Three-dimensional data of wirecut surface scans under the confocal microscope (110 character maximum, inc. spaces)

Yuhang Lin[1,2,*] **and Heike Hofmann**[2,3]

[1]Iowa State University, Department of Statistics, Ames,
[2]Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, Ames,
[3]University of Nebraska-Lincoln, Department of Statistics, Lincoln,
[*]corresponding author(s): Yuhang Lin (yhlin@iastate.edu)

ABSTRACT

Update later
Wire cut data is important in forensic investigations but lacks a systematic way of analyzing the data. We created a data set of 120 scans of aluminum wire cut in `x3p` format, using 5 wire cutters and 3 locations along the 4 blades, with 2 replicates for each combination. A systematic pipeline with multiple analysis plots was developed to analyze the data and draw conclusions based on numerical measures.
(maximum 170 words) This is a manuscript template for Data Descriptor submissions to *Scientific Data* (http://www.nature.com/scientificdata). The abstract must be no longer than 170 words, and should succinctly describe the study, the assay(s) performed, the resulting data, and the reuse potential, but should not make any claims regarding new scientific findings. No references are allowed in this section.

Please note: Abbreviations should be introduced at the first mention in the main text no abbreviations lists or tables should be included. Structure of the main text is provided below.

1 Background & Summary

Wire cut data is a type of forensic tool mark data used to identify the source of a wire cutter based on the striations left on the surface. There have been cases where the evidence and testimony on wire cut evidence played a crucial role in the criminal investigation and conviction of a defendant.

However, there is a lack of a standardized method to analyze it except for visual comparison. Although the Association of Firearm and Toolmark Examiners (AFTE) developed the Theory of Identification (AFTE 1998) incorrect citation format, which outlines the process of comparing tool marks, it is still subjective and relies on the examiner's experience, making results hard to reproduce and validate. Several reports, such as those from the National Research Council (NRC 2009) and the President's Council of Advisors on Science and Technology (President's Council of Advisors on Science and Technology 2016), pointed out the abovementioned issues and called for more objective and reproducible methods to analyze tool marks. Early research by (Ma et al. 2004) and (Zheng, Soons, and Thompson 2016) has focused on collecting and distributing datasets for this purpose and providing a foundation for future advancements in tool mark analysis. Studies such as those by Chu et al. (2013) and Vorburger et al. (2011) have demonstrated the efficacy of using numerical methods to improve accuracy and consistency in tool mark analysis. Hare et al. (2017) and Ju and Hofmann (2022) have explored methods for quantifying the similarity between representative signals, but alignment remains a major hurdle.

In this study, we would like to follow the same path and provide a data set of wire cut scans, and also discuss a systematic pipeline to analyze the data and draw conclusions based on numerical measures. Here, we provide a data set containing multiple files, as described in Table 1.

Table 1. Overview of the whole data set.

	Description	Section
Raw data		

x3p files	120 scans of aluminum wire cut in x3p format	Cutting wires
metadata	metadata of the scans in 1 CSV	Cutting wires
Manually processed derivatives		
profiles	profiles extracted from scans in 120 CSVs	Extract profiles
Computational processed derivatives		
signals	signals filtered from profiles in 1 CSV	Filtered signals
aligned signals	pictures of pairwise aligned signals from same sources in PNGs	Align signals
CCF values	CCF values of all pairwise aligned signals in 1 CSV	Align signals

breakable table, cannot cross-reference inside table

For the reproducibility of all our data and alignment results, we introduce in detail in Section 2.1 (cross-reference not working without section number, number-sections: false not working) how we cut the wire and collect the 120 scans with 5 tools on 3 locations, in Section 2.2 how we extract profiles from the scans, in Section 2.3 how we filter signals from the profiles, in Section 2.4 how we align signals from different scans and optimize the alignment with the cross-correlation function (CCF) values. In Section 3, we discuss where our data is held. Then, in Section 4, a technical validation was conducted to further compare signals from different sources also match our assumption, together with visual aids for drawing conclusions. Finally, in Section 5, we provide available codes for creating the data set and conducting technical validation, as discussed in Section 2 and Section 4. Section 6 discusses where these codes can be found online. We hope this pipeline developed using this data set can be further generalized and applied to real crime scenes to help investigators draw conclusions based on real wire cut data.

(unlimited length) An overview of the study design, the assay(s) performed, and the created data, including any background information needed to put this study in the context of previous work and the literature. The section should also briefly outline the broader goals that motivated the creation of this dataset and the potential reuse value. We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design. The Background & Summary should not include subheadings. This section and the other main body sections of the manuscript should include citations to the literature as needed.

2 Methods

full descriptions of the experimental design, data acquisition assays, and any computational processing

In this study, aluminum wire was used to create an optimal scenario where the most amount of information could be transferred from the tool to the substrate, despite the wire in some real cases being made of lead. The physical property of aluminium wire make it an excellent candidate for keeping marks while being relatively easy to bend and non-toxic.

