

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title: Improving the Understanding and the Reliability of the Concept of “Sufficiency” in Friction Ridge Examination

Author(s): Cedric Neumann, Christophe Champod, Mina Yoo, Thibault Genessay, Glenn Langenburg

Document No.: 244231

Date Received: December 2013

Award Number: 2010-DN-BX-K267

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this Federally-funded grant report available electronically.

<p>Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.</p>

National Institute of Justice – Office of Justice Program
**Fundamental Research to Improve Understanding of the Accuracy,
Reliability, and Measurement Validity of Forensic Science Disciplines**

Solicitation # SL000909

Award 2010-DN-BX-K267

**Improving the Understanding and the Reliability of the Concept of
"Sufficiency" in Friction ridge Examination**

**By Cedric Neumann, Christophe Champod, Mina Yoo,
Thibault Genessay, Glenn Langenburg**

Revised Final report submitted July 12th 2013

From:

Dr. Cedric Neumann
Statistics Department
107 Whitmore Laboratory
The Pennsylvania State University
University Park PA 16802
Email: cedric.neumann@me.com

To:

Gerry LaPorte
National Institute of Justice
Office of Investigative and Forensic Sciences
810 Seventh Street, N.W.
Washington, DC 20531
Phone: (202) 305 1106
Email: gerald.laporte@usdoj.gov

1. Executive Summary

This document reports on a 2 year research project sponsored by the National Institute of Justice of the Department of Justice of the United States of America under contract 2010-DN-BX-K267. The aim of the project was to study the concept of *sufficiency* associated with the decisions made by latent print examiners at the end of the various phases of the examination process. During this 2 years effort, a web-based interface was designed to capture the observations made by 146 latent print examiners and latent print trainees on a set of 15 pairs of latent/control prints. The variables of interest ranged from demographics data on the participants through to the type of features, their quality, their level of agreement between the latent and control prints, and their decisions at the end of each phase of the examination process. A statistical model was also developed to quantify the specificity of the configurations of minutiae annotated by the participants on the prints. Random Forest classifiers were used to measure the importance of the different variables on the decisions made by the participants. Random Forest classifiers were used as rational proxies of the decision-making process of human examiners based on the observations of the latent/control prints.

Two main findings resulted from our study:

- 1) The concept of *sufficiency* is mainly driven by the number and spatial relationships between the minutiae observed on the latent and control prints. Our data indicate that demographics (training, certification, years of experience) or non-minutiae based features (such as level 3 features) do not play a major role in the making of decisions by examiners;
- 2) Our results show a significant variability between the detection and interpretation of friction ridge features. This has been observed at all levels of details, as well as for factors potentially influencing the examination process, such as degradation, distortion, or influence of the background and the development technique. There is an urgent need for development of standards and training to ensure consistency in the definition, selection, interpretation and use of observations made on friction ridge impressions.

2. Table of Contents

1. Executive Summary.....	2
2. Table of Contents	3
3. Introduction	4
4. Purpose, objectives and general design of the project	8
5. Material and methods	9
5.1. Trial images	9
5.2. Examiners contacted and initial survey.....	9
5.3. PiAnoS4 platform.....	11
6. Variables extracted from PiAnoS	17
6.1. Quality consensus and its divergence.....	18
6.2. Minutiae consensus and it divergence	19
7. Development of a statistical model for the quantification of sufficiency in latent print examination.....	22
7.1. Model.....	24
7.2. Method for quantitative observations on fingerprints.....	27
7.3. Shape element of the model	30
7.4. Direction element of the model	32
7.5. Type element of the model	34
7.6. Datasets.....	36
7.7. Model performances.....	37
8. Descriptive statistics of the 15 trials.....	50
8.1. Descriptive statistics related to the examiners.....	50
8.2. Comfort and coherence levels of the participants with the interface.....	51
8.3. Descriptive statistics of trials results	52
8.4. Descriptive statistics of the weight of evidence for the trial results.....	60
9. Relationships between participants' annotations and sufficiency	69
9.1. Sufficiency in relation to the Analysis phase.....	71
9.2. Sufficiency in relation to the Comparison phase	75
9.3. Analysis of the annotations made in two cases	79
Trial 08 (same source)	79
Trial 12 (different sources).....	81
10. Implications of the main findings for practice and conclusion.....	84
10.1. Concept of sufficiency	84
10.2. Consistency in the definition, observation and use of friction ridge skin features ..	85
11. Bibliography.....	86
12. Appendix A - Trial images	89

3. Introduction

The skin of the digits (fingers and toes), palms and soles of human beings is formed of papillary ridges, also known as friction ridges. Fingerprint is commonly used as a generic term to describe the impression of a friction ridge skin area on a given surface.

Fingerprints have been used with considerable success over the past century to determine or verify the identity of individuals using finger impressions taken under controlled conditions, or from friction ridge impressions left inadvertently on crime scenes. In particular, latent print examiners are concerned with the determination of the identity of criminals through the examination of partial, potentially distorted and degraded friction ridge impressions recovered on crime scenes. These impressions will be designated in this report either as latent prints (to follow the practice in the US) or as marks (in line with the European terminology).

Recently, Daubert and Frye hearings have brought to light the need for improving the understanding of the accuracy and reliability of friction ridge examination. The recent review of the state of forensic science in the United States by the National Research Council of the National Academies [1] has also stressed the need to develop quantifiable measures for methods that are currently qualitative in nature, such as the examination of fingerprints (and other impressions): current protocols and procedures to perform these examinations heavily rely on a succession of subjective decisions, from the initial acceptance of evidence for probative value to the final assessment of forensic results.

The FBI/NIJ-sponsored Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) defines these subjective decisions by a generic term [2,3]:

Suitable (Sufficient): the determination that there is adequate quality and quantity of detail in an impression for further analysis, comparison or to reach a conclusion.

Currently, one general protocol is accepted as guiding fingerprint examination: ACE-V (analysis, comparison, evaluation and verification). Albeit this acronym is not always used, this protocol is the most commonly referred to by the different professional bodies [4,5], discussed in the relevant literature [6,7], and cited in US courts when examiners reported

fingerprint evidence [8-11].

The practical implementation of this protocol may vary between agencies. However, fingerprint professionals, and scientific and legal scholars, generally accept that it aims at minimizing the risk of errors and provides a measure of quality assurance. When following this protocol, examiners are requested to make decisions after each phase:

Analysis: The purpose of the analysis stage is to assess the usefulness of recovered latent prints. In order to avoid being influenced by the comparison exemplar prints (which are typically clearer and taken under controlled conditions from a known source), it is recognized in the literature that the analysis of latent prints needs to be carried out in isolation, without referring to the inked (or known) impression [3,6,7].

The assessment of the expected value or potential of the latent prints is based on the observation of the quantity and quality of the characteristics available on the latent prints, on the determination of whether distortion effects (or lack of clarity) are present, what is their impact on the reliability of the observed characteristics and what tolerance levels need to be set when subsequently comparing the print to a control print.

Ultimately, examiners need to decide whether a latent print bears *sufficient* quantitative and qualitative information for further comparison, or at least for exclusion purposes [3].

Three outcomes are generally made:

1. Value for Exclusion Only (VEO): the latent prints can be used to exclude or potentially associate an individual, but is insufficient to individualize;
2. Value for Identification (VID): if a corresponding control print from a known individual is provided, an individualization will be declared. Among these, some marks will be declared to be searchable in a fingerprint database (AFIS).
3. No Value (NV): The latent prints cannot be used further in the process; it is insufficient for comparison.

Depending on training, experience and several other factors, a significant variability between different examiners may be observed at this stage of the protocol. A given latent print may be deemed suitable for comparison by some examiners, while it may only be considered suitable for exclusion purposes by others, or not usable by others.

Comparison: During the comparison phase, examiners search control prints, for the characteristics observed in a latent print during the analysis phase. For each characteristic observed on the latent print, a decision is taken with respect to its presence on the control print. These decisions are made based on the features' location, type, orientation and spatial relationships with other features.

Some characteristics of the latent print may not be clearly defined on the control print and examiners need to weight clarity and distortion factors to make the requested decision of correspondence. The tolerance levels defined during the analysis phase and, thus, decisions of correspondence, are mostly based on examiners' training and experience, decisions may differ between examiners for a given feature and this may lead to different decisions being taken during the evaluation stage of the examination protocol [12].

Evaluation: The evaluation phase requires examiners to attribute a weight to the correspondences and differences found between the two impressions examined in the previous stages, in order to infer, or not, the commonality of source of the latent and control prints.

At present, fingerprint examiners are required to express their conclusion in one of the three following ways: the outcome of a fingerprint comparison can be an identification (the term individualization is also used here synonymously), an exclusion or the comparison is said to be inconclusive with respect to the source attribution of the latent print [3,13]:

1. An identification (ID) decision is formed when two impressions contain sufficient quality and quantity of friction ridge detail in agreement to declare that the impressions share a common source of friction ridge skin.
2. An exclusion (EXC) decision is reached when sufficient quality and quantity of friction ridge detail are not in agreement to the point that both impressions cannot be from the same source.
3. An inconclusive (INC) decision is made when there is no sufficient detail in agreement or disagreement to justify either of the two previous decisions.

Currently, no transparent system exists to assign weight to correspondences/differences between ridge friction features. The concept of *sufficiency* has no clear definition and does

not relate to any objectively measurable quantity. The assignment of the weights and the decision to identify, exclude or otherwise is therefore often described as an holistic informed judgment and may be subject to differences between examiners.

Verification: Finally, the verification phase consists in the repetition of the previous tasks by one or several other examiners to confirm the initial conclusion.

Without doubt, forensic fingerprint examination has an extremely low rate of misidentification [14] and has demonstrated a tremendous contribution to criminal investigations. Nevertheless, the inherent subjectivity and lack of transparency of the decision-making at each stage of the ACE-V process exposes it to constant challenges and criticisms [15,16].

4. Purpose, objectives and general design of the project

The purpose of this research project is to gather data informing on the robustness and transparency of fingerprint examination, and to identify areas of improvement for preventing divergent decisions between two examiners considering the same latent print. The objective of this project is also to provide the fingerprint community with a body of research, tools and data allowing examiners to better understand the concept of *sufficiency*, in order to define better protocols for expressing and supporting the conclusions of fingerprint examinations.

More specifically, the project has been designed to study the relationships between the observations made by examiners on pairs of latent/control prints and the decisions reached at the end of the different phases of the examination of those prints. A web-based system (called PiAnoS) has been used to capture the observations and the decisions made by a group of examiners on a set of paired latent/control prints (section 5). The observations were summarized using different types of variables, some derived directly from the web-based system (section 6), and some assigned by a statistical model quantifying the weight of fingerprint evidence (section 7). A statistical analysis was conducted to measure the respective importance of the different variables in the decision-making process (sections 8 and 9). Finally, a series of recommendations were derived from our findings (section 10).

5. Material and methods

The study of the concept of *sufficiency* requires the study of the boundaries of the decision thresholds (both at the end of the Analysis and Evaluation phases). Indeed, the study of examinations resulting in clear identifications, or clear exclusions conclusions would not be very informative. In addition, the expected variability in the decisions made by examiners at those decision thresholds requires the collection of data from a large sample of examiners. Therefore, the research team decided to select a limited number of challenging cases, and to gather data from the largest possible number of examiners.

5.1. Trial images

15 latent prints were selected to represent challenging cases, which would maximize the variability between the decisions made by examiners. Among these cases, 12 latent prints were presented with control prints originating from the same source, while 3 latent prints were presented with prints originating from different sources. For those 3 cases, the control prints were specifically selected to display friction ridge details as similar as possible as the ones observed in the latent prints. The control prints were selected using a regional fingerprint database available to one of the authors. Images of the latent and control prints were available at 1000 dpi. The images associated with these cases can be found in appendix A of this report.

5.2. Examiners contacted and initial survey

About 600 U.S. latent print examiners were contacted to participate to the study. Examiners conducting casework were targeted although participants who were currently training to become latent print examiners were also accepted. The list of examiners was built based on contacts established through agencies and organizations such as the IAI (International Association for Identification) and SWGFAST (Scientific Working Group on Friction Ridge Analysis, Study and Technology).

The nature and purpose of the study was disclosed in the following terms:

The aim of the study is to understand what does a latent print examiner consider to be "sufficient"? We are looking at sufficiency in the Analysis phase of ACE-V for determining "value" of a latent print. We are also exploring sufficiency during the Evaluation phase for the determination of "individualization" and "exclusion" decisions.

Participants could freely accept to be enrolled in the study. Since the study was conducted through a web-based platform (see section 5.3), it was possible to guarantee to the participants that they would remain anonymous; each examiner would receive a randomly generated user name and a password. The research team has no mechanism to associate the username with the individuals enrolled in the study.

Initial survey questions	Possible answers
Sex	1. Male 2. Female
Expert Status	1. Certified Latent Print Examiner (i.e. IAI certified, FBI certified, or other governmental certification) 2. Latent Print Examiner - trained to competency and actively working cases 3. Latent Print Examiner - trained to competency but no longer actively working cases (e.g. manager, crime scenes only, or other duties that do not require latent print case work) 4. Latent Print Trainee - currently in training and not responsible for reporting case results 5. Other, please explain:
Year of experience performing Latent Print examination (you may include your training period)	Integer
Approximately how many hours per week would you estimate that you spend analyzing and comparing latent prints	Integer
Approximately how many latent print cases per month would you estimate that you complete	"0-10" ; "11-20" ; "21-30" ; "31-40" ; "41-50" ; "51-60" ; "> 60"
Which approach is your laboratory using for the determination of suitability?	1. Approach #1 (commonly referred to as "of value for identification"): Only impressions of value for individualization are compared. If a latent print cannot be individualized when presented with the correct (corresponding) exemplars from the same source as the latent print, then the latent print is deemed "no value". Under this approach, when an "inconclusive" opinion is rendered, it means "I need additional exemplars to complete the comparison". 2. Approach #2 (commonly referred to as "of value for comparison"): Impressions of value for individualization (and possibly for exclusion value only) are considered. If a latent print bears some corresponding characteristics to a clear, known exemplar, but insufficient to individualize, I would report "inconclusive". Under this approach, when an "inconclusive" opinion is rendered, it may be for several reasons (e.g. quality or completeness of the exemplars, insufficient characteristics to individualize, unable to locate in the exemplars, etc.)
Does your SOP have defined criteria to determine whether a print is suitable for further examination?	1. Yes, clearly defined 2. Yes, but criteria not necessarily well defined 3. No
What is the most common type of case that you work on a daily basis ?	Free text
Do you also process evidence (exhibits) for latent prints ?	1. Yes, always 2. Yes, often 3. Yes, rarely 4. No
In your practice do you frequently use 3rd level details for identification ?	1. Yes, always 2. Yes, often 3. Yes, rarely 4. No

Table 1: Initial survey taken by each examiner.

Once enrolled, examiners could work at their own pace, pausing and resuming as needed over a couple of sessions. They were not required to complete all trials, but encouraged to do so by offering them a number of "goodies" in the form of a compilation of scientific papers, transcript from court hearings and training images.

During their first login on the platform, the examiners were asked a series of 10 demographic questions (Table 1), which allowed for gathering information related to their training, experience and work practices.

5.3. PiAnoS4 platform

A dedicated platform allowing for conducting the study was designed (Picture Annotation Software 4 – PiAnoS4). PiAnoS4 is a free software package, released under the GNU Affero GPL license. Documentation and downloads can be found on the PiAnoS website. [17].

The platform offers an environment that allows examiners to conduct each trial separating the Analysis from the Comparison phases. Dedicated tools are offered to conduct the documentation of the observations made on the prints during both phases. A full description and user manual of the software was distributed to each participant before conducting the trial [18]. Some of the key elements of this software are summarized below.

During the Analysis phase of each latent print, examiners were asked to (at a minimum):

1. Annotate their perception of the quality of the print using a quality tool with three levels of quality. The tool allows for annotating separately different regions of each print. Examiners were not requested to annotate areas of the print that do not have visible ridge detail (i.e. highly smudged, smear/drag marks, etc.). The three levels are presented in Table 2.
2. Annotate **all observed minutiae** using the minutiae tool. Minutiae can be assigned as *ridge ending* or *bifurcation* when their type and location are discernable on the latent print. When the type (but not location) is uncertain, a specific annotation, called *Type unknown*, should be used. By extension when the location (and de facto type) is unclear, a fourth type of minutiae, called *Position unknown*, should be used.

The choice of the visual marker was made in order to reflect the decreasing levels of certainty. The four markers are shown in Figure 1.




<p>Quality tool</p> <p>Standard ▾</p> <p> High</p> <p> Medium</p> <p> Low</p>	<p>An area is annotated of high quality if: Level 1 is distinct; Level 2 details are distinct; There are distinct Level 3 details.</p> <p>An area is annotated of medium quality if: Level 1 is distinct; Most of the Level 2 details are distinct; There are minimal distinct Level 3 details.</p> <p>An area is annotated of low quality if: Level 1 may not be distinct; Most of the Level 2 details are indistinct; There are no distinct Level 3 details. Low quality (RED) is used only when you can see ridges in the degraded areas of the latent print, but indistinct minutiae. It should not be used to indicate areas without any ridges (such as a drag mark of a finger)</p>
<p>Following SWGFAST [3]: Level 1 detail refers to the overall ridge flow. Level 2 detail refers to individual friction ridge paths, friction ridge events (e.g., bifurcations, ending ridges, dots, and continuous ridges) and their relative arrangements. Level 3 detail refers to ridge structures (edge shapes, and pores) and their relative arrangements. Creases, scars, warts, incipient ridges, and other features may be reflected in all three levels of details.</p>	

Table 2: Definition of the Standard three-level system used in PiAnoS4.

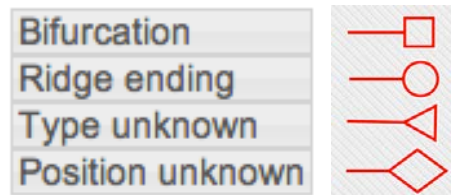


Figure 1: Illustration of the four types of minutiae (in order from type to bottom).

3. Provide some of their observations/decisions on the suitability of the latent print for further examination using four dialogue boxes (including a free text for additional note taking). The four inputs for concluding the Analysis phase are shown in Figure 2. Note that examiners were encouraged to report all adverse factors potentially affecting their examination. The possibilities for the conclusions on suitability depend on the choice of approach to suitability made during the survey (Table 1), as defined in Table 3

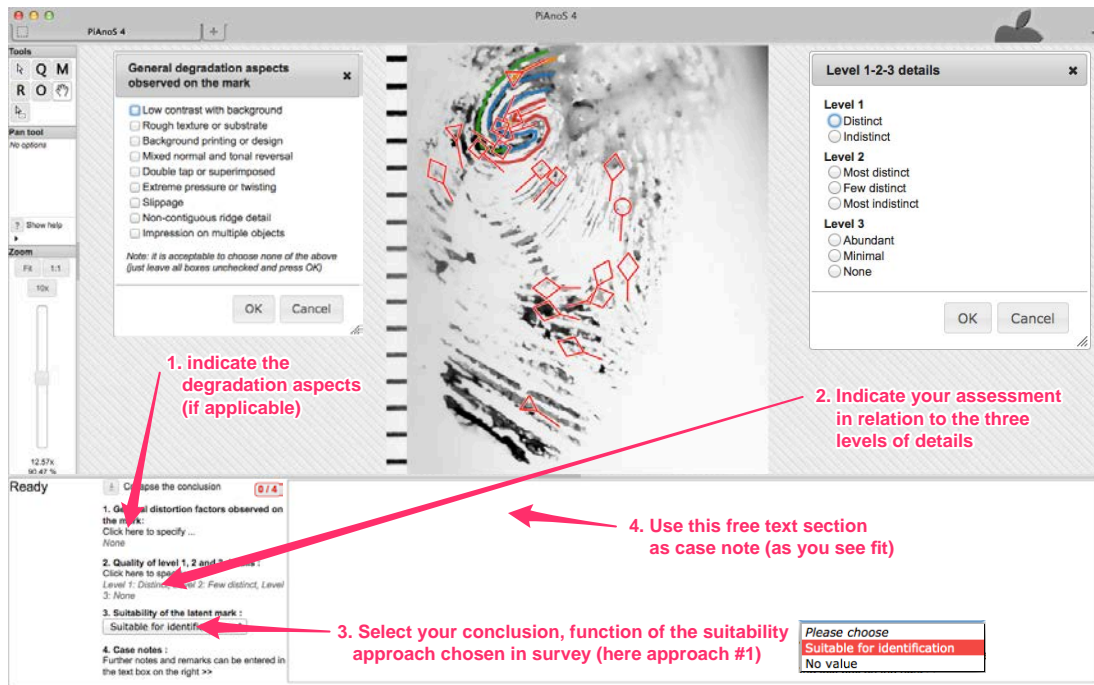


Figure 2: Four inputs involved in the conclusion section associated the Analysis phase.

Suitability approach	Possible conclusions following the Analysis phase
Approach #1	<p>1. Suitable/value for identification (VID) 2. No value (NV)</p> <p>Note: The term "suitable" indicates that the mark is or may be identifiable. Practice has shown that most examiners will mean "is", but it was felt important to recognized that the conclusion following analysis may be subject to revision. The label "No Value" to a mark refers only to its potential to be individualized. Marks allowing potential exclusion but failing the individualization threshold will be qualified as "No Value" in that approach #1. The term "identification" is used for all conclusions as meaning "individualization".</p>
Approach #2	<p>1. Suitable/ value for identification (VID) 2. Suitable only for exclusion (but not for identification) (VEO) 3. No value (NV)</p> <p>Note: The term "suitable for identification" indicates that the mark is or may be identifiable as before. The second option "suitable only for exclusion" indicates that the mark is not expected to be individualized but have sufficient features to allow an exclusion or an association of a strength that is less than an individualization. The term "No Value" is reserved to marks of quality that is insufficient either to associate or to exclude.</p>

Table 3: Possible conclusions following the Analysis phase depending on the approach adopted by the examiner¹

¹ Note that the website used "fingerprint" and "mark" instead of "latent prints"

During the Analysis phase, examiners could also trace ridges and annotate other features, such as scars, wrinkles or creases, using dedicated tools.

For the purpose of this study, examiners, who reached a conclusion of "No Value" at the end of the Analysis phase, were invited to process further with the Comparison phase. Examiners were only allowed to move to the Comparison phase once all observations and decisions for the Analysis phase were submitted to the system without possibilities of further modification.

During the Comparison phase, examiners were presented with the latent prints examined during the preceding Analysis phase and with paired control prints². The annotations made during the Analysis were displayed on the latent print as a starting point. Examiners were allowed to modify them³ as required using the same annotation tools as in the previous phase. During the Comparison phase, examiners were invited to annotate relevant corresponding and discordant minutiae according to the following guidelines:

1. Only the minutiae that were considered to be corresponding between the latent and the control print had to be annotated. This implied that (a) if a minutia was observed on the latent print, but was not available on the control print (due to a lack of clarity, or an area that is not available), the minutia on the latent print needed to be removed; (b) if a minutia was observed on the control print but had not been indicated on the latent print (e.g. missed) during the Analysis phase, the minutia on the latent print had to be annotated only if it could have reasonably been indicated during the Analysis phase;
2. All corresponding minutiae had to be annotated, even if the total amount of information was overwhelming and an examiner, in casework condition, would have stopped earlier.

² As explained in section 5.1, 3 of the 15 control prints originated from a different donor than the paired latent print.

³ Note that the annotations made during the Analysis phase are kept completely separated from the ones made during the Comparison phase. Any addition, modification, or subtraction of information occurring during the Comparison phase does not affect the observations collected during the Analysis phase.

3. Discordant minutia types between paired minutiae (e.g. viewed as a bifurcation on the latent print and ridge ending on the control print) had to be left unchanged, unless an obvious misjudgment had occurred;
4. Unexplainable differences had to be indicated using a specific type of minutia called "Difference".

Figure 3 illustrates the annotations of matching minutiae on a latent (left) and control print (right) originating from the same source. Note that in the context of the study, the indication of a perceived difference does not mean de facto that an exclusion conclusion will be reached. The purpose is to transparently indicate the observations made.

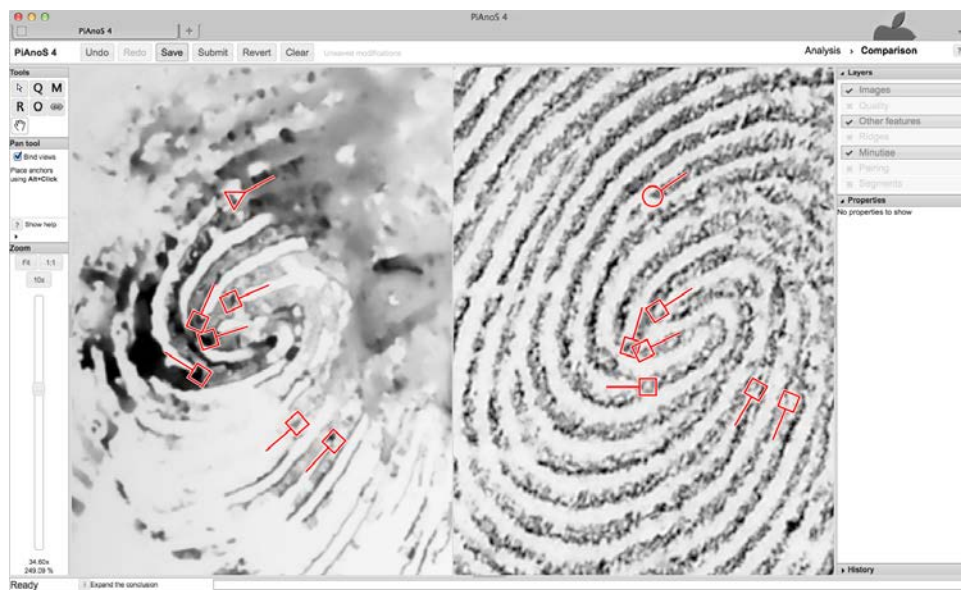


Figure 3: Annotation of the minutiae on the latent and control prints using the Minutiae tool (M).

5. All corresponding minutiae between latent and control prints had to be paired using a specific tool designed for that task (Figure 4).
6. The decisions of the examiners after the Comparison (and Evaluation) phase(s) had to be provided using the choice of options shown in Figure 5. When examiners reached an "inconclusive" decision, they were asked a few additional questions to help clarify their exact opinion on the source of the latent print (Figure 6).

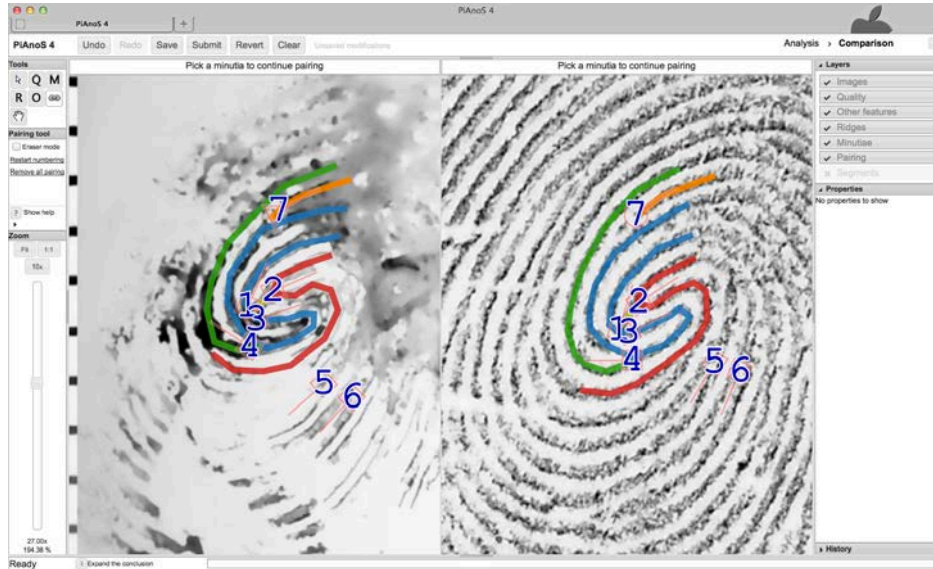


Figure 4: Pairing of all “matching” minutiae using the Pairing tool (P).

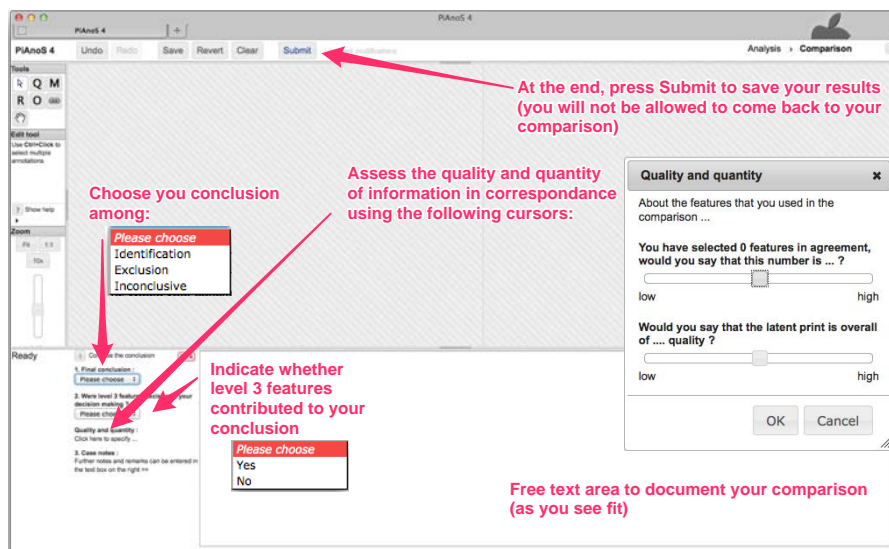


Figure 5: Conclusions options following the Comparison phase.

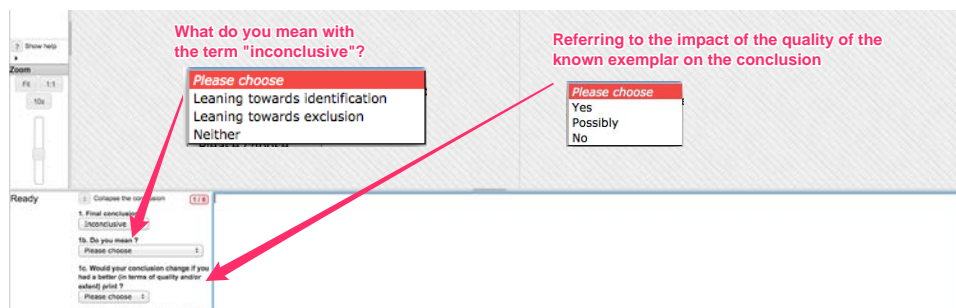


Figure 6: Additional questions put forward when the conclusion of the Comparison phase is “inconclusive”.

