Multiple Comparisons in Toolmark Evidence

Susan Vanderplas, Heike Hofmann

Introduction

- Problems with database searches and multiple comparisons in forensics
- Fingerprint - madrid bombing
- DNA databases
- Firearms and toolmarks?? NIBIN/IBIS issues, but more broadly there are issues

One of the common suggestions in the 2016 PCAST report across pattern disciplines in forensic science was the importance of creating huge databases of e.g. known prints and firearms to support research into matches and closest non-matches(PCAST 2016, fingerprints pg 11, 88, firearms pg 12). As researchers, we concur with this suggestion, as it is critical for developing objective algorithms and creating useful black-box studies. We have spent considerable time assembling large data sets of tool marks, bullets, and cartridge cases to support our research as well as create public resources to accelerate future research (Center for Statistics and Applications in Forensic Evidence 2023). However, we are wary about the implications of database searches across forensic disciplines, for reasons also mentioned in the PCAST report: the necessity of understanding the random match probability for searches against these databases, in particular when using degraded evidence often found in crime scenes(PCAST 2016, pg 52).

This issue has been raised in the past with regard to DNA (Thompson, Taroni, and Aitken 2003) and fingerprints (Fine 2006); one of the many issues identified in the aftermath of the 2004 Madrid bombing case was that the IAFIS database used to locate similar prints is extremely large and thus it is possible for the database to locate unusually similar non-matches. This issue will almost certainly become an issue in other pattern disciplines as databases of firearms evidence such as NIBIN Bureau of Alcohol, Tobacco, Firearms, and Explosives (2021) become more commonly utilized by law enforcement. In this paper, however, we examine a more subtle issue present in tool mark examinations: searches within evidence collected as part of a single case. We begin with a hypothetical set of three scenarios involving the collection of tool mark evidence and then assess the statistical issues involved with each scenario and explore possible resolutions to those issues.

Thought Experiment

Consider, for example a situation where someone builds a bomb, and a wire fragment is recovered from un-detonated explosive. We will examine how the forensic examination might proceed in three different cases:

1. Examiners identify a suspect and locate a single tool which is believed to have been used to construct the bomb. Examiners would like to determine whether this tool is a forensic match to wires within the undetonated bomb.
2. Examiners identify a suspect and locate a garage full of tools, one of which may have been used to construct the bomb. Examiners would like to identify which tool(s) were used to cut the wires during bomb construction.
3. Examiners identify several suspect, each of whom has a set of tools, and wish to determine which suspect(s) may have constructed the bomb based off of the toolmark comparisons.

We will consider wires ranging in diameter from 2mm (solid 12 gauge wire) to 0.0116 mm (stranded 12 gauge wire may be composed of 49 strands of 0.0116 mm diameter wire) for the purposes of this thought experiment; these values represent types of wire easily obtainable at a local hardware store.

Single Tool Comparisons

Let us first consider the situation where a single tool is recovered from the suspect's house that is viewed as likely to have made the wire cut. Then, a forensic examiner might take the tool and use it to cut a piece of metal, known as a blade, in order to record the markings of the entire cutting surface simultaneously. This process might be repeated at multiple angles in order to record the effect of different contact surfaces on the resulting striations. For simplicity, let us assume that angles are not a factor. Let us further define the blade to be of length ℓ and the wire to have diameter d which is fully covered with striations that are untouched by the blast.

At minimum, the examiner will have to do ℓ/d comparisons in order to assess whether the tool matches the wire; when a comparison algorithm is used, and both the blade and the wire are scanned at the same resolution of r mm, there may be as many as $\ell/r - d/r + 1$ comparisons performed automatically in order to find the optimal alignment between the two samples[1].

To make this more concrete, consider the pair of pliers shown in Figure 1, which has a 1.5 cm cutting surface machined on both sides to produce a peaked cross-section. When used to cut a wire, the wire rests against a rectangular surface used to hold it in place; the blade is pushed

[1]This assumes we restrict the wire to be completely contained within the bounds of the blade surface. If we allow for partial matches and restrict these matches to require at least k points of overlap, then this becomes $\ell/r - d/r + k + 1$.

into the wire, producing striae resulting from imperfections in side A on one half of the wire and striae from imperfections and wear in side B on the other.