2.1 Cutting wires

The aluminum wire used was 16 Gauge/1.5 mm, anodized. In order to cut the wire, 4-inch pieces were unspooled and cut using Kaiweets wire cutters, model KWS-105, as shown in Figure 1(a), for 1 blade location, either inner, middle, or outer, which gives us 1 replicate. Each piece was then cut into half to create 2-inch pieces for each side, AB and CD, with a sharpie line marking the cut ends, giving us 4 samples. Here, we are showing AB sides only in Figure 1(b) (need a different tent figure), and the CD sides are similar from the other side of the cut, with the back of A being C and the back of B being D. Both AB and CD sides form tent structures on the tips of the wire, and we can separate each side of the tent into 2 pieces along the bending position, resulting in 8 scans. We repeated this process for all 3 locations along the blade and 5 wire cutters, with 2 replicates for each tool-edge-location combination, resulting in 120 scans. Each piece was labeled with the naming conventions, T(ool) 1/2/3/4/5 (Edge) A/B/C/D W(ire) - L(ocation) I(nner)/M(iddle)/O(uter) - R(epetition) 1/2, with T1AW-LI-R1 being the piece cut by tool 1 on the A edge at the inner location for the first repetition. Then, we can use the standard scanning protocols for the confocal microscope, shown in Figure 1(c) (need an extra pic of the very tip), to scan the wire tip surfaces. The scanned surfaces are saved in a resolution of $0.645\mu m \times 0.645\mu m$ per square pixel in an `x3p` file format.

XXX figure 1 - generally, zoom into these images - we do not want to have a hand in the image, nor a view of the crafting aluminum :) - what are the exact rules on visuals in Scientific Data ? XXX hard to put the full requirements here, see https://www.nature.com/sdata/publish/submission-guidelines#figures

put into quarto layout with (a) (b) (c) on the top left, no tent, add blade C & D

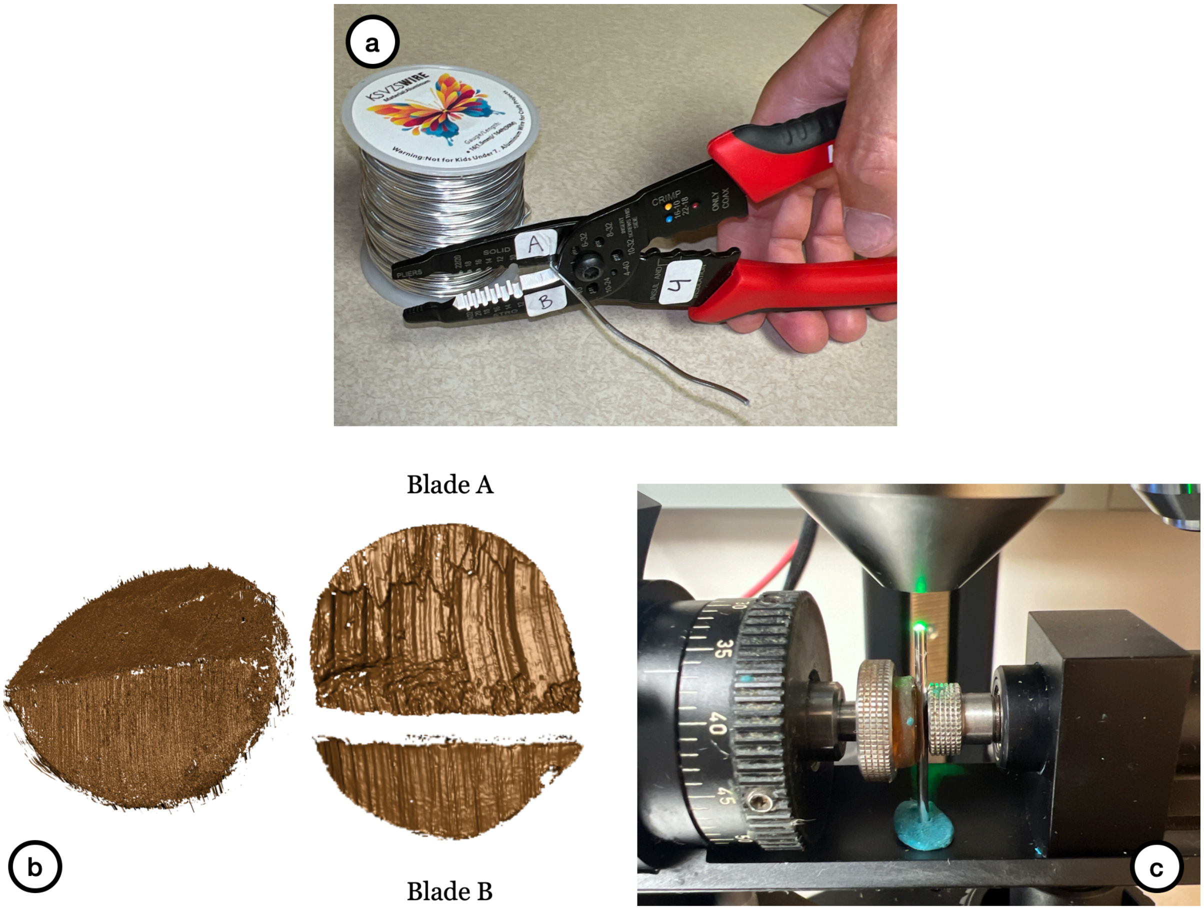

Figure 1. (a) A Kaiweets wire cutter of model KWS-105 was used to cut the wire. (b) A tent structure created by blade AB. After separating 2 tent structures by the connecting position, we obtained 2 samples - 1 sample from blade A and B. (c) A confocal microscope was used to scan the wire surfaces.

