

## AUTOMATIC MATCHING OF BULLET LANDS

\*

BY ERIC HARE<sup>†,‡</sup>, HEIKE HOFMANN<sup>‡</sup> AND ALICIA CARRIQUIRY<sup>‡</sup>

*Iowa State University<sup>‡</sup>*

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. We propose an automated framework for the analysis of the 3D surface measurements, that first transcribes the markings into a 2D plotting framework. 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 lands (the surface between two bullet grooves) 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, 2009](#)).

**1. Introduction.** Firearm examination is a forensic tool used to 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 vs. 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](#)

---

\*Resources for this manuscript, including R code to reproduce all of the models, tables and figures, can be found at <https://github.com/erichare/imaging-paper>.

†This research was partially funded 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.

*Keywords and phrases:* 3D Topological Surface Measurement, Data visualization, Machine Learning, Feature importance, Cross-correlation function

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 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.

Current standard practice [AFTE Criteria for Identification Committee \(1992\)](#) relies in part on the assessment of the so-called maximal number of consecutively matching striae (CMS), first defined by [Biasotti \(1959\)](#). One of the problems is that a human inspection to determine CMS is subjective ([Miller, 1998](#)).

Here, we focus on the question of defining what constitutes a match, since there is no agreement in the forensics community. 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 2D plotting framework. This allows for an objective and quantitative assessment of striae-based bullet matches.

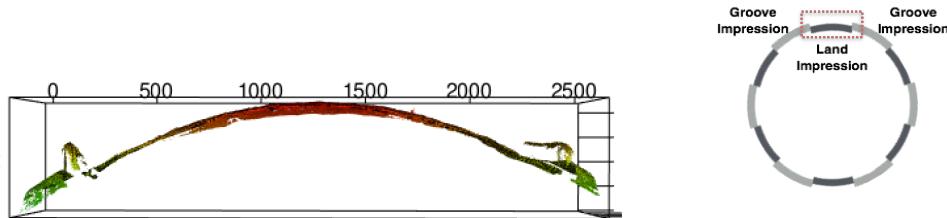
Throughout this paper, we work with images from the James Hamby Consecutively Rifled Ruger Barrel Study ([Hamby, Brundage and Thorpe, 2009](#)). Ten consecutively rifled Ruger P-85 pistol barrels were obtained from the manufacturer and fired to produce 20 known test bullets and 15 unknown bullets for 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<sup>1</sup> in a format called x3p (XML 3-

---

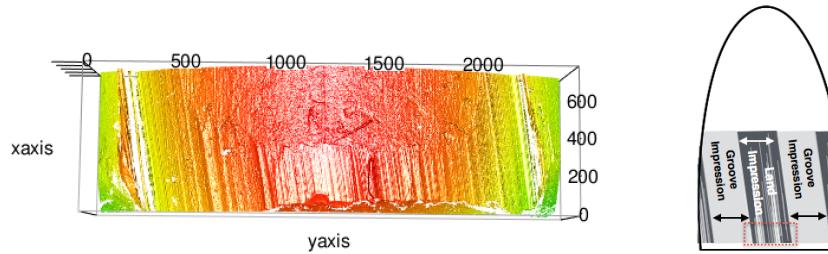
<sup>1</sup><http://www.nist.gov/forensics/ballisticsdb/hamby-consecutively-rifled-barrels.cfm>

50 D Surface Profile). The x3p format conforms to the ISO5436-2 standard<sup>2</sup>,  
 implemented to provide a simple and standard conforming way to exchange  
 2D and 3D profile data. It was adopted by the OpenFMC (Open Forensic  
 Metrology Consortium<sup>3</sup>), a group of academic, industry, and government  
 55 firearm forensics researchers whose aim is to establish best practices for  
 researchers using metrology in forensic science.

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



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



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

Each fired bullet is provided in the form of a set of six x3p files, where each file is a surface scan between adjacent grooves on the bullet, called a “land”. In the Hamby data, typical length (groove-to-groove) of a land is about 2005.9 micrometers or 2.01 millimeters. For notational simplicity, we  
 60 refer to a particular land of some bullet as bullet X-Y, where X is the bullet identifier, and Y is the land number. An example of plotting one of these lands is given in Figures 1a and 1b. These figures show side and top profiles of the land respectively. The tilt of the lines to the left in Figure 1b is not

<sup>2</sup><http://sourceforge.net/p/open-gps/mwiki/X3p/>

<sup>3</sup><http://www.openfmc.org/>

an artifact, but a direct and expected consequence of the spin induced by  
 65 the rifling during the firing process. Depending on whether a barrel is rifled clockwise or counter-clockwise, the striations have a left or right tilt. The direction of the rifling is a class characteristic, i.e. a feature that pertains to a particular class of firearms, and is not unique at the individual barrel/bullet level.

70 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](#)).

