

Adaption of the Chumbley Score to matching of bullet striation marks

Ganesh Krishnan *

Department of Statistics, Iowa State University
and

Heike Hofmann

Department of Statistics and CSAFE, Iowa State University

May 22, 2018

Abstract

Keywords: forensic science, toolmark, cross-correlation, Mann-Whitney U statistic, ,

*The authors gratefully acknowledge ...

Contents

1 Introduction and Background

1.1 Motivation

Same-source analyses are a major part of a Forensic Toolmark Examiner’s job. In current practice, examiners make these comparisons by visual inspection under a comparison microscope and come to one of the following four conclusions: identification, inconclusive, elimination or unsuitable for examination (?). These conclusions are made on the basis of “unique surface contours” of the two toolmarks being in “sufficient agreement” (?). AFTE describes the term “sufficient agreement” as the possibility of another tool producing the markings under comparison, as practically impossible (?). Potential subject bias in the assessment as well as the lack of specified error rates are the main points of criticisms first raised by the National Research Council in 2009 (?) and later emphasized further by the President’s Council of Advisors on Science and Technology (?).

Technological advances, such as profilometers and confocal microscopy allow to measure 3D surfaces in a high-resolution digitized form. This technology has become more accessible over the last decade, and has made its way into topological images of ballistics evidence, such as bullet lands and breech faces (????). Digitized images of 3D surfaces of form the data basis of statistical analysis of toolmarks. A statistical approach based on data removes both subjectivity from the assessment and allows a quantification of error rates for both false positive and false negative identifications.

Methods for matching marks for a variety of tools have been studied in the literature (see ?? for an overview): ? and ? have been analyzing screwdriver marks digitized using a profilometer; ? have investigated 3D marks from screwdriver, tongue and groove pliers captured using a confocal microscope; ? have been investigated digitized marks from slip-joint pliers generated by a surface profilometer.

Analysis of these digitized markings require the use of statistical methods which can quantify the scientific mechanism of comparing markings and serve as basis for an error rate calculation. ? define a relative distance metric and use it as similarity measure between two toolmarks. ? extract many small segments in the markings of two toolmarks and compare similiarity using a maximum pearson correlation coefficient. The Chumbley scor-

ing method, first introduced by ?, uses a similar but more extensive framework based on a Mann-Whitney U test of the resulting correlation coefficients. This approach is non-deterministic, because segments are chosen randomly. (?) make the score deterministic for each pair of toolmarks by choosing segments for comparison systematically. This approach also ensures independence between segments of striae. In this paper, we are investigating the applicability of the Chumbley scoring method by ? to assess striation marks on bullet lands for same-source identification.

Research paper	Method	Data Source	False Positives	False Negatives
	? Maximized Correlation	Screwdrivers	-	-
(Same-Surface Same-Angle)	? Randomized Chumbley Score	Screwdrivers	2.3%	8.9%
	? Randomized Chumbley Score	Slip-joint	-	-
(Same-Surface Same-Angle)	? Deterministic Chumbley Score	Screwdrivers	0%	6%
(Different Surfaces-same angle)	? Similarity Measure Relative Distance Metric	Screwdrivers	5.9%	9.4%
	(Same Surfaces-same angle)		0.22%	0%

Table 1: Error Rates in same-source toolmark analysis reported in the literature. All reported papers use some variation of the Chumbley-score method.

Striation marks on bullets are made by impurities in the barrel. As the bullet travels through the barrel, these imperfections leave “scratches” on the bullet surface (see top of ??). Typically, only striation marks in the land engraved areas (LEAs) are considered (?). Bullet lands are depressed areas between the grooves made by the rifling action of the barrel. Compared to toolmarks made by screwdrivers striation marks on bullets are typically much smaller, both in length and in width. Bullets also have a curved cross-sectional topography. In same-source comparisons this curvature is usually removed using some form of Gaussian filter (?) or non-parametric smoothing (?). ? use an automatic method for counting consecutive matching striae (CMS). The authors report an error rate of 52% for known same-source land comparisons to be (incorrectly) identified as different-source (false neg-

ative) and zero false positives for known different-source lands. [?] and [?] discuss CCF (cross-correlation function) and its discriminating power and applicability for same-source analyses of bullets, but do not provide any error rates in their discussion. [?] use multiple features, such as CCF, CMS, D (distance measure), etc. in a random forest based method and compare every land against every other land of digitised versions of Hamby 252 and Hamby 44 ([?]) published on the NIST Ballistics Database ([?]). The authors report an out-of-bag overall error rate of 0.46%, comprised of an error rate of 30.05% of same-source pairs that were not identified and an error rate of 0.026% of different-source pairs that were incorrectly identified as same-source.

The Chumbley score provides us with another approach in the same-source assessment of bullet striation marks. [?] compare two toolmarks for same-source. The data for this study was obtained from 50 sequentially manufactured screwdriver tips. [?] report error rates for markings made by the tips at different angles. For markings made at 30 degree the authors report an average false negative error rate of 0.089 and an average false positive error rate of 0.023. For marks made under angles of 60 and 85 degrees, respectively, the false negatives error rate is 0.09 while the rate of false positives decreases to 0.01. The paper by [?] is based on the same data but the authors focus on markings made under the same angle. The error rates associated with the deterministic version of the score are reported as 0.06 for false negatives and 0 for false positives.

1.2 Scans for land engraved areas

Comparisons of striae from bullets are usually based on comparisons of striae in land engraved areas, which are extracted in form of cross sections, called *profiles* ([?]). From profiles bullet *signatures* ([?]) are extracted as residuals of a loess fit or Gaussian filter. This effectively removes topographic structure from the data in the attempt to increase the signal to noise ratio. [?] shows how the signature from a bullet land (bottom) lines up with the image of the land (top) from which it was extracted. We can see in the figure how the depth and relative position of the striation markings seen in the image are interpreted as peaks and valleys in the signature.

There are two sources of scans for sets from the Hamby study available to us: scans of



Figure 1: Image of a bullet land from a confocal light microscope at 20 fold magnification (top) and a chart of the corresponding signature of the same land (bottom). The dotted lines connect some peaks visible in both visualizations.

