Applications of the Gauge Repeatability and Reproducibility Framework to Complex Data Science Structures

Kiegan Rice

The title is a work in progress. Don't love it yet.

1 Introduction

Gauge Repeatability and Reproducibility (Gauge R&R) studies are traditionally used in engineering fields as part of measurement systems analysis. Gauge R&R studies focus on measuring the repeatability and reproducibility of a measurement process within and across environmental conditions [Vardeman and VanValkenburg, 1999]. Their study design assumes a univariate measurement value resulting from some measurement process, and aims to quantify sources of variability around a single measurement average.

As data collection and analysis become more complex, the Gauge R&R framework must be adapted to allow for a wider definition of measurement data. We focus on adapting the traditional Gauge R&R framework to complex data structures. In the following, we describe the complex nature of modern measurement data and methods for adapting the Gauge R&R framework to quantify measurement repeatability and reproducibility.

Consider the length of a manufactured screw, y. If the length of that screw y is measured by multiple operators using multiple tools, a Gauge R&R model can be used to determine to what degree the variability in the set of resulting measurements can be attributed to differences across operators or tools.

Gauge R&R studies utilize a basic random-effects model, assuming one overall measurement mean and a set of random effects corresponding to study design factors, such as operator and measurement device. The quantities of interest are primarily the variance components associated with each random effect, which describe both the overall measurement process variability as well as the degree to which measurement variability is associated with each study factor.

There are several useful applications of variance components obtained from a Gauge R&R study. One major outcome is the ability to compare any two objects for similarity. Consider the difference in measured length between two manufactured screws, say y_a and y_b: $y_a - y_b = d$. We can assess whether the magnitude of

d falls within a reasonable range of measurement variability given previously quantified repeatability and reproducibility components. If the difference in measured length d falls within an acceptable range, the two objects can be considered quantitatively similar after allowing for variability in measurement.

The length of the screw in our example is a univariate response measurement. Quantifying the similarity of two objects and accounting for measurement variability is more challenging when working with complex data structures.

1.1 Complex Data Structures

The birth of data science and significant improvements in computing power have opened up new avenues in measurement systems analysis. Data collection and analyses are much quicker and easier today than ever before. Measurement systems can collect data in greater detail, often taking on complex forms with unique structures.

Consider our example of screw length. We may want to measure multiple characteristics of that screw to more completely capture the object as a whole. To get a sense for whether any two screws are truly similar we can compare more than just a singular characteristic. The problem of comparing the set of screw characteristics measured on any two manufactured screws to determine whether they are sufficiently similar is then more difficult. Screws are a simple traditional example which naturally applies in engineering fields. However, there are many non-traditional data forms which can be captured using measurement systems, due in part to the growth of data science, a field which often leverages complex data structures.

The field of data science often leverages complex data structures. Data science deals in "pipelines" that begin with complex data structures, apply sequential actions such as data processing and data transformation to the data to extract useful information, and achieve some quantitative result [Donoho, 2017]. In many applications, the pipeline involves data captured using a measurement system, and there is measurement uncertainty associated with the resulting data.

In order to leverage complex data structures while taking measurement variability into account, we must adapt the definition of distance between any two objects with complex structure, and carefully consider how we apply statistical models to quantify measurement variability.

There have been several approaches which apply Gauge R&R models to multivariate data. Sweeney [2007] considers two-dimensional response data and accounts for the correlation between the two responses of interest when estimating variance components. Peruchi et al. [2013] proposes a weighted principal components (WPC)

approach to multivariate Gauge R&R and compares it with estimates found using MANOVA and traditional principal components analysis (PCA). WPC, MANOVA, and PCA are all applied to data with either four- or five-dimensional response measurements which each represent a different characteristic of an object of interest.

Many data structures utilized in data science include much higher dimensionality than the four- or five-dimensional data utilized in Peruchi et al. [2013] and the bivariate response in Sweeney [2007]. In addition, there are often structures which inherently differ from traditional multi-dimensional response data. We focus on an application with a more complex structure, and as such we must consider a different modeling approach.

