The Annals of Applied Statistics
2017, Vol. 11, No. 4, 2332–2356
https://doi.org/10.1214/17-AOAS1080
© Institute of Mathematical Statistics, 2017

AUTOMATIC MATCHING OF BULLET LAND IMPRESSIONS

BY ERIC HARE[1], HEIKE HOFMANN AND ALICIA CARRIQUIRY

Iowa State University

In 2009, the National Academy of Sciences published a report questioning the scientific validity of many forensic methods including firearm examination. Firearm examination is a forensic tool used to help the court determine whether two bullets were fired from the same gun barrel. During the firing process, rifling, manufacturing defects, and impurities in the barrel create striation marks on the bullet. Identifying these striation markings in an attempt to match two bullets is one of the primary goals of firearm examination. We propose an automated framework for the analysis of the 3D surface measurements of bullet land impressions, which transcribes the individual characteristics into a set of features that quantify their similarities. This makes identification of matches easier and allows for a quantification of both matches and matchability of barrels. The automatic matching routine we propose manages to (a) correctly identify land impressions (the surface between two bullet groove impressions) with too much damage to be suitable for comparison, and (b) correctly identify all 10,384 land-to-land matches of the James Hamby study (Hamby, Brundage and Thorpe [*AFTE Journal* **41** (2009) 99–110]).

1. Introduction. Firearm examination is a forensic tool used to help the court determine whether two bullets were fired from the same gun barrel. This process has broad applicability in terms of convictions in the United States criminal justice system. Firearms identification has long been considered an accepted and reliable procedure, but in the past ten years has undergone more significant scrutiny. In 2005, in *United States versus Green*, the court ruled that the forensic expert could not confirm that the bullet casings came from a specific weapon with certainty, but could merely "describe" other casings which are similar. Further court cases in the late 2000s expressed caution about the use of firearms identification evidence [Giannelli (2011)].

In 2009, the National Academy of Sciences published a report [National Research Council (2009)] questioning the scientific validity of many forensic methods including firearm examination. The report states that "[m]uch forensic evidence—including, for example, bite marks and firearm and toolmark

Received February 2017; revised June 2017.

[1]Supported in part by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California, Irvine, and University of Virginia.

Key words and phrases. 3D topological surface measurement, data visualization, machine learning, feature importance, cross-correlation function.

identification—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."

Rifling, manufacturing defects, and impurities in a barrel create striation marks on the bullet during the firing process. These marks are assumed to be unique to the barrel, as described in a 1992 AFTE article [AFTE Criteria for Identification Committee (1992)]. "The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the *unique surface contours* of two toolmarks are in sufficient agreement." The article goes on to state that "Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges, and furrows."

From a statistical standpoint, identification of the gun that fired the bullet(s) requires that we compare the probabilities of observing matching striae under the competing hypotheses that the gun fired, or did not fire, the crime scene bullet. If indeed the uniqueness assumption is plausible, the latter probability approaches zero and the former approaches (but never reaches) one.

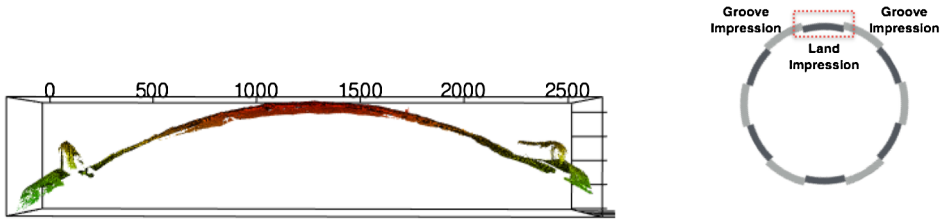
Current firearm examination practice relies mostly on visual assessment and comparison of striation. Indeed, the AFTE Theory of Identification (https://afte.org/about-us/what-is-afte/afte-theory-of-identification) explicitly requires that examiners evaluate the strength of similarity between two samples relative to other comparisons they may have carried out in the past. An attempt to quantify the degree of similarity consists in counting the number of consecutively matching striae (CMS) between two bullets, first proposed by Biasotti (1959). This approach has two drawbacks, however. First, determining matching striae is still a subjective activity. Second, as discussed by Miller (1998), the number of CMS may be high even if the bullets were not fired by the same gun.

Here, we focus on the question of defining a metric that can be used to objectively compare two bullets. We propose a framework which allows for the automatic analysis of the surface topologies of bullets, and the transcription of the individual characteristics into a set of features that quantify their similarities. This allows for an objective and quantitative assessment of striae-based bullet matches.

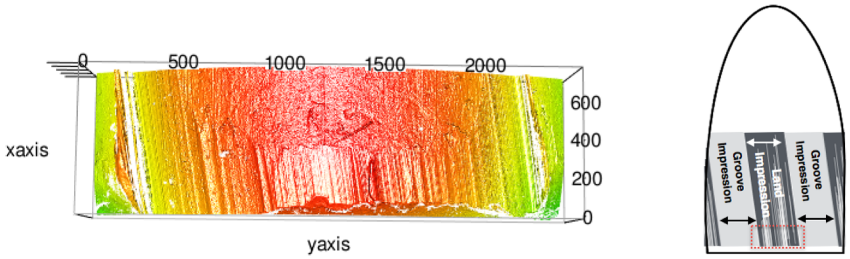
We work with images from the James Hamby Consecutively Rifled Ruger Barrel Study [Hamby, Brundage and Thorpe (2009)]. Ten consecutively rifled Ruger P-85 pistol barrels were obtained from the manufacturer and fired to produce 20 known test bullets and 15 unknown bullets for comparison. 3D topographical images of each bullet were obtained using a NanoFocus lens at 20x magnification and made publicly available on the NIST Ballistics Database Project[2] in a format called x3p (XML 3-D Surface Profile). The x3p format conforms to the ISO5436-2 standard[3] and is implemented to provide a simple and standard conforming way to

[2] http://www.nist.gov/forensics/ballisticsdb/hamby-consecutively-rifled-barrels.cfm

[3] http://sourceforge.net/p/open-gps/mwiki/X3p/



(a) View of the data along the circumference of the bullet (circular segment of about 30 degrees).



(b) Frontal view of a bullet land impression (lower end of the view is the bottom of the bullet).

FIG. 1. *Example of a groove-to-groove scan of a bullet land impression. The red-dotted rectangle on the right shows the location and orientation of the segment.*

exchange 2D and 3D profile data. It was adopted by the OpenFMC (Open Forensic Metrology Consortium[4]), a group of academic, industry, and government firearm forensics researchers whose aim is to establish best practices for researchers using metrology in forensic science. We have developed an open-source package for analyzing bullet land impressions written in R [R Core Team (2016)]. This package is called bulletr [Hofmann and Hare (2016)] and enables direct reading and manipulation of x3p files. It also implements all of the methods we propose in this paper. A different package exists for reading x3p files called x3pr [OpenFMC (2014)], developed by Petraco (2014), but it is not designed to carry out calculations like the ones we propose after the x3p files have been read.

Each fired bullet is provided in the form of a set of six x3p files, where each file is a surface scan between adjacent groove impressions on the bullet, called a "land" or land impression. The shoulders are the raised portion of the surface between the groove impression and the land impression. In the Hamby data, typical length (shoulder-to-shoulder) of a land impression is about 1998.28 micrometers or 2 millimeters. For notational simplicity, we refer to a particular land impression of a bullet as bullet X-Y, where X is the bullet identifier, and Y is the land number. An example of plotting one of these land impressions is given in Figure 1(a) and (b). These figures show side and top profiles of the land, respectively. The tilt of the lines to the left in Figure 1(b) is not an artifact, but a direct and expected consequence of the spin induced by the rifling during the firing process. Depending on

[4]http://www.openfmc.org/

whether a barrel is rifled clockwise or counterclockwise, the striations have a left or right tilt. The direction of the rifling is a class characteristic, that is, a feature that pertains to a particular class of firearms, and is not unique at the individual barrel/bullet level.

The typical number and width of striation markings on bullets varies significantly depending on the gun barrel. For instance, a Smith and Wesson barrel with a land-width of 2.4 millimeters contained an average 60 striae, with an average width of about 0.08 millimeters [Chu et al. (2011)].