Hamby 44 and Hamby 252 are available from the NIST database (?). The physical Hamby 44 set has also been made available to us and has been scanned locally for CSAFE at the Roy J. Carver High Resolution Microscopy Facility using a Sensofar confocal light microscope. Scans in the NIST database are made with a NanoFocus at 20x magnification. The resolutions of the two instruments are different: the NIST scans are taken at a resolution of $1.5625 \mu m$ per pixel, while the CSAFE scans are available at a resolution of $0.645 \mu m$ per pixel. The length of an average bullet land from Hamby (9 mm Ruger P85) is about 2 millimeter, resulting in signatures of about 1200 pixels for NIST scans, and about 3000 pixels for CSAFE scans.

In comparison, scans from the profilometer used by ?? were taken at a resolution of about $0.73 \mu m$ per pixel. The screw driver toolmarks are about 7 mm in length (?), for a total of over 9000 pixels for the width of these scans.

This severe limitation in the amount of available data poses the main challenge in adapting the Chumbley score to matching bullet lands, because of the resulting loss in power.

1.3 The Chumbley Score Test

A digitized toolmark forms a spatial process $z(t)$ with location indexed by t . t here denotes equally spaced pixel locations for the striation marks under consideration. For a toolmark consisting of t pixels, $t = 1, \dots, T$. Let further $z^s(t)$ denote a vector of markings of length s starting in location t .

The Chumbley score algorithm takes input in form of two digitized toolmarks:

Let $x(t_1)$, $t_1 = 1, 2, \dots, T_1$ and $y(t_2)$, $t_2 = 1, 2, \dots, T_2$ be two digitized toolmarks (where T_1 and T_2 , the lengths of the two marks, are not necessarily equal). The toolmarks under consideration are potentially from two different-sources or the same-source.

In a pre-processing step the two markings are smoothed using a lowess (?) with coarseness parameter c . Originally, this smoothing is intended to remove drift and (sub)class characteristics from individual markings, however, in the setting of matching bullet striae, we can also make use of this mechanism to separate bullet curvature in profiles from signatures before matching signatures. ?? shows an example of a bullet land profile (left) and the corresponding signature (right).

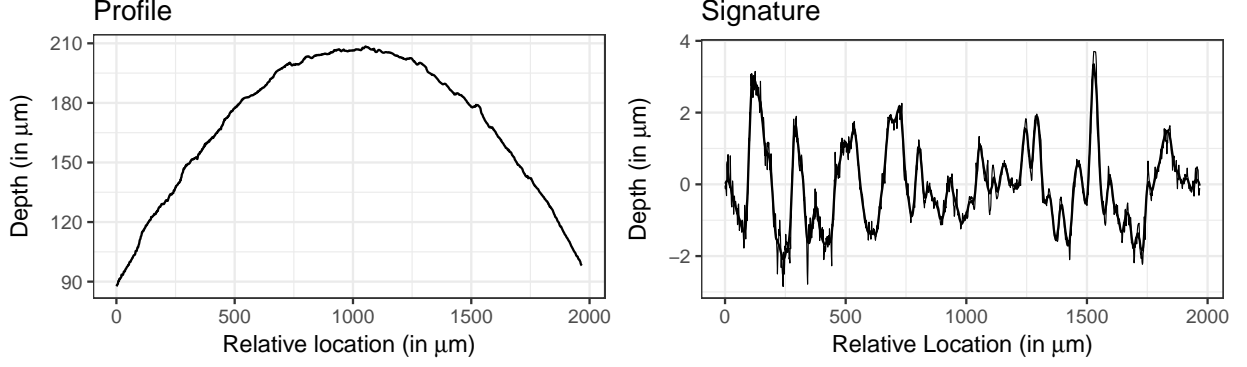


Figure 2: Bullet land profile (left) and the corresponding signature (right) for one of the lands of Hamby-44.

After removing sub-class structure, the Chumbley score is calculated in two steps: an optimization step and a validation step. In the optimization step, the two markings are aligned horizontally such that within a pre-defined window of length w_o the correlation between $x(t_1)$ and $y(t_2)$ is maximized:

$$(t_1^o, t_2^o) = \arg \max_{1 \leq t_1 \leq T_1 - w_o, 1 \leq t_2 \leq T_2 - w_o} \text{cor}(x^{w_o}(t_1), y^{w_o}(t_2))$$

This results in an optimal vertical (in-phase) shift of $t_1^o - t_2^o$ for aligning the two markings. We will denote the relative optimal locations as t_1^* and t_2^* , where $t_i^* = t_i^o / (T_i - w_o)$ for $i = 1, 2$, such that $t_1^*, t_2^* \in [0, 1]$. Once (sub-)class characteristics are removed, the relative optimal locations should be distributed according to a uniform distribution in $[0, 1]$.

In the validation step, two sets of windows of size w_v are chosen from both markings, see fig:win-comparison. In the first set, pairs of windows are extracted from the two markings using the optimal vertical shift as determined in the first step, whereas for the second set the windows are extracted using a different (out-of-phase) shift.

More precisely, let us define starting points $s_i^{(k)}$ for each signature $k = 1, 2$ as

$$s_i^{(k)} = \begin{cases} t_k^* + iw_v & \text{for } i < 0 \\ t_k^* + w_o + iw_v & \text{for } i \geq 0, \end{cases} \quad (1)$$

for integer values of i with $0 < s_i^{(k)} \leq T_k - w_v$.

Same-shift pairs of length w_v are defined in ? as all pairs $(s_i^{(1)}, s_i^{(2)})$ with integer values i for which both $s_i^{(1)}$ and $s_i^{(2)}$ are defined. Similarly, different-shift pairs are defined as

$(s_i^{(1)}, s_{-i-1}^{(2)})$ for all i where both $s_i^{(1)}$ and $s_{-i-1}^{(2)}$ are defined (see ??).

