	0.994	0.986	0.964	0.920

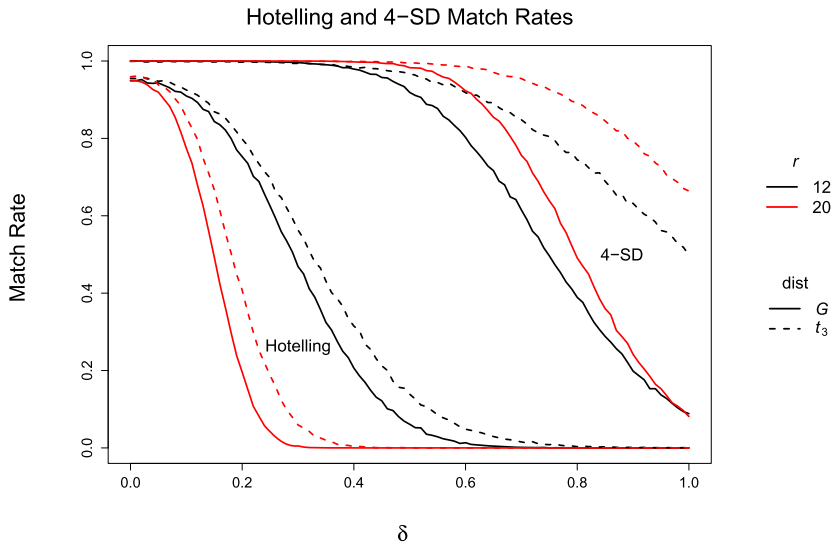

FIG. 4. *Match rates for Hotelling's T^2 vs. 4-SD approach for G and t_3 distributed data. Distribution has a much larger effect on the 4-SD approach.*

ror at various α. Because LA-ICP-MS measurement error is generally less than 5%, one might consider samples that differ by 5% or less "indistinguishable" and would want high match rates. Accordingly, the vertical axis plots the match rate of the procedure at $\delta = 0.05$ ("sensitivity") while the horizontal axis plots the (false) match rate when $\delta = 0.1, 0.2, 0.3, 0.6$, for $n = 0, \ldots, 5$, when the number of replicates is $r = 3$ (ASTM standard), 9 (Canadian data set), or 12 (minimum