The purpose of our paper is to present an automatic matching routine that allows for a completely objective assessment of the strength of a match between two bullet land impressions. While we assess the performance of the algorithm in terms of a binary decision of match versus nonmatch using a 50% probability cut-off, our primary goal is to highlight the features that are statistically associated with matches and nonmatches, and to provide a quantitative assessment of this association. In a real-world application of our algorithm, the raw scores would need further analysis and scrutiny, and it is likely that a 50% cut-off would be an inappropriate choice on the basis of reasonable doubt.

Our algorithm is fully open source and available on GitHub [Hofmann and Hare (2016)]. This transparency allows for a greater understanding of the individual steps involved in the bullet matching process, and allows other forensic examiners, as well as outside observers, to examine the factors that discriminate between known bullet matches and nonmatches. We have chosen to perform the matching on a land-to-land level, rather than bullet-to-bullet level. Although doing so introduces an implicit assumption of independence between land impressions, assuming independence only serves to make the task more challenging.

The remainder of this paper is structured as follows: We first briefly review some earlier work. We then discuss two methods for modeling the class structure of the bullet surfaces. Finally, we proceed to describing an automatic matching routine which we evaluate on the bullets made available through the Hamby study.

2. Previous work. There have been attempts to develop automatic or semi-automatic matching protocols, but most have focused on breech face and firing pin marks [e.g., Riva and Champod (2014)] or discuss a single attribute for comparison [e.g., Chu et al. (2011), Vorburger et al. (2011)]. Still others refer to proprietary algorithms [Roberge and Beauchamp (2006)]. We briefly review some of this earlier work in what follows.

The original paper on the complete Hamby study already reports the successful use of several computer-assisted methods. However, aside from a zero false positive rate, false-negative error rates for bullets are not given nor are error rates for land-to-land matches mentioned.

Lock and Morris (2013) proposed an approach to quantify similarity of toolmarks. Their algorithm determines an optimal matching window between two toolmark signatures, and then performs a set of both coordinated and independent

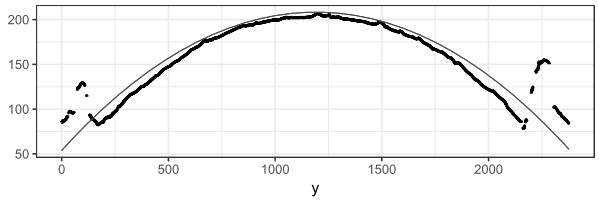


FIG. 2. *Side profile of the surface measurements (in μm) of a bullet land impression at a fixed height x. Note that the global features dominate any deviations, corresponding to the individual characteristics of striation marks.*

shifts. Given a match, the coordinated shifts would be expected to yield correlation values higher than those obtained from independent shifts. This is assessed using a Mann–Whitney U-Statistic.

A procedure for bullet matching using the BulletTrax3D system is described in Roberge and Beauchamp (2006). Their study used a different set of ten consecutively rifled barrels; matches are identified based on a bullet-to-bullet correlation score. The authors state that this process "could be automated," but no implementation of the algorithm is available.

Modern automated techniques using 3D images have also been proposed by Riva and Champod (2014). However, the authors focused on cartridge cases and not bullets. This might seem like a trivial distinction, but it has implications for the development of the algorithm. Their algorithm performs alignment of striae by rotation of the XY plane, which is not generalizable to bullets in which the XY plane is not flat.

Other work on 3D images has been described by Petraco and Chan (2012), who also focus on cartridge cases, as well as screwdriver striation patterns, and by others [e.g., Chu et al. (2010, 2011), Vorburger et al. (2011)].

3. Bullet signatures. To analyze the striation pattern, we extract a *bullet profile* [Ma et al. (2004)] by taking a cross section of the surface measurements at a fixed height x along the bullet land impression, as previously illustrated in Figure 1. Figure 2 shows a plot of the side profile of a bullet land impression. It can be seen that the global structure of the land dominates the appearance of the plot. The shoulders can be clearly identified on the left and right side, and the curvature of the surface is the most visible feature in the middle.

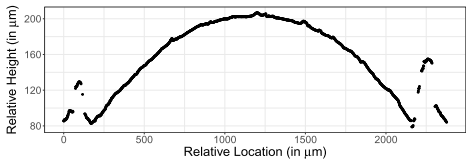
The smooth curve on the plot represents a segment of a perfect circle with the same radius as the bullet. While the circle is an obvious first choice for fitting the structure, it does not completely capture the bullet surface after it was fired. A discussion of a circular fit and the remaining residual structure can be found in the Supplementary Material, Section 1 [Hare, Hofmann and Carriquiry (2017)].

Instead of a circular fit, we use multiple loess fits to model the overall structure and extract the bullet markings.

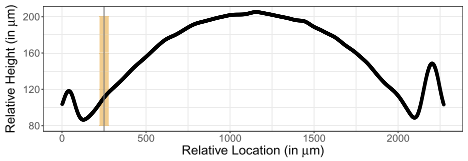
3.1. *Identifying shoulder locations.* We first identify the location of the left and right shoulders in the image. The groove impressions are assumed to contain no information relevant for determining matches, and the shoulders identify the location at which the land impression begins. The shoulders also dominate the structure and, therefore, need to be removed. Fortunately, the location and appearance of the shoulders in the surface profiles is quite consistent. Surface measurements reach local maxima around the peak of the shoulder at either end of the range of y, and we can then follow the descent of the surface measurements inwards to the valley of the shoulder. The location of the valleys mark the points at which we trim the image. The procedure can be described as follows:

1. At a fixed height x, extract a bullet's profile [Figure 3(a), with $x = 243.75$ μm].

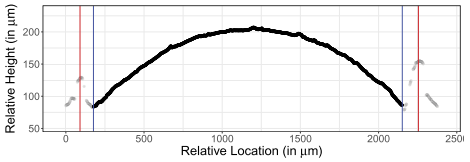
2. For each y value, smooth out any deviations occurring near the minima by twice applying a rolling average with a pre-set *smoothing factor* s. [Figure 3(b), smoothing factor $s = 35$ (data points) corresponding to 55 μm].



(a) Step 1 of identifying shoulder locations: For a fixed height ($x = 243.75$ μm) surface measurements for bullet 1-5 are plotted across the range of y.



(b) Step 2 of identifying shoulder locations: The surface measurements are smoothed twice with a smoothing factor of $s = 35$. The orange rectangle shows an example of the smoothing window. Valleys and peaks are detected, if they are not within the same window.



(c) Steps 3–6 of identifying shoulder locations: After smoothing the surface measurements extrema on the left and right are detected (marked by vertical lines, red indicating peaks and blue indicating valleys). Values outside the blue boundaries are removed (shown in grey).

FIG. 3. *Overview of all six steps of the smoothing algorithm to identify and remove shoulders and groove impressions from the bullet images.*

3. Determine the location of the peak of the left shoulder by finding the first doubly-smoothed value y_i that is the maximum within its smoothing window (e.g., such that $y_i > y_{i-1}$ and $y_i > y_{i+1}$, where i is between 1 and $\lfloor s/2 \rfloor$). We call the location of this peak p_ℓ [see Figure 3(c)].

4. Similarly, determine the location of the valley of the left shoulder by finding the first double-smoothed y_j that is the minimum within its smoothing window. Call the location of this valley v_ℓ.

5. Reverse the order of the y values and repeat the previous two steps to find the peak and valley of the right shoulder, (p_r, v_r).

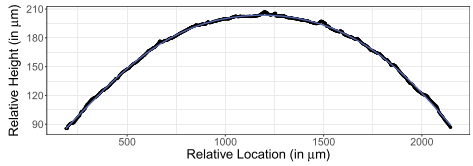
6. Trim the surface measurements to values within the two shoulder valleys (i.e., remove all records with $y_i < v_\ell$ and $y_i > v_r$) [see Figure 3(c)].

The smoothing factor s introduced in the algorithm represents the window size to use for a rolling average. Higher values of s therefore lead to more smoothing. Empirically, a value of $s = 35$ for the smoothing factor seems to work well (the smoothing factor is further discussed in Section 4.4). It is important to note that the smoothing pass is done twice, that is, the smoothed data are once again smoothed by computing a new rolling average with the same smoothing factor. This bears some similarities to the ideas of John Tukey in his book Exploratory Data Analysis, where he describes a smoothing process called "twicing" in which a second pass is made on the residuals computed from the first pass and then added back to the result [Tukey (1977)]. This has the effect of introducing a bit more variance back into the smoothed data. We instead performed a second smoothing pass on the smoothed data, which has the effect of weighting observations near the center of the window the highest, with the weights linearly dropping off as we reach either end of the smoothing window.