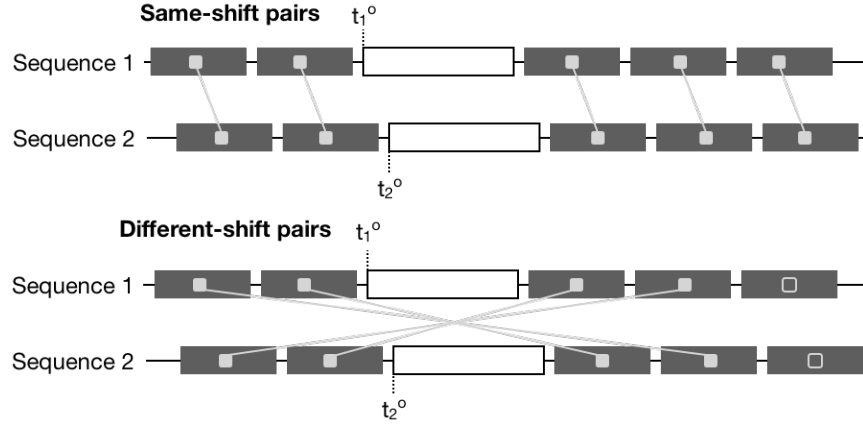


Figure 3: Sketch of same-shift pairings (top) and different-shift pairings (bottom). Filled in rectangles show pairings resulting in correlations, unfilled rectangles are segments without a match.

For both same- and different-shift pairs correlations between the markings are calculated. The intuition here is that for two markings from the same-source the correlation for the in-phase sample should be high, while the correlations of the out-of-phase sample provide a measure for the base-level correlation for non-matching marks of a given length w_v . The Chumbley score is then computed as a Mann Whitney U statistic to compare between in-phase sample and out-of-phase sample. In the original method proposed in ? both in-phase and out-of-phase sample are extracted randomly, whereas ? proposed the above specified deterministic rules for both samples to make the resulting score deterministic while simultaneously avoiding overlaps within selected marks to ensure independence.

1.4 A problem with failed tests

Looking closer at ??, we see that by definition, some number of tests will fail to produce a result. The problem of failed tests is first mentioned in ?. Unfortunately, the authors do not provide any percentage of how many tests failed for their data. The algorithm fails to produce for two reasons: either the number of eligible same-shift pairs is zero, or the number of different-shift pairs is zero.

The number of same-shift pairs will be zero, if the optimal locations t_1^o and t_2^o are so far

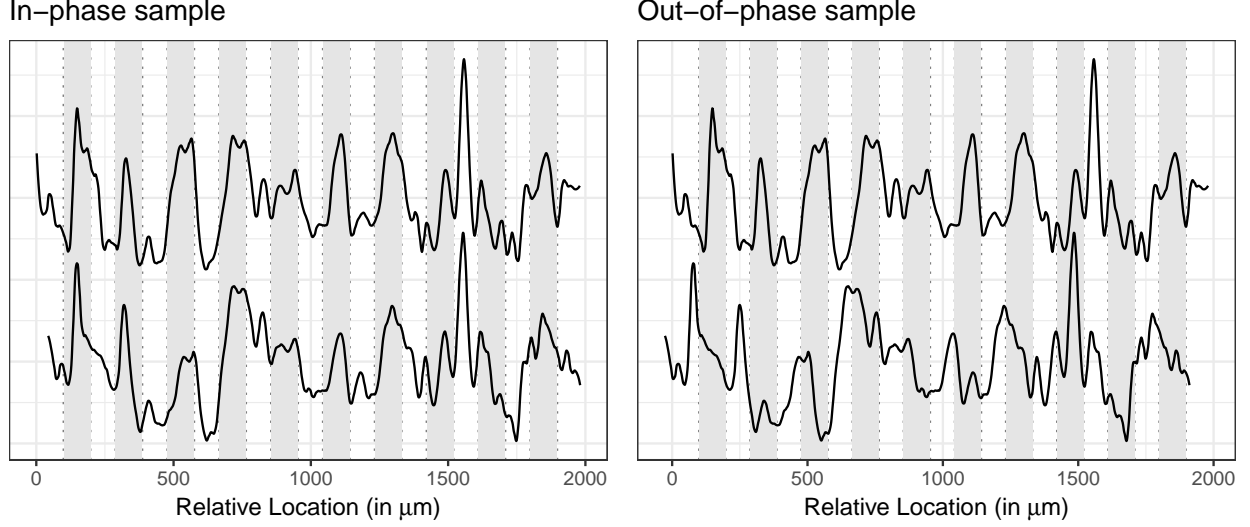


Figure 4: Two same-source markings. For convenience, the markings are moved into phase on the left and out-of phase on the right. In-phase (left) and out-of-phase (right) samples are shown by the light grey background. The Chumbley-score is based on a Mann-Whitney U test of the correlations derived from these two sets of samples.

apart, that no segments of size w_v are left on the same sides of the optimal locations, i.e. $t_1^o < w_v$ and $t_2^o > T_2 - w_o - w_v$ or $t_1^o < T_1 - w_o - w_v$ and $t_2^o < w_v$, i.e. we have a failure rate of

$$P(t_1^o < w_v \cap t_2^o > T_2 - w_o - w_v) + P(t_1^o < T_1 - w_o - w_v \cap t_2^o < w_v).$$

While we can assume that once (sub-)class characteristics are removed, optimal locations t_i^o are uniformly distributed across the length of the profile, we cannot assume that t_1^o and t_2^o are independent of each other. In particular, for same-source profiles, we would expect a strong dependency between these locations, in which case a large difference between locations is unlikely. However, for different-source matches, we can assume that locations are independent. In that case, we expect a test to fail with a probability of $\frac{2w_v^2}{(T_1 - w_o)(T_2 - w_o)}$. For an average length of T_i of 1200 pixels, $w_o = 120$ pixels and $w_v = 30$ pixels this probability is about 0.0015.

The number of possible different-shift pairs also depends on the location of the optimal locations t_1^o and t_2^o . Whenever the optimal locations are close to the boundaries, the number of possible pairings decreases and reaches zero, if $t_i^{(o)} < w_v$ or $t_i^{(o)} > T_i - w_o - w_v$. Assuming a correlation between optimal locations t_1^o and t_2^o of close to one for same-source profiles,

this results in an expected rate of failure of $2w_v/(T_i - w_o)$, or about 5.6% for an average length of T_i of 1200 pixels, $w_o = 120$ pixels and $w_v = 30$ pixels. Assuming independence in the optimal locations for different-source profiles the expected probability for a failed test is, again, $\frac{2w_v^2}{(T_1 - w_o)(T_2 - w_o)}$.