6. Variables extracted from PiAnoS

Different variables (Table 4) were extracted to describe and summarize the observations made by the examiners on the trial images. These variables can be automatically extracted from PiAnoS.

Variable code	Variables summarizing examiners' interactions with PiAnoS	Description of the metric extracted from PiAnoS
M1a	Number of minutiae outside quality zone	The number of minutiae annotated outside quality zones.
M1b1 / M1b2 / M1b3	Ratio of the number minutiae with declared type (RE or BIF) to the total number of minutiae in the quality zone	Sum of minutiae designated as ridge ending or bifurcation / total number of annotated minutiae. The metric is available for each quality zone separately: 1: green (high quality), 2: orange (medium quality) and 3: red (low quality)
Variable code	Variables summarizing the annotations during Analysis	Description of the metric extracted from PiAnoS
M1c1 / M1c2 / M1c3	Number of minutiae per area	Number of minutiae in a given quality zone / surface of the given quality zone (respectively 1, 2 and 3)
M4	Total number of minutiae annotated in Analysis	Total number of minutiae annotated in Analysis
M1d	Ratio of the number minutiae with declared type (RE or BIF) to the total number of minutiae	Sum of minutiae designated as ridge ending or bifurcation / total number of minutiae
M5b	Divergence from minutiae consensus	Distance between the user's minutiae map the the minutiae consensus map
M1e1/ M1e2 / M1e3	Relative proportion of the area of the quality zone	Surface of a given quality zone / total work surface (respectively 1, 2 and 3)
M2b75	Quality of the mark	$(a * (\# \text{ of green pixels}) + b * (\# \text{ of orange pixels})) / \text{total work surface}$ with $a > b$ Values used: $a = 1$, $b = 0.5$
qs2	Degradation aspects	Quality Score based on the degradation aspects indicated by the user. It counts the number of degradation factors ticked by the user, the higher the more complex the mark is from 0 to 6
M3b75	Divergence from quality consensus	Distance between the user's quality maps and the quality consensus map
Variable code	Variables summarizing the annotations during Comparison	Description of the metric extracted from PiAnoS
M6	Number of paired minutiae	Total number of paired minutiae annotated in comparison
Diff	Number of differences	Total number of differences (star) indicated

Table 4: Variables extracted from PiAnoS and summarizing the annotations provided by the examiners for each trial.

The first series of variables is designed to measure the level of comfort of the examiners with PiAnoS. The second series of variables is designed to capture the information provided by the examiners during the Analysis phase. The third series of variables is designed to summarize the information provided by the examiners during the Comparison phase. Most variables intend to capture the information provided by the examiners in absolute terms. However, two metrics (M5b, M3b75 in Table 4) were created to measure the differences between the annotations of any examiner and a consensus obtained from all examiners who

completed the trial. The first metric (M5b) captures the divergence from the consensus in terms of the quality of the latent print, while the second (M3b75) measures the divergence in terms of the annotated minutiae. These two metrics are described in more details in the next sections.

6.1. Quality consensus and its divergence

Section 5.3 presents the three levels that can be used to annotate the quality on the latent prints. For a given examiner, each pixel of a trial image can then take one of four values: green for high quality, orange for medium quality, red for low quality and N/A for pixels that were not annotated.

The quality annotations across all examiners for a given trial can be compiled (by superimposition) to reflect the variability in quality assessment at each pixel of the trial image, resulting in a pixel quality distribution (PQD) for any given pixel.

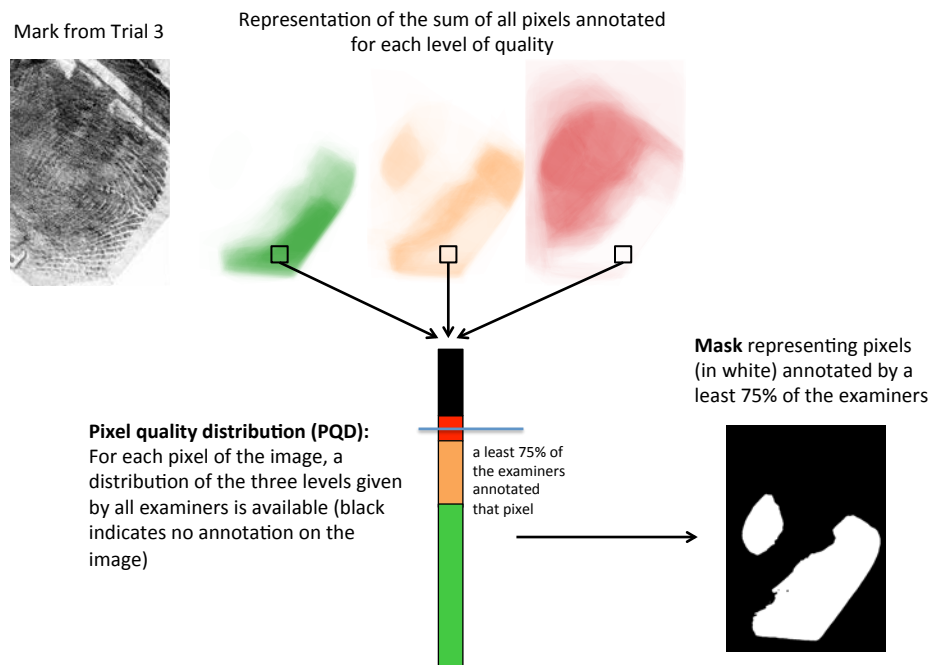


Figure 7: Derivation of the pixel quality distribution (PQD) and the Mask (here for trial 3) representing the pixels annotated by a least 75% of the examiners.

By selecting the pixels that were annotated by certain percentile (here we chose 75%) of the examiners (regardless of the level of quality), we derive a mask that can be used to

define the area of the image, which contains the majority of the relevant features for that trial. An example is shown in Figure 7.

The mask can be used to normalize the observations made within a given trial, but also across the different trials. In addition, the mask can be used to study the divergence of an examiner to the consensus of all examiners taking part in the study. Given an examiner's annotations, the divergence from the consensus is computed by comparing his/her assessment of the quality at each pixel of the trial image with the PQD for that pixel, as shown in Figure 8. The magnitude of the divergence is proportional to the weighted difference between the examiner's assessment and the PQD.

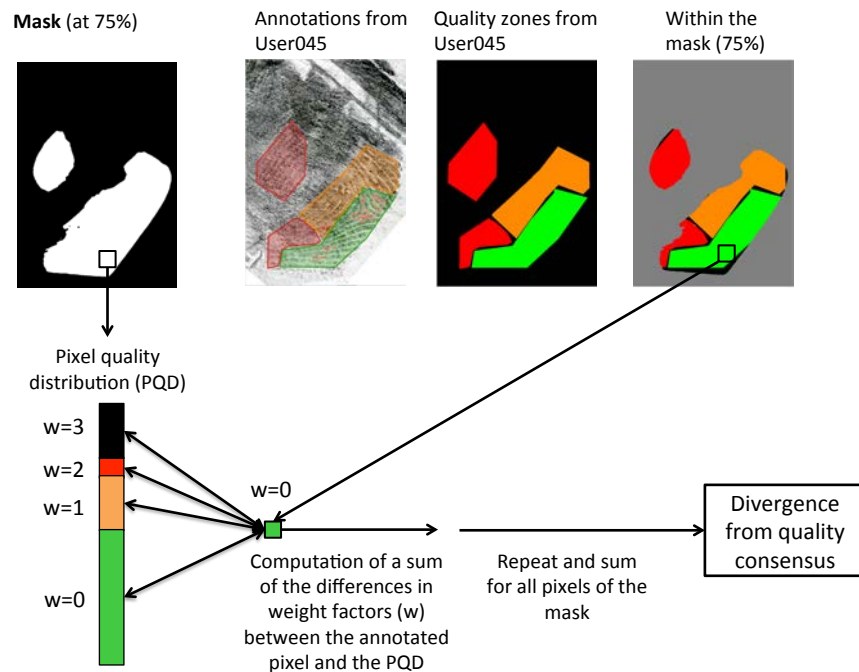


Figure 8: Computation of the divergence for a given examiner from the quality consensus of a trial (here for User045).

6.2. Minutiae consensus and its divergence

A similar approach is taken for the minutiae annotations. Each minutia can be represented by an ellipse, which orientation is dictated by its direction as indicated by the examiner and size is proportional to its type. Larger ellipses, representing larger uncertainty on the type or location of the minutiae are assigned as we progress from minutiae, which types are

declared (ridge ending and bifurcation), to minutiae, which types or positions are unknown (Figure 9).

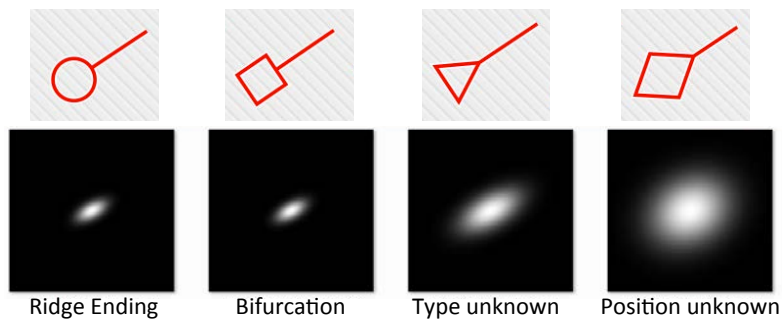


Figure 9: Elliptical representation of the minutiae. The orientation of an ellipse is defined by the orientation of the corresponding minutia and its size is proportional to the minutia type.

The minutiae annotations can be compiled across all examiners for a given trial to obtain its minutiae map. The intensity of each pixel of the minutiae map is a function of the number of time it falls within the boundary of an ellipse in the examiners' individual annotations. The combination of all examiners' annotations results in a pixel minutiae distribution (PMD) for any given pixel of the minutiae map (Figure 10).

The distance between each examiner's individual annotations of minutiae to the consensus is obtained by computing the weighted distance between his/her entries (given each type of minutiae) and the PMD as illustrated in Figure 11.

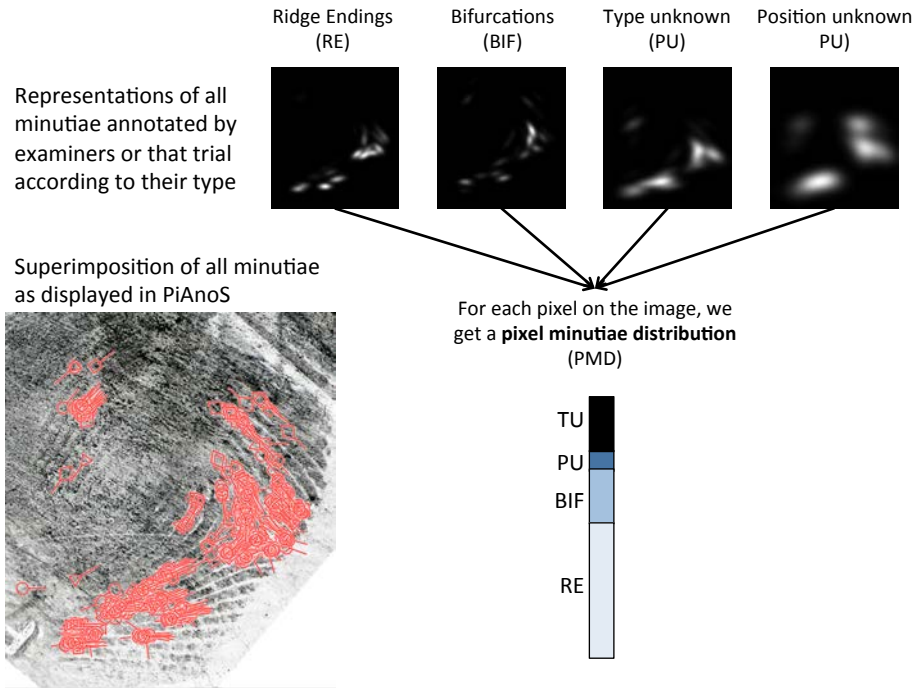


Figure 10: Derivation of the pixel minutiae distribution (PMD) (here for trial 3).

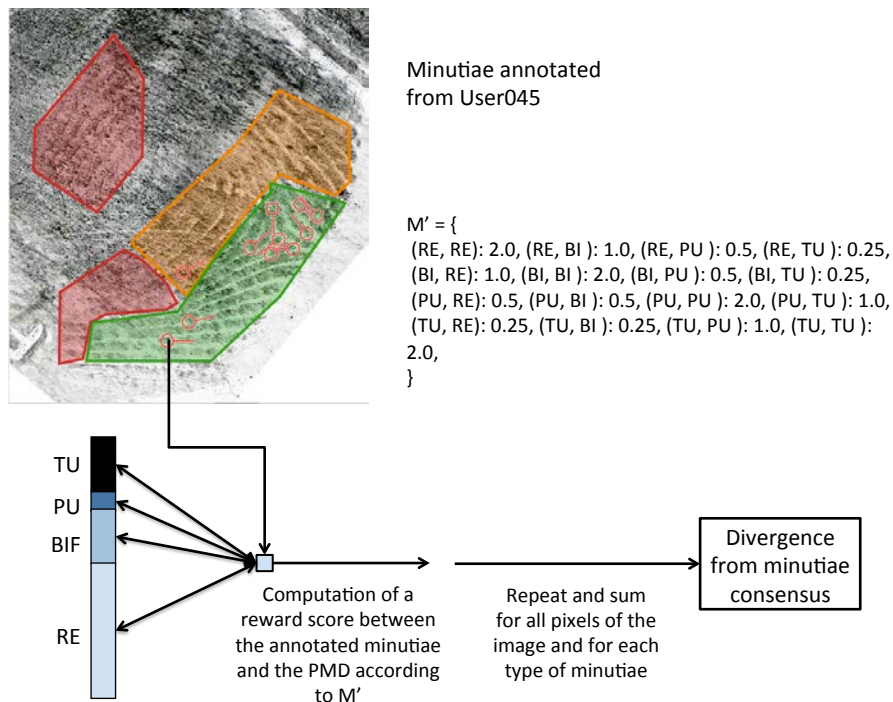


Figure 11: Computation of the divergence for a given examiner's minutiae annotation from the minutiae consensus of a trial (here for User0045).

7. Development of a statistical model for the quantification of sufficiency in latent print examination.

Section 3 outlines that the decisions/conclusions reached during the Analysis, Comparison and Evaluation phases are based on each examiner's personal training and experience. Several authors [see 19-20 for a review] have argued that these decisions should be supported by a probabilistic framework, and possibly by the use of a statistical model enabling the quantification of fingerprint evidence, in a similar fashion as this is done for DNA evidence. In addition, the fingerprint community has long claimed that the spatial relationships between friction ridge features were equally (if not more) important as their number when determining *sufficiency*. We chose to quantify this aspect of *sufficiency* by developing a fingerprint statistical model that would provide some measurement of the rarity of spatial configurations of fingerprint features.

Several models have been proposed during the past century to quantify the weight of fingerprint evidence and provide support for or objectivize the conclusions reached during fingerprint examinations. Models pre-dating 2001 have been reviewed by Stoney [21]. More recent models were reviewed in [20,22]. These models can be classified in two groups: (1) score based models and (2) so-called generative models.

Score-based models: Contrary to DNA, there is no easily definable and quantifiable set of features that can be used to characterize friction ridge skin. Indeed, while DNA can be described using alleles at given loci, which are easily measurable, friction ridge skin contains patterns with many different levels of details that cannot be readily summarized by discrete variables. In addition, impressions from these patterns are affected by numerous factors (such as distortion, substrate, detection technique), which lowers the reproducibility of their characteristics and increases the complexity of their modeling. Several research projects attempted to capture both the multi-dimensionality and heterogeneity of pattern variables by measuring the similarity between pairs of impressions and summarizing it typically as a univariate score. Score-based models assign the probability of the score resulting from the comparison of a latent print with a control

print under two mutually exclusive hypotheses to generate a measure of the weight of the evidence.

However, score-based statistical models have intrinsic limitations: the integration of the score in the statistical model is not well understood [20, 23]; the need to compute a score between trace and control prints prevents from measuring the specificity of the features observed on the trace (and thus, precludes from providing information at the end of the Analysis phase); and adding new features to an existing model requires the redevelopment and re-optimization of the scoring algorithm.

Generative models: Other researchers attempted to model the underlying distributions of some of the features that can be observed on friction ridge skin impressions. In theory, these models can assign the probability of observing any constellation of fingerprint features detected on a latent print. However, these models were developed on datasets that were too limited in size to account for the dependencies between the hundreds of fingerprint features (in particular minutiae) that can be observed on any given impression and to account for the variability between impressions from different fingers; the models used to describe the underlying distributions do not fit well the empirical distributions of the features, especially when it comes to model the dependency between neighboring minutiae; and, those models do not provide a satisfactory mean of accounting for the level of similarity between the trace and the considered control prints (and thus, limits the support that those models can provide during the Comparison and Evaluation phases of the examination process).

The next sections describe a novel approach for the quantification of the weight of fingerprint evidence. The idea behind this model is to (a) reduce the complexity of the problem while accounting for the dependencies between fingerprint features, as in score-based models, and (b) provide a measure of the specificity of the crime scene print without reference to the control print, as in generative models. This new approach is designed to provide support to the decisions made during all phases of the examination process.

In this new approach, we attempt to reduce the dimensionality of the sets of variables used to describe minutiae configurations by using shape variables as proposed in [24]. In our model, the probability of observing a particular minutiae configuration shape is assigned by

modeling the distribution of the shapes of similar constellations retrieved from a large dataset of reference impressions, which helps preventing the common issues of generative models.

The next sections present the general framework of the model (section 7.1); the “radial triangulation” used in this research project to measure variables on minutiae configurations (section 7.2); the individual components of the model for shapes of configurations (section 7.3), minutiae directions (section 7.4) and types (section 7.5); the datasets used to develop, support and test the model (section 7.6); and data on the performance of the model when tested using pairs of latent and control prints originating from the same source and from different sources (section 7.7).

7.1. Model

The general framework of the model is similar to the one described by Neumann et al. [25]. We denote the entire collection of observations made on the latent print by the multi-dimensional quantity Y . We denote the observations made on corresponding properties on the control print by X . The model uses Y and X to address the following propositions:

H_p : the latent print comes from the same finger as the control;

H_d : the latent print comes from some other, unknown finger, from a different person⁴.

Following Lindley [26] and many others, the objective is to assign a value to the likelihood ratio (LR), which we write here, after some simplifications [25], as:

$$LR = \frac{p_{X,Y}(X,Y|H_p)}{p_{X,Y}(X,Y|H_d)} = \frac{p_{Y|X}(Y|H_p)}{p_Y(Y|H_d)} \quad (1)$$

In [25], we explained that the number of minutiae k recorded on the latent print defines the dimensionality of the problem. We denote the vector of observations made on the latent print by $y^{(k)}$ and on the control print by $x^{(k)}$.

⁴ While the model described in this report addresses propositions at the finger level, the model can be extended at the person level as proposed in [27]

When comparing the features observed on a latent print with the ones on a control print, an examiner will attempt to select the subset $x^{(k)}$ of X that corresponds best to the observations $y^{(k)}$ made on the trace. The examiner first verifies that the general pattern of the ridge flow on the latent and control prints are similar. Secondly, the examiner focuses on the general location within the ridge flow (i.e., core, delta, periphery) of the control impression where the minutiae were observed on the latent print. Thirdly, the examiner determines whether a set of features $x^{(k)}$ on the control print resemble the set $y^{(k)}$ observed on the trace at the corresponding location within the ridge flow. Finally, the examiner compares the details of the features between both prints. Mathematically, this process corresponds to the selection of a single k minutiae configuration out of the $\binom{n}{k}$ possible configurations on the control print, such that it is the most similar one to the k minutiae configuration observed on the latent print. We denote this configuration by $x_{min}^{(k)}$. Equation (1) can be rewritten as follows:

$$LR = \frac{P_{Y|X_{min}}(y^{(k)}|H_p)}{P_Y(y^{(k)}|H_d)} \quad (2)$$

At this point in the development of the model, it is critical to realize that the k minutiae on the trace, and the corresponding k minutiae on the control print are paired during the comparison process (as explained in section 5.3): the i^{th} minutia on the latent print is associated to one and only minutia on the control print. This information is implied in the model.

Assigning a probability to the numerator of the model in Equation (2) involves the comparison of the latent print with a single control print. However, assigning a probability to the denominator requires a model of the distributions of fingerprint features in a relevant population determined by H_d [25]. This model can be parametric, as in the generative fingerprint models mentioned above, or can be data-driven as proposed in [25]. In this project, we use a fingerprint matching algorithm as a proxy for the human-based comparison process described above. The matching algorithm is used to search a large dataset of reference finger impressions, and select, on each finger, the set of k minutiae configurations that is most similar to $y^{(k)}$ in terms of general pattern, location on the ridge

flow, and general appearance (i.e. shape). We denote these configurations by $Z_{\min}^{(k)}$. As mentioned above, and similarly to a human examiner, it is important to realize that the matching algorithm pairs the minutiae between the latent and reference prints.

We define V as the existence of such minutiae configuration on a control/reference print. V is a dichotomous variable indicating the presence ($v = 1$) or absence ($v = 0$) of a compatible set of k features⁵ on a given control/reference print.

We include the additional information provided by V in Equation (2) as follows:

$$LR = \frac{p_{Y|X_{\min},V}(y^{(k)}|H_p)p_V(v|H_p) + p_{Y|X_{\min},\bar{V}}(y^{(k)}|H_p)p_{\bar{V}}(\bar{v}|H_p)}{p_{Y|V}(y^{(k)}|H_d)p_V(v|H_d) + p_{Y|\bar{V}}(y^{(k)}|H_d)p_{\bar{V}}(\bar{v}|H_d)} \quad (3)$$

Equation (3) can be simplified by making the following assumptions:

1. $p_{Y|\bar{V}}(y^{(k)}|H_d)$ tends to zero when the examiner/matching algorithm cannot find compatible configurations in the control/reference prints;
2. $p_V(v|H_p)$ tends to one when H_p is true⁶.

The remaining $p_V(v|H_d)$ can easily be assigned by calculating the relative frequency of reference fingers containing a configuration of k minutiae, which is compatible with the configuration observed on the latent print.

The terms $p_{Y|X_{\min},V}(y^{(k)}|H_p)$ and $p_{Y|V}(y^{(k)}|H_d)$ are estimated by characterizing k minutiae configurations using three different types of variables: shape of configuration S , minutiae direction D and minutiae type T . Rewriting Equation (3), we obtain:

$$LR \approx \frac{p_{Y|X_{\min},V}(y_S^{(k)}, y_D^{(k)}, y_T^{(k)}|H_p)}{p_{Y|V}(y_S^{(k)}, y_D^{(k)}, y_T^{(k)}|H_d)} \frac{1}{p_V(v|H_d)} \quad (4)$$

⁵ Note that the term "compatible" depends on the performance of the human examiner under H_p and the selected algorithm under H_d . In this study, we used a latent/tenprint matching algorithm provided by 3M Cogent.

⁶ This is not strictly true, in particular when the latent print is heavily distorted. But this assumption has no significant impact on the rest of the mathematical development.

In Equation (4), we consider that the shapes of minutiae configurations, and the types and directions of the minutiae are influenced by the general pattern of the prints and by the location of the configurations on the ridge flow. This dependency is included in the variable V . However, we make the assumption that within a particular location (i.e., core, delta or periphery) of a particular pattern, shapes of configurations, and minutiae types and directions are independent of each other.

Using this assumption, we obtain:

$$LR \approx \frac{P_{Y_S|X_{\min},V}(y_S^{(k)}|H_p) P_{Y_D|X_{\min},V}(y_D^{(k)}|H_p) P_{Y_T|X_{\min},V}(y_T^{(k)}|H_p)}{P_{Y_S|V}(y_S^{(k)}|H_d) P_{Y_D|V}(y_D^{(k)}|H_d) P_{Y_T|V}(y_T^{(k)}|H_d)} \frac{1}{P_V(v|H_d)} \quad (5)$$

Our model has three conditionally independent components and a given event V . The first component is based on the shape of the configuration, the second component is based on the directions of the minutiae in the configuration, and the third component is based on their types. Note that the design of the model enables the consideration of additional fingerprint features, without the need for changing the existing elements of the model.

It would then be possible to consider other elements commonly used by latent print examiners, such as the presence of differences between the features observed on the trace and control prints, the presence/absence of scars, warts and creases, as well as the presence/absence of impressions from sweat pores on the prints, or the shape of the ridges.

To ease the description of the model, the three components of the model are described in separated sections. However, we first describe a method for quantifying the observations made on the fingerprints and for reducing the dimensionality of the problem.

7.2. Method for quantitative observations on fingerprints

The process of extracting features from friction ridge impressions is image dependent: minutiae locations and directions are measured relatively to a coordinate system defined by the image. Figure 12 displays a set of 7 features on a crime scene impression and the corresponding features on a control print. Figure 12 shows that the locations and directions

of corresponding minutiae are different in the two images, and that it would be inappropriate to build a statistical model relying directly on these measurements.

Following Neumann et al. [25], we propose to describe any configuration of k minutiae as a set of k triangles, which vertices are defined by pairs of consecutive minutiae and the virtual centroid of the k configurations. This design enables the capture of the spatial relationship between minutiae, provides some robustness to the distortion affecting impressions when finger pads are pressed against a surface, and allows for measuring variables with respect to the triangles, thus breaking their dependency to the images.

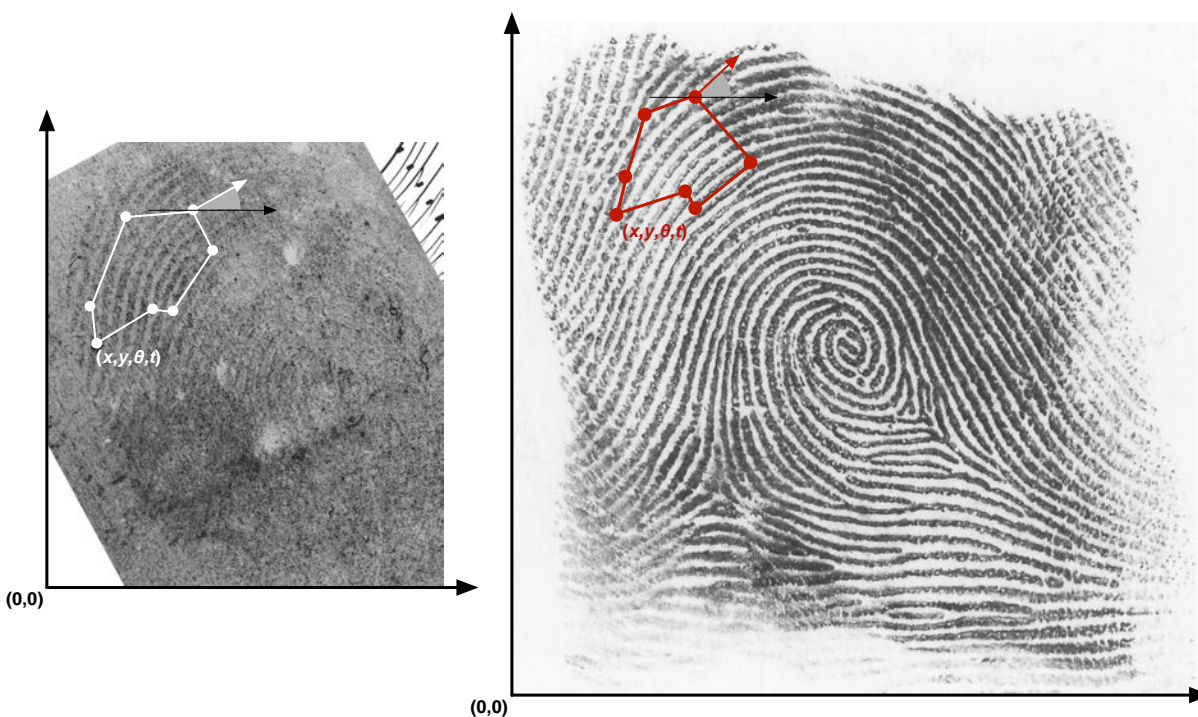


Figure 12: Raw information extracted from minutiae location and direction, with indication of the image defined axes.

Figure 13 shows how the considered variables are extracted from a given configuration. At first, the minutiae are annotated on the finger impression using markers indicating their locations, types and directions (section 5.3). This image dependent information is used to organize the minutiae around a virtual centroid, defined by the arithmetic mean of the spatial coordinates of the minutiae. This process creates a series of triangles, which vertices are defined by pairs of consecutive minutiae and the centroid. The triangulation is rotationally independent: the minutiae will be organized in the same order, irrespectively

of the angle between the impression and the axes of the image. The triangulation also enables to measure the considered variables according to the triangles, and thus to break their dependency to the images.

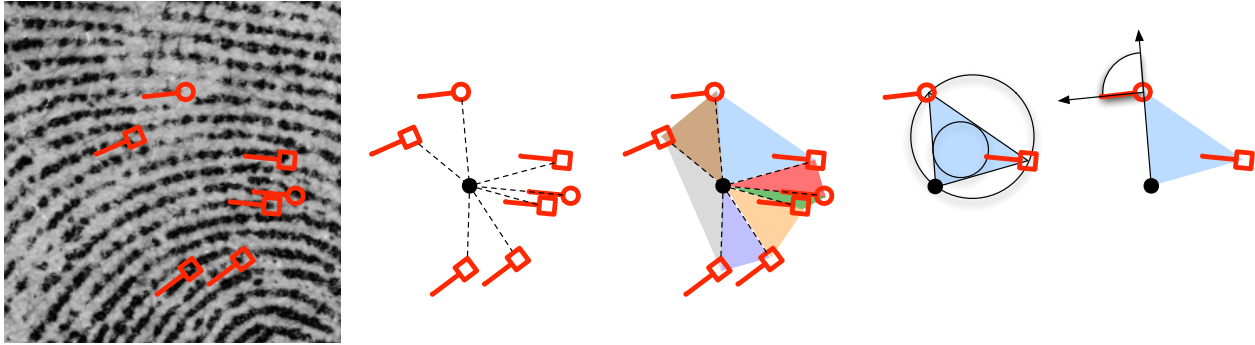


Figure 13: Extraction of the variables considered by the model from the raw information available on the image of a fingerprint impression. From left to right: (a) annotation of the minutiae on the fingerprint image – (b) definition of the centroid and organization of the minutiae with respect to the centroid – (c) creation of the triangle – (d) extraction of the shape variables for one triangle – (e) extraction of the type and direction variables of the minutiae for one triangle (the variables for all triangles are similarly extracted).

