

Case study validations of automatic bullet matching^{*,**}

Heike Hofmann^{a,b,*}, Susan Vanderplas^{a,b}

^a*Statistics Department, Iowa State University
2438 Osborne Dr, Ames, IA 50011*

^b*Center for Statistical Applications in Forensic Evidence, Iowa State University
613 Morrill Rd, Ames, IA 50011*

Abstract

Recent advances in microscopy have made it possible to collect 3D topographic data, enabling virtual comparisons based on the collected 3D data next to traditional comparison microscopy. Automatic matching algorithms have been introduced for various scenarios, such as matching cartridge cases (Tai and Eddy 2018) or matching bullet striae (Hare et al. 2017b, Chu et al 2013, De Kinder and Bonfanti 1999). One key aspect of validating automatic matching algorithms is to evaluate the performance of the algorithm on external tests. Here, we are presenting a discussion of the performance of the matching algorithm (Hare et al. 2017b) in three studies. We are considering matching performance based on the Random forest score, cross correlation, and consecutive matching striae (CMS) at the land-to-land level and, using Sequential Average Maxima scores, also at the bullet-to bullet level. Cross correlation and Random Forest scores both result in perfect discrimination of same-source and different-source bullets. At the land-to-land level, discrimination (based on area under the curve, AUC) is excellent (≥ 0.90).

^{*}This work 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.

^{**}The authors wish to thank Alan Zheng (NIST), Tylor Klep (Phoenix PD), and Melissa Nally and Kasi Kirksey (FSI Houston) for access to test set bullets. We would also like to thank the efforts of the Roy J Carver High Resolution lab in scanning the bullet sets and providing the scans to us. Finally, thanks to Prof James Hamby for his advice on XXX.

^{*}Corresponding author

Email address: hofmann@iastate.edu (Heike Hofmann)

Keywords: forensic science, toolmark, cross correlation, random forest, 3D microscopy, land engraved areas (LEAs)

In current practice, firearms and toolmark examiners (FTE) evaluate the similarity of striae on bullets by placing the evidence together with another bullet under a comparison microscope. The second bullet could be a test fire from a weapon recovered during the investigation, or it could be a second bullet from the crime scene. Examiners visually classify similarity according to the theory of firearms identification [1] as one of identification, inconclusive or exclusion. Exact guidelines for this classification vary from lab to lab; some labs will exclude only on the basis of non-matching class characteristics, such as direction of the twist in rifling, land length or number of lands, or type of rifling. In other labs, CMS (consecutively matching striae) as defined by Biasotti [2] is used as a measure to quantify the strength of a match. In virtually all labs, individual characteristics used to identify matching bullets are derived from visual assessment; some class characteristics may be directly measured, but these are not sufficient for individualization.

More explicit characterization of bullet surfaces has been discussed since at least 1958 by Davis [3], but at the time technology was not sufficiently advanced to make an analysis based on 3D measurements or surface traces a viable option. Gardner [4] demonstrated use of a scanning electron microscope to quantitatively examine and compare bullet striae. More recently, approaches using 3D measurement data were explicitly described in 1999 by De Kinder and Bonfanti [5], and have been further developed in [6, 7, 8]. These approaches utilize 3D surface measurements directly to characterize the topology of land engraved areas (LEAs), rather than using visual or image comparison techniques. In many cases, these approaches also provide some level of automation of the comparison process, with the goal of reducing human biases by augmenting the visual information with 3D measurements. Utilizing the 3D measurements allows for examination of both peaks and valleys in LEAs, it also allows us to take the depth of striae into account, something that is difficult, if not impossible, from

visual inspection.

Commonly, approaches derived from 3D surface measurements use some features which are similar to those visually assessed by FTEs [9]. Class characteristics, which are shared by a group of firearms with the same rifling design, manufacturer, and tooling process, are typically evaluated first, as a mismatch on class characteristics is sufficient for an exclusion. Automated approaches to estimation of width of the land engraved area and twist angle were described in Chu et al. [8]. Individual characteristics, which are not shared by all members of a class, can also be automatically assessed from 3D surface measurements. One of the most common features used to assess the strength of a match based on individual characteristics is the cross correlation function, which is utilized in several studies [10, 11, 8]. Additional features proposed for automatically assessing match strength also include signature distance [10], striae depth and width [4], and consecutive matching striae (CMS) [12].

1. Methods

In the matching algorithm proposed in Hare et al. [13], several features are combined using a random forest [14] to produce a similarity score based on 3D topographic scans of land engraved areas (LEAs). In order to generate these features, some pre-processing is necessary in order to transform the 3D surface measurements into ‘signatures’ which can be compared.

1.1. Automated Processing of 3D Scans

From each land engraved area a signature is derived using the process described in detail in Hare et al. [13]:

- (a) Identify an area on the LEA with expressed striae. Locate a stable crosscut in the identified area.
- (b) Discard extraneous/contaminated data, such as data from groove engraved area and areas affected by break-off or contact with objects after the bullet exited the barrel (“tank rash”).

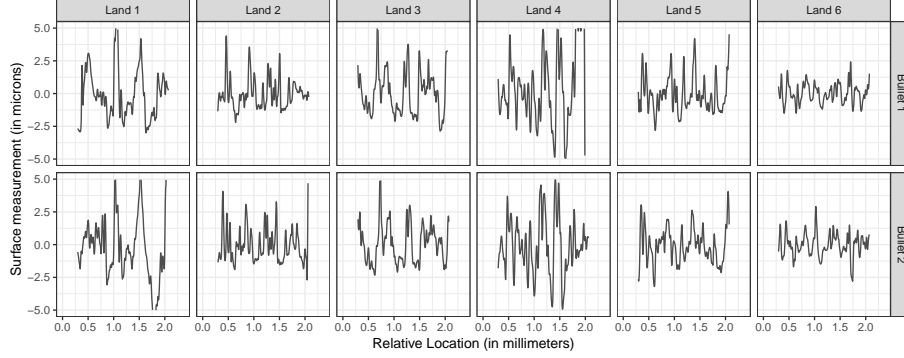


Figure 1: Signatures of all six lands of bullets 1 and 2 from barrel 1 of the Hamby set 44. Matching lands between the two bullets have been placed above each other.

(c) Remove bullet curvature using a non-parametric smooth.

A signature for a land engraved area is then defined as the sequence $S(x)$, $x = 1, \dots, I$, where I is the number of observed locations across the base of the bullet. Figure 1 shows a set of six signatures corresponding to the six land engraved areas of bullets 1 and 2 from barrel 1 of set 44 of the Hamby study [15].

Once signatures have been extracted from stable regions with expressed striae, they must be aligned in order to assess their match strength. Maximized cross correlation is used to pair-wise align signatures [13, 11].