2.2 Extract profiles

Numerical comparisons between 2 replicates cannot be done directly on the `x3p` files. We need to extract representative functions from the scans first. A representative function with the most information is considered as a signal for one scan, which can be used for comparison later. To obtain this function, we first need a profile of the scan, which is a sequence of values along a user-drawn line on the surface. The profile should capture most features of the scan, and be orthogonal to the striation marks of the scan, which are formed by ups and downs of grooves. So, we draw the line across the wide region of the scan to maximize the feature captured, as shown in dark blue in Figure 2(a). We can then investigate the values under this profile line. The profile function is along the line is plotted in Figure 2(b).

2.3 Filtered signals

With the profile extracted, we can then obtain the signal. Two Gaussian filters are applied to these resulting profiles. In particular, we first used a large low-pass filter with bandwidths of 400 microns to remove large trend, as it can overwhelm the signals, and then used a small high-pass filter of 40 microns to average across noise and remove spikes, as shown in Figure 2(c). (Cleveland, Grosse, and Shyu 1992) (add reference: W. S. Cleveland, E. Grosse and W. M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). Finally, the extreme tail values are removed.

2.4 Align signals

Signals extracted from different scans can be put together for comparison, and we maximize the corss-correlation function (CCF) values between the signals to numerically find the best alignment. For example, we compare T1AW-LI-R1 to T1AW-LI-R2, T1CW-LI-R1 to T1CW-LI-R2, and so on. That is comparing each row in Figure 3. We know that signals from two replicates with the same tool-edge-location combination should yield similar signals as in the first and second column of

Figure 4, which will give alignments of massive overlapping and high CCF values close to 1. The alignments and values we got in the rightmost column of Figure 4 fulfill our expectations.

(unlimited length) The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the detailed section in our submission guidelines for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section. Subheadings should not be numbered.

Authors should review the transparent methods checklist below, and ensure that their manuscript complies with any relevant points. Authors are also encouraged to search FAIRsharing.org for community reporting standards that may be relevant to their specific data-type.

Transparent Methods Checklist

- Materials & reagents: Identify commercial suppliers of reagents, instrumentation or kits, when the source is critical to the outcome of the experiments. Declare any restrictions on the availability of unique materials (more information here). Provide catalogue or clone numbers for all antibodies (if available). For primary antibodies, provide proof of validation for the relevant species and applications.

- Exclusion criteria: If any data or samples were excluded, explain the exclusion criteria and state in the methods whether the criteria were established before the study was conducted.

- Randomization & blinding: For any studies that involve assigning samples, animals or participants into different groups: State clearly whether randomization methods were used. If randomization was not employed, this should be clearly stated. State clearly whether blinding was employed during data collection. If blinding was not employed, this should be clearly stated.

- Animal & human studies (full journal policies here): Experiments involving human participants must identify the committee approving the experiments, and include a statement confirming that informed consent was obtained from all participants. Studies employing nonhuman animals should ensure that methods descriptions comply with the ARRIVE checklist.

- Cell lines: For each eukaryotic cell line used, state the source and whether the cell line has been authenticated or otherwise tested for integrity. If any commonly misidentified cell lines were used (see ICLAC or NCBI Biosample), justify their use. Report whether the cell lines were tested for mycoplasma contamination.

- Chemistry & materials science: Manuscripts describing chemical syntheses, or characterizing new chemicals or materials should refer to the guidance at Nature Chemistry.

3 Data Records

The complete data set is available on the ISU DataShare repository at https://iastate.figshare.com/, which is public and open access for every interested researcher. The data set consists of 120 scans in the `x3p` file format with the naming convention as described before. (Explain the x3p header info?)

(unlimited length) The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited numerically in the text of this section, for example [?], and included in the main reference list as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record. Providing a direct link to the dataset may also be helpful to readers (https://doi.org/10.6084/m9.figshare.853801).

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each (please see 'Tables' below). They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

4 Technical Validation

a picture of alignment with ccf from different sources to show if different source, our evaluation returns small ccf, which matches what we thought.

For the data collection process, two team members did the cutting and labeling together, then one person did the scanning and named according to the naming convention above. The scanning was done in a specific order to ensure consistency across all scans. The data was saved in a consistent format to ensure they could be easily accessed and analyzed. A third person then checked the data to ensure that the data was consistent in naming and accurate.

again - the website is not the right place for the validation - instead, move parts from the website here.

For the validation of the scans and their processing we investigate the correlation scores of pair-wise aligned signals. For signals from scans of wires cut with a different tool, we would expect a low correlation score. Large scores between signals are indicative of being made by the same tool. Show boxplots and roc curve.