1.5 A modified approach

While failures due to missing correlations from same-shift pairs are unavoidable by definition of the Chumbley score, failures due to missing correlations from different-shift pairs can be prevented by using a different strategy in assigning pairs.

Using the same notation as in ??, we define same-shift pairs identical to ? as pairs $(s_i^{(1)}, s_i^{(2)})$ for all i where the boundary conditions of both sequences are met simultaneously. Let us assume that this results in I pairs. Define $s_{(j)}^{(k)}$ to be the j th starting location in sequence $k = 1, 2$, i.e. $s_{(1)}^{(k)} < s_{(2)}^{(k)} < \dots < s_{(I)}^{(k)}$.

We then define the pairs for different-shifts as

$$\left(s_{(j)}^{(1)}, s_{(I-j+1)}^{(2)}\right) \text{ for } j = \begin{cases} 1, \dots, I & \text{for even } I \\ 1, \dots, (I-1)/2, (I-1)/2 + 2, \dots, I & \text{for odd } I \end{cases}, \quad (2)$$

i.e. for an odd number of same-shift correlations, we skip the middle pair for the different-shift correlations (see ??). This pairing ensures that the number of different-shift pairings is the same or at most one less than the number of same-shift pairings in all tests. In the remainder of the paper, we will refer to the algorithm defined by ? as **(CS1)** and the suggested modified algorithm as **(CS2)** and compare their performance on the available scans of the Hamby study.

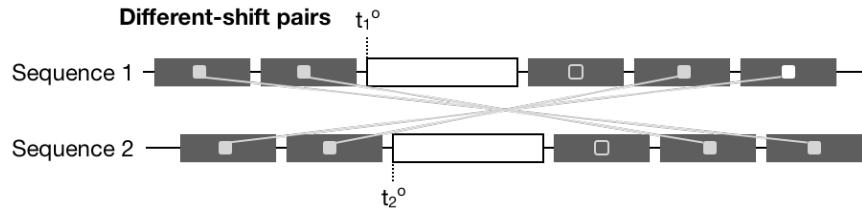


Figure 5: Sketch of adjusted different-shift pairings. At most one of the same-shift pairings can not be matched with a different-shift pair under algorithm (CS2).

2 Testing setup

2.1 The Data

Lands for all Hamby-44 and Hamby-252 scans are made available through the NIST ballistics database (?) and are considered, here. Both of these sets of scans are part of the larger Hamby study (?). Each set consists of twenty known bullets (two each from ten consecutively rifled Ruger P85 barrels) and fifteen questioned bullets (each matching one of the ten barrels). Ground truth for both of these Hamby sets is known and was used to assess correctness of the tests results.

Profiles for each bullet land were extracted from scans close to the heel of the bullet while avoiding break-off as described in ?.

2.2 Setup

Both algorithms (CS1) and (CS2) are implemented in R (?). (CS1) is available from package `toolmaRk` (?), (CS2) is available from a modified version of the `toolmaRk` package available from GitHub (<https://github.com/heike/toolmaRk>). We applied both methods to all pairwise land-to-land comparisons of the Hamby scans provided by NIST for a total of 85,491 land-to-land comparisons.

3 Results

3.1 Failed Tests

As described above, the Chumbley-score is based on three parameters: coarseness c and the sizes of the optimization window w_o and validation window w_v . In a first run of results, we applied default settings for the parameters, as suggested in ?: $w_o = 120$ pixels or about $190 \mu m$ (ten percent of the average length of profiles) and coarseness $c = 0.25$, and varied the size of the validation window w_v in steps of 10 from 10 pixels to 60 pixels. Based on a significance level α of 0.05 for the test, this results in a correct identification of same-source and different-source toolmarks of 93.5% to 94.1%, corresponding to a rate of false negatives

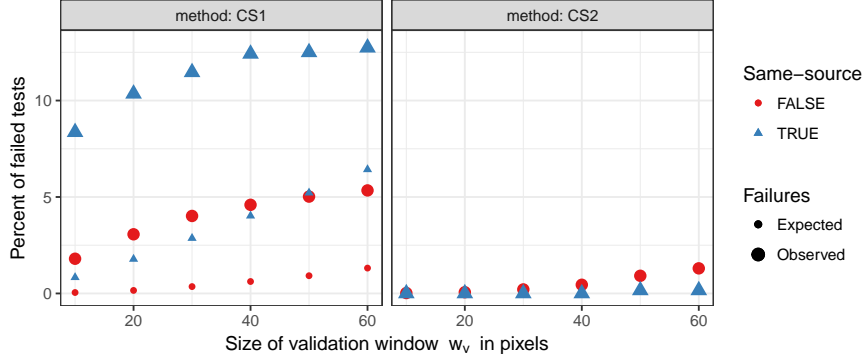


Figure 6: Percent of failed land-to-land comparisons using an optimization window $w_o = 120$ and a coarseness of $c = 0.25$. With an increase in the size of the validation window a higher percentage of tests fails under both methods (CS1) and (CS2), but the percentage of failed tests is much smaller under (CS2). Observed failure rates are higher than expected rates.

between 0.28 and 0.36 and a rate of false positives between 0.05 and 0.06. However, the most prominent result we encountered, are the high number of failed tests, i.e. the number of instances, in which CS1 did not return any result. ?? shows the percentage of failed tests among the 85,491 land-to-land comparisons of the NIST data for different values of the validation window size w_v . For same-source lands up to 12.5 percent of the tests fail using CS1. The highest percentage of failed tests under CS2 is 1.3% for different-source tests using a validation window size w_v of 60 pixels. Rates of expected failures are based on simulation runs using covariances between locations of same-source profiles of 0.854, and 0.120 for locations from different-source profiles, matching observed covariances for the Hamby scans. Observed failure rates are higher than expected. This might be due to remaining sub-class structure at a coarseness of 0.25 resulting in a distribution of optimal locations different from the assumed uniform.