The cross correlation (CC) function between two signatures $S_1(x)$ and $S_2(x)$ is defined as

$$CC(S_1, S_2, z) = \text{cor}(S_1(x), S_2(x + z))$$

where x and $x + z$ are integer values appropriately defined within the domains of S_1 and S_2 , z is the lag between the two sequences, and $\text{cor}(\cdot, \cdot)$ is the Pearson correlation coefficient. When signatures contain missing values, pairwise complete observations are used to calculate the cross correlation. The lag z used to achieve maximum CC is used to determine the best alignment.

1.2. Statistics for Matching Aligned Signatures

Using the lag determined for aligning two signatures, other quantitative features describing the similarity of the two signatures can be extracted. Numerical features such as cross correlation and Euclidean distance can be computed from the aligned signatures alone. Other features include characteristics of the signature that firearms and toolmark examiners base their visual assessment on, such as striae depth, number of matching striae, and the number of consecutively matching striae (CMS), one of the measures firearms examiners use to describe the strength of a match quantitatively.

In order to evaluate CMS, we must first identify peaks and valleys in each of the signatures, then determine whether these peaks and valleys overlap sufficiently. Features which firearms examiners assess visually depend on the identification of these extrema. If there are at least six consecutive matching striae, the bullets are considered to be a good match in practice [16, 17]. CMS in this setting are counted as the number of consecutively matching peaks in signatures of two aligned lands. Peaks – rather than peaks and valleys – are used for this determination in part because it is difficult to visually assess valleys except as relative to peaks. Note that with automatic alignment based on the maximum cross correlation rather than matching striation marks, it is possible for two lands to have a consecutive matching striae of zero.

Cross correlation can also be used to assess match strength between two aligned signatures. Cross correlation varies between -1 and 1, but as the alignment is based on the maximized cross correlation value, in practice the cross correlations for aligned signatures are generally positive. Cross correlation values which are close to 1 indicate similarity between signatures and with it may indicate that the two signatures come from the same source. Low cross correlation values indicate signatures which may originate from different sources.

The random forest presented in Hare et al. [13] is based on a combination of multiple characteristics, including cross correlation, number of matching striae, and number of consecutively matching striae. The output of the random forest is a score between 0 and 1 representing the algorithm’s assessment of the strength

of the match, where scores close to 1 indicate it to be highly likely that the aligned signatures come from the same source, while scores close to 0 indicate that the aligned signatures are highly likely to originate from different sources.

1.3. Interpreting the Algorithm Score

The features described above are each individually designed to separate same-source and different-source signatures: for instance, same-source signatures would have high cross correlation values and different-source signatures would have low cross correlation values. An automated algorithm selects a cutoff value, introducing a barrier between identification and exclusion.

This threshold system introduces a binary classification: identification or exclusion. This definition of error is more stringent than the AFTE Theory of Identification, in that it does not allow for inconclusive results. This increased rigor may increase the error rate of the model when compared to examiner error rates, but we expect that the increased information available from the 3D measurement data will compensate for this loss. A binary decision model is also easier to interpret and provides more clear-cut, definitive results.

We can enumerate characteristics of an ideal similarity scoring mechanism:

(R1) Monotonicity: a higher score is indicative of higher similarity between a pair of bullets, in particular, similarity scores of same-source pairs of bullets are higher than different-source pairs.

(R2) Stability: the same score leads to the same conclusion.

In particular, requirement (R2) would imply that the same threshold value would be used for all comparisons of a certain scoring mechanism. If the same-source and different-source similarity score distributions overlap, setting a threshold value will introduce classification errors.

1.4. Identification Errors and Algorithm Evaluation

When evaluating algorithm performance, it is useful to systematically assess the set of possible outcomes. If ground truth and the algorithm's prediction

Results	Ground Truth	
	Same Source	Different Source
Exclusion	Missed Identification	Correct Exclusion
Identification	Correct Identification	Wrong Identification

Table 1: Possible outcomes of an examination of two pieces of evidence. Correct decisions are shown in the top-right and bottom-left corners; incorrect decisions in the top-left and bottom-right corners.

match, we have either a correct identification or a correct exclusion. If ground truth and the algorithm’s prediction do not match, we distinguish between two types of errors, which we will refer to as **wrong identification** and **missed identification**. Wrong identifications are those in which two bullets from different sources are determined to be from the same source. Missed identifications are those in which two bullets from the same source are determined to be from different sources. The full range of possible outcomes is shown in Table 1.

In machine learning and statistics, it is common to evaluate algorithms using sensitivity and specificity; these concepts are related to the error rate. **Sensitivity** is defined as the proportion of actual positives that are correctly identified; that is, when ground truth is same source, the sensitivity is the proportion of correct identifications. **Specificity** is defined as the proportion of actual negatives that are correctly identified; that is, when ground truth is different source, the specificity is the proportion of correct identifications. The combination of sensitivity and specificity is sufficient to describe the reliability of an algorithm.

When evaluating the Hare et al. [13] algorithm, we will describe errors as a percentage of correct evaluations. Statistics that do not translate to percent-form interpretation, such as area under the curve (AUC), will be presented as a decimal.

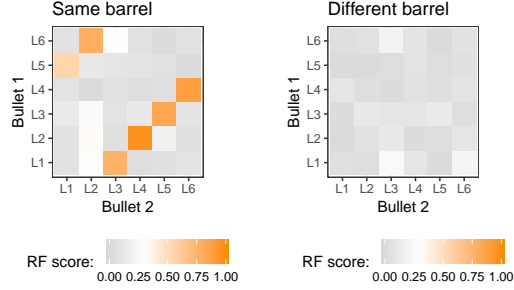


Figure 2: Overview of land-land matching scores for two pairs of bullet-bullet comparisons. On the left the two bullets are known to come from the same source (barrel) on the right, the bullets are from two different sources (barrels).

1.5. Assessing Bullet-to-Bullet Match Strength

Each of the scoring methods we have described is computed on a land-to-land basis. While these comparisons are useful, the question of interest typically involves the entire physical object (e.g. all lands on a single bullet), and a conclusion on same or different source should be reached based on the evidence of all lands of the bullets.

Land-to-land comparisons lead to a whole set of scores for bullet-to-bullet comparisons. Figure 2 shows two matrices of scores for two pairs of bullet-to-bullet matches. On the left, a matrix is shown that is typical for scores from two bullets from the same barrel. On the right, values for a pair of known non-matching bullets are shown.