For validation of all other tools and locations of scan replicates, see the detailed report.

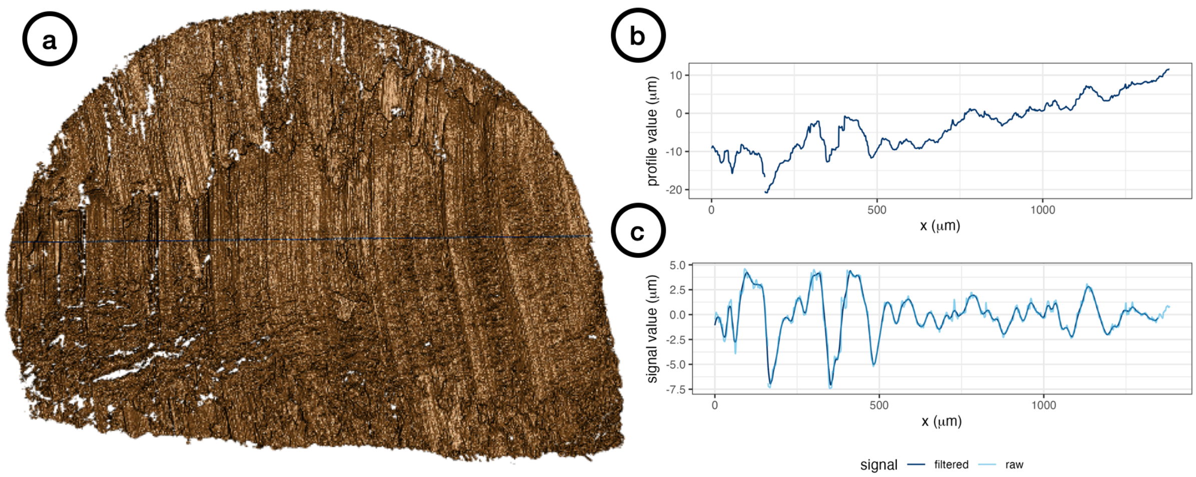

Figure 2. (a) A profile line in dark blue was drawn across the striations of the scan. (b) The profile function extracted along the profile line in (a). (c) The raw signal in light blue is obtained by using the low-pass filter on the profile function in (b) and the filtered signal is obtained by using the high-pass filter on the raw signal.

(unlimited length) This section presents any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. This is a required section; authors must present information justifying the reliability of their data.

- Measurement of data quality?
  - Numeric measurements / tests: ?
  - Visualizations: ?
  - Check with existing data: ?
  - Questionable / slur procedures:
    * AidDatas Geospatial Global Chinese Development Finance Dataset: Second, all data collected is reviewed by at least two individuals. Although this is not a double-blind review procedure, the use of satellite imagery to verify project locations results in far less uncertainty when compared to previous approaches to geocoding where locations were selected entirely based on text descriptions.
    * A large open access dataset of brain metastasis 3D segmentations on MRI with clinical and imaging information: A medical student (D.R.) double checked and adjusted the revised NIfTI segmentation masks and manually counted the number of lesions with contrast-enhancement, necrosis, and peritumoral edema for each patient.

* Time series of freshwater macroinvertebrate abundances and site characteristics of European streams and rivers: Technical validation of the TREAM dataset was achieved through exclusion of time series data that did not match our inclusion criteria and data standardisation steps (outlined in Methods above). Any noted issues that did not adhere to the outlined standardisation within the datasets from the 41 independent projects included in this dataset were checked with data providers and corrected or removed when standardisation was not achievable (e.g., when collection methods changed over the course of the time series).
* 3D surgical instrument collection for computer vision and extended reality: The main issue...Since we store our models in a standard format (STL), they are compatible with a large variety of visualisation and processing software.
* Three-dimensional reconstruction of high latitude bamboo coral via X-ray microfocus Computed Tomography : Regular quality assurance inspections are carried out on the ţ-CT scanner to verify its metrological and geometrical (alignments) accuracy for conducting the scans. The geometry of source to object and source to detector distances are verified whenever there is any significant physical interaction with the source such as re-alignment, change of filament, or source anode change. This calibration process involves scanning a specially designed phantom known as an hourglass36, which consists of three pairs of high-sphericity spheres. The sphere sizes are as follows: two spheres with a diameter of 3.000 mm, two spheres with a diameter of 6.000 mm, and two spheres with a diameter of 9.525 mm, and each sphere is kept in contact with its size-counterpart. By using this phantom, it becomes possible to accurately determine a known distance, specifically the centre-to-centre distance of the spheres, in a threshold-independent manner. If the measured distance deviates beyond the acceptable limits of metrological accuracy, the systems calibration parameters are adjusted to ensure agreement between the measured distance and the actual distance.

5 Usage Notes

The `R` package `x3ptools` (available from `CRAN`) supports working with files in x3p format.

Sample scripts in R for processing scans from x3p format to their signal are available from ... github.

