Algorithmic Approaches to Match Degraded Land Impressions

Eric Hare, Heike Hofmann, Alicia Carriquiry

10/05/2017

Abstract

Bullet matching is a process used to determine whether two bullets may have been fired from the same gun barrel. Historically, this has been a manual process performed by trained forensic examiners. Recent work however has shown that it is possible to add statistical validity and objectivity to the procedure. In this paper, we build upon the algorithms explored in Automatic Matching of Bullet Lands (Hare, Hofmann, and Carriquiry 2016) by formalizing and defining a set of features, computed on pairs of bullet lands, which can be used in machine learning models to assess the probability of a match. We then use these features to perform an analysis of the two Hamby (Hamby, Brundage, and Thorpe 2009) bullet sets (Set 252 and Set 44), to assess the presence of microscope operator effects in scanning. We also take some first steps to address the issue of degraded bullet lands, and provide a range of degradation at which the matching algorithm still performs well. Finally, we discuss generalizing land to land comparisons to full bullet comparisons as would be used for this procedure in a criminal justice situation.

1 Background

Intense scrutiny has been focused on the process of bullet matching in recent years (e.g., Giannelli 2011). Bullet matching, the process of determining whether two bullets could have been fired from the same gun barrel, has traditionally been performed without meaningful determination of error rates or statistical assessments of uncertainty (National Research Council 2009). There have been some attempts towards developing mathematical and statistical approaches to bullet matching. One such attempt was the definition of CMS, the Consecutively Matching Striae (Biasotti 1959), with a cutoff of six to separate matches from non-matches. Still, rigorous assessments of the applicability of such cutoffs have not to this point been described (President's Council of Advisors on Science and Technology 2016).

Recently, several authors have addressed these well-known shortcomings. Focusing on firing pin impressions and breech faces, Riva and Champod (2014) have described an automated algorithm using 3D images that enables comparison between pairs of exemplars. Other examples of work in this and related areas include Petraco and Chan (2012), W. Chu et al. (2011), T. Vorburger et al. (2011), and others. In our approach to this problem, Automatic Matching of Bullet Lands, we used the Hamby 252 set (Hamby, Brundage, and Thorpe 2009) to train and develop a random forest in order to provide a matching probability for two bullet lands (Hare, Hofmann, and Carriquiry 2016). While the algorithm had a very strong performance on this set, some limitations were immediately clear. For instance, performance was assessed only on this single set of 35 bullets fired from a consecutively manufactured set of only ten known and 15 unknown gun barrels. Each of these bullets was part of controlled study, and the full lands were available for matching. While there were some data quality issues, this was still a near ideal test case for the algorithm.

Real world applications of bullet matching often involve the recovery of fragments of bullets from the crime scene (National Research Council 2004). Traditional features used in forensic examination work well for a full land, but there has been less investigation into their performance in the case of a fragmented land. For example, the CMS is naturally limited by the portion of the land that can be recovered and varies across manufacturers (W. Chu et al. 2011).

In this paper, we take steps to address these and other concerns. Specifically, we begin by reviewing features from the literature, computed on pairs of bullet lands, and presenting some of our own features. We propose an approach to standardize the featues, to account for the fact that only a portion of the land impression may be recovered from the crime scene. With the standardized features, we tackle two issues that were not addressed in Hare, Hofmann, and Carriquiry (2016). The first is the effect of the microscope operator on the resulting images and consequent algorithm performance. The second issue has to do with the robustness of the land matching algorithm in Hare, Hofmann, and Carriquiry (2016) relative to the degree of degradation of the questioned land impression. Finally, we describe some of the initial steps toward generalizing a matching algorithm based on land-to-land comparisons, to one based on bullet-to-bullet comparisons, as would be of interest in a real world application of these ideas.

2 Feature Standardization

To start, we introduce a standardized version of each of the features used in the matching routine proposed by (Hare, Hofmann, and Carriquiry 2016). These features are computed on *aligned pairs of bullet land impressions* rather than on individual lands. This enables us, for instance, to compute the number of matching striae between two lands. We generalize the definitions of these features to account for the possibility that we may be handling degraded bullet lands, where only fragments can be recovered. The definition of each feature is given below, where $f(t)$ represents the height values of the first profile at position t along the signature, and $g(t)$ the height values of the second. An indication of whether the feature is new since Hare, Hofmann, and Carriquiry (2016) is also given:

- **ccf** (%) is the maximum value of the Cross-Correlation function evaluated at the optimal alignment. The Cross-Correlation function is defined as $C(\tau) = \int_{-\infty}^{\infty} f(t)g(t+\tau)dt$ where τ represents the the lag of the second signature (T. Vorburger et al. 2011).
- **rough_cor** (new) (%) feature that quantifies the correlation between the two signatures after performing a second LOESS smoothing stage and then subtracting the result

from the original signatures. This attempts to model the roughness of the surface after removing structure such as waviness.
- **lag** (mm) Is the optimal lag for the ccf value.
- **D** (mm) is the Euclidean vertical distance between each height value of the aligned signatures. This is defined as $D^2 = \frac{1}{\#t}\sum_t [f(t) - g(t)]^2$. This is a measure of the total variation between two functions (Clarkson and Adams 1933).
- **sd_D** (mm) provides the standard deviation of the values of D from above.
- **signature_length** (mm) is the overall length of the smallest of the two aligned signatures.
- **overlap** (new) (%) provides the percentage of the two signatures that overlap after the alignment stage.
- **matches** (per mm) is the number of matching peaks/valleys (striae) per millimeter of the overlapping portion of the aligned signatures.
- **mismatches** (per mm) is the number of mismatching peaks/valleys (striae) per millimeter of the overlapping portion of the aligned signatures.
- **cms** (per mm) is the number of consecutively matching peaks/valleys (striae) per millimeter of the overlapping portion of the aligned signatures (Biasotti 1959, Wei Chu et al. (2013)).
- **non_cms** (per mm) is the number of consecutive mismatching peaks/valleys (striae) per millimeter of the overlapping portion of the aligned signatures.
- **sum_peaks** (per mm) is the the sum of the average heights of matched striae.

The features that are expressed on a per millimeter level are intended to support the degraded land case, as discussed earlier. Note that the computation differs slightly depending on the feature. For example, to standardize the number of matches, the first count the raw number of matching striae, and then divide this number by the length of the overlapping region of the two lands (`overlap` from above). In most cases, the overlapping region will be very close to the length of the smaller signature. But depending on the alignment, this may not always be true. This ensures that we do not punish a particular cross-comparison for having a smaller region in which matches could occur. On the other hand, the number of mismatches is divided by the total length of the two aligned signatures, since mismatched striae can occur even in the non-overlapping region of the two signatures.

The `rough_cor` or Roughness Correlation is derived by performing a second smoothing step, and subtracting the result from the original signatures. This creates a new signature which eliminates some of the overall structure, allowing global deformations to have less of an influence on the model output. Where the roughness correlation is most useful is in a scenario like Figure 1. This figure shows the alignment of profile 40977 with 47600. The top panel shows the smoothed signatures. The middle panel overlays a LOESS fit to the average of the two signatures. Finally, to derive the roughness correlation, this LOESS is subtracted from the original signature to create a new set of roughness residuals, which are then given in the bottom panel. Note that these two profiles do not match, yet the ccf is 0.7724. The roughness correlation (-0.0324) correctly indicates the lack of matching. The roughness correlation acts as a check against false positives which can arise when there are significant deformations in the overall structure, as in the case with both these profiles.