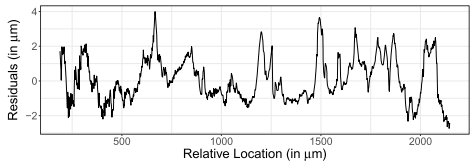
3.2. *Removing curvature.* Next, we fit a loess regression to the data. Loess regression [Cleveland (1979)] is based on the assumption that the relationship between two random variables X and Y can be described in the form of a smooth, continuous function f with $y_i = f(x_i) + \varepsilon_i$ for all values $i = 1, \ldots, n$. The function f is approximated via locally weighted polynomial regressions. Parameters of the estimation are α, the proportion of all points included in the fit (here, $\alpha = 0.75$), the weighting function and the degree of the polynomial (here, we fit a quadratic regression).

The main idea of locally weighted regression is to use a weighting routine that emphasizes the effect of points close to the fitting location and n-emphasizes the effect of points as they are further away. The weighting function used here is the tricubic function $w(d) = (1 - d^3)^3$, for $d \in [0, 1]$ and $w(d) = 0$ otherwise. Here, d is defined as the distance between x_i and the location of the fit x_o, divided by the overall range of the included data, so as to map the results to a $[0, 1]$ range as the definition requires.

Figure 4(a) shows the loess fit, in blue, overlaid on the processed image of bullet 1-5. The fit seems to do a reasonable job of capturing the structure of the



(a) Loess fit for bullet 1-5.



(b) Residuals of loess fit for bullet 1-5.

FIG. 4. *Fit and residuals of a loess fit to bullet* 1-5 (*Barrel* 1). *The residuals define the signature of bullet* 1-5.

image. Figure 4(b) shows the residuals from this fit. These residuals are called the *signature* of bullet 1-5.

4. Automatic matching. Applying the loess fit to a range of different signatures (see Figure 5 for signatures extracted at heights between 50 μm and 150 μm) shows the 3D striation marks from two bullets. Signatures of bullet 1 are shown on the left (all extracted from heights below 100 μm) and signatures of bullet 2 are shown on the right (extracted at heights above 100 μm). Signatures are manually aligned, resulting in many of the striation marks to continuously pass from one side to the other. Visually, this allows for an easy assessment of these two bullet land impressions as a match. However, this match relies on visual inspection and is therefore subjective. The goal of this section is to eliminate the need for a visual inspection during the matching process and replace it by an automatic algorithm. This also allows for a quantification of the strength of the match.

In this section, we describe the algorithm for matching signatures first, and the impact of parameter choices in the subsections thereafter.

4.1. *Algorithm*. Figure 6 gives an overview of the automated matching routine: We first identify a stable region for each bullet land and extract the signature at the lowest height in this region, because typically, individual characteristics are best expressed at the lower end of the bullet (see the Supplementary Material, Section 3, for a more detailed discussion).

All of the other steps are done on pairs of bullet land impressions:

1. *Smooth the two signatures* using a loess with a very small span [see Figure 6(a)].
2. Use cross-correlation to *find the best alignment* of the two signatures: shift one of the signatures by the lag indicated by the cross-correlation function [see

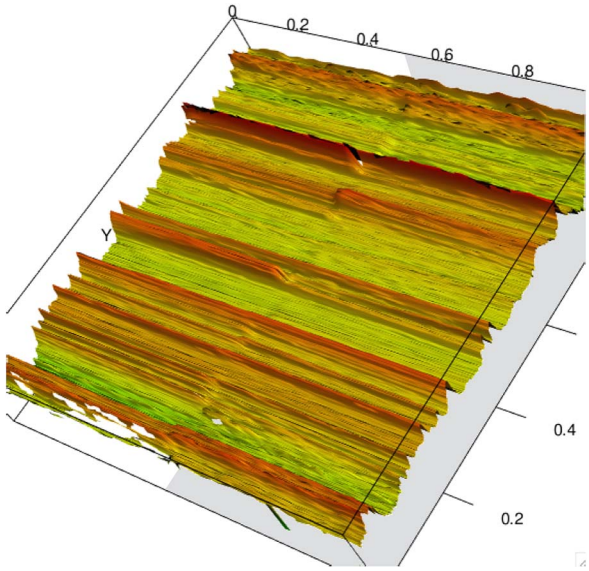
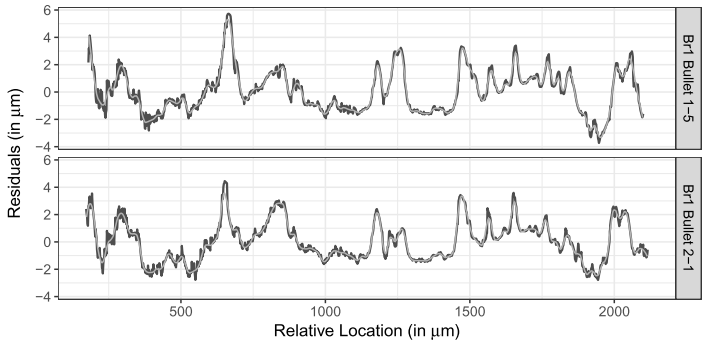


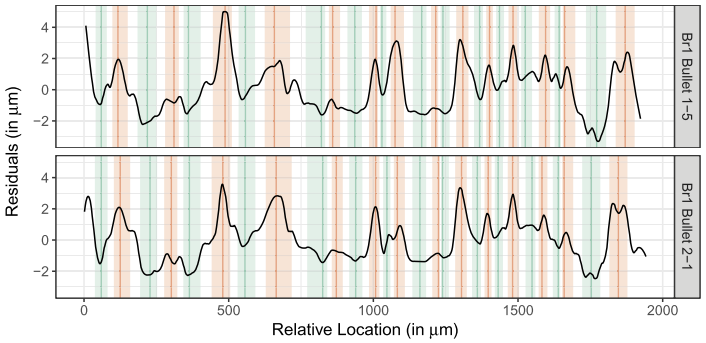
FIG. 5. *3D view of the manually adjusted side-by-side comparison of bullet* 1-5 *and bullet* 2-1 *after removing the curvature. Bullet* 2-1 *is shaded light grey in the background.*

Figure 8 for the cross-correlation function and Figure 7(b) for the resulting shift].

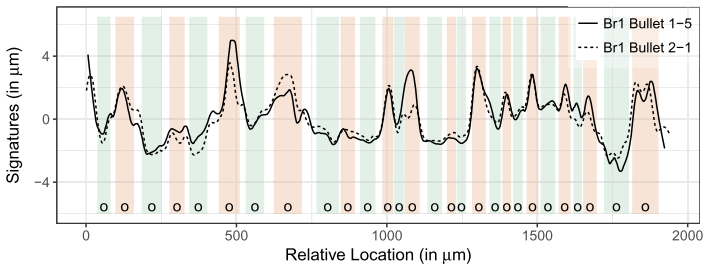
3. Using a rolling average, *identify peaks and valleys* for each of the signatures by identifying points at which the derivative of the signature is equal to zero. We then define an interval around the location of the extrema on each side as one third of the distance to the location of the next extrema [see Figure 6(b)]. Peaks and valleys constitute the *striation marks* on the bullet.
4. *Match striations across signatures*: based on the intervals around the extrema as defined above, we identify common intervals as the areas in which two or more of the individual intervals overlap: a joint interval is defined as the smallest interval that encompasses all of the overlapping intervals. A joint interval is then called a match(ing stria) between the signatures, if all of the intervals are of the same type of extrema, that is, they are either all peaks or all valleys. In Figure 6 all matches are shown as color-filled rectangles corresponding to their type of extrema (peaks are shown in orange, and valleys in green). Nonmatching intervals are left grey.
5. *Extract features from the aligned signatures and the matches between them*: many different features can be extracted from the aligned signatures. Here, we describe a few of the ones that can be found in the literature and some that we found to be of practical relevance:
   (i) Maximal number of CMS (consecutive matching striae), and, similarly, the number of consecutively nonmatching striae (CNMS),



(a) Loess smooth of signatures at a height of $x = 100$ μm (span is 0.03).