3.2 Coarseness

The purpose of the coarseness parameter is to remove (sub-)class characteristics from profiles before comparisons for matching. ? suggest a coarseness parameter of 0.25 in the setting of toolmark comparisons. For bullet lands, coarseness might need to be adjusted

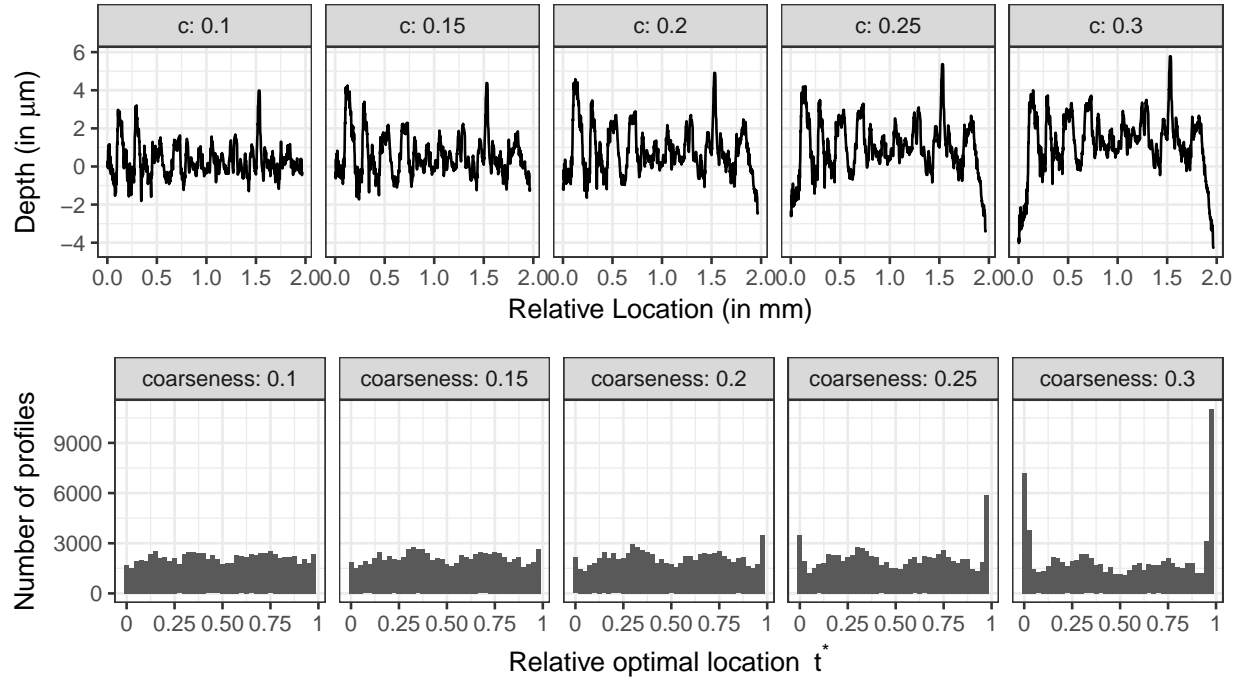


Figure 7: Overview of the effect of different coarseness parameters c on the profile shown in Figure ?? (top). The bottom row shows histograms of the (relative) optimal locations t^o identified in the optimization step for different values of the coarseness parameter c .

because of the strong effect bullet curvature has on profiles.

?? gives an overview of the effect of different coarseness parameters: from left to right, coarseness levels c are varied in steps of 0.05 from 0.1 to 0.3. The top row shows resulting signatures after smoothing the profile shown in Figure ?? with different levels of coarseness. The histograms in the bottom row show the relative optimal location t^* . Optimal locations are distributed uniformly once (sub-)class characteristics are removed. However, for coarseness values of $c > 0.20$ we see quite distinct boundary effects: optimal locations t^* are found at the very extreme ends of a profile more often than one would expect based on a uniform distribution.

The key effect of the optimal locations and thereby the coarseness is seen in the number of failed tests. Irrespective of whether CS1 or CS2 is being used, if the relative optimal locations are at the boundaries we will see an increase in the number of failed tests. A balance is therefore needed in the selection of the coarseness parameter which reduces the boundary effect but also does not remove important individual characteristics

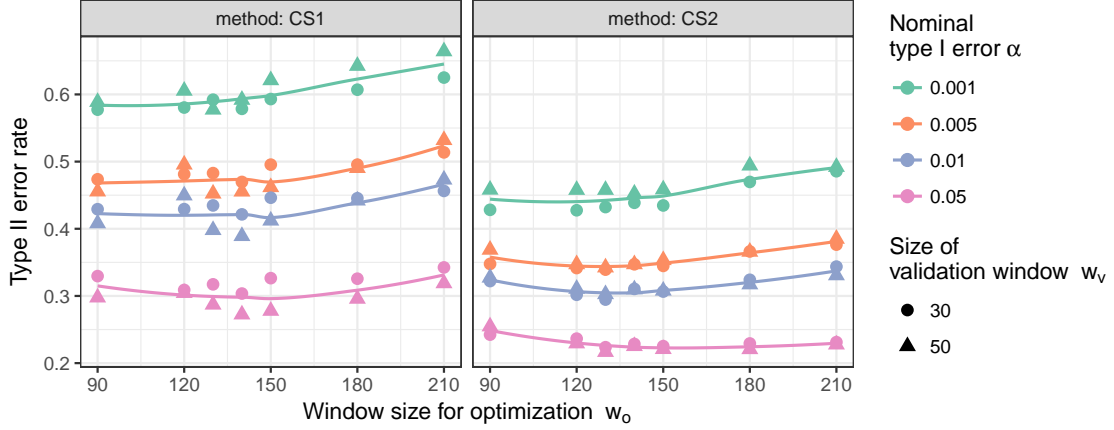


Figure 8: Type II error rates observed across a range of window sizes for optimization w_o . For a window size of $w_o = 130$ we see a minimum in type II error rate across all type I rates considered. Smaller validation sizes w_v are typically associated with a smaller type II error.

3.3 Type II error rates

?? gives an overview of the type 2 error of methods CS1 and CS2 across a range of different optimization windows w_o .

?? gives an overview of ROC (Receiver operating characteristic) curves for methods CS1 and CS2 over a range of different optimization window sizes w_o . The different color hues represent the two methods CS1 (red) and CS2 (blue). The ROC curves show the superior performance of CS2 over CS1. Generally, an optimization window w_o of 150 pixels or more leads to the best performance with respect to ROC curves.