When imaging bullets, operators scan one land at a time in a clockwise (left twisted rifling) or anti-clockwise (right twisted rifling) sequence. The order in which scans are acquired is kept as meta-information. Let us assume that lands on a bullet are labelled ℓ_i with $i = 1, \dots, p$, where p indicates the number of lands a bullet has, as determined by the rifling of a barrel. For all of the bullets considered here the number of lands p is 6. A match between a pair of lands from two bullets therefore results in an expected additional $p - 1$ matches between pairs of lands. These lands are also expected to be in a sequence, i.e. if there is a match between lands ℓ_i on bullet 1 and ℓ_j on bullet 2, we also

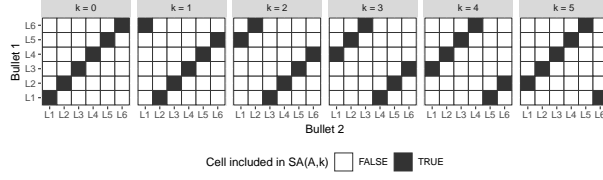


Figure 3: Sketch of all six land-to-land sequences between two bullets with six lands.

expect lands $\ell_{i \oplus s}$ and $\ell_{j \oplus s}$ to match for all integers s , where \oplus is defined as $a \oplus b \equiv ((a + b - 1) \bmod p) + 1$. This relationship gives rise to the sequence average maximum (SAM) to quantify a bullet-to-bullet match.

Definition (Sequence Average and its Maximum)

Let A be a square real-valued matrix of dimensions $p \times p$. For the purpose of this paper, A consists of scores describing the similarity between two sets of land engraved areas. The k th sequence average $SA(A, k)$ for $k = 0, \dots, p-1$ is defined as

$$SA(A, k) = \frac{1}{p} \sum_{i=1}^p a_{i, i \oplus k}, \text{ where } i \oplus k := ((i + k - 1) \bmod p) + 1.$$

The Sequence Average Maximum [SAM, 18] of square matrix A of scores is defined as

$$SAM(A) = \max_{k=1}^p SA(A, k).$$

Looking back at Figure 2, we see that for the two bullets from the same barrel, the sequence average for $k = 2$ is higher than the other sequence averages, and also higher than the sequence averages for the other pair of bullets shown on the right of the figure. SAM scores allow us to define a single quantity for each pair of bullets that describes the similarity between these two bullets.

The sequence average maximum of the correlation between lands is used in SensoComp [19] to capture the similarity between bullets. The correlation based SAM score has also been called the 'average correlation calculated at the max phase' in Chu et al. [8].

In this paper, we validate the algorithm described in Hare et al. [13] on three external test sets. We assess the results on each set and evaluate whether the minimal requirements described above are fulfilled for bullet-to-bullet SAM scores. We compare the random forest algorithm’s performance to the performance of other suggested measures for quantitative assessment of match strength, such as cross correlation and consecutive matching striae.

2. Validation Sets

Random forests [14] have a built-in internal testing mechanism to prevent potential overfitting. Errors reported by random forest algorithms are based on these internal test sets. However, there is some discussion on whether or in which situations these errors are biased [20, 21]. Neither of these papers addresses another issue with internal test sets: internal test sets are constructed to have the same distribution as the training data (apart from sampling variability). For any machine learning method, a true benchmark of the performance of an algorithm requires testing its performance on external test data. Good performance (in terms of wrong and missed identifications) on external test data validates a ML algorithm. External data also allows an assessment of the algorithm’s sensitivity to distributional changes as well as testing the algorithm’s robustness by going outside the parameters of the training data.

The algorithm in Hare et al. [13] is trained on scans from Hamby sets 252 and 173 made available through Zheng [22]. Set 173 was originally published as Hamby set 44. This mis-labeling has been corrected now. The scans used to train the model were taken at 20 fold magnification for a resolution of 1.5625 microns per pixel. While magnification is generally of interest in microscopy, ****resolution**** – usually measured in microns per pixel – is of more interest for 3D topographic measurements, as it determines the operative level of available data.

For the validation of the automatic matching algorithm we are considering three validation sets of – what should be – increasing difficulty level:

- **Hamby set 44** is one set of the Hamby study [15]. Each Hamby set consists of a total of 35 bullets fired through ten consecutively manufactured barrels of Rugers P85. Each set consists of 20 known bullets (two from each of the ten barrels) and 15 questioned bullets of unknown origin. Note that all Hamby sets are closed sets; that is, all questioned bullets are fired through one of the ten barrels. The ammunition used for this set were 9 mm Luger 115 Grain Full Metal Jacket from the Winchester Ammunition Company.
- **Phoenix PD** Tylor Klep from Phoenix PD provided sets of known test fires and questioned bullets: the set of known bullets consists of three test fires (B1, B2, B3) from each of eight different, consecutively manufactured Ruger P-95 barrels (A9, C8, F6, L5, M2, P7, R3, U10). Ten questioned bullets were provided (B, E, H, J, K N, Q, T, Y, Z). This set is an open set; that is, it is not known in advance whether all (or any) of the questioned bullets are fired from the known barrels. In fact, the results will show that three of the questioned bullets were fired from three different barrels not included in the knowns. These three barrels correspond to an eleventh Ruger P-95, a Ruger P-95C and a Ruger P-85. All bullets fired for this study are American Eagle 9mm copper jacketed Luger. Land engraved areas for each of the six lands of each bullet were scanned by Bill Henderson (Sensofar).
- **Houston FSI** This study was set up by Melissa Nally and Kasi Kirksey from FSI Houston. Three test sets based on ten consecutively rifled Ruger LCP barrels (A, B, C, D, E, F, G, H, I, J) and three other, non-consecutively rifled Ruger LCP barrels (R1, R2, R3). Each test set consists of three test fires each from five consecutively rifled barrels. Additionally, ten questioned bullets are provided for each kit. The ammunition used in both test fires and the questioned bullets were Remington UMC 9mm Luger Full Metal Jackets. All three of the test sets are open; that is, not every one of the questioned bullets is fired from the five known barrels

in each of the test set. The structure of these three sets is similar to a forthcoming study by Nally and Kirksey, but the results here come from preliminary test sets made available to us.

Scans of all land engraved areas for the validation data were taken on a Sensofar Confocal Light microscope at 20x magnification resulting in a resolution of 0.645 microns per pixel. If not indicated otherwise, scans were taken at the Roy J Carver high resolution microscopy lab at Iowa State University.

Case studies were chosen such that difficulty for the matching algorithm increases with each case study: Hamby set 44 is part of the Hamby study. The algorithm in Hare et al. [13] was trained on sets 173 and 252, so the bullets in Hamby 44 are of the same type of ammunition and are fired through the same barrels as the bullets in the training set. The Phoenix PD set uses Ruger P-95 barrels, which are different from the barrels in the Hamby sets used for training the algorithm, but the barrels are rifled similarly to the Ruger P-85 barrels. The Houston sets use Ruger LCP barrels. These barrels are first rifled traditionally, i.e. similar to the Ruger P-85, but are treated to a secondary round of heating after rifling. This second round of heat exposure has the potential to introduce some subclass characteristics, i.e. anomalies that are shared between different barrels manufactured in this manner. This should make the automatic matching harder, and in particular, is expected to complicate the classification as different-source land engraved areas/bullets. In all three studies, the scan resolution is much higher than the scans used to train the algorithm in Hare et al. [13]; this difference also provides a test of the algorithm’s ability to generalize to scans taken at different resolutions.