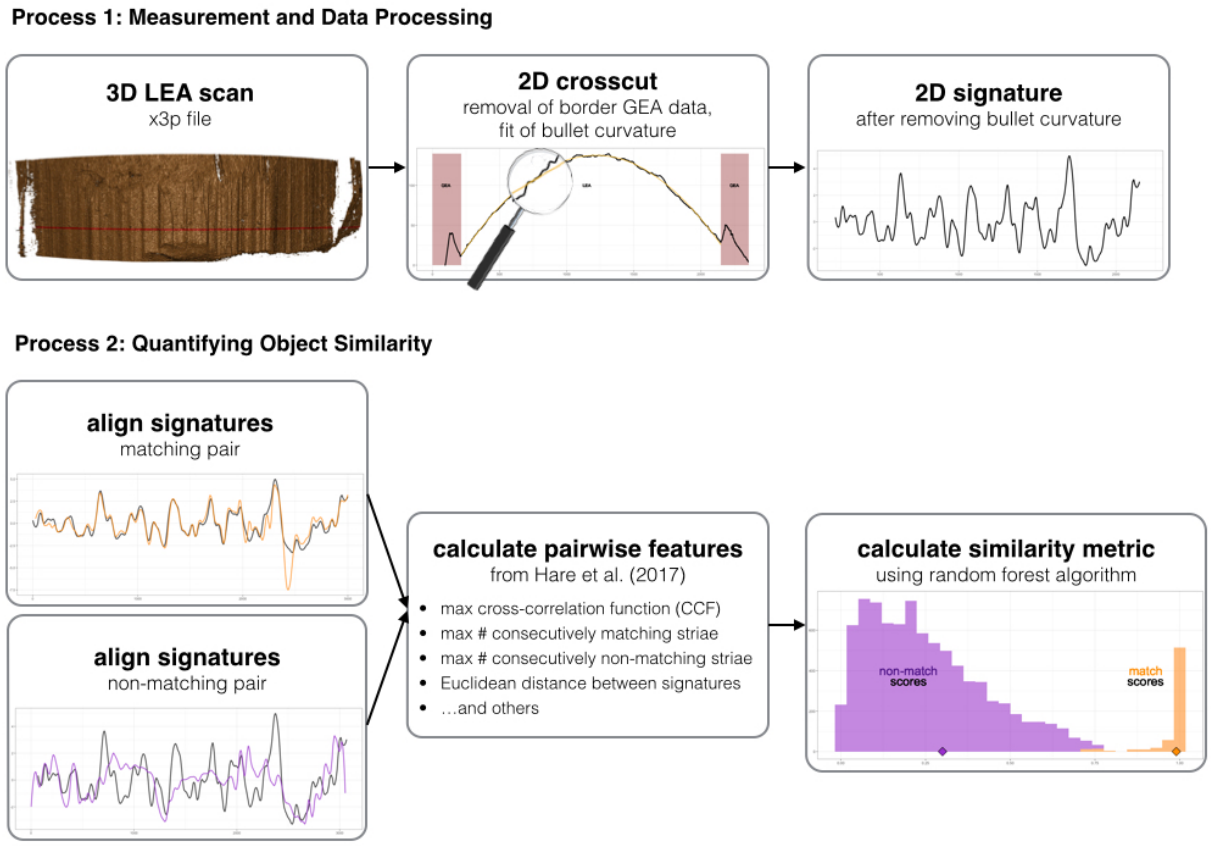

Figure 1: Process of automated firearms comparison as proposed by Hare et al. [2017]. There are two stages of measurement within the full process. First, 3D microscopy is used to capture a three-dimensional surface topography of a bullet LEA and automated data processing techniques are used to remove bordering GEA data as well as bullet curvature. The result is a two-dimensional LEA signature. The second process involves pairing and aligning two data objects, extracting pairwise features, and calculating a pairwise similarity metric.

1.2 Automated Forensic Firearms Analysis

We consider the application of data science to automated forensic firearms comparison. When a bullet is fired through a gun barrel, imperfections on lands inside the barrel engrave striation patterns on the surface of the bullet in alternating sections called land engraved areas (LEAs). Striation patterns, which visually appear as a series of peaks and valleys, are the primary evidence used to answer the forensic question of interest: whether two bullets originated from the same source or different sources.

In recent years, researchers have proposed several approaches to automating the comparison of bullet striation patterns using surface topographies of bullet LEAs captured using high-resolution microscopes (De Kinder and Bonifanti [1999], Chu et al. [2010], Chu et al. [2013], Hare et al. [2017]). An overview of the comparison process proposed by Hare et al. [2017] is shown in Figure 1.

Automated bullet comparison uses a data science approach to translate the complex striation pattern on bullet LEAs into measured data and subsequently calculate similarity metrics for any two bullet LEA striation patterns. Two LEA patterns originating from the same source (the same land in the same barrel) should have high measures of similarity, while two LEA patterns originating from different sources should receive low measures of similarity.

The data pipeline developed by Hare et al. [2017] relies on the high-resolution microscopy measurement system. The repeatability and reproducibility of captured data and resulting similarity metrics are important to establish. However, the complex nature of striation patterns does not naturally lend itself to the traditional Gauge R&R framework for univariate measurement data.

This paper presents two adaptations to the Gauge R&R framework as they apply to two distinct data structures within the forensic firearms analysis data science pipeline. We use automated forensic firearms analysis as a motivating example for the two approaches described hence.

The structure of this paper is as follows: we first describe the general three-factor Gauge R&R model and assumptions, then describe two forensic firearms analysis data structures in detail and propose meaningful adaptations to the Gauge R&R framework to best leverage repeated measurement data. We then describe the scope of data collected in our study, present results from both model frameworks, and discuss the implications of our work.

2 Methodology

The traditional three-factor gauge R&R model is defined as follows. For parts $p_j : j = 1, \ldots, n_p$, operators $o_k : o = 1, \ldots, n_o$, devices $d_m : m = 1, \ldots, n_d$, and repetitions $r_n : n = 1, \ldots, n_r$, let y_{jkmn} be the measured response value for part j, operator k, device m, and repetition n. Define a mixed-effects model,

$$y_{jkmn} = \mu + p_j + o_k + d_m + po_{jk} + pd_{jm} + od_{km} + pod_{jkm} + \epsilon_{jkmn}, \tag{1}$$

where μ is a fixed, unknown measurement mean and all other model components are random effects MONTGOMERY and RUNGER [1993]. Assume each p_j, o_k, d_m, po_{jk}, pd_{jm}, od_{km}, and pod_{jkm} are independent and identically distributed random variables which follow normal distributions centered at zero, but with respective different variances; for example, $p_j \overset{iid}{\sim} N(0, \sigma_p^2)$. We also assume $\epsilon_{jkmn} \overset{iid}{\sim} N(0, \sigma^2)$.

The model estimates of primary importance in measurement systems analysis are the variance components associated with each random effect. For example, the estimated value of σ_d^2, $\hat{\sigma}_d^2$, estimates the variance in the measurement process associated with differences across measurement devices. In addition, we consider two summary values:

$$\sigma_{repeatability} = \sqrt{\sigma^2} \quad \text{and} \tag{2}$$

$$\sigma_{reproducibility} = \sqrt{\sigma_o^2 + \sigma_d^2 + \sigma_{po}^2 + \sigma_{pd}^2 + \sigma_{od}^2 + \sigma_{pod}^2}. \tag{3}$$

$\sigma_{repeatability}$ estimates variability of measurements taken under the same environmental conditions, while $\sigma_{reproducibility}$ estimates variability of measurements across differing environmental conditions.

To adapt this traditional three-factor model to our firearms analysis data pipeline, we must carefully consider the structure of our data.

Automated forensic firearms analysis utilizes data to answer the forensic question of whether two bullets were fired through the same gun barrel. Determination of same-source or different-source origin for any two bullets is achieved by measuring the degree of similarity between the two objects. Within the automated analysis pipeline, this is achieved with two stages of measurement, both depicted in Figure 1. The first measurement process involves capturing the surface topography of bullet LEAs and extracting a two-dimensional "signature" which represents the striation patterns engraved by the gun barrel. The second measurement process deals

with pairs of signatures by extracting pairwise features and subsequently calculating a pairwise similarity metric. Both measurement processes result in data structures which are too complex for the scope of the traditional Gauge R&R model defined in Equation 1.

Measurement process 1 translates a physical object – the surface of a bullet LEA – to a complex data structure. Trained microscope operators stage bullets under high-resolution microscopes to capture a set of relative height values for a grid of equidistant (x, y) locations. The result is an x3p (XML 3-D Surface Profile) file which contains a matrix of height measurements $\mathbf{Z} = z_{i,q}$ where $i = 1, \ldots, n_x$ and $q = 1, \ldots, n_y$. The pattern of interest on the bullet surface is the striation pattern, a series of peaks and valleys corresponding to scratches engraved by gun barrel imperfections.

We extract the striation pattern using the two-step procedure proposed in Hare et al. [2017], shown in Figure 1. First, we identify an optimal cross-section height, y_{opt}, and extracting ten consecutive rows of height measurements corresponding to the y_q indices $q = (opt - 4, \ldots, opt + 5)$. The resulting height measurement following the first step is the averaged measurement z across ten rows by x location i, $\bar{z}_{i\cdot}$. Second, we remove extraneous data on the edges originating from neighboring groove engraved areas (GEAs), and model and remove bullet curvature, resulting in a LEA "signature" capturing the striation pattern of peaks and valleys as deviations from global bullet structure. For additional details on signature extraction, see Hare et al. [2017].