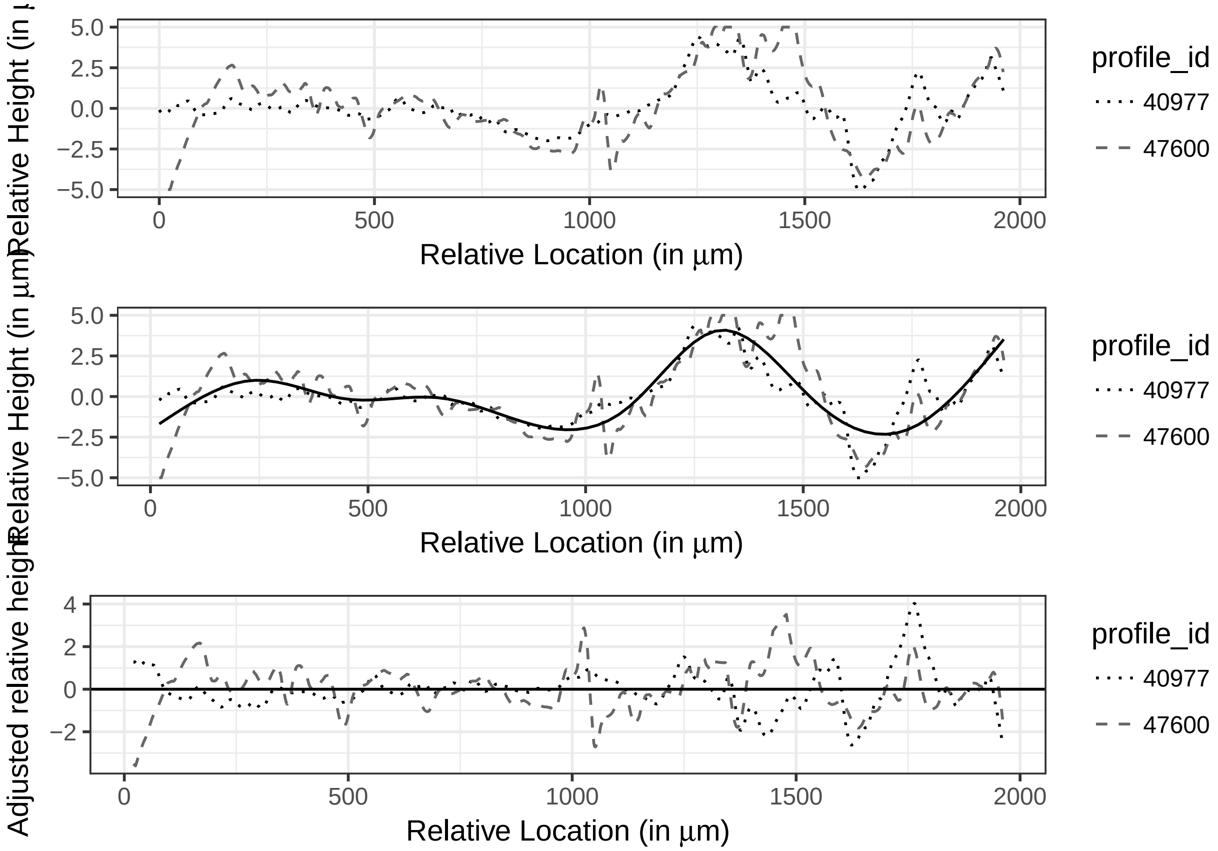

Figure 1: Alignment of profile 40977 with 47600. The top panel shows the smoothed signatures. The middle panel overlays a LOESS fit to the average of the two signatures. Finally, to derive the roughness correlation, this LOESS is subtracted from the original signature to create a new set of roughness residuals, which are then given in the bottom panel. Note that these two profiles do not match, yet the ccf is 0.7724. The roughness correlation (-0.0324) correctly indicates the lack of matching.

In a typical comparison between two profiles, such as in Figure 2, the roughness correlation does not meaningfully impact the matching probability given the presence of the ccf in the model. In this figure, we see the alignment of profile 8752 with profile 136676. In this case, the waviness or the deformation pattern in the signatures is less pronounced, and hence the resulting roughness signature resembles the original signature more closely. These profiles match, and both ccf (0.6891) and rough_cor (0.7980) provide values indicative of matching.

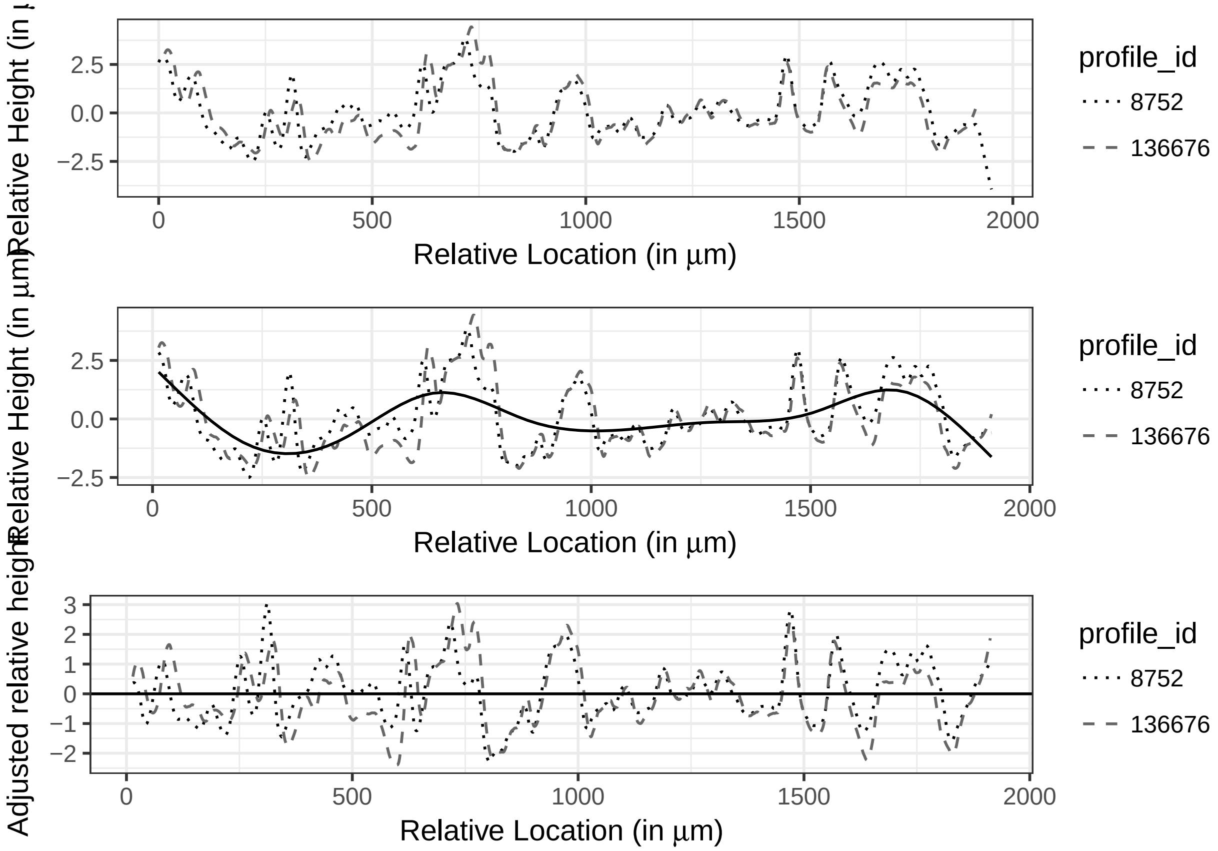

Figure 2: Alignment of profile 8752 with profile 136676. In this case, the waviness or the deformation pattern in the signatures is more minor, and hence the resulting roughness signature resembles the original signature more closely. These profiles match, and both ccf (0.6891) and rough_cor (0.7980) provide values indicative of matching.

We can observe the distributions of both CCF and the Roughness Correlation side by side, differentiating between known matches and known non-matches. Figure 3 displays this as an empirical CDF plot. It can be seen that the separation of known matches and known non-matches along both the CCF and the Roughness Correlation is quite strong and follows similar distributions (the known non-matches are relatively symmetric, while the known matches are very skewed left). However, some known non-matches with CCF values that would typically be indicative of a match have relatively lower values for the Roughness Correlation, which indicates that this feature could provide some added value when it comes to discriminating between matches and non-matches.

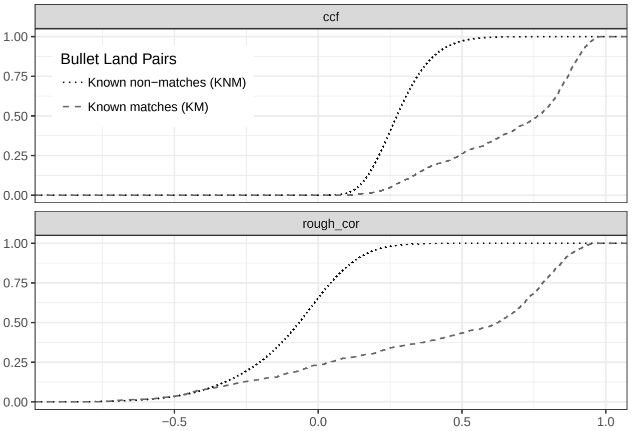

Figure 3: Empirical CDFs of the Roughness Correlation compared to the CCF for known matches and known non-matches. It can be seen that the distribution of each feature for the known non-matches is quite symmetric, while the distribution for each feature for the known matches is skewed left.

3 Model Training

Using these features, we can train a randomForest (Liaw and Wiener 2002) model which attempts to predict whether two lands match given the value of the features. There are currently three studies for bullets included in the NIST Ballistics Toolmark Research Database. Those are Hamby (Set 252), Hamby (Set 44), and Cary. For purposes of the analysis we describe in this paper, we exclude the Cary bullets from consideration, because the study was designed to assess the persistence of striation markings over a series of fires from the same barrel. Thus, every Cary bullet is a known match to every other Cary bullet. Hence, we will consider Hamby (Set 252) and Hamby (Set 44) only. This leaves us a total of 83,028 land-to-land comparisons, of which 1208 are among known matching land impressions and 81,820 are among known non-matching land impressions.

We can now train the forest using the features we defined earlier. Using the `caret` package (Jed Wing et al. 2016), we perform the following partitioning scheme. Out of the 50 barrels total, ten knowns and fifteen unknowns from each of the two Hamby sets, we hold out ten barrels randomly as a testing set, and use the remaining 40 to train the model. We repeat this procedure ten times and average the confusion matrix in order to assess the model accuracy with different holdout samples. Table 1 displays the results in the form of a confusion matrix on the test set, averaged over these ten independent random forests trained on ten random barrel subsets. It can be seen that false positives are exceedingly rare, but false negatives occur more frequently (approximately 21‘ false negative land to land comparisons on the test set, compared with an average of less than two false positives).

Result	Count
False Negative	20.6
False Positive	2.0
True Negative	3716.7
True Positive	56.6

Table 1: The average confusion matrix for the 10 random forests. It can be seen that false positives are exceedingly rare, but false negatives occur more frequently.