3.4 Observed versus Nominal Type I error rates

?? shows observed type I error rates across a range of optimization windows w_o . Generally, observed type I errors are higher than expected. Method CS1 shows in this instance slightly better performance than method CS2, but for both an increase in the size of the optimization window leads to a decrease in the observed type I error rates.

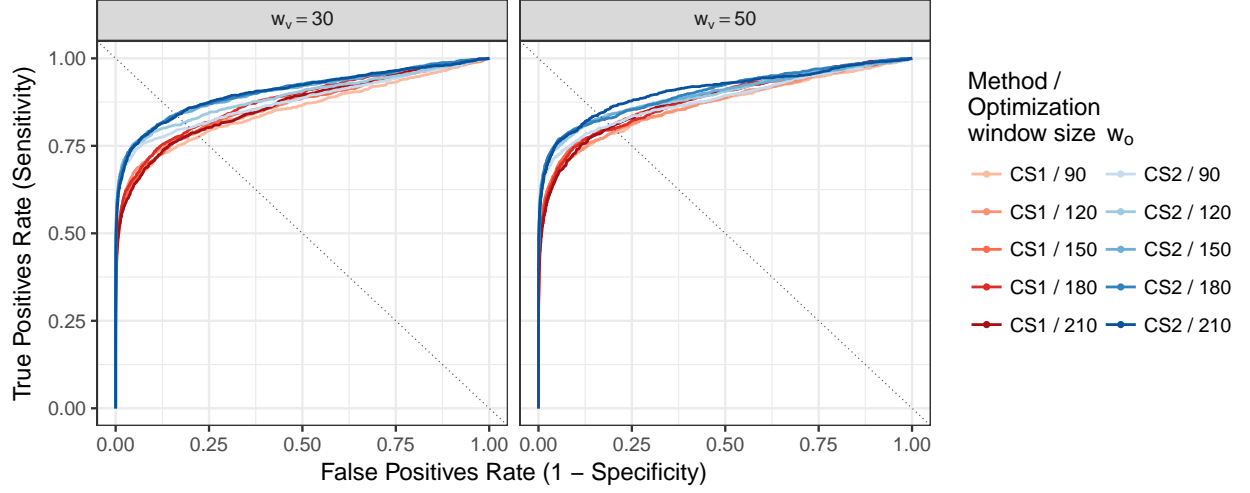


Figure 9: ROC curves of methods CS1 and CS2 for different sizes of optimization window w_o . Best performances with respect to ROC curves are reached for optimization windows of sizes 150 and higher. Points of equal error rates (EERs) can be found at the intersection of the dotted line and the ROC curves.

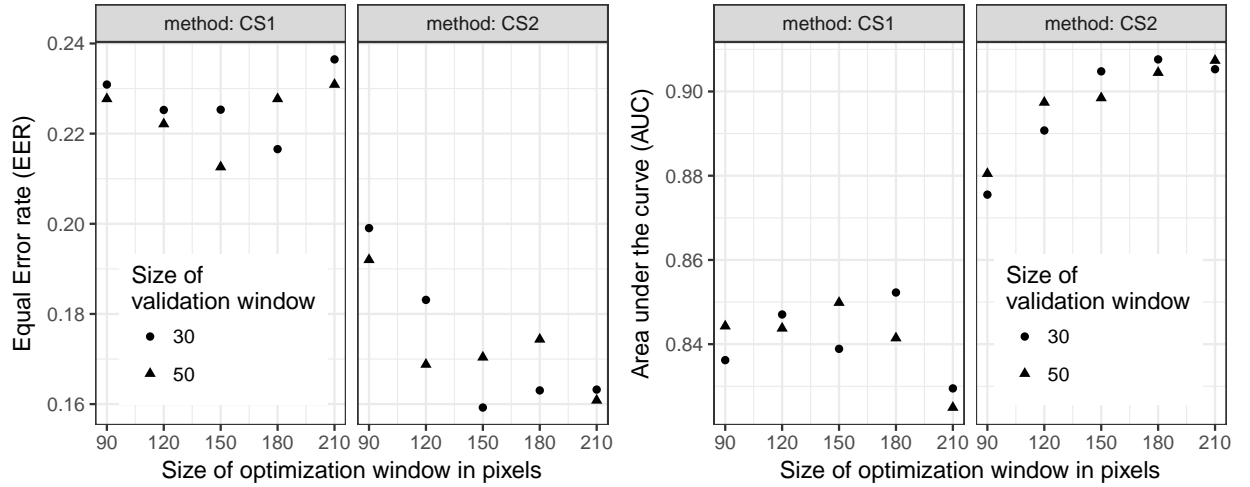


Figure 10: Comparison of results for CS1 and CS2 using area under the curve (AUC) and equal error rates (EER).