3. Results

3.1. Hamby Set 44

Figure 4 shows an overview of all scores from pairs of questioned bullets with all other bullets. On the left of Figure 4 there are ten strips labelled 1 through 10. These strips correspond to known barrels 1 through 10. Each barrel

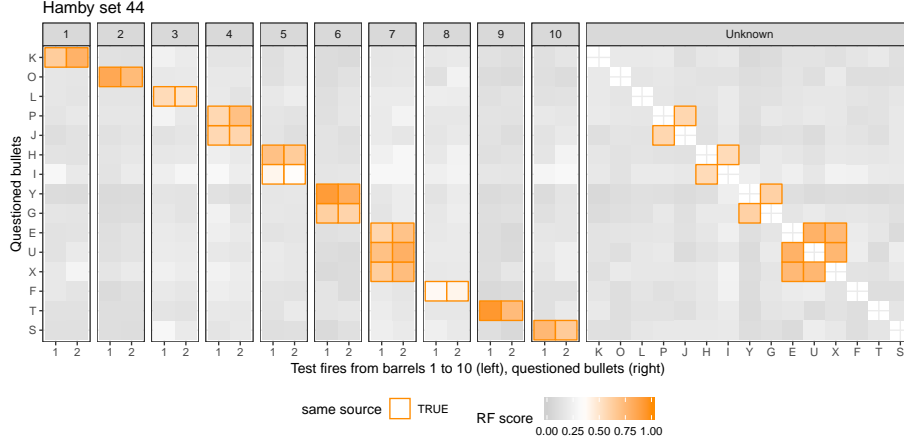


Figure 4: Hamby set 44: overview of all bullet-to-bullet matches between all pairs of questioned bullets (y-axis) and known test fires from ten barrels (x axis).

was test-fired twice, so each of the questioned bullets (shown along the y axis) matches two bullets fired from a known barrel. Colored tiles are used to show the strength of the comparison; light grey colors correspond to low similarity scores, dark colors correspond to high similarity scores. Ground truth is encoded in this figure as a thin, dark, colored frame for all pairs of same-source bullets. Ideally, we want to see one barrel with two dark-filled, dark-framed tiles for each questioned bullet, and light grey tiles for all other barrels, indicating a match between a questioned bullet and a single barrel. This expectation is met for all questioned bullets, i.e. the automatic matching identifies the correct barrel for all questioned bullets. For two of the questioned bullets, 'I' and 'F', the similarity scores to the matching barrels are considerably smaller than for the other questioned bullets.

The right side of Figure 4 shows the relationship between all pairs of questioned bullets. Note that questioned bullets are not compared to themselves, leaving white squares on the diagonal. Some of the questioned bullets match the same barrel, e.g. questioned bullets 'P' and 'J' both match barrel 5. Therefore bullets 'P' and 'J' also match each other in the square on the right hand side.

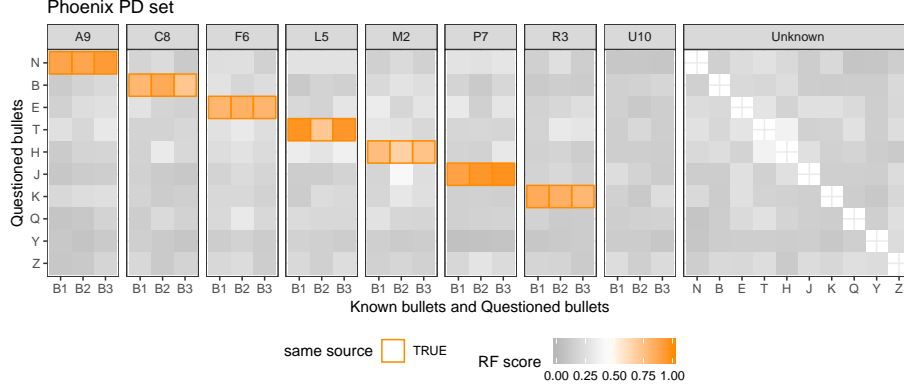


Figure 5: Phoenix study: overview of all bullet-to-bullet matches between all pairs of questioned bullets (y-axis) and known test fires from ten barrels (x axis). The order of the bullets on the y axis is determined by matching barrel.

3.2. Phoenix PD

Figure 5 shows an overview of similarity scores for all pairs of questioned bullets and test fires. The color encoding is the same as in the previous figures. Here, we see that questioned bullets either match all 3 bullets fired out of the same barrel for exactly one of the barrels or none of the known barrels. Questioned bullets 'Q', 'Y', and 'Z', do not match any of the known barrels. None of the test fires from barrel U10 match any of the questioned bullets. This is a sign of the previously mentioned open set characteristic of the study. Once again, the results from automatic matching correctly pair questioned bullets with their corresponding barrels.

3.3. Houston FSI

Figure 6 shows an overview of the three sets of scores for all pairs of questioned bullets with all other bullets in the set. In set 1, five of the questioned bullets can be matched to known bullets in the set. None of the questioned bullets in set 1 match bullets fired from barrels KC or KD. Additionally, two of the questioned bullets which do not match any of the barrels in the set match each other (U28 and U37). In set 2, four of the questioned bullets can be matched

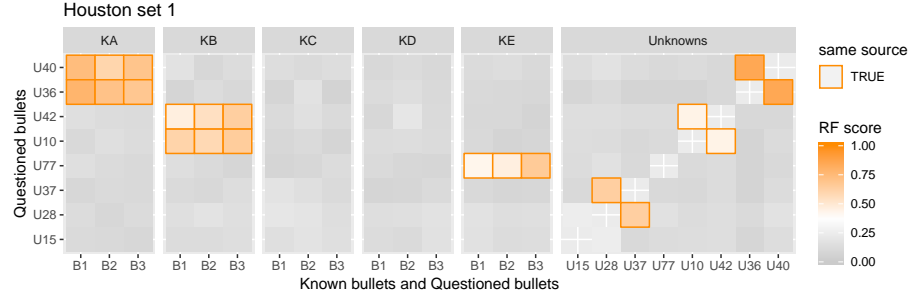
to known bullets in the set; of the additional four questioned bullets which do not match any known bullets, two were fired from the same barrel (bullets U34 and U73). In set 3, six of the questioned bullets match known bullets, and the remaining two questioned bullets match each other (bullets U14 and U45) but do not match any known bullets in the set. While all bullet-to-bullet matches are correctly identified, and no known matches are missed, it is obvious from the generally lighter shades of the tiles corresponding to matching bullets in Figure 6(b) that the algorithm is not performing as well on set 2 as it performs on sets 1 and 3.