Further analysis can be conducted with the `GitHub` `R` package `wire` and the `GitHub` `R` shiny app `wireShiny` (citation?). We already conduct between-replicate comparisons in the technical validation section, and we can also conduct across-replicate comparisons to establish error rates threshold and produce other analysis plots.

Suppose we put the CCF values in a tilemap with different tool, location and edge combinations. In that case, we expect only the diagonal to have high CCF values, close to 1 and marked as orange in the tilemap, as the diagonal represents the same source, and the rest of the matrix to have low CCF values, close to 0 and marked as gray. In Figure 5, the behavior is consistent with our expectation overall, except for some rare cases with tool 5 edge D, which is caused by ?????. We also put the resulting CCFs in the boxplot, as in Figure 6. We can see that the CCF values for the same sources are close to 1, while the CCF values for different sources are much lower than expected. This difference can be used to establish a threshold for CCF and help us draw conclusions about the similarity between wire cut scans numerically, which can be used in real crime scenes. The density plot in Figure 7 shows the distribution of the CCF values with the same sources and different sources. The overlapping points between the tails of these two distributions can be a rough threshold. Furthermore, the receiver operating characteristic (ROC) curve in Figure 8 shows the sensitivity / true positive rate against the false positive rate (FPR) (1 - specificity). The curve is very close to the upper left corner, which is excellent for classification and drawing conclusion. It gives us a true threshold of 0.589 to control the FPR to be less than 0.05 with false negative rate (FNR) to be 0, (false positive rate (FPR) / false discovery rate (FDR) -> define the H0 or call it false identification rate (FIR)???), and 0.658 to control the FPR to be less than 0.01, with FNR to be 0.02.

(unlimited length) The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our code availability policy for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

6 Code availability

table of code-manual?

no, we can't use the website as a place for more detailed procedures. This paper is the detailed procedure.no more README, scanning procedures in another HTML

We put together the cutting and the standard scanning procedures mentioned in Section 2.1 cross-ref not working with more pictures for each step into a README of the GitHub repository `heike/Wirecuts` (High-res pics needed in the README).

The data set can be easily accessed with the `CRAN R` package `x3ptools`. Further analysis can be conducted with the `GitHub R` package `wire` and the `GitHub R` shiny app `wireShiny` (citation) (again????).

For all studies using custom code in the generation or processing of datasets, a statement must be included under the heading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

7 End of Body

(Note that the bibliography style and the name of the bib-file are hard coded in the template file right now.)

AFTE. 1998. "The Association of Firearm and Tool Mark Examiners: Theory of Identification as It Relates to Toolmarks." *AFTE Journal* 30 (1): 86–88.

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. https://doi.org/gn65cz.

Cleveland, William S., Eric Grosse, and William M. Shyu. 1992. "Local Regression Models." In *Statistical Models in S*. Routledge.

Hare, Eric, Heike Hofmann, Alicia Carriquiry, et al. 2017. "Automatic Matching of Bullet Land Impressions." *The Annals of Applied Statistics* 11 (4): 2332–56. https://doi.org/10.1214/17-AOAS1080.

Ju, Wangqian, and Heike Hofmann. 2022. "The R Journal: An Open-Source Implementation of the CMPS Algorithm for Assessing Similarity of Bullets." *The R Journal* 14 (2): 267–85. https://doi.org/10.32614/RJ-2022-035.

Ma, Li, John Song, Eric Whitenton, Alan Zheng, Theodore Vorburger, and Jack Zhou. 2004. "NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets." *Journal of Forensic Sciences* 49 (4): 1–11. https://doi.org/cdsbv8.

NRC. 2009. *National Research Council: Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press.

President's Council of Advisors on Science and Technology. 2016. *President's Council of Advisors on Science and Technology: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, D.C.: Executive Office of the President of the United States, President's Council. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

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. https://doi.org/dt4kzk.

Zheng, Xiaoyu A., Johannes A. Soons, and Robert M. Thompson. 2016. "NIST Ballistics Toolmark Research Database | NIST." https://www.nist.gov/publications/nist-ballistics-toolmark-research-database.

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g. [**? ? ? ?**]. For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file. For journal articles, DOIs should be included for works in press that do not yet have volume or page numbers. For other journal articles, DOIs should be included uniformly for all articles or not at all. We recommend that you encode all DOIs in your bibtex database as full URLs, e.g. https://doi.org/10.1007/s12110-009-9068-2.

Acknowledgements

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Y.L. did all of the work, H.H. made him do the work. But seriously, this paper is the one where we need to cite everybody: Eden Amin, Curtis Mosher, Jeff Salyards. Alicia? Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Competing interests

(mandatory statement)

H.H. is a technical advisor to AFTE (Association of Firearms and Toolmarks Examiners), fellow of the ASA (American Statistical Association), and committee member of the ASA Forensic Science Committee. H.H. has testified as court witness on behalf of judge April Neubauer, NY State Supreme Court Criminal Term in New York City. The corresponding author is responsible for providing a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Figures & Tables

Figures, tables, and their legends, should be included at the end of the document. Figures and tables can be referenced in LaTeX using the ref command, e.g. Figure 9 and Table 2.