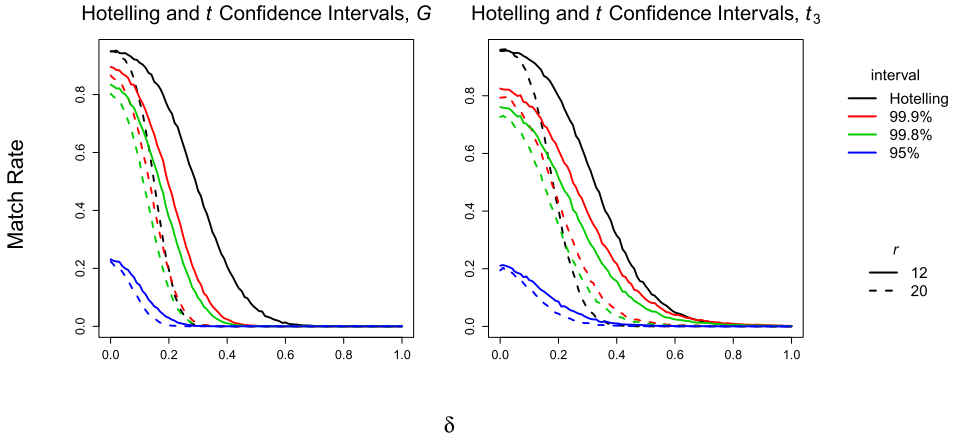

FIG. 5. *Match rates for Hotelling T^2 and t (95%, 99.8%, 99.9%) intervals.*

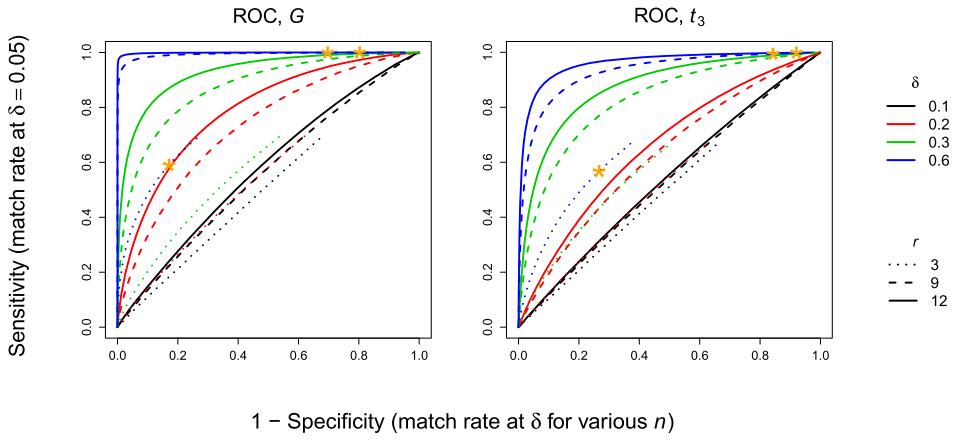

FIG. 6. *ROC curves for n-SD simulations using Canadian covariance matrix for four δ values with* $n = 0, \ldots, 5$. *The three orange stars denote the* $n = 4$ *point on each* $\delta = 0.6$ *curve for the three replicate levels.*

for Hotelling's T^2). The orange stars mark the $n = 4$ point on the $\delta = 0.6$ curves ($r = 3, 9, 12$).

Due to the popularity of the "n-SD" procedure, we simulated error rates using the exact same procedure, including using the maximum of the minimum and measured SD, but with different multipliers n, for $r = 3$ and $r = 6$. Assuming Gaussian data with 17 elements and $\delta = 0.1$, to ensure a match rate of 5%, $n = 1.29$ or $n = 0.74$ should be used ($r = 3$ or 6, respectively). For $\delta = 0.2$, the values of n are 1.41 or 0.88. These multipliers decrease under the t_3 assumption: $\delta = 0.1$: 1.21, 0.72 and $\delta = 0.2$: 1.28, 0.78, respectively. Figure 7 and Table 5 show detailed

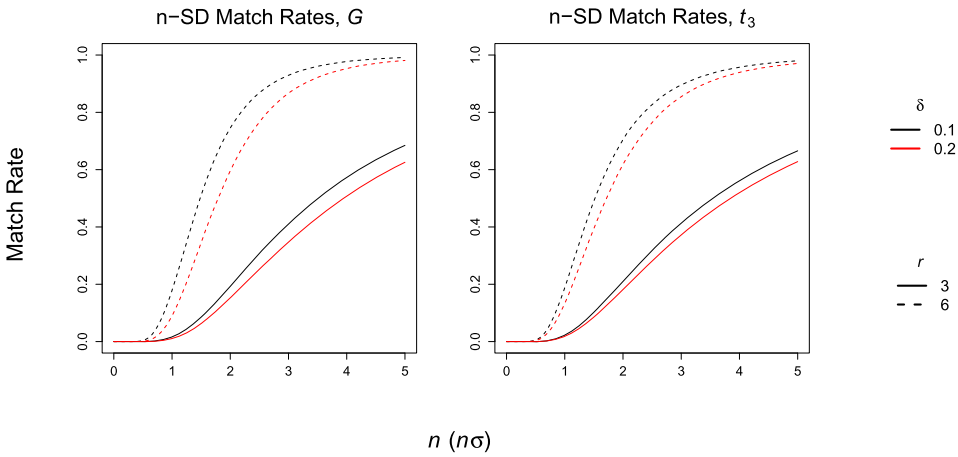

FIG. 7. *n-SD approach match rates where* $\delta = 0.1, 0.2$ *for* $n = 0, \ldots, 5$.