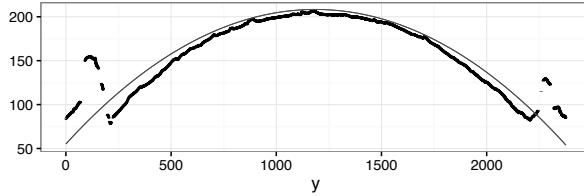
75 The purpose of our paper is to present an automatic matching routine that allows for a completely objective assessment of the quality of a match. Our algorithm is fully open source and available on GitHub ([Hofmann and Hare, 2015](#)). This transparency allows for a greater understanding of the individual steps involved in the bullet matching process, and allows other forensic examiners, as well as outside observers, examine the factors that separate known bullet matches from non-matches. We have chosen to perform  
 80 the matching on a land-to-land level, rather than bullet-to-bullet. Although doing so introduces an implicit assumption of independence between lands, assuming independence only serves to make the task more challenging. The original paper on the complete Hamby study already reports the successful  
 85 use of several computer-assisted methods; aside from a zero false positive rate, no further details on a false-negative error rate for bullets are given nor are error rates for land-to-land matches mentioned.

A procedure for bullet matching using the BulletTrax3D system is described in [Roberge and Beauchamp \(2006\)](#). This study is based on another  
 90 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 conducted by [Riva and Champod \(2014\)](#), but this work applies to striae found on cartridge  
 95 cases and not bullets.

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

100 **2. Bullet signatures.** In an approach 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. Figure 2

105 shows a plot of the side profile of a bullet land. It can be seen that the global structure of the land dominates the appearance of the plot. The grooves can be clearly identified on the left and right side, and the curvature of the surface is the most visible feature in the middle.



**Fig 2:** Side profile of the surface measurements (in  $\mu\text{m}$ ) of a bullet land at a fixed height  $x$ . Note that the global features dominate any deviations, corresponding to the individual characteristics of striation marks.

110 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 Supplement section 1.

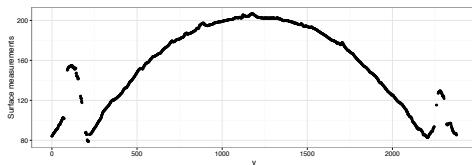
Instead of a circular fit, we use multiple loess fits to model the overall structure and extract the bullet markings.

115 **2.1. Identifying groove locations.** We first have to identify the locations of the left and right groove in the image. The grooves are assumed to contain no information relevant for determining matches. They also dominate the structure, and therefore need to be removed. Fortunately, the location and appearance of the grooves in the surface profiles is quite consistent. Surface measurements reach local maxima around the peak of the groove at either 120 end of the range of  $y$ , we can then follow the descent of the surface measurements inwards to the valley of the groove. The location of the valleys mark the points at which we should trim the image. The procedure can be described as follows:

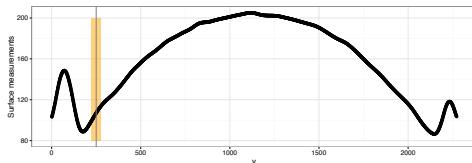
- 125 1. At a fixed height  $x$  extract a bullet's profile (Figure 3a, with  $x = 243.75\mu\text{m}$ ).
2. For each  $y$  value, smooth out any deviations occurring near the minima by twice applying a rolling average with a pre-set *smoothing factor*  $s$ . (Figure 3b, smoothing factor  $s = 35$  corresponding to  $55\mu\text{m}$ ).
3. Determine the location of the peak of the left groove by finding the first 130 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  $p_\ell$  (see Figure 3c).
4. Similarly, determine the location of the valley of the left groove by finding the first double-smoothed  $y_j$  that is the minimum within its smoothing window. Call the location of this valley  $v_\ell$ .
  - 135 5. Reverse the order of the  $y$  values and repeat the previous two steps to find the peak and valley of the right groove,  $(p_r, v_r)$ .
  6. Trim the surface measurements to values within the two grooves (i.e. remove all records with  $y_i < v_\ell$  and  $y_i > v_r$ ) (see Figure 3c).

- (a) Step 1 of identifying groove locations: For a fixed height ( $x = 243.75\mu\text{m}$ ) surface measurements for bullet 1-5 are plotted across the range of  $y$ .



- (b) Step 2 of identifying groove 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 groove 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 grooves from the bullet images.

140 The smoothing factor  $s$  introduced in the algorithm represents the window

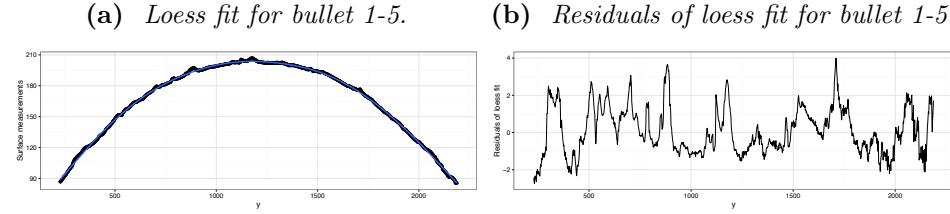
size to use for a rolling average. Higher values of  $s$  therefore lead to more smoothed results. In the groove detection portion of this algorithm, we wish to remove most of the deviation but still maintain enough signal in the groove to detect it. Empirically, a value of  $s = 35$  for the smoothing factor seems  
145 to work quite well (the smoothing factor is further discussed in Section 3.4). It is important to note that the smoothing pass is done twice. That is, the smoothed data is 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  
150 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  
155 the center of the window the highest, with the weights linearly dropping off as we reach either end of the smoothing window.

2.2. *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  
160 form of a smooth, continuous function  $f$  with  $y_i = f(x_i) + \varepsilon_i$  for all values  $i = 1, \dots, n$ . The function  $f$  is approximated via locally weighted polynomial regressions. Parameters of the estimation are  $\alpha$ , 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).