Authors are encouraged to provide one or more tables that provide basic information on the main inputs to the study (e.g. samples, participants, or information sources) and the main data outputs of the study. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Tables may be provided within the LaTeX document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included here. Generally, a Data Descriptor should have fewer than ten Tables, but more may be allowed when needed. Tables may be of any size, but only Tables which fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as Online-only tables and numbered separately from the main table list e.g. Table 1, Table 2, Online-only Table 1 etc.

Condition	n	p
A	5	0.1
B	10	0.01

Table 2. Legend (350 words max). Example legend text.

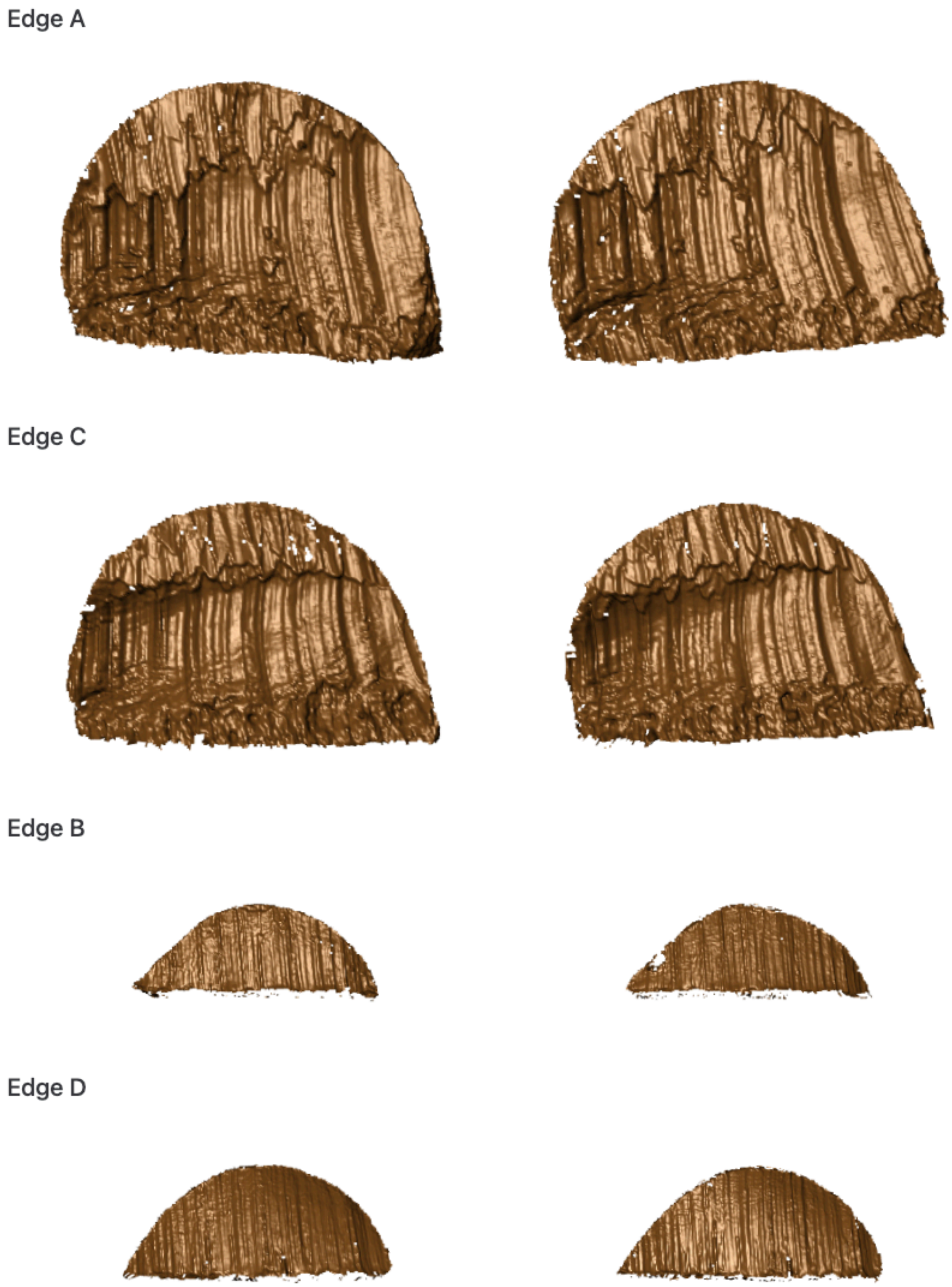

Figure 3. Scans from different sides of tool 1 at the inner location.