In this research project, we decided to characterize each configuration by the following variables:

S The shape of each triangle in the configuration is described by two popular quantitative measurements: (a) the ratio between its area and perimeter (form factor), and (b) the ratio between the diameters of its circumcircle and incircle (aspect ratio). The shape of a latent print configuration can be formally represented by $Y_S = [Y_{S,1}, \dots, Y_{S,k}]$.

D The direction of each minutia in the configuration is described by the angle between the direction of the minutia and an axis defined by the centroid and the minutiae location. The angle is measured counterclockwise from the axis to the minutiae. The directions of the minutiae in a latent print configuration can be formally represented by $Y_D = [Y_{D,1}, \dots, Y_{D,k}]$.

T The type of each minutia in the configuration is described by a nominal variable, which can take the following values: *RE* for ridge ending minutiae; *BI* for bifurcation minutiae; *UK* for minutiae which type is unknown. The types of the minutiae in a latent print configuration can be formally represented by $Y_T = [Y_{T,1}, \dots, Y_{T,k}]$.

7.3. Shape element of the model

From Equation 5 and section 7.2, we obtain the shape element of the model as:

$$LR_S = \frac{p_{Y_S|X_{\min},V}(y_S^{(k)}|H_p)}{p_{Y_S|V}(y_S^{(k)}|H_d)} = \frac{p_{Y_S|X_{\min},V}(y_{S,1}^{(k)}, \dots, y_{S,k}^{(k)}|H_p)}{p_{Y_S|V}(y_{S,1}^{(k)}, \dots, y_{S,k}^{(k)}|H_d)} \quad (6)$$

In order to simplify the k dimensionality of the modeling of the ratio in Equation 6, we assume that the shape of triangle i is mostly influenced by its immediate neighbors. This assumption is fairly reasonable as adjacent triangles share one side with each other, while non-adjacent triangles share only one vertex (Table 5). However, removing possible dependencies between non-adjacent triangles forces us to set the first triangle, and to assign a marginal probability to this triangle, rather than a joint probability.

Shape	$Z_{S,1}$	$Z_{S,2}$	$Z_{S,3}$	$Z_{S,4}$	$Z_{S,5}$	$Z_{S,6}$	$Z_{S,7}$	$Z_{S,8}$	$Z_{S,9}$	$Z_{S,10}$	$Z_{S,11}$	$Z_{S,12}$
$Z_{S,1}$	1											
$Z_{S,2}$	-.211	1										
$Z_{S,3}$.021	-.278	1									
$Z_{S,4}$	-.040	-.055	-.188	1								
$Z_{S,5}$.042	.008	.008	-.308	1							
$Z_{S,6}$.113	.001	-.027	-.037	-.156	1						
$Z_{S,7}$.028	.074	-.004	.020	-.033	-.326	1					
$Z_{S,8}$	-.009	.047	.041	.072	-.004	.008	-.185	1				
$Z_{S,9}$	-.021	-.001	-.035	-.041	-.023	.000	.054	-.194	1			
$Z_{S,10}$.009	.100	.059	.119	.018	-.029	-.000	.066	-.221	1		
$Z_{S,11}$.007	-.014	.052	.051	.019	.033	-.020	-.067	-.016	-.119	1	
$Z_{S,12}$	-.180	-.031	-.009	.046	.022	.043	-.005	.051	-.069	-.012	-.325	1

Table 5: Spearman rank correlation coefficients between the form factors of approximately 100,000 reference prints paired with a example latent print with 12 minutiae

Each $Y_{S,i}$ is a bi-dimensional variable containing the form factor and the aspect ratio of triangle i . The form factor and the aspect ratio are functionally independent and may capture the shape of a triangle in complimentary ways. To select the first triangle, we use the aspect ratio information of the k triangles in the configuration. We set $Y_{S,1} = \min_{1 \leq i \leq k} Y_{S,i}$ based on the aspect ratio variable of each i triangle, and then register the remaining $k-1$ triangles clockwise. Since the k minutiae in the latent configuration are paired with the k minutiae in the control and reference prints, the triangles in these prints are reordered according to Y_S .

Based on this assumption, Equation 6 can be rewritten as:

$$LR_S = \frac{P_{Y_S|X_{\min},V}(y_{S,1}^{(k)}|H_p) P_{Y_S|X_{\min},V}(y_{S,2}^{(k)}|y_{S,1}^{(k)}, H_p) \dots P_{Y_S|X_{\min},V}(y_{S,k}^{(k)}|y_{S,k-1}^{(k)}, H_p)}{P_{Y_S|V}(y_{S,1}^{(k)}|H_d) P_{Y_S|V}(y_{S,2}^{(k)}|y_{S,1}^{(k)}, H_d) \dots P_{Y_S|V}(y_{S,k}^{(k)}|y_{S,k-1}^{(k)}, H_d)} \quad (7)$$

LR_S does not have an analytical solution and needs to be estimated from samples. The numerator describes the probability of observing the configuration on the latent print if this impression originates from the same finger as the control print. Ideally, assigning this probability would require the source of the control print to generate multiple pseudo-traces in various conditions of distortion and pressure, and to model the distribution of the shape of the considered k minutiae configurations across these pseudo-traces. In practice, this is unrealistic and we used the same distortion model as in Neumann et al. [25], based on Bookstein [28], to generate pseudo-traces from the control print.

The denominator describes the probability of observing the configurations on the latent print in a relevant population defined by H_d . As mentioned in section 7.1, in the absence of satisfactory theoretical model describing the joint distributions of various fingerprint features, we use a fingerprint matching algorithm to search the latent print configuration in a large database and to retrieve the most similar k configuration on each reference finger.

Let $f_{X_S}(x)$ and $f_{Z_S}(z)$ be the density functions of the shapes of the control print X_S and reference prints Z_S respectively. The joint density function of X_S can be written by $f_{X_S}(x) = f_{X_{S,1}}(x_1)f_{X_{S,2}|X_{S,1}}(x_2)\dots f_{X_{S,i}|X_{S,i-1}}(x_i)$. We propose to use $f_{X_S}(x)$ and $f_{Z_S}(z)$ as an estimator for the numerator and denominator of LR_S respectively.

The histogram estimates of $f_{X_{S,i}}(x_i)$, with $i = 1, \dots, k$ are reasonably symmetrical, unimodal and similar to a normal distribution. However, the histogram estimates of $f_{Z_{S,i}}(z_i)$, with $i = 1, \dots, k$ are moderately to highly skewed on their right or left tails. Thus, we decided to model $f_{X_{S,1}}(x_1)$ and $f_{X_{S,i+1}|X_{S,i}}(x_i)$ using uni- and bivariate normal densities; and we choose not to impose any parametric assumption on the structure of the densities for $f_{Z_{S,1}}(z_1)$ and $f_{Z_{S,i+1}|Z_{S,i}}(z_i)$ by learning the density function from the data.

7.4. Direction element of the model

From Equation 5 and section 7.2, we obtain the direction element of the model as:

$$LR_D = \frac{P_{Y_D|X_{\min},V}(y_D^{(k)}|H_p)}{P_{Y_D|V}(y_D^{(k)}|H_d)} = \frac{P_{Y_D|X_{\min},V}(y_{D,1}^{(k)}, \dots, y_{D,k}^{(k)}|H_p)}{P_{Y_D|V}(y_{D,1}^{(k)}, \dots, y_{D,k}^{(k)}|H_d)} \quad (8)$$

As explained in section 7.2, the direction of each minutia in a configuration is described with respect to an axis defined by the centroid of the configuration and the minutia itself. This allows for obtaining directional information on the minutiae in the configuration regardless of the rotation and location of the impression on the image. This transformation also enables the reduction of the dependency between the directions measured for neighboring minutiae (Tables 6 and 7). We observed that few neighboring minutiae have moderately correlated directions; however most show low correlation.

Direction	$Z_{D,1}$	$Z_{D,2}$	$Z_{D,3}$	$Z_{D,4}$	$Z_{D,5}$	$Z_{D,6}$	$Z_{D,7}$	$Z_{D,8}$	$Z_{D,9}$	$Z_{D,10}$	$Z_{D,11}$	$Z_{D,12}$
$Z_{D,1}$	1											
$Z_{D,2}$.115	1										
$Z_{D,3}$.690	.147	1									
$Z_{D,4}$.672	.172	.749	1								
$Z_{D,5}$.046	.728	.115	.189	1							
$Z_{D,6}$.620	.183	.699	.766	.235	1						
$Z_{D,7}$.548	.239	.601	.687	.283	.729	1					
$Z_{D,8}$.548	.241	.564	.638	.239	.653	.686	1				
$Z_{D,9}$.496	.225	.500	.558	.223	.578	.611	.617	1			
$Z_{D,10}$.673	.174	.690	.725	.160	.722	.686	.666	.593	1		
$Z_{D,11}$.644	.113	.623	.624	.056	.598	.559	.566	.545	.638	1	
$Z_{D,12}$.109	.746	.049	.075	.644	.083	.156	.168	.156	.101	.078	1

Table 6: Spearman rank correlation coefficients between the directions of neighboring minutiae in 100,000 reference prints paired with a example latent print with 12 minutiae – **before** transformation described in section 7.2

Direction	$Z_{D,1}$	$Z_{D,2}$	$Z_{D,3}$	$Z_{D,4}$	$Z_{D,5}$	$Z_{D,6}$	$Z_{D,7}$	$Z_{D,8}$	$Z_{D,9}$	$Z_{D,10}$	$Z_{D,11}$	$Z_{D,12}$
$Z_{D,1}$	1											
$Z_{D,2}$.022	1										
$Z_{D,3}$	-.056	.022	1									
$Z_{D,4}$.003	-.004	-.006	1								
$Z_{D,5}$.015	-.003	-.026	-.062	1							
$Z_{D,6}$.020	.049	-.048	-.167	-.101	1						
$Z_{D,7}$.098	.164	.039	-.084	-.168	.093	1					
$Z_{D,8}$	-.020	.128	.064	-.003	.015	-.015	.030	1				
$Z_{D,9}$	-.042	.077	.056	.059	.169	-.069	-.068	.059	1			
$Z_{D,10}$	-.012	.053	-.015	.002	.127	-.016	-.038	.065	.122	1		
$Z_{D,11}$	-.044	.046	.005	.027	.152	-.037	-.019	.020	.142	.118	1	
$Z_{D,12}$	-.083	-.000	-.016	-.069	.018	.063	.070	.082	.047	.085	.049	1

Table 7: Spearman rank correlation coefficients between the directions of neighboring minutiae in 100,000 reference prints paired with the same latent print with 12 minutiae as in Table 6 – **after** transformation described in section 7.2

By taking advantage of the low correlation, we make the assumption of independence between the minutiae direction (as specifically measured in our project) and obtain the following ratio:

$$LR_D = \prod_{i=1}^k \frac{p_{Y_{D,i}|X_{\min},V}(y_{D,i}^{(k)}|H_p)}{p_{Y_{D,i}|V}(y_{D,i}^{(k)}|H_d)} \tag{9}$$

In a similar fashion as in section 7.3, we define $f_{X_D}(x)$ and $f_{Z_D}(z)$ as the density functions of the directions of the minutiae in the control print X_S and reference prints Z_S respectively. As explained in section 7.3, we use a distortion model to generate pseudo-traces for the estimation of the density function under H_p and a large reference dataset for the estimation of the density function under H_d . The histogram estimates for the density function of minutiae directions in the control print show that they tend to be skewed to the right, while the estimates for the reference prints show multiple modes. We decided to estimate $f_{X_D}(x)$ and $f_{Z_D}(z)$ using non-parametric distributions based on von Mises kernels.

7.5. Type element of the model

From Equation 5 and section 7.2, we obtain the type element of the model as:

$$LR_T = \frac{P_{Y_T|X_{\min},V}(y_T^{(k)}|H_p)}{P_{Y_T|V}(y_T^{(k)}|H_d)} = \frac{P_{Y_T|X_{\min},V}(y_{T,1}^{(k)}, \dots, y_{T,k}^{(k)}|H_p)}{P_{Y_T|V}(y_{T,1}^{(k)}, \dots, y_{T,k}^{(k)}|H_d)} \quad (10)$$

In order to simplify the dimensionality of the probabilities in Equation 10, we assume that minutiae types are influenced by the location of the minutiae within the pattern of the ridge flow (accounted for in V) but not by each other. Thus, given V , we can make the following simplification:

$$LR_T = \prod_{i=1}^k \frac{P_{Y_{T,i}|X_{\min},V}(y_{T,i}^{(k)}|H_p)}{P_{Y_{T,i}|V}(y_{T,i}^{(k)}|H_d)} \quad (11)$$

We have defined previously minutiae type as nominal variable such that any i minutia can take one of the following values $y_{T,i}^{(k)} = \{RE, BI, UK\}$. That said, the observation of the type of a minutia on a potentially distorted and degraded latent print is not only conditioned by the true type of that minutia, but also by the ability of the examiner to correctly interpret the ridge flow. Therefore, we have for the numerator:

$$\begin{aligned} P_{Y_{T,i}|X_{\min},V}(y_{T,i}^{(k)}|H_p) &= \sum_j^{\{RE, BI, UK\}} P_{Y_{T,i}|X_{\min},V}(y_{T,i}^{(k)} = j|H_p) \\ &= \sum_l^{\{RE, BI\}} \sum_j^{\{RE, BI, UK\}} P_{Y_{T,i}|X_{T,i}=l,V}(y_{T,i}^{(k)} = j|H_p) P_{X_{T,i}|V}(x_{T,i}^{(k)} = l|H_p) \end{aligned} \quad (12)$$

Ideally, the $P_{X_{T,i}|V}(x_{T,i}^{(k)} = l|H_p)$ terms should be assigned by having the examiner annotate the type of the i^{th} minutia on series of pseudo-traces generated by the source of the control impression. Indeed, $P_{X_{T,i}|V}(x_{T,i}^{(k)} = l|H_p)$ can be developed in a similar fashion as $P_{Y_{T,i}|X_{\min},V}(y_{T,i}^{(k)} = j|H_p)$ using by conditioning on the true type of the minutiae observed on the friction ridge skin of the finger pad. However, for all intents and purposes of this project, we consider that there is no uncertainty affecting the determination of the type of a given

minutia when observed on control prints or pseudo-traces. Thus, in our model, $p_{X_{T,i}|V}(x_{T,i}^{(k)} = l | H_p)$ takes values $\{0,1\}$ depending on whether the l^{th} type is observed by the examiner on the i^{th} minutia of the k configuration present on the control print.

The $p_{Y_{T,i}|X_{T,i}=l,V}(y_{T,i}^{(k)} = j | H_p)$ terms take different values depending on the type observed on the latent print for the i^{th} minutia. A survey of a series of 82 minutiae, each annotated by more than 200 latent prints examiners on 12 pairs of latent and control prints, reveals that (for any i):

1. When the i^{th} minutia on the latent print is deemed to be a ridge ending, $p_{Y_{T,i}|X_{T,i}=RE,V}(y_{T,i}^{(k)} = RE | H_p) = 0.76$ and $p_{Y_{T,i}|X_{T,i}=BI,V}(y_{T,i}^{(k)} = RE | H_p) = 0.16$; all other terms are set to 0;
2. When the i^{th} minutia on the latent print is deemed to be a bifurcation, $p_{Y_{T,i}|X_{T,i}=RE,V}(y_{T,i}^{(k)} = BI | H_p) = 0.16$ and $p_{Y_{T,i}|X_{T,i}=BI,V}(y_{T,i}^{(k)} = BI | H_p) = 0.75$; all other terms are set to 0;
3. When the type of the i^{th} minutia on the latent print is unknown, $p_{Y_{T,i}|X_{T,i}=RE,V}(y_{T,i}^{(k)} = UK | H_p) = 0.08$ and $p_{Y_{T,i}|X_{T,i}=BI,V}(y_{T,i}^{(k)} = UK | H_p) = 0.09$; all other terms are set to 0.

Note that under H_p only one of the $p_{Y_{T,i}|X_{T,i}=l,V}(y_{T,i}^{(k)} = j | H_p) p_{X_{T,i}|V}(x_{T,i}^{(k)} = l | H_p)$ terms is non-null depending on the types observed on the i^{th} pair of minutiae on the latent and control prints.

Similarly for the denominator, we have:

$$\begin{aligned}
 p_{Y_{T,i}|V}(y_{T,i}^{(k)} | H_d) &= \sum_j^{\{RE, BI, UK\}} p_{Y_{T,i}|V}(y_{T,i}^{(k)} = j | H_d) \\
 &= \sum_l^{\{RE, BI\}} \sum_j^{\{RE, BI, UK\}} p_{Y_{T,i}|V}(y_{T,i}^{(k)} = j | H_d) p_V(l | H_d)
 \end{aligned} \tag{13}$$

The $p_V(l | H_d)$ terms can be assigned by using the distribution of the type of the i^{th} minutia in the k minutiae configurations retrieved by the matching algorithm as described in the previous sections.

The $p_{Y_{T,i}|V}(y_{T,i}^{(k)} = j | H_d)$ terms are assigned using the same values as for the numerator depending on whether a ridge ending, bifurcation or unknown type was observed on the i^{th} minutia of the latent print.

7.6. Datasets

Several datasets were available to the research team. A dataset was used to study the different probability densities during the development of the model; a second dataset was used to test the model; and a third dataset was directly used to estimate the different probability density functions of the denominator when quantifying the weight of fingerprint evidence.

Development dataset: The model was developed using 48, 45 and 33 configurations of respectively 4, 8 and 12 minutiae sampled from latent and corresponding control prints obtained from archive casework [25]. The histogram estimates of the numerators of the different elements of the model were obtained by generating 2,500 pseudo-traces from each control print, using the distortion model mentioned in section 7.3. The dataset used for studying the histogram estimates of the denominators of the different elements of the model contained the same 12,000 impressions as in [25].

Reference dataset: A reference dataset of approximately 4,000,000 control finger impressions from approximately 400,000 anonymous donors was used to support the assignment of the probability density functions in the denominator of the model.

Test dataset: The model was tested using 565 latent prints: the first 364 latent prints originate from casework and correspond to the data used to test the model in [25]; an additional 201 latent prints, developed in casework-like conditions, were added to complete the test dataset. The minutiae on the 565 latent prints and the corresponding control prints were annotated using PiAnoS (section 5.3). Each minutia was paired between the latent and control prints using PiAnoS's pairing feature.

The following numbers of configurations of 3 to 12 minutiae (Table 8) were sampled from the 565 impressions and used to test the model⁷:

# minutiae	3	4	5	6	7	8	9	10	11	12
All regions	96	99	98	97	97	100	100	96	93	89
Core	151	170	159	144	125	97	72	60	47	33
Delta	61	70	66	57	61	29	25	24	17	14
Periphery	159	180	125	142	101	76	53	39	27	16

Table 8: Configurations used to test the model, presented by number of minutiae and region.

To test the model under the most difficult conditions, the latent print configurations were searched against the reference dataset. For each latent configuration, the most similar k configuration in the database was retrieved and used to test the model in the situation when the latent and control configurations do not originate from the same source.

For each configuration, we therefore obtained a triplet of data: (1) a latent print configuration, (2) a configuration on the corresponding control print, (3) a configuration on the most similar non-corresponding print out of 4,000,000 impressions.

7.7. Model performances

The performances of the model were considered with respect to the overarching goals of the project. The proposed model should not only be able to quantify the weight of fingerprint evidence, but should allow for supporting the decisions made by latent print examiners at the different stages of the examination process:

1. The decision of suitability is associated with the denominator probability of observing the minutiae configuration, detected on a latent print during the analysis phase, in a relevant population: common configurations, represented by a higher probability of being observed in the relevant population, may not be worth examining further due to their low potential evidential value.
2. The level of resemblance of latent and control prints is associated with the numerator probability of observing the minutiae configuration on a latent print if it

⁷ Note that more configurations were sampled, but, when searched in the reference datasets, some latent configurations were not associated with a sufficient numbers of reference configurations to assign the density functions. In essence, $p_v(v|H_d) = 0$ was assigned for these configurations and it was not possible to calculate their LR.

truly originates from the same source as the control print: higher level of agreement (and absence of differences) will result in higher probability values.

3. The final conclusion of the examination during the Evaluation phase is associated with the ratio of the numerator probability assigned based on the similarities and differences observed between the latent and control prints, and with the probability of observing these minutiae in a relevant population. Higher ratios will be more supportive of the hypothesis that the two prints come from the same source than ratio values below 1.

In addition, the data on the performances of the model will also inform us on the robustness and discriminative abilities of each of its three components and indicate which feature(s) may be the most appropriate to support the decisions made by examiners practice.

Figure 14a-d presents the data obtained for the denominator of the three components (*S*, *D* and *T*) of the model, separately (a, b and c), and jointly (d). We observe that the contribution of the shape variable is much larger than the other ones. We also observe the lack of contribution of the direction component of the model. Overall, Figure 14a-d shows that the model has the ability to assess the specificity of minutiae configurations; that this specificity increases with the number of minutiae; but that it also varies between configurations with a given number of minutiae, as these configurations have different shapes and combinations of minutiae types.

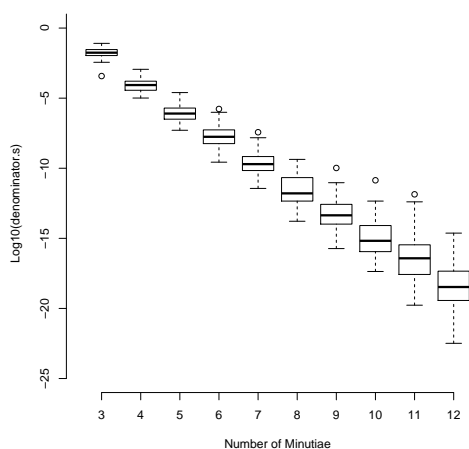


Figure 14a: Values for the denominators of the shape component of the model – All regions

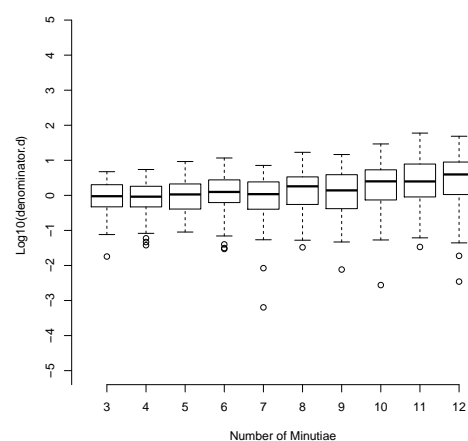


Figure 14b: Values for the denominators of the direction component of the model – All regions

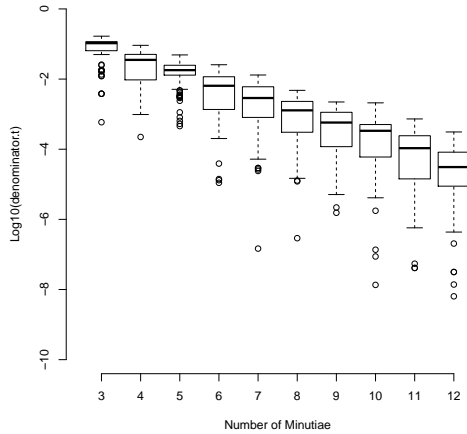


Figure 14c: Values for the denominators of the type component of the model – All regions

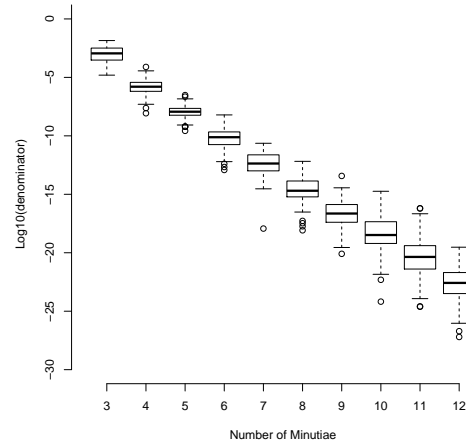


Figure 14d: Values for the denominators of the model – All regions

Figure 15a-d presents the data obtained for the numerators of the three components (S , D and T) of the model, separately (a,b and c), and jointly (d). The Figures on the left hand-side present the data obtained for latent print configurations compared with control prints provided by the true source, while Figures on the right hand-side present data obtained for the same latent configurations when compared with prints provided by different sources. Figure 15a-d shows that the expected probability of observing the features on a latent print, based on potentially corresponding features observed on the control print, decreases with the number of minutiae. This is not surprising as the increase in the number of minutiae induces increasing variability between multiple impressions of the same set of k minutiae due to pressure, distortion and other factors.

When comparing the left (same source) to the right (different sources) columns of Figure 15a-d, we realize that the expected numerator probability decreases faster when latent prints are compared to control prints originating from different sources. This effect is the result of the added discrimination introduced by the increasing number of features. The somewhat large ranges of values calculated for the numerator of the model can be explained by two elements: (1) the distortion model used in this project is not providing enough variability in the set of pseudo-traces generated from the control print, and thus cannot compensate for medium to large distortion effects of the latent print; (2) the model is affected by a lack of accuracy from the users annotating the prints in PiAnoS.

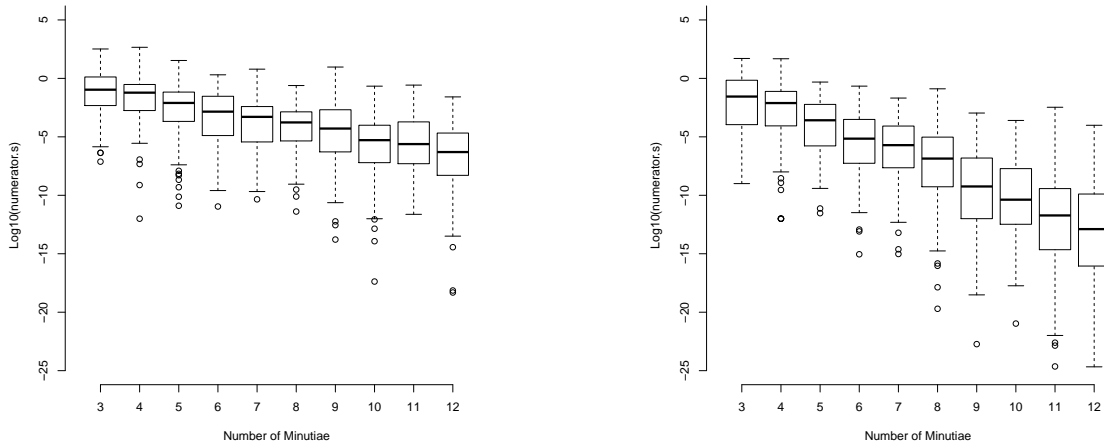


Figure 15a : Values for the numerators of the shape component of the model – All regions - Left: same source control print – Right: difference source control print.

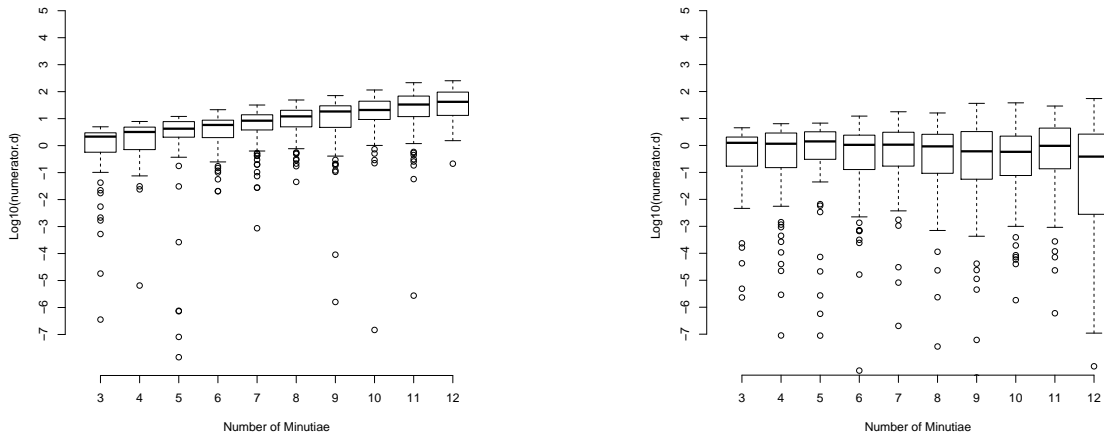


Figure 15b: Values for the numerators of the direction component of the model – All regions - Left: same source control print – Right: difference source control print

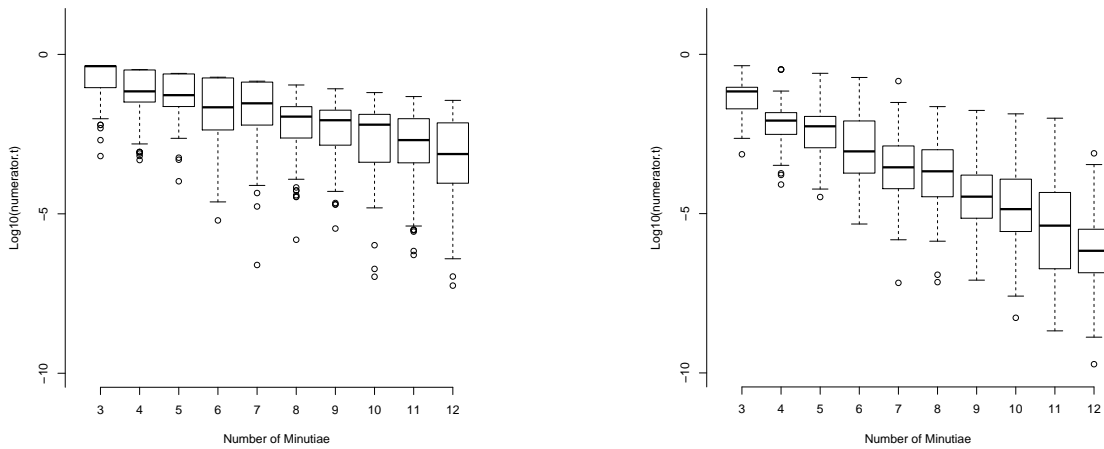


Figure 15c: Values for the numerators of the type component of the model – All regions - Left: same source control print – Right: difference source control print

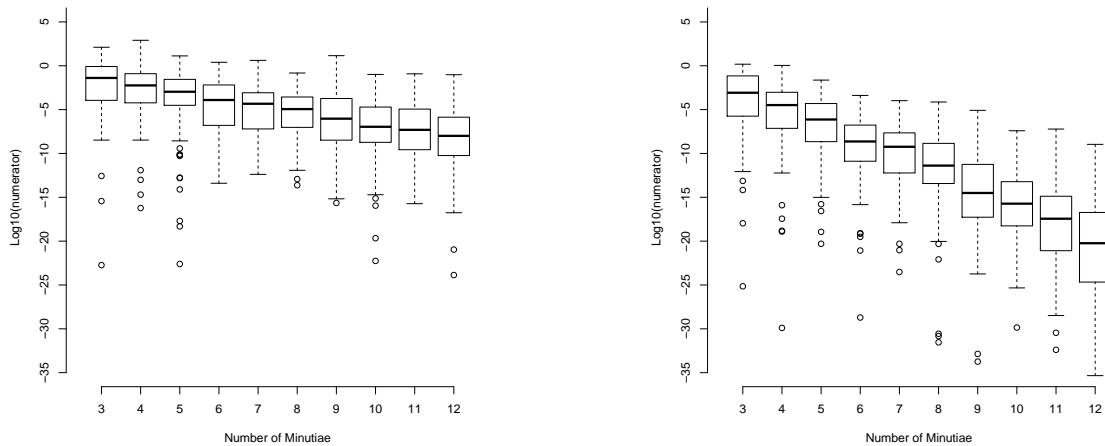


Figure 15d: Values for the numerators of the model – All regions - Left: same source control print – Right: difference source control print

Figure 16a-d presents the data obtained for the LR of the three components (S , D and T) of the model, separately (a,b and c), and jointly (d). The Figures on the left hand-side column present the data obtained for latent print configurations compared with control prints provided by the true source, while the Figures on the right hand-side column present data obtained for the same latent configurations when compared with prints provided by different sources. As expected, we observe that the LRs calculated for pairs of latent and control prints originating from the same source increases with the number of minutiae, while it remains centered around $LR=1$ for pairs of latent and control prints originating from different sources. These results are similar to the ones obtained by Neumann et al. [25]: (1) the expected value of the LR increases with the number of minutiae, (2) a range of LR values is observed for each number of minutiae, indicating that each comparison needs to be evaluated based on its own merits.