These results suggest that our algorithm is too conservative in predicting a match when in fact the bullets were fired from the same gun barrel. We can break down the confusion matrix by the study from which each of the two land impressions originated. Table 2 shows the average confusion matrix for the 10 random forests, broken down by study. It can be seen that Hamby252 to Hamby252 comparisons exhibit the fewest errors, while Hamby44 to Hamby44 comparisons exhibit the most errors on average. This intuitively makes some sense given the potential presence of scanner operator effects, which we address further in this section.

Study	False Negative	False Positive	True Negative	True Positive
Hamby252_Hamby252	0.38%	0.02%	97.56%	2.04%
Hamby252_Hamby44	0.45%	0.05%	98.3%	1.19%
Hamby44_Hamby44	0.88%	0.09%	97.68%	1.35%

Table 2: The average confusion matrix for the 10 random forests, broken down by study. It can be seen that Hamby252 to Hamby252 comparisons exhibit the fewest errors, while Hamby44 to Hamby44 comparisons exhibit the most on average.

4 Feature Robustness

Our goal is to assess the robustness of the previously defined features as it pertains to our bullet matching routines. This goal is both a backward looking assessment of our previous results for full land-to-land comparisons, and a forward looking one to help support the case that these can be used in the degraded land case. As a first stage to assessing this robustness, we produce parallel coordinate plots of the various features based on true positive, true negative, false positive, and false negative land-to-land matches. Figure 4 displays these plots. The means of the true positive and the true negative groups are shown as thick blue lines, respectively, in the two panels. The dashed lines represent individual land to land comparisons, with errors highlighted larger in red. It can be seen that the few false positives tendt o have anomalously high `ccf` or `matches`, while the false negatives tend have a lot of variability, though tending to also have a high `ccf` value.

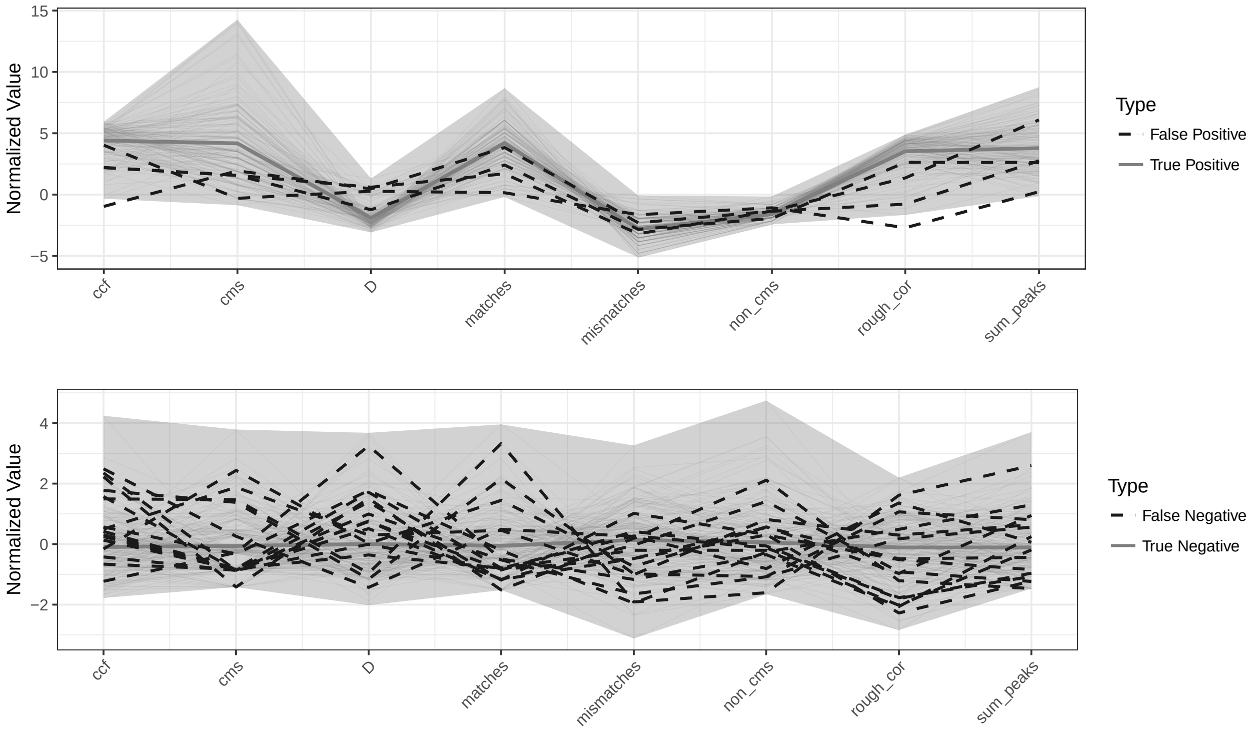

Figure 4: Parallel coordinate plot of the features based on the random forest confusion matrix for true and false positives (above), and true and false negatives (below). False positives tend to have some feature anomalously high, while false negatives exhibit quite a spread, sometimes having very large values of CCF or CMS, for example.

4.1 Operator Effects

We attempt to quantify the effect of the study on the matching probability by fitting a new random forest which is designed to predict the study from which the scans came from based on the derived features. It should be noted that the two sets of scans were performed by two trained professionals at NIST, and therefore this analysis is intended to shed light on how slightly different operating procedures for the scans may lead to varying results in algorithms derived from these scans. Ideally, if the assumption of independence between lands holds across different operators, this forest should have poor performance - The set of derived features should be relatively consistent among known matches and known non-matches regardless of the study since the Hamby data in both sets originated from the same gun barrels.

Table 3 shows the confusion matrix, with column proportions, for the random forest with study as the response. It can be seen that indeed the random forest performs poorly, as hoped, indicating that a simple model to predict the study using the features available is not enough to detect the operator effects.

Prediction \ Actual	Hamby252_Hamby252	Hamby252_Hamby44	Hamby44_Hamby44
Hamby252_Hamby252	09.93%	08.47%	011.3%
Hamby252_Hamby44	81.24%	80.82%	78.27%
Hamby44_Hamby44	08.83%	010.7%	10.43%

Table 3: Confusion Matrix (Column Proportions) for the random forest with study as the response. It can be seen that the random forest performs poorly, as hoped, indicating that a simple model to predict the study using the features available is not enough to detect the operator effects.

Figure 5 shows the distributions of the land-to-land features, faceted by whether the lands are known to be fired from the same gun barrel, across different study to study comparisons. The distributions among the known non-matches seem relatively consistent across study based on visual inspection. On the other hand, among known matches, Hamby252 to Hamby252 comparisons exhibit more pronounced features, including a higher average ccf, higher number of matches, and higher value of sum_peaks.

Though visual inspection clearly shows differences, we can more formally assess the differences between distributions with a Kolmogorov-Smirnov test. Table 4 gives the results of pairwise tests, for each feature, between different set comparisons, and between known matches compared with known non-matches. Although most of the tests are significant, looking at the raw values of the D statistic suggest that the largest effect sizes do in fact occur in comparisons with two Hamby252 lands, as the visual inspection of the boxplots also suggested.