The resulting LEA signatures are a series of residual values $e_i = \hat{z}_{i\cdot} - \bar{z}_{i\cdot}$, where $\hat{z}_{i\cdot}$ are predicted height values for location x_i from a LOESS model fit to the bullet curvature in the second step. The series of relative height measurements and their corresponding locations x_i are the structure of interest, so for clarity we will denote LEA signatures as a set of measurements (x_i, z_i), $i = 1, \ldots, \ell$ where ℓ is the number of x locations on a signature originating from land L. Figure Figure 2 depicts the structure of a single signature emphasized in the foreground, as well as repetitions of matching signatures from three same-source bullets captured by multiple microscopes and operators. The repetitions are shown in grey, and demonstrate variability within the measurement process of translating striation patterns on bullets to signature data objects.

Measurement process 2 deals with answering the forensic question: whether two bullets were fired from the same gun barrel. The forensic question is addressed at the LEA level, comparing pairs of LEAs from two bullets to determine whether any of the LEA striation patterns match. Determination of whether two LEA striation patterns match is based on a similarity metric which aims to separate measures of similarity from "matching" LEA pairs from measures of similarity of "non-matching" LEA pairs. In the Hare et al. [2017] data pipeline, calculating a similarity metric begins with extracting a set of pairwise features which represent

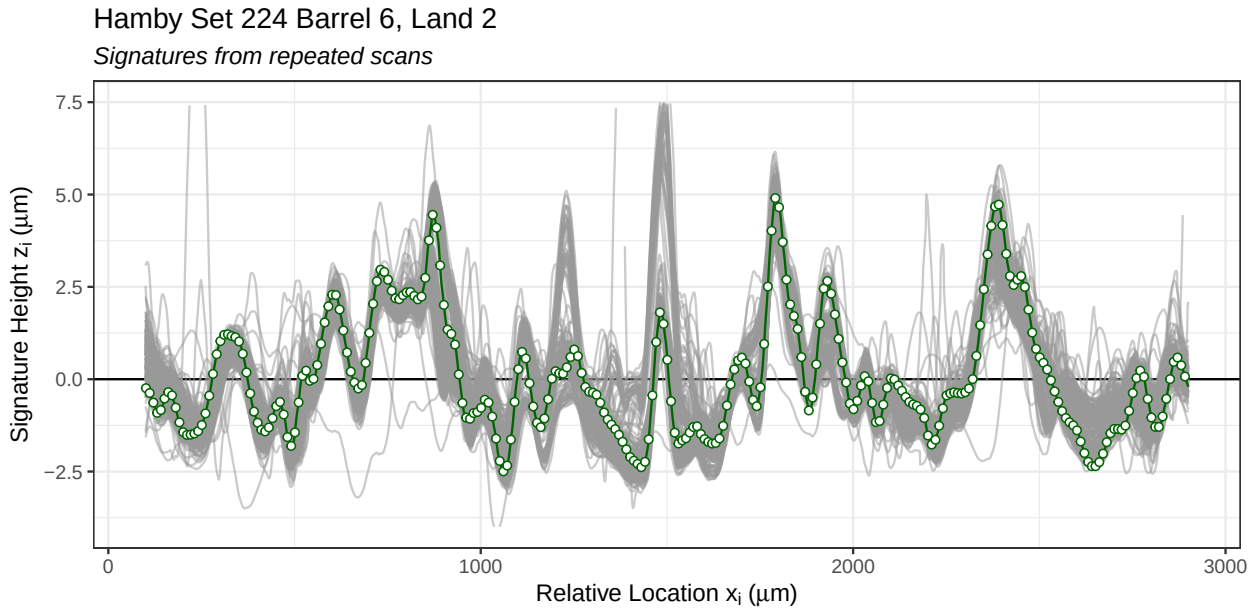

Figure 2: Two-dimensional signature data utilized in automated forensic firearms analysis. The data structure, emphasized in green for every 10^{th} data point in one single signature, consists of equidistant x_i locations with corresponding relative height values z_i. Most signatures contain several missing z_i values, which are recorded as 'NA'. Repeated signature captures gathered for land engraved areas originating from the same gun are shown in grey.

paired characteristics. These features, such as the maximum cross-correlation function and the number of consecutively matching striae, provide information about the level of similarity between any two objects. The set of paired features are then combined using a random forest algorithm, resulting in a similarity score which is then compared to distributions of similarity metrics for matching pairs and non-matching pairs. The similarity score for any two signatures takes possible values in $[0, 1]$. For additional information regarding the pairwise features and resulting similarity score, see Hare et al. [2017]. The pairwise nature of comparison introduces complexity to measurement analysis; modeling variability in relation to environmental conditions requires a redefinition of the framework traditionally used to consider environmental conditions in Gauge R&R.

We next propose adaptations to the traditional Gauge R&R framework that accommodate the complex data structures resulting from both Measurement Process 1 and 2: LEA signature data and the pairwise similarity score.

2.1 Reframing Gauge R&R for Signature Data

The structure of LEA signature data introduces two primary obstacles to utilizing the traditional Gauge R&R framework. First, the measured response is a set of structured measurements rather than a single, univariate

measurement. Secondly, the structure of peaks and valleys in LEA signatures violates the assumption of independence for measured response values. To address these incongruities, we make two major adaptations to the traditional three-factor Gauge R&R model. The first adaptation focuses on the model's fixed effect structure, while the second focuses on data reduction for compliance with independence assumptions.