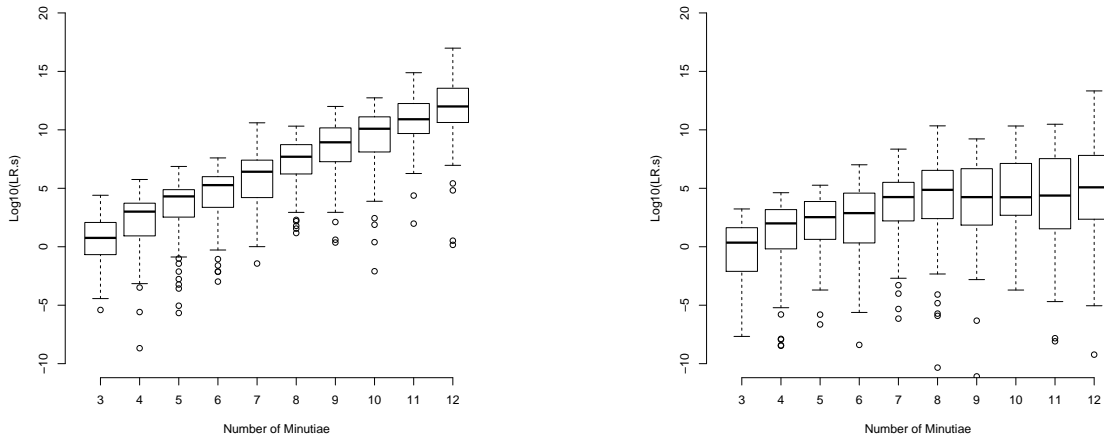


Figure 16a: Values for the LR.s of the shape component of the model – All regions - Left: same source control print – Right: difference source control print.

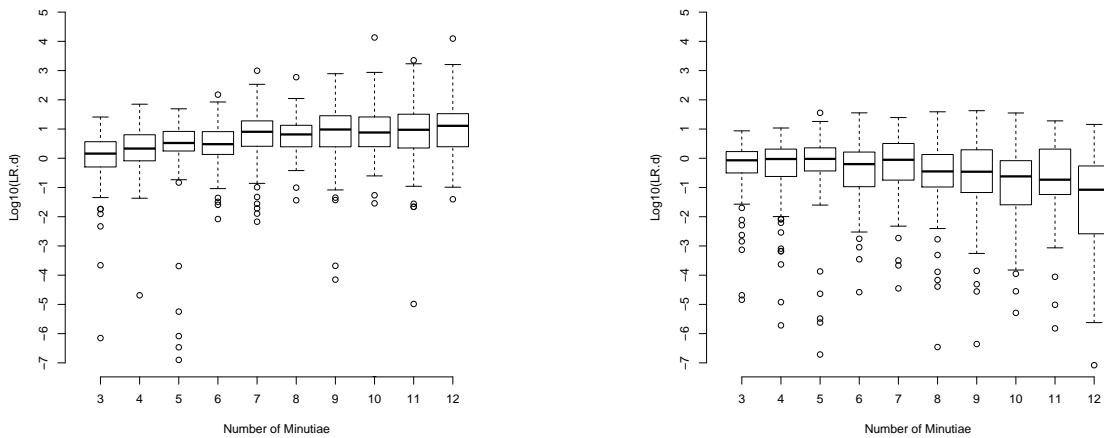


Figure 16b: Values for the LR.d of the direction component of the model – All regions - Left: same source control print – Right: difference source control print.

The slight increase in the LR.s computed for the shape component of the model for pairs of latent and control prints originating from different sources can be explained by the method used to select the control prints, since our method involved selecting the most similar control prints out of a reference dataset of 4,000,000 prints using a fingerprint matching algorithm relying heavily on configuration shapes. Nevertheless, we can observe that the increase is minimal compared to the increase in the expected value of LR.s for pairs of prints originating from the same source.

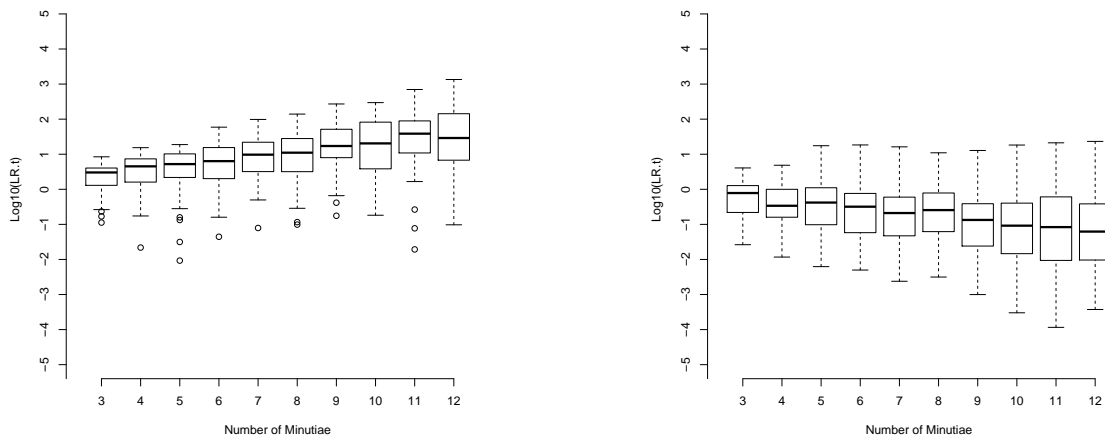


Figure 16c: Values for the LR_t of the type component of the model – All regions - Left: same source control print – Right: difference source control print.

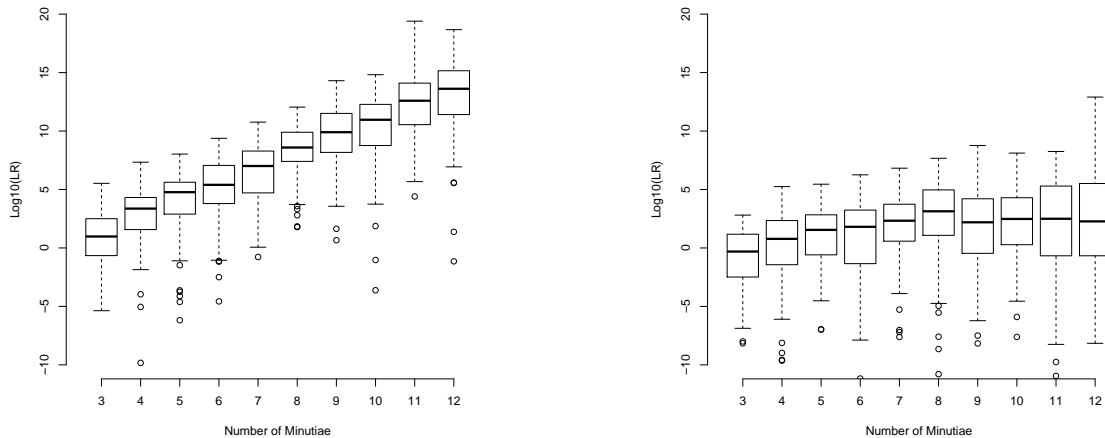


Figure 16d: Values for the LR_t – All regions - Left: same source control print – Right: difference source control print.

Figure 17a-c, Figure 18a-c, Figure 19a-c and Figure 20a-c present the denominators, numerators and LR_s obtained for the three components of the model, separately, and jointly for the configurations sampled in three different regions of fingerprint impressions: core, delta and peripheral regions. The behavior of the model for these specific configurations is similar to the behavior reported above.

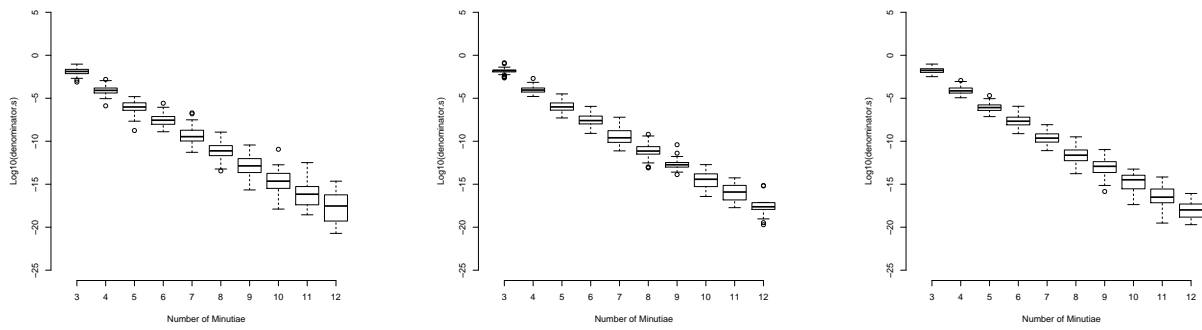


Figure 17a : Values for the denominators of the shape component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

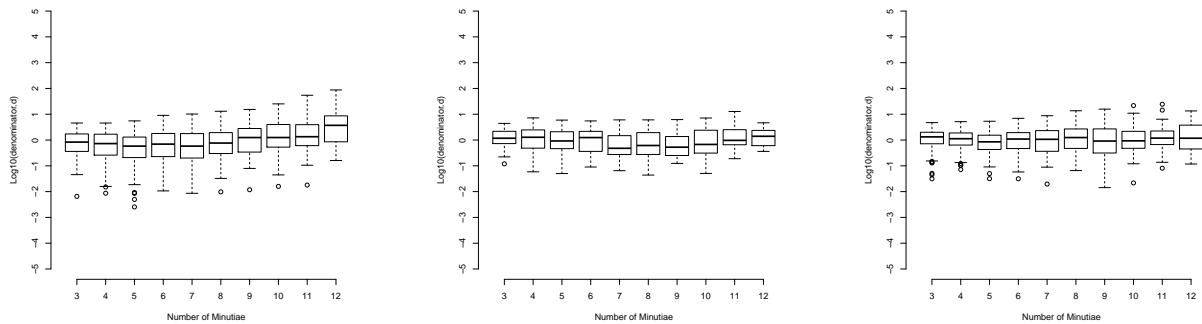


Figure 17b: Values for the denominators of the direction component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

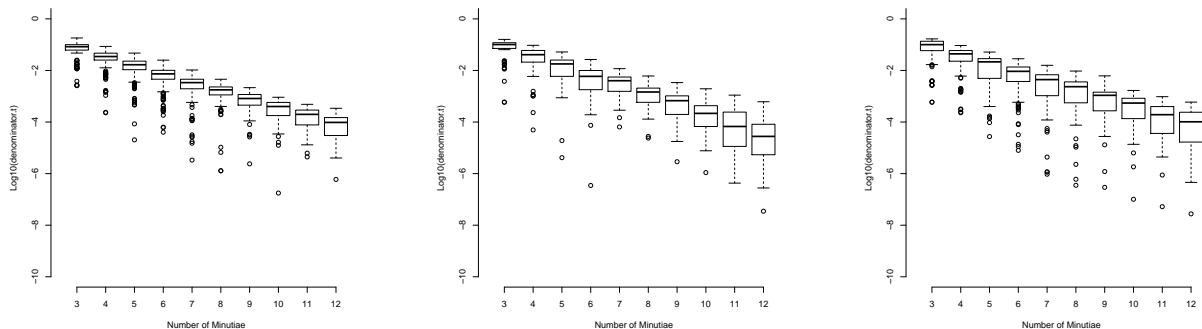


Figure 17c: Values for the denominators of the type component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

Interestingly, it seems that the expected weight of the evidence is not different between regions of the friction ridge skin. This seems counterintuitive: we were expecting the configurations in core and delta regions to be less discriminative than the configurations in the periphery of the prints (at least shape-wise), and therefore, have lower weight of

evidence. One explanation of this result may be that the model over-relies on the ability of the fingerprint matching algorithm to only retrieve k minutiae reference configurations in the same regions as the latent print configurations⁸. While the algorithm relies partly on the ridge flow detected on the image of the impression (and thus on the region), it may not be set up to completely ignore reference configurations in other regions than the targeted one if those configurations are similar enough. Thus, the algorithm may retrieve more reference configurations than it should, which, in turns, increases the spread of the probability densities assigned to the denominators, and eventually results in low denominators.

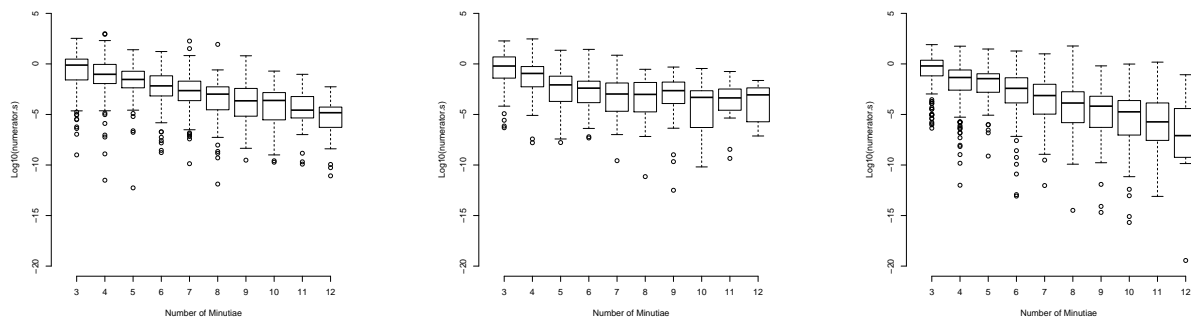


Figure 18a: Values for the numerators of the shape component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

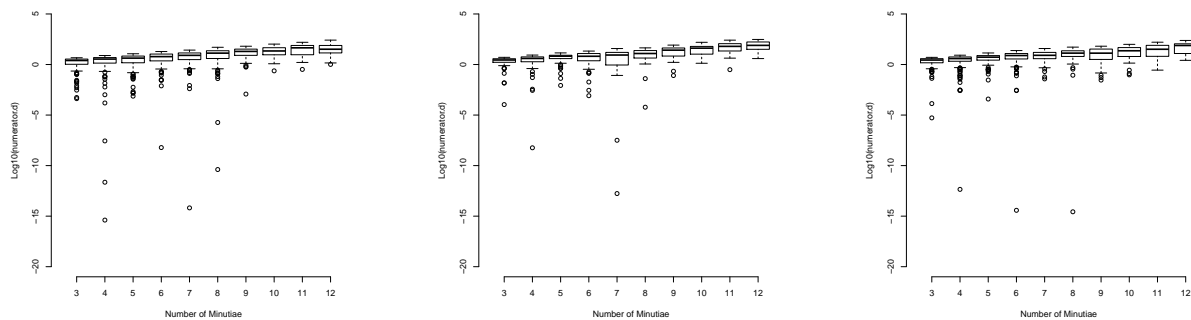


Figure 18b: Values for the numerators of the direction component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

⁸ This project uses a commercial grade algorithm and it is not possible to precisely understand its behavior.

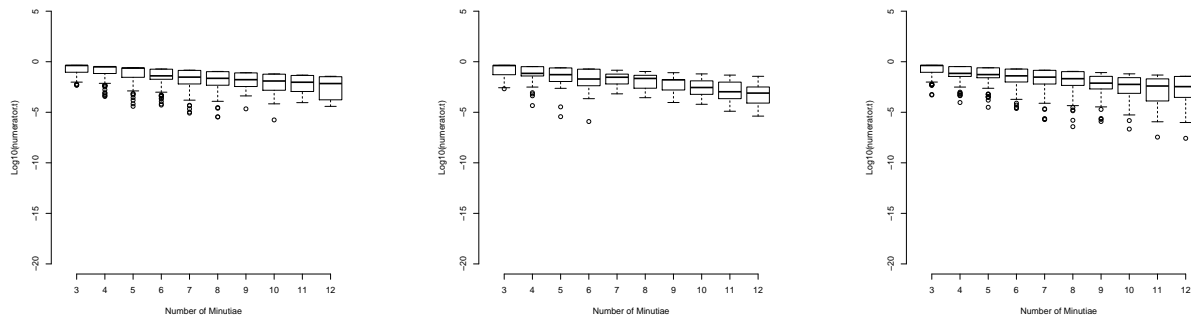


Figure 18c: Values for the numerators of the type component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

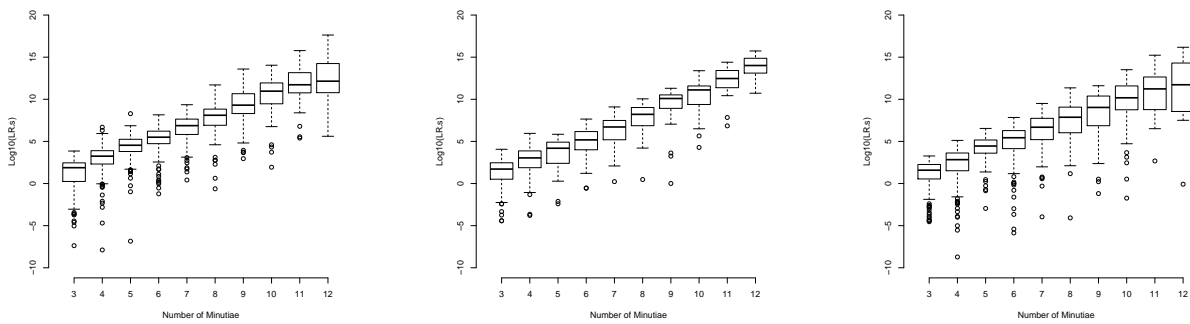


Figure 19a: Values for the LRs of the shape component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

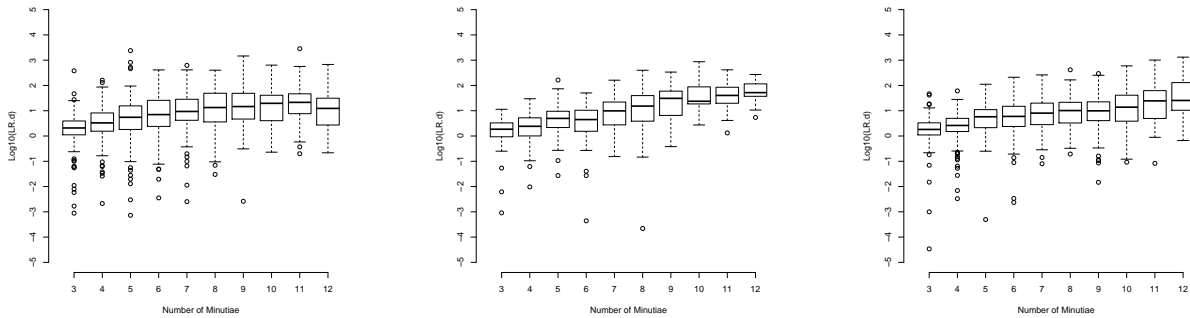


Figure 19b: Values for the LRs of the direction component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

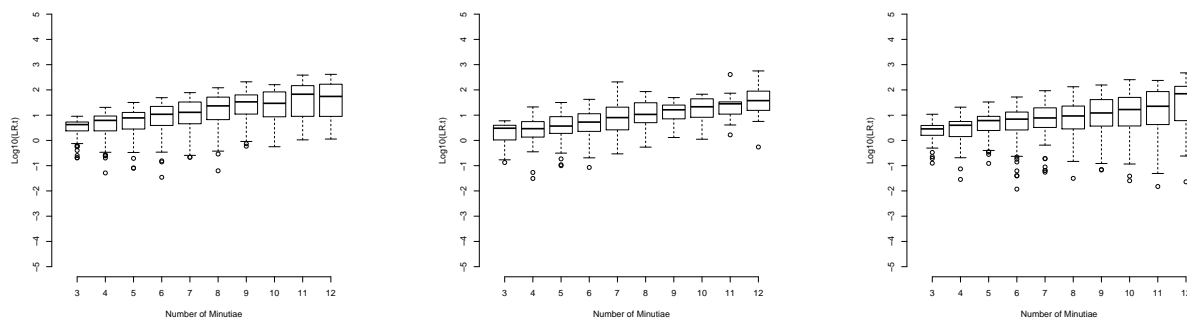


Figure 19c: Values for the LR_is of the type component of the model – Same source - Left: core region – Middle: delta region - Right: peripheral region

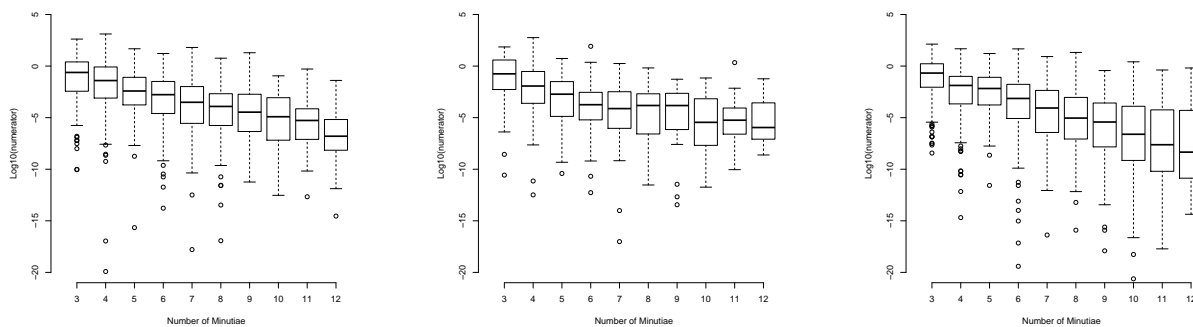


Figure 20a: Values for the numerators the model – All components - Same source - Left: core region – Middle: delta region - Right: peripheral region

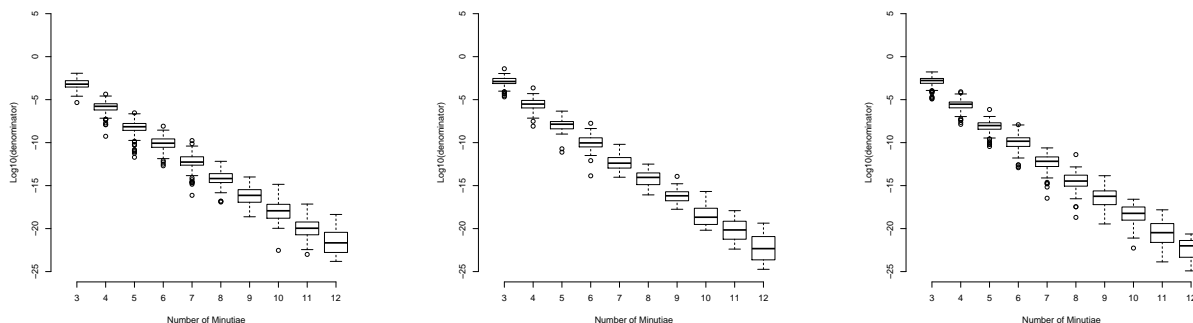


Figure 20b: Values for the denominators the model – All components - Same source - Left: core region – Middle: delta region - Right: peripheral region

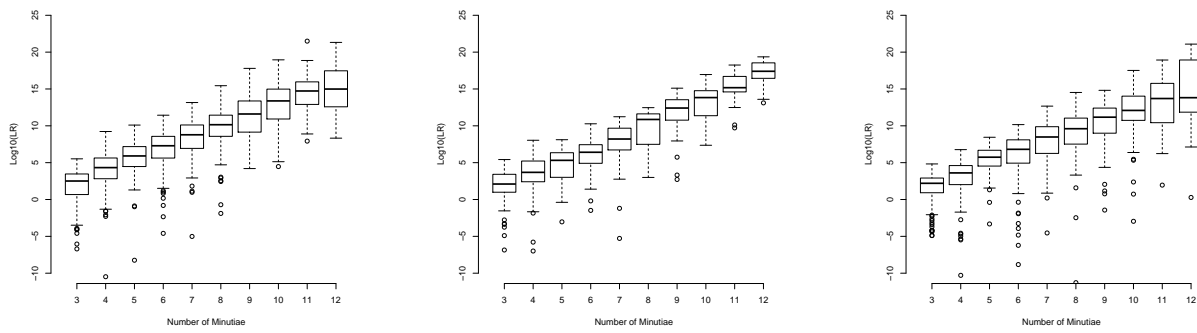


Figure 20c: Values for the LRs – All components – Same source - Left: core region – Middle: delta region - Right: peripheral region

Figure 21a-d present the values for $p_V(v|H_d)$ observed for the core (a), delta (b) and peripheral regions (c), and for all regions together (d). On the one hand it appears that the fingerprint matching algorithm is not entirely accounting for the region in which the latent print has been observed: the $p_V(v|H_d)$ values for the core and peripheral regions are very similar, while more variability between the configurations (i.e. lower $p_V(v|H_d)$ values) would be expected for the peripheral region.

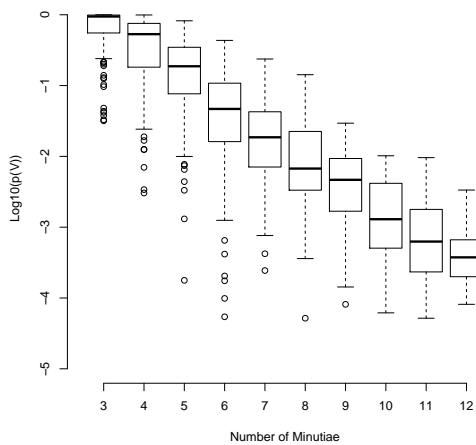


Figure 21a: Values for the $p_V(v|H_d)$ component of the model – Same source – Core region

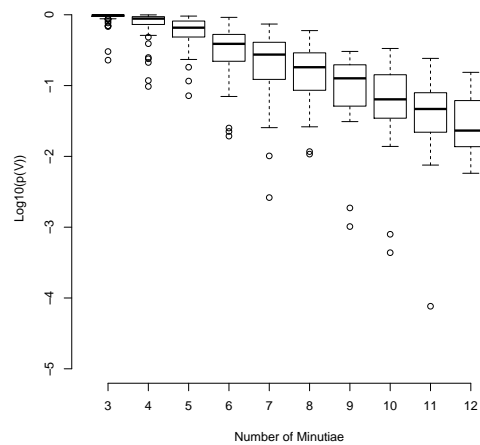


Figure 21b: Values for the $p_V(v|H_d)$ component of the model – Same source – Delta region

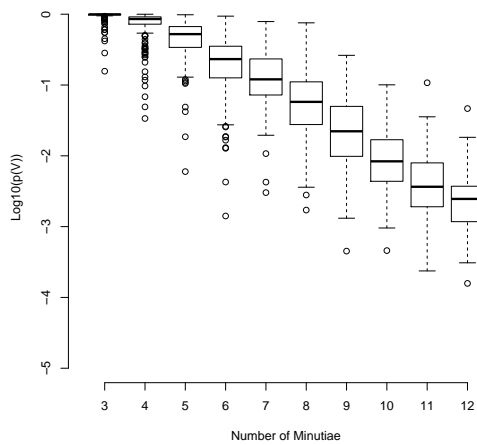


Figure 21c: Values for the $p_V(v|H_d)$ component of the model – Same source – Peripheral region

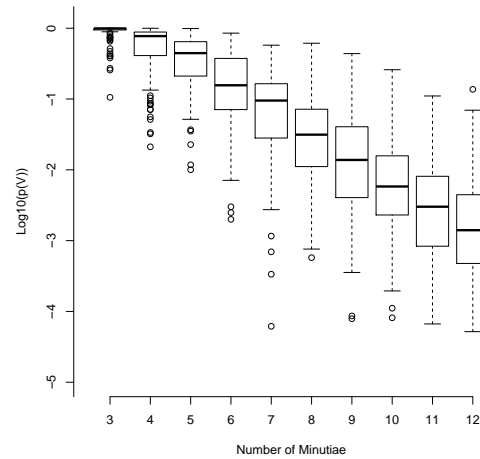


Figure 21d: Values for the $p_V(v|H_d)$ component of the model – Same source – All regions

On the other hand, the values for the delta region seems to indicate that configurations in this region are not very discriminative and that the matching algorithm is able to retrieve an abundant number of reference configurations for any given latent configuration. These results are confusing, especially in the light of the results presented in Figure 21a-d, which seems to indicate that the shape, minutiae direction and type of the configurations in the delta region are equally as discriminative as in the other regions.

Overall, the performance tests show that it is possible to design a model that provides support to the decisions made during all three phases of the examination process. In particular, it is possible to design a model that captures the spatial relationships between minutiae in any given configuration, and that can quantify the specificity of those configurations based on the number of minutiae, the shape of the configuration, the minutiae types and the minutiae directions. The observation of the $p_V(v|H_d)$ quantity also shows that a suitably configured commercial fingerprint matching algorithm can readily provide information on the suitability of latent prints during the Analysis phase, and support the forming of the conclusions during the Evaluation phase.