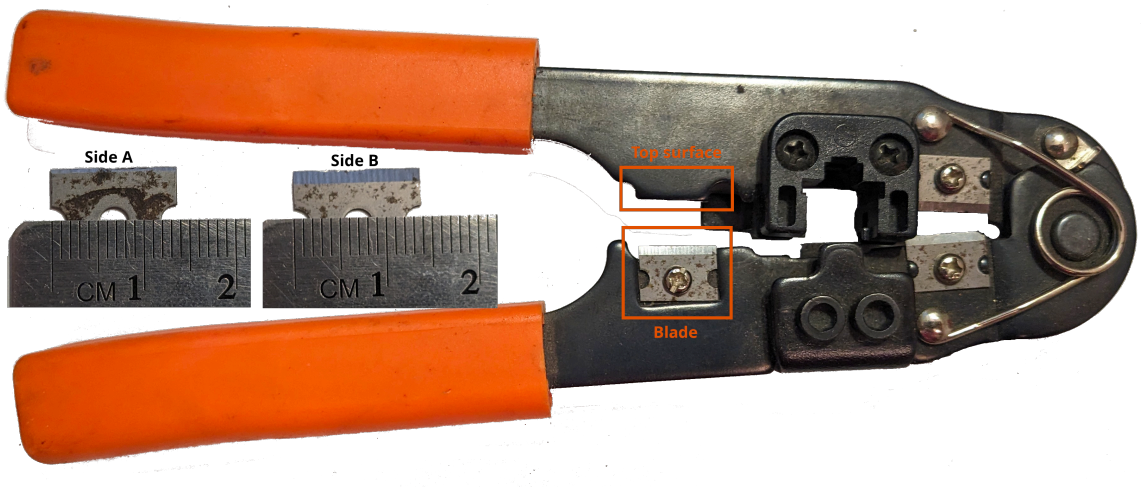


Figure 1: Wire cutter with 1.5 cm cutting surface beveled on both sides

A scan of a cut made by the blade at $0.645\mu m$ resolution will produce approximately $15 \times 1000/0.645 = 23,256$ observations; a scan of a 2mm wire would produce $2 \times 1000/0.645 = 3,101$ observations, and computationally aligning these two data sequences using maximum cross-correlation would require $2.3256 \times 10^4 - 3101 + 1 = 20,156$ (obviously dependent) comparisons. Under an optical microscope, an examiner might move the wire at 0.2mm increments along the blade cut surface for a total of $15/0.2 - 2/0.2 + 1 = 66$ distinct comparisons. Then, we must consider that the examiner or algorithm must compare the wire to both side A and side B; we then double the number of total comparisons made to $40,312$ and 132 for the algorithm and the examiner, respectively.

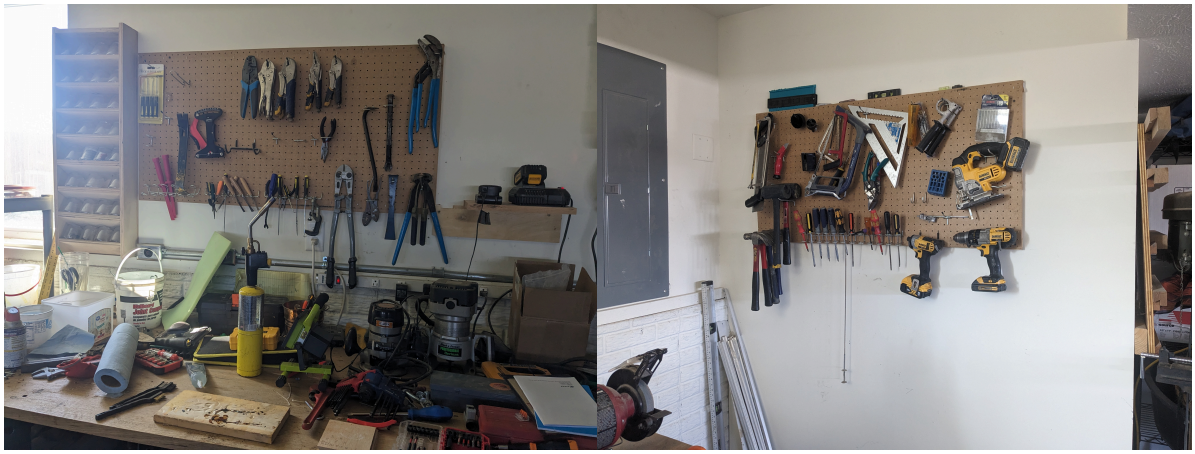
Ultimately, in order to compare tool marks in this scenario, the examiner or algorithm completes the equivalent to a database search resulting from a single piece of evidence. The dangers of such searches are well known and have been cited in forensic missteps such as the Madrid bombing case, as discussed in Li et al. (2023); Newman (2007); Inspector General Oversight and Review Division (2006, Pg 137).

