Document Title:	**Benefits and Limitations of the Statistical Methodology in Determining Evidentiary Value of Latent Print Identification: A Prospective Evaluation**
Author(s):	**Denver Police Department Crime Laboratory**
Document Number:	**255944**
Date Received:	**December 2020**
Award Number:	**2015-DN-BX-K027**

Benefits and Limitations of Statistical Methodology in Determining Evidentiary Value of Latent Print Identifications: A Prospective Evaluation
Final Summary Overview
Award - 2015-DN-BX-K027

Project Purpose:

The purpose of this project is to study the benefits and limitations of using objective statistical models to determine the sufficiency of a latent print at the analysis stage. We also examine the benefits and limitations of using statistical models for calculating likelihood ratios (LR) of close non-matches (CNM) and verified latent print identifications at the evaluation stage.

Project Design and Methods:

During this project, statistical information resulting from the model was compared to conclusions made by examiners on latent prints from current casework. This project relied on a large dataset of casework latent prints maintained and examined by the Denver Police Department Crime Laboratory (DPDCL) containing latent prints collected as evidence within the City and County of Denver. Analytical information such as the presence of pattern type, ridge characteristics, movement distortion effects, and substrate interference was documented. This produced a research dataset of latent prints with differing levels of quality observed in actual casework that expands on previously conducted research in controlled conditions.

The DPDCL partnered with Dr. Cedric Neumann of South Dakota State University and Two N's Forensics, Inc. The model used during this research project was developed under NIJ Award 2014-IJ-CX-K088. A scientific paper describing the model has been submitted to the Journal of the Royal Statistical Society: Series C and is currently undergoing revisions. The submitted version of the paper is freely available as a technical report [Hendricks et al. 2018].

To calculate the weight of evidence of a given fingerprint comparison, the model uses information relative to the ridge events (i.e., minutiae) that are believed (by a latent print examiner) to correspond between the latent and the known prints. The model compares the relative locations (expressed in x, y-coordinates and measured in pixels) and directions (measured in radians) of the paired minutiae between the latent and the known prints.

Due to the lack of user-friendly interface with the model, the calculation of the weight of evidence of the fingerprint data collected is relatively involved. Recent literature provided guidance on how LRs can be used to estimate the suitability of latent print evidence at the analysis stage and the weight, or evidentiary value of the evidence, at the evaluation stage. Results can be supported by a numerical value of the LR. The current model combined these two measures and enabled (a) the prediction of whether any given latent print should be further examined based on the quality and quantity of minutiae observed during the analysis stage; and (b) the determination of the probative value of the latent print based on the quality and quantity of information in agreement and disagreement observed during the comparison stage. In the latter instance, the probative value was calculated using the alternative hypothesis (a) that an individual is not the origin of the latent print but is a CNM resulting from an AFIS submission, and (b) that the individual is not the true source (TS) but has been randomly selected from the general population. These are the two most common types of defense arguments in court and the model provides a calculation for both.

To determine if proposed statistical information can aid a latent print examiner, the DPDCL examined approximately 611 latent prints from casework in a prospective manner. Examiners encoded each latent print in AFIS using a combination of manual and automated encoding. Upon review of the AFIS candidate list, examiners determined the presence or absence of an identification to any of the candidates. Based on the resulting identification of the latent print submission to AFIS, one of two procedures were followed. If a candidate was identified as the source of the latent, the examiner recorded the mated minutiae as assigned by AFIS between the latent and the known print. The examiner then evaluated the best four CNMs in the candidate list. The mated minutiae were recorded as assigned by AFIS between the latent print and the four CNMs. This information was documented in the ULW software for determination of the weight of evidence of the “identified source” comparisons and for the CNM comparisons. Secondly, if no identification was made from the AFIS submission, no additional data collection was completed for the latent print. These data sets were submitted to Dr. Neumann for statistical calculations. Examiners documented the candidate ranking of the “identified” known print and the four CNMs as noted in the candidate list. The data submitted to Dr. Neumann consisted strictly of the mated minutiae intentionally excluding all potential bias information as to which of the five data sets were identified to the latent print as recorded by the examiner. These data sets were submitted for each latent in a random order to ensure that Dr. Neumann was not aware of the original candidate ranking as returned from AFIS. Dr. Neumann assigned a Bayes factor statistic to each latent/known pair. He used these data to measure the likelihood of finding CNM based on the minutiae previously marked by the examiner. He then assessed the ability of the examiner to discriminate these CNMs and relate that to the statistical weights derived by the statistical model. The collected data enabled the research team to assess whether an LR threshold can be defined to objectively discriminate between identifications and CNMs, whether examiners are making identifications that would be better reported as corroborative evidence, and whether examiners are accurately assigning conclusions. Overall, the data obtained during this project can be used to propose guidelines on the operational requirements, benefits, and limitations of the use of statistical models in casework. In addition, the comparison of LRs calculated for identifications and CNMs will provide information on whether such information should be used as part of the scientific validation of LR models.