8. Descriptive statistics of the 15 trials

8.1. Descriptive statistics related to the examiners

146 examiners completed at least one trial. The respondents are 48 men and 98 women. Most of them are certified or active in casework (Table 9).

	0-5 years	6-10 years	11-15 years	16-20 years	20+ years
Man	18	7	8	1	14
Women	53	25	9	5	6
Total	71	32	17	6	20

All Examiners	LPE certified	LPE active	LPE nonactive	Trainee	Other
Man	19	15	2	11	1
Women	44	37	2	11	4
Total	63	52	4	22	5

Approach #1 (VID)	LPE certified	LPE active	LPE nonactive	Trainee	Other
Man	6	4	2	1	1
Women	11	7	1	3	1
Total	17	11	3	4	2

Approach #2 (VFC)	LPE certified	LPE active	LPE nonactive	Trainee	Other
Man	13	11	0	10	0
Women	33	30	1	8	3
Total	46	41	1	18	3

Use of Level 3 features	LPE certified	LPE active	LPE nonactive	Trainee	Other
Yes always	2	4	1	0	0
Yes often	23	15	2	12	3
Yes rarely	31	23	1	8	1
No	7	10	0	2	1
Total	63	52	4	22	5

Table 9: Summary of the 146 examiners enrolled in the study

Examiners' work experience (numbers of years of experience versus the number of hours per week) is presented in Figure 22 using four panels for the reported use of Level 3 features.

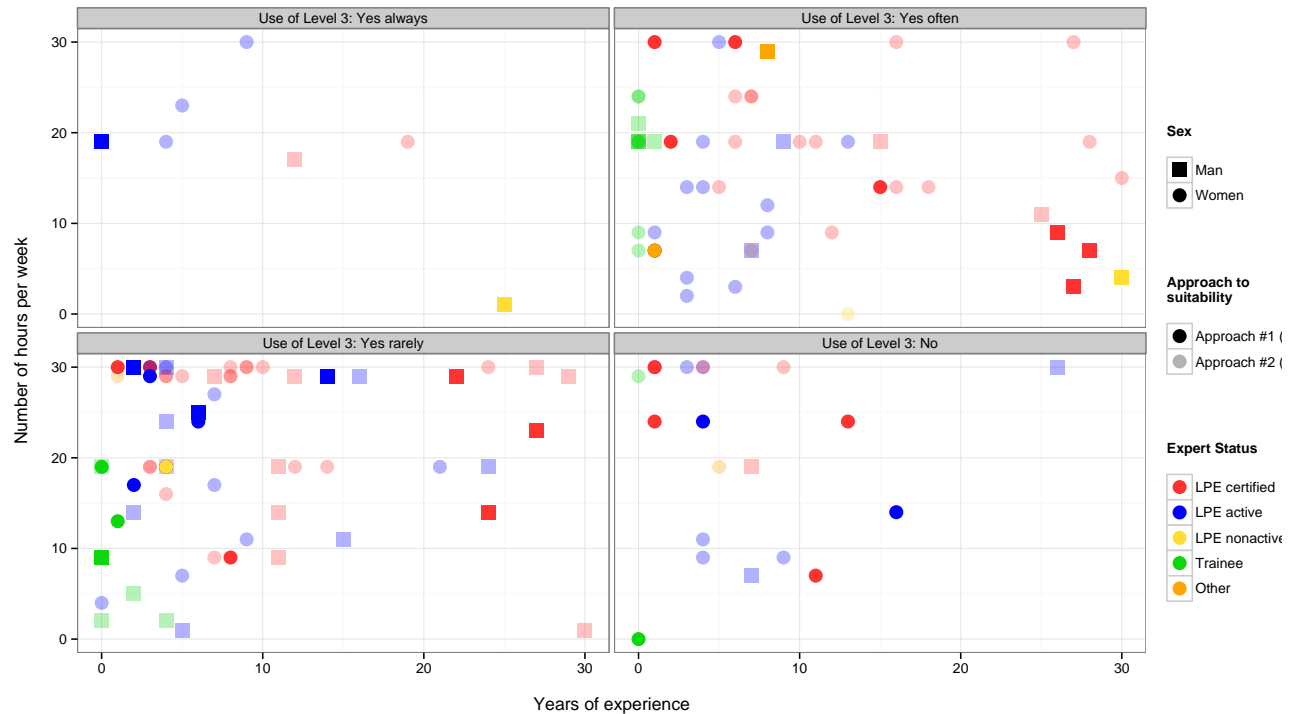


Figure 22: Characteristics of the examiners who participated to the study (N = 146).

8.2. Comfort and coherence levels of the participants with the interface

The observation of variables M1a, M1b and M1c show a large variability in the level of coherence between the participants in the study. For example, we can observe some examiners annotating minutiae with great confidence in areas of poor quality. This result is quite surprising. While we appreciate that some examiners are not entirely comfortable with onscreen examination of latent prints, we expected a better coherence between quality and quantity, as these concepts are commonly claimed in practice.

In addition, it was surprising to see the lack of documentation for a large part of the exclusion decision. This lack of documentation has prevented us to fully exploit the data to understand the concept of sufficiency when exclusion decisions are made. It may also show a lack of understanding and standardization, among the community, of the concept of exclusion.

8.3. Descriptive statistics of trials results

The decisions made by the participants following the **Analysis phase** are presented in Table 10, Figure 23 and Figure 24. Regardless of the approach adopted by the examiners, most of the latent prints led to split decisions between examiners, hence offering the potential for exploring underpinning factors that may explain such variation.

Conclusions reached after the Analysis phase all examiners																
	Trial01	Trial02	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09	Trial10	Trial11	Trial12	Trial13	Trial14	Trial15	Total
VID	126	137	103	13	46	33	105	75	114	121	34	115	95	109	36	1262
VEO	13	1	24	54	44	34	19	34	7	1	18	8	17	7	21	302
NV	6	0	10	69	44	66	7	19	6	3	73	1	10	7	66	387
Total	145	138	137	136	134	133	131	128	127	125	125	124	122	123	123	1951

Conclusions reached after the Analysis phase, examiners under Approach #1																
	Trial01	Trial02	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09	Trial10	Trial11	Trial12	Trial13	Trial14	Trial15	Total
VID	34	32	27	5	12	10	23	14	22	26	7	26	19	22	9	288
NV	3	0	5	27	18	19	5	13	4	0	20	0	6	4	17	141
Total	37	32	32	32	30	29	28	27	26	26	27	26	25	26	26	429

Conclusions reached after the Analysis phase, examiners under Approach #2																
	Trial01	Trial02	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09	Trial10	Trial11	Trial12	Trial13	Trial14	Trial15	Total
VID	92	105	76	8	34	23	82	61	92	95	27	89	76	87	27	974
VEO	13	1	24	54	44	34	19	34	7	1	18	8	17	7	21	302
NV	3	0	5	42	26	47	2	6	2	3	53	1	4	3	49	246
Total	108	106	105	104	104	104	103	101	101	99	98	98	97	97	97	1522

Table 10: Reported outcomes following the Analysis phase. Outcomes are: VID (value for identification) VEO (value for exclusion only) and NV (no value).

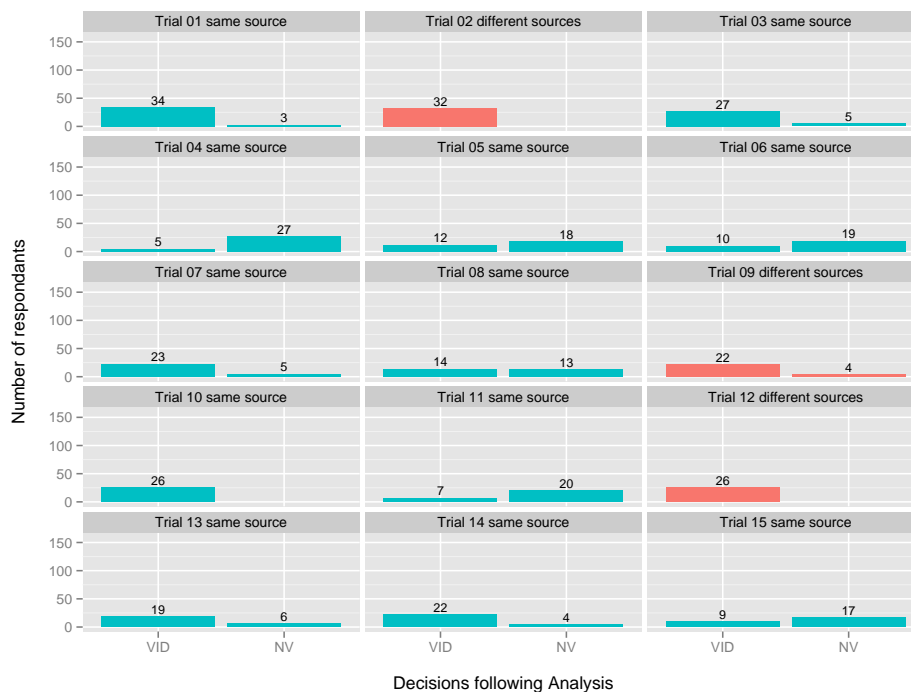


Figure 23: Reported conclusions following the Analysis phase for each trial for examiners using Approach #1. Outcomes are: VID (value for identification) and NV (no value). The results for same source comparisons are presented in aqua; the results for different sources comparisons are presented in red.

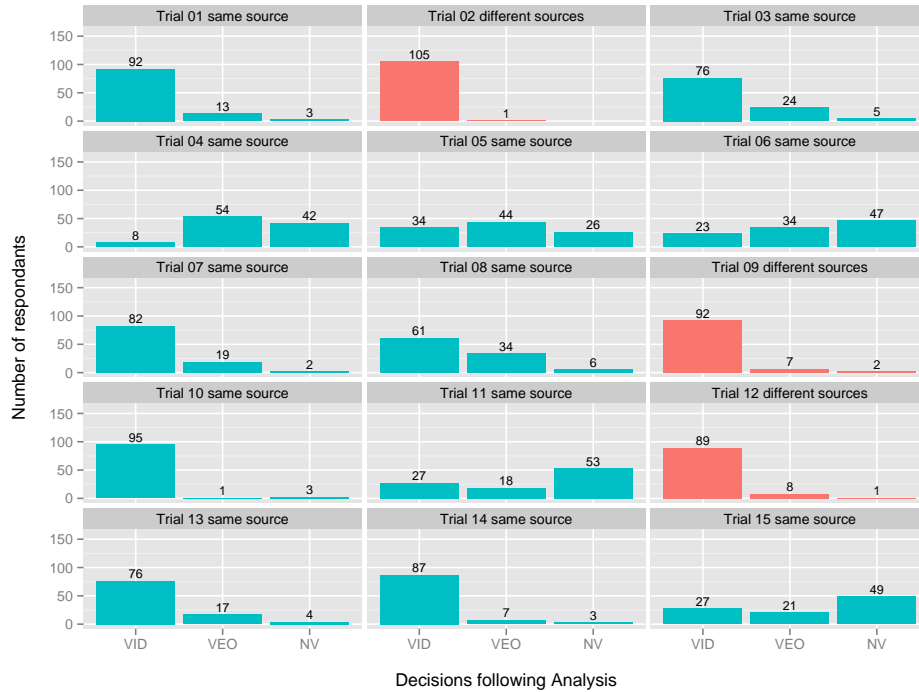


Figure 24: Reported conclusions following the Analysis phase for each trial for examiners using Approach #2. Outcomes are VID (value for identification) VEO (value for exclusion only) and NV (no value). The results for same source comparisons are presented in aqua; the results for different sources comparisons are presented in red.

The latent prints were characterized by the examiners with regards to the visibility of the three levels of details (L1, L2 and L3) as shown in Figure 25.

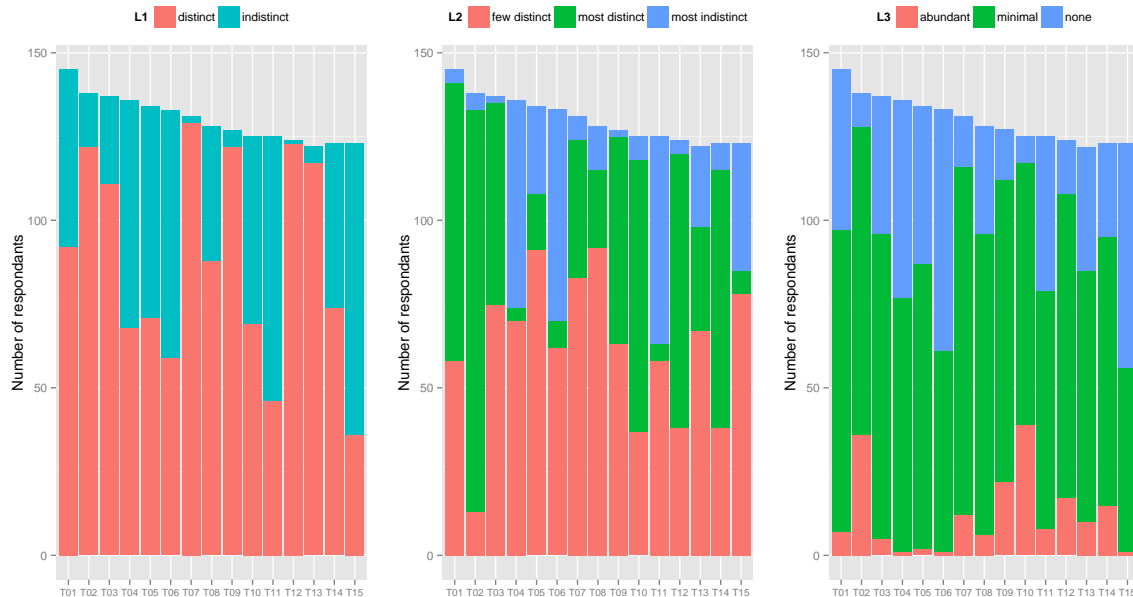


Figure 25: Characterization of the latent prints in relation to the visibility of the each level of details (L1, L2 and L3) as reported at the end of the Analysis phase.

Figure 26 shows the distribution of the general degradation aspects associated with each latent print of the trials (multiple selections allowed for each examiner). We can observe the significant variability in the perception of the presence/absence of factors potentially affecting the examination across all participants for any given trial.

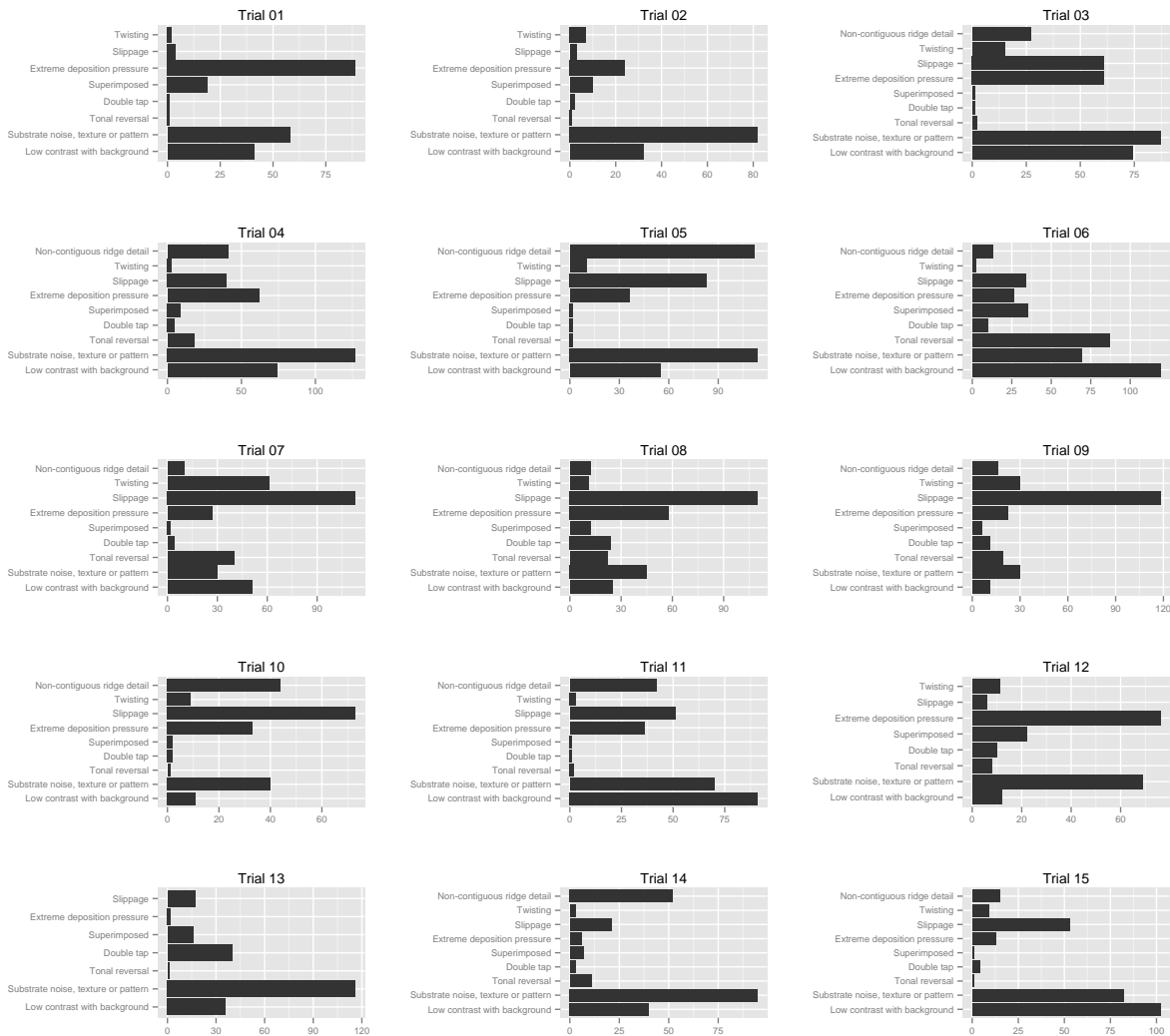


Figure 26: Characterization of the latent prints in relation to the general degradation aspects reported at the end of the Analysis phase.

Finally, Figure 27 shows the relationship between the perceived quality of the latent prints, the quantity of minutiae observed and the decision reached during the Analysis phase. In other words, Figure 27 present the data in a way that corresponds to the claims of the fingerprint community that conclusions are reached based on the quantity/quality of features observed on the prints, following Ashbaugh [5] and the latest SWGFAST standard

[3]. A large variability in the reported quality judgments and number of minutiae is observed between examiners for all trials. In addition, we can also observe differences in the decisions made by examiners perceiving similar quality and quantity of information. The variability can fairly be explained by the lack of consistency in the definition and understanding of these concepts within the fingerprint community.

Nevertheless, it can be seen that, overall, both the quality and quantity of information observed on the latent prints are driving the decision-making process – since there are more green data points in the upper right corner.

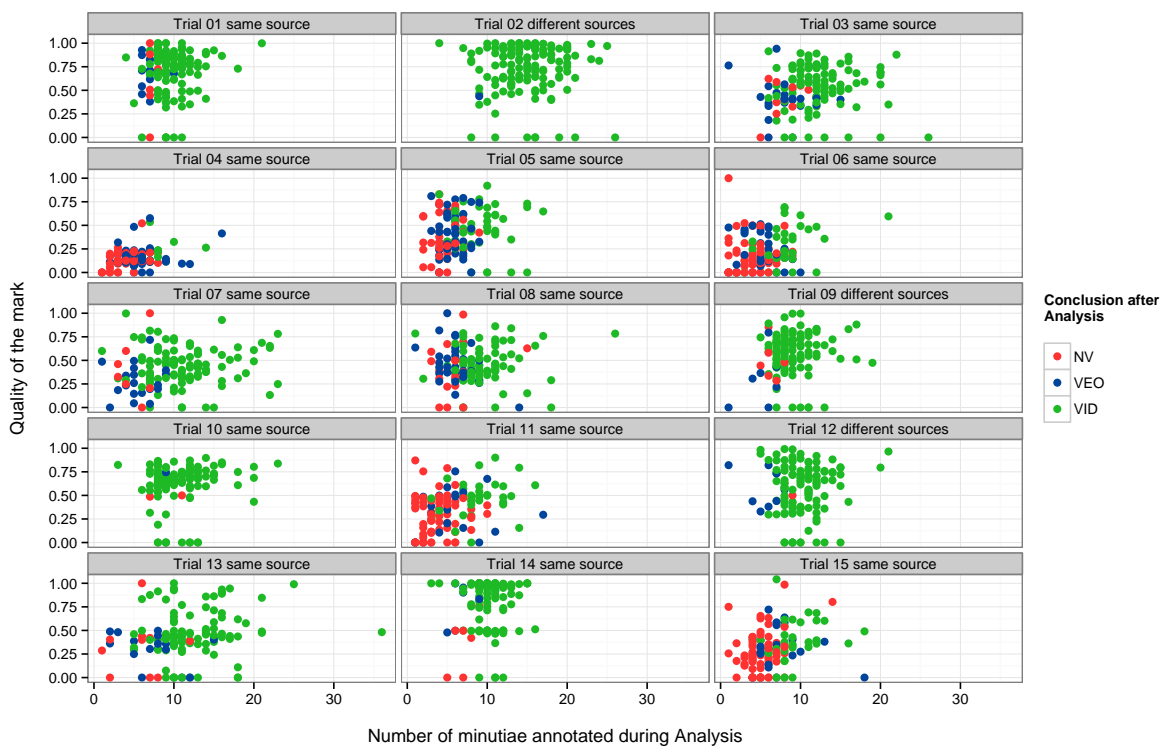


Figure 27: Reported conclusions following the Analysis phase for each trial given the number of annotated minutiae (x-axis) and the quality of the latent print (y-axis).

The conclusions reported after the **Comparison and Evaluation phases** are presented in Table 11 and Figure 28. Paralleling the observations made on the variability of the conclusions reached during the Analysis phase, we see (Figure 28) that examiners diverge substantially on the conclusions reached for these trials.

Conclusions reached after the Comparison phase

	Trial01	Trial02*	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09*	Trial10	Trial11	Trial12*	Trial13	Trial14	Trial15	Total
ID	120	0	120	7	34	29	73	43	2	100	13	11	76	72	28	728
INC	21	11	17	126	93	98	45	66	19	21	103	40	38	33	90	821
EXC	4	127	0	3	7	6	13	19	106	4	9	73	8	18	5	402
Total	145	138	137	136	134	133	131	128	127	125	125	124	122	123	123	1951

Nature of the INCONCLUSIVE decision

	Trial01	Trial02*	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09*	Trial10	Trial11	Trial12*	Trial13	Trial14	Trial15	Total
TOWARDS ID	14	2	11	52	42	37	24	25	8	12	22	18	19	24	26	336
TOWARDS EXC	1	5	0	4	7	6	3	7	8	2	4	10	4	4	1	66
NEUTRAL	6	4	6	70	44	55	18	34	3	7	77	12	15	5	63	419
Total	21	11	17	126	93	98	45	66	19	21	103	40	38	33	90	821

Table 11: Reported conclusions following the Comparison phase. Outcomes are: ID (Identification / Individualization), EXC (Exclusion) and INC (Inconclusive). The star (*) indicates for that trial that the latent and control prints originated from different sources.

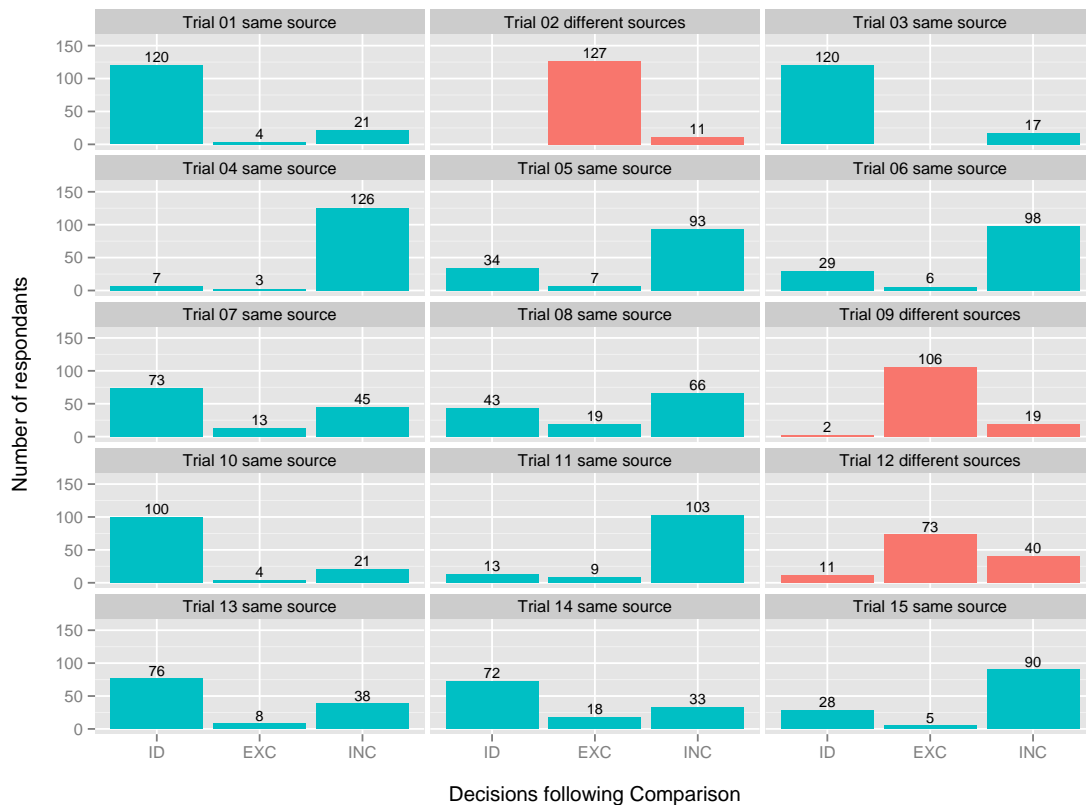


Figure 28: Reported conclusions following the Comparison phase for each trial. Outcomes are: ID (Identification / Individualization), EXC (Exclusion) and INC (Inconclusive). The results for same source comparisons are presented in aqua; the results for different sources comparisons are presented in red.

As explained in section 5.3, examiners, who reported an inconclusive conclusion (INC), were invited to position themselves as to the level of support (if any) towards identification (ID) or exclusion (EXC).

The results with this additional level of detail are given in Figure 29, which also shows the large variability between the conclusions reached by the participants for any given trial. A conclusion is considered to be misleading⁹ when the examiners guided (either categorically or by degree) against the ground truth. Among 146 examiners, 79 provided at least one misleading conclusion, whereas 67 provided all their conclusions in agreement with the ground truth. All cases of misleading conclusions are detailed in Table 12.



Figure 29: Reported conclusions following the Comparison phase for each trial. The x-axis translates the confidence attached by the examiner to the conclusion (ID and EXC being considered as categorical). The y-axis translates the reliability of the conclusion compared to the ground truth.

⁹ Here the word “misleading” is used literally for the purpose of this report. We realize that examiners currently do not testify as to which way they are leaning and that therefore they would not have misled anyone should they have reported these comparisons.

Examiner	Erroneous IDENTIFICATION	Misleading towards ID	Erroneous EXCLUSION	Misleading towards EXC	Total number of misleading cases	Number of trials finished	Examiner's declared status
user228	2	1	0	0	3	15	Trainee
user450	1	1	1	0	3	15	Other
user471	1	1	0	0	2	15	LPE certified
user425	0	2	1	1	4	15	Trainee
user334	1	0	1	0	2	15	LPE active
user384	1	0	1	0	2	15	LPE certified
user436	1	0	0	1	2	15	LPE certified
user024	1	0	0	0	1	15	LPE certified
user214	1	0	0	0	1	15	Trainee
user312	1	0	0	0	1	15	LPE certified
user342	1	0	0	0	1	15	LPE active
user373	1	0	0	0	1	15	Other
user395	0	2	0	0	2	15	LPE active
user481	1	0	0	0	1	15	LPE certified
user158	0	1	3	0	4	15	LPE active
user370	0	1	2	0	3	15	LPE active
user255	0	1	1	0	2	15	LPE active
user332	0	1	1	0	2	15	Trainee
user361	0	1	1	0	2	15	LPE active
user388	0	1	1	0	2	15	LPE active
user454	0	1	1	0	2	15	LPE active
user056	0	1	0	1	2	15	LPE certified
user393	0	0	5	0	5	15	LPE certified
user119	0	1	0	0	1	15	LPE active
user150	0	1	0	0	1	15	LPE active
user213	0	1	0	0	1	14	LPE active
user309	0	1	0	0	1	15	LPE certified
user328	0	1	0	0	1	15	LPE active
user378	0	1	0	0	1	15	LPE certified
user379	0	1	0	0	1	15	LPE active
user393	0	0	5	0	5	15	LPE certified
user399	0	1	0	0	1	15	LPE certified
user401	0	1	0	0	1	2	LPE active
user412	0	1	0	0	1	15	LPE certified
user431	0	1	0	0	1	15	LPE active
user458	0	1	0	0	1	15	LPE active
user466	0	1	0	0	1	15	Other
user324	0	0	3	1	4	15	Trainee
user344	0	0	3	0	3	15	LPE active
user348	0	0	3	0	3	15	LPE certified
user424	0	0	2	3	5	15	Trainee
user432	0	0	2	2	4	15	LPE active
user045	0	0	2	1	3	15	LPE certified
user248	0	0	2	1	3	15	LPE certified
user329	0	0	2	1	3	15	LPE certified
user027	0	0	2	0	2	10	LPE certified
user357	0	0	2	0	2	15	LPE certified
user404	0	0	2	0	2	15	LPE certified
user421	0	0	2	0	2	15	Trainee
user440	0	0	2	0	2	15	LPE certified
user161	0	0	1	2	3	7	LPE active
user128	0	0	1	1	2	10	LPE active
user306	0	0	1	1	2	15	LPE active
user372	0	0	1	1	2	15	LPE active
user430	0	0	1	1	2	15	LPE active
user118	0	0	1	0	1	15	Trainee
user143	0	0	1	0	1	9	Trainee
user151	0	0	1	0	1	8	Trainee
user184	0	0	1	0	1	15	Trainee
user278	0	0	1	0	1	15	LPE certified
user341	0	0	1	0	1	15	LPE certified
user356	0	0	1	0	1	15	LPE certified
user363	0	0	1	0	1	15	LPE active
user371	0	0	1	0	1	15	LPE certified
user434	0	0	1	0	1	15	Trainee
user435	0	0	1	0	1	9	LPE certified
user484	0	0	1	0	1	15	LPE certified
user350	0	0	0	3	3	15	LPE certified
user351	0	0	0	2	2	15	LPE certified
user428	0	0	0	2	2	15	LPE certified
user053	0	0	0	1	1	15	Trainee
user209	0	0	0	1	1	15	LPE certified
user323	0	0	0	1	1	15	LPE certified
user354	0	0	0	1	1	15	LPE active
user368	0	0	0	1	1	15	LPE active
user382	0	0	0	1	1	15	LPE active
user439	0	0	0	1	1	15	Trainee
user442	0	0	0	1	1	15	Other
user474	0	0	0	1	1	15	LPE active

Table 12: List of the examiners who reported at least one misleading conclusions. They are ordered as follows: “Erroneous ID”, “Misleading towards ID”, “Erroneous EXC” and last “Misleading towards EXC”.