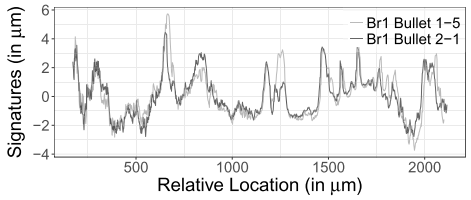
(b) Using a rolling median peaks and valleys are identified for each signature. Peaks and valleys on the signature correspond to striation marks on the bullet's surface.



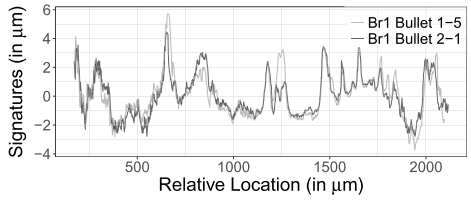
(c) Rectangles in the back identify a striation mark on one of the bullets. Matching striation marks are indicated by color filled rectangles and marked by an "o".

FIG. 6. *Matching striation marks*: *smooth* (a), *identify peaks and valley* (b), *and match peaks and valleys between signatures* (c).

(ii) Number of matches and nonmatches,
(iii) The value of the cross-correlation function (ccf) between the aligned signatures [Vorburger et al. (2011)],



(a) Raw bullet land impression signatures.



(b) Aligned signatures.

FIG. 7. *Signatures of bullets* 1-5 *and* 2-1 *taken at heights of* $x = 100$ μm. *A horizontal shift of the values of bullet* 1-5 *to the right shows the similarity of the striation marks.*

(iv) Average difference D between signatures, defined as the Euclidean vertical distance between surface measurements of aligned signatures. Let $f(t)$ and $g(t)$ be smoothed, aligned signatures:

$$D^2 = \frac{1}{\#t} \sum_t [f(t) - g(t)]^2,$$

(v) The sum S of average absolute heights of matched extrema: for each of the two matched stria, compute the average of the absolute heights of the peaks or valleys. S is then defined as the sum of all these averages.

The difference D between signatures is here defined as the Euclidean distance (in μm). In the paper by Ma et al. (2004), distance is defined as a measure relative

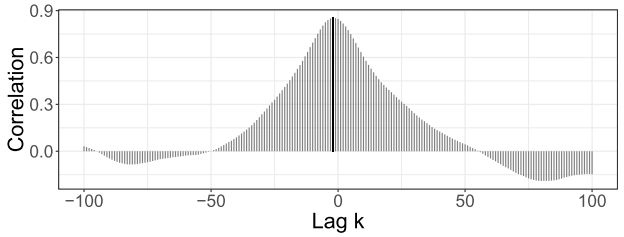


FIG. 8. *Cross-correlation function between the two signatures shown in Figure* 7(a) *at lags between* -100 *and* 100. *The correlation is maximized at a lag of* -2, *indicating the largest amount of agreement between the signatures. Figure* 7(b) *shows the slight change resulting from the lag-shifted signatures.*

to the first signature, which serves as a comparison reference and is therefore a unitless quantity.

Counting the maximal number of CMS is part of the current practice to identify bullet matches [Nichols (1997, 2003a, 2003b)]. In the example of Figure 6, the number of consecutive matching striations (CMS) is fifteen, a high number suggestive of a match between the land impressions. Note that the definition of CMS we use does not match the one given in Chu et al. (2013). There, CMS is defined only in terms of matching peaks without regarding valleys. Additionally, peaks in Chu et al. (2013) are used only if they can be identified and matched "within a tolerable range" between land impressions. The definition given here is computationally less complex, but should yield highly correlated values, because of the requirement to only consider signatures from a stable region in the land (see Section 4.3 for further details on stability of regions). In the Hamby study, the definition of CMS by Chu et al. (2013) leads to approximately half of the values of CMS defined in this paper (with a correlation coefficient between the values of the two definitions of about 0.92). For lead bullets, such as used in the Hamby study, Biasotti (1959) considered four or more consecutive peaks (corresponding to eight or more consecutive lines in our definition) to be sufficient evidence of a match.

Determining a threshold such that CMS values above the threshold indicate a match with high reliability is beyond the scope of this work, even though it is critically important in practice. We provide some ideas in the next section, but first we assess the robustness of the matching algorithm to different choices of the parameter values.

4.2. *Horizontal alignment*. Signatures of each of the two land impressions, 1-5 and 2-1, in Figure 5 are shown in Figure 7 extracted at a height of $x = 100$ μm. Striation marks show up in these representations as peaks and valleys. The individual characteristics are prominent and, again, suggest a match between the land impressions. A horizontal shift of one of the signatures [result shown in Figure 7(b)] emphasizes the strong similarities between signatures. For this alignment, we use the cross-correlation function to find a maximal amount of agreement between the signatures [Bachrach (2002), Vorburger et al. (2011), Chu et al. (2010, 2013)]. This horizontal shift is based on the cross-correlation between the two signatures: let $f(t)$ and $g(t)$ define the signature values at t, where t are locations between 0 μm and about 2500 μm, 1.5625 μm apart. The cross-correlation between f and g at lag k is then defined as

$$(f * g)(k) = \sum_t f(t+k)g(t),$$

with suitably defined limits for the summation.

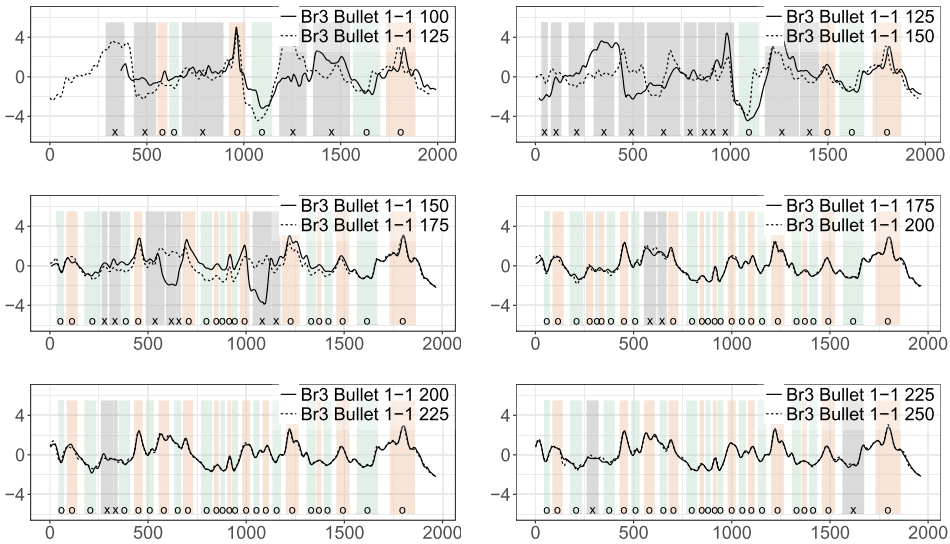


FIG. 9. *Signatures for barrel* 3, *bullet* 1-1 *extracted from varying heights. Initially, the match between signatures taken at heights* 25 μ*m apart is affected strongly by some break off at the bottom of the bullet. At a level of* 175 μ*m the bullet's signature stabilizes. For this land impression, matches should not be attempted at lower heights.*

4.3. *Impact of bullet height.* The height at which signatures are extracted for a comparison between bullet land impressions matters—signatures taken from heights that are further apart, show more pronounced differences between the signatures. This poses both a caveat to matching attempts as well as an opportunity for quality control: we have to be aware of the height that was used in a matching. Visually, matches degrade if the signatures upon which the match is based are from heights further than 200 μm apart (see the Supplementary Material, Section 2, for more discussion). However, we can extract signatures from multiple heights of the same bullet land impression for an initial assessment of its quality. By comparing signatures from heights that are not too far apart—25 μm to 50 μm—we get an indication whether the signatures come from a rapidly changing section of the surface, indicative of a break-off or some other damage, or from a stable section, where we have a reasonable expectation of finding matches to other signatures. In the approach here, we keep increasing the height x at which the signature is taken until we find a section with a stable pattern. This process is shown in Figure 9 at the example of bullet 1-1 from barrel 3, where "stability" is defined as two aligned signatures from heights chosen 25 μm apart having a cross-correlation of at least 0.95.