TABLE 5
n-SD approach match rates where $\delta = 0.1, 0.2$ for certain values of n

(*n*)	**0.50**	**0.75**	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**
					(a) $r = 3$				
$\delta = 0.1\ G$	0.000	0.003	0.016	0.044	0.086	0.136	0.193	0.251	0.307
$\delta = 0.1\ t_3$	0.000	0.005	0.022	0.056	0.102	0.155	0.210	0.264	0.318
$\delta = 0.2\ G$	0.000	0.002	0.011	0.031	0.063	0.105	0.153	0.202	0.253
$\delta = 0.2\ t_3$	0.000	0.004	0.018	0.046	0.085	0.133	0.182	0.231	0.281
					(b) $r = 6$				
$\delta = 0.1\ G$	0.004	0.053	0.179	0.346	0.511	0.645	0.745	0.817	0.868
$\delta = 0.1\ t_3$	0.006	0.061	0.188	0.344	0.490	0.611	0.705	0.775	0.827
$\delta = 0.2\ G$	0.001	0.021	0.090	0.207	0.345	0.479	0.596	0.691	0.766
$\delta = 0.2\ t_3$	0.003	0.041	0.133	0.261	0.394	0.515	0.618	0.699	0.764

match rates for some values of n. They allow us to determine the closest value of n to ensure a desired "false match rate" (say, 0.01) when the means of the samples really differ by $\delta = 0.10$ or 0.20 (10% or 22% relative change in means). For example, using the ASTM standard suggested $r = 3$, if the samples really do differ by 22%, we should use a "1-SD" approach to ensure a false match rate of no more than 1.8%; a 4-SD approach will have much higher "match rates."

6.3. *FIU simulations*. The FIU data set was divided into six main categories—Container, Float Architecture, Float Autowindow (CFS and non-CFS), Headlamp, Laboratory, and "Rare." Simulations were run using covariance matrices estimated from the first four categories. Only 10 samples were labeled "Laboratory" and the 43 samples labeled "Rare" consisted of a highly diverse collection (candlesticks, stained glass, etc.), so we did not attempt to estimate covariance matrices from these two sets of samples. In Figure 8, the higher set of four dashed lines used either 25 (Float Autowindow CFS, Headlamp) or 75 (Container, Float Architecture, Float Autowindow non-CFS) degrees of freedom when sampling from the Wishart distribution based on sample size, and the bottom set of four solid lines used a more conservative 10–13, or the number of elements, as the degrees of freedom.

The data tell us that Container and Float Architecture still contain a diverse population of glass samples. The samples labeled "Container" include bottles and jars of multiple types, the main groups of which can be separated into alcoholic beverages, other beverages, and other (including food). Samples from Float Architecture come from four main manufacturers: Cardinal, Guardian, PPG, and TempGlass. Figures 9 and 10 show that the correlation matrices for these are considerably different, which may explain the much larger match rates seen in these two categories—these populations actually consist of three or four different subpopulations. This mixture of subpopulations inflates the covariance matrix and results

