VALIDATION OF LIKELIHOOD RATIO METHODS USED FOR FORENSIC EVIDENCE EVALUATION:

APPLICATION IN FORENSIC FINGERPRINTS

Rudolf Haraksim
Enschede, The Netherlands, 2014

VALIDATION OF LIKELIHOOD RATIO METHODS USED FOR FORENSIC EVIDENCE EVALUATION:

APPLICATION IN FORENSIC FINGERPRINTS

Rudolf Haraksim

Enschede, The Netherlands, 2014

PhD dissertation committee:

Chairman and Secretary	**Peter M.G Apers**, Professor University of Twente, Netherlands
Promotor	**Didier Meuwly**, Professor University of Twente, Netherlands Netherlands Forensic Institute
Promotor	**Raymond N.J. Veldhuis**, Prof. Dr. Ir. University of Twente, Netherlands
Committee Members	**Daniel Ramos**, Dr. Universidad Autonoma de Madrid, Spain
	Christophe Champod, Professor Université de Lausanne, Switzerland
	Richard Boucherie, Professor University of Twente, Netherlands
	Marianne Junger, Professor University of Twente, Netherlands

CTIT

CTIT PhD Dissertation series No. 13-302
Centre for Telematics and information Technology
P.O.Box 217, 7500 AE, Enschede, The Netherlands

The research has been funded by the Marie Curie ITN grant (FP7-PEOPLE-ITN-2008, grant number 238803), within the scope of the BBfor2 project.

Printed and bound by: www.ipskampdrukkers.nl
Cover designed by: Rudolf Haraksim
ISSN: 1381-3617
ISBN: 978-90-365-3648-6
http://dx.doi.org/10.3990/1.9789036536486

VALIDATION OF LIKELIHOOD RATIO METHODS USED FOR FORENSIC EVIDENCE EVALUATION:

APPLICATION IN FORENSIC FINGERPRINTS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, 18th June 2014 at 12:45 hours

by

Rudolf Haraksim

born in Kosice, Slovak Republic
on 23rd December 1981

This thesis has been approved by:

Promotor: Prof. Didier Meuwly

and

Promotor: Prof. Dr. Ir. Raymond N.J. Veldhuis

“The opposite of a correct statement is a false statement. But the opposite of a profound truth may well be another profound truth”

- Niels Bohr

Mojim rodičom, ktorí pri mne vždy stoja.

A ma fille Auróra bien-aimée.

A Daniel – por me motivar :)

Table of Contents

Thesis Motivation

Traditional biometric systems, set to distinguish between a genuine user and an impostor can follow for example the Neyman-Pearson approach, in which for instance a threshold is set to the false acceptance rate. Some forensic biometric systems aim at producing decisions based on probability of two mutually exclusive propositions, knowing the evidence and some assigned prior as an end result. Methods used for forensic evidence evaluation are based on similar technology as the standard biometric systems, but they aim to evaluate the probability of the evidence knowing the propositions in a logically correct framework. This form of evaluation, derived from the Bayes theorem, is called likelihood ratio (LR) approach. It allows for the evaluation of the strength of evidence of different observations independently of the prior probabilities of the propositions tested. During the 20th century probabilities of the propositions or even categorical decisions were reported by forensic practitioners in forensic evidence evaluation, but a critical review of the logic (hard decisions made by forensic practitioners based on their subjective personal probabilities) applied in forensic evidence evaluation showed, that the LR approach presents a logically correct way to evaluate and report the strength of forensic evidence.

In the absence of data and statistical models, the LR is usually assigned using a human based method, by a forensic practitioner on the basis of personal probabilities and to the best of her/his knowledge and experience. The strength of evidence assigned in this way is arbitrary and the resulting LR values are relative. Human-based methods are also rather difficult to test and validate. A long-term ambition in forensic biometric evaluation is to embed the LR inference model and biometric core technology into automatic procedures to support and complement human-based methods. It consists of developing more objective methods for the calculation of strength of forensic evidence based on data and statistical models and the validation of these methods. This thesis focuses on the latter topic: *the validation of LR-based methods used for forensic evidence evaluation*.

Automated systems evaluating the strength of evidence are usually set to produce discriminating scores, describing the similarities or differences between 2 objects – a test specimen (for example a fingermark recovered from the crime-scene [1,2,3]) and a reference specimen (for example from a 10-print card of the suspected individual) using feature comparison algorithms (in fingerprints the features mostly constitute of minutiae position and orientation). The strength of evidence of these scores is then evaluated in favour of both of the prosecution and of the defence propositions within the LR framework.

Biometric core technologies used for forensic evidence evaluation are usually developed (and their performance evaluated) on standardized datasets, which may not reflect to full extent the forensic conditions of the specimens-to-evaluate (distortion or reduced quality to name a few). While the biometric core technologies may provide exceptional performance in their application domain

(security systems or other), when used in forensic conditions their performance may decrease severely.

In the identity verification, when for example comparing two full and high quality fingerprints, it is possible to achieve a near 100% success rate using the automatic feature extraction and comparison algorithms when discriminating between the genuine user and an impostor. On the other hand in the process of forensic fingermarks investigation about 62% of the marks encoded automatically are automatically linked to a candidate from the database in the Netherlands [4]. Even though the technology has made a significant leap forward, a decrease in performance of state-of-the-art built-for-purpose biometric systems can be observed when comparing forensic fingermarks with fingerprints. Therefore validation of these methods (using forensically relevant datasets) is necessary to quantify and to make explicit the limitations of the LR-based methods (for example as a function of the quality of the specimens, quantity of the material, representativeness of the data). The final result of the validation procedure is then a binary decision regarding the suitability of the LR methods developed in forensic research and development (R&D) process for the use in forensic casework. According to the ISO 17025:2005 section 5.4.5.2 [11] *"…The laboratory shall record the results obtained, the procedure used for the validation, and a statement as to whether the method is fit for the intended use."*

In the scope of this thesis a validation framework will be proposed for the validation of semi-automatic LR methods for forensic evidence evaluation. In [5], Mansfield and Wayman have devised a methodology for assessing the performance of biometric systems. In scope of their work they proposed to split the evaluation of a biometric system into three phases – technology, scenario and operational evaluation. A the three-way evaluation is a standard practice across a whole range of industries and we intend to keep the format proposed for validation of the forensic LR-based methods.

The main contribution of this thesis is in the domain of the scenario evaluation. In order to perform the scenario validation in the forensic evaluation, one should start with answering following questions: "***Which criteria should be used to validate a LR-based inference model*?**", "***What performance characteristics and metrics should be used to report the findings?***". All of these questions help in the development of ***validation framework*** *for LR methods used for fingerprint evidence evaluation*. The performance characteristics and metrics for validation of LR-based methods are motivated by the research carried out in speaker recognition, inspired by the work of N. Brümmer [6], D. Ramos [7,8], D. van Leeuwen [12] and others.

A technology evaluation is out of the scope of the thesis, due to the fact that the algorithms in use have been subject to extensive benchmark tests and

evaluation by third parties. Standardized datasets play a significant role in the technology evaluation for example available through the National Institute of Standards and Technology (NIST). The operational evaluation is also out of the scope of the thesis and rests within the competence of the operational units responsible for implementation of LR-based methods in the casework processes.

Thesis contributions

As the title of the thesis "Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation: *Application in Forensic Fingerprints*" suggests, this thesis mainly deals with the forensic interpretation of discriminating scores produced by Automated Fingerprint Identification System (AFIS). Hence despite the fact that the validation framework for LR methods used for forensic evidence evaluation was in theory developed for application across the whole range of biometric modalities, its applicability is presented in the area of forensic fingerprints.

As a part of this thesis several literature surveys were conducted, addressing issues of guidelines and standards for validation of LR methods used for forensic evidence evaluation; measures of accuracy, discriminating power and calibration in (forensic) biometrics; use of Bayesian Networks for fingerprint evidence evaluation and evidential value of the first level detail fingerprint evidence.

A theoretical framework has been proposed for validation of LR methods used for forensic evidence evaluation. Different methods were used to calculate the LR's from the fingerprint AFIS scores and their performance evaluated using the performance metrics proposed in the theoretical framework.