While 132 visual comparisons might not seem so terrible, this is the simplest of the 3 scenarios we consider; each scenario presents a progressively more complex problem.

A Garage of Comparisons

In our next scenario, investigators identify a suspect and collect all of the tools in the suspect's residence which might have been used to construct the bomb.

Let us consider the garage of one of the authors of this paper; Figure 2 shows the garage work bench and tool storage in typical condition. While there is no reason to think the author's garage is representative of all garages, the general setup and number of tools in the garage is not unusual for a suburban household where residents may dabble in automotive repair, home improvement, or woodworking.



(a) Shop Bench Tool rack (b) Shop Side Tool rack

Figure 2: Layout of tool storage in the author's garage. For this study, we considered only tools which might reasonably be used to cut wires.

We assembled the easily available tools from the house and garage which might reasonably be used to cut wires[2], including pliers, wire cutters and strippers, scissors, tongs, and utility knives. Estimated cutting surface length and quantity of each type of tool is provided in Table 1. The total length of all cutting surfaces of relevant tools in the author's garage is 982 cm; primary contributions to the total length are the 50 pack of utility knife blades and the 14 pairs of full-size scissors which could be located throughout the house. Of course, a thorough search of the garage and house, as one would expect from professionals investigating an actual crime, would yield far more cutting surfaces to compare against.

As in Scenario 1, the wire fragment recovered from the un-detonated explosive would be compared to blade surface test impressions generated from each cutting surface in the garage. However, unlike in Scenario 1, we have 982 cm of cutting surfaces to work with. We can apply our comparison formula on a by-surface basis to yield the digital and analog comparison estimates shown in Table 2. Overall, our algorithm would make 14,592,452 (dependent) comparisons and an examiner doing a visual inspection would make 1,527 comparisons under the assumptions we made in Scenario 1.

[2]Reasonably is defined as we (the authors) have used a tool like this to cut wires while doing home improvement projects in the past.

Table 1: Shop cutting surfaces. All lengths in cm.

Type	Length (cm)	Cutting Surfaces	# Tools	Total Cutting Length (cm)
Pliers	1.5	2	10	30
Pliers	1.0	4	3	12
Side cutters	1.0	4	1	4
Side cutters	2.0	4	2	16
Scissors	8.0	2	14	224
Tin snips	4.0	2	4	32
Lineman pliers	2.0	4	2	16
Lineman pliers	1.0	4	1	4
Wire strippers	1.0	2	2	4
Long handled cutting pliers	2.0	4	1	8
Long handled cutting pliers	3.0	4	1	12
Utility knife blades	6.2	2	50	620

Table 2: Comparions performed by tool type, assuming digital scans at a resolution of $0.625\mu m$ and visual comparisons performed at $0.2mm$ intervals along the test blade impression.