165 The main idea of locally weighted regression is to use a weighting routine that emphasizes the effect of points close by and de-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  $x_o$   
170 and the maximum distance of the range of the  $x$ -values for span  $\alpha$  in  $x_o$ .

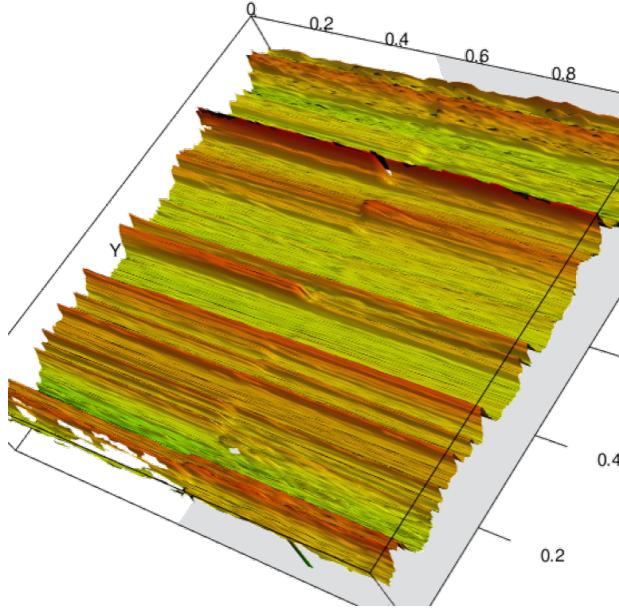
Figure 4a shows the loess fit, in blue, overlaid on the processed image of bullet 1-5. The fit seems to do a reasonable job of capturing the structure of the image. Figure 4b shows the residuals from this fit. These residuals are called the *signature* of bullet 1-5.

175 **3. Automatic matching.** Applying the loess fit to a range of different signatures (see Figure 5 for signatures extracted at heights between  $50\mu m$  and  $150\mu m$ ) shows the 3D striation marks from two bullets. Signatures of bullet 1 are shown on the left (all extracted from heights below  $100\mu m$ ) and signatures of bullet 2 are shown on the right (extracted at heights above



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

180      $100\mu\text{m}$ ). 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 lands as a match. However, this  
 185     match relies on visual inspection and can therefore not be completely objective. 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 quality of the match.

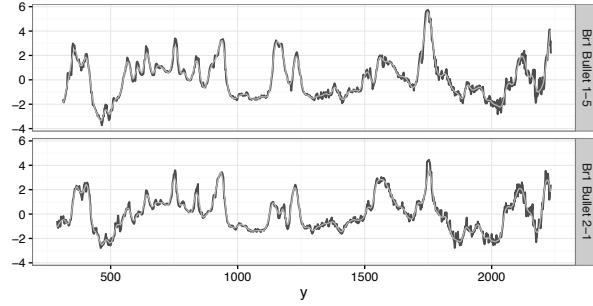


**Fig 5:** 3D view of the manually adjusted side-by-side comparison of bullet 1-5 and bullet 2-1 after removing the curvature. Bullet 2-1 is shaded light grey in the background.

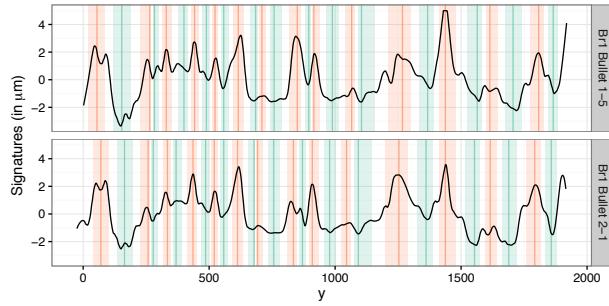
In this section, we describe the algorithm for matching signatures first,

and the impact of parameter choices in the subsections thereafter.

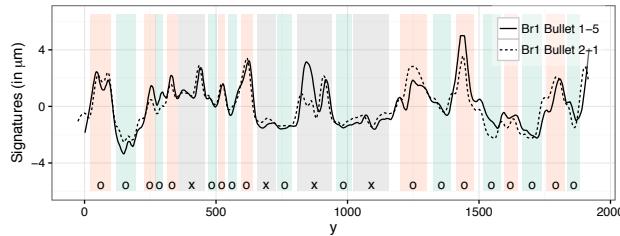
(a) Loess smooth of signatures at a height of  $x = 100\mu\text{m}$  (span is 0.03).



(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'. Mismatches are filled in grey and marked by an 'x'.



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

3.1. 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 Supplement section 3 for a more detailed discussion).