4. Evaluating the Random Forest Algorithm

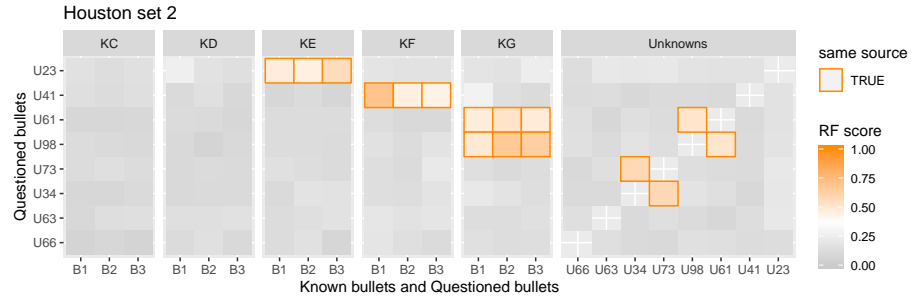
The random forest SAM scores correctly separate known matches from known non-matches, using SAM scores to aggregate land-to-land scores into a bullet-to-bullet comparison. With no errors at the bullet-to-bullet level, we can now evaluate the scores in light of (R1) and (R2). In this section, we will compare the random forest scores to cross correlation scores; cross correlation is one of the components of the random forest, but has also been used to quantify the strength of a match in its own right [8]. In addition, both scoring methods use the same scale (0 to 1) and are continuous along that interval; other evaluation methods, such as consecutive matching striae are discrete and more difficult to directly compare to continuous measurements.

4.1. Comparison of SAM scores

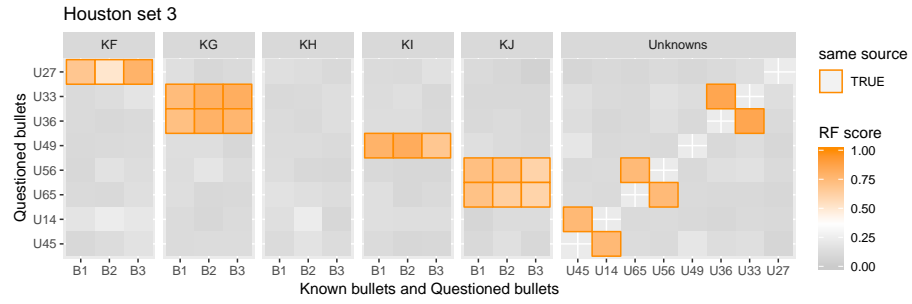
Figure 7 shows density curves for the scores bullet-to-bullet matches of each of the studies, comparing the SAM scores based on cross correlation (top) and the random forest score (bottom). Color indicates ground truth. Small vertical lines below the x -axis indicate observed scores in a **rug-plot**. The plot shows that for all three case studies and both continuous measures, i.e. SAM scores based on cross correlation as well as SAM scores based on the random forest scores, all similarity scores are higher for pairs of bullets from the same source,



(a) Set 1



(b) Set 2



(c) Set 3

Figure 6: Overview of matching scores for all pairs of questioned bullets to known bullets from five barrels (on the left) and questioned bullets to themselves (tiles on the right).

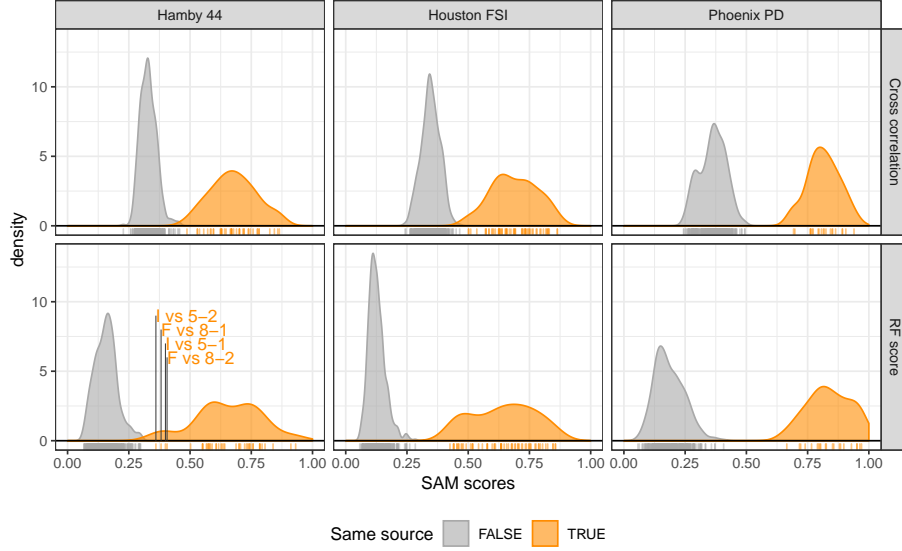


Figure 7: Density curves of similarity scores from cross correlation (top) and RF scores (bottom). Different colors indicate same source versus different source. Ideally all scores of different source pairs should be much lower than scores for same source pairs.

i.e. bullets fired through the same barrel, than pairs of bullets from different sources (fired through different barrels).

According to Figure 7, neither cross correlation nor random forest score fulfill both of the minimal requirements laid out in the introduction. Both measures fulfill (R1), i.e. all scores of different-source pairs are lower than scores of same-source pairs for SAM scores based on cross correlation and RF scores. However, SAM scores based on the multivariate random forest score show better separation between same-source and different-source scores than SAM scores based on cross correlation. This can be seen from the larger horizontal distance between the modes (peaks) of the density curves of RF based SAM scores compared to the cross correlation SAM scores.

Regarding requirement (R2), Figure 7 shows that a single cutoff value does not exist that separates same-source scores from different-source scores across all three studies without introducing errors.

There are several approaches for potential improvements: we can try to fine tune the matching algorithm to take make and model of the firearm and ammunition used into account or expand the training base of the algorithm to include a wider variety of makes and models. The downside of either of these options is that we would need to considerably expand the database used for training. Another solution would be to augment the aggregation used to get from land-to-land scores to bullet scores: SAM scores only take the scores of the maximum sequence into account and ignore scores associated with pairs of land engraved areas from different sources. Those scores might be useful in establishing a baseline that better expresses similarity between pairs of bullets. One approach that uses scores for both same source pairs and different source pairs are score-based likelihood ratios [23, 24].