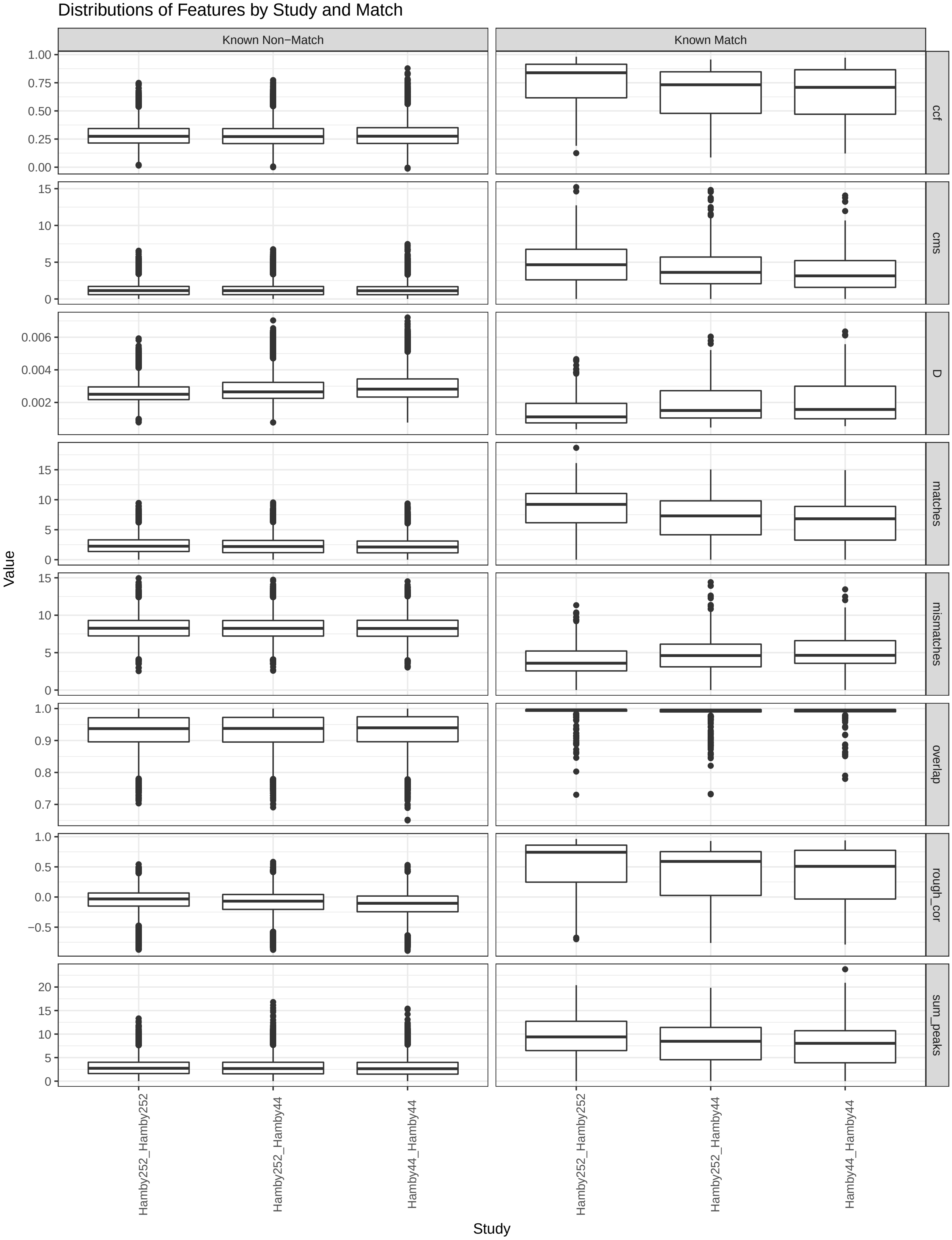

Figure 5: Distribution of the features, facetted by match, for different study to study comparisons of lands.

set1	set2	feature	matchtest	matchd	nonmatchtest	nonmatchd
H252_H252	H252_H44	ccf	< 0.0001	0.2723	1e-04	0.0189
H252_H252	H252_H44	cms	< 0.0001	0.1751	< 0.0001	0.0245
H252_H252	H252_H44	D	< 0.0001	0.2567	< 0.0001	0.1049
H252_H252	H252_H44	matches	< 0.0001	0.1933	< 0.0001	0.0327
H252_H252	H252_H44	mismatches	< 0.0001	0.2015	0.3537	0.0079
H252_H252	H252_H44	overlap	0.0492	0.0984	< 0.0001	0.0276
H252_H252	H252_H44	rough_cor	< 0.0001	0.2647	< 0.0001	0.0970
H252_H252	H252_H44	sum_peaks	8e-04	0.1426	0.0015	0.0162
H252_H252	H44_H44	ccf	< 0.0001	0.2160	< 0.0001	0.0257
H252_H252	H44_H44	cms	< 0.0001	0.2515	< 0.0001	0.0467
H252_H252	H44_H44	D	< 0.0001	0.2342	< 0.0001	0.1946
H252_H252	H44_H44	matches	< 0.0001	0.2770	< 0.0001	0.0713
H252_H252	H44_H44	mismatches	< 0.0001	0.2505	0.0414	0.0138
H252_H252	H44_H44	overlap	0.2432	0.0906	< 0.0001	0.0408
H252_H252	H44_H44	rough_cor	< 0.0001	0.2242	< 0.0001	0.1718
H252_H252	H44_H44	sum_peaks	1e-04	0.1926	< 0.0001	0.0289
H252_H44	H44_H44	ccf	0.1149	0.0883	< 0.0001	0.0259
H252_H44	H44_H44	cms	0.111	0.0888	< 0.0001	0.0262
H252_H44	H44_H44	D	0.2923	0.0724	< 0.0001	0.0906
H252_H44	H44_H44	matches	0.0603	0.0977	< 0.0001	0.0423
H252_H44	H44_H44	mismatches	0.3301	0.0700	0.1633	0.0096
H252_H44	H44_H44	overlap	0.8231	0.0465	1e-04	0.0190
H252_H44	H44_H44	rough_cor	0.2671	0.0741	< 0.0001	0.0769
H252_H44	H44_H44	sum_peaks	0.047	0.1011	0.006	0.0147

Table 4: Results for the Kolmogorov-Smirnov distributional test.

These results strongly suggest the need for controlling for more effects when performing the analysis. Specifically, microscope operator effects resulting in variations in scan quality and scan parameters seem to play a role in the utlimate performance of the matching algorithm. Land to land comparisons from Hamby252 consistently result in more pronounced expression of features among known matches, and therefore result in higher accuracy in the random forest. Rigorous procedures to ensure scan quality and consistency across operators need to be in place to minimize the effect of the study and ensure that the assumption of land to land independence is satisfied.

Another way to demostrate the study/operator effect is by observing the distribution of our algorithm's ideal cross section by study. Figure 6 gives the distributions of the ideal cross sections by study. It can be seen that the Hamby44 ideal cross sections are more likely to be close to the base of the bullet when compared to the position of the ideal cross sections in Hamby252.

Indeed, another Kolmogorov-Smirnov test confirms a significant difference in the distributions of these values ($D = 0.6239, p < 0.0001$). This result strongly suggests that the operator

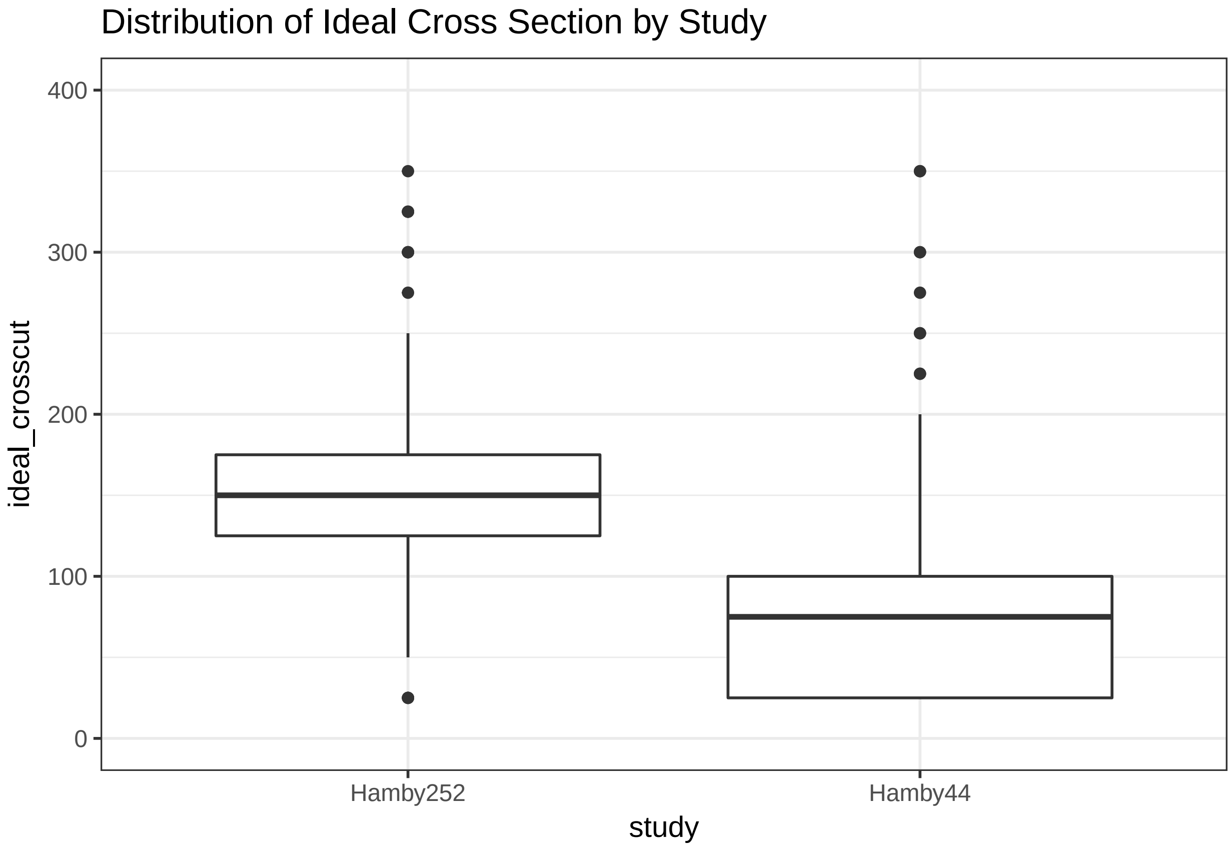

Figure 6: Distributions of the ideal cross sections by study. It can be seen that the Hamby44 ideal cross sections are much more likely to be close to the base of the bullet compared to Hamby252.

effect in the bullet scanning procedure must be taken into account in order to assume pairwise independence of bullet land scans between Hamby sets 252 and 44.

4.2 Degraded Lands

We now turn our attention to matching degraded bullet lands, in which only fragments of the land can be recovered. Because the NIST database currently contains only full bullet lands, we artificially degrade bullets under some simplifying assumptions. Essentially, we delete portions of lands to simulate the situation where we only recover a fragment from the crime scene. We simulate various levels of degradation from the left, right, and middle of the land impression. We vary the proportion of the land impression that is recovered, between 100% (no degradation) and 25% (significant degradation). For example, a left-fixed 75% scenario implies that the left hand portion of the land was recovered, and the 25% rightmost portion was lost. We will do this by subsetting the signatures. Note that this is a a simplified scenario because the signatures themselves are somewhat dependent on the data that are missing because of the properties of the LOESS smoother.