When combining the data on the experience of the participants presented in Figure 22 with the data on the conclusions reported for the trials by these examiners presented in Figure 29 and Table 12, we observe that there is no clear relationship between the reliability of an examiners' conclusions and experience or workload (Figure 30).

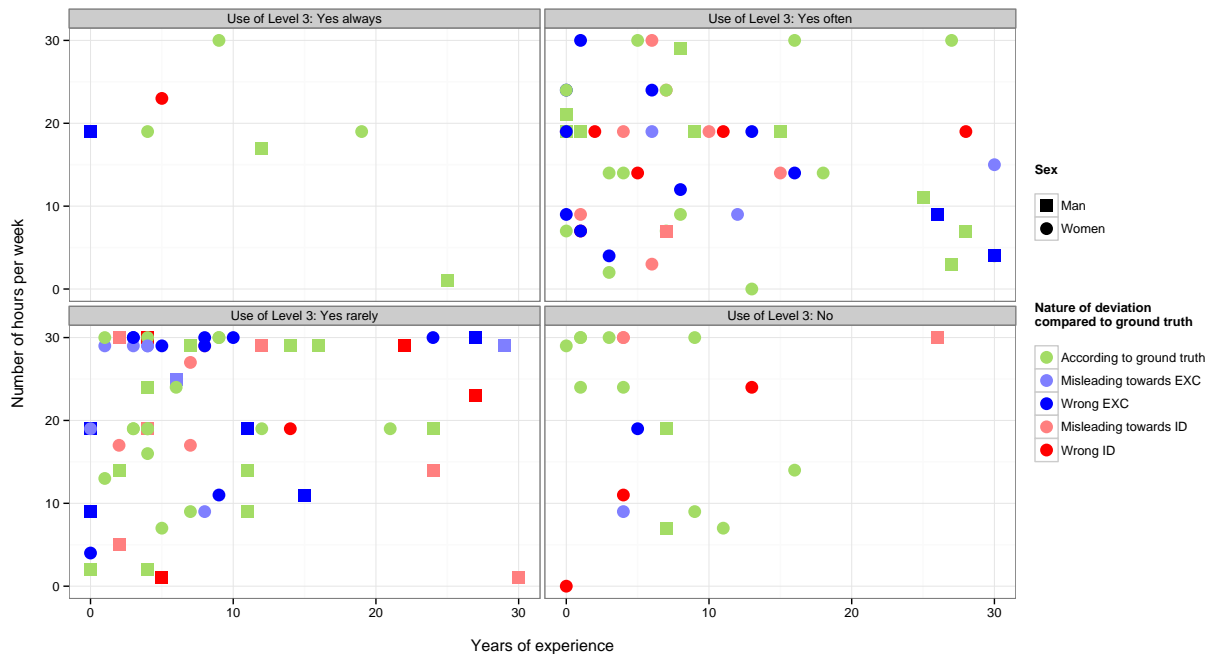


Figure 30: Reliability of the examiners who participated to the study (N = 146). Note that some examiners may have reported more than one misleading/or erroneous conclusion. For the display, priorities have been given in the following order: “Erroneous ID”, “Misleading towards ID”, “Erroneous EXC” and last “Misleading towards EXC”.

Figure 31 shows the relationship between the perceived quality of the latent prints, the quantity of corresponding minutiae observed on the latent and control prints, and the decision reached during the Evaluation phase (following SWGFAST [3] and Ashbaugh [5]). Similarly to the Analysis phase (Figure 27), a large variability in the observations and conclusions reported by the participants is observed for all trials. Evett and Williams [29] already observed such results. We note that for the same given pair of prints, some participants will see as few as no minutiae in common, while other can see up to 28. In terms for the conclusions reached by the participants, it is equally worthwhile to realize that some of them reached identification decisions with as low as 3 minutiae in common

between the latent and control prints, while some other remains indecisive with 10 or more minutiae (for the same pair of latent/control print)¹⁰.

Nevertheless, as mentioned before, we observe that, overall, **both** the **quality** and **quantity** of information in agreement impact the decision-making process.

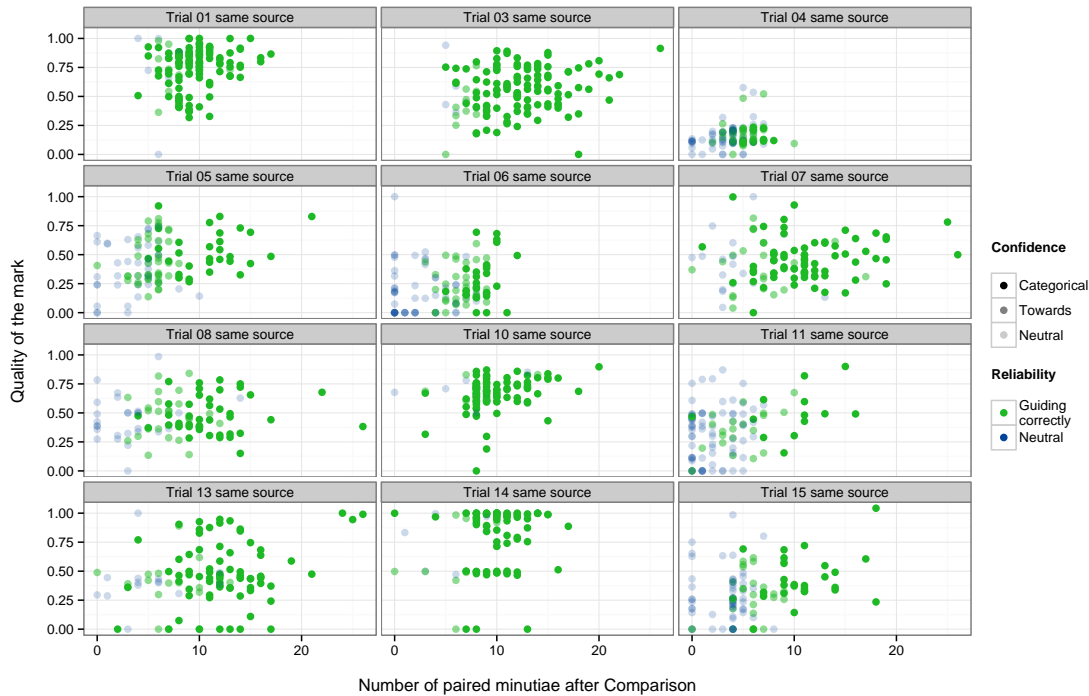


Figure 31: Reported conclusions following the Comparison phase for each same source comparison trial given the number of paired minutiae between the latent and control print (x-axis) and the quality of the latent print given during the Analysis phase (y-axis).

8.4. Descriptive statistics of the weight of evidence for the trial results

The statistical model described in section 0 was used to quantify the weight of the evidence for the data gathered during this project (section 5.1). The minutiae annotations provided by the participants for each trial were extracted from PiAnoS and used to extract minutiae configurations as described in section 7.2. Tables 13 and 14 present the number of configurations extracted from the study data and processed using the model.

¹⁰ The reader is reminded that the participants were explicitly asked to report all corresponding minutiae, irrespective of their use of additional features to reach their conclusion

No Value																											
# of minutiae	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	36	
Trial 01 same source	0	0	0	0	0	0	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 02 different sources	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 03 same source	0	0	0	0	1	1	3	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 04 same source	0	2	11	10	14	9	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 05 same source	0	0	2	10	8	5	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 06 same source	0	0	6	10	10	4	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 07 same source	0	0	2	2	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 08 same source	0	0	0	0	2	4	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Trial 09 different sources	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 10 same source	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 11 same source	2	2	0	4	4	4	1	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 12 different sources	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 13 same source	0	1	0	0	0	3	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 14 same source	0	0	0	0	1	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 15 same source	0	0	3	15	8	11	4	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Count per # of minutiae	2	5	24	51	48	44	33	16	7	2	2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Value for Exclusion																											
# of minutiae	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	36	
Trial 01 same source	0	0	0	0	0	5	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 02 different sources	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 03 same source	1	0	0	0	1	4	2	5	2	2	1	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Trial 04 same source	0	0	2	10	16	8	6	1	2	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Trial 05 same source	0	0	2	6	6	9	7	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 06 same source	0	1	3	1	6	7	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 07 same source	0	0	1	1	6	1	3	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 08 same source	0	0	0	4	4	7	4	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 09 different sources	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 10 same source	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 11 same source	0	0	1	1	2	3	2	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Trial 12 different sources	0	0	0	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 13 same source	0	1	1	0	2	1	1	3	2	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Trial 14 same source	0	0	0	0	1	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 15 same source	0	0	0	1	5	4	4	2	2	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Count per # of minutiae	1	2	10	24	50	51	40	24	20	7	4	5	1	0	2	1	1	1	1	0	0	0	0	0	0	0	0
Value for Identification																											
# of minutiae	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	36	
Trial 01 same source	0	0	0	1	1	1	5	24	32	16	20	8	4	4	1	1	0	1	0	0	1	0	0	0	0	0	0
Trial 02 different sources	0	0	0	1	0	0	1	2	4	3	2	5	6	2	2	5	6	4	0	5	2	0	1	1	0	0	0
Trial 03 same source	0	0	0	0	0	2	4	3	10	10	11	18	11	6	5	4	3	1	0	5	1	1	0	0	0	1	0
Trial 04 same source	0	0	0	0	0	0	6	4	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 05 same source	0	0	0	1	2	5	4	7	2	7	6	1	1	0	2	0	1	0	0	0	0	0	0	0	0	0	0
Trial 06 same source	0	0	1	2	0	2	5	10	4	2	1	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Trial 07 same source	1	0	1	2	0	6	9	7	6	14	8	9	9	4	1	5	1	3	1	2	1	3	2	0	0	0	0
Trial 08 same source	1	1	0	0	1	2	9	7	11	4	11	4	4	0	1	1	1	1	0	0	0	0	0	0	0	0	0
Trial 09 different sources	0	0	0	0	0	2	15	17	15	3	5	3	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0
Trial 10 same source	0	0	1	0	0	1	6	23	16	12	12	15	7	4	5	2	0	2	0	3	0	0	1	0	0	0	0
Trial 11 same source	0	0	1	2	0	0	4	1	5	2	2	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Trial 12 different sources	0	0	0	0	2	2	8	11	11	8	12	11	4	3	3	1	0	0	0	0	1	0	0	0	0	0	0
Trial 13 same source	0	0	0	0	1	2	2	3	8	17	11	7	2	8	8	7	3	4	1	0	3	0	0	0	1	0	1
Trial 14 same source	0	0	0	1	0	1	3	6	21	23	25	12	6	3	3	1	0	0	0	0	0	0	0	0	0	0	0
Trial 15 same source	0	0	0	1	1	4	2	7	6	1	3	5	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Count per # of minutiae	2	1	4	11	8	30	83	132	152	123	129	104	57	38	32	29	16	17	2	15	10	4	4	2	1	1	
Total Count	5	8	38	86	106	125	156	172	179	132	135	110	58	39	35	30	17	18	2	15	10	4	4	2	1	1	

Table 13 : Configurations of minutiae from the study processed using the model after the analysis phase, presented by number of minutiae, trial and decision reached.

The data are presented (in Figures 32 to 37) for each trial image, according to the decision reached at the end of the Analysis and Comparison phases. The data in Tables 13 and 14 correspond to the data graphically represented in Figures 27 and 29 in the previous sections.

Exclusion																										
# of minutiae	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	24	25	26			
Trial 01 same source	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 02 different sources	22	13	4	7	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 03 same source	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 04 same source	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 05 same source	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 06 same source	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 07 same source	1	2	4	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 08 same source	0	2	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 09 different sources	35	9	5	1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 10 same source	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 11 same source	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 12 different sources	7	7	10	6	8	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 13 same source	1	1	3	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 14 same source	3	1	3	3	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 15 same source	0	3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Count per # of minutiae	71	42	34	21	19	8	5	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0		

Inconclusive																										
# of minutiae	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	24	25	26			
Trial 01 same source	0	1	2	6	6	3	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 02 different sources	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 03 same source	0	0	3	5	4	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 04 same source	18	28	25	21	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 05 same source	7	13	20	23	4	3	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 06 same source	9	9	7	13	10	9	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 07 same source	6	10	2	6	7	2	2	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0		
Trial 08 same source	10	6	5	8	5	4	4	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 09 different sources	1	1	4	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 10 same source	1	0	1	1	4	7	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 11 same source	8	13	11	6	4	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 12 different sources	1	3	9	4	6	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 13 same source	3	5	3	6	2	3	2	4	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 14 same source	2	2	0	4	4	3	4	5	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 15 same source	2	31	15	10	5	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Count per # of minutiae	70	122	108	118	71	45	24	16	6	5	5	4	0	1	0	0	0	0	0	0	0	0	0	0		

Identification																										
# of minutiae	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	24	25	26			
Trial 01 same source	0	1	2	4	6	27	21	22	11	5	6	5	1	2	1	0	0	0	0	0	0	0	0	0		
Trial 02 different sources	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 03 same source	0	0	1	3	2	8	10	12	9	16	11	10	9	6	4	3	3	2	2	1	0	0	1	1		
Trial 04 same source	0	0	1	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 05 same source	0	0	1	3	1	5	2	0	5	5	2	2	2	0	1	0	0	0	1	0	0	0	0	0		
Trial 06 same source	0	0	1	0	4	8	8	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 07 same source	0	2	0	5	2	7	8	5	12	8	4	2	3	3	1	3	3	0	0	0	0	0	1	1		
Trial 08 same source	0	1	1	0	5	2	5	8	4	5	4	4	1	0	1	0	0	0	0	1	0	0	0	1		
Trial 09 different sources	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 10 same source	2	0	0	0	3	31	19	10	5	6	6	3	4	1	0	1	0	1	0	0	0	0	0	0		
Trial 11 same source	0	0	0	1	2	0	1	2	3	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0		
Trial 12 different sources	0	2	0	2	2	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
Trial 13 same source	1	1	0	0	2	5	5	6	7	12	6	8	6	7	2	0	1	0	1	0	1	1	1	1		
Trial 14 same source	0	1	0	0	5	11	9	16	9	8	5	3	2	1	1	0	0	0	0	0	0	0	0	0		
Trial 15 same source	0	1	1	1	2	1	5	3	5	0	2	3	0	0	1	2	0	0	0	0	0	0	0	0		
Count per # of minutiae	3	9	9	22	39	106	95	88	72	66	48	41	29	20	12	9	7	3	4	2	1	2	4	4		

Total Count	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	24	25	26	
Total Count	144	173	151	161	129	159	124	104	79	72	54	45	29	21	12	9	7	3	4	2	1	2	4	4

Table 14: Number of configurations from the study extracted from PiAnoS, processed using the model after the Comparison phase, presented by number of minutiae, trial and conclusion reached.

The specificity of the configurations of minutiae observed during the Analysis phase was computed using the denominator of the model. The results are reported in Figure 32 for the denominator of the model, and Figure 33 for $p_V(v|H_d)$ (i.e., the probability of finding a similar configuration in the reference database). Note that no result could be obtained for configurations of less than 3 minutiae (since the model is based on triangles) and for most configurations of more than 12-13 minutiae (since $p_V(v|H_d)$ tends to 0 for large k). Figure 32 and Figure 33 show that the specificity of the configurations increases with the number

of annotated features. Unsurprisingly, participants that have determined that the latent prints in the trials were of value for identification (VID) have reported more features. Those configurations lead to higher specificity values. The output of the model is clearly dependent on the information provided by users and, thus sensitive to their perception of the suitability of the latent prints. Nevertheless, it can be seen that the model indicates that some configurations, which were reported as having little to no value, contain valuable information (Figures 32 and 33, configurations reported VEO and NV with higher number of minutiae). Overall, the model proves useful to support the decision made by examiners during the Analysis phase.

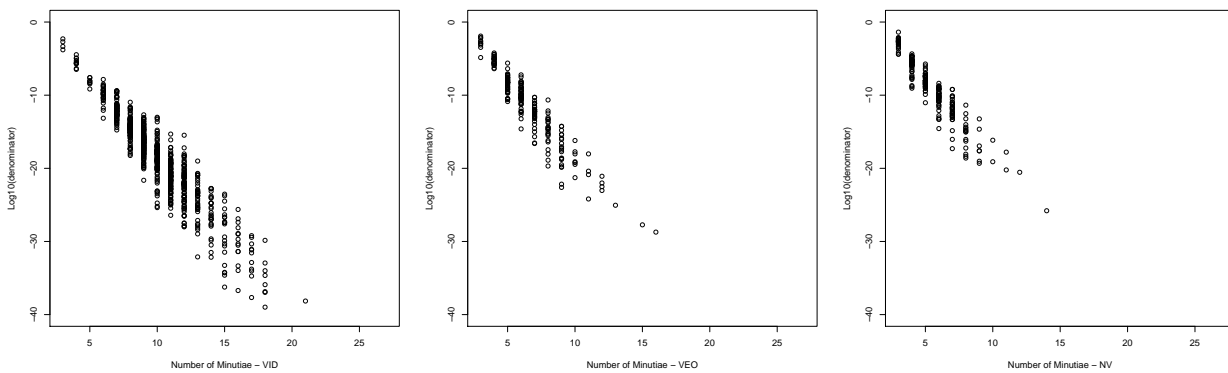


Figure 32: Values for the denominators of the model calculated for the study dataset based on the observations reported after the analysis stage of the examination process – Left: analyses leading to VID decisions – Middle: analyses leading to VEO decisions – right: analyses leading to NV decisions.

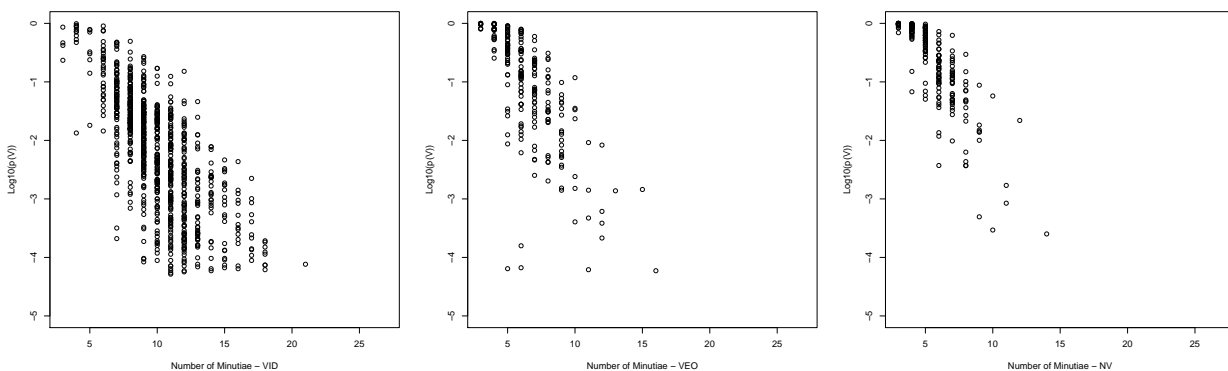


Figure 33: Values for the $p_V(v|H_d)$ components of the model calculated for the study dataset based on the observations reported after the analysis stage of the examination process – Left: analyses leading to VID decisions – Middle: analyses leading to VEO decisions – right: analyses leading to NV decisions.

The numerator of the model expresses the level of resemblance between pairs of latent and control prints. Figure 34 presents the results calculated based on the paired minutiae annotated by the participants in the study. As for the annotations made during the Analysis phase, we can see that the participants reported more paired minutiae for the comparisons that resulted in identification conclusions, or that were deemed inconclusive, than in exclusion decisions. We also observe that while the majority of the numerator values are relatively high (say above 10^{-10}), a significant proportion of data points indicate a very low resemblance between the latent and control prints. This low resemblance can be due to high distortion effects that are beyond the capability of the distortion model used to create pseudo-traces (section 7.3), but may also indicate that the latent and control prints are not from the same source. In any case, such low numerator values can be used as a warning signal, which could help examiners during their assessment of the similarity/dissimilarity between pairs of latent and control prints.

Figures 35 and 36 present the data on the specificity of the k configurations annotated on the latent prints by the participants during the Comparison phase. Similarly to the Analysis phase, these data are based on the denominator of the model and on $p_V(v|H_d)$. The similar behavior of the data for all three possible conclusions is explained by the fact that configuration' specificity is only dependent on the information observed on the latent print, and not on the resemblance with the control print. Overall, the same observations, already made previously for Figures 32 and 33 (Analysis) and in section 7.7 can be made in this case.

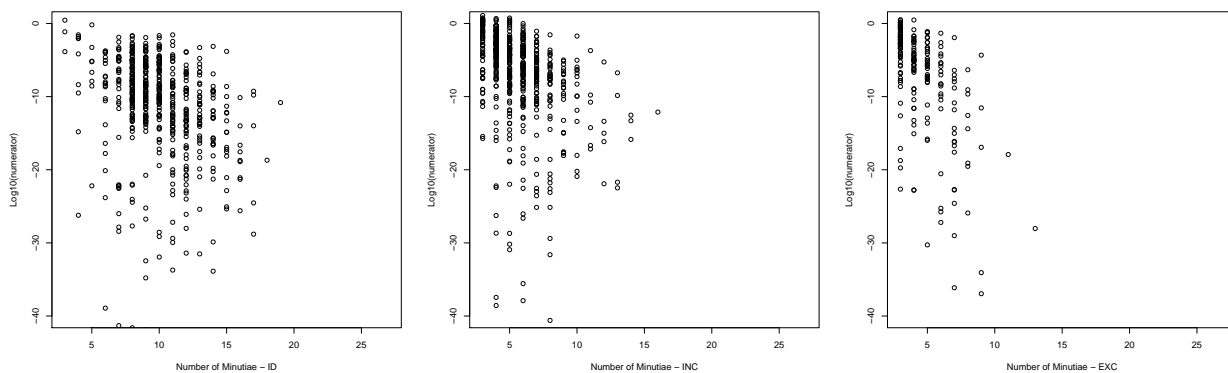


Figure 34: Values for the numerators of the model calculated for the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions –

Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

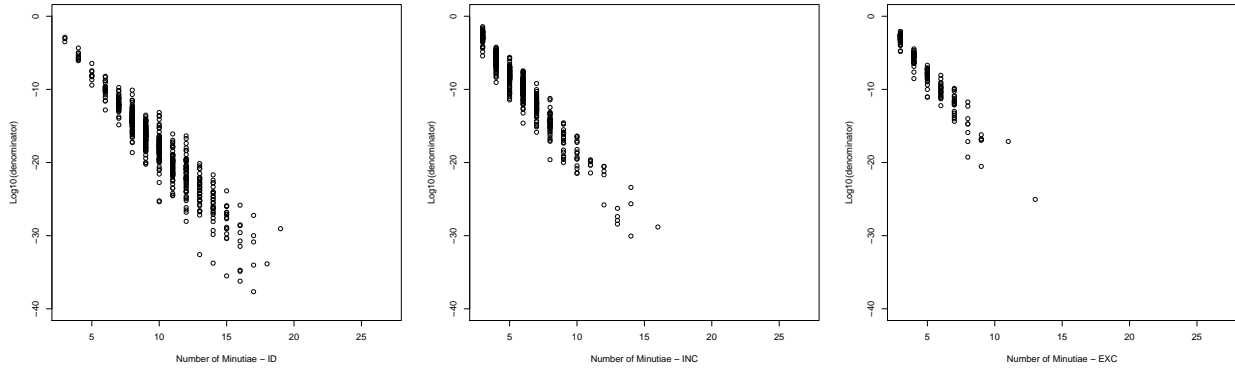


Figure 35: Values for the denominators of the model calculated for the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

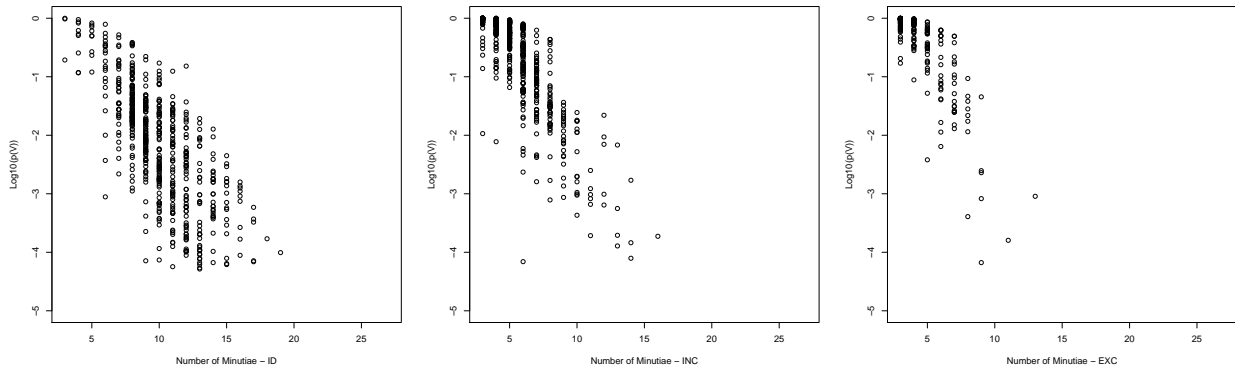


Figure 36: Values for the $p_V(v|H_d)$ components of the model calculated for the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

Finally, Figure 37 present the values calculated for the LR's of the comparisons performed by the participants of the study.

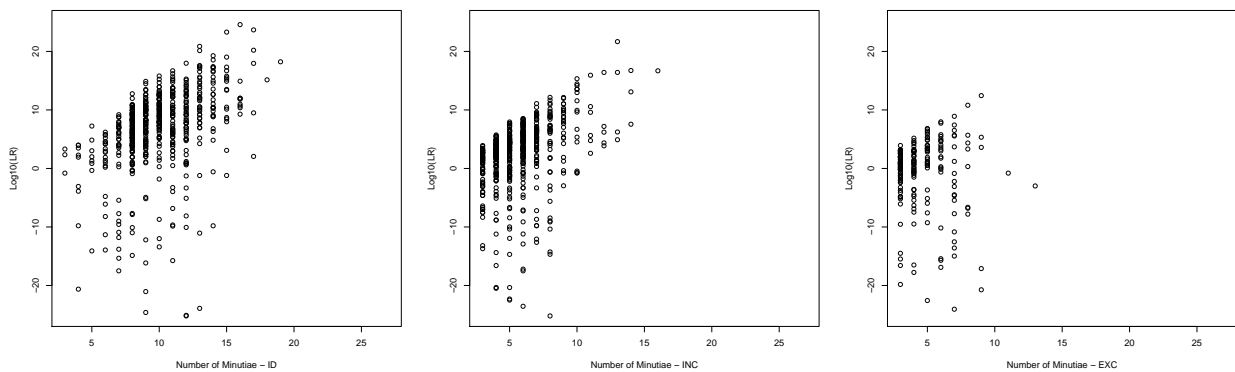


Figure 37: values for the LR's calculated for the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC

decisions – right: comparisons leading to EXC decisions.

It is interesting, but not surprising given the design of the study, to observe in Figure 37 that relatively high LR's are calculated for the majority of comparisons resulting in exclusion decisions. This result can be explained by two main elements:

1. The model is currently not taking into account differences. It only relies on annotated pairs of corresponding minutiae. This is not necessarily a concern: the model is only designed to support examiners' decisions, therefore the different statistics calculated needs to be weighted by the other observations made during the examination.
2. The trials containing pairs of latent and control prints originating from different sources were specifically designed to have prints as similar as possible. The results presented in Figure 37 indicate that it is possible to find pairs of latent and control prints originating from different sources that are reasonably similar up to a certain point.

The following figures present an example of the same data reported above for all participants but for a single trial (Trial 01 – same source). It is interesting to observe the variability between the results calculated for the different participants when examining the same prints. These data, and the results reported in the previous sections, support the call for better definition of what constitutes a feature, and for accurately and exhaustively documenting observations.

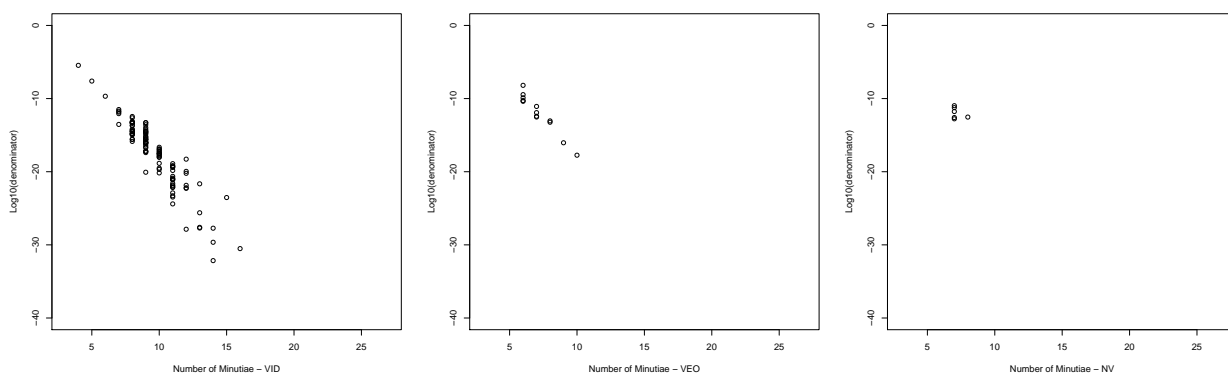


Figure 38a: values for the denominators of the model calculated for **Trial 01** of the study dataset based on the observations reported after the analysis stage of the examination process – Left: analyses leading to VID decisions – Middle: analyses leading to VEO decisions – right: analyses leading to NV decisions.