4.2. Digging deeper: comparison of land-to-land scores

Examining land-to-land scores provides more details about the matching algorithm’s performance. Figure 9 shows density curves for each study. There is some overlap between known matching and known non-matching land-to-land scores; in many cases the random forest scores for these overlapping values are more extreme than the corresponding cross correlation scores. This suggests that the matching algorithm is sensitive to the presence of LEAs which have low land-to-land match scores, such as the matches of bullets ‘I’ and ‘F’ in the Hamby 44 study. Comparing this to the bullet-to-bullet score density plot shown in Figure 7, we see that there is much greater separation between the same-source and different-source densities for each study in the bullet-to-bullet comparisons than in the land-to-land comparisons, that is, one or two poorly matching lands does not prevent the bullet from showing a matching score, though if there are weak matches on several lands, such as the marked Hamby 44 comparisons in Figure 7, the overall score may be affected.

Several lands in Hamby set 44 have major deficiencies, such as ‘tank rash’ (a collision of the bullet after exiting the barrel with a surface causing markings on top of the striations from the barrel) or extreme pitting (holes caused by direct

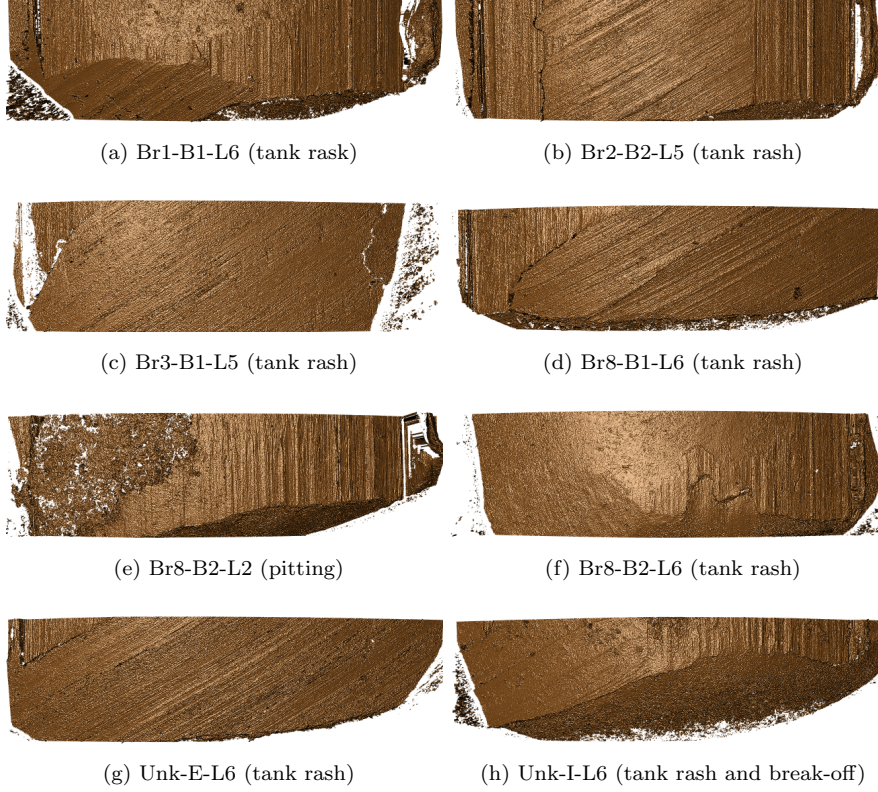


Figure 8: Overview of bullet lands with prominent deficiency such as tank rash or extreme pitting in the Hamby-44 study.

contact with burning gun powder). Figure 8 gives an overview of all lands of affected bullets in the Hamby-44 set. These issues affect the algorithm's ability to identify a stable signature, which impacts the subsequent match scores at the land-to-land and bullet-to-bullet level.

Three lands from bullets known to be fired from barrel 8 are affected, producing low scores for barrel 8 in Figure 4. We also see that questioned bullet 'T' is affected, explaining some of the low similarity scores for this bullet. None of the lands of any of the bullets in the other studies are affected by tank rash in a similar manner.

When evaluating classifier performance, it is common to use a receiver op-

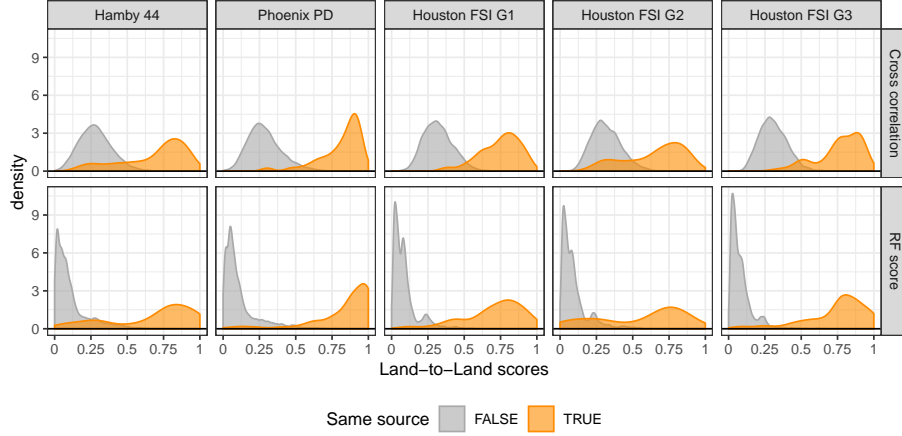


Figure 9: Density curves of land-to-land similarity scores from cross correlation (top) and RF scores (bottom). Different colors indicate same source versus different source for each land. RF scores for different source comparisons are generally well below 0.5. Some same-source comparisons are also below 0.5 for RF scores, indicating potential problems with at least one of the lands involved in the comparison.

erating characteristic (ROC) curve, which plots the percent of wrong identifications against the percent correct identifications for each possible value of the cutoff between the two classes (in this case, known match and known non-match). Land-to-land scores are not perfectly separated, we can therefore use ROC curves to distinguish between the performance of the different methods and studies. The ROC curves for the land-to-land scores are shown in Figure 10 (a). A perfect classifier would have 100% correct identifications and 0% wrong identifications, e.g. the ROC curve would be a right angle at (0, 100). Classifiers with better performance will be closer to this corner of the plot. A random classifier would have an ROC curve that was a straight diagonal line through (0, 0) and (100, 100).