Figure 8 gives the sensitivity (true positive rate) and specificity (true negative rate) of the random forest predictions for given levels of degradation. It can be seen that the sensitivity drops a bit until 50% of the land is available and then rises again. This occurs because the algorithm begins producing more positive predictions in general, likely as the result of the ccf being arbitrarily higher for known non-matches due to the small signature. On the other hand, the specificity drops dramatically for left, middle, and right fixed degraded lands when less than 50% of the land impression is available for examination. For a more in-depth exploration of the matching probabilities, Figure 7 provides histograms of the matching probability by degradation level and by known match versus known non-match categories. The matching probabilities suffer when compared with the probablities obtained from comparisons between full land impressions in all cases. The jump seems to be most noticeable beginning at about 25% degradation (75% land recovered), and the algorithm struggles beyond 50%.

Figure 9 gives feature expression for known matches, as a function of the proportion of land impression recovered. It is immediately obvious that the variability in feature expression is large when only a small fraction of the land is recovered, such as 25%. For instance, `sum_peaks` and `cms` both drop, while `D` rises. Interestingly, some of the features are better expressed for the middle-fixed case. Overall, feature expression remains relatively consistent as long as we recover 50% or more of the land impression. Feature 10 shows the feature expression for known non-matches by comparison. The non-matches don't exhibit the same pattern of better feature expression for the middle-fixed case, except perhaps for very low degradation levels. However, feature expression rises as a function of the land proportion, which indicates why the random forest begins predicting more positives, raising the sensitivity, but drastically lowering the specificity.

To come full circle, we now attempt to match a particular land which exhibits bad tank rash. Figure 11 provides an image of the surface of this land impression. Due to the tank rash, this particular land impression was originally excluded from consideration (see (Hare,

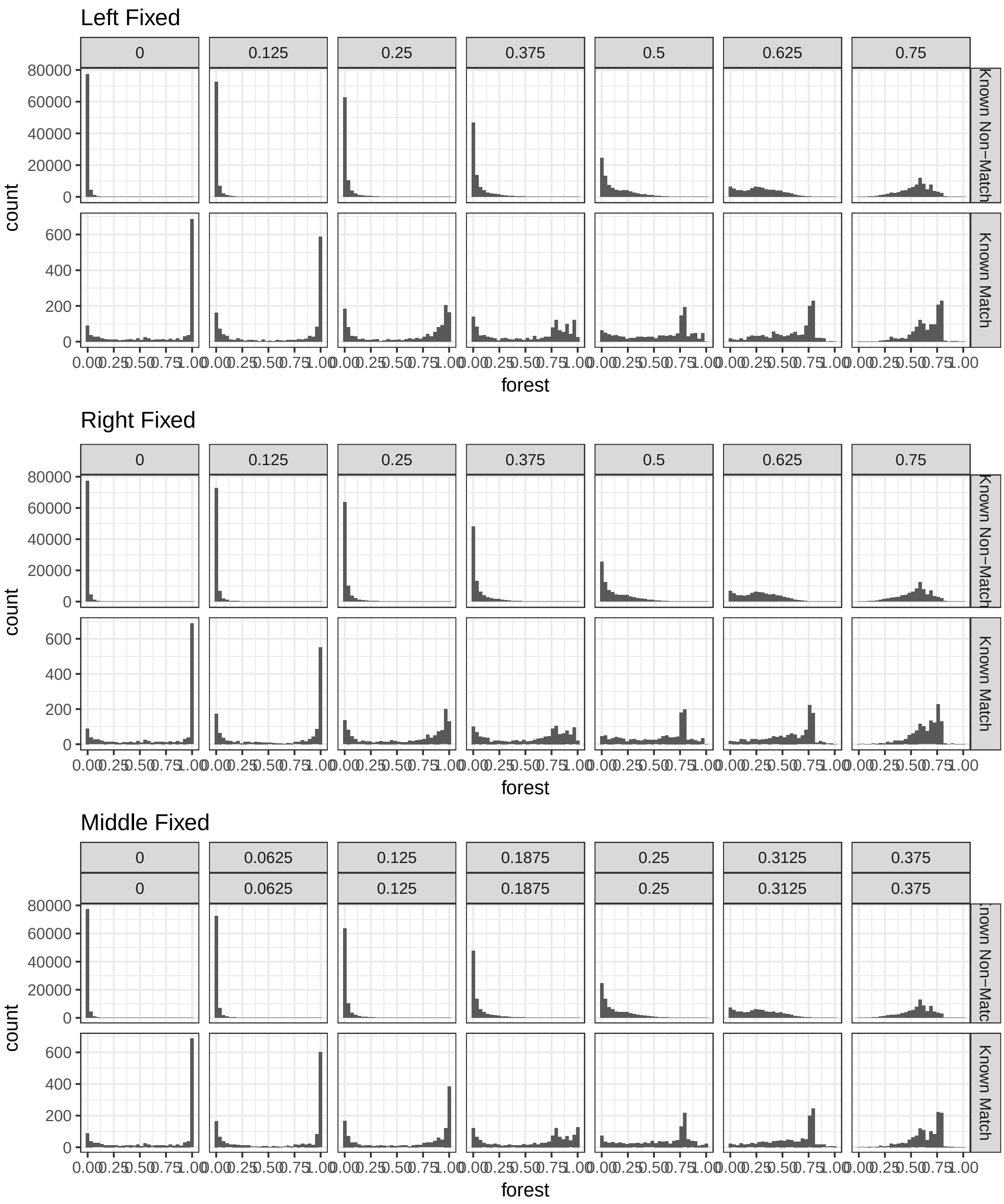

Figure 7: Histograms of matching probability, facetted by the degradation level and known match versus known non-match.

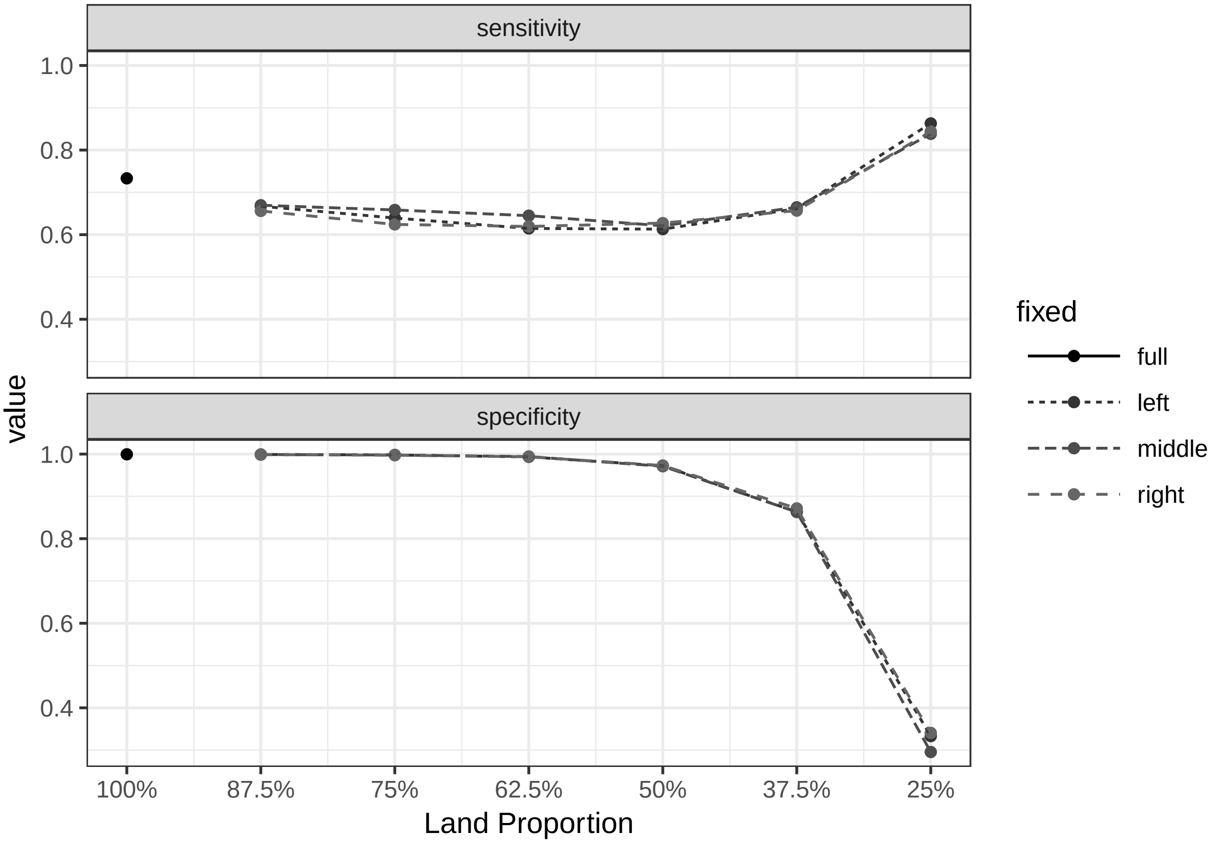

Figure 8: Sensitivity and specificity of the random forest for given levels of degradation. It can be seen that both metrics decline as a function of the land proportion, except for the sensitivity, which rises for very low levels of the land proportion due to an increase in the amount of positive predictions.