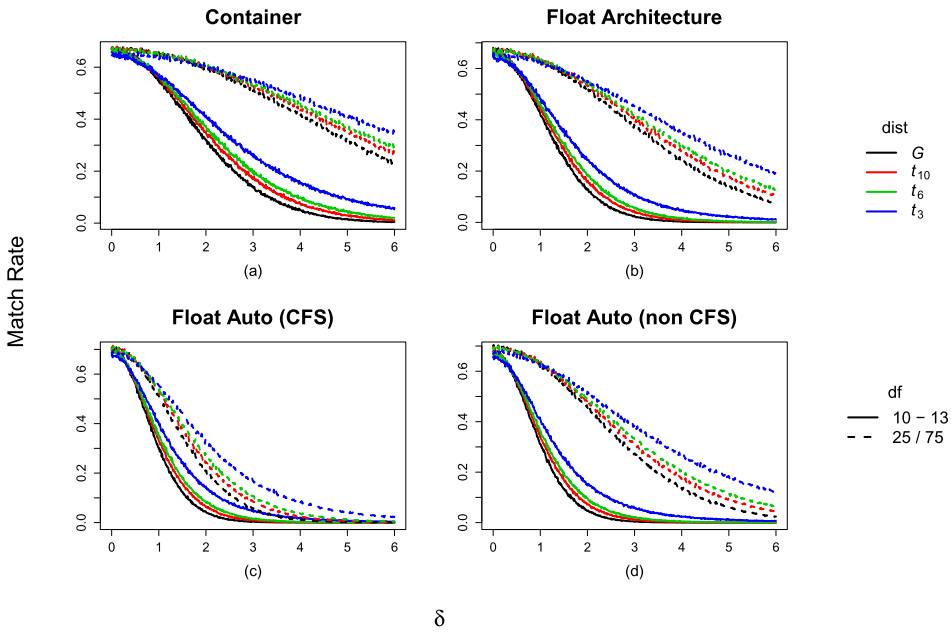

FIG. 8. *Match rates for FIU (ICP-MS) data for four different glass categories. The number of Wishart degrees of freedom by category are* (*solid*, *dashed*): *Container* (13, 75), *Float Architecture* (12, 75), *Float Autowindow non CFS* (12, 75), *Float Autowindow CFS* (10, 25), *Headlamp* (12, 25).

in more false matches in comparison to Figure 8(c), which is just one population with lower overall match rates.

The data set also contained a marker that split Float Autowindow into two groups: CFS (casework data from the Centre of Forensic Sciences in Canada) and non-CFS. Similar to the Container and Float Architecture data, Figure 11 clearly indicates two different correlation matrices, and as such match rates for the two groups are calculated separately. The match rates for Headlamp are not shown, but are similar to those in Figure 8(c).

7. ASTM standard E2926-13 for XRF. Conducting this study to evaluate the procedure in Section 10, "Calculation and Interpretation of Results" in ASTM 2926-13 for XRF, was more problematic for three reasons:

1. The standard recommends that the "signature" include six (or more) "peak intensity ratios": Ca/Mg, Ca/Ti, Ca/Fe, Sr/Zr, Fe/Zr, and Ca/K (calcium/manganese, calcium/titanium, calcium/iron, strontium/zirconium, iron/zirconium, calcium/potassium), rather than single elements.

2. The inference procedure from the "*Peak Intensity Ratio Comparisons*" in Section 10.6.2.1 is less specific; it states only examples of "ratios for evaluations" that can be considered: *"Ratios for evaluation* can include*: Ca/Mg, Ca/Ti, Ca/Fe,*