All of the other steps are done on pairs of bullet lands:

- 195 1. **Smooth the two signatures** using a loess with a very small span  
(see Figure 6a).
- 200 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 Figure 8 for the cross-correlation function and Figure 7b for the resulting shift).
- 205 3. Using a rolling average, **identify peaks and valleys** for each of the signatures. 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 6b). Peaks and valleys constitute the *striation marks* of the bullet.
- 210 4. **Match striations across signatures:** based on the intervals around the extrema as defined above, we identify joint intervals between two signatures 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, i.e. 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). Non-matching intervals are left grey.
- 215 5. **Extract features from the aligned signatures and the matches between them:** many different features could be extracted. 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:
  - 220 (i) Maximal number of CMS (consecutive matching striae), and, similarly, the number of consecutively non-matching striae (CNMS),
  - (ii) Number of matches and non-matches,
  - (iii) The value of the cross-correlation function (ccf) between the aligned signatures Vorburger et al. (2011),
  - 225 (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:

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

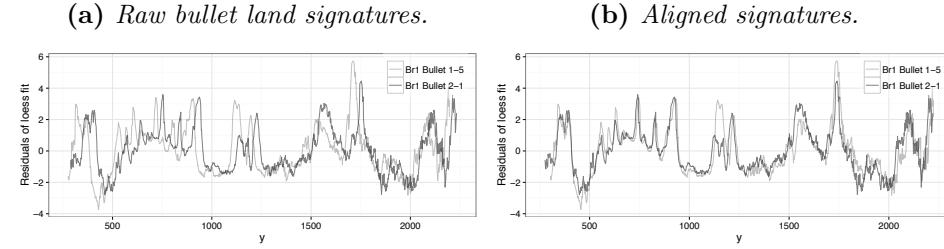
- 230 (v) The sum  $S$  of average absolute heights of matched extrema: for  
 each of the two matched stria, compute the average of the abso-  
 lute heights of the peaks or valleys.  $S$  is then defined as the sum  
 of all these averages.

235 The difference  $D$  between signatures is here defined as the Euclidean distance  
 (in  $\mu m$ ). In the paper by [Ma et al. \(2004\)](#), distance is defined as a measure  
 relative to the first signature, which serves as a comparison reference and is  
 therefore a unitless quantity.

240 Counting the maximal number of CMS is part of the current practice to  
 identify bullet matches ([Nichols, 1997, 2003a,b](#)). In the example of Figure 6,  
 the number of consecutive matching striations (CMS) is fifteen, a high num-  
 ber indicative of a match between the bullets. Note that the definition of  
 245 CMS given in this paper 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 lands. The defi-  
 nition given here is computationally less complex, but should yield highly  
 250 correlated values, because of the requirement to only consider signatures  
 from a stable region in the land (see Section 3.3 for further details on sta-  
 bility 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 pa-  
 per (with a correlation coefficient between the values of the two definitions  
 255 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.

255 Determining a threshold such that CMS values above the threshold in-  
 dicate 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.

260 **3.2. Horizontal alignment.** Signatures of each of the two lands, 1-5 and  
 2-1, in Figure 5 are shown in Figure 7 extracted at a height of  $x = 100\mu m$ .  
 Striation marks show up in these representations as peaks and valleys. The  
 individual characteristics are quite prominent and, again, suggest a match  
 between the lands. A horizontal shift of one of the signatures (result shown  
 in Fig 7b) emphasizes the strong similarities. For this alignment we use the  
 cross-correlation function to find a maximal amount of agreement between  
 265 the signatures [Bachrach \(2002\)](#); [Chu et al. \(2010\)](#); [Vorburger et al. \(2011\)](#);  
[Chu et al. \(2013\)](#). This horizontal shift is based on the cross-correlation

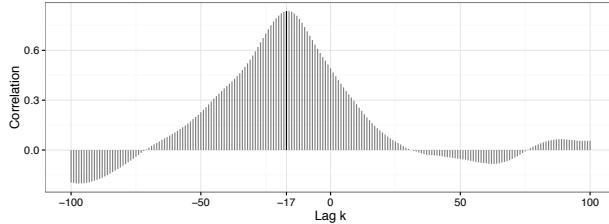


**Fig 7:** Signatures of bullets 1-5 and 2-1 taken at heights of  $x = 100\mu\text{m}$ . A horizontal shift of the values of bullet 1-5 to the right shows the similarity of the striation marks.

between the two signatures: let  $f(t)$  and  $g(t)$  define the signature values at  $t$ , where  $t$  are locations between  $0 \mu\text{m}$  and about  $2500 \mu\text{m}$ ,  $1.5625 \mu\text{m}$  apart. The cross-correlation between  $f$  and  $g$  at lag  $k$  is then defined as

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

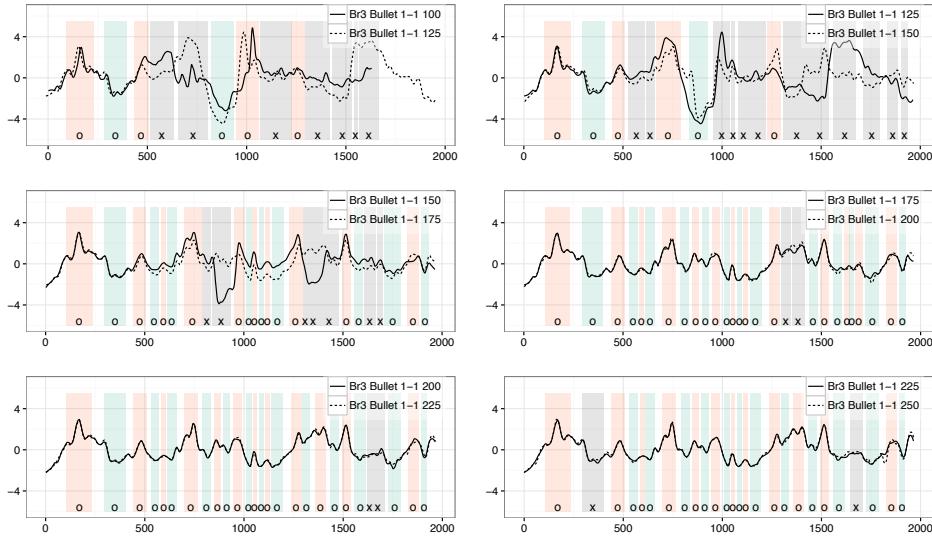
270 with suitably defined limits for the summation.