Bayes factor likelihood ratio results

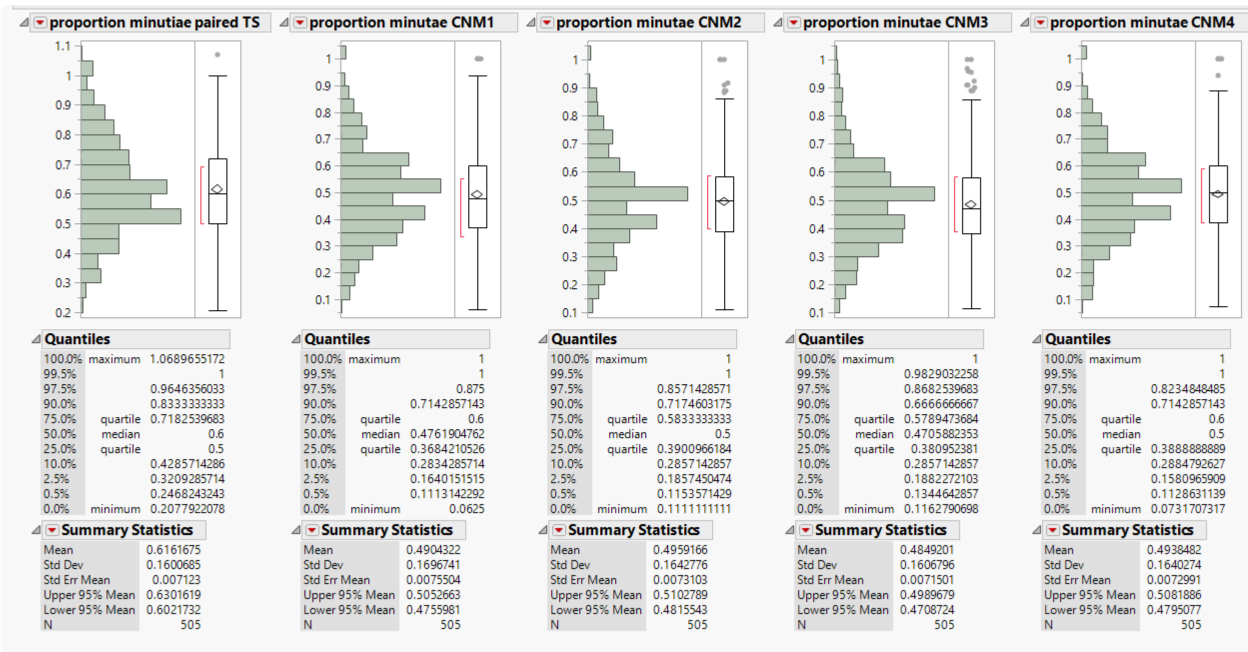

Figure 1. Proportion of latent print minutiae considered for each candidate by examiners.

During the project, we analyzed each latent print and documented the number of unique minutiae. Figure 1 summarizes our findings when comparing the proportion of minutiae used for comparison in each pair to the unique minutiae marked during the original latent analysis as a Bayes factor LR-rescaled by taking the $\log_{10}$ of each value. Each column of data in Figure 1 represents a summary of each of the five pairs for each data sample set (one TS, four CNMs). On average, examiners utilized 62% of originally analyzed unique minutiae as compared to each CNM which averaged 49%, 49%, 48%, and 49% respectively. Each difference was statistically significant ($p<0.01$) indicating that for those candidates chosen by examiners as the TS, about 62% of minutiae were used, which is significantly more than used for any of the CNM candidates.