Tool type	Algorithm	Examiner
Lineman pliers	272,880	132
Long handled cutting pliers	285,279	232
Pliers	551,970	107
Scissors	3,386,075	391
Side cutters	272,880	132
Tin snips	471,326	191
Utility knife blades	9,302,426	301
Wire strippers	49,616	41
Total	14,592,452	1,527

Table 3: Total cutting length of easily accessible wire-cutting tools for 4 individuals associated with the author.

Individual	Total Cutting Length (cm)	Algorithm	Examiner
0	982.0	14,592,452	47,264
1	431.6	6,288,502	20,410
2	897.0	13,287,022	43,050
3	826.4	12,186,249	39,502
4	2,243.0	33,584,880	108,694
Total	5,380.0	79,939,104	258,920

Multiple Suspects

To fully understand the magnitude of this problem, suppose that we have a group of 4 additional individuals who may have contributed to building the un-detonated bomb. These individuals are each associated with the author and have hobbies that include building electronic devices, home improvement, welding, and woodworking (thus, they each have sets of tools). As part of the investigation, each individuals' tools are confiscated and examined; blade cuts are made of each cutting surface to be matched to the wire in evidence. Table 3 shows the number of comparisons which would need to be performed to ensure that all wire-cutting tools in each suspect's garage are compared to the evidence from the scene.

If investigators were to attempt to determine which of the suspected individuals cut the wire in evidence, they would need to do 258,920 distinct visual comparisons. Even if there is an incredibly small false positive error rate, the odds of a false identification are much more likely in this scenario than either of the previous two scenarios. But how likely? In the next section, we describe two different scenarios for calculating the overall likelihood of a false identification using simple probability calculations and estimated coincidental match probabilities as well as a theoretical approach designed to mimic the approach used with database searches that return the N most similar results.

The Problems with Database Searches

Random Match Probability

Score-based Distributions

Potential Solutions

We need to have an explicit relationship between random match probability and amount of signal (assessed by length, number of features, overall striae depth, etc.). Until we have this kind of relationship, it will be hard to use error rates (even if they were known for toolmarks) from algorithms or black-box studies in practical situations.

We need much more study of real data, examiner evaluations, and performance of algorithms in situations which mimic real casework. We also need to quantify the strength, distinctiveness, and total number of striations to ensure that comparisons are being made only in situations where random matches are unlikely to occur.

Bureau of Alcohol, Tobacco, Firearms, and Explosives. 2021. "National Integrated Ballistic Information Network (NIBIN)." https://www.atf.gov/firearms/national-integrated-ballistic-information-network-nibin.

Center for Statistics and Applications in Forensic Evidence. 2023. "CSAFE Forensic Science Dataset Portal." *Open Data Portal*. https://forensicstats.org/data/.

Fine, Glenn A. 2006. "A Review of the FBI's Handling of the Brandon Mayfield Case Unclassified Executive Summary." Washington DC. http://www.latent-prints.com/images/Final%20OIG%20Executive%20Summarylow.pdf.

Inspector General Oversight, Office of the, and Review Division. 2006. "A Review of the FBI's Handling of the Brandon Mavfield Case." https://oig.justice.gov/sites/default/files/archive/special/s0601/final.pdf.

Li, Shuo, Kang Li, Jun Yang, Yiwen Liu, Wenqiang Han, and Yaping Luo. 2023. "Research on the Local Regional Similarity of Automatic Fingerprint Identification System Fingerprints Based on Close Non-Matches in a Ten Million People Database—Taking the Central Region of Whorl as an Example." *Journal of Forensic Sciences* 68 (2): 488–99. https://doi.org/10.1111/1556-4029.15196.

Newman, Drew. 2007. "The Limitations of Fingerprint Identifications." *Criminal Justice* 22 (1): 36–41. https://heinonline.org/HOL/P?h=hein.journals/cjust22&i=38.

PCAST. 2016. *President's Council of Advisors on Science and Technology: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President of the United States, President's Council.

Thompson, William C., Franco Taroni, and Colin G. G. Aitken. 2003. "How the Probability of a False Positive Affects the Value of DNA Evidence." *Journal of Forensic Sciences* 48 (1): 2001171. https://doi.org/10.1520/JFS2001171.