**Fig 8:** Cross-correlation function between the two signatures shown in Figure 7a at lags between -100 and 100. The correlation is maximized at a lag of -17, indicating the largest amount of agreement between the signatures. Figure 7b shows the lag-shifted signatures.

3.3. *Impact of bullet height.* The height at which signatures are extracted for a comparison between bullet lands 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  $200\mu\text{m}$  apart (see Supplement section 2 for a more thorough discussion). However, we can extract  
275

signatures from multiple heights of the same bullet land for an initial assessment of its quality. By comparing signatures from heights that are not too far apart –  $25\mu m$  to  $50\mu m$  – 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  $25\mu m$  apart having a cross-correlation of at least 0.95.



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

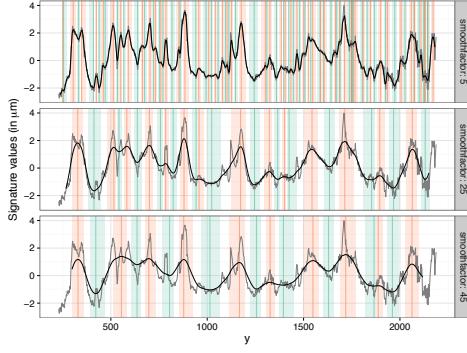
3.4. *Varying smoothing factor.* As mentioned earlier, the algorithm for detecting peaks and valleys depends on the selection of a smoothing window, called the smoothing factor or span. A smoothing factor of  $k$  means that the  $k$  closest observations to  $x_o$  are considered for a fit for  $x_o$ . Because surface measurements are recorded at an equidistant resolution (here, of  $1.5625\mu m$ ), 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 grooves 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 – in order 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.

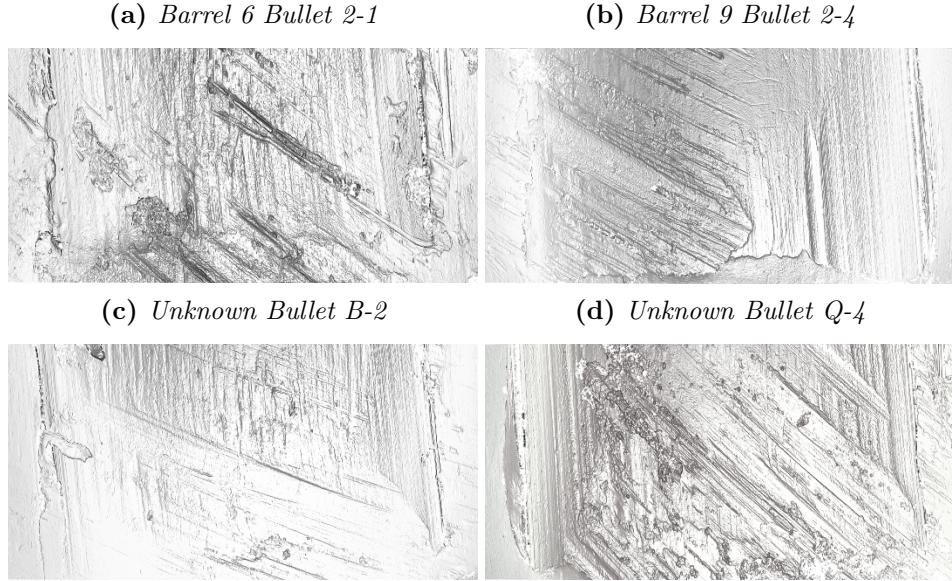
**4. Evaluation.** In order to get a better understanding of how the matching algorithm works in known matches and non-matches, we investigate its performance in a test scenario based on the James Hamby study. As a first step, we automatically assess the quality of each of the lands by checking that we can identify a stable region on each land. For this, we compute the cross-correlation of signatures extracted from heights  $25\mu m$  apart. For a stable region, we require a minimum of 0.95 for the cross correlation. Four lands from different bullets are flagged as problematic in this respect. A visual inspection (see Figure 11) shows that each one of these lands has scratch marks across the surface, also known as ‘tank rash’. We exclude these four lands from further matching considerations and run all remaining lands from the unknown bullets against all remaining lands from known



**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.

335 bullets for matches, i.e. we are comparing  $15 \times 6 - 2 = 90 - 2 = 88$  lands from unknown bullets against  $2 \times 10 \times 6 - 2 = 120 - 2 = 118$  lands 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 non-matches (KNM). When things go well, they look like the results  
 340 in Figure 12: Figure 12a shows the distribution of the number of maximum consecutive striae between land C-3 and all 118 lands from known bullets. Two lands show a high CMS. These correspond to the known matches with C-3, shown in Figures 12b and 12c. Unfortunately, not all results are as clear cut. It might not be reasonable to assume that we can match all lands, but the idea is to try to maximize the number of matches  
 345 to get an overview of what we might be able to expect from an automated match.