Evaluation of Bayes factor likelihood ratio as a function of AFIS rank

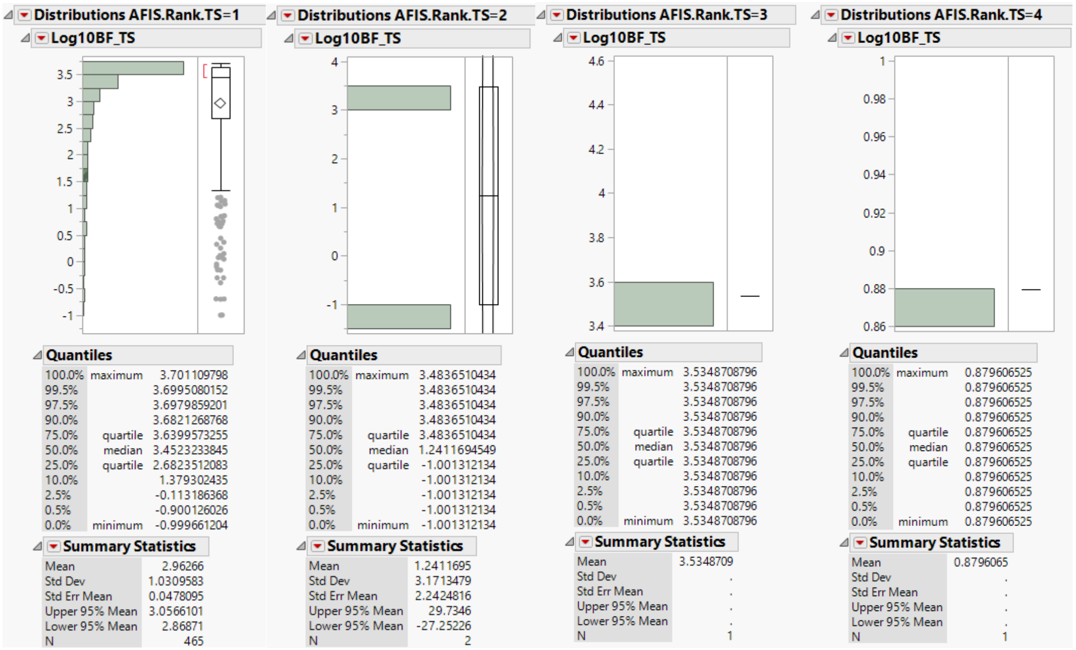

Figure 2. Log$_{10}$ Bayes factor likelihood ratio distribution for each level of AFIS rank.

An evaluation of the efficacy of the Bayes factor score was conducted by examining the distribution of scores after rescaling by taking the log$_{10}$ of each score and ranking by AFIS rank. It is clear from these results that the Bayes factor LR for the prints scored as TS (left column in Figure 2) follow a left skewed distribution with a mean of 2.96, median 3.45. Contrast this against any other AFIS ranks where the scores become either bimodal or unimodal and are significantly lower. These results indicate that the Bayes factor LR score is somewhat predictive of the top AFIS ranked print that was chosen as the TS, however, the range in results do indicate that low Bayes LR scores were identified for pairs of latent and known prints determined as true sources. As a population, the approach seems predictive, but there are TS pairs that would be eliminated if reliance was only on a threshold Bayes LR chosen as either the mean or median.

Bayes factor likelihood ratio and number of paired minutiae

We examined the relationship between the Bayes factor LR and the number of paired minutiae as a test between an independent evaluation of the latent print paired to the TS and the number of paired minutiae used in the determination by the latent print examiners.