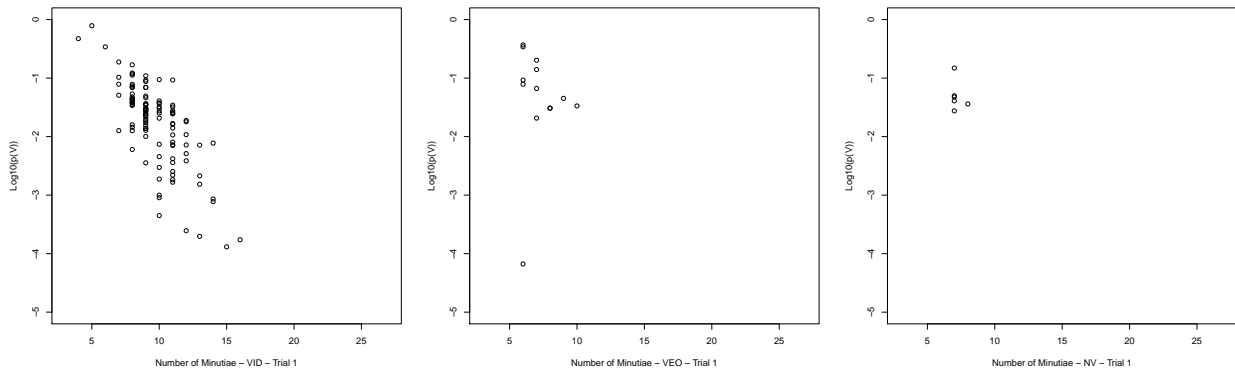


Figure 38b: values for the $p_V(v|H_d)$ components of the model calculated for **Trial 01** of the study dataset based on the observations reported after the analysis stage of the examination process – Left: analyses leading to VID decisions – Middle: analyses leading to VEO decisions – right: analyses leading to NV decisions.

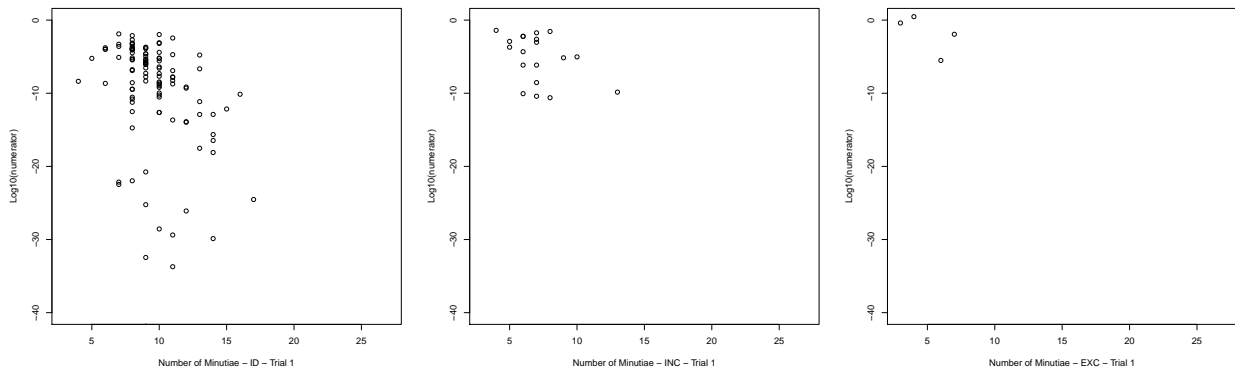


Figure 38c: values for the numerators of the model calculated for **Trial 01** of the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

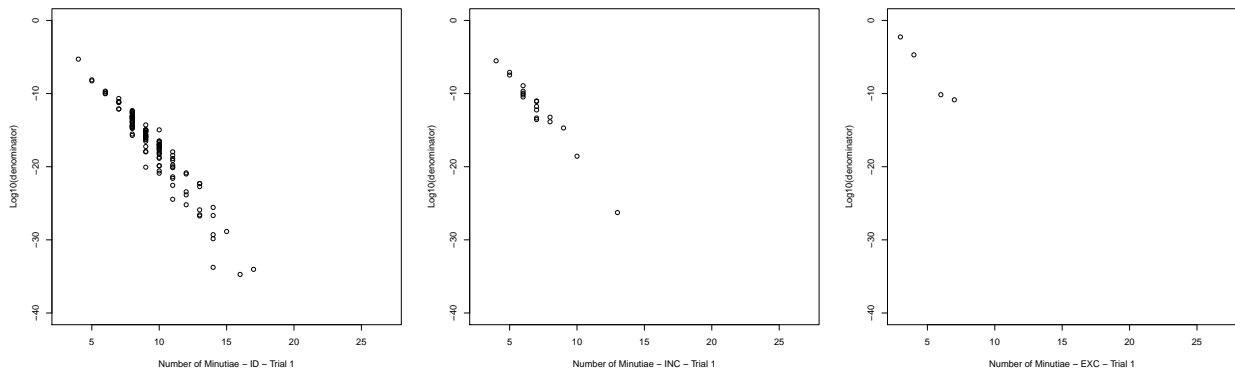


Figure 38d: values for the denominators of the model calculated for **Trial 01** of the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions

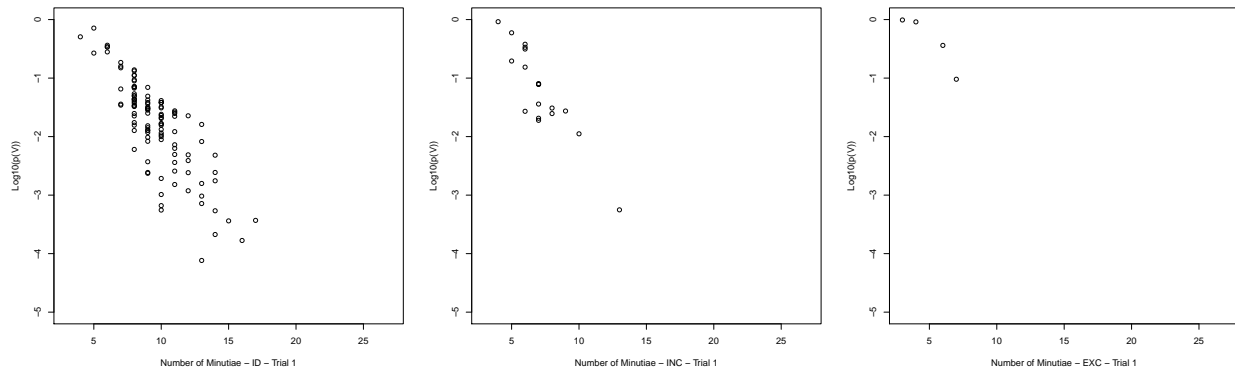


Figure 38e: values for the $p_V(v|H_d)$ components of the model calculated for **Trial 01** of the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

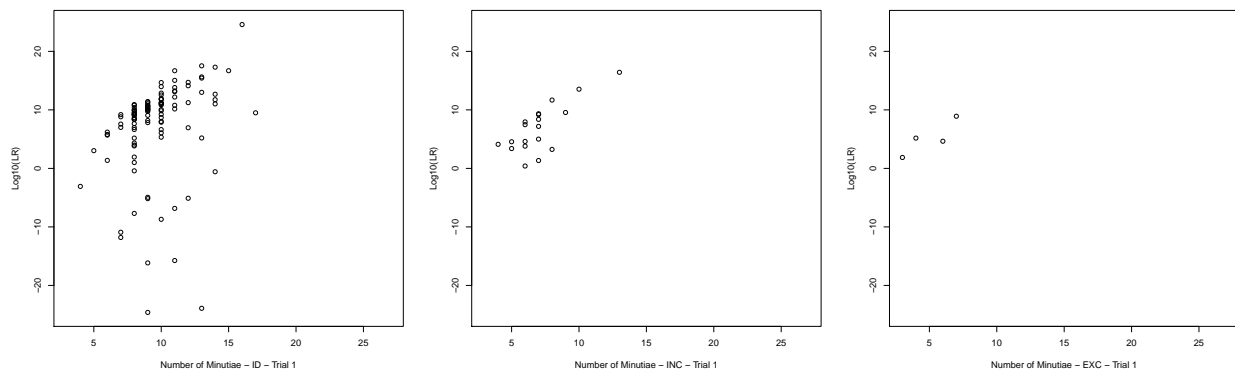


Figure 38f: values for the LRs calculated for **Trial 01** of the study dataset based on the observations reported after the comparison stage of the examination process – Left: comparisons leading to ID decisions – Middle: comparisons leading to INC decisions – right: comparisons leading to EXC decisions.

9. Relationships between participants' annotations and sufficiency

The objectives of this section are to explore:

1. If a relationship can be established between the observations and annotations provided by the participants to document their examinations of the trial images, and the decisions made during the different phases of the process;
2. The influence of demographic data, such as age, gender, personal training and experience, on these decisions;
3. If a relationship can be established between the consensus variables (sections 6.1 and 6.2) and the misleading conclusions reported by the examiners; and if it could be possible to predict misleading conclusions from divergent observations between individuals examining a given print.

A large range of statistical tools can be deployed to study the relationships presented above. We chose to take advantage of a classification machine learning technique called Random Forest [30]. Random Forest family of algorithms is used extensively in data mining and has the following advantages:

1. It allows to handle both categorical and continuous data input;
2. It allows to handle variables that are correlated;
3. The training mechanism of the classifier is based on bootstrapping techniques, and prevents over-fitting the model to the data. The classification error can be reasonably estimated on the entire dataset;
4. Once trained, the model ranks the predictive variables in terms of importance, hence allowing distinguishing the variables that have a strong impact on the classification performances;
5. A significance test of each contributing variable can be made using ad hoc procedures.

A Random Forest is a collection of unpruned classification trees. Each tree is individually trained on a sample of the entire dataset, while the remaining data (Out-Of-Bag samples - OOB) is used to measure its classification ability. The OOB sample is randomly drawn for each tree, which makes it unlikely for two trees to be trained on the same sample of the original dataset (and thus prevents over-fitting the forest to the data). The training process involves feeding the tree with predictive variables and adapting the internal variables of the classifier to reach the expected output variable. Each tree has a classification error rate corresponding to the difference between its expected and actual outputs (i.e., the proportion of OOB samples, which are not classified in their respective expected class). The performance of the forest is measured by the aggregated OOB classification errors of the individual trees. These classification errors are reported in confusion matrices.

The importance of each predictive variable in the classification process can be measured. However, it is necessary to realize that, with Random Forest classifiers, the importance of a given variable can only be measured with respect to the other predictive variables. The measures of importance reported in the following sections are relative.

The importance of a variable can be defined in several ways:

1. It can be defined in terms of its influence on the accuracy of the classifier and on its ability to create homogeneous sub-datasets. In other words, a variable is considered important if a significant decrease in the accuracy of the classifier is observed when that variable is removed, and if that variable is able to part the training dataset into homogenous sub-categories [31];
2. *Ad hoc* procedures can also be used to express the contribution of each variable to the reduction of the classification error rate of the forest;
3. The significance of each variable can finally be measured by comparing its classifying abilities when compared to a shadow variable making random classification decisions. This is performed using the Boruta feature selection algorithm. [32]

All computations were carried out using R [33].

During this study, we use Random Forest classifiers as a rational proxy for human decision-making, in order to find a set of reasonable predictors for the decisions made by the study participants based on their annotation. We use the variables extracted from the annotations made by the examiners as input variables to the classifier (see Table 1, Table 4 and section 7). We study the importance of these variables against (1) the individual decisions made by the participants, and (2) case-appropriate decisions defined based on the consensus of the group; we use these two variables as the expected output of the classifier. The results are presented below. We will take the Analysis phase and the Comparison phase in turn.

9.1. Sufficiency in relation to the Analysis phase

The analysis of the results obtained during the Analysis phase of the examination process is carried out by separating the examiners using Approach #1 and Approach #2. The output variable is the decision reached by the examiners at the end of the Analysis phase. Under Approach #1, only VID and NV decisions are possible, whereas under Approach #2, VEO is a potential output.

The Random Forest (RF) classifiers trained under Approaches #1 and #2 give the confusion matrices shown in Table 15. Table 15 essentially shows that our classifier can reasonably predict the decisions made by the examiners based on their annotations, but that it is not entirely accurate.

Approach #1		OOB estimate of error rate: 10.49%			
		Predicted states			
		NV	VID	class.error	
True states	NV	113	28	0.20	
	VID	17	271	0.06	

Approach #2		OOB estimate of error rate: 20.63%			
		Predicted states			
		NV	VOE	VID	class.error
True states	NV	158	65	23	0.36
	VOE	62	129	111	0.57
	VID	10	43	921	0.05

Table 15: RF model confusion matrices for the Analysis decisions, respectively for the examiners using Approach #1 and Approach #2.

This lack of accuracy may result from examiners taking into account features observed on the latent prints that were not captured during our project, or from some part of lack of

rationality in the decision-making process of the participants. This observation is similar to the ones reported in [34,35].

Table 16 ranks the predictive variables in terms of importance (refer to section 6 for the code of the predictive variables).

Approach #1				Approach #2			
Predictive variables	Importance	Error rate	Drop	Predictive variables	Importance	Error rate	Drop
denominator of LR	42.3	0.22	0.28	denominator of LR	142.9	0.34	0.16
$p(v H_d)$	36.6	0.13	0.09	M4	135.6	0.25	0.09
M4	35.2	0.12	0.00	$p(v H_d)$	96.0	0.24	0.01
M2b75	15.8	0.13	0.00	M2b75	76.6	0.24	0.00
M1c1	13.5	0.12	0.00	M1c1	53.3	0.24	0.00
L2	11.3	0.13	0.00	M1c2	50.0	0.24	0.00
M1c2	8.0	0.13	0.00	L2	47.6	0.24	0.01
Years_of_Exp	5.3	0.11	0.01	Years_of_Exp	38.6	0.23	0.01
Expert_Status	4.1	0.12	0.00	qs2	30.0	0.22	0.00
M1c3	4.1	0.12	0.00	M1c3	29.5	0.23	0.00
Use_of_level3	3.9	0.10	0.02	Use_of_level3	23.0	0.22	0.01
qs2*	3.4	0.10	0.00	L1	19.5	0.21	0.01
L3	2.2	0.10	0.00	Expert_Status	18.4	0.20	0.00
SOP*	1.4	0.10	0.00	L3	17.7	0.21	0.00
L1*	1.2	0.10	0.00	SOP*	11.7	0.21	0.00
Sex*	0.6	0.10	0.00	Sex	6.8	0.20	0.00

Table 16: List of the considered variables ranked according to their importance (first column) for Approach #1 and Approach #2. The second and third columns indicate the contribution of each variable to the reduction of the error rate. Non-significant variables (according to the Boruta significance test) are indicated with a *.

Table 16 shows that the variables of importance are all related to the number of minutiae annotated during the Analysis phase, namely M4, and to the specificity of their spatial relationship, expressed by the match probabilities associated with the statistical model (denominator of LR) and the probability of our matching algorithm retrieving a similar configuration in the reference database ($p_V(v|H_d)$). When the individual contribution of the variables to the reduction of the error rate of the classifier is measured, it emerges that the importance of variables that are not related to minutiae (e.g. the years of experience, the expert status, or the observation of level 3 features) is very limited. During the test, most of the variables remained significant, but it is fair to say that the most decisive driving force is associated with the number of minutiae, the certainty associated with their type and location, and the specificity of their spatial relationships.

We note that the number of minutiae annotated during the Analysis phase, or the decision made at the end of that phase, show no correlation with the number of hours spent

examining latent print per week, or with the number of years of experience (Figure 39). This corroborates the results presented in [35]. The large variability between the number of minutiae observed by the various participants on a given latent print was already discussed in relation with Figure 27.

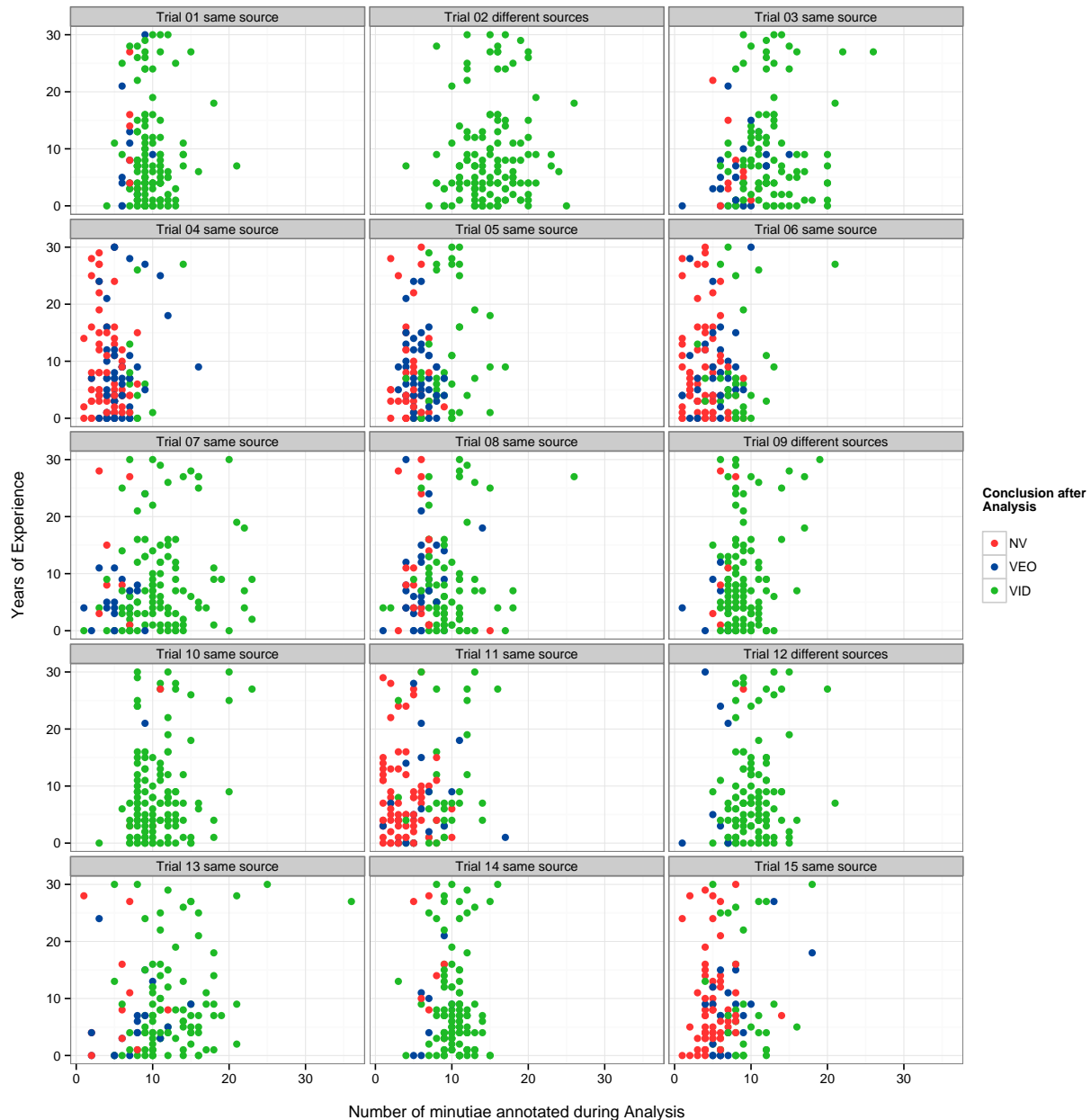


Figure 39: For each trial, the number of minutiae annotated during the Analysis phase against the number of years of experience.

Finally, we have also explored the relationship between (a) the deviation of each examiner from the consensus of the participants for a given case, in terms of quality (M3b75) and

number of minutiae (M5b), and (b) the level of agreement between the decision made by that examiner and the appropriate decision that should have been reached in that case. The hypothesis is that large deviations from the consensus in terms of quality and quantity may result in unexpected deviations from the appropriate response.

RF classifiers incorporating these two additional variables and other relevant variables were trained using an output variable representing the divergence (if any) between the decision reached by each examiner and the appropriate decision on the suitability of the latent print. For each trial, this appropriate decision was set based on the opinion of the majority of the participants (for each trial/approach). Hence, based on the decisions presented in Figure 23 and Figure 24, the following consensus decisions were deemed appropriate for the purpose of this project (Table 17).

	Trial01	Trial02	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial09	Trial10	Trial11	Trial12	Trial13	Trial14	Trial15
Approach # 1	VID	VID	VID	NV	NV	NV	VID	VID	VID	VID	NV	VID	VID	VID	NV
Approach #2	VID	VID	VID	VEO	VEO	NV	VID	VID	VID	VID	NV	VID	VID	VID	NV

Table 17: Appropriate decisions associated with each trial, obtained by a majority vote among all examiners, respectively for Approach #1 and Approach #2.

The ranked list of the importance of the variables is given in Table 18.

Approach #1				Approach #2			
Predictive variables	Importance	Error rate	Drop	Predictive variables	Importance	Error rate	Drop
M5b	36.8	0.28	0.22	M5b	176.4	0.38	0.12
DevLevels	27.5	0.20	0.08	DevLevels	122.8	0.28	0.10
M3b75	25.0	0.21	-0.01	DevDegradationFactors	107.5	0.29	-0.01
DevDegradationFactors*	21.2	0.21	0.00	M3b75	106.0	0.30	0.00
Hours_per_week	10.6	0.21	0.00	Years_of_Exp	66.2	0.29	0.01
Years_of_Exp*	9.7	0.21	0.00	Hours_per_week	59.2	0.29	0.00
Expert_Status*	6.5	0.19	0.01	Expert_Status	23.6	0.29	0.01
no.of.difference*	2.3	0.19	0.00	no.of.difference*	11.4	0.29	0.00
Sex*	0.9	0.20	0.00	Sex*	11.1	0.29	0.00

Table 18: List of the considered variables ranked according to their importance for Approach #1 (left) and Approach #2 (right). Non-significant variables (according to the Boruta significance test) are indicated with a *.

The results show that the variables expressing a deviation from consensus are decent predictors of the deviation from the appropriate decision. In particular, the deviation from the minutiae consensus (M5b) is the strongest predictor. The other variables reflecting the participants' deviation from consensus are significantly weaker in their ability to predict a deviation from the appropriate decision. It means that such participants observing more (or less) minutiae than the consensus of their peers have a greater tendency to form a different opinion than the decisions deemed appropriate in a given case. This result is not surprising,

but confirms that (a) the observation of more features correspond to a higher likelihood of a VID decision; and (b) that there is significant variability between the sets of features considered by each participants in a given case, thus indicating a lack of standardization of these features.

9.2. Sufficiency in relation to the Comparison phase

A similar approach is used to study the relationships between the annotations made during the Comparison phase, the variables associated with the examiners and the conclusion reached after the Evaluation phases. RF classifiers were trained to study the importance of each variable towards the conclusions reached by the participants. For this part of our analysis, we have grouped all inconclusive decisions together, and only consider ID, EXC and INC as possible values for the output variable.

The confusion matrix of the RF classifier is given in Table 19. Exclusions are difficult to predict. This is mainly due to the fact that examiners were not adopting any form of consistent mechanism to document cases where they reached exclusion decisions (i.e., some participants reported observed differences, some reported observed concordances, and some did not report anything). We observed here a lack of consistency in the documentation of these decisions by the participants.

OOB estimate of error rate: 18.96%					
		Predicted states			
		EXC	ID	INC	class.error
True states	EXC	289	17	96	0.28
	ID	10	631	87	0.13
	INC	49	111	661	0.19

Table 19: Confusion matrix for the classifier trained with Comparison annotation and Evaluation decisions.

The study of the importance of each variable (Table 20) shows that the most important variables (in terms of general importance and contributing to the reduction in classification errors) are the number of annotated paired minutiae and their spatial relationships (as expressed by the denominator of the LR and by $p_v(v|H_d)$). Some other variables related to the minutiae and their quality (C_QQ1, C_QQ2, M2b75, L2) and the observation of (the absence of) differences between the latent and control prints also play an important role in the decision-making. All other variables have much less impact. Typically, we observe no

relationship between the decisions reached by the participants and the variables associated with their practice, or with the presence/absence of features such as sweat pores, or shape of ridge edges.

Predictive variables	Importance	Error rate	Drop
M6	184.1	0.34	0.16
denominator of LR	183.7	0.28	0.05
$p(v Hd)$	143.0	0.27	0.01
C_QQ2	109.8	0.24	0.03
C_QQ1	90.5	0.24	0.01
M2b75	86.1	0.23	0.01
L2	59.4	0.22	0.01
M1c1	55.0	0.21	0.01
M1c2	48.7	0.21	0.01
no.of.difference	42.7	0.20	0.01
Years_of_Exp	37.4	0.19	0.00
M1c3	29.3	0.19	0.00
qs2	27.2	0.19	0.00
M1d	26.2	0.19	0.00
Expert_Status	18.6	0.19	0.00
Use_of_level3	18.2	0.19	0.00
Processing*	17.8	0.19	0.00
L3	17.2	0.19	0.00
L1	16.6	0.19	0.00
SOP*	12.0	0.19	0.00
C_L3	11.6	0.19	0.00
Sex	7.2	0.19	0.00
Suitability_approach	6.4	0.19	-0.01

Table 20: List of the considered variables ranked according to their importance (first column). The second and third columns indicate the contribution of each variable to the reduction of the error rate. Non-significant variables (according to the Boruta significance test) are indicated with a *.

Similarly to the observations made during our exploration of the Analysis phase, there is no observed correlation between the number of minutiae paired during the Comparison phase and the number of cases worked per week or the number of years of experience. This is illustrated in Figure 40. Note that due to the lack of consistency in the annotations associated with the exclusions (or towards them), Figure 40 focuses only on same source trials and with the number of paired minutiae associated with the identification and the inconclusive decisions. The significant variability in the minutiae paired by the participants is discussed in relation to Figure 31.

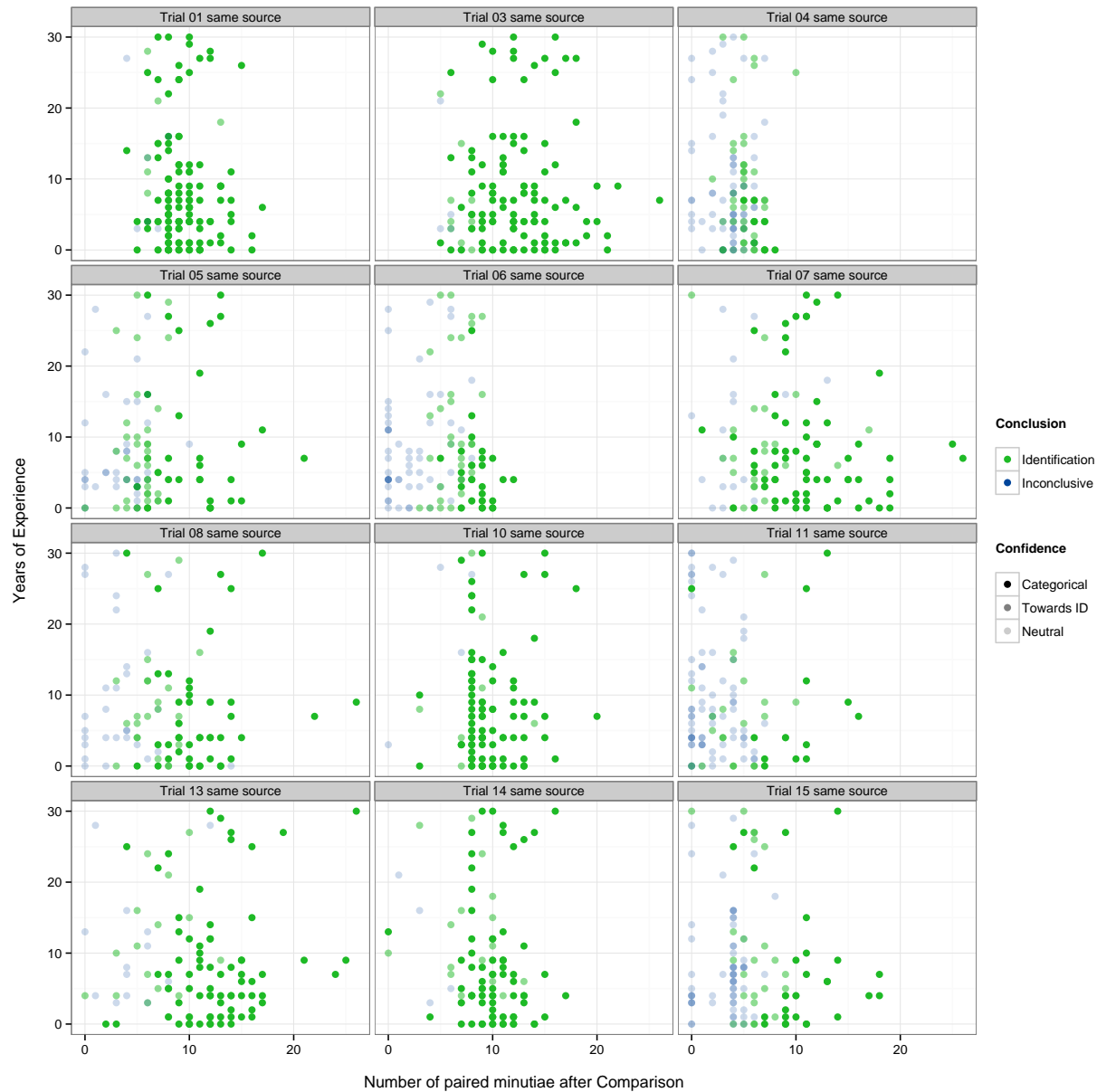


Figure 40: For each trial with a corresponding source, the number of minutiae annotated and paired during the Comparison phase against the number of years of experience. The cases with conclusions towards exclusion have been removed from the graph.

We analyze the potential relationship between the divergence between the participants in their observations and conclusions in a similar fashion as we did for the Analysis phase. We measure the contribution of the demographic variables associated with the users and their deviation from the consensus (typically M5b and M3b75) to the divergences of their answers to decisions that are deemed appropriate for the trials. Once again, due to the difficulties associated with the documentation of exclusions, attention was given only to the

cases where the latent and control prints shared a common source. The appropriate decisions following the Comparison phase for these trials was set according to the opinion of the majority of the participants and are reported in Table 21.

Trial01	Trial03	Trial04	Trial05	Trial06	Trial07	Trial08	Trial10	Trial11	Trial13	Trial14	Trial15
ID	ID	INC	INC	INC	ID	INC	ID	INC	ID	ID	INC

Table 21: Consensus decision deemed appropriate for each trial (same source) following the Comparison phase, obtained by a majority vote among all examiners.

A RF classifier was trained to measure the deviation between reported and appropriate conclusions, as a function of the users' demographic variables and their deviation from the consensus in terms of quality and quantity of their reported observations (typically M5b and M3b75). The importance attributed to each variable by the classifier is shown in Table 22.