350 Figure 13 shows the strong connection between the maximal number of consecutive striae and matches in the Hamby study. All 42 pairs of lands 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 non-matches. On average, there are 2.31 maximal CMS between known non-matches (with a standard deviation of 1.4). Known matches share on average 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  
 355



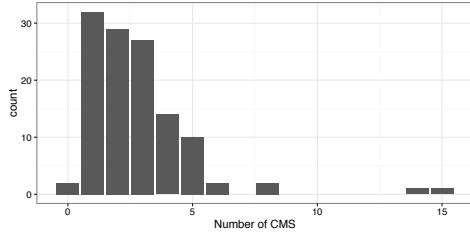
**Fig 11:** *Images of the four lands that got flagged during the quality assessment. All of them show scratch marks (tank rash) across the striation marks from the barrel. They are excluded from the remainder of the analysis.*

known mismatch between two lands that share twelve consecutively matched striae. Visually we can easily tell that these two lands do not match well.

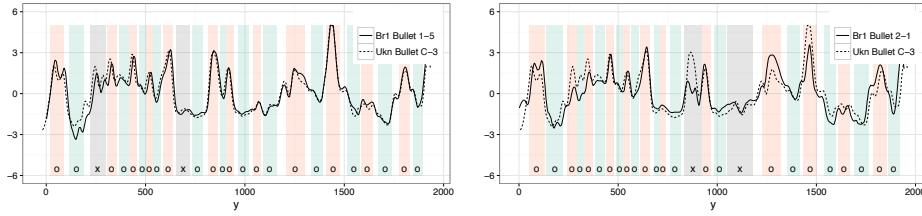
360 For smaller numbers of CMS, the percentage of false positives quickly increases. However, if we take other features of the image into account, we can increase the number of correct matches considerably: Figure 15 gives an overview of the densities of all of the features derived earlier, distinguishing between known matches (KM) and known non-matches (KNM). The densities of almost all of the features show strong differences between matches 365 and non matches. 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.

370 All of the features in Figure 15 show large, if not significant, differences 375 between matches and non-matches. 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, i.e. 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

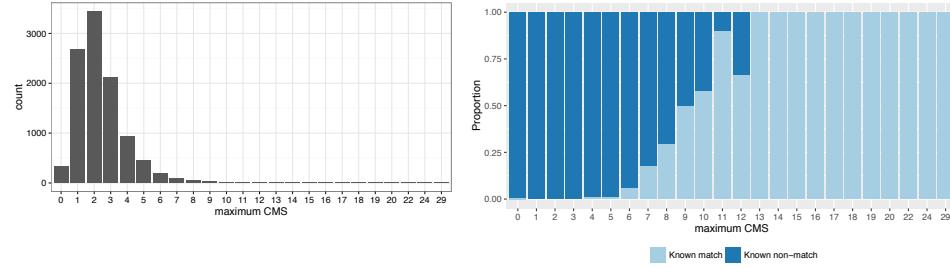
(a) Maximal number of CMS between unknown bullet C-3 and all of the other 118 considered (known) lands. For two lands the number of maximum CMS is high.



(b) Overlaid signatures of C-3 and the land with the top matching CMS. (c) Top 2 match with C-3 based on CMS.

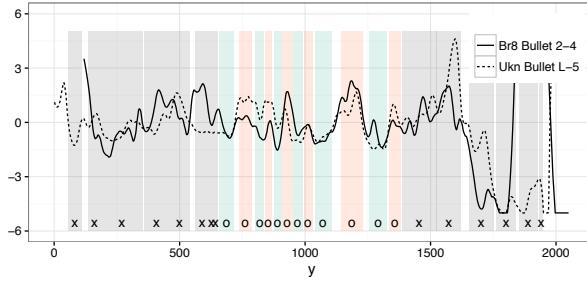


**Fig 12:** Showcase scenario when matching with CMS works very well. Unfortunately the matches are not always that convincing.

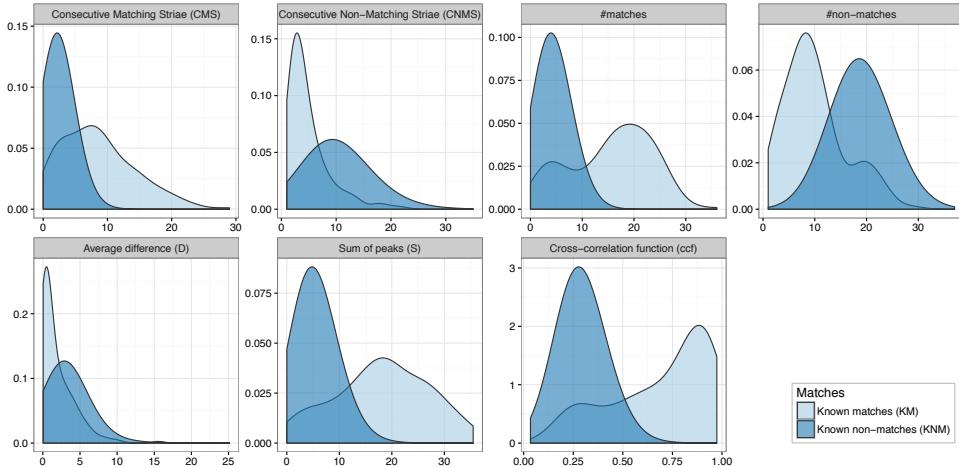


**Fig 13:** Distribution of maximal CMS (left). Conditional barchart (Hummel, 1996) on the right: heights show probability of match/non-match given a specific CMS. All land-to-land comparisons with at least 13 CMS are matches.

that in this instance a low 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 non-matches.