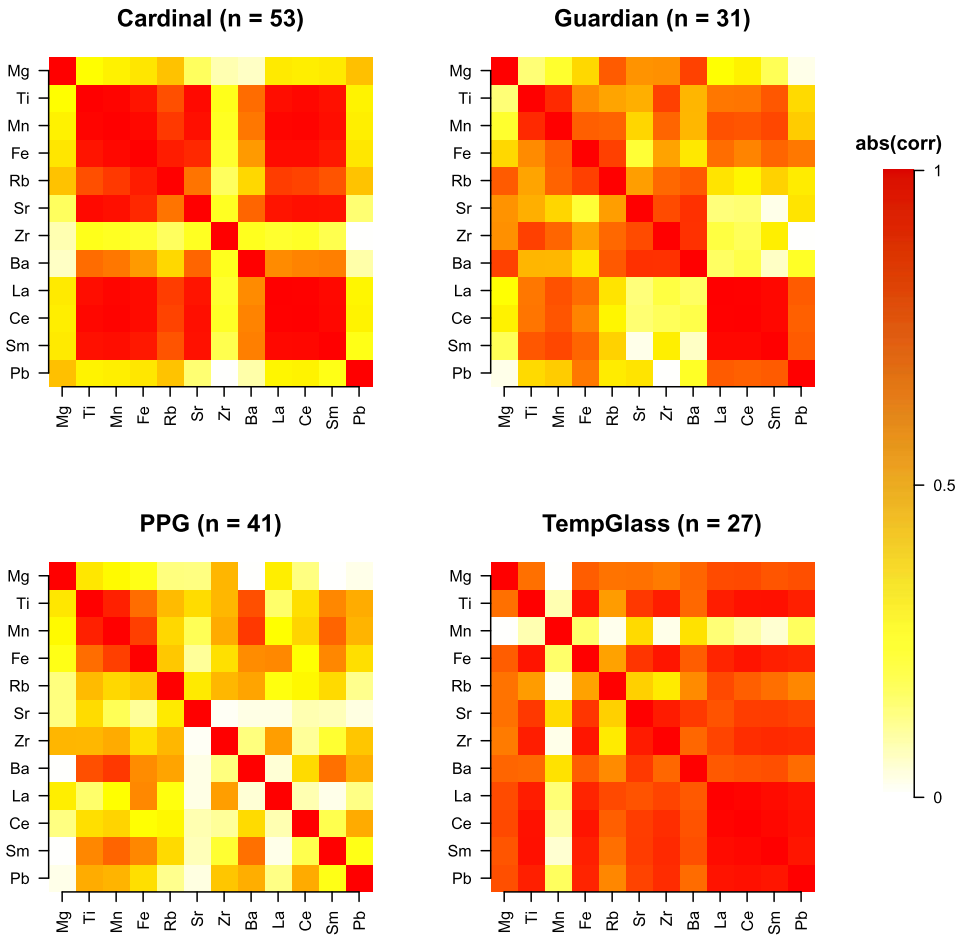

FIG. 9. *Float Arch correlations*: *Cardinal*, *Guardian*, *PPG*, *TempGlass*.

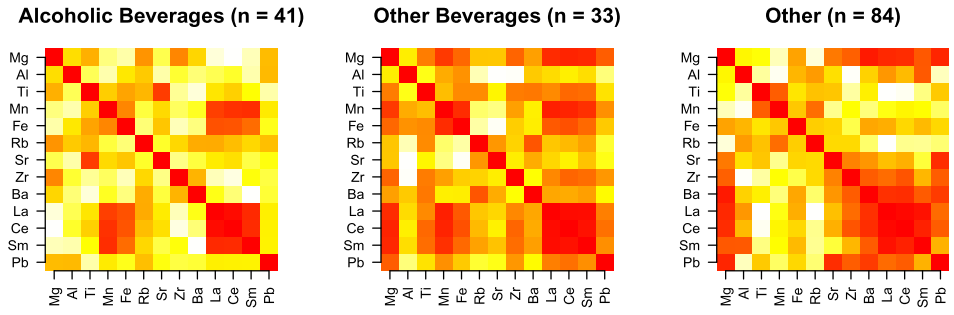

FIG. 10. *Container correlations*: *alcoholic beverages*, *other beverages*, *other*.

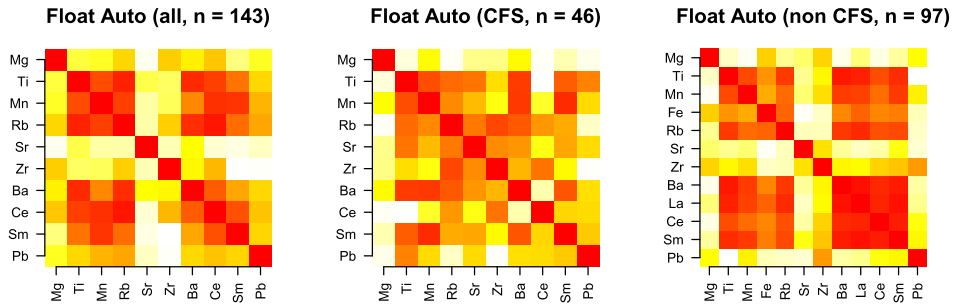

FIG. 11. *Float Autowindow*: *all*, *CFS*, *non-CFS*.