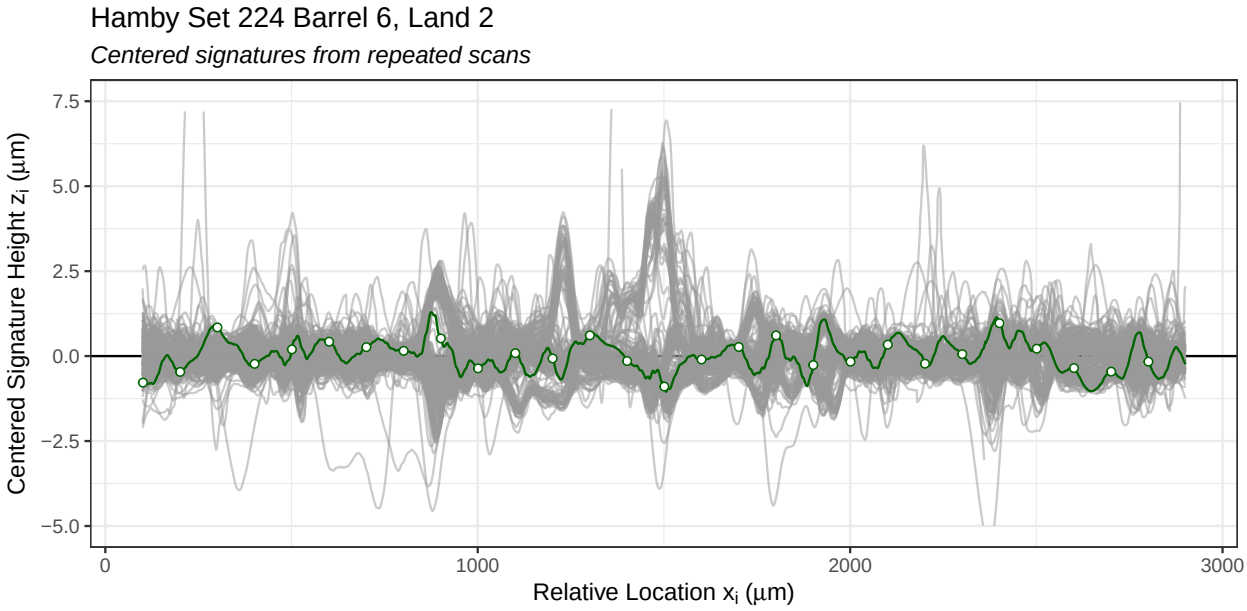

Figure 3: Two-dimensional signature data, centered by subtracting the mean by location x_i. One exemplar signature is shown in green, with points emphasized at each 100^{th} x_i location.

The three-factor model defined in Equation 1 defines all effects as random save for a single fixed effect: a fixed, unknown measurement average, μ. LEA signature data inherently contain a more complex structure of center, as signatures are not made up of a univariate response corresponding to a single measurement value. Failing to account for differences in structural height by location will result in mischaracterization of variance components within the model; variability estimates will include the object's structural variability rather than measurement variability. To address this, we define the model's fixed effect structure by location x_i on each land L: μ_{Li}.

The practical implication of including a fixed effect by location in the model is demonstrated in Figure 3. After centering signature height values by the mean height measurement from all repetitions at location x_i, remaining variability around the center is greatly reduced, and more adeptly captures measurement variability rather than structural variability.

Despite this adjustment, there is still significant structure remaining in the LEA signature data which may cloud estimation of quantities of interest, such as the variance components associated with operator and device. One approach to mitigating this misrepresentation is to carefully define the grouping structure used

to estimate variance components. We consider, for example the variability of measurements across devices. It is of interest whether one device captures more extreme values, such as higher peak measurements and lower valley measurements. If we define a random effect for device and group by device along, without accounting for structural differences by location, extreme highs and extreme lows will average out across the entire group and the variability will be artificially deflated. We address this by defining our random effect groups by both study factor and location Li. This is accomplished by interacting location with study factor in a random effect.

Adaptations to the fixed effect and grouping structure ensure that the estimated variance components associated with our model accurately represent quantities of interest. However, the independence assumption for response values is violated and needs to be addressed before a random effects model can be properly applied to the LEA signature data structure.

Estimating covariance matrices for signature data and incorporating data covariance into the model would be computationally intensive and unnecessarily complex for answering questions of repeatability and reproducibility in bullet LEA data. We therefore focus on a different approach. To uphold the traditional model structure and satisfy the independence assumption, we propose data reduction through subsampling at equidistant x_i locations. The resulting data is then a set of measurements

$$(x_i, z_i) : i = 1, 1 + w, 1 + 2w, \ldots, 1 + cw \tag{4}$$

where $w \in \mathbb{N}$ is some subsampling window size, with positive integer c such that $1 + cw < \ell$, where ℓ is the maximum index i for x_i locations for signatures from land L. The necessary window between sampling locations, w, will differ based on application and data structure. Here, we propose a window of $w = 100$ based on the autocorrelation functions (ACFs) of signature data, which are depicted in Figure 4. Signature data structure is not a traditional dependence data structure, such as time series or spatial data, with monotonic decrease of dependence with increase in lag. We also completed a sensitivity analysis for window size w, and on average estimated variance components remained consistent for window sizes between $w = 10$ and $w = 150$.

The resulting model we use to estimate measurement variability of LEA signatures is then as follows. Let z_{Lijkmn} be the measured signature height value for land $L = 1, \ldots, 6$ at location $x_i : i = 1, 101, \ldots, 1 + 100 * c$, on part j, scanned by operator k, on device m, and at repetition n.

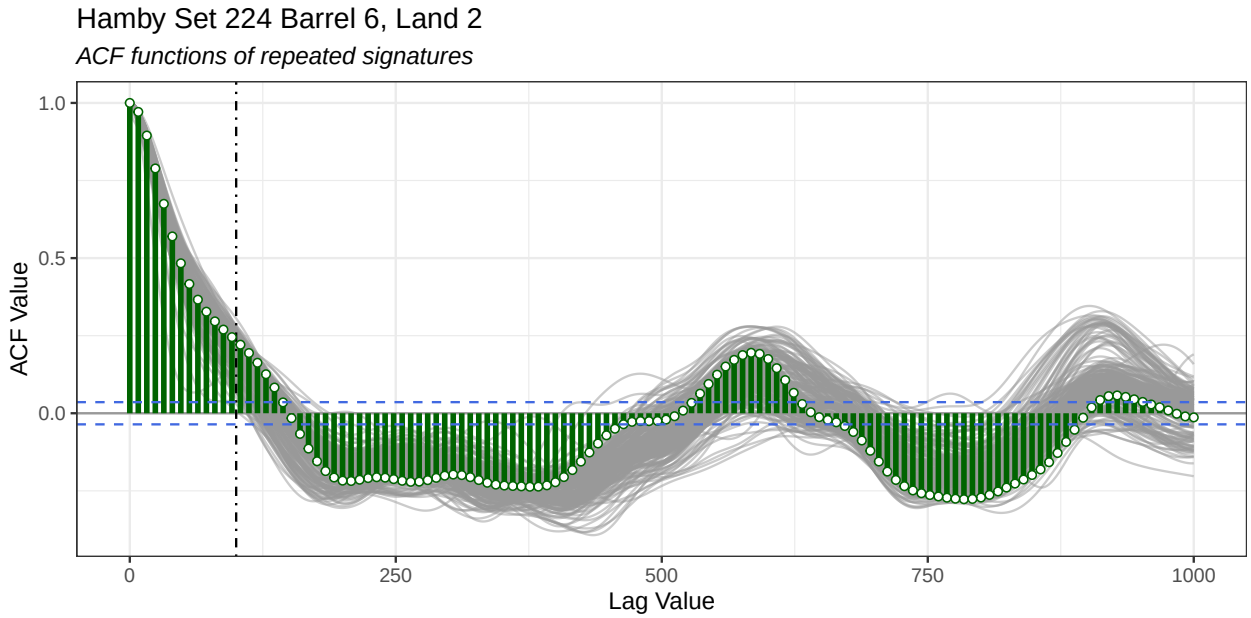

Figure 4: Autocorrelation functions (ACFs) for each repeated signature of Barrel 6, Land 2, shown for lags 0 to 1000. One exemplar ACF is shown in green in the foreground. Autocorrelation drops of significantly for the signature structure by lag 100, in this example here crossing zero near lag 150. When considering all six lands in the barrel, the ACF is typically low by lag 100.