4.4. *Varying smoothing factor.* As mentioned earlier, the algorithm for detecting peaks and valleys depends on the selection of a smoothing window, called the smoothing factor or span. A smoothing factor of k means that the k closest observations to x_o are considered for a fit for x_o. Because surface measurements are

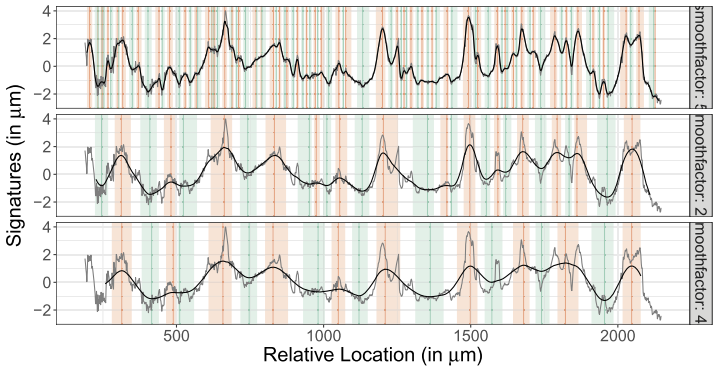
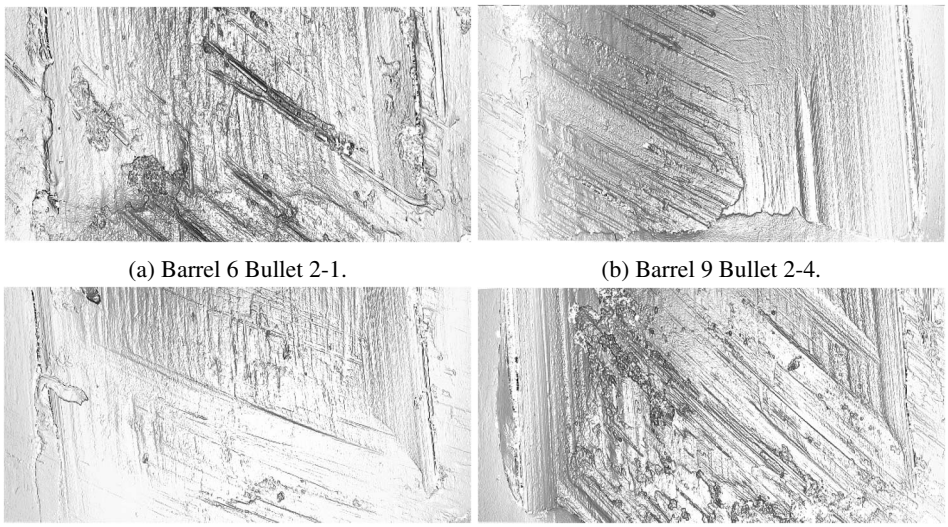


FIG. 10. *Peak/valley detection at smoothing factors of* 5, 25, *and* 45, *respectively. Note that a smoothing factor of* 5 *yields enough noise that many very minimal overlapping peaks and valleys are detected, while a smoothing factor of* 45 *might over-smooth and cause the peaks/valleys to either end disappear or shift horizontally from their original position in the signature.*

recorded at equal sampling intervals (here, of 1.5625 μm), we decided to only consider odd smoothing factors $2k + 1$, which means that the k observations to the left and right of x_o are considered for a local fit of x_o. For detecting and removing the shoulders prior to fitting a loess regression, we selected a smoothing factor of 35, while for detecting the peaks/valleys of the loess residuals a smoothing factor of 25 seems more appropriate.

Figure 10 displays the peaks and valleys detected in the same signature at smoothing factors of 5, 25, and 45, respectively. The dark line corresponds to the smoothed values, while the grey line in the back shows the raw signature. The choice of smoothing factor is a classical decision of a bias/variance trade-off. It is immediately clear that a small smoothing factor like 5 is a poor choice. It results in a significant amount of noise in the data such that even just a point or two can skew the rolling average enough for a peak or valley to be detected. Given that striation widths are typically much larger, we are in effect muddying the waters by performing such minimal smoothing. Another consideration is that the smoothing should not fall below the resolution of the equipment at which the surface measurements are taken—so as to not introduce artifacts in the analysis.

A larger smoothing factor on the other hand (like 45), seems to be a more plausible option. Most of the peaks/valleys present which are detected by a smoothing factor of 25 are also detected at 45. However, some notable issues arise. Notice that the valley on the right-hand side of the image is smoothed out, and thus not detected. On the left-hand side, a double peak is detected—that might be a questionable decision—but there are several peaks in the middle, that are smoothed out, for example, the peak at around $y = 750$. That is, in many cases, large windows are smoothing out some of the structure that we wish to see. Furthermore, it can be seen that the peaks/valleys are often shifted relative to their position in the original loess residuals, or in the smoothed data with smaller smoothing factors.

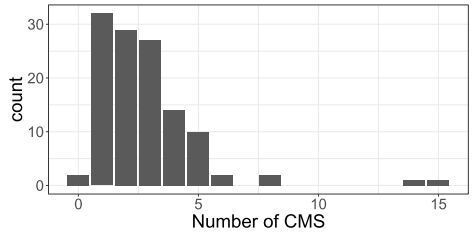


(a) Barrel 6 Bullet 2-1. (b) Barrel 9 Bullet 2-4.

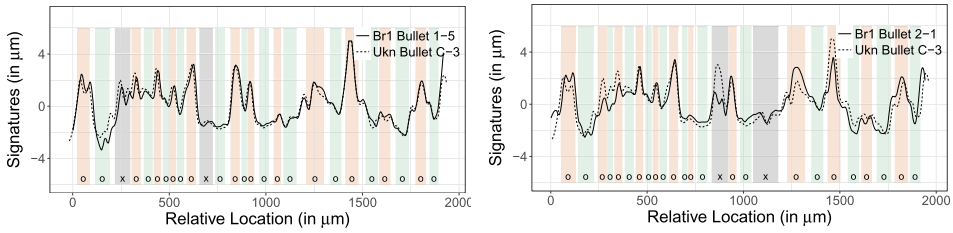
(c) Unknown Bullet B-2. (d) Unknown Bullet Q-4.

FIG. 11. *Images of the four land impressions that got flagged during the quality assessment. All of them show scratch marks (tank rash) across the striation marks from the barrel. They are excluded from the remainder of the analysis.*

5. Evaluation. In order to get a better understanding of how the matching algorithm works in known matches and nonmatches, we investigate its performance using the James Hamby study data. As a first step, we automatically assess the quality of each of the land impressions by checking that we can identify a stable region. For this, we compute the cross-correlation of signatures extracted from heights 25 μm apart. For a stable region, we require a minimum of 0.95 for the cross correlation. Four land impressions from different bullets are flagged as problematic in this respect. A visual inspection (see Figure 11) shows that each one of these land impressions has scratch marks across the surface, also known as "tank rash" [Hamby, Brundage and Thorpe (2009)]. We exclude these four land impressions from further matching considerations and run all remaining land impressions from the unknown bullets against all remaining land impressions from known bullets for matches, that is, we are comparing $15 \times 6 - 2 = 90 - 2 = 88$ land impressions from unknown bullets against $2 \times 10 \times 6 - 2 = 120 - 2 = 118$ land impressions from known bullets, yielding a total of 10,384 land-to-land comparisons. Out of these comparisons, there are 172 known matches (KM), while the rest are known nonmatches (KNM). Ideally, results look like the results in Figure 12: Figure 12(a) shows the distribution of the number of maximum consecutive matching striae between land impression C-3 and all 118 land impressions from known bullets. Two land impressions show a high CMS. These correspond to the known matches with C-3, shown in Figure 12(b) and (c). Unfortunately, not all results are as clear cut. It might not be reasonable to assume that we can match all



(a) Maximal number of CMS between unknown bullet land impression C-3 and all of the other 118 considered (known) land impressions. For two land impressions, the number of maximum CMS is high.



(b) Overlaid signatures of C-3 and the land impression with the top matching CMS.

(c) Top 2 match with C-3 based on CMS.