Sr/Zr, Fe/Zr, and Ca/K, if those elements are present above the limit of quantitation (LOQ). These peak intensity ratio comparisons have been shown to provide the best discrimination among different sources of soda-lime glasses. Additional ratios should be chosen *based on the elements present in the specimens*" (our emphasis). The *LOQ* is not further defined in the standard.

3. As with the elemental concentrations, ratios of peak areas are more suitably analyzed on a logarithmic scale, especially as the logarithm of the ratio of two peak areas becomes the difference in the logs of the two areas. See Kafadar and Eberhardt (1983, 1984).

For these reasons, and in the absence of publicly available data with XRF measurements on glass, we do not pursue further an evaluation of the ASTM E2926-13 standard (XRF).

8. Conclusions. We have calculated false positive probabilities (FPPs) for two ASTM procedures designed to determine whether forensic glass samples are "distinguishable" (or "indistinguishable") based on measured trace element concentrations. Using simulations based on estimates of means and covariances between elements from real data sets, the estimated FPPs are free from the specific characteristics of the samples in a study where error rates are calculated from all possible pairs of samples. That is, the estimated FPPs do not depend on a specific data set where the different samples may have very different concentrations (and therefore low estimated FPPs) or very similar concentrations (higher FPPs), because we estimate the FPPs for a wide range of possible concentrations. The approach extends one used to calculate FPPs for a very similar procedure that had been used by the FBI in its Compositional Analysis of Bullet Lead [discontinued in September 2005 following the NRC (2004) report]. We simulate distributions of trace element concentrations with a known difference δ_0 and evaluate the FPP as δ_0 varies from small to large, with many more simulation runs than exist in typical glass data bases. Dettman et al. (2014) also used a simulation procedure to estimate FPPs for a similar n-SD procedure on trace element concentrations in copper wire, but without simulating the covariance matrix with each run as we did here.

The three data sets that we used here (from Germany, Canada, and FIU) may well contain some different-sourced pairs that "match" using the ASTM standard's "4-SD interval" method, but such a comparison would not provide as complete an assessment of the method's performance as the extensive simulations that we conduct here. These three data sets led to very different covariance matrices, in part because the FIU measurements were made via ICP-MS (versus the German and Canadian measurements made via LA-ICP-MS), but also because of differences in the variety of the samples in the three data sets and because of individual laboratory protocols that can affect the measurement process of 17 elemental concentrations. Because a forensic glass analysis is likely to be conducted in a single laboratory, the importance of a reliable estimate of the covariance matrix and the FPP curve for that lab cannot be overstated.

There is little doubt that, when fragments come from very different sources [$\delta_0 = 3.0$; relative change in means is $\exp(3.0) - 1 = 19$], the ASTM procedures are very likely to declare the two samples as "distinguishable." The bigger challenge arises when two samples come from sources with a much smaller difference in relative means, say 3.5 ($\delta_0 = 1.5$). In this case, the probability of declaring the two samples as "distinguishable" is much lower; that is, the "match interval" criterion is satisfied sometimes 10–20% of the time.

The advantage of the "n-SD match interval" approach is its simplicity. For that reason, Table 5 enables one to choose a multiplier that will have a desired error rate (probability of failing to claim "distinguishable") when the true difference is at least δ_0 [where δ_0 on the log scale, or $\exp(\delta_0) - 1$ on the raw scale, is prespecified, depending on the expected level of change in means that one expects to see between fragments from the *same* pane]. In this paper, we focused on the probability of false inclusions ("Type I" error rate); alternatively, one can choose n to ensure a low false exclusion rate ("Type II" error). Either way, our population-based statistical modeling approach permits the study of this or alternative "match" procedures with this objective. Note that, even with a revised n-SD approach, the failure to claim "distinguishable" does *not* mean that the samples came from the same source, due to (potentially high) variability in the measured concentrations on the same piece of glass, and (potentially low) manufacturing variability (many samples may have "indistinguishable" concentrations in all elements). For example, companies constructing neighborhood tracts may purchase architectural float glass in bulk from a single manufacturer, resulting in many houses containing glass from possibly the same batch or batches produced on the same day. Using error rates determined from a diverse population could be extremely misleading given the recovered glass fragment could have come from one of many window panes in the neighborhood.[12] Thus, the probative value of glass evidence needs to be assessed in light of the size of the population that may have "indistinguishable" concentrations in all elements.