$$z_{Lijkmn} = \mu_{Li} + p_{Lij} + o_{Lik} + d_{Lim} + po_{Lijk} + pd_{Lijm} + od_{Likm} + pod_{Lijkm} + \epsilon_{Lijkmn}, \tag{5}$$

where μ_{Li} is a fixed, unknown measurement average by location index i on land L, and all other model components are random effects. Assume each p_{Lij}, o_{Lik}, d_{Lim}, po_{Lijk}, pd_{Lijm}, od_{Likm}, and pod_{Lijkm} are independent and identically distributed random variables which follow normal distributions centered at zero and have respective differing variances; for example, $p_{Lij} \overset{iid}{\sim} N(0, \sigma_p^2)$. Also assume $\epsilon_{Lijkmn} \overset{iid}{\sim} N(0, \sigma^2)$.

One additional aspect to consider is the significant reduction in data caused by our subsampling scheme. To address this concern, we fit a series of phased models for each set of repeated signatures which use data subsampled using the same window size w, but whose starting index is staggered throughout the width of a single window. To fit ten phased models with a window size $w = 100$, we stagger the starting index at each 10^{th} location:

$$
\begin{aligned}
\text{Phase 1:} &\quad 1,\ 101,\ \ldots,\ 1 + 100 * c \\
\text{Phase 2:} &\quad 11,\ 111,\ \ldots,\ 11 + 100 * c \\
&\quad \vdots \quad\ \vdots \\
\text{Phase 10:} &\quad 91,\ 191,\ \ldots,\ 91 + 100 * c.
\end{aligned} \tag{6}
$$

The phased approach incorporates a larger proportion of the original data points, and with ten individual estimates for each variance component, provides a range of uncertainty for our estimates while still satisfying the independence assumption needed for each individual model. The combination of subsampling and phased models helps achieve a balance between over-downsampling and reducing dependency.

We next consider the second data format in the Hare et al. [2017] data pipeline.

2.2 Reframing Gauge R&R for Similarity Scores

The structure of pairwise similarity scores presents an obstacle to the Gauge R&R framework which differs from the LEA signatures. Similarity scores resulting from Measurement Process 2 are univariate, singular measurement values. They do not pose the challenge to independence that LEA signatures present. However, the levels of study factors parts, operators, and devices no longer have a one-to-one relationship with the levels of study factors represented in a single response measurement. In addition, a singular measurement average μ as the fixed effects structure (as in Equation 1) does not fully capture the mean structure present.

We first consider the problem of paired study factors. When any two pairs of measured LEA signatures are compared, the signatures may have been collected under two differing sets of environmental conditions. For example, two same-source LEA signatures may have been scanned on the same machine, but by different operators. When considering the response value, z, we must carefully consider how we index the environmental conditions.

The primary values of interest are variance components associated with differing environmental conditions. In order to determine whether the pairwise measurement process is reproducible across measurement conditions, we investigate variability of pairwise scores resulting from scans in a variety of environmental conditions. To capture this mechanism, we consider each possible pair of environmental conditions as a grouping. For example, instead of considering whether the object was measured by Operator A, B, or C, we consider the

pair of operators represented in the pairwise score: Operator A and Operator A (A-A), Operator A and Operator B (A-B), Operator A and Operator C (A-C), et cetera. This grouping structure provides the closest parallel to traditional R&R grouping structure: all pairs of environmental conditions are represented as groups, and we measure variability of resulting scores across those groups. We can then estimate how much of the variability in pairwise scores can be attributed to differences between any pair of operators. If one operator has scans that differ from other operators, that mechanism will be captured in their pairings with each of the other operators.

We therefore re-index our response data in the following way. Let $z_{(L)(j)(k)(m)(n)}$ be the pairwise similarity score resulting from Measurement Process 2 for land pairing $(L) : 1-1, 1-2, \ldots, 6-6$, part pairing $(j) : 1-1, 1-2 \ldots, 3-3$, operator pairing $(k) : A-A, A-B, \ldots, G-H, H-H$, device pairing $(m) : 1-1, 1-2, 2-2$, and repetition pairing $(n) : 1-1, 1-2, \ldots, 5-5$. The updated indexing addresses our first modeling concern.

The second major adaptation we make to the Gauge R&R framework redefines the fixed effects. When considering pairwise scores from pairs of lands, $(L) : 1-1, 1-2, \ldots, 6-6$, we must account for the nature of the similarity metric. As shown in Figure 1, similarity scores for matching pairs fall within a small spread towards the top of the range, while non-matching pairs have a much more diffuse distribution which falls to the middle and lower end of the range. We expect matched pairings $(1-1, 2-2, \ldots, 6-6)$ to have similar mean structures; however, we do not expect the same property of the non-matching pairings. Non-matching pairs are not necessarily different *in the same way*. That is, scores for the pairing $1-2$ may be very low and close to 0.1, whereas scores for the pairing $1-4$ may be higher and fall near 0.4. For this reason, we consider the land-to-land pairing index to be a critical component of the fixed effects structure. We expect scores for each land-to-land pairing to vary around a common mean $\mu_{(L)}$, and aim to measure that variability.

The spread of scores also differs by land pairing: non-matching land pairs have more diffuse distributions, and some land-to-land pairs may have more or less diffuse score distributions than others by nature. To account for these distributional differences, we also interact land pairing index with our random effect grouping structures to ensure we fully capture variability due to environmental conditions. The resulting model for pairwise scores is then:

$$\begin{aligned} z_{(L)(j)(k)(m)(n)} = \mu_{(L)} + p_{(L)(j)} + o_{(L)(k)} + d_{(L)(m)} + po_{(L)(j)(k)} + pd_{(L)(j)(m)} + od_{(L)(k)(m)} \\ + pod_{(L)(j)(k)(m)} + \epsilon_{(L)(j)(k)(m)(n)}, \end{aligned} \tag{7}$$

We next describe the scope of data collected, present results from modeling both LEA signature data and pairwise similarity scores, and discuss conclusions reached from adapting the modeling framework.