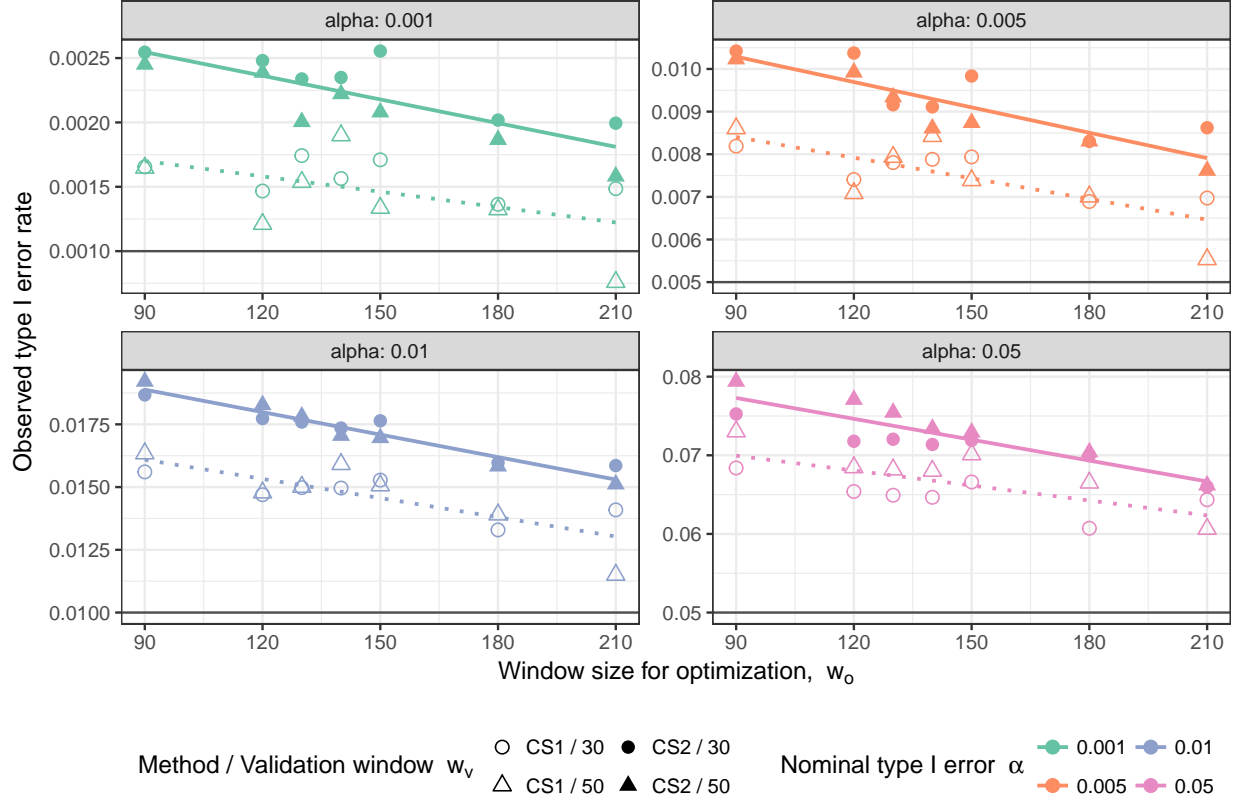


Figure 11: Comparison of observed and nominal type I error rates across a range of window sizes for optimization w_o . The horizontal line in each facet indicates the nominal type I error rate.

4 Conclusions

We started out trying to assess the suitability of the Chumbley Score’s suitability for matching striae on bullet lands. Besides using recommended defaults, we also proposed ways to optimize parameters: based on the assumption that once sub-class characteristics are removed, optimal locations are distributed uniformly across the profile, we found for bullets the smaller coarseness value of $c = 0.15$ to be suitable. Optimized window sizes were found based on cross-validation to minimize type 2 error rates. Method CS1 proposed by ? has a minimal type 2 error rate of XXX for an optimized window size of YYY - which is considerably higher than the error rates achieved on matching toolmarks. Unfortunately, method CS1 also has a high rate of failed tests – situations, in which the algorithm does not provide a result, due to the way different-shift pairs are constructed. XXX The algorithm CS2 introduced here is constructed in a way that reduces the number of failures by a magnitude of about XXX. While reducing the failure rate, the algorithm also shows an increase in the power of the test. Type II error rates reach a minimum of XXX for an optimized window size of YYY. While significantly reduced over CS1, CS2 still has type 2 error rates on bullet lands that are higher than the error rates achieved on the -much larger- toolmarks. However, bullets usually have multiple lands - in the case of Ruger P85s as used in the Hamby study, there are six lands for each bullet. We might be able to get more power out of the test by adapting CS2 to deal work in a bullet-to-bullet comparison.

Fix in CS2 should also improve power for matching toolmarks.

5 Appendix

A Scenarios of failed tests

Figures ?? and ?? show scenarios in which the deterministic Chumbley score can fail. In ?? both CS1 and CS2 fail, because the lag between optimal locations is so large that no same-shift pairs can be found once the signatures are aligned. These failures are inherent to the Chumbley Score and cannot be prevented.

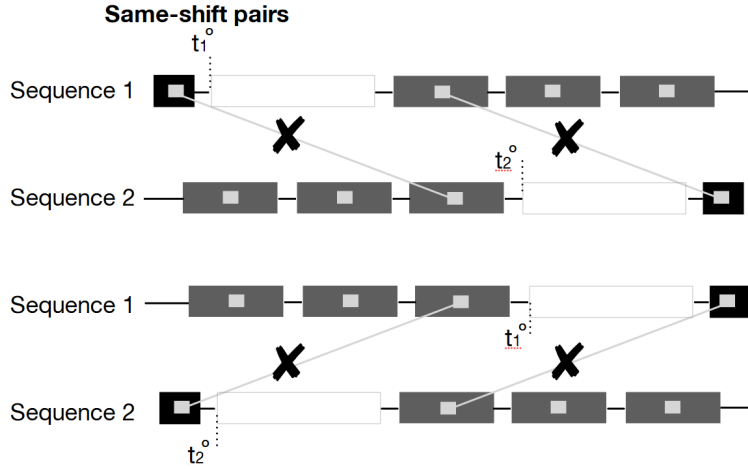


Figure 12: Sketch of same-shift pairings. When the lag between optimal reference points is too large to accommodate a validation window in both signatures, both CS1 and CS2 fail.

Figure 12 shows two situations, in which CS1 fails to identify any different-shift pairings. This happens, when both relative locations are close to either end of the signature. CS2 can still be computed in this situation as long as there are at least two same-shift pairs identified.

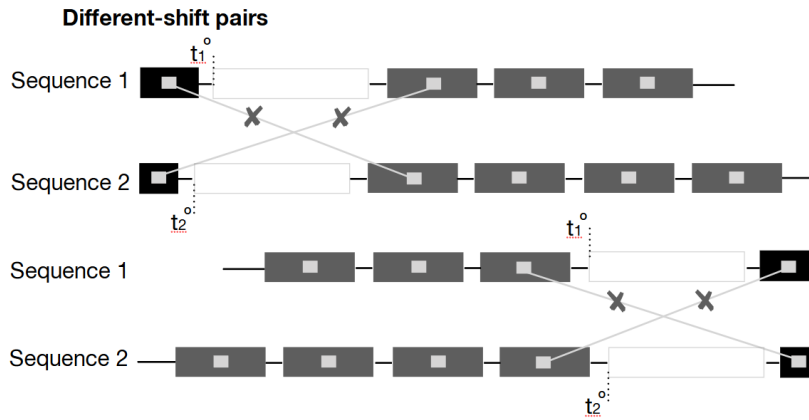


Figure 13: Sketch of different-shift pairings. For CS1 no different-shift pairs can be identified, resulting in a failed test.