The theoretical framework developed was applied to validate fingerprint LR method based on the AFIS scores. Several issues have been addressed in the course of the LR method development, namely robustness to the dataset shift (generalization), robustness to the lack of data (data sparsity) and coherence.

Somewhat remotely stands the development of the Bayesian Network for the first level detail (General Pattern) fingerprint evidence evaluation. Original objective to use the metrics proposed in the theoretical framework to measure the performance of the Bayesian Networks developed was unfortunately not met within the thesis timeframe.

Thesis outline

The thesis constitutes of the validation framework, proposed for the validation of LR methods used for forensic evidence evaluation; a collection of published articles dedicated to the performance characteristics defined in the validation framework, such as stability of the LR, robustness of the LR, measuring the coherence of the LR, Bayesian networks for the fingerprint evidence evaluation; the validation report and the appendix, in which the performance characteristics are used to evaluate the performance of human examiners. The thesis, is structured in the following way:

Chapter 1 is dedicated to the introduction of the problem of validation. In this chapter the general validation criteria, as well as the performance characteristics and performance metrics are defined and summarized in a validation framework. A validation report is presented independently as an example in chapter 7.

Chapter 2 focuses on stability of the LR's in the lower region of the within source distribution and the direct dependence on the size of the population datasets. This region is particularly interesting, since the resulting LR's are spread around the "LR = zero" value, which in the Bayes theorem represents a decision boundary between the two propositions – supporting either the prosecution or defence.

Chapter 3 is dedicated to the topic of conditioning in the fingerprint evidence evaluation addressed for example in [3], by looking at the robustness to lack of data of two different approaches: the source independent and the source dependent. For a comparison of the two approaches, the size of the datasets used to produce the same source (SS) and different source (DS) distributions was limited to 100, 500, 1000 and 2000 score samples.

Chapter 4 studies in detail the discriminating scores produced by an automatic fingerprint feature comparison algorithm. This chapter handles issues of data sparsity (especially in the tails of the SS and DS score distributions), multimodality of the resulting discriminating scores and dataset shift. The baseline LR method for producing the LR values from the similarity scores is established using the Kernel Density Function (KDF). An outcome of this work is a multimodal LR method, which unlike the KDF baseline method is robust to the above-mentioned issues. The performance of the two methods is evaluated using the Log Likelihood Ratio Cost [6], Equal Error Rates [9] and presented using the Empirical Cross-Entropy plots [7,8], Tippett plots [10] and Decision Error Trade-off plots [9].

The issue of coherence is addressed in **chapter 5**. Coherence is defined as "*the variation of some measurable parameters in the features studied, perceived as influencing the strength of evidence"*. In this chapter coherence is

observed by introducing additional features (e.g. minutiae points). Multimodal LR method defined in the Chapter 4 is used to produce LR's for 5 – 12 minutiae configurations. Performance of the LR method for different minutiae configurations is evaluated using the Log Likelihood Ratio Cost [6], Equal Error Rate [9] and presented using the Empirical Cross Entropy [7,8], Tippett [10] and Decision Error Trade-off [9] plots.

Automated systems used for fingerprint evidence evaluation consider the second level fingerprint details (mostly minutiae position and orientation). The first level details (General Pattern, ridge count to name a few) are nowadays at best used by the forensic practitioners for exclusion of not-relevant candidate. In **chapter 6** we attempt to quantify the strength of evidence of the General Pattern fingerprint evidence using a Bayesian Network. Even though the strength of evidence of the General Pattern alone is limited, the use of Bayesian Networks brings transparency in the inference process (despite the fact that the validation of Bayesian Networks is not trivial task). Two "data driven" and "built for purpose" Bayesian Networks – graphical models – are proposed in this chapter.

Following the validation framework introduction in chapter 1, the empirical validation report for the multimodal LR model used for fingerprint evidence evaluation (developed in chapter 4) is presented in **chapter 7**.

Thesis conclusions are presented in **the epilogue**, in which the work presented within the scope of this thesis is summarized and the main contributions are highlighted.

In the **appendix A** a subset of the performance characteristics defined in the chapter 1 is used to evaluate the performance of human practitioners in fingerprint evidence evaluation.

References

[1] – C. Neumann, I. Evett et al., *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. 2011, 207(1-3), pp. 101-5

[2] – N. Egli et al, *Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability*, Forensic Sci. Int. 2007, 167, pp. 189-195

[3] – I. Alberink, A. Jongh, C. Rodriguez, Fingermark Evidence Evaluation Based on Automatic Fingerprint Identification System Matching Scores: The Effect of Different Types of Conditioning on Likelihood Ratios, Journal of Forensic Sci, online, DOI: 10.1111/1556-4029.12105, 2013

[4] – D. Meuwly, Friction Ridge Skin – Automated Fingerprint Identification System (AFIS), Wiley – Encyclopedia of Forensic Science. 1–8, 2013

[5] – A.J. Mansfield, J.L. Wayman, Best Practices in Testing and Reporting Performance of Biometric Devices, ISSN 1471-0005, 2002

[6] – N. Brümmer, J. du Preez, *Application independent evaluation of speaker detection*, Computer Speech Lang 2006, 20(2-3):230-75
[7] – D. Ramos, J. Gonzales-Rodriguez, *Reliable support: Measuring calibration of likelihood ratios,* Forensic Sci. Int. 2013
[8] – D. Ramos, J. Gonzales-Rodriguez, G. Zadora, C. Aitken, *Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods*, J. Forensic Sci 2013
[9] – A. Martin et al., *The DET Curve in Assessment of Detection Task Performance*, National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940; 1997
[10] – D. Meuwly, *Forensic Individualization from Biometric Data*, Science & Justice 2006, 46, pp. 205-213
[11] – International Organization for Standardization EN ISO/IEC 17025, General requirements for the competence of testing and calibration laboratories, ICS: 03.120.20, stage 90/93 (2010-12-15)
[12] – D. van Leeuwen and N. Brümmer, *An Introduction to Application-independent Evaluation of Speaker Recognition Systems*, in *Speaker Classification I: Fundamentals, Features, and Methods*, Christian Müller (Ed.),Springer 2007.

Chapter 1

A Framework for Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation

ABSTRACT

In this chapter the Likelihood Ratio (LR) inference model will be introduced, the theoretical aspects of probabilities will be discussed and the validation framework for LR methods used for forensic evidence evaluation will be presented. Prior to introducing the validation framework, following questions will be addressed: ***"which aspects of a forensic evaluation scenario need to be validated?"***, ***"what is the role of the LR as part of a decision process?"*** and ***"how to deal with uncertainty in the LR calculation?"*** The answers to these questions are necessary to define the validation strategy based on the validation criteria. The questions: ***"what to validate?"*** focusing on defining validation criteria and methods, and ***"how to validate?"*** dealing with the implementation of a validation protocol, form the core of this chapter.

The validation framework described can be used to provide assistance to the forensic practitioners, when determining the suitability and applicability of the LR method developed in the forensic practice by introducing performance characteristics, performance metrics, validation criteria and the decision thresholds. This chapter will start with the introduction of the LR inference model, followed by the validation framework proposed.

1. The LR method as a part of the inference model

The likelihood ratios in this chapter and throughout this thesis are computed from biometric scores following the Bayesian inference model, hereafter inference model.