3 Data Collected

The data collected for this study were x3p (XML 3-D Surface Profile) data objects capturing the surface topography of bullet LEAs. The resulting data structures, as described previously, are two-dimensional LEA signatures resulting from data processing during Measurement Process 1, and pairwise similarity scores for pairs of LEA signatures resulting from Measurement Process 2. We next describe the scale of the study and environmental conditions.

3.1 Study design

Our Gauge R&R study focused on varying three environmental conditions: parts, devices, and operators. In our study, parts are considered to be bullets of origin. We focus on three bullets originating from the same Ruger P-85 gun barrel. The bullets, which originate from Hamby set 224, were physically fired and collected as part of a large-scale study on accuracy of firearms examination conclusions [Hamby et al., 2009]. Each bullet contains six LEAs, each engraved by contact with one of the six lands inside the same gun barrel.

Each LEA was scanned on two devices by eight operators at least three times. Each operator was tasked with collecting between three and five repeated scans of each bullet on each device (high-resolution microscope), spread out over time as a series of "Rounds". A single round consisted of a capture of each LEA on each bullet once on Microscope 1 and once on Microscope 2. Operators were directed to complete each round in its entirety before moving to the next round. Seven operators completed five rounds, while one operator completed three rounds. This collection scheme resulted in a total of 228 repeated captures of each of the six LEA signatures, with 1368 LEA signatures in total.

While the three bullets originate from the same barrel, pairwise comparisons are completed at the land-to-land level, resulting in a set of pairwise comparisons which represent a mix of same-source and different-source pairs. For example, each LEA on Bullet 1 only originates from the same source (here land) as one of the six LEAs on Bullet 2. All other comparisons are considered different-source pairs. The result is over 930,000 paired comparisons, with over 100,000 same-source comparisons and over 700,000 different-source comparisons.

4 Results

Within our collected data, there were several LEAs with surface damage. In forensic science practice, firearms examiners deem LEAs with damage unsuitable for comparison and they are removed from consideration. Under each model, we fit models which include all available data as well as a set of models which removed all signatures originating from damaged LEAs. Both sets of results are reported for both LEA signatures and pairwise scores.

Models were fit using the lmer() function in the `lme4` package in R, and Restricted Maximum Likelihood (REML) was used for parameter optimization.

4.1 LEA signatures

Variance components were estimated for each individual land L as well as for a set of data pooled across all six lands $L = 1, \ldots, 6$. The model in Equation 5 was used for all seven sets of estimates, with land index L held constant for models on each individual land. Ten phased models were fit for each data set with window size $w = 100$. Results are reported in Table 1 by variance component as a mean, min and max estimate per component across the ten phased models. Results are also shown visually in Figure 5.

The variance components with the largest magnitude are those associated with parts and residual error. In our study, parts are bullets of origin. This result is not surprising, due to the fact that striation patterns contain some amount of structural variability originating from the engraving process, as opposed to measurement variability due to differences in operators or microscopes.

Removing damaged LEAs from consideration reduced the magnitude of estimated variance components, most noticeably for the part (bullet) effect and residual error, though a noticeable difference is also seen for the part-operator interaction.

Summary values $\sigma_{repeatability}$ and $\sigma_{reproducibility}$, reported in Table 2, demonstrate the same behavior, with models having lower summary values when damaged LEA data is removed. The value of $\sigma_{reproducibility}$ is lower in magnitude than that of $\sigma_{repeatability}$ for all six lands and the pooled model, which indicates systematic differences in environmental condition (operator and device) contribute less to overall measurement variability than that of inherent measurement error or variability in how the same striation patterns are engraved across different bullets.

Table 1: LEA signature R&R model estimates for variance components using the model defined in Equation 5. Mean estimate from 10 phased models (min estimate, max estimate).

Land	σ_p	σ_o	σ_d	σ_{po}
1	0.51 (0.46, 0.57)	0.07 (0.06, 0.08)	0.04 (0.00, 0.06)	0.14 (0.12, 0.17)
damage excluded	*0.24 (0.20, 0.30)*	*0.08 (0.07, 0.10)*	*0.09 (0.07, 0.11)*	*0.09 (0.05, 0.13)*
2	0.44 (0.32, 0.55)	0.08 (0.05, 0.11)	0.08 (0.08, 0.09)	0.04 (0.00, 0.08)
3	0.55 (0.44, 0.62)	0.00 (0.00, 0.02)	0.08 (0.07, 0.10)	0.13 (0.11, 0.15)
4	0.80 (0.65, 0.98)	0.00 (0.00, 0.03)	0.03 (0.00, 0.06)	0.18 (0.16, 0.20)
damage excluded	*0.16 (0.12, 0.19)*	*0.06 (0.06, 0.08)*	*0.05 (0.03, 0.07)*	*0.10 (0.07, 0.11)*
5	0.45 (0.38, 0.53)	0.06 (0.00, 0.08)	0.06 (0.03, 0.07)	0.27 (0.23, 0.44)
6	1.30 (1.27, 1.36)	0.20 (0.09, 0.28)	0.05 (0.00, 0.10)	0.38 (0.32, 0.44)
damage excluded	*0.33 (0.15, 0.46)*	*0.05 (0.00, 0.11)*	*0.03 (0.00, 0.06)*	*0.04 (0.00, 0.10)*
Pooled	0.70 (0.66, 0.74)	0.09 (0.00, 0.06)	0.05 (0.00, 0.10)	0.21 (0.20, 0.24)
damage excluded	*0.44 (0.37, 0.51)*	*0.06 (0.05, 0.07)*	*0.08 (0.07, 0.08)*	*0.16 (0.14, 0.17)*