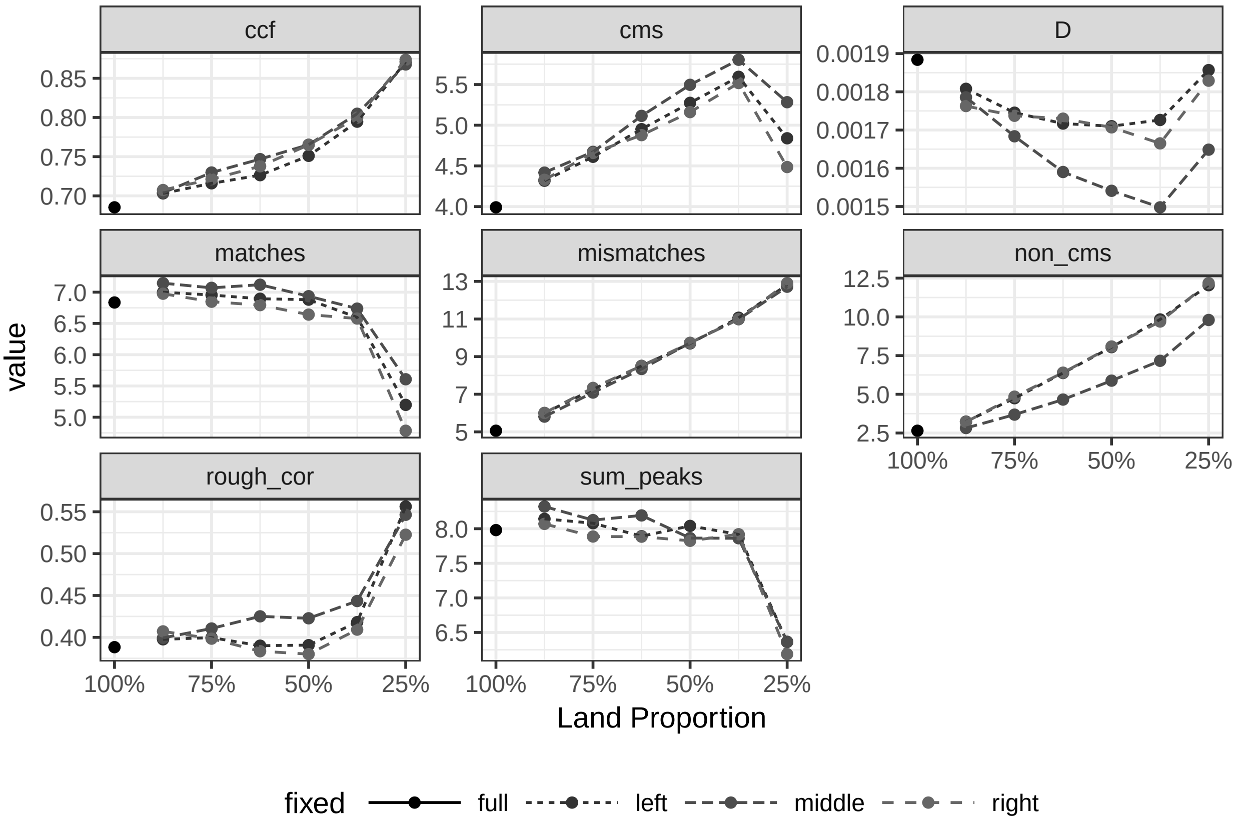

Figure 9: Feature expression for known matches, as a function of land proportion. It can be seen that when we fix the middle portion of the bullet land, the features tend to be better expressed.

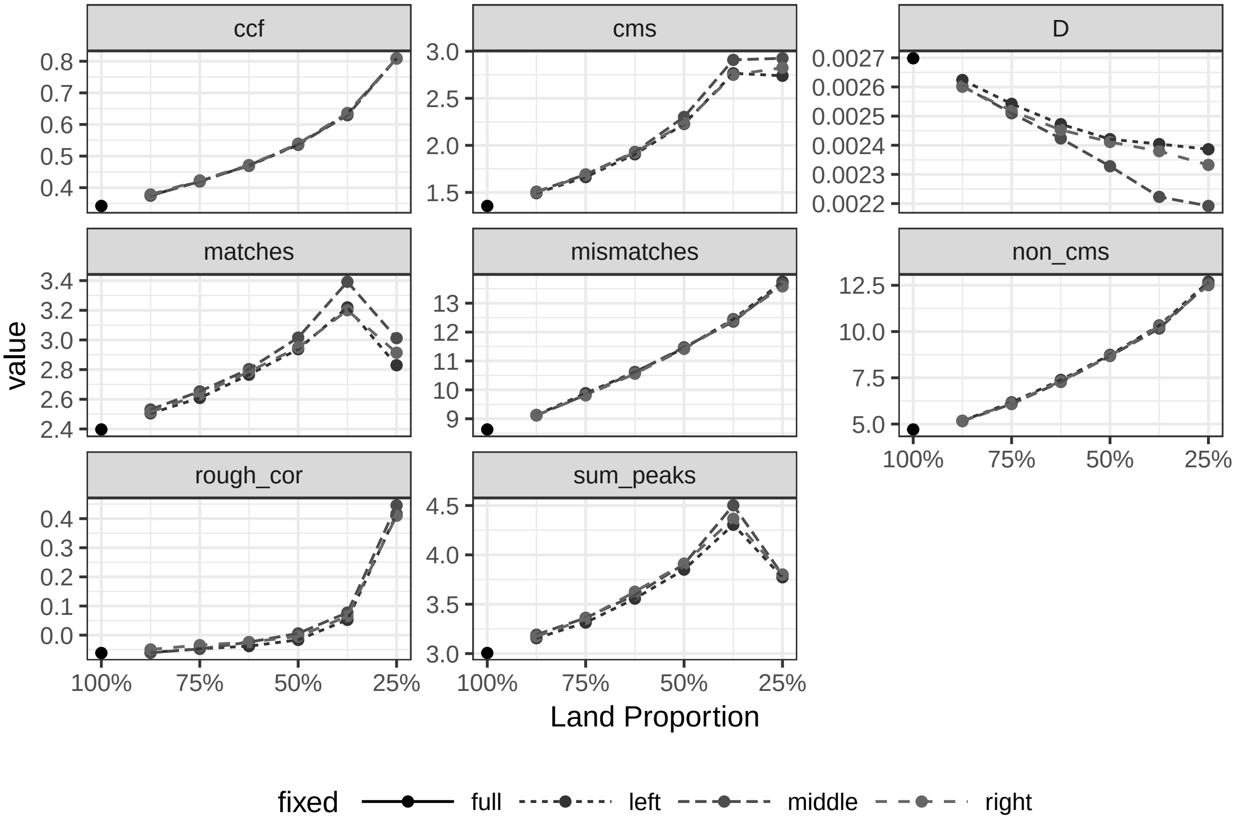

Figure 10: Feature expression for known non-matches, as a function of land proportion. As expected by the fact that the false positive rate increases as the land proportion decreases, so too does the feature expression. However, unlike for the known matches, fixing the middle portion of the land does not seem to lead to more expressed features, except perhaps for very low degradation levels.

Hofmann, and Carriquiry 2016)) based on our subjective assessment of the quality of the land. However, it appears that approximately half of the bullet land remains relatively unaffected. We extract a signature from the unaffected half and attempt to match this signature to its full known match.

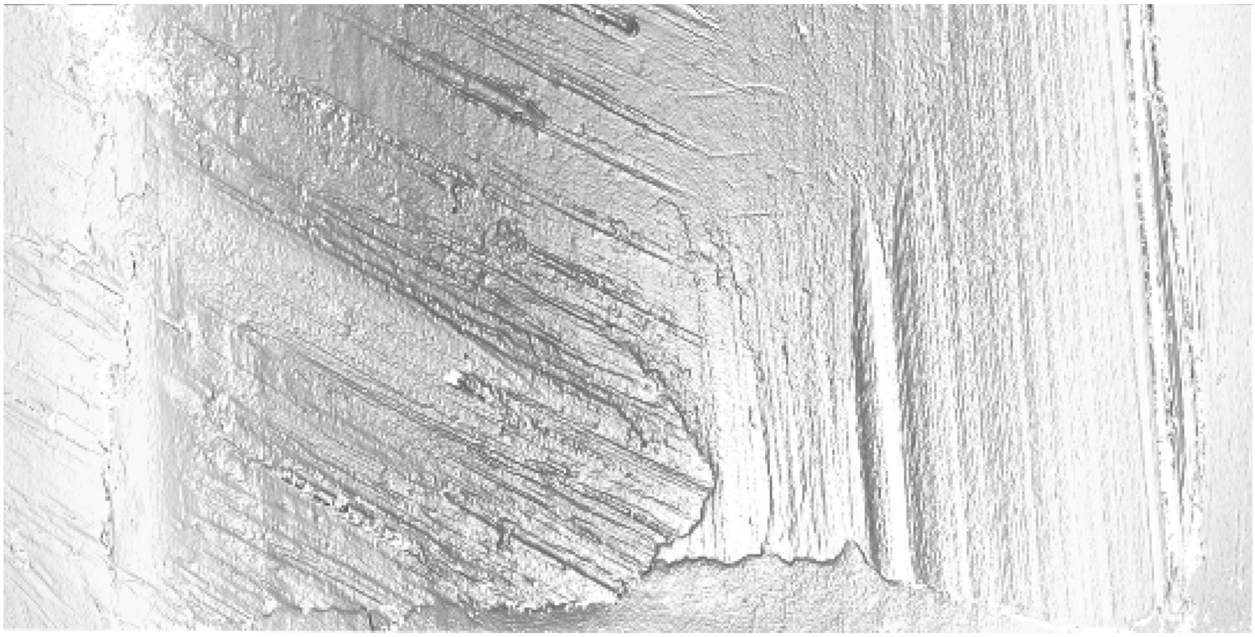

Figure 11: Land 4 of Bullet 2, from Barrel 9 of Hamby Set 252. It can be seen that this particular land exhibits some major tank rash on the left half.