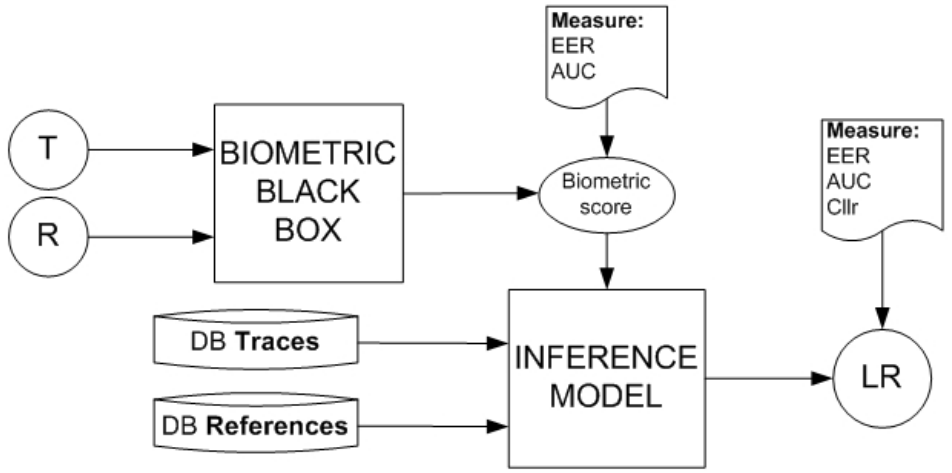


Figure 1 – LR as a part of the decision process

Biometric scores, as presented in *figure 1* are the result of the trace-to-reference sample comparison. Throughout the biometric modalities, this comparison can be performed using off-the-shelf commercial automated systems (in fingerprint modality the Automatic Fingerprint Identification System – AFIS). It is common that the forensic practitioner has very little control over the resulting biometric score (in speaker recognition these biometric scores take the form of a LR). These systems are commonly referred to as Biometric Black Box (BBB).

In the fingerprint modality a fingermark (trace - T) and a fingerprint (reference – R) under evaluation are presented to the biometric black box. The AFIS system performs the feature extraction and comparison and produces a discriminative score of a certain magnitude. The performance of the BBB can be evaluated based on the scores produced. Typical tools include the Decision Error Trade-off (DET) plots where the Equal Error Rates (EER) can be measured or Receiver Operating Characteristics (ROC) from which the Area Under Curve (AUC) can be calculated.

Consecutively the score feeds the inference model together with the database of traces (DB Traces) and database of references (DB

References), where depending on certain assumptions (different aspects discussed below) a LR method is used to produce the likelihood ratio. In ideal conditions DET plots and the EER of the biometric scores and the resulting LR after applying the inference model should be the same.

1.1 Aspects influencing the choice of the LR method

There are several aspects that need to be taken into consideration when choosing a LR method. Good examples of these aspects are the generation of the propositions, the calculation of the evidence, the evaluation of the evidence under the propositions, the choice of the evaluation datasets (in fingerprints modality commonly referred to as conditioning/anchoring).

1.1.1 Generation of the propositions

The propositions under evaluation are usually generated during the case pre-assessment phase, in which the likely and relevant propositions are distinguished from the less relevant ones. It is worth mentioning, that there might be more than two propositions that are relevant to the case [1,2].

In order to evaluate the strength of the evidence ***E*** in a LR framework, we need a pair of mutually exclusive propositions – one for the prosecution $\boldsymbol{H_p}$ and one for the defence $\boldsymbol{H_d}$:

- $\boldsymbol{H_p}$: The trace originates from the individual/object suspected to be the source
- $\boldsymbol{H_d}$: The trace originates from another individual/object than the one suspected

The LR (equation 1) compares the probability of observing the evidence (*E*) under either of these propositions:

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} \quad \text{(eq. 1)}$$

The LR is derived from the Bayes theorem (equation 2) in a following way:

$$\frac{P(H_p \mid E)}{P(H_d \mid E)} = LR \times \frac{P(H_p)}{P(H_d)} \quad \text{(eq. 2)}$$

in which the LR, multiplied by the prior probability ratio $P(H_p) / P(H_d)$ is equal to the posterior probability ratio $P(H_p|E) / P(H_d|E)$.

Probabilities vs. probability densities – the main difference is in the type of data used. When dealing with discrete data we express the LR using probabilities.

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} \text{ (eq. 3)}$$

where again the E denotes the evidence and the prosecution and defence propositions are abbreviated using the H_p and H_d.

When dealing with continuous data we express the LR for example using a probability density function *f*.

$$LR = \frac{f(E \mid H_p)}{f(E \mid H_d)} \text{ (eq. 4)}$$

1.1.2 Calculation of the evidence

The evidence is most of the time calculated as a discriminative score resulting from the comparison of the features extracted from the crime-scene trace and a reference specimen collected from the suspect (biometric score of a fingermark and fingerprint comparison in case of the automatic fingerprint feature comparison algorithm). For automatic methods, this calculation is made using feature extraction and feature comparison algorithms.

1.1.3 Evaluation of the evidence under the propositions

A LR method is used to interpret the discriminative score as strength of evidence. Since the LR method can, in the simplest case, consist of parametric modelling the same-source – SS and different source – DS distributions, it can be referred to as a LR model. A detailed description of the LR method used, derived from [3], is beyond the scope of this chapter and more details on LR models can be found in chapters 4 and 5. Recall that the set of propositions from the section *2.1* is important to select the most relevant data and conditioning to fit with the circumstances of the case.

Having defined a set of propositions against which the biometric score will be evaluated, one can proceed to build the LR model for example in a following way:

- use the minutiae comparison algorithm to compare a crime-scene fingermark to a fingerprint of the suspect to compute the evidence score (E)
- use the minutiae comparison algorithm to compare the fingermarks of the suspect with the fingerprint of the suspect, obtaining a same-source score distribution (SS)
- use the minutiae comparison algorithm to compare a crime scene fingermark to a database of fingerprints of "other than the suspected" individuals, obtaining a different source score distribution (DS)
- model the SS and DS distributions reflecting the conditions set by the two propositions
- compute the strength of the evidence using *equation 2*

It should be noted here that the evidence evaluation following the procedure described above represents the simplest case from the modelling point of view, in which the LR values are calculated parametrically based on the score distributions in the numerator and denominator of the LR. In reality more complex models (for example non-parametric) could be used instead. In this case the procedure described above may involve more steps. In reality the simplest approach would be the one containing the least amount of assumptions in the inference process.

1.1.4 Choice of datasets

Conditioning on different types of data (in fingerprints commonly used conditioning configurations are *person dependent / person independent*) can be defined as a result of LR method using different sets of data, satisfying the propositions H_p and H_d. Since the majority of the LR methods used for evidence evaluation are data-driven/dependent, conditioning on different types of data will affect the resulting LR [1]. This issue has also been described in [2,4]. Also there may be several types of data satisfying the propositions – some more general and some more case specific.

Datasets chosen for the validation of LR methods can be real or simulated forensic data. For the validation experiments the choice of the data is made according to their properties, such as known ground truth, quantity and quality. The data are constituted of pairs of specimens, the reference material and the trace material.

Some concerns have been expressed regarding the use of simulated data. Real data are preferred[1] over simulated data, but simulated data brings

[1] The preference for the real forensic data is solely based on the ambiguity linked to the origin of the simulated data and t he way the simulated data was produced. Establishing a degree of similarity/divergence between the simulated and the forensic datasets has been

considerable added value to real data, especially in the LR method development, when for example the variation in the data is extremely difficult to model (such as modelling distortion in fingerprints). The fact that real data are often limited in number, representativeness (sample bias) and may present outliers or missing values also advocate for the use of simulated data. A good practice and a minimum requirement for the use of simulated data would be to establish a degree of similarity to which the simulated data corresponds to the real forensic data (for example using methods such as Kullback-Leibler divergence, visual representations or other).

- **Ground truth** – The ground truth regarding the origin of the data is usually known for simulated forensic data and according to their source we can label the datasets as originating from the same source (SS) or different source (DS). For the real forensic data the ground truth is per definition unknown, but in some particular cases a ground truth *by proxy* can be assigned. This pragmatic approach is only satisfactory from a methodological point of view, if there are reliable indicators of the similarity between the ground truth by proxy and the reference. These indicators can be intrinsic to the data, for example when this data, and particularly the trace material, are of such high quality that there is extremely strong evidence for the trace to belong to a given source. The indicators can also be external to the data, for example the existence of case information related to the data allowing to induce their origin.
- **Quality of the data**: the quality of the data can be understood as a value that has no information about the proposition, which is true in a comparison, but despite of this, it can predict performance of that comparison. In other words, samples of high quality to compare in a forensic case predict good performance of that comparison, and low quality predicts bad performance. Under this definition, more robustness to variation or degradation indicates less loss of performance measure as the quality decreases.
- **Quantity**[2] **of the data**: the quantity of the data is a value or a component that may be expressed in numbers (Oxford dictionary /

deemed desirable (for example using the Kullback-Leibler divergence, however other measures can be used instead).