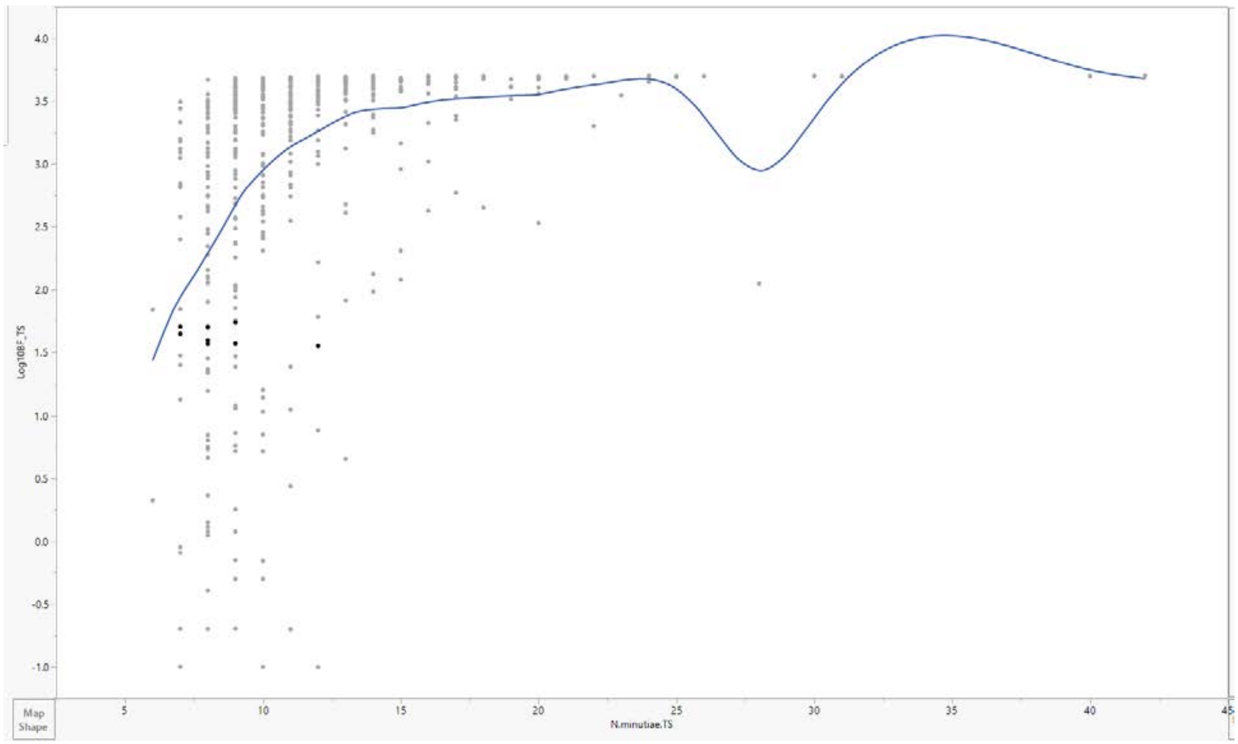

Figure 3. Scatterplot with best fit curve for all true source print pairs.

A further examination of the Bayes factor LR is presented in Figure 3. As the number of paired minutiae increase, the $\log_{10}$ Bayes factor LR also increases. This indicates a strong relationship between the confidence each examiner had in the TS evaluation as mirrored by the Bayes factor LR calculated independently. These data suggest that the independent evaluation of the strength of the latent print pair (for those pairs deemed TS) by the Bayes factor LR is reflective of examiner conclusions. There do exist some lower Bayes factor LR scores, however, most of the data corresponds well with examiner conclusions.

Levels of Bayes factor likelihood ratios-true source versus closest non-matching prints

We evaluated the distribution of the difference in the Bayes factor LR for TS pairs versus the CNM pairs to measure if the LR was always larger for the pairs determined to be TS versus those of the highest ranked CNM.