Table 5 shows the values of the features, after extracting only the last 50% of the Hamby Barrel 9 Bullet 2, 4th land (and hence, simulating a right-fixed 50% degraded scenario), compared with a feature comparison between both full lands (and hence, including the tank rash striae). The features are derived in a comparison with its known match, the complete Bullet 1 third land fired from Barrel 9. The features, including the ccf and the matches, are expressed enough to (barely) indicate a match in the case of the degraded bullet. Using the pre-trained random forest, the predicted matching probability is 52%. This is encouraging in that attempting to match the full bullet land, by comparison, yields a matching probability of 0.0067%. This is due to the relatively higher values of the ccf, cms, and matches for the degraded comparison, and suggests that the feature standardization is working as intended.

5 From Lands to Bullets

Another area that deserves more study is the question of generalizing these algorithms to matching entire bullets rather than indvidual lands, as would be done in a criminal justice application. One such approach is to recognize that (at least for the Hamby bullets) there should be six matching pairs of lands for any two bullets that were fired from the same

Feature	Degraded Land	Full Land
ccf	0.6004	0.4442
rough_cor	0.3671	0.1633
D	0.0018	0.0023
overlap	0.9968	0.9968
matches	10.2236	5.6275
mismatches	7.5949	5.0713
cms	9.2013	4.6043
non_cms	6.5823	2.5357
sum_peaks	12.0020	6.3148
matchprob	0.5200	0.0067

Table 5: Features extracted for a comparison of the full Hamby Barrel 9 Bullet 1 Land 3, with a left-fixed 50 percent degraded portion of Hamby Barrel 9 Bullet 2 Land 4. These two lands are known matches, and indeed the random forest does predict a match.

gun barrel. Therefore, for each pair of bullets, we can extract the six highest matching probabilities and average them. If we do so, we obtain a clear separation between the scores that are obtained when matching bullets known to be matches and the scores obtained from known non-matches. This is shown in Figure 12. No known-matches have a score below 50%, while all known non-matches have a score below 10%.

We can improve on this approach by exploiting the rotation of the bullet to compute a score. Under the assumption of land to land independence, we can define the probability that two bullets match (M) as one minus the probability that the two bullets do not match (NM). Exploiting the idea that when two bullets do not match, none of the individual lands match either, we can write the matching probability as the probability that at least one land pair in the matrix matches. Specifically,

$$\begin{aligned} P(M) &= 1 - P(NM) \\ &= 1 - (P(NM1) \times P(NM2) \times ... \times P(NM6)) \\ &= 1 - ((1 - P(M1)) \times (1 - P(M2)) \times ... \times (1 - P(M6))) \end{aligned}$$

where M is the event that two bullets match, NM is the event that two bullets do not match, $M1$, $M2$, ..., $M6$ are the probabilities of land one, land two, ..., land six matching, and $NM1$, $NM2$, ..., $NM6$ are the probabilities that land one, land two, ..., land six do not match. However, to compute this probability, we need to know the alignment of the two sets of lands. Fortunately, the consistent rotation of the bullet permits this. For instance, if we knew that land 1 of bullet 1 matches land 4 of bullet 2, then we immediately know that land 2 of bullet 1 matches to land 5 of bullet 2, land 3 of bullet 1 matches to land 6 of bullet 2, etc. Hence, we can take look across six diagonals of the $6 \otimes 6$ matrix containing match probabilities. Table 6 gives an example of the matrix of matching probabilities between two sets of six lands from bullets that are known matches. The matching diagonal is clear based on the high probabilities (cell $(1, 3)$, cell $(2, 4)$, cell $(3, 5)$, etc.) although it can be seen that

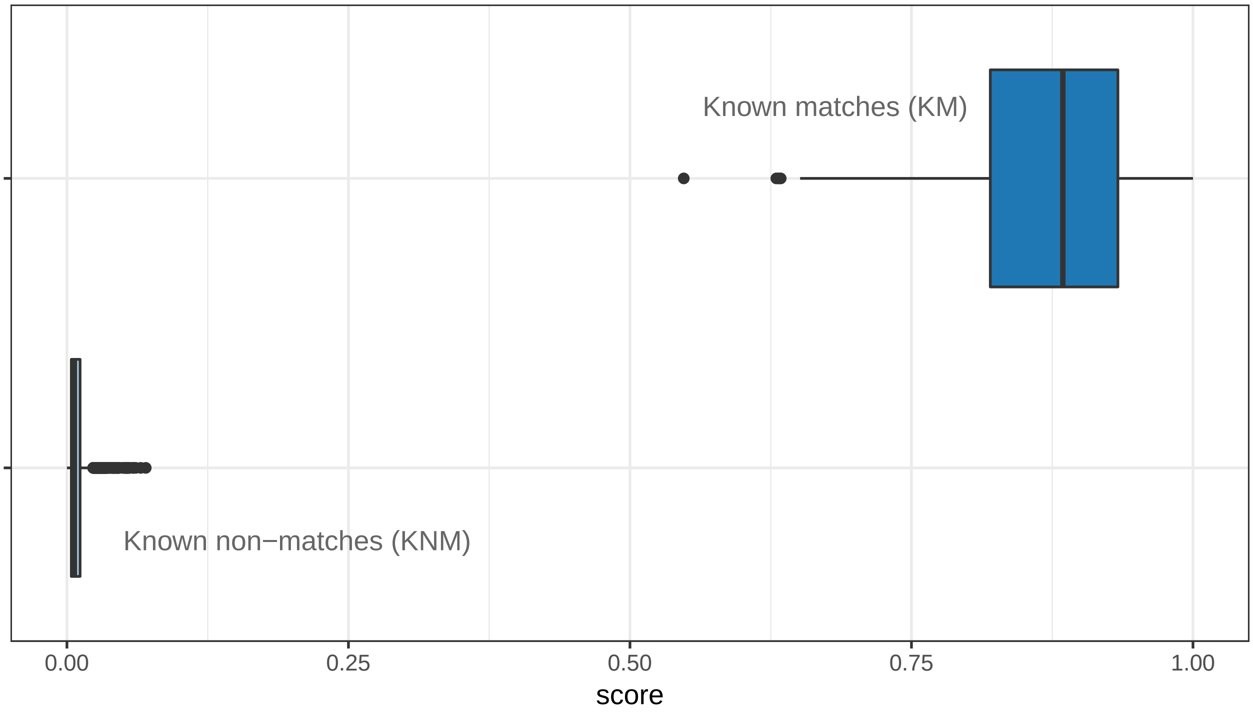

Figure 12: Score distributions for the naive approach to bullet matching, for known matches and known non-matches.

one of the six comparisons has a relatively lower matching probability. This procedure is based on the Sequence Average Maximum (SAM) by Sensofar (2017) in their bullet matching software application `SensoMatch`. A similar approach using the cross correlation maximum was first proposed by W. Chu et al. (2010). Compared to that approach, ours uses random forest based probabilities compared to correlation values allowing for elements of probability theory to help determine the resulting bullet match probability.

profile1_id	45604	46104	46601	47069	47600	48069
42594	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
43063	0.0000	0.0000	0.0000	1.0000	0.0067	0.0000
43581	0.0000	0.0000	0.0000	0.0000	0.8433	0.0000
44211	0.0000	0.0000	0.0133	0.0000	0.0000	0.6700
44568	1.0000	0.0000	0.0033	0.0000	0.0000	0.0000
45070	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000

Table 6: Matrix of matching probabilities between two sets of six lands from bullets that are known matches.

We now describe four methods for deriving a score from this matrix. The results derived from these methods are shown in Figure 13. In **Method 1**, we derive a score by computing the bullet matching probability on each set of six matrix diagonals using the previously defined formula, under the assumption of land to land indepndence. Finally, we take the maximum score obtained out of the six results as the final matching score for a bullet pair. After doing so, we can plot the scores for known matches and known non-matches separately. It can be seen that the known matches all have scores of around 100%, while no non-match achieves a score of above 30%, and hence this procedure provides perfect discrimination between all pairs of bullets between and within the two Hamby datasets.

On the other hand, **Method 2** is obtained by flipping this procedure around by assuming that a match occurs if and only if all six lands match. As it turns out, this does not discriminate quite as well. Every known bullet non-match achieves a score of about zero, but so do about 15 known bullet matches. This method performs poorly because our matching algorithm exhibits a larger false negative rate than the rate of false positives. Multiplying the probabilities together compounds the issue of false negatives and leads to some misidentification of matching bullets.

Method 3 is a hybrid of these two approaches, where we average the probabilities along the diagonal rather than multiplying those probabilities. Now, we once again differentiate the two groups well with no known non-match achieving a score above 10%, and no known match with a score below 40%.