[2] Quality is not an intrinsic factor; it should always be evaluated relative to the purpose. In general, one can speak about quantity of information (with respect to the coherence performance characteristic) and ability to exploit, extract, compare and evaluate this information (with respect to the robustness). We can use a fingermark example - it is different to have a partial fingermark with 5 minutiae visible or a partial fingermark with 12 minutiae visible, from which only 5 can be used with the state of the art technology. The strength of evidence of the 1st one is intrinsically limited to 5 minutiae; the strength of the evidence of

Mathematics & Physics), e.g. length of the speech fragment, number of minutiae in a fingermark, etc.

- **Representativeness of the data**: the representativeness of the data refers to the variation of the performance characteristics to a change in the data used to measure such performance. Therefore, a LR method will be more representative if the performance varies less when two different datasets are used.

2. The LR in the forensic evaluation process

LR methods are used across multiple forensic disciplines and the LR approach is being extensively used for example for the interpretation the DNA profiles. Some recommendations on the interpretation of the DNA mixtures have been issued in 2006 [5] stating that *"The court may be unaware of the (LR) method if the scientist does not attempt to introduce it"*, meaning that an attempt should be made by a scientist to explain the LR method in the simplest way possible to the court of justice. Recommendation 1 of this article it states that ***"LR is the preferred approach to (DNA) mixture interpretation"***, indicating that there are other methods (Random Man Not Excluded) which don't possess the same qualities as the LR approach, while in recommendation 2 of this article it states that ***"Even if the legal system does not implicitly appear to support the use of the likelihood ratio, it is recommended that the scientist is trained in the methodology and routinely uses it in case notes, advising the court in the preferred method before reporting the evidence in line with the court requirements"***.

Forensic research makes progress in the field of evaluation of forensic evidence. Currently a uniform and logical inference model is used for evaluating and reporting forensic evidence [6]. It uses a likelihood ratio (LR) approach based on the Bayes Theorem. Standards have been proposed for the formulation of evaluative forensic science expert opinion in UK [7]. A similar initiative is in progress in Europe, within the European Network of Forensic Science Institutes (ENFSI), the ENFSI Monopoly Project M1-2010 entitled "*The development and implementation of an ENFSI standard for reporting evaluative forensic evidence*" [8].
Computer-assisted methods also have been developed to compute LRs, assisting the forensic practitioners in their role of forensic evaluators to perform inferences at source level [9]. Very early principles for using the LR approach in forensic evaluation can be found in the analysis of glass microtraces [10]. It has also been used in forensic evaluation fields focusing

the 2nd one is limited by the current state-of-the-art (the impossibility to exploit 7 minutiae because of lack of robustness of the minutiae comparison algorithm).

on human individualization, such as fingermark [11,12], earmark [13], speaker recognition [14] and hair [15]; or object individualization such as toolmarks [16], envelopes [17], fibre [18] and glass microtraces [19] (which represents a very early practical example of the use of the LR approach). But the LR approach has been firstly implemented in a casework process as a standard for the evaluation of DNA profiles [6].

2.1 Validation of LR methods

The EU Council Framework Decision 2009/905/JHA [20] on the ***"Accreditation of forensic service providers carrying out laboratory activities"*** regulates issues related to the quality standards in two forensic areas: DNA-profile and fingerprint/fingermark data. This decision framework seeks to ensure that the results of laboratory activities carried out by accredited forensic service providers in one member state are recognized by the authorities responsible for the prevention, detection and investigation of criminal offences within any other member state. Equally reliable laboratory activities carried out by forensic service providers are sought to be achieved by the EN ISO/IEC 17025 accreditation of these activities [21]. For this reason, this framework focuses on the *General requirements for the competence of testing and calibration laboratories* as described in the EN ISO/IEC 17025 norm, and particularly on the requirements for the validation of non-standard methods in the section 5.4.4, as we consider the LR methods used for forensic evaluation as non-standard methods.

To foster cooperation between police and judicial authorities across the European Union member states, the ***"Vision for European Forensic Science 2020"*** of the Council of the European Union DS 1459/11 [22] proposes to create a European Forensic Science Area. Member States and the Commission will work together to make progress in several areas, aiming to ensure the even-handed, consistent and efficient administration of justice and the security of citizens. Amongst them several are related to the validation of the methods used for forensic evaluation:

- accreditation of forensic science institutes and laboratories
- establishment of common best practice manuals and their application in daily laboratory work
- application of the principle of mutual recognition of law enforcement activities with a forensic nature with a view to avoiding duplication of effort through cancellation of evidence owing to technical differences, and achieving significant reductions in the time taken to process crimes with a cross-border component
- research and development projects to promote further development of the forensic science infrastructure

2.2 Necessity for guidelines

Because the computer-assisted methods for forensic evaluation are still very new, the EN ISO/IEC17025 [21] and the ILAC-G19:2002 guideline for forensic laboratories [23] do not address the question of their validation. They mainly address the question of the validation of instrumental methods used for analytical purpose. More recently an explanatory document of the Dutch accreditation body, RvA-T015 issued in 2010 [24], provided some guidelines for the validation of the opinions and interpretations of forensic practitioners. In short, the criteria proposed for the validation of instrumental analytical methods are based on performance and the approach for the validation of the human-based methods used for interpretation is based on competence assessment.
As the existing criteria used for interpretation only focus on human-based methods, they are not suitable for the validation of computer-assisted methods developed for forensic evaluation.

2.3 Preliminary consideration

In the forensic community there are major differences in the understanding of the concept of probability and of the LR were observed, which has direct consequences on the definition of the criteria for the validation of computer-assisted LR methods developed for forensic evaluation. Therefore some of the points of view regarding the concept of probability and of the LR are discussed prior to the main discussion about the performance characteristics and criteria.

3. The LR as a part of the decision process

Several roles are devoted to the forensic scientists. The first role is dedicated to the forensic methodology. The forensic methodologists conceive new approaches and solutions to specific forensic open questions, for example the current attempt to find an adequate approach for the validation of computer-assisted methods developed for forensic evaluation.

The second role focuses on the development. In the forensic research and development stages, a part of the role of the forensic developer is to test methods for forensic evaluation in the whole range of their application. In the validation stage, the range of validity of the LR method is tested in a Full Bayesian inference model, taking into account the prior probabilities of the propositions, the LR, the posterior probabilities of the propositions and the decision thresholds. The forensic developers create new technologies or adapt existing technologies for some specific forensic purpose, like for example the development of computer-assisted methods for forensic

evaluation. In these circumstances (development and validation) the forensic developer will consider the LR as part of a decision process and simulate the functionality of the methods developed for the whole range of decision thresholds (whole range of priors and decision costs or utilities).

The third role focuses on the forensic practice. The forensic practitioners introduce new methods and use them for casework, for example using computer-assisted LR methods in their forensic evaluator role. In casework the forensic evaluator plays a role of neutral facilitator. The purpose is to consider the strength of the evidence regarding the alternative propositions provided by the criminal justice system, at least one proposition from the prosecution and one from the defence. Therefore as an evaluator, a responsibility for the forensic practitioner is to obtain the most relevant alternative propositions to be considered in the case; to provide the most correct strength of the evidence in form of a LR. In some particular cases the forensic practitioner can also supply relevant forensic information unknown from the trier of the fact to help to assign the prior probabilities. The forensic evaluator has also the responsibility to understand the scope and limitations of the method used, which are described in the validation report. The forensic evaluator should be careful not to be too prescriptive towards the trier of the fact, since there are legal standards and laws that are out of the scope and competence of the forensic evaluator.

The trier of the fact has also the *"freedom of proof"*, meaning (s)he can, in some legal systems with due motivation, decide not to follow the statement of the forensic practitioner. In that sense the forensic evaluator remains an advisor, while the assignment of the prior and posterior probabilities and the decisions made on this basis are the responsibility of the criminal justice system, or the court in general.