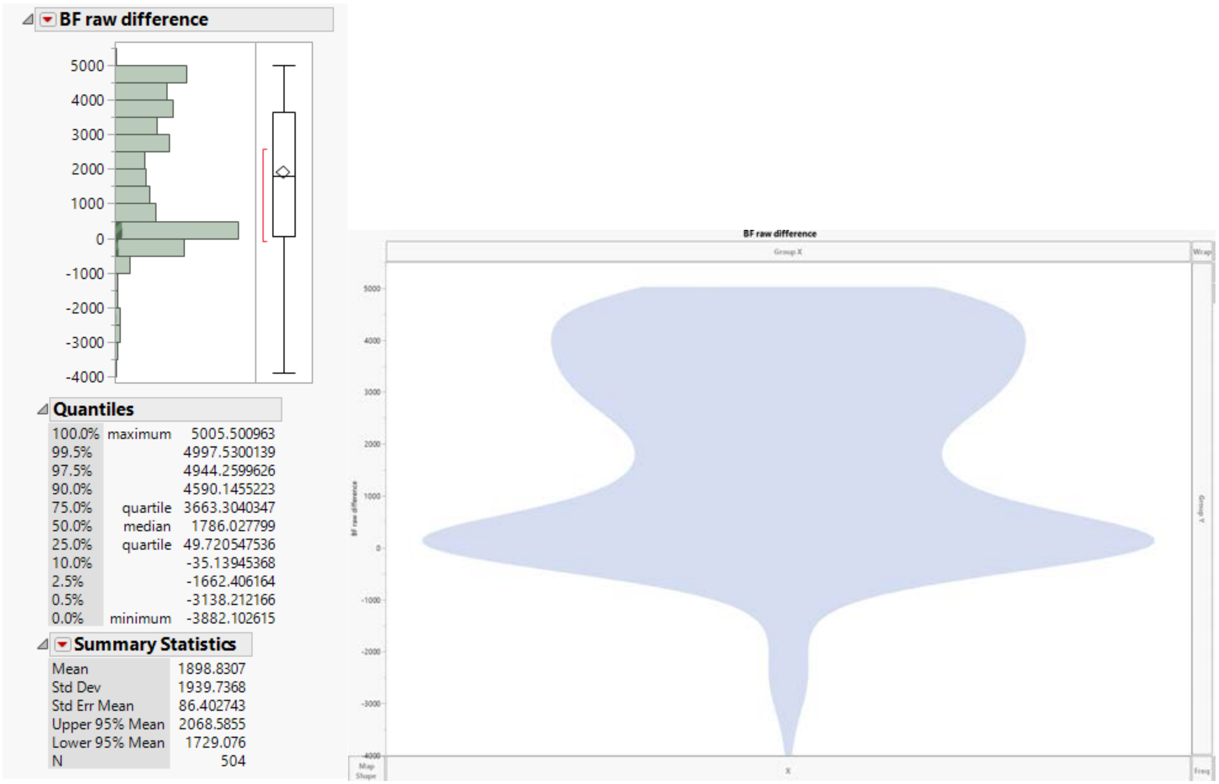

Figure 4. Difference between 'true source' and 'close non-match' Bayes factor LR raw scores.

In most cases, the Bayes factor LR was larger in magnitude (84%) for TS versus the highest ranked CNM, further indicating the utility of LR ranking of latent prints independent of examiner results. In some cases, scores were larger for the CNM pair indicating a difference in examiner conclusions and those of the scoring algorithm. One possibility for this is that examiners make determinations based on similarities and differences in the pair being compared and in this experimental design, only similar minutiae were evaluated by Dr. Neumann's software model. Those differences could be significant if included in statistical LR evaluations.

Testing a threshold in Bayes factor across the study

We examined a series of threshold values of the $\log_{10}$ Bayes factor in terms of performance against known 'gold standard' outcomes of latent print TS and the four CNM prints.

Threshold	True Positives	False Positives	False Negatives	Positive Predictive Value	Specificity
0.43	448	451	22	50%	39.37%
1.65	411	274	59	60%	62.97%
2.63	358	154	112	70%	79.16%
3.21	292	73	178	80%	90.13%
3.54	189	21	278	90%	97.15%

Table 1. Thresholds in $\log_{10}$ Bayes factor and test performance

These data indicate that over our experimental cases, that a $\log_{10}$ Bayes factor of about 2.63 (424) balances the test. Higher thresholds have more predictive value and specificity but that comes at the cost of a large increase in false negatives if judged by the Bayes factor alone, indicating that the Bayes factor is an interesting way to express latent print results but not a replacement technology at present. We intend to continue working on refining the model to include 'difference' data in addition to pairs of minutiae and determine the impact on the predictive performance of the Bayes factor and statistical model.

Implications for Criminal Justice Policy and Practice in the United States:

It would be inappropriate to surmise whether LR information would be considered beneficial to criminal justice policy without additional research and consultation with the court system regarding the findings of this project. This additional research would assist in determining how critical the court deems LRs for latent print examiner testimony versus current testimony practices given the significant increase in time spent per case. The development of our project design and methods indicated the necessity for the development of laboratory Standard Operating Procedures as a critical component of casework application. Specifically, we encountered various processes that needed adjustment during the data collection phase to ensure the specifications for the software model were met. We believe additional research is required to expand upon the capabilities of available LR software, particularly the inclusion of differences between a latent print and a known print to give a more accurate LR. If this additional research is successful, we are confident that latent print examiners would be able to report identifications, and possibly inconclusive decisions, with a LR. The data that has been collected during this project can be used to test any updated or new LR software by providing casework data. We plan to publish our project findings in three peer reviewed forensic journals: The Journal of Forensic Sciences, Forensic Science International, and The Journal of Forensic Identification to reach a broad multi-disciplinary audience. We also plan to present our project findings at two forensic educational conferences: the American Academy of Forensic Sciences and the International Association for Identification.

References

Sorkine-Hornung, O., Rabinovich, M. (2017) Least-Squares Rigid Motion Using SVD. Department of Computer Science, ETH Zurich, Switzerland. https://igl.ethz.ch/projects/ARAP/svd_rot.pdf Last Verified October 25th 2018

Hendricks, J., Neumann, C., Saunders, C.P. (2018) Quantifying the weight of fingerprint evidence using an ROC-based Approximate Bayesian Computation algorithm. Technical Report. https://arxiv.org/abs/1803.10121 Last Verified 23 March 2020