One more approach to generate bullet matching scores, which we call **Method 4**, would exploit the SAM procedure on individual features. For each diagonal in the $6 \otimes 6$ matrix, we can compute an average value for each feature in our model. This yields six sets of feature values for all six diagonals. We can then feed all six sets of features into the random forest in order to obtain a matching probability for each, taking the highest resulting probability to locate the diagonal and thus identify land to land alignment. It can be seen that while this procedure does discriminate well, it yields some false negatives (matching bullets that our forest identifies as a non-match).

6 Conclusion

In this paper, we have introduced a set of robust features that can be used to train bullet matching models. We have used these features to train a random forest and assess its out-of-sample accuracy. In doing so, we noted strong evidence of operator effects that resulted in differences in the quality of the microscope scans. These effects were noted despite the experience of the individuals conducting the scans, which implies that such effects could quite likely be more pronounced when scans are done by those more inexperienced, or with fewer standard operating procedures in place.

While these effects were clearly identified, the best approach to account for them in practice is less clear. In the ideal case, bullets fired from a particular gun barrel should yield surface scans that are of identical quality and properties, regardless of the operator performing the scan. To achieve this, rigorous standards may need to be put in place with regards to the alignment of the bullet under the objective, and the procedure used to scan the bullet surface. To appropriately design a set of best practices requires more research. For instance, because of the significant difference between the placement of the ideal cross section across the two

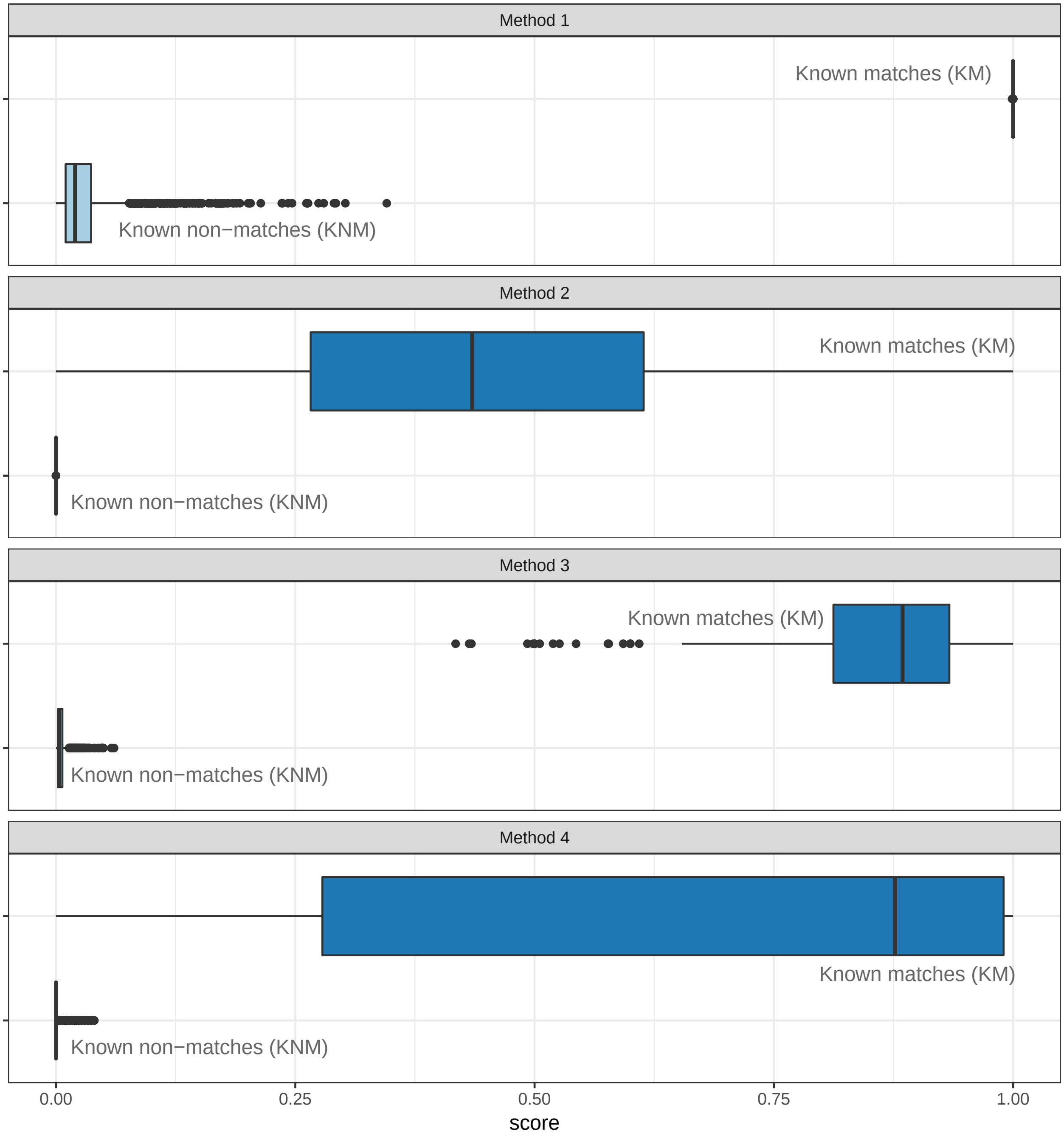

Figure 13: Distribution of matching scores using four methods. Method 1 assumes a match if at least one pair of lands match. Method 2 assumes a match if all pairs of lands match. Method 3 averages the probabilities instead of multiplying them. Finally, Method 4 uses a SAM procedure on the feature values for known matches compared to known non-matches. While these methods have various levels of discriminatory power, and rely on slightly different assumptions, they do show clear and significant separation between matching bullets and non-matching bullets in general.

studies, a best practice may specify the margin from the edge of the objective at which the bullet can be placed.

We began exploring the robustness of the matching algorithm proposed in Hare, Hofmann, and Carriquiry (2016) to land degradation. As suspected, the algorithm performance declines as a function of the rate of degradation. However, there is a relatively clear threshold around about 50%; if 50% of the land or more is recovered, the algorithm still performs reasonably well. When the proportion of the land that is recovered is below 50%, the accuracy with which we can compare land impressions is low.

Finally, one pleasing conclusion from these results is the fact that generalizing them to full bullet comparisons rather than the land to land level appears to work quite well. Depending on the assumptions made, the out of sample accuracy of bullet to bullet comparisons can range from nearly perfect to perfect. This result is encouraging in that real world use of these algorithms would be done on the bullet level, assuming enough of the bullet was recovered to make these procedures possible.

As we have stated before, the lack of 3D images of bullets available in the public domain limits the extent to which these algorithms can be tested and validated. The degraded land simulation itself may be too simplistic and not faithfully represent realistic scenarios. However, as more data are collected, we can continue to update, train, and test the matching algorithm in order to improve its performance in real datasets.

7 Acknowledgment

We wish to acknowledge the anonymous reviewers, whose quick yet thorough and immensely helpful reviews helped us improve this manuscript. We greatly appreciate the extra effort put forth for us.

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.

References

Biasotti, Alfred A. 1959. "A Statistical Study of the Individual Characteristics of Fired Bullets." *Journal of Forensic Sciences* 4 (1): 34–50.

Chu, W., J. Song, T. Vorburger, R. Thompson, and R. Silver. 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 (3): 649.

Chu, W., J. Song, T. Vorburger, J. Yen, S. Ballou, and B. Bachrach. 2010. "Pilot study of automated bullet signature identification based on topography measurements and correlations." *J. Forensic Sci.* 55 (2): 341–47.

Chu, Wei, Robert M Thompson, John Song, and Theodore V Vorburger. 2013. "Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria." *Forensic Science International* 231 (1–3): 137–41.

Clarkson, James A, and C Raymond Adams. 1933. "On Definitions of Bounded Variation for Functions of Two Variables." *Transactions of the American Mathematical Society* 35 (4). JSTOR: 824–54.

Giannelli, Paul C. 2011. "Ballistics Evidence Under Fire." *Criminal Justice* 25 (4): 50–51.

Hamby, James E., David J. Brundage, and James W. Thorpe. 2009. "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 (2): 99–110.

Hare, E., H. Hofmann, and A. Carriquiry. 2016. "Automatic Matching of Bullet Lands." *ArXiv E-Prints*, January.

Jed Wing, Max Kuhn. Contributions from, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2016. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." *R News* 2 (3): 18–22. http://CRAN.R-project.org/doc/Rnews/.

National Research Council. 2004. *Forensic Analysis: Weighing Bullet Lead Evidence*. National Academies Press.

———. 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: The National Academies Press. doi:10.17226/12589.

Petraco, Nicholas, and Helen Chan. 2012. *Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons*. Mannheim, Germany: Bibliographisches Institut AG.

President's Council of Advisors on Science and Technology. 2016. "Report on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods."

https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Riva, Fabiano, and Christophe Champod. 2014. "Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases." *Journal of Forensic Sciences* 59 (3): 637–47. doi:10.1111/1556-4029.12382.

Sensofar. 2017. *SensoMATCH Bullet Comparison Software*.

Vorburger, T.V., J.-F. Song, W. Chu, L. Ma, S.H. Bui, A. Zheng, and T.B. Renegar. 2011. "Applications of Cross-Correlation Functions." *Wear* 271 (3–4): 529–33. doi:http://dx.doi.org/10.1016/j.wear.2010.03.030.