In Figure 10a, the Houston FSI G1 and G3 curves and the Phoenix PD curve show excellent performance; Houston FSI G2 and Hamby 44 show still a very good performance. Area under the curve (AUC) values, which summarize ROC curves, are shown in Figure 10b. AUC values are useful for differentiating

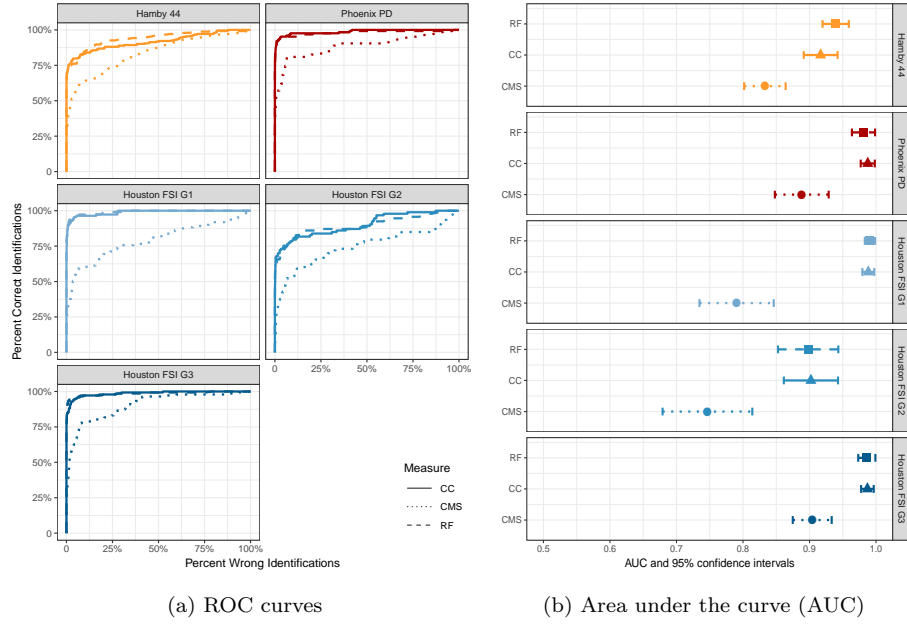


Figure 10: ROC curves and AUC for each (set) of the studies for the random forest score (RF), cross correlation (CC) and consecutively matching striae (CMS). As seen in the ROC curves, the algorithm performs the least well on set 2 of the Houston FSI study. Based on AUC, the overall performance on all sets is very good to excellent for cross correlation and the random forest score. At the land level there is no significant difference in prediction power between RF score and cross correlation.

between poor, good, and excellent model performance, but are not particularly useful when determining which of several models with approximately the same level of performance should be used [25]. Hamby 44 and Phoenix PD are the only studies with a somewhat noticeable difference between the AUC score for RF compared to cross correlation, but even those differences are not statistically significant.

Figure 11 shows an overview of Equal Error and their thresholds based on the ROC curves of Figure 10. Equal Errors are errors when sensitivity and specificity of a test are equal, i.e. we see the same percentage of missed and wrong identifications for a land-to-land comparison. Equal errors based on CMS of two lands are significantly higher than equal errors based on cross correlation or the RF score. At best, RF score and cross correlation have an equal error of around 5% for land-to-land comparisons. Interestingly, for consecutively matching striae, equal errors are reached for just two matching striae, which from a practical point of view is not sustainable. A CMS of 6 or above is an established cutoff used in identifications; when applied to the land-to-land comparisons, it is much more likely to result in a missed identification than a wrong identification. A threshold of 6 leads to 21 wrong identifications out of 49,286 different-source comparisons (0.04% errors, or about 4 errors for each 10,000 evaluations); when applied to same-source comparisons, 517 of 715 same-source lands were incorrectly classified as exclusions, for an error rate of 80.6%.

5. Discussion and Conclusions

At the beginning of this study, we anticipated Hamby 44 would be the easiest set to work with because of its similarity to the sets used to train the random forest algorithm. In a surprise turn of events, it was the hardest to solve, in part because of damage to the bullets that obscured LEA striae. It has been shown Hare et al. [26] that parts of lands can be used for successful identifications. The algorithm proposed by Hare et al. [13] is not capable of automatically detecting parts of lands with well-expressed striae. It may be useful to couple

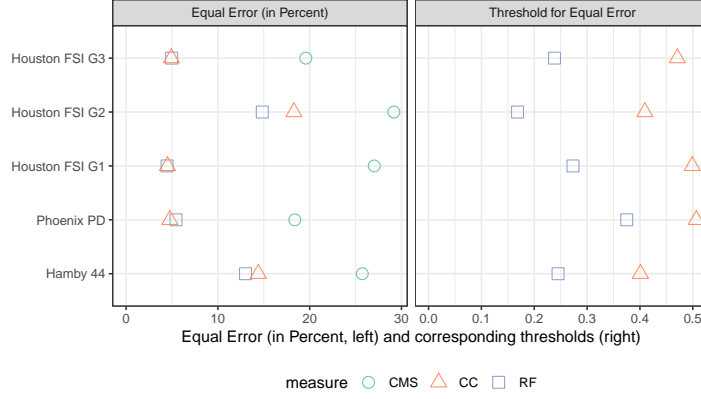


Figure 11: Equal Error (in Percent) and corresponding thresholds for all measures (RF score, Cross correlation and CMS) under all five (sub-)studies. CMS has an equal error threshold of 2 for all studies. CMS is significantly worse than both Cross correlation and the RF score. At best, Cross correlation and RF score have equal errors for land-to-land comparison of around 5%.

the Hare et al. [13] algorithm with an algorithm which assesses the quality of the input data and determines which portions of the data to use for comparison. This would emulate the process used by examiners, who first assess the quality of the evidence and whether there is enough information present to attempt a match. One major disadvantage of an automated algorithm is that all of the decisions humans make (recognizing degraded land areas, excluding those areas from consideration, matching only the remaining areas) must be explicitly characterized; however, this explicit characterization means that the process can be scientifically validated to a much higher degree than the human perceptual process.

All of the studies presented here involve the same type of firearm. We see some variations in scores across different models of Rugers, and it is probable that we will see even bigger variations with different brands of firearms that vary in number of lands and land length. We also know that different firearms and ammunition combinations mark differently well [27, 28]. Rugers are among the firearms that mark very well, so we would correspondingly expect to see lowered

similarity scores for other firearms. The matching algorithm works on matching lands. Barrels with polygonal rifling, such as Glockes, do not introduce well defined lands on bullets, which makes an application of the matching algorithm impossible.

The random forest matching algorithm presented in Hare et al. [13] does perform well on three different external test sets. While a single cutoff value cannot be used to distinguish matches and non-matches across the different test sets, the algorithm makes no errors when a set-specific cutoff value is used on bullet-to-bullet scores aggregated using sequence average maximum. This performance is on par with the performance of forensic examiners. For a future round of the Houston study, we are planning to compare the algorithmic performance directly with scores given by forensic toolmark examiners. **XXX need to ask XXX** The algorithm's score is not intended to replace an examiner's judgment; instead, it provides a complementary tool to measure and quantify the strength of an identification. By validating the algorithm on external test sets, we have demonstrated that the method can be generalized to different types of ammunition and is not overly sensitive to small differences in rifling procedure. Future studies can and should generalize this to a wider range of external test sets to establish the limits of the algorithm's generalizability.