FIG. 12. *Showcase scenario when matching with CMS works very well. Unfortunately the matches are not always that convincing.*

land impressions, but the idea is to try to maximize the number of matches to get an overview of what we might be able to expect from an automated match.

Figure 13 shows the strong connection between the maximal number of consecutive striae and matches in the Hamby study. All 42 pairs of land impressions with at least thirteen CMS in common are matches. There are two things that should be noted at this point: the automated algorithm finds a relatively high number of CMS even for nonmatches. On average, there are 2.31 maximal CMS between known nonmatches (with a standard deviation of 1.4). Known matches share on average

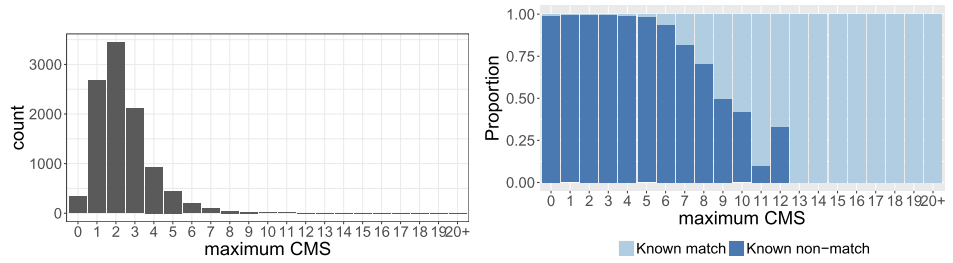


FIG. 13. *Distribution of maximal CMS* (*left*). *Conditional barchart* [*Hummel* (1996)] *on the right*: *heights show probability of match/nonmatch given a specific CMS. All land-to-land comparisons with at least* 13 *CMS are matches.*

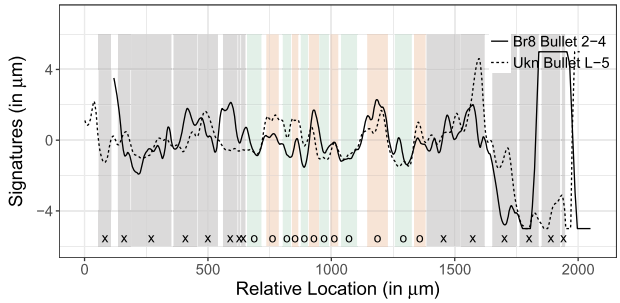


FIG. 14. *Known mismatch with a relatively large number of maximal consecutive matching striae (twelve) in the middle. The pattern in the middle does look surprisingly similar; however, the outer ends of the signatures easily reveals this comparison as mismatch.*

8.49 maximal CMS, with a standard deviation of 5.65. While the probability for a match increases with the number of maximal CMS, a large number of maximal CMS by itself is not indicative of a match, as was previously pointed out by Miller (1998). Figure 14 shows a known mismatch between two land impressions that share twelve consecutively matched striae. Visually we can easily tell that these two land impressions do not match well.

For smaller numbers of CMS, the percentage of false positives quickly increases. However, if we take other features of the image into account, we can increase the number of correct matches considerably: Figure 15 gives an overview of the densities of all of the features derived earlier, for known matches (KM) and

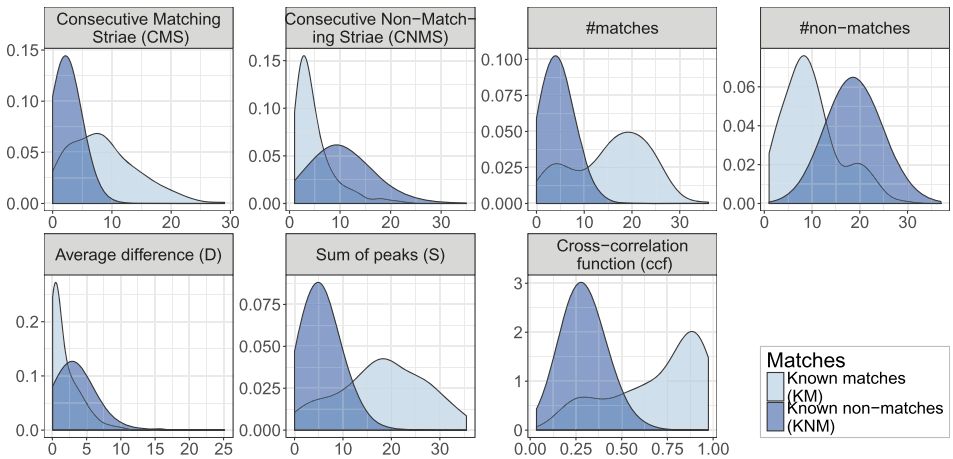


FIG. 15. *Overview of all the marginal densities for features described in Section 4.1. Shifts in the mode of the density functions between known matches and known nonmatches indicate the variable's predictive power in distinguishing matches and nonmatches. Predictive power is shown in more detail in Figure 16.*

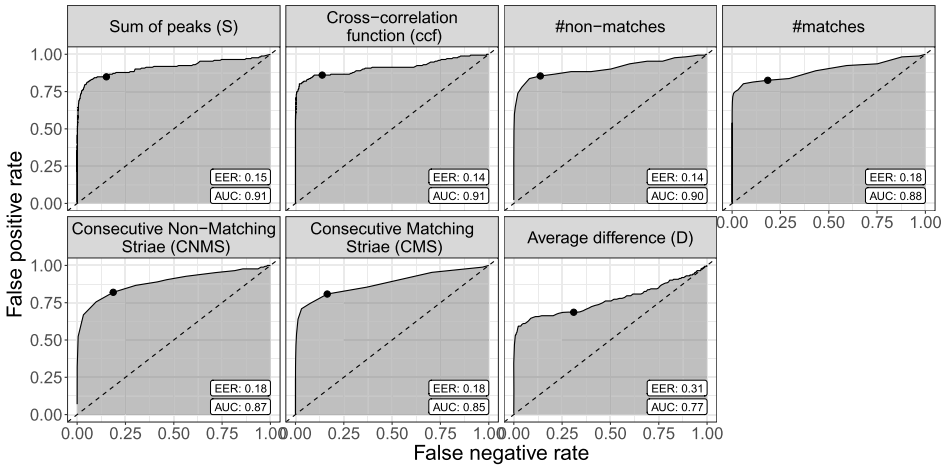


FIG. 16. *ROC curves for all of the features described in Section 4.1. Variables are sorted according to their area under the curve (AUC). The equal error rate (EER) is marked by a point on the ROC curve. Except for the distance D between signatures, all individual features derived from the surface measurements and the aligned striation marks are more predictive than the maximal CMS.*

known nonmatches (KNM). The densities of almost all of the features show strong differences between matches and nonmatches. For example, a high amount of cross-correlation between two signatures is indicative of a match—in the Hamby study, only known matches have a cross-correlation of 0.75 or higher. There are 97 land-to-land comparisons with a cross-correlation that high.

All of the features in Figure 15 show large, if not significant, differences between matches and nonmatches. The predictive power of each one of these features is shown in the form of the Receiving Operating Characteristic (ROC) curves in Figure 16. The features are arranged in descending order according to the area under the curve (AUC). The dots mark the equal error rate, that is, the location on the ROC curve, where false positive and false negative error rates are the same. The smaller the value, the better. We see that in this instance a low equal error rate (EER) goes hand in hand with high predictive power as measured in AUC. The feature with the highest individual predictive power is S, the sum of the average heights of two signatures at peaks and valleys. The maximal number of CMS is only in the seventh position here. The overall high AUC values indicate that we can successfully employ machine learning methods to distinguish matches from nonmatches.