4. Validation strategy

Two important components, identified for the validation of computer-assisted LR methods used for forensic evaluation are a theoretical validation and an empirical validation of the inference model. The theoretical validation of the BBB rests upon the mathematical proof or falsification (not handled in this thesis) and the empirical validation of the LR method rests upon the acceptance or rejection of validation criteria.

4.1 Theoretical validation

Where applicable, the theoretical validation is handled using the falsifiability approach [25], focusing on proving / disproving mathematical formulae, propositions, lemmas and theorems, in general assuming that there is a

ground truth (trueness) of a given statement that can be falsified (disproved or nullified). This part of the validation is deductive (deductive reasoning), since it relies on mathematical properties and does not imply assumptions. The choice of any (LR) method needs to be validated empirically using appropriate measure of performance, even if it seems ***"theoretically so well grounded"*** that it may appear as mathematically correct. The term **"*theoretically so well grounded*"** should be approached with moderation; it refers to situations where the choices within a method are solidly grounded, for example based on deductive reasoning, justifying its use by proofs and mathematical rigor.

4.2 Empirical validation

The empirical validation focuses on the acceptance or rejection of chosen validation criteria. This part of the validation is inductive as it implies assumptions regarding the inference model(s) used for the evidence evaluation.

The empirical validation incorporates a definition of the validation protocol and experiments, in order to demonstrate the acceptance / rejection of the chosen validation criteria. Where a validation process leads to quantitative results, a range of variable in which the LR method gives acceptable performance will be presented. The following elements have been deemed important and determine the structure of the validation protocol:

- performance characteristics
- performance metrics
- graphical representations
- validation criteria
- experiments
- datasets
- analytical results
- validation decision

The order of the elements determines the structure of the protocol. The performance characteristics and the related performance metrics need to be identified. The validation criteria need to be established, such as the numerical threshold expressed in terms of the performance metrics chosen. An experiment (or series of experiments) has to be designed for the LR method under evaluation and appropriate sets of data have to be chosen for each step of the validation protocol. Each result produced on this basis is confronted with the appropriate validation criterion, in order to achieve a validation decision which would ideally take a binary form – favour either the acceptance / rejection of the LR method validated. The conclusion of an

empirical validation should be conditioned by all the assumptions made in the validation protocol, which should be mentioned explicitly at the beginning of the validation report.

The scope of validation should be defined prior to the empirical validation of a LR method. Where applicable, requirements should be described in a form of thresholds for each validation criterion and overall desired functionality of this LR method. These thresholds can for example obtained by a comparison with the ***"state-of-the-art"***. In absence of existing thresholds due to the novelty of the LR method, the thresholds can be specified based on the functionality of a ***"baseline method"***. Such requirement can be formulated for example in a following way for a fingerprint LR method: "***Equal Error Rate of LR method under evaluation using a NIST SD27 database <= 5%***"[3] or "***CLLR of LR method under evaluation smaller than the baseline LR method***". Different aspects of empirical validation, broken down into necessary steps and categories are structured in the table 1 below:

Table 1: Aspects of empirical validation

Validation Aspects	Performance Characteristic	Performance Metric	Graphical Representation
Primary performance characteristics	Accuracy	Cllr	ECE plot
	Discriminating power	EER, $Cllr^{min}$	ECE^{min} plot DET plot
	Calibration	$Cllr^{cal}$	Tippett plot
Secondary performance characteristics	Robustness	LR range	ECE plot DET plot Tippett plot
	Coherence	Cllr, EER	ECE plot DET plot Tippett plot
	Generalization	Cllr, EER	ECE plot DET plot

5. Propopsed performance characteristics

As an outcome of the validation workshop, several performance characteristics have been identified for the validation of computer-assisted LR methods developed for forensic evaluation. Some of these were already defined, though the workshop helped to structure them and to clarify their

[3] As mentioned in the first paragraph of this chapter, the EER can be measured already at the biometric score level. Propagation of the discriminating properties of the Biometric Black Box is a desirable property of a *good* inference model.

role. They are now structured in primary and secondary characteristics. The primary characteristics of the LR method under evaluation are related directly to performance metrics and focus on desirable properties (e.g. goodness of a set of LR values, in which we are assessing whether a set of LR values is good or bad, adequate or non-adequate, whether it has desirable properties or not). The secondary characteristics describe how the primary metrics behave in different situations, in some cases simulating the typical forensic casework conditions (e.g., specimens of degraded quality, varying quality conditions between the training data and the crime scene samples, etc.). The difference between the "primary" and "secondary" metrics is that the primary ones directly measure desirable properties of the LR, while the secondary complement the primary ones, and measure/present how the primary measures vary in different conditions (for instance quality of the data or quantity of information). The secondary characteristics may relate to a single primary metric. For instance, generalization may refer to the variation of Cllr (primary metric) when varying the amount of data.

Originally performance characteristics have been defined in the context of validation of analytical methods for the measurement of physical and chemical quantities (metrology). The definitions of these performance characteristics can be found in the International Vocabulary of Metrology (VIM) [26]. The performance characteristics proposed for the forensic evaluation methods (shown below in table 1) have been chosen on the basis of their similarity with the original performance characteristics defined for the validation of analytical methods. To prevent confusion between the original and newly defined performance characteristics, we present both definitions in parallel in the sections 5.1 to 5.3. Where the VIM does not provide an exact definition, analogous definitions are extracted from sources cited in the ENFSI 2013 Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science [27], keeping in mind that the fact that the two documents do not have the same status.

5.1 Proposed primary performance characteristics

For forensic evaluation methods, three primary performance characteristics have been identified (presented below in table 2):

Table 2: Definitions of the primary performance characteristics for LR methods

Performance characteristics	VIM definition or other authoritative definition	New definitions for forensic evaluation methods
Accuracy[4]	*"Closeness of agreement between a measured quantity value and a true quantity value of a measure"* Closely linked to the accuracy is the precision, in VIM defined as follows: "Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions"[5]	Closeness of agreement between a LR computed by a given method and the ground truth status of the proposition in a decision-theoretical inference model. The LR is accurate if it helps to lead to a decision that is correct according to the ground truth of the propositions. In case of source level inference, the ground truth relates to the following pair of propositions: • H_p: the pair of samples tested originate from the same source (SS) • H_d: the pair of samples tested originate from different sources (DS) If an experimental set of LR values is to be evaluated, and the corresponding ground-truth labels of each of the LR values are known, then a given LR value is evaluated as more accurate if it supports the true (known) proposition to a higher degree, and vice-versa

[4] In analytical methods accuracy and precision imply the existence of a true magnitude of certain physical phenomena that is to be measured. One can for instance measure the short side of a standard credit card, and performing 100.000 measurements arrive to a certain probability density. There is a "true" (exact) value in this case – the exact value of the short side of a credit card is in reality 53.98mm. By performing additional measurement (obtaining a size of 63.98mm) the accuracy/trueness then relates to the systematic error and represents distance (10mm in this case) between the reference value and the "true" value.
On the other hand we understand, that due to the definition of the LR as being the result of a ***probabilistic inference*** and not a ***measurement***, no quantitative ground truth exists for the LR because of the ***"Bayesian interpretation of probabilities as a degree of belief"*** [4]. Therefore it is **not possible** to establish **univocal relation** between **a pair of samples** and a numerical **likelihood ratio value.**

[5] In [30] the accuracy is deemed equal to validity and presision is deemed equal to reliability. In this work the validation is regarded as a process, rather than a single measurable entity.

Discriminating Power	*"Discriminating power of a series of k attributes is defined as probability that the two distinct samples selected at random from the parent population would be discriminated in at least one attribute if the series of attributes were determined. The distribution of each attribute over the population is assumed to be known from a study of a large number of samples" [28]*	Performance property representing the capability of a given method to distinguish amongst forensic comparisons under each of the propositions involved
Calibration (Calibration loss)	*"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.*" The concept of calibration used in the context of analytical methods has nothing to do with the definition of calibration used in statistics.	In probabilistic terms can be defined as the property of a set of LRs. Perfect calibration of a set of LR's means that those LR's can probabilistically be interpreted as the strength of evidence of the comparison result for either proposition. Under those conditions the LR is exactly as big or small as is warranted by the data. The strength of evidence of well-calibrated LRs tends to increase with the discrimination power for a given method [32].