6. References

References

- [1] AFTE Criteria for Identification Committee, Theory of identification, range striae comparison reports and modified glossary definitions, AFTE Journal 24 (1992) 336–340.
- [2] A. A. Biasotti, A statistical study of the individual characteristics of fired bullets, Journal of Forensic Sciences 4 (1959) 34–50.
- [3] J. Davis, An Introduction to Tool Marks, Firearms and the Striagraph,

- Charles C. Thomas Publisher, Limited, 1958. URL: <https://books.google.com/books?id=a70ZHBJYIAQC>.
- [4] G. Y. Gardner, Computer Identification of Bullets, *IEEE Transactions on Systems, Man, and Cybernetics* 8 (1978) 69–76. doi:10.1109/TSMC.1978.4309834.
 - [5] J. De Kinder, M. Bonfanti, Automated comparisons of bullet striations based on 3D topography, *Forensic Science International* 101 (1999) 85–93. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0379073898002126>. doi:10.1016/S0379-0738(98)00212-6.
 - [6] B. Bachrach, Development of a 3D-based Automated Firearms Evidence Comparison System, *Journal of Forensic Sciences* 47 (2002) 1555–1571. URL: <http://www.astm.org/doiLink.cgi?JFS15557J>. doi:10.1520/JFS15557J.
 - [7] F. Xie, S. Xiao, L. Blunt, W. Zeng, X. Jiang, Automated bullet-identification system based on surface topography techniques, *Wear* 266 (2009) 518–522. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0043164808002998>. doi:10.1016/j.wear.2008.04.081.
 - [8] W. Chu, J. Song, T. Vorbuerger, J. Yen, S. Ballou, B. Bachrach, Pilot Study of Automated Bullet Signature Identification Based on Topography Measurements and Correlations, *Journal of Forensic Sciences* 55 (2010) 341–347. URL: <http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1556-4029.2009.01276.x>. doi:10.1111/j.1556-4029.2009.01276.x.
 - [9] J. Lu, S.-H. Wu, K.-C. Yang, M. Xia, Automated bullet identification based on striation feature using 3D laser color scanner, *Optik - International Journal for Light and Electron Optics* 125 (2014) 2270–2273. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0030402613014253>. doi:10.1016/j.ijleo.2013.10.065.

- [10] L. Ma, J. Song, E. Whitenton, A. Zheng, T. Vorburger, J. Zhou, NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets, *Journal of Forensic Sciences* 49 (2004) 1–11. URL: <http://www.astm.org/doiLink.cgi?JFS2003384>. doi:10.1520/JFS2003384.
- [11] T. Vorburger, J.-F. Song, W. Chu, L. Ma, S. Bui, A. Zheng, T. Renegar, Applications of cross-correlation functions, *Wear* 271 (2011) 529–533. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0043164810001407>. doi:10.1016/j.wear.2010.03.030.
- [12] W. Chu, R. M. Thompson, J. Song, T. V. Vorburger, Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria, *Forensic Science International* 231 (2013) 137–141. URL: <http://linkinghub.elsevier.com/retrieve/pii/S037907381300248X>. doi:10.1016/j.forsciint.2013.04.025.
- [13] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, *Ann. Appl. Stat.* 11 (2017) 2332–2356. URL: <https://doi.org/10.1214/17-AOAS1080>. doi:10.1214/17-AOAS1080.
- [14] L. Breiman, Random Forests, *Machine Learning* 45 (2001) 5–32. URL: <https://doi.org/10.1023/A:1010933404324>. doi:10.1023/A:1010933404324, 40064.
- [15] J. E. Hamby, D. J. Brundage, J. W. Thorpe, The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries, *AFTE Journal* 41 (2009) 99–110.
- [16] A. Biasotti, J. Murdock, Firearms and toolmark identification: Legal issues and scientific status, *Modern Scientific Evidence: The Law and Science of Expert Testimony* (1997) 124–151. Citation Key: biasotti1997firearms bibtex[publisher=West Publishing Co. St. Paul].

- [17] R. Nichols, The scientific foundations of firearms and tool mark identification???a response to recent challenges, California Association of Criminalists News (2006) 8–27. 00013.
- [18] Sensofar, SensoMATCH bullet comparison software, 2017.
- [19] {Sensofar Metrology}, A Step Forward in 3d Firearms Identification, 2018. URL: <https://www.sensofar.com/wp-content/uploads/2018/06/PR180601-SensoCOMP-SensoMATCH-Forensics.pdf>, 12640.
- [20] M. W. Mitchell, Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters, Open Journal of Statistics (2011) 205–211. doi:10.4236/ojs.2011.13024.
- [21] S. Janitza, R. Hornung, On the overestimation of random forest’s out-of-bag error, PLOS ONE 13 (2018) 1–31. URL: <https://doi.org/10.1371/journal.pone.0201904>. doi:10.1371/journal.pone.0201904.
- [22] X. A. Zheng, NIST Ballistics Toolmark Research Database (NBTRB), 2016. URL: <https://tsapps.nist.gov/NRBD>.
- [23] S. Bunch, G. Wevers, Application of likelihood ratios for firearm and toolmark analysis, Science & Justice 53 (2013) 223–229. URL: <http://dx.doi.org/10.1016/j.scijus.2012.12.005>. doi:10.1016/j.scijus.2012.12.005.
- [24] G. S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors, Science & Justice 58 (2018) 200–218. URL: <https://doi.org/10.1016/j.scijus.2017.12.005>. doi:10.1016/j.scijus.2017.12.005.
- [25] C. Marzban, The ROC Curve and the Area under It as Performance Measures, Weather and Forecasting 19 (2004) 1106–1114. URL: <https://journals-ametsoc-org.proxy.lib.iastate.edu/doi/abs/10.1175/825.1>. doi:10.1175/825.1, 00162.

- [26] E. Hare, H. Hofmann, A. Carriquiry, Algorithmic approaches to match degraded land impressions, *Law, Probability and Risk* 16 (2017) 203–221. URL: <http://dx.doi.org/10.1093/lpr/mgx018>. doi:10.1093/lpr/mgx018.
- [27] R. S. Bolton-King, Preventing miscarriages of justice: A review of forensic firearm identification, *Science & Justice* 56 (2016) 129–142. URL: <http://www.sciencedirect.com/science/article/pii/S1355030615001185>. doi:10.1016/j.scijus.2015.11.002.
- [28] M. Bonfanti, J. De Kinder, The Influence of the Use of Firearms on their Characteristic Marks., *Association of Firearms & Tool Mark Examiners (AFTE) Journal* 31 (1999) 318–323.