Using recursive partitioning, we fit a decision tree [Breiman et al. (1984), Therneau, Atkinson and Ripley (2015), Milborrow (2015)] to predict matches between land impressions based on features derived from the image files. The resulting tree is shown in Figure 17. A total of 132 land impressions is being matched correctly. Interestingly, the number of consecutive matching striae does not feature

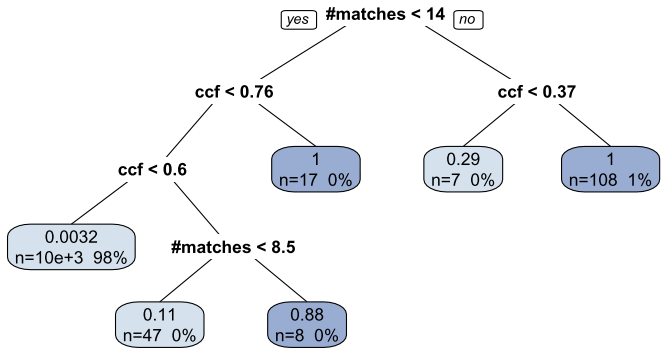


FIG. 17. *Decision tree of matching bullets based on recursive partitioning. The rectangular nodes are the leaves, giving a short summary consisting of the number of observations in the leaf (bottom left), the corresponding percentage of the total (bottom right). The number at the top shows the fraction of these observations that are a match. A* 1 *or a* 0 *therefore indicate a homogeneous (or perfect) node.*

in this evaluation. Instead of CMS, cross-correlation (ccf) between the signatures is very important in the matching process by the decision tree. Aside from cross correlation, the total number of matches is also included in the decision rule. Between cross-correlation and CMS, cross-correlation has higher predictive power. This does not contradict earlier findings emphasizing the value of CMS on visual assessments of bullet matches: in those papers, assessments were based on purely visual inspection of either actual bullets or 2D microscopic images of bullets. Neither one of these methods allows for an assessment of cross-correlations. This is one of the benefits of switching to a digitized version of the images that preserves the 3D surface structure. The findings about the discriminating power of cross-correlation are consistent with the results of the study by Ma et al. (2004). However, in that study, the authors did not consider the number of matches and nonmatches.

Another benefit of the digitized version of the images is that we can apply several hundred decision trees to combine in a random forest [Breiman (2001), Liaw and Wiener (2002)]. For each of the trees in a random forest, only two-thirds of the observations are used for fitting, while the remaining third is used to evaluate the tree's predictive power and accuracy, or its reverse, the error rate. Because errors are determined from the one-third of held-back observations, this error rate is called the out-of bag (OOB) error. Figure 18 shows the cumulative out-of-bag error (OOB) rate for 300 trees.

After about 100 trees, the error rate of land-to-land comparisons stabilizes at 0.0039. This is a weighted average between false positive error rate of 0.0001 and an error rate of false negatives of 0.2267. This out-of-bag error rate is overestimating the actual error in the Hamby study: here, the final random forest based on 300 trees is able to correctly predict all known matches and nonmatches (see

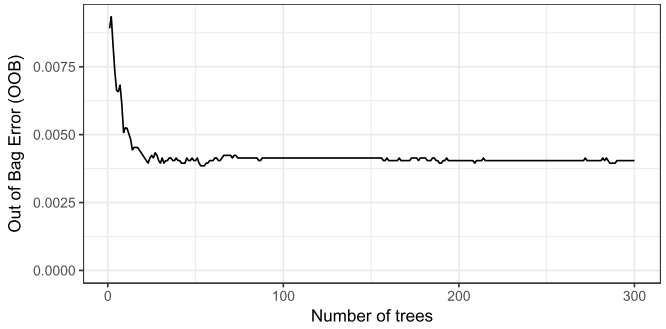


FIG. 18. *Cumulative out-of-bag error rate of a random forest fit to predict land-to-land matches from image features.*

Figure 19). Note that this error rate is based on land-to-land comparisons and is expected to be much lower for bullet-to-bullet comparisons. In the case of the Hamby data, even a single tree results in an overall error rate of zero, if we require that a match of two bullets occurs when at least two of the bullet's land impressions are matched. This makes the errors in the automated approach smaller than the human error in the Hamby study. Out of the 507 participants who returned results, eight (out of $15 \times 507 = 7605$) bullets were not matched conclusively, corresponding to a rate of 0.0011.

For the Hamby data, error rates based on bullet-to-bullet matches do not carry a lot of weight because of the small size of the study: fifteen unknown bullets are successfully matched to two pairs of ten bullets. Matching bullets can only be tested realistically in a much bigger experiment. Another thing to note about the random forest's error rates is that they are based on probability cutoffs of 0.5, that is, whenever the predicted probability of a match exceeds 0.5, a match is declared. Basing this decision on a threshold fixed at 0.5 may not be the best approach. In practice, examiners are allowed a third option of "inconclusive." On a probability

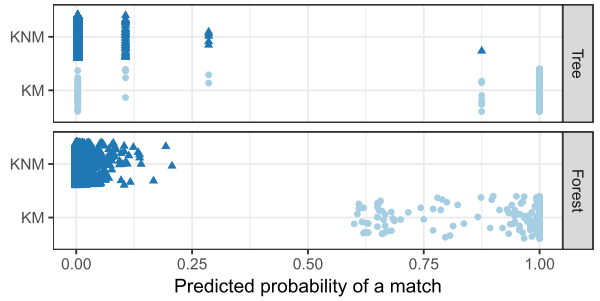


FIG. 19. *Prediction results from the tree and the forest. Using a cut-off probability of* 0.5 *the forest correctly predicts every single comparison. Compared to the tree, the forest's prediction probabilities are shrunk toward either end of the prediction range.*

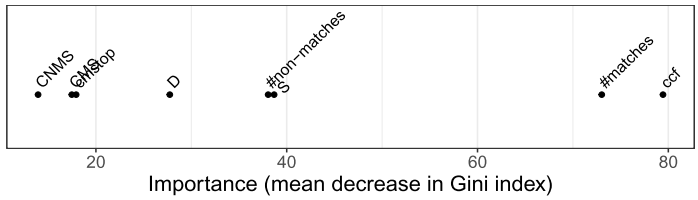


FIG. 20. *Importance of features in the random forest. Importance is measured in terms of mean decrease in gini index when including the variable in a decision tree.*

spectrum of outcomes, we could therefore introduce an interval of "inconclusive" results in the middle of the spectrum—which turns out to be unnecessary in the Hamby study, because, here, the results from the random forest are very clear cut. Figure 19 shows a comparison of the predicted probabilities of a match by the tree and the random forest. As expected, the random forest provides a more realistic estimate of the uncertainty in the classification.

Besides resulting in a probabilistic quantification of matches, random forests also provide an assessment of the importance of each of the features derived from the bullets' 3D topological surface measurements. Figure 20 shows an overview of the importance of each variable measured as the mean decrease in the Gini index when the variable in question is included in a tree (for the exact values please refer to the Supplementary Material, Section 5).

The variables with the most predictive power are cross-correlation and the overall number of matching extrema, followed by the total depth of joint striations S and total number of nonmatches. CMS is found only in sixth place.

Besides including results from known matches against known nonmatches, we can increase the number of comparisons in the Hamby study to include all possible land-to-land comparisons. This effectively doubles the number of data points available. Comparisons not previously included in fitting the random forest can also be used as an additional source for assessing error rates. Results for this and a more detailed discussion can be found in the Supplementary Material, Section 4.

6. Discussion. We present an algorithm which detects the most prominent but least relevant structure of a bullet from a firearms identification perspective, removes these features, and produces residuals which allow for the easy identification of markings. We have generalized this algorithm to align the residuals from two bullets to automatically determine whether they are indistinguishable. A random forest model provides a probabilistic assessment of the strength of a match, along with an ordering of the relevance of features. Matching bullets is clearly not a one-step process, but rather a sequence of data analysis tasks each deserving attention. As there is no scientific standard in place at this point in time, our intent is to explain an approach to addressing these tasks, while documenting all steps and

providing all code so other researchers and forensic scientists can reproduce and expand on our findings.

The matching algorithm is sensitive to the parameter choices made. The heights at which signatures are extracted (currently 25 μm apart) to evaluate stability, as well as the cross-correlation factor (currently 0.95) we set as a minimum threshold do affect the final outcome. Another parameter that must be selected is the amount of smoothing when identifying peaks and valleys (currently, a window of 23.4375 μm is used, corresponding to a window of 7 values to the left and the right of an observation). We try to lay out in the paper the impact that each of the parameter choices has on the matching performance, but more research and better data are needed to define an optimized scenario.

The Hamby study serves as our evaluation "database." It consists of only 35 bullets—this is obviously not a particularly realistic scenario for an automatic matching procedure, but for now we are unaware of other databases containing bullets in the x3p format that we could add to our study.