Predictive variables	Importance	Error rate	Drop
M5b	133.04	0.42	0.08
DevLevels	129.08	0.31	0.12
M3b75	117.20	0.29	0.02
DevDegradationFactors	110.79	0.28	0.01
Years_of_Exp	67.84	0.28	0.00
Hours_per_week	62.54	0.27	0.00
no.of.difference	29.19	0.27	0.00
Expert_Status	24.69	0.27	0.00
Sex*	9.17	0.27	0.00

Table 22: List of predicting variables for the prediction of the deviation of the reported conclusions from the postulated ground truth state in the trials with latent and print originating from the same source, ranked according to their importance. Non-significant variables (according to the significant test from the Boruta procedure) are indicated with a *.

Table 22 shows that the variable translating the deviation from the consensus in terms of the number of minutiae annotated during the Analysis phase (M5b) has the larger bearing on the divergence from the appropriate decision. Note that the variables translating the deviation from the consensus in terms of overall quality of the latent print (M3b75) and in terms of degradation factor and quality/quantity of non-minutiae features (DevLevels and DevDegradationFactors) are also useful to explain diverging conclusions. As before, participants observing more/less minutiae than their peers are more likely to reach a different conclusion than the conclusion deemed appropriate for the considered trial.

9.3. Analysis of the annotations made in two cases

It is important to go beyond the statistical analysis and explore the annotations provided by the participants in relation to their decision. This enables a finer understanding of the motivations behind the decisions made by the examiners. In particular, we focus our attention on two cases that show significant variability in the conclusions reached by the examiners after the Evaluation phases.

Trial 08 (same source)

The results associated with this trial are recalled in Figure 41.

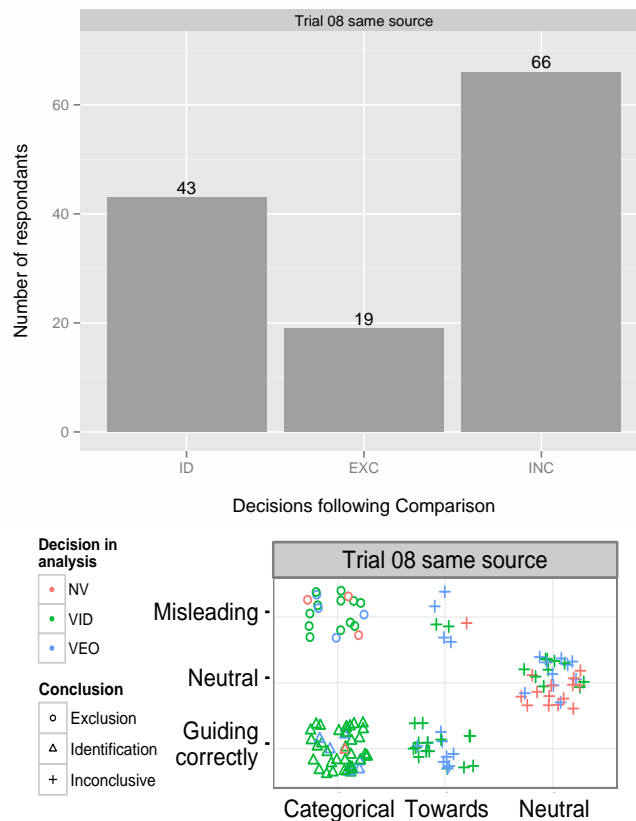


Figure 41: Summary of the results associated with trial 08.

A surprising result of this study, which is enabled by the documentation features in PiAnoS, is the range of variations among the observations reported by the examiners who formed an ID conclusion for Trial 08. For example, User06, as shown in Figure 42, kept all the annotations from the Analysis phase to conduct the comparison with only one minutia

added (in yellow) during the Comparison phase. In total, 10 minutiae are paired to support the identification conclusion.

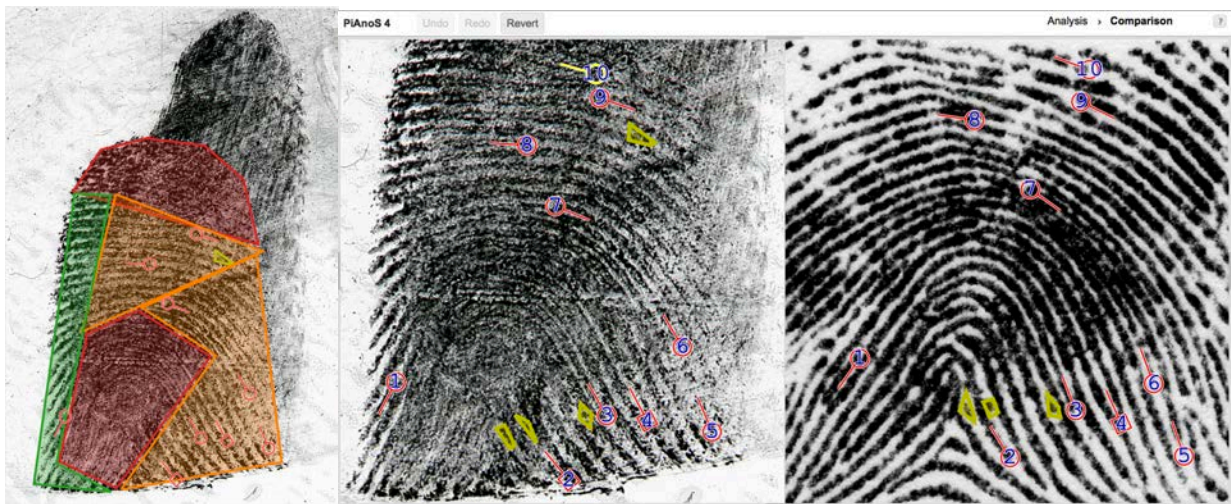


Figure 42: Annotations made by User06 both in Analysis phase (left) and Comparison phase (middle and right).

On the contrary, User481 (Figure 43) annotated 6 minutiae during the Analysis phase, which were erased and replaced by 22 matching minutiae. These minutiae were all annotated on the latent print during the Comparison phase (in yellow). In our opinion the working protocol of User06 has safer practices than User481, although they both reach the same conclusion, in line with the consensus decision for that case.

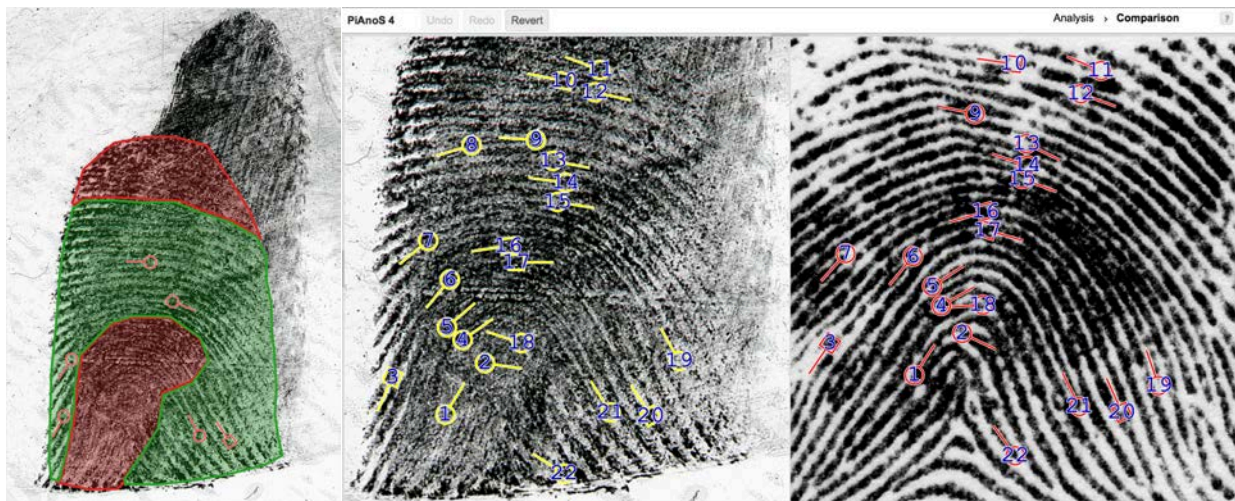


Figure 43: Annotations made by User481 both in Analysis phase (left) and Comparison phase (middle and right).

Trial 12 (different sources)

The results obtained for that trial are recalled in Figure 44.

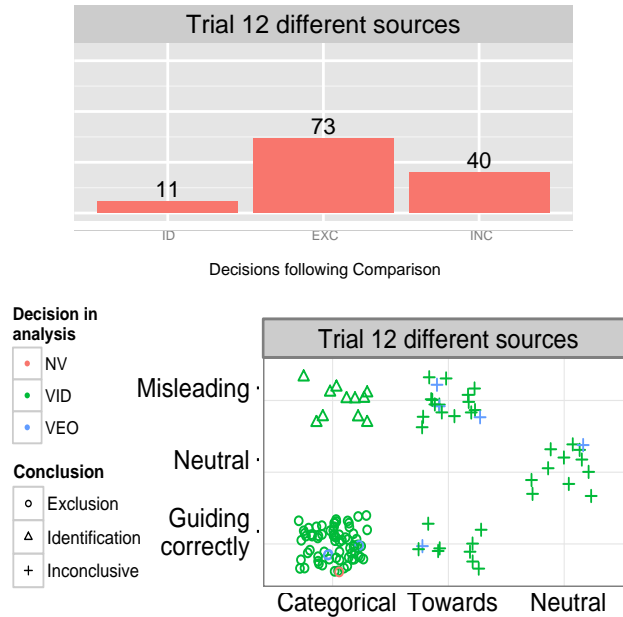


Figure 44: Summary of the results associated with trial 12.

The observation of the annotations provided by a sample of the examiners who wrongly identified the source of the latent print is revealing. The following figures (Figures 45 to 48) present the observations reported by some participants, ordered according to the number of years of experience.

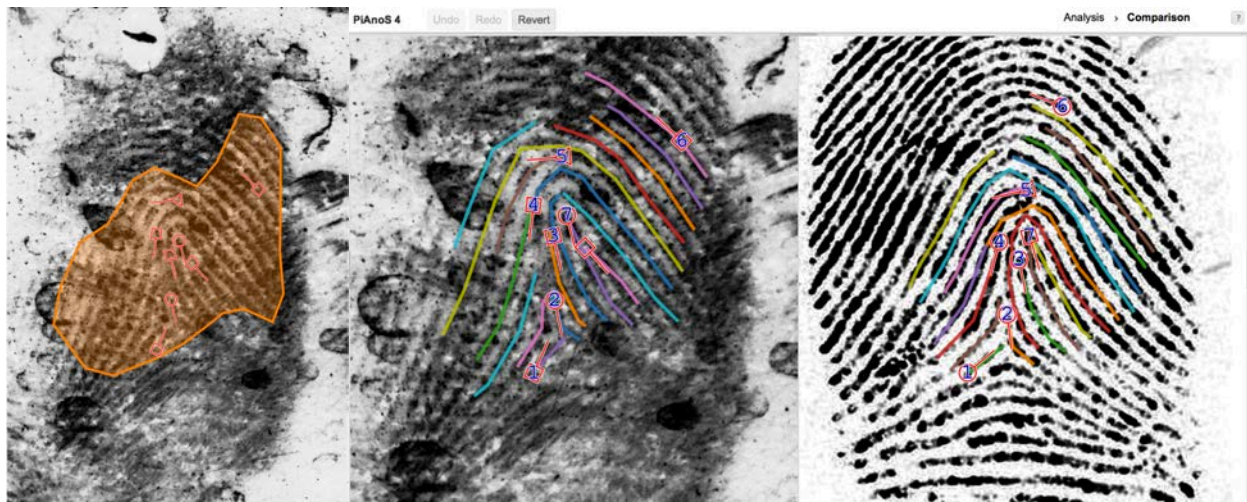


Figure 45: Annotations made by User342 – not certified, 3 years of experience in Analysis phase (left) and Comparison phase (middle and right).

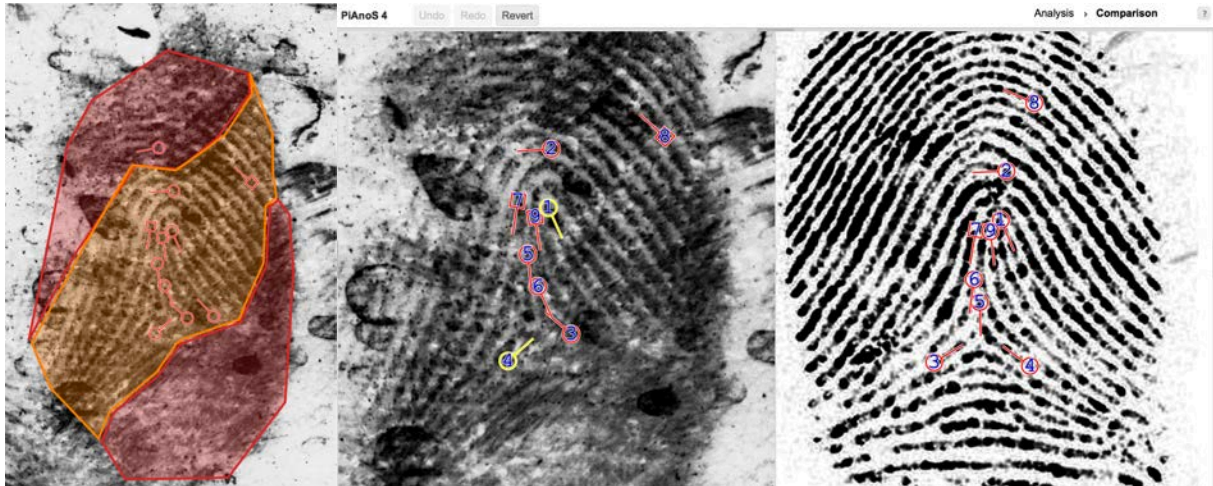


Figure 46: Annotations made by User436 – certified, 5 years of experience in Analysis phase (left) and Comparison phase (middle and right).

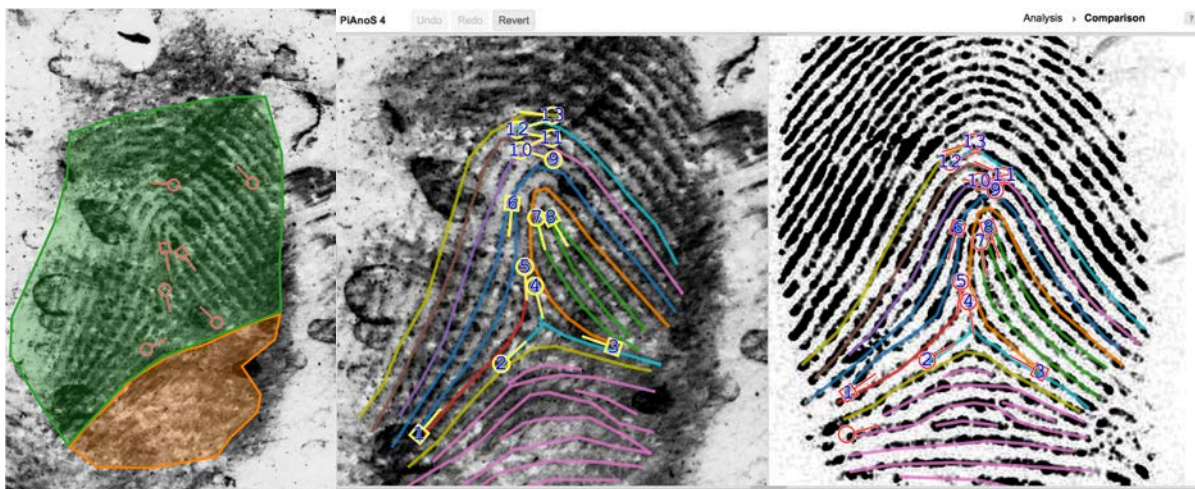


Figure 47: Annotations made by User481 – certified, 7 years of experience in Analysis phase (left) and Comparison phase (middle and right).

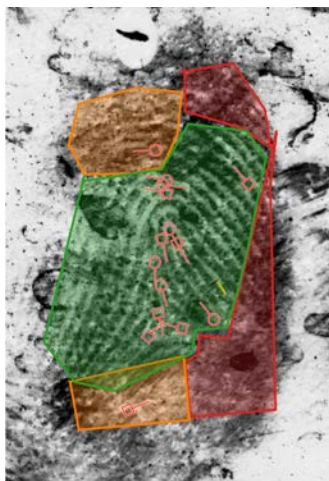


Figure 48: Annotations made by User024 – certified, 19 year of experience in Analysis phase (left) and Comparison phase (middle and right).

No documentation provided for the Comparison phase.

The first two examiners (User342 and User436) used the annotations in the Analysis phase as a basis for their comparison, albeit User436 adapted the position of two minutiae (number 1 and 4 in yellow). Both users have annotated one pair of minutiae that seems out of tolerance (number 6 for User342, number 8 for User346). The likelihood ratios computed on these annotations are $3.17e^{-18}$ for User342 and $9.03e^{-22}$ for the second. These LRs represent very strong evidence that the latent and control prints do not come from the same source. Thus, a probabilistic assessment could have functioned as a quality assurance mechanism in these cases to strengthen the mismatch in the pairing of minutiae 6 and 8. For the two last examiners, the situation is different. The last examiner (User024) did not offer any annotation in the comparison phase and it is impossible to determine which features were used to reach the erroneous identification. However, the third examiner (User481) literally re-annotated and paired all the minutiae in agreement from the information observed on the provided comparison print provided (all minutiae are in yellow). When marshalled through to provide the illusion of a correspondence, there is no surprise to have a corresponding likelihood ratio equals to 723,657 (very strongly in favour of a common source).

User 342 The known is of very poor quality. Making the comparison more difficult.

User436 My minutiae didn't number correctly. Sorry

User481 *This latent print was very complex. At first analysis it appears to be a pretty clear and straightforward impression; however, upon comparison to the known print it was obvious there were several distortion issues at play in both impressions. In the latent impression the ridges are being spread apart at the lower portion of the print due to pressure distortion. There is also some distortion factors at play toward the tip above the core of the latent impression. In the known print, there is a surface scar radiating from the tip of the core moving outward toward the right side which is causing a pulling effect on the surrounding ridges due to the healing of the scar tissue tightening around the surface of the ridged skin. There are also some areas of concern toward the tip and the left side of the latent impression where the print detail becomes less visible and also in the poor tonal quality of the known impression causing some red flags; however, with the amount of 2nd level detail in agreement and 3rd level ridge shapes (particularly the trifurcating area at the delta of the loop) there is sufficiency for a conclusion of identification. This conclusion did take an enormous amount of time to reach due to distortion and quality issues in both impressions.*

User024 *Annotating this comparison was especially difficult. I spent much time on it and finally did the comparison after removing all markings.*

Table 23: Narratives put forward by four of the examiners who wrongly identified the latent print in trial 12.

The notes associated with the comparison phase are equally representative of questionable practices (Table 23). Only User481 provided a detailed justification for its conclusion (with the caveat that all observations on the latent print are driven by the observations made on the control print).

10. Implications of the main findings for practice and conclusion

The main findings of this study are twofold:

10.1. Concept of sufficiency

The concept of sufficiency at the end of the Analysis phase is mostly driven by the quality, number and spatial relationships of the minutiae observed on the latent print during the Analysis phase. The concept of sufficiency at the end of the Evaluation phase is mostly driven by the quality, number and spatial relationships of the minutiae observed in agreement between the latent and control prints during the Comparison phase.

That said, the data collected during this research project and presented in this report do not allow us to observe a consensus between the participants on a quantitative measure of the quality and quantity of features that could (a) help define the concept of sufficiency and (b) serve as a quality assurance standard for declaring latent print of value and differentiating between inconclusive and identification conclusions. To some extent these results are similar to the conclusions reached by [34,35].

Other features indicated during the Analysis phase (such as the qualifiers associated with each level of features or the degradation factors), and variables capturing the demographics of the users, do not seem to particularly influence the decision process. This is not to say that they do not carry any weight, but we have not found any strong relationship between these variables and the conclusions reached.

The use of statistical tools (quality metrics or statistical models) may prove to be useful in the long term to support and strengthen the decisions made by examiners; however, these tools are subject to the same variability in the observations between different examiners as the current decision-making process (see below). It appears from our results that the use of such tools is not creating new needs, but is rendering more apparent the shortcomings of current practice. The use of such tools is conditioned on addressing the current needs in terms of standardization and documentation of working practices.

Finally, our results show that there is some lack of understanding of the concept of sufficiency when it comes to exclusion conclusions. This may reflect shortcomings in the training of latent print examiners.

10.2. Consistency in the definition, observation and use of friction ridge skin features

While our data clearly show that the concept of sufficiency is associated with the observation of minutiae, our results also present a picture of the lack of standardization in the definition of what constitutes a minutia, and when it is appropriate to take it into account in the decision-making process. More generally, our results show that the lack of standardization extends to other aspects of friction ridge skin examination, including factors such as distortion, degradation, and influence of background and development techniques, which are heavily used to *explain* the differences observed between the latent and control prints. Overall, it seems that examiners are coherent with their own internal appreciation of the quality/quantity of features related to the notion of sufficiency (i.e., the observation of more features leads to more categorical decisions); however, there is low consistency between examiners in a given case (at least in complex cases).

These results are clearly mitigated by our research design: we specifically chose trial cases that would generate as much variability as possible between participants. Such variability may not exist often in casework, where it seems that most of the identifications decisions are made based on latent prints of much higher quality and presenting more features [36]. Nevertheless, it appears urgent to develop and provide guidelines and training defining more robustly the concept of minutia, both in the Analysis phase and in the Comparison phase (as described in [35] for the Netherlands).

In addition, improvement towards consistency could be achieved through rigorous definitions of the conditions under which minutiae are annotated, supplemented by working protocols for auditing and reviewing these annotations. It is only with such documentation that a relevant assessment of the quality of the supporting information in a case can be made.

11. Bibliography

[1] National Research Council of the National Academies (2009). Strengthening Forensic Science in the United States: A Path Forward. The National Academies Press, Washington, D.C.

[2] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) (2011). Terminology (Latent/Tenprint), version 4.0.
http://www.swgfast.org/documents/terminology/121124_Standard-Terminology_4.0.pdf
(last verified April 29th 2013)

[3] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) (2013). Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint), version 2.0.
http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf (last verified April 29th 2013)

[4] Interpol European Expert Group on Fingerprint Identification II - IEEGFI II (2004). Part 2: Detailing the Method Using Common Terminology and Through the Definition and Application of Shared Principles. Lyon: Interpol.

[5] Ashbaugh DR. (1999). Qualitative-Quantitative Friction Ridge Analysis – An Introduction to Basic and Advanced Ridgeology. Geberth VJ, ed., Boca Raton: CRC Press.

[6] Champod C, Lennard CJ, Margot PA, Stoilovic M. (2004). Fingerprints and other Ridge Skin Impressions. Boca Raton: CRC Press.

[7] US v Llera Plaza, Acosta and Rodriguez, US District Court of the Eastern District of Pennsylvania, Criminal No. 98-362-10,11,12.

[8] Evidence. Fingerprint Experts. Seventh Circuit Upholds the Reliability of Expert Testimony regarding the Source of a Latent Fingerprint. United States v. Havvard, 260 F.3d 597 (7th Cir. 2001). Harvard Law Review. 2002;115(8):2349-56.

[9] United States v Byron Mitchell, Court of Appeals for the Third Circuit, No. 02-2859 (April 29, 2004).

[10] State of Maryland v. Bryan Rose, The Circuit Court for the Baltimore County, K06-0545.

[11] State of Minnesota v. Jeremy Jason Hull, District Court - Seventh Judicial District, No. 48-CR-07- 2336.

[12] Langenburg G., Champod C., Wertheim P. (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. 54 (3), 583-590.

[13] International Association for Identification (1980). Resolution VII Amended. Identification News 30(3).

[14] Ulery BT., Hicklin RA., Buscaglia J., Roberts MA. (2011). Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. 108(19), 7733-8.

- [15] Haber L., Haber RN. (2008). Scientific Validation of Fingerprint Evidence under Daubert. *Law Probability and Risk* 7(2), 87-109.
- [16] Cole S. (2005). More than zero, accounting for error in latent fingerprint identification. *Journal of Criminal Law and Criminology* 95 (3), 985–1078.
- [17] Picture Annotation Software 4 – PiAnoS (2012). University of Lausanne, Switzerland. <https://ips-labs.unil.ch/pianos/index.html> (last verified April 29th 2013)
- [18] User Instructions for PiAnoS 4 (2012). Penn State, USA – University of Lausanne, Switzerland. https://ips-labs.unil.ch/pianos/downloads/Pianos4_Instruction_Manual_V1.2.pdf (last verified April 29th 2013)
- [19] Champod C. (1995). Reconnaissance Automatique et Analyse Statistique des Minuties sur les Empreintes Digitales. Ph. D. thesis, University of Lausanne, Switzerland.
- [20] Neumann C. (2012). Statistics and Probabilities as a Means to Support Fingerprint Examination, IN Lee and Gaensslen's *Advances in Fingerprint Technology* (ed. R. Ramotowski), 3rd ed., CRC Press, 419-466.
- [21] Stoney D. (2001). Measurement of fingerprint individuality. IN *Advances in Fingerprint Technology* (eds H. Lee and R. Gaensslen), 2nd ed., CRC Press, 327–388.
- [22] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) (2011). SWGFAST Response to the Research, Development, Testing & Evaluation Inter-Agency Working Group of the National Science and Technology Council, Committee on Science, Subcommittee on Forensic Science. <http://swgfast.org/Resources/111117-ReplytoRDT&E-FINAL.pdf> (last verified April 29th 2013)
- [23] Helper AB., Saunders CP., Davis LJ., Buscaglia J. (2012). Score-based likelihood ratios for handwriting evidence, *Forensic Sci. Int.* 219, 129-40.
- [24] Hotz T., Munk. A. (2012). Comments on Neumann C., Evett I., Skerrett J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, *J. R. Statist. Soc. A* 175, 371-415.
- [25] Neumann C., Evett I., Skerrett J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, *J. R. Statist. Soc. A* 175, 371-415.
- [26] Lindley D. (1977). A problem in forensic science. *Biometrika* 64, 207–213.
- [27] Neumann C., Evett I., Skerrett J., Mateos-Garcia I. (2011). Quantitative assessment of evidential weight for a fingerprint comparison I: Generalisation to the comparison of a mark with set of ten prints from a suspect. *Forensic Sci Int.* 207, 101-105.
- [28] Bookstein F. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 567–585.
- [29] Evett IW., Williams RL. (1996). A review of the sixteen point fingerprint standard in England and Wales. *J. Forens. Ident.* 46, 49–73.

- [30] Liaw A., Wiener M. (2002). Classification and Regression by RandomForest. R News 2(3), 18-22.
- [31] Ishwaran H., Kogalur UB. (2007). Random survival forests for R. Rnews 7(2), 25-31
- [32] Kursu MB., Rudnicki WR. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software 36(11), 1-13.
- [33] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
<http://www.R-project.org/> (last verified April 29th 2013).
- [34] Ulery BT., Hicklin AR., Kiebusinski GI., Roberts MA., Buscaglia J. (2013). Understanding the sufficiency of information for latent fingerprint value determinations, Forensic Sci. Int., <http://dx.doi.org/10.1016/j.forsciint.2013.01.012>
- [35] Langeburg G. (2012). A critical analysis and study of the ACE-V process, Ph. D. Thesis, University of Lausanne, Switzerland.
- [36] Neumann C., Mateos-Garcia I., Langenburg G., Kostroski J., Skerrett JE., Koolen M. (2011) Operational benefits and challenges of the use of fingerprint statistical models: a field study. Forensic Sci. Int. 212, 32-46.

12. Appendix A – Trial images



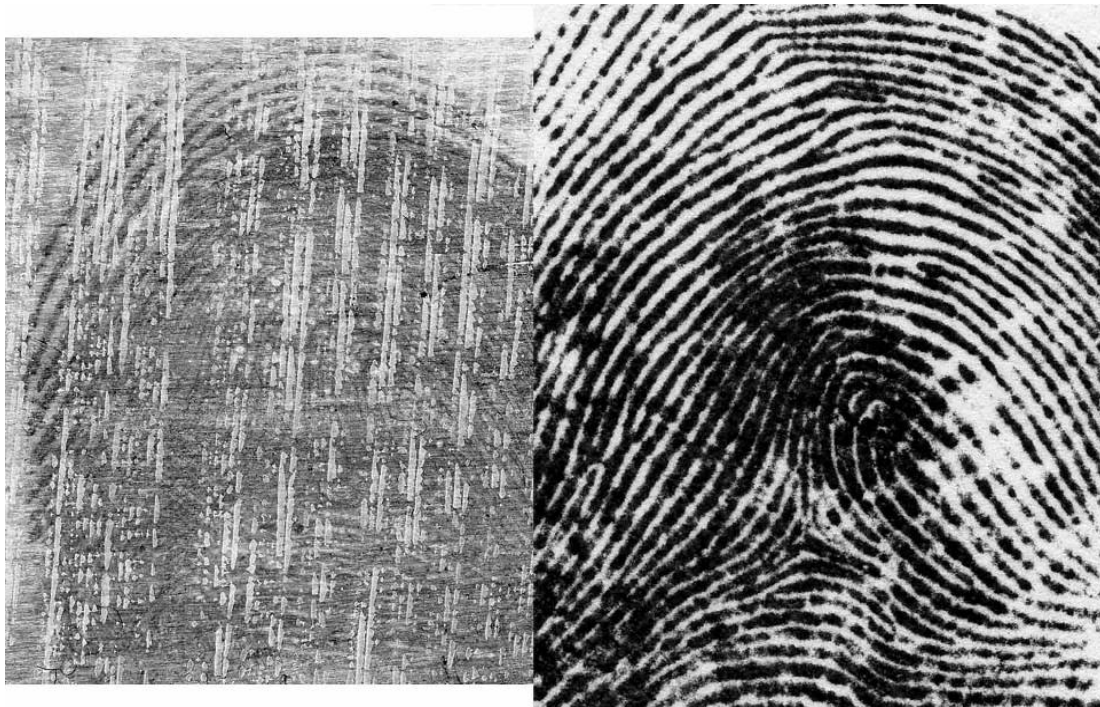
Trial 1: Same source comparison



Trial 2: Different sources comparison



Trial 3: Same source comparison



Trial 4: Same source comparison



Trial 5: Same source comparison



Trial 6: Same source comparison



Trial 7: Same source comparison



Trial 8: Same source comparison



Trial 9: Different sources comparison



Trial 10: Same source comparison with an apparent discrepancy



Trial 11: Same source comparison



Trial 12: Different sources comparison



Trial 13: Same source comparison



Trial 14: Same source comparison with connective ambiguities



Trial 15: Same source comparison