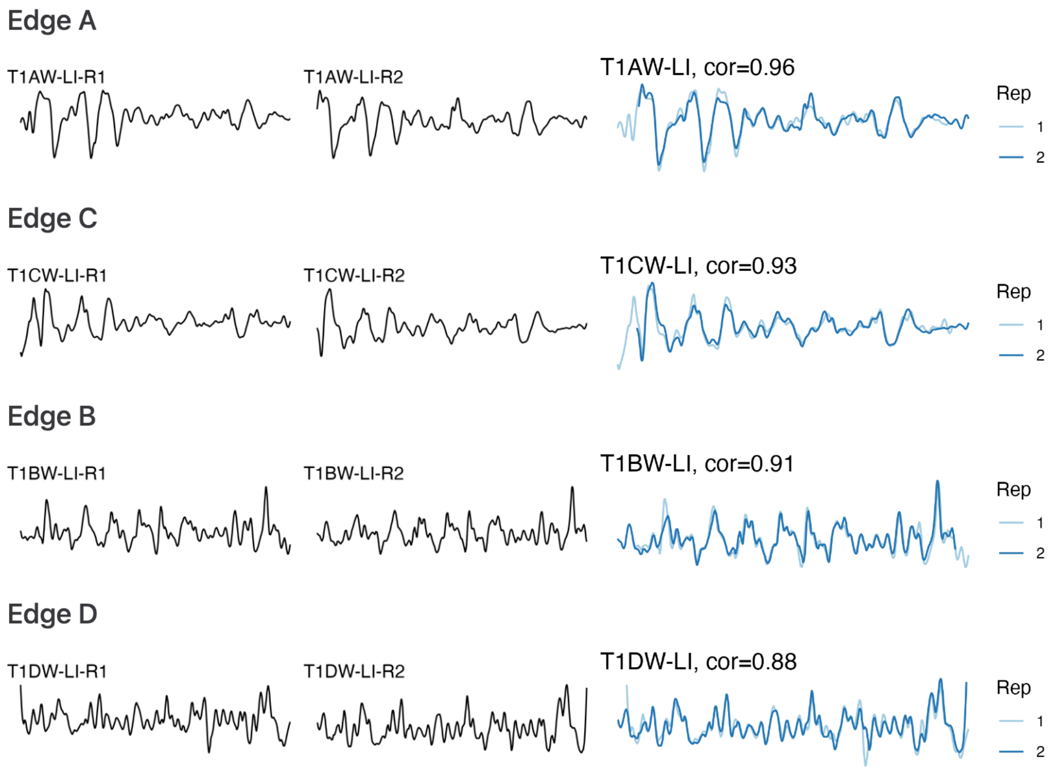

Figure 4. The first and second columns show the signals extracted from Figure 3, and the third column shows the alignments and CCF values between pairs of signals.

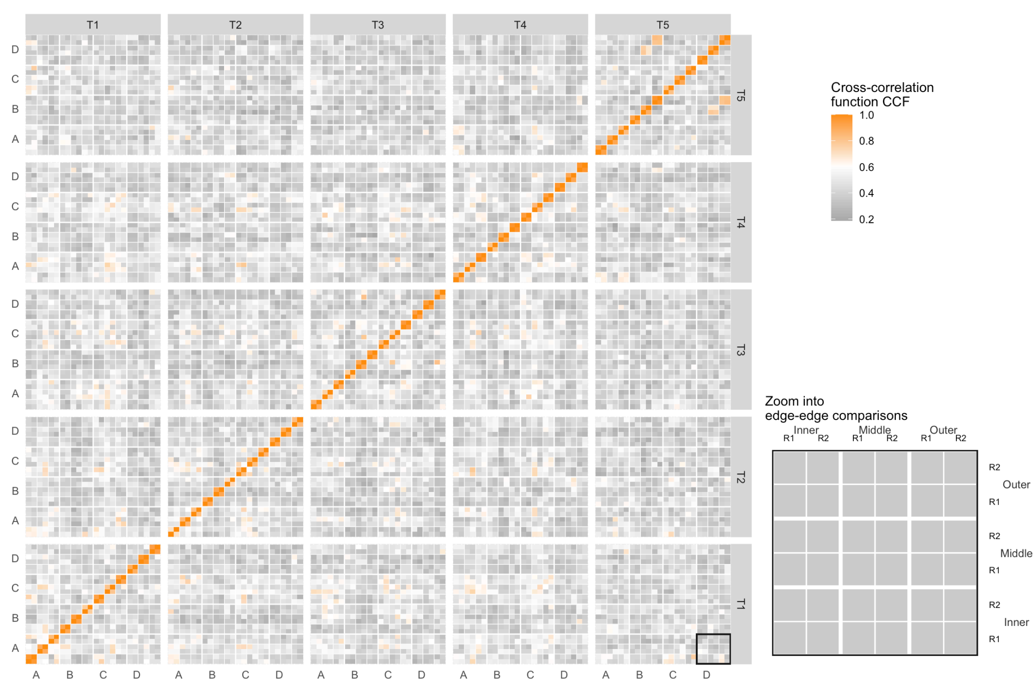

Figure 5. The tilemap shows signals from the same source have CCFs close to 1.