5.2 Proposed secondary performance characteristics

The following secondary characteristics have been identified (presented below in table 3):

Table 3 - Definitions of the secondary performance characteristics

Performance characteristics	VIM definition or other authoritative definition	New definitions for forensic evaluation methods
Robustness	The robustness / ruggedness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal use" definition given in [27]	The ability of the method to maintain performance (e.g., Cllr) when a measurable property in the data changes. For instance, method A is more robust to the lack of data than method B if, as the data gets sparser, the performance of method A degrades relatively less than the performance of method B. Note 1: Good indicator of LR method not being robust to the lack of data is when the LR method produces LR's of unreasonable and not explicable magnitudes (e.g. LR = infinity). Note 2: When talking about robustness in forensic science, most of the time we speak about the stability of the method to the forensic conditions detrimental to the quality/quantity of data that prevent reliable measurement of the information, or of the features carrying the information.

Coherence	Not defined in the VIM Oxford Dictionary: • The quality of being logical or consistent The quality of forming a unified whole	The ability of the method to yield LR values with better performance with the increase of intrinsic quantity/quality of the information present in the data. It focuses on the variation of some measurable parameters[6] in the features[7] studied, perceived as influencing the strength of evidence, like the quantity of minutiae in the fingerprint field or the signal to noise ratio in the speaker recognition field.
Generalization	Any statement ascribing a property to every member of a class (universal generalization) or to one or more members (existential generalization) Example: Every function is a relation but not every relation is a function. **Collins English Dictionary: Logic**	Property of a given method to maintain its performance under dataset shift. *A dataset shift occurs when the joint distributions of inputs and outputs differs between the training data (used to build the LR methods) and the testing data (previously unseen)* [29] used to compute LRs in operational conditions. **For instance, LR method trained on a dataset A generalizes well to a dataset B if the LR method maintains its performance.**

6. Performance metrics and their corresponding graphical representations

For each performance characteristics, the performance metrics and the associated graphical representations will be presented in this section.

6.1 Decision Error Trade-off (DET) plot and Equal Error Rate (EER)

The main idea behind the DET plot is linked to "thresholding" of a biometric score (or a LR) and ability of the BBB (or an inference model) to make decisions based on the decision errors – the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). In biometric terms, the FAR refers to a likelihood of a biometric system (or an inference model) to accept an

[6] Parameter can be seen as a measurable value of the degradation of the extracted features due to forensic conditions (signal to noise ratio, distortion, clarity). LR method can be the robust to these parameters.

[7] Feature is to be understood as a carrier of information extracted from raw data. Coherence is related to the information carried by the features.

unauthorized claim, while the FRR refers to a likelihood of a biometric system (or an inference model) to reject authorized claim. DET plot then represents a trade-off between these decision errors.

DET plot defined in [31] is a 2 dimensional plot in which the FAR is plotted as a function of the FRR. The error rates are consecutively plotted on a Gaussian-warped scale. Thus, linearity of the DET curves happens when de distribution of the log(LR) values is normal. The closer the curves to the coordinate origin, the better are the discriminating capabilities of the method. The intersection of a DET curve with the main diagonal of the DET plot marks the Equal Error Rate (*EER*), which will be used as a performance measure to show the *coherent* behaviour of the LR method (for example when comparing forensic fingermarks in different minutiae configurations *EER_5minutiae* < *EER_10minutiae* as presented in figure 2 below). Even if the DET plot is meant to characterize a discrimination system (implying a decision) the information provided indirectly informs about the coherence of the LR method when evaluating datasets with different quantity of information (for example different number of minutiae in fingermark evidence evaluation).

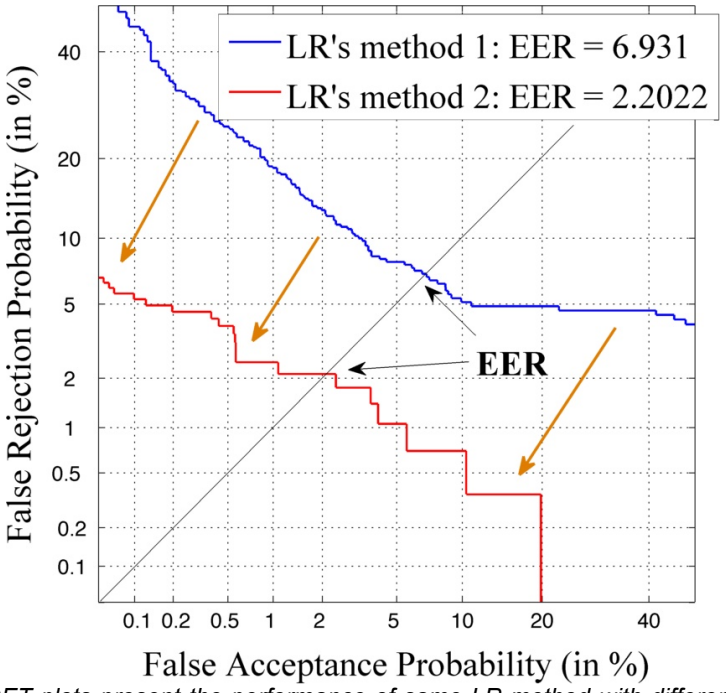


Figure 2 - DET plots present the performance of same LR method with different quantity of information. Blue line represents a method showing less evidential information captured in the LR of fingermark to fingerprint comparison for 6 minutiae, while the red line shows more evidential information captured in the LRs of fingermark to fingerprint comparison for 10 minutiae configuration.

6.2 Tippett plots

The Tippett plots [3] are representations of inverse cumulative density functions of LR's. Each of the curves represents the decay of the proportion of the LR's supporting one of the competing propositions. In the Tippett plots rates of misleading evidence can be observed when either of the proposition is true. These rates are visible at the intersection of each of the inverse cumulative density lines for either LR same source or LR different source and the imaginary line going through value zero on the X-axis. The log(LR) value zero on the X-axis on the log scale corresponds to the LR value of 1.

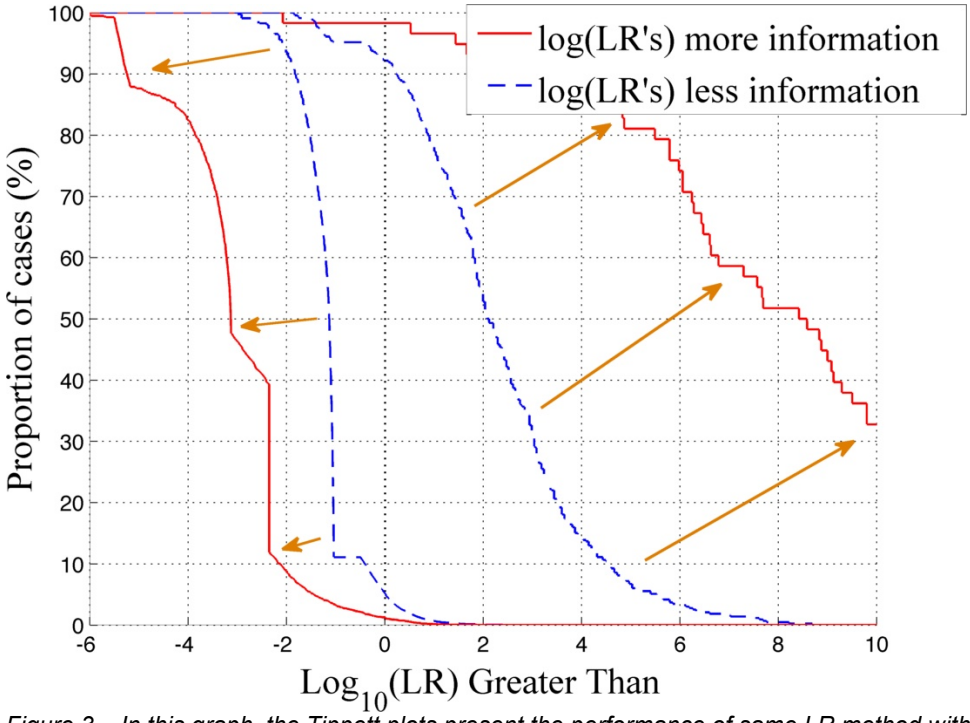


Figure 3 – In this graph, the Tippett plots present the performance of same LR method with different quantity of information. Dashed blue line represents a method showing less evidential information captured in the LR of fingermark to fingerprint comparison for 5 minutiae, while the solid red line shows more evidential information captured in the LRs of fingermark to fingerprint comparison for 10 minutiae configuration.