[12]We kindly thank Reviewer 2 for this example.

9. Final thoughts: Statistics, forensic science, criminal justice, and Stephen Fienberg. Steve Fienberg had a vast array of interests and a great influence on statistical theory and practice, much of which can be seen in this article. It was not only his research in multivariate statistics that influenced our approach to this problem. Steve advocated tirelessly for data sharing, because new data types and structures are key to driving research in statistics. Without the willingness of the forensic scientists mentioned in this article to share their data, this article could not have been written. When the data involved potentially sensitive variables, he helped develop methods for ensuring privacy and confidentiality. But most relevant to this topic, Steve was a real visionary for research needed to advance forensic science and ensure sound scientific methods, even before he chaired the National Academy of Sciences (NAS) Committee that issued the report *The Polygraph and Lie Detection* [NRC (2003)]. He continued to serve on many NAS Committees and Panels, and, when he didn't, often was part of the committees that reviewed the reports; many of these reports involved proper use of statistical evidence in the legal system. The second author especially benefited from his generous advice and his strategic leadership that led to the Center for Statistical Applications in Forensic Evidence (CSAFE), which he co-founded and which partially supported the present research. At the time of his death, his service on the National Commission on Forensic Science was acknowledged by a Certificate of Appreciation signed by Deputy Attorney General Sally Yates and NIST Director Willie May: "*In grateful appreciation of your unique contributions to the National Commission on Forensic Science, your lifelong public service and unwavering pursuit of science and justice as a pioneering statistician and applying scientific principles of great public importance, we hereby recognize your many accomplishments and achievements.*" We hope that this research will benefit both statistics and society in a way that Steve would appreciate.

Acknowledgments. We are grateful to Dr. David Ruddell, Centre for Forensic Science, Toronto, for sharing his laboratory's data and for many helpful discussions throughout our project. We also thank Dr. Peter Weis, Bundeskriminalamt/Federal Criminal Police Office, Forensic Science Institute, Wiesbaden, Germany, for providing an electronic version of data published in Weis et al. (2011). The FIU data were obtained from the Technical Support Working Group via Jeff Huber (`jeff.huber.ctr@cttso.gov`). We also are grateful to the anonymous reviewers who provided useful feedback that led to improvements in this article. The second author also thanks the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, for its hospitality during the program *Probability and Statistics in Forensic Science* which was supported by EPSRC Grant Number EP/K032208/1. Finally, we gratefully acknowledge the late Stephen Fienberg, whose life-long commitment to rigorous science in many disciplines of great public importance, including those arising in criminal justice and forensic science, inspired us to pursue this work.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical modeling and analysis of trace element concentrations in forensic glass evidence." (DOI: 10.1214/18-AOAS1180SUPP; .pdf). We provide additional plots of match rates under certain different simulation conditions.

REFERENCES

ASTM INTERNATIONAL (2012). *ASTM E2330-12 Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*. Retrieved from https://www.astm.org/Standards/E2330.htm. DOI:10.1520/E2330-12.

ASTM INTERNATIONAL (2013). *ASTM E2926-13 Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*. Retrieved from https://www.astm.org/Standards/E2926.htm. DOI:10.1520/E2926.

ASTM INTERNATIONAL (2016). *ASTM E2927-16e1 Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons*. Retrieved from https://www.astm.org/Standards/E2927.htm. DOI:10.1520/E2927-16E01.

BRILLINGER, D. R. and TUKEY, J. W. (1985). Spectrum analysis in the presence of noise: Some issues and examples. In *The Collected Works of John W. Tukey II. Time Series*: 1965–1984 (D. R. Brillinger, ed.) 1001–1141. Wadsworth, Monterey, CA.