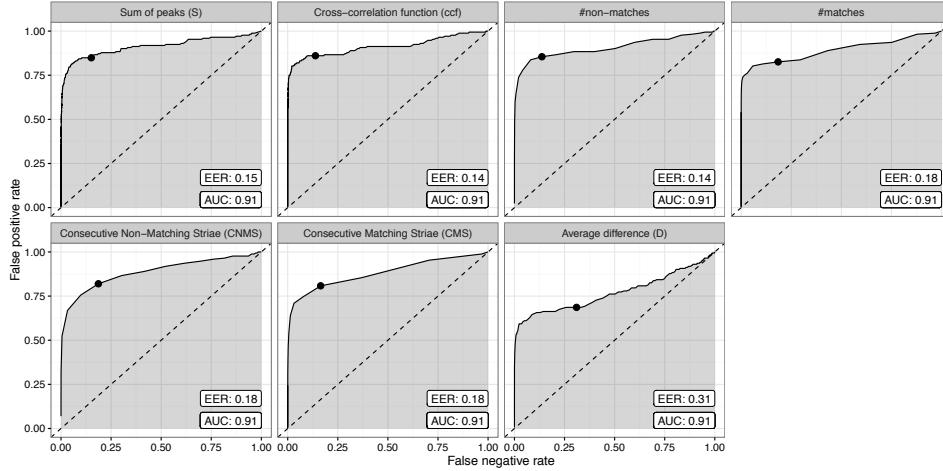


**Fig 14:** Known mismatch with a relatively large number of maximal consecutive matching striae (twelve) in the middle. The pattern in the middle does look surprisingly similar, however the outer ends of the signatures easily reveals this comparison as mismatch.



**Fig 15:** Overview of all the marginal densities for features described in section 3.1. Shifts in the mode of the density functions between known matches and known non-matches indicate the variable's predictive power in distinguishing matches and non-matches. Predictive power is shown in more detail in Figure 16.

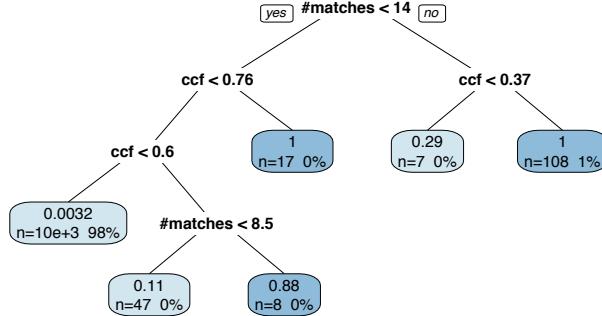
Using recursive partitioning, we fit a decision tree (Breiman et al., 1984; Therneau, Atkinson and Ripley, 2015; Milborrow, 2015) to predict matches between lands based on features derived from the image files. The resulting tree is shown in Figure 17. A total of 132 lands is being matched correctly. Interestingly, the number of consecutive matching striae does not feature in this evaluation. Instead of CMS, cross-correlation (ccf) between the signatures is very important in the matching process by the decision tree. Aside



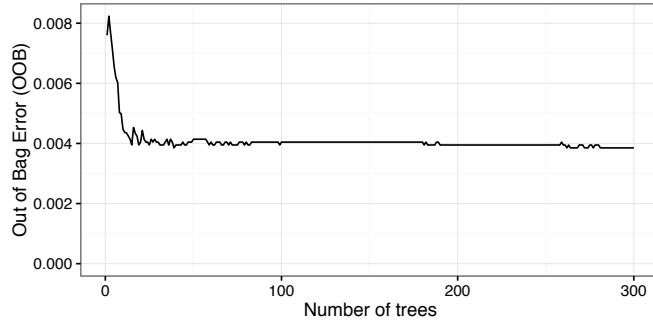
**Fig 16:** ROC curves for all of the features described in section 3.1. Variables are sorted according to their area under the curve (AUC). The equal error rate (EER) is marked by a point on the ROC curve. Except for the distance D between signatures, all individual features derived from the surface measurements and the aligned striation marks are more predictive than the maximal CMS.

390 from cross correlation, the total number of matches is also included in the de-  
 cision 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, as-  
 sessments were based on purely visual inspection of either actual bullets or  
 395 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,  
 400 the authors did not consider the number of matches and non-matches.

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  
 405 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



**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.

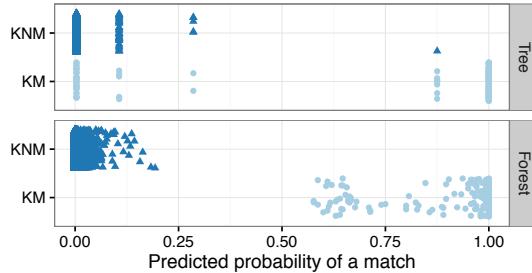


**Fig 18:** Cumulative out-of-bag error rate of a random forest fit to predict land-to-land matches from image features.

100 trees, the error rate of land-to-land comparisons stabilizes at 0.0039. This  
 410 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 over-estimating  
 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 non-matches (see  
 Figure 19). Note that this error rate is based on land-to-land comparisons  
 415 and is much lower for bullet-to-bullet matches. In the case of the Hamby  
 study, even for the single tree the overall error rate is zero, if we require for  
 a bullet match that at least two of a bullet's lands 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  $15 \times 507 = 7,605$ ) bullets were not matched conclusively, corresponding to a rate of 0.0011.