On the Tippett plots, it is relatively easy to distinguish the quantity of the evidential information within the LR values captured by the LR method presented with datasets in different conditions. Tippett plots of a LR method evaluating the strength of evidence in fingermarks with 6 minutiae configuration (blue dashed line) and 10 minutiae configuration are presented in figure 3. Orange arrows indicate the increase of the surface of

the two curves in the Tippett plots when LR method is presented with additional information (here additional minutiae).

6.3 Empirical Cross-Entropy (ECE) plot and the Log likelihood ratio cost (Cllr)

The ECE plot [32,33] has been deemed *"a useful representation of the performance and calibration of the LR values"* and *"an excellent complement of other already established methods (e.g. Tippett plots or DET plots)"* [32].

ECE and Cllr tend to get lower when the likelihood ratio leads to the correct decision. The difference relies on the interpretation of both measures. Cllr is interpreted as an average decision cost for all prior probabilities and costs involved in the decision process. On the other hand, ECE has an information-theoretical interpretation as the information needed to reach the correct value of the proposition, on average in a given set of LR values. Cllr is an average over costs and priors, and therefore is not giving the performance for a given value of the prior, but for an average of all possible priors. ECE can be represented as an ECE-plot, showing its value for a certain range of priors [32,33]. In fact, both measures are related, and it can be easily shown that Cllr is ECE at the prior probability of 0.5. In this sense, ECE seems a more general and interpretable performance metric than Cllr in a forensic context in which no decision is to be made by the forensic evaluator and in which the value of the prior changes very much from one case to another one. It also appears to be more suitable for the forensic practice, in which the aim is to show the range of application (scope of validity) of the LR method over a relevant set of priors, which are in general unknown to the forensic evaluator. On the other hand, Cllr is a single scalar measure, useful for ranking and comparison, and it in fact summarizes ECE.

In [34] the Cllr is defined in a following way:

$$Cllr = \frac{1}{2 \cdot N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d} \sum_{j_d} \log_2 \left(1 + LR_j\right) \qquad \textit{(eq. 5)}$$

where the N_p and N_d are the number of target (same source) / non-target (different source) scores under evaluation, while the i_p and j_d indices present sum over the target / non-target set of LR's.

In [32,33] the ECE is defined in a following way:

$$ECE = \frac{Q(\theta_p)}{N_p}\sum_{i_p}\log_2\left(1+\frac{1}{LR_i\cdot\frac{P(\theta_p)}{P(\theta_d)}}\right)+\frac{Q(\theta_d)}{N_d}\sum_{j_d}\log_2\left(1+LR_j\cdot\frac{P(\theta_p)}{P(\theta_d)}\right) \quad \textit{(eq. 6)}$$

where the $P(\theta_p)$ and the $P(\theta_d)$ represent the prior probabilities of the propositions under evaluation and the $Q(\theta_p)$ and $Q(\theta_d)$ denote the reference probabilities.

Closely related to the ECE plot are the measures of accuracy *Cllr*, discrimination $Cllr^{min}$ and calibration $Cllr^{cal}$ [34,35]. The *Cllr* can be found on the intersection of the red (solid) curve in the ECE plot with the $Prior_{log}odds = 0$ (the lower the *Cllr* the better performance of the system); the $Cllr^{min}$ can be found on the intersection of the blue (dashed) curve with the $Prior_{log}odds = 0$ (the lower the $Cllr^{min}$ the better the discrimination of the LR method – see [31,32] for details); while the difference between these two lines on the intersection with the $Prior_{log}odds = 0$ represents the $Cllr^{cal}$ (the smaller the distance, the better the calibration of the LR method).

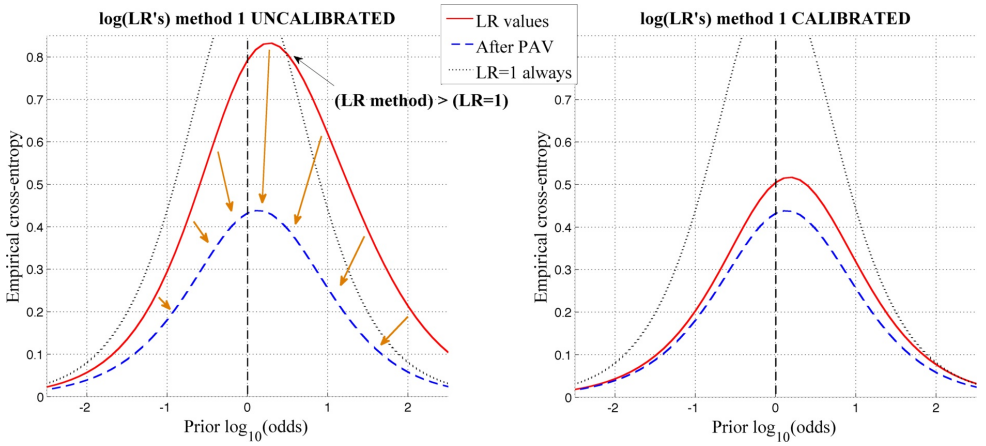


Figure 4 – ECE plots of the same method, using the same data. On the left-hand-side the LR method is uncalibrated while on the right-hand-side calibrated. The main drawback of the uncalibrated LR method, apart from obviously higher Cllr value and greater calibration loss, is the fact that around Prior log_{10}(odds) = 0.5 the uncalibrated LR method crosses the reference method (outputting always LR = 1). Loosely translated for Prior log_{10}(odds) > 0.5 the uncalibrated method performs worse than a method constantly returning the "I don't know" answer.

Besides the information-theoretical aspect, the ECE provides another interesting insight – that is the "range of application" of the LR method under

evaluation. We can safely assume that one of the most desirable properties of a LR method should be to obtain “good” performance for the whole range of priors comparing to the reference method producing LR = 1 all the time (equivalent to I don’t know). Such a reference method has an interesting property – in long term it is perfectly calibrated, it is however as well completely useless for making predictions. Since the accuracy of a LR method in terms of *Cllr* represents “goodness” of predictions of the LR method under evaluation, a LR method can be deemed “good” if the *Cllr* values produced by a LR method don’t exceed the ones of the reference method.

Figure 4 presents the ECE plots of the LR method using the fingermarks in 5 minutiae configuration in 2 different settings – uncalibrated and calibrated. The event of LR method calibration not only minimizes the calibration loss of the LR method (here measured by the *Cllr*), it also extends the range of application of this method. While the range of application of the uncalibrated LR method in terms of *Prior* log_{10}*(odds)* is <-2.5,0.5> (intersection of the red solid line and the black dotted line in the ECE plot in figure 4 left), the range of application of the calibrated LR method is <-2.5,2.5> (figure 4 right).

7. Validation Experiment

Before entering validation experiments, a set of validation requirements should be established. This can be done in two ways – either by examining the current state of the art or by establishing a baseline LR method from which the initial set of performance measures will be compared.

Validation experiment itself should be divided into two stages – the method development stage and the validation stage. In the method development stage we propose to deal with processes related to the method selection, method training and method testing and measure primary performance characteristics as well as the generalization factor.

In the validation stage we evaluate the LR method performance on the validation dataset (with a known ground truth) and measure the method response to the previously unseen data by measuring both – primary and secondary performance characteristics. An example of a validation procedure is shown in *figure 5*.