Land	σ_{pd}	σ_{od}	σ_{pod}	σ
1	0.07 (0.00, 0.10)	0.03 (0.00, 0.05)	0.00 (0.00, 0.00)	0.53 (0.52, 0.54)
damage excluded	*0.04 (0.03, 0.04)*	*0.04 (0.03, 0.06)*	*0.01 (0.00, 0.05)*	*0.32 (0.29, 0.34)*
2	0.00 (0.00, 0.00)	0.02 (0.00, 0.06)	0.06 (0.00, 0.09)	0.55 (0.52, 0.59)
3	0.04 (0.03, 0.09)	0.02 (0.00, 0.04)	0.02 (0.00, 0.05)	0.42 (0.39, 0.46)
4	0.04 (0.00, 0.09)	0.02 (0.00, 0.10)	0.14 (0.08, 0.18)	0.59 (0.56, 0.61)
damage excluded	*0.01 (0.00, 0.02)*	*0.07 (0.06, 0.09)*	*0.01 (0.00, 0.05)*	*0.39 (0.35, 0.44)*
5	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.61 (0.51, 0.76)
6	0.03 (0.00, 0.05)	0.04 (0.00, 0.19)	0.00 (0.00, 0.00)	1.92 (1.90, 1.95)
damage excluded	*0.02 (0.00, 0.05)*	*0.01 (0.00, 0.04)*	*0.01 (0.00, 0.04)*	*0.27 (0.33, 0.42)*
Pooled	0.02 (0.00, 0.05)	0.01 (0.00, 0.06)	0.00 (0.00, 0.00)	0.87 (0.86, 0.88)
damage excluded	*0.00 (0.00, 0.11)*	*0.00 (0.00, 0.03)*	*0.00 (0.00, 0.02)*	*0.48 (0.46, 0.52)*

Table 2: LEA signature R&R model estimates for summary quantities, $\sigma_{repeatability}$ and $\sigma_{reproducibility}$.

Land	$\sigma_{repeatability}$	$\sigma_{reproducibility}$
1	0.53	0.18
damage excluded	*0.32*	*0.16*
2	0.55	0.14
3	0.42	0.16
4	0.59	0.23
damage excluded	*0.39*	*0.15*
5	0.61	0.28
6	1.92	0.43
damage excluded	*0.37*	*0.08*
Pooled	0.87	0.245
damage excluded	*0.48*	*0.18*

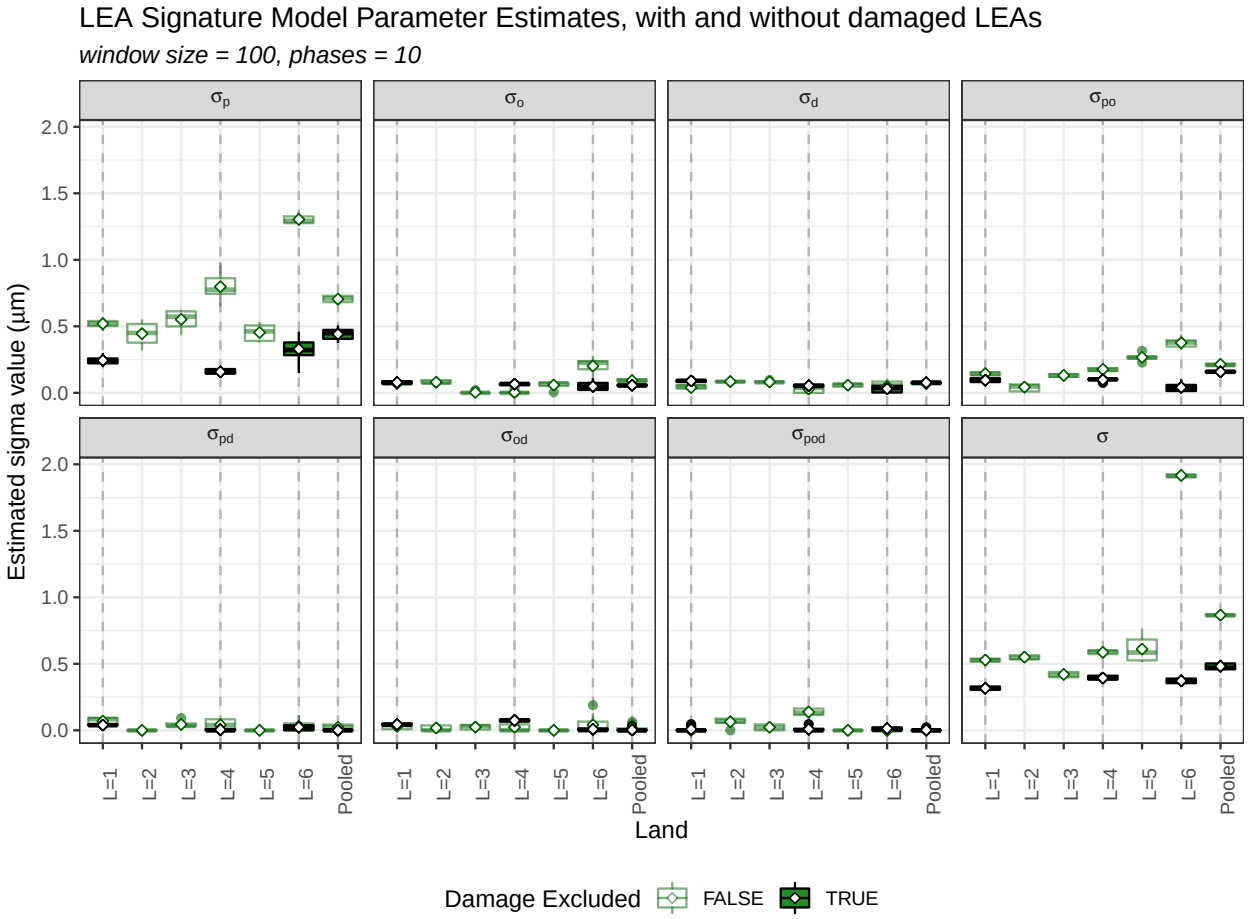

Figure 5: Distribution of variance component estimates resulting from ten phased models applied individually to each of six lands and pooled signature data. Results presented here were modeled using Equation 5. Estimates emphasized in dark green represent estimates after removing damaged LEAs from consideration.

Table 3: Pairwise score R&R model estimates for variance components, reported with 95% bootstrap confidence intervals.