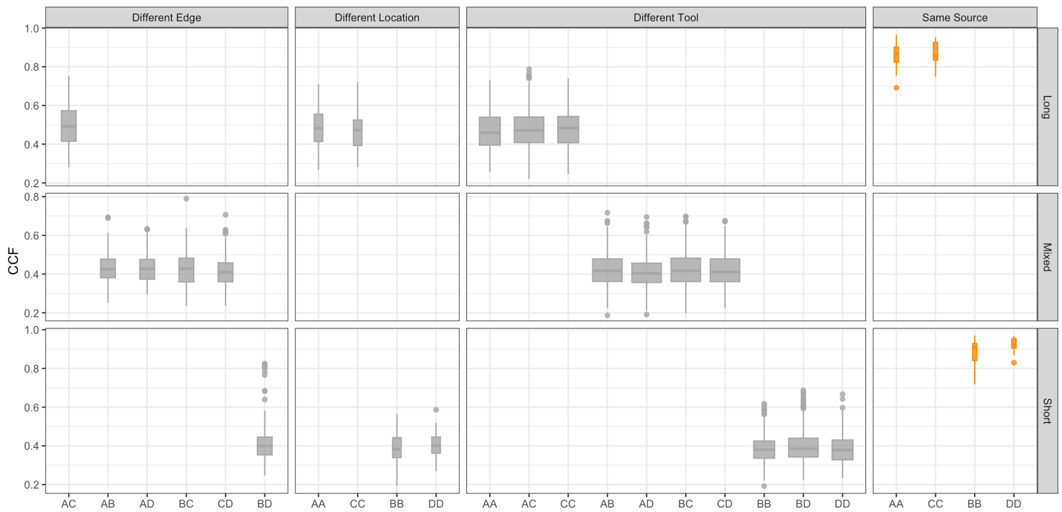

Figure 6. The boxplot shows that signals from the same sources have higher CCfs than those from different sources.

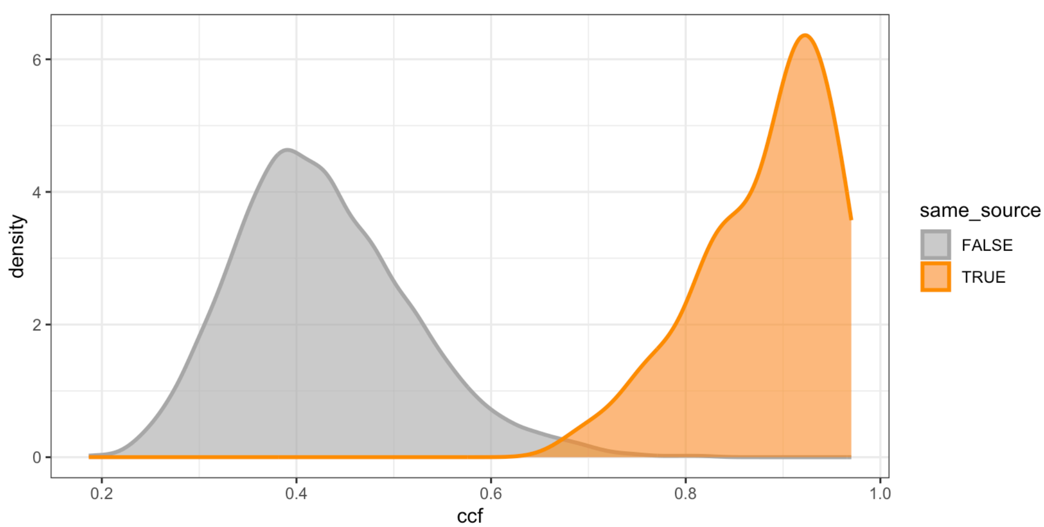

Figure 7. The density plot shows tails of distributions overlap, which can be used as a rough threshold for drawing conclusions.

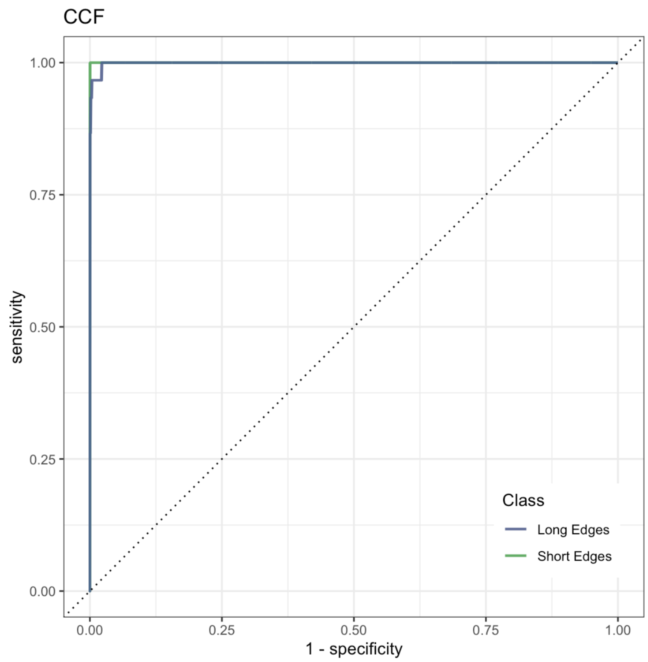

Figure 8. The ROC curve is bending very close to the upper left corner, which means excellent in classification and drawing conclusions.

Figure 9. Legend (350 words max). Example legend text.