Figure 13

References

AFTE Criteria for Identification Committee (1992), ‘Theory of identification, range striae

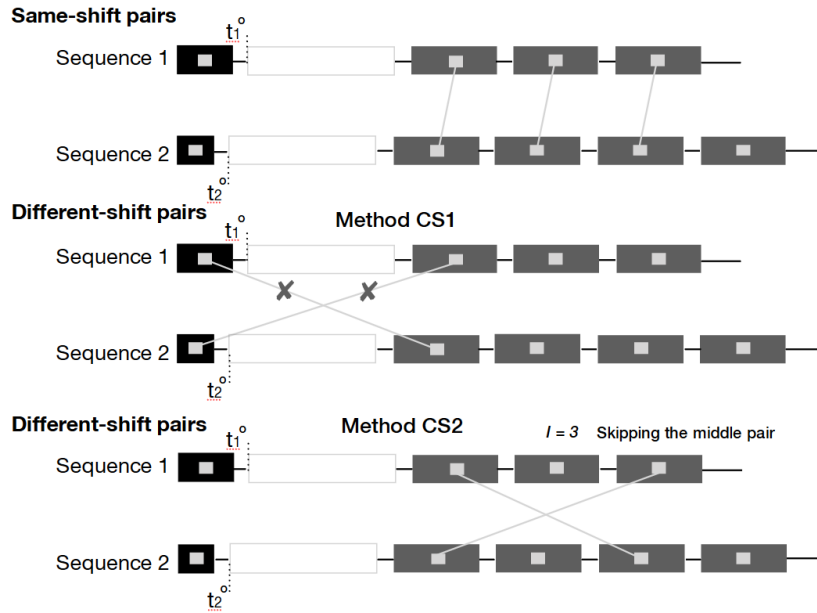


Figure 14: Sketch of a case where CS1 fails but CS2 does not fail.

comparison reports and modified glossary definitions', *AFTE Journal* **24**, 336–340.

AFTE Glossary (1998), 'Theory of identification as it relates to toolmarks', *AFTE Journal* **30**, 86–88.

Bachrach, B. (2002), 'Development of a 3d-based automated firearms evidence comparison system', *Journal of Forensic Sciences* **47**(6), 1253–1264.

Bachrach, B., Jain, A., Jung, S. and Koons, R. (2010), 'A statistical validation of the individuality and repeatability of striated tool marks: Screwdrivers and tongue and groove pliers', *Journal of Forensic Sciences* **55**(2), 348–357.

Chu, W., Thompson, R. M., Song, J. and Vorburger, T. V. (2013), 'Automatic identification of bullet signatures based on consecutive matching striae (cms) criteria.', *Forensic Science International* **231**, 137–141.

Chumbley, L. S., Morris, M. D., Kreiser, M. J., Fisher, C., Craft, J., Genalo, L. J., Davis, S., Faden, D. and Kidd, J. (2010), 'Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm', *Journal of Forensic Sciences*

55(4), 953–961.

URL: <http://dx.doi.org/10.1111/j.1556-4029.2010.01424.x>

Cleveland, W. S. (1979), ‘Robust locally weighted regression and smoothing scatterplots’, *Journal of the American Statistical Association* **74**(368), 829–836.

De Kinder, J. and Bonifanti, M. (1999), ‘Automated comparison of bullet striations based on 3d topography’, *Forensic Science International* **101**, 85–93.

De Kinder, J., Prevot, P., Pirlot, M. and Nys, B. (1998), ‘Surface topology of bullet striations: an innovating technique’, *AFTE Journal* **30**(2), 294–299.

Faden, D., Kidd, J., Craft, J., Chumbley, L. S., Morris, M. D., Genalo, Lawrence J. and Kreiser, M. J. and Davis, S. (2007), ‘Statistical confirmation of empirical observations concerning toolmark striae’, *AFTE Journal* **39**(2), 205–214.

Grieve, T., Chumbley, L. S., Kreiser, J., Ekstrand, L., Morris, M. and Zhang, S. (2014), ‘Objective comparison of toolmarks from the cutting surfaces of slip-joint pliers’, *AFTE Journal* **46**(2), 176–185.

Hadler, J. (2017), ‘toolmaRk: Tests for Same-Source of Toolmarks’. R package version 0.0.1, [Online; accessed 19-March-2018].

URL: <https://github.com/heike/toolmaRk>

Hadler, J. R. and Morris, M. D. (2017), ‘An improved version of a tool mark comparison algorithm’, *Journal of Forensic Sciences* pp. n/a–n/a.

URL: <http://dx.doi.org/10.1111/1556-4029.13640>

Hamby, J. E., Brundage, D. J. and Thorpe, J. W. (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., Hofmann, H. and Carriquiry, A. (2016), ‘Automatic Matching of Bullet Lands’, *Annals of Applied Statistics* .

Ma, L., Song, J., Whitenton, E., Zheng, A., Vorburger, T. V. and Zhou, J. (2004), ‘NIST bullet signature measurement system for RM (Reference Material) 8240 standard bullets’, *Journal of Forensic Sciences* **49**, 649–659.

National Research Council (2009), *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, DC. [Online; accessed 25-April-2018].

URL: <http://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward>

President’s Council of Advisors on Science and Technology (2016), ‘Report on forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods’. [Online; accessed 25-April-2018].

URL: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: <https://www.R-project.org/>

Vorburger, T., Song, J., Chu, W., Ma, L., Bui, S., Zheng, A. and Renegar, T. (2011), ‘Applications of cross-correlation functions’, *Wear* **271**(3-4).

URL: <https://doi.org/10.1016/j.wear.2010.03.030>.

Vorburger, T., Song, J. and Petraco, N. (2016), ‘Topography measurements and applications in ballistics and tool mark identifications. surface topography: metrology and properties’, *Surface topography: metrology and properties* **4**(1).

Zheng, X. A. (2016), ‘NIST Ballistics Toolmark Research Database (NBTRB)’. [Online; accessed 19-March-2018].

URL: <https://tsapps.nist.gov/NRBTD>