Model	$\sigma_{(p)}$	$\sigma_{(o)}$	$\sigma_{(d)}$	$\sigma_{(po)}$
Same-Source	0.212 (0.154, 0.272)	0.027 (0.022, 0.032)	0.020 (0.011, 0.027)	0.050 (0.047, 0.052)
damage excluded	*0.067 (0.045, 0.088)*	*0.030 (0.025, 0.036)*	*0.023 (0.013, 0.034)*	*0.042 (0.039, 0.045)*
Different-Source	0.035 (0.029, 0.040)	0.011 (0.010, 0.012)	0.003 (0.000, 0.005)	0.016 (0.015, 0.016)
damage excluded	*0.034 (0.027, 0.041)*	*0.010 (0.008, 0.011)*	*0.000 (0.000, 0.003)*	*0.015 (0.014, 0.016)*
All Pairings	0.117 (0.101, 0.133)	0.017 (0.016, 0.019)	0.011 (0.008, 0.014)	0.029 (0.028, 0.030)
damage excluded	*0.045 (0.038, 0.054)*	*0.018 (0.016, 0.019)*	*0.011 (0.009, 0.015)*	*0.025 (0.025, 0.026)*

Model	$\sigma_{(pd)}$	$\sigma_{(od)}$	$\sigma_{(pod)}$	σ
Same Source	0.013 (0.011, 0.016)	0.022 (0.020, 0.024)	0.003 (0.000, 0.007)	0.159 (0.159, 0.160)
damage excluded	*0.009 (0.006, 0.012)*	*0.029 (0.027, 0.032)*	*0.000 (0.000, 0.007)*	*0.155 (0.154, 0.156)*
Different-Source	0.008 (0.007, 0.009)	0.007 (0.007, 0.008)	0.003 (0.002, 0.004)	0.104 (0.104, 0.104)
damage excluded	*0.008 (0.006, 0.009)*	*0.006 (0.005, 0.006)*	*0.005 (0.005, 0.006)*	*0.087 (0.087, 0.087)*
All Pairings	0.010 (0.009, 0.011)	0.013 (0.012, 0.013)	0.004 (0.003, 0.005)	0.115 (0.115, 0.115)
damage excluded	*0.008 (0.007, 0.009)*	*0.015 (0.014, 0.016)*	*0.003 (0.000, 0.004)*	*0.103 (0.103, 0.103)*

Table 4: Pairwise score R&R model estimates for summary quantities, $\sigma_{repeatability}$ and $\sigma_{reproducibility}$.

Model	$\sigma_{repeatability}$	$\sigma_{reproducibility}$
Same-Source	0.159	0.065
damage excluded	*0.155*	*0.064*
Different-Source	0.104	0.022
damage excluded	*0.087*	*0.021*
All Pairings	0.115	0.039
damage excluded	*0.103*	*0.037*

4.2 Pairwise Similarity Scores

Three models were fit using different sets of data. Similarity scores for all same-source pairs were included in one model, with a second model fit for scores from all different-source pairs. A model which pooled all scores regardless of origin was also fit. Results from these three sets of models are reported in Table 3.

We observe the largest variance components associated with residual error and part (bullet), with other variance components having much smaller magnitude. We also observe that after excluding pairs that include a damaged LEA, the magnitude of our estimated variance components decreases. Summary values, $\sigma_{repeatability}$ and $\sigma_{reproducibility}$, are also reported in Table 4. The magnitude of the summary values lessen when damaged LEAs are removed, but not as drastically as the reduction seen in the signature modeling results.

5 Conclusions

Our approach to adapting the Gauge R&R framework to two complex data structures was focused on preserving the model assumptions and structure of traditional Gauge R&R mixed-effects models. Adapting

data science problems which leverage complex data structures to fit the traditional Gauge R&R framework has several advantages, the first of which is the preservation of the data format and units. Interpretability of estimated variance components is immediate and does not require any additional transformation or maneuvering.

In addition, the framework provides us with a way to estimate quantities relevant to the measurement process we are working inside of. Summary values such as $\sigma_{repeatability}$ and $\sigma_{reproducibility}$ provide insight about the overall reproducibility of the measurement process, while variance components for study factors provide detailed insight on sources of variability. In our study, the variance components associated with device and operator were much smaller than those associated with differences across parts and residual error, which provides assurance that a minimal amount of measurement variability is due to procedural or mechanical differences across operators and devices. We are able to immediately gain this insight after modeing because of the R&R framework.

Another advantage to the Gauge R&R framework is that it provides actionable items within a measurement process. Our results indicated that operator variability was larger for lands with one or more unsuitable LEAs, and was greatly reduced after removal of the offending LEAs from modeling. This insight about the measurement process informed a larger focus on unsuitable LEAs and damaged bullets in training materials and Standard Operating Procedures for microscope operators.

In addition, we have also established what levels of variability are typically observed when repeated scans are gathered by trained operators in our measurement process. These standards can be used to assess scanning consistency for trainees as well as identify LEAs with damage or low scan quality.

Our approach to signature modeling emphasized subsampling to remove dependence and comply with Gauge R&R model independence assumptions. Taking this approach with similar data structures requires considering window sizes and carefully balancing data reduction with reducing dependency. Future work should include investigating the possibility of incorporating dependency into the model, perhaps by considering a small slice of signature data, such as twenty consecutive x_{Li} locations. We may still be slightly underestimating variability to due small positive correlation remaining present in the data.

We focused here on a specific application to a forensic bullet matching data pipeline; however, similar approaches can be applied to other data science pipelines. These adaptations are most directly applicable to structured response data or pairwise comparisons applications. Paired data are a commonly considered structure in data science, and this approach could be very advantageous to analyzing sources of variability in a variety of paired comparison frameworks.

The study design and results presented here resulted in actionable items for the automated firearms analysis measurement process. It is our hope that similar approaches with analogous or new data structures can yield researchers in other application fields similar results and insights.

Bibliography

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

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

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

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

James E. Hamby, David J. Brundage, and James W. Thorpe. The identification of bullets fired from 10 consecutively rifled 9mm ruger pistol barrels: A research project involving 507 participants from 20 countries. *AFTE Journal*, 41(2):99–110, 2009.

Eric Hare, Heike Hofmann, and Alicia Carriquiry. Automatic matching of bullet land impressions. *The Annals of Applied Statistics*, 11:2332–2356, 12 2017.

DOUGLAS C. MONTGOMERY and GEORGE C. RUNGER. Gauge capability analysis and designed experiments. part ii: Experimental design models and variance component estimation. *Quality Engineering*, 6(2):289–305, 1993. doi: 10.1080/08982119308918725. URL https://doi.org/10.1080/08982119308918725.

Rogerio Santana Peruchi, Pedro Paulo Balestrassi, Anderson Paulo de Paiva, Joao Roberto Ferreira, and Michele de Santana Carmelossi. A new multivariate gage r&r method for correlated characteristics. *International Journal of Production Economics*, 144(1):301 – 315, 2013. ISSN 0925-5273. doi: https://doi.org/10.1016/j.ijpe.2013.02.018. URL http://www.sciencedirect.com/science/article/pii/S0925527313000856.

Shannon Sweeney. Analysis of two-dimensional gage repeatability and reproducibility. *Quality Engineering*, 19(1):29–37, 2007. doi: 10.1080/08982110601057641. URL https://doi.org/10.1080/08982110601057641.

Stephen B. Vardeman and Enid S. VanValkenburg. Two-way random-effects analyses and gauge r&r studies. *Technometrics*, 41(3):202–211, 1999. ISSN 00401706. URL http://www.jstor.org/stable/1270565.