The feasibility of creating a database of images that could be used to identify guns used in crimes was evaluated in a 2008 report [Committee to Assess the Feasibility, Accuracy and Technical Capability of a National Ballistics Database (2008)] by the National Research Council. The committee investigated the scalability of NIBIN (National Integrated Ballistic Information Network), which uses proprietary matching algorithms provided by IBIS. The bottom line of the report was that in spite of the many technical and practical hurdles, solutions to all but one problem could be found. The problem that remained is that statistically, the quality of the matching algorithm (in this case, of breech-face marks and firing pin impressions) could not withstand a hugely increased number of records without overwhelming forensic examiners, who have to examine possible matches suggested by the system. The findings of the NRC report on imaging are based on two-dimensional greyscale images, which the committee argued were not reliable enough for distinguishing between fine marks. This finding coincides with the assessment by De Kinder, Tulleners and Thiebaut (2004) based on the IBIS Heritage system. A further reassessment by De Ceuster and Dujardin (2015) came to the same conclusions based on the EvoFinder system. The NRC report also found that results from 2D images can be improved when matches are based on 3D images. This is consistent with the importance of features found here: out of the top five features (see Figure 20), only the total number of matches and mismatches are available for a match based on 2D features.

By suggesting an automated algorithm that first removes class characteristics, such as the groove impressions, shoulders, and the curvature of the bullet to reveal the region of the land impression, then identifies peaks and valleys on this land impression, we reduce subjectivity and with it possible sources of bias. In particular, "the concept of counting striations is subjective and based on experience" [Miller (1998)]. The steps outlined in this paper could also help explore other important

forensic science problems. In particular, more general toolmark examination can benefit from the approach we discuss.

For a fair assessment of the performance of an algorithm, we need transparency. Our matching algorithm is open: the code is readily available in form of the R package bulletr [Hofmann and Hare (2016)], and the code to produce this paper is available at http://www.github.com/erichare/imaging-paper. To understand whether an automated approach along the lines of the one, we propose can accurately identify sets of bullets with undistinguishable markings, it will be necessary to assemble a much larger database that includes a wide range of ammunition types, degrees of damage, gun makes, etc. We are unaware of the existence of any such database. In addition to serving as a realistic testbed for the performance of the automated matching algorithm, such a database would also permit testing the underlying, as of yet untested, assumptions of uniqueness and reproducibility of the markings left by a gun on bullets.

Acknowledgments. Thanks to David Baldwin for pointing us to the NIST database and doing a Firearms 101 for us. Thanks to the men and women behind the software R [R Core Team (2016)], and the authors of the R packages knitr [Xie (2015)] and ggplot2 [Wickham (2009)]. We also wish to thank an anonymous reviewer, whose constructive and insightful comments contributed greatly to the improvement of the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Automatic matching of bullet land impressions" (DOI: 10.1214/17-AOAS1080SUPP; .pdf). Supplementary derivations for Automatic matching of bullet land impressions.

REFERENCES

AFTE CRITERIA FOR IDENTIFICATION COMMITTEE (1992). Theory of identification, range striae comparison reports and modified glossary definitions—AFTE criteria for identification committee report. *AFTE Journal* **24** 336–340.

BACHRACH, B. (2002). Development of a 3D-based automated firearms evidence comparison system. *J. Forensic Sci.* **47** 1253–1264.

BIASOTTI, A. A. (1959). A statistical study of the individual characteristics of fired bullets. *J. Forensic Sci.* **4** 34–50.

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

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHU, W., SONG, J., VORBURGER, T., YEN, J., BALLOU, S. and BACHRACH, B. (2010). Pilot study of automated bullet signature identification based on topography measurements and correlations. *J. Forensic Sci.* **55** 341–347.

CHU, W., SONG, J., VORBURGER, T., THOMPSON, R. and SILVER, R. (2011). Selecting valid correlation areas for automated bullet identification system based on striation detection. *J. Res. Natl. Inst. Stand. Technol.* **116** 649.

CHU, W., THOMPSON, R. M., SONG, J. and VORBURGER, T. V. (2013). Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria. *Forensic Science International* **231** 137–141.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836. MR0556476

COMMITTEE TO ASSESS THE FEASIBILITY, ACCURACY AND TECHNICAL CAPABILITY OF A NATIONAL BALLISTICS DATABASE (2008). *Ballistic Imaging*. The National Academies Press, Washington, DC.

DE CEUSTER, J. and DUJARDIN, S. (2015). The reference ballistic imaging database revisited. *Forensic Science International* **248** 82–87.

DE KINDER, J., TULLENERS, F. and THIEBAUT, H. (2004). Reference ballistic imaging database performance. *Forensic Science International* **140** 207–215.

GIANNELLI, P. C. (2011). Ballistics evidence under fire. *Criminal Justice* **25** 50–51.

HAMBY, J. E., BRUNDAGE, D. J. and THORPE, J. W. (2009). The identification of bullets fired from 10 consecutively rifled 9 mm ruger pistol barrels: A research project involving 507 participants from 20 countries. *AFTE Journal* **41** 99–110.

HARE, E., HOFMANN, H. and CARRIQUIRY, A. (2017). Supplement to "Automatic matching of bullet land impressions." DOI:10.1214/17-AOAS1080SUPP.

HOFMANN, H. and HARE, E. (2016). bulletr: Algorithms for Matching Bullet Lands. R package version 0.1.

HUMMEL, J. (1996). Linked bar charts: Analysing categorical data graphically. *Journal of Computational Statistics* **11** 23–33.

LIAW, A. and WIENER, M. (2002). Classification and regression by randomForest. *R News* **2** 18–22.

LOCK, A. B. and MORRIS, M. D. (2013). Significance of angle in the statistical comparison of forensic tool marks. *Technometrics* **55** 548–561. MR3176558

MA, L., SONG, J., WHITENTON, E., ZHENG, A., VORBURGER, T. and ZHOU, J. (2004). NIST bullet signature measurement system for RM (Reference Material) 8240 standard bullets. *Journal of Forensic Sciences* **49** 649–659.

MILBORROW, S. (2015). rpart.plot: Plot "rpart" Models: An Enhanced Version of "plot.rpart." R package version 1.5.3.

MILLER, J. (1998). Criteria for identification of toolmarks. *AFTE Journal* **30** 15–61.

NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC.

NICHOLS, R. G. (1997). Firearm and toolmark identification criteria: A review of the literature. *Journal of Forensic Sciences* **42** 466–474.

NICHOLS, R. G. (2003a). Consecutive matching striations (CMS): Its definition, study and application in the discipline of firearms and tool mark identification. *AFTE Journal* **35** 298–306.

NICHOLS, R. G. (2003b). Firearm and toolmark identification criteria: A review of the literature: Part II. *Journal of Forensic Sciences* **48** 318–327.

OPENFMC (2014). x3pr: Read/Write functionality for X3P surface metrology format. R package version 1.0.

PETRACO, N. and CHAN, H. (2012). *Application of Machine Learning to Toolmarks*: *Statistically Based Methods for Impression Pattern Comparisons*. Bibliographisches Institut AG, Mannheim, Germany.

R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

RIVA, F. and CHAMPOD, C. (2014). Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. *Journal of Forensic Sciences* **59** 637–647.

ROBERGE, D. and BEAUCHAMP, A. (2006). The use of BulletTrax-3D in a study of consecutively manufactured barrels. *AFTE Journal* **38** 166–172.

THERNEAU, T., ATKINSON, B. and RIPLEY, B. (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

VORBURGER, T. V., SONG, J. F., CHU, W., MA, L., BUI, S. H., ZHENG, A. and RENEGAR, T. B. (2011). Applications of cross-correlation functions. *Wear* **271** 529–533.

WICKHAM, H. (2009). *ggplot2*: *Elegant Graphics for Data Analysis*. Springer, New York.

XIE, Y. (2015). *Dynamic Documents with R and Knitr*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.

E. HARE
H. HOFMANN
A. CARRIQUIRY
DEPARTMENT OF STATISTICS AND STATISTICAL LABORATORY
IOWA STATE UNIVERSITY
SNEDECOR HALL
AMES, IOWA 50011-1210
USA
E-MAIL: erichare@iastate.edu
hofmann@iastate.edu
alicia@iastate.edu