DETTMAN, J. R., CASSABAUM, A. A., SAUNDERS, C. P., SNYDER, D. L. and BUSCAGLIA, J. (2014). Forensic discrimination of copper wire using trace element concentrations. *Anal. Chem.* **86** 8176–8182.

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. DOI:10.1080/00085030.2015.1019224.

GABEL-CINO, J. (2017). Expert witnesses and lawyers: Can we all get along? Presentation to the Second Annual Conference of the National Center for Forensic Science, Orlando, Florida, October 17, 2017.

GENZ, A., BRETZ, F., MIWA, T., MI, X., LEISCH, F., SCHEIPL, F., BORNKAMP, B., MAECHLER, M. and HOTHORN, T. (2016). Multivariate normal and t distributions, R package mvtnorm. R package version 1.0-5.

GIANNELLI, P. C. (2010). Comparative bullet lead analysis: A retrospective (September 1, 2011). *Crim. Law Bull.* **47** 306. Case Legal Studies Research Paper No. 2011-21.

KAFADAR, K. and EBERHARDT, K. R. (1983). Statistical analysis of some gas chromatographic measurements. *NBS J. Res.* **88** 37–46.

KAFADAR, K. and EBERHARDT, K. R. (1984). Some basic statistical methods for chromatographic data. In *Advances in Chromatography*, *Chapter* 1 (J. C. Giddings, E. Grushka, J. Cazes and P. R. Brown, eds.) **24** 1–34. Dekker, New York.

KOONS, R. D. (2003). Personal communication to K. Kafadar.

KOONS, R. D. and BUSCAGLIA, J. A. (2001). Interpretation of glass composition measurements: The effects of match criteria on discrimination capability. *J. Forensic Sci.* **47** 505–512.

KOONS, R. D. and BUSCAGLIA, J. (2005). Forensic significance of bullet lead compositions. *J. Forensic Sci.* **50** 341–351.

NATIONAL RESEARCH COUNCIL (2003). *The Polygraph and Lie Detection* (*Committee to Review the Scientific Evidence on the Polygraph*, *Division of Behavioral and Social Sciences and Education*). The National Academies Press, Washington, DC. DOI:10.17226/10420.

NATIONAL RESEARCH COUNCIL (2004). *Forensic Analysis*: *Weighing Bullet Lead Evidence* (K. O. MacFadden, Chair). The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward* (The Honorable H. T. Edwards and C. Gatsonis, Co-Chairs). The National Academies Press, Washington, DC. Available at http://books.nap.edu/catalog.php?record_id=12589.

PAN, K. D. and KAFADAR, K. (2018). Supplement to "Statistical modeling and analysis of trace element concentrations in forensic glass evidence." DOI:10.1214/18-AOAS1180SUPP.

RIPLEY, B. (2015). MASS: Support functions and datasets for venables and Ripley's MASS. R package version 7.3-45.

ROUSSEEUW, P. and VAN DREISSEN, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

SPIEGELMAN, C. H. and KAFADAR, K. (2006). Data integrity and the scientific method: The case of bullet lead data as forensic evidence. *Chance* **19** 17–26 (with discussion). MR2247019

TREJOS, T., KOONS, R., WEIS, P., BECKER, S., BERMAN, T., DALPE, C., DUECKING, M., BUSCAGLIA, J., ECKERT-LUMSDON, T., ERNST, T., HANLON, C., HEYDON, A., MOONEY, K., NELSON, R., OLSSON, K., SCHENK, E., PALENIK, C., POLLOCK, E. C., RUDELL, D., RYLAND, S., TARIFA, A., VALADEZ, M., VAN ES, A., ZDANOWICZ, V. and ALMIRALL, J. (2013). 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. DOI:10.1039/c3ja50128k.

WEIS, P., DÜCKLING, 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.

DEPARTMENT OF STATISTICS
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE, VIRGINIA 22904-4135
USA
E-MAIL: kdp4be@virginia.edu
kk3ab@virginia.edu