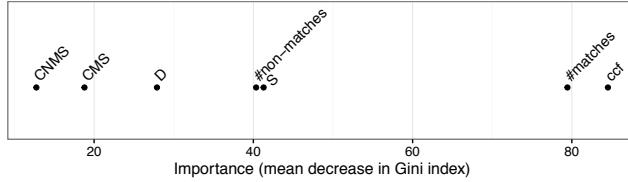
However, error rates based on bullet-to-bullet matches do not carry a lot of weight because of the small size of the study: fifteen unknown bullets are successfully matched to two pairs of ten bullets. Matching bullets can only be tested realistically in a much bigger experiment. Another thing to note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



**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 towards either end of the prediction range.

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 Supplement section 5).

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



**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.

Besides including results from known matches against known non-matches,  
 445 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  
 450 Supplement section 4.

**5. Conclusion.** We have presented 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  
 455 to align the residuals from two bullets to automatically determine whether they are matches. We created a random forest model using many of the features to provide a probabilistic assessment of the quality of a match, along with the most relevant features. Matching bullets is clearly not a single step, but rather a sequence of a number of data analysis tasks which each deserve  
 460 attention. As there is no scientific standard in place at this point in time, our intent is to explain one 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  
 465 heights at which signatures are extracted (currently  $25\mu m$  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 grooves (currently, a window of  $23.4375\mu m$  is used, corresponding to a window of 7 values to the left and the right of an observation). While we tried  
 470 to lay out in the paper the impact that each of the parameter choices has on the matching performance, we are still far from 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.

The feasibility of creating a database of ballistic 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 evaluation was investigating 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 while still maintaining a reasonable workload for forensic examiners, who have to examine possible matches suggested by the system. The findings of the NRC report on ballistic 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, Tulленers and Thiebaut \(2004\)](#) based on the IBIS Heritage system. A further re-assessment 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 grooves and the curvature of the bullet to reveal the region of the land, then identifies peaks and valleys on this land, 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 this detection and removal of the overall structure and the extraction of features from deviations of this structure.

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 x3prplus ([Hofmann and Hare, 2015](#)), 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.

**Acknowledgment.** 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, 2015](#)), and the authors of the R packages knitr ([Xie, 2015](#)) and ggplot2 ([Wickham, 2009](#)).

## References.

- 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. *Journal of Forensic Sciences* **4** 34–50.
- BREIMAN, L. (2001). Random Forests. *Machine Learning* **45** 5–32.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Inc.
- 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. *Journal of Research of the National Institute of Standards and Technology* **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 Scatter-plots. *Journal of the American Statistical Association* **74** 829–836.
- COMMITTEE TO ASSESS THE FEASIBILITY, ACCURACY AND TECHNICAL CAPABILITY OF A NATIONAL BALLISTICS DATABASE (2008). *Ballistic Imaging*. The National Academies Press, Washington, DC.
- NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward*. 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.
- 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.

- 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  
560 Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project  
Involving 507 Participants from 20 Countries. *AFTE Journal* **41** 99–110.
- HOFMANN, H. and HARE, E. (2015). x3prplus: Simple Models for x3d surface topology R  
package version 0.0.1.
- HUMMEL, J. (1996). Linked Bar Charts: Analysing Categorical Data Graphically. *Journal  
565 of Computational Statistics* **11** 23–33.
- LIAW, A. and WIENER, M. (2002). Classification and Regression by randomForest. *R  
News* **2** 18–22.
- 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  
570 bullets. *Journal of Forensic Sciences* **49** 649–659.
- MILBROWNE, 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.
- NICHOLS, R. G. (1997). Firearm and Toolmark Identification Criteria: A Review of the  
575 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  
580 Literature: part II. *Journal of Forensic Sciences* **48** 318–327.
- 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 con-  
secutively manufactured barrels. *AFTE Journal* **38** 166–172.
- 585 R CORE TEAM (2015). R: A Language and Environment for Statistical Computing R  
Foundation for Statistical Computing, Vienna, Austria.
- 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.
- 590 VORBURGER, T. V., SONG, J. F., CHU, W., MA, L., BUI, S. H., ZHENG, A. and RENE-  
GAR, T. B. (2011). Applications of cross-correlation functions. *Wear* **271** 529–533. The  
12th International Conference on Metrology and Properties of Engineering Surfaces.
- WICKHAM, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New  
York.
- 595 XIE, Y. (2015). *Dynamic Documents with R and knitr*, 2nd ed. Chapman and Hall/CRC,  
Boca Raton, Florida. ISBN 978-1498716963.

## ADDRESS OF THE AUTHORS

DEPARTMENT OF STATISTICS AND STATISTICAL LABORATORY  
SNEDECOR HALL

AMES IOWA 50011-1210

E-MAIL: erichare@iastate.edu  
hofmann@iastate.edu  
alicia@iastate.edu