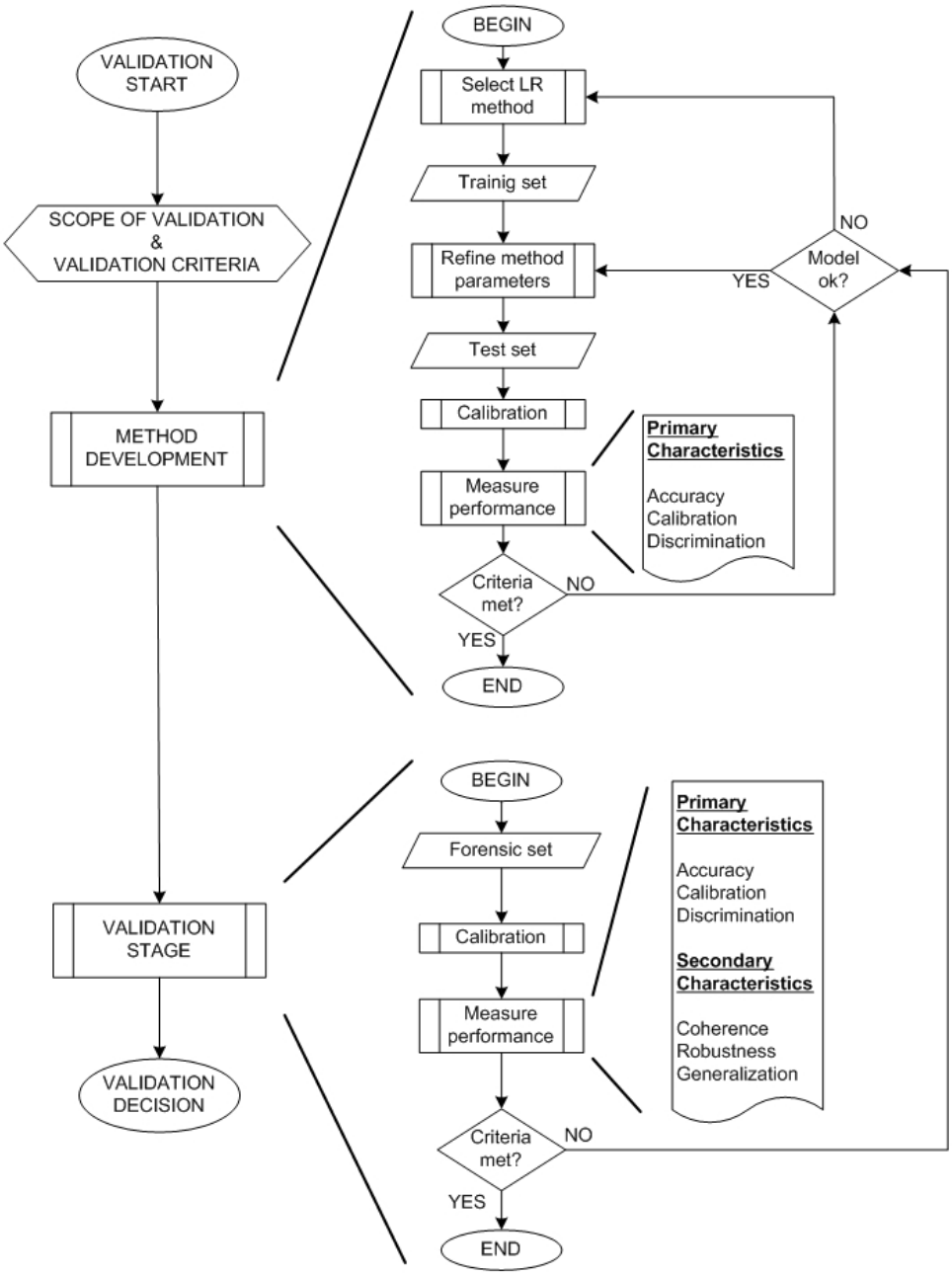


Figure 5 – Example validation procedure

7.1 Method development stage

Recall that the main objective of the LR method development stage is to establish inference models with the most relevant data and the most robust statistical models in order to provide the most correct LRs in the widest scope of conditions possible. We use the primary performance characteristics applying the measures proposed, against which the LR method will be validated in the validation stage using the real forensic data. The training dataset is used to calculate the parameters of the LR method, while the test dataset is used to establish the robustness of the LR method to the previously unseen data.

The LR method development stage of the validation framework uses independent datasets for the training and the test phase, in order to confront the method to previously unseen data of the same quality. The training dataset is used to define the parameters of the given method, while the testing dataset is used to establish the robustness, coherence and generalization of the method to previously unseen data. It is common practice in biometrics to test the robustness of a method based by using different training and testing datasets. The real difficulty is to determine a priori, whether the "previously unseen" test data has similar properties as the training data. This is easy to accomplish when splitting one dataset, however can pose a significant challenge when using two datasets acquired in different conditions (e.g. simulated and real data). A single dataset can be split into a training and test sub-sets, which should be non-overlapping, independent (previously unseen) and representative. Inadequate split of one single dataset can cause issues known as under / over fitting. In under-fitting the LR method will be a bad representation of the dataset, while in over-fitting the LR method will fit too closely to the training dataset and will be less robust to the previously unseen data. Using the training and test set the cross-validation of the LR method developed in the method development stage is guaranteed. The validation of the LR method then follows the same logic – the dataset used for the validation should be independent and representative with respect to the dataset used in the method development stage. Ideally the real forensic data should be kept for the validation stage, ensuring the functionality of the method developed in real forensic conditions. It is common that well-performing methods in the method development stage lose some of their performance when subjected to real forensic data.

8.1.1 Training dataset

In a simple case where we aim for example to fit a normal distribution to the set of scores, our objective in the training phase is to use the training

dataset to obtain the parameters of the normal function (mean and variance). Usually more complex methods are used instead, such as linear logistic regression, beta or gamma distribution to name a few. However the principle is the same – use the training dataset to obtain the parameters for the LR method.

7.1.2 Test dataset

The test dataset is intended as the sanity check regarding the basic functionality of the LR method. It is used in the method development stage to evaluate the robustness of the LR method to the previously unseen data. In a simple case, we take the LR method developed on the testing dataset (using the parameters calculated on the testing dataset) and measure how closely these parameters fit the test dataset. This process can be repeated several times and is usually referred to as cross-validation ("*n-fold*" cross validation if repeated "*n*" times).

Since the test dataset appears to the LR method as previously unseen, sub-optimal performance of the LR method is expected. Therefore prior to evaluating the performance of the LR method we can calibrate the LR values produced by the LR method on the test dataset (usefulness of the LR method calibration is highlighted in the section 6.3 of this chapter).

If the performance of the LR method is satisfactory and the validation criteria[8] are met, one can proceed further to the validation stage. If following the calibration step the LR method shows sub-optimal performance and fails to meet the requirements set by the validation criteria following steps are possible:

- refine the training parameters of the LR method
- use alternative LR method
- relax the validation criteria

The order in which the steps should be applied should be critically assessed – based on the time / budget constraints. One can spend months trying to refine parameters of a completely ill-performing LR method, whereas an alternative LR method might give comparable (if not better) results. Relaxing the validation criteria should be used as a last resource, however this should be critically addressed in the validation report. It might be therefore a good idea to evaluate the performance of different LR methods in the training stage prior to moving to the test stage.

[8] Validation criteria defined previously in the "validation scope" either in comparison to the state-of-the-art or to the baseline method.

7.2 Validation stage

In the validation stage the LR method developed in the method development stage uses the forensic dataset, in order to evaluate the primary and secondary performance characteristics. As mentioned earlier, the LR method developed on the training dataset may show sub-optimal performance on the previously unseen dataset, mainly due to the dataset shift between the datasets used in the method development stage and the validation stage. Calibration is therefore a mandatory requirement for the validation stage prior the LR method performance evaluation (usefulness of the LR method calibration is highlighted in the section 6.3 of this chapter).
Should the validation criteria not be met by the LR method on the forensic dataset, a logical step is to move back to the method development stage and either refine the existing LR method (taking into account the specific settings of the forensic dataset) or develop alternative LR method.

8. Validation criteria

The validation criteria should address questions like: ***"What to measure?"***, ***"How to measure?"*** as well as ***"What values should be observed or deemed satisfactory?"***

For example, newly developed biometric technologies used as black boxes are subject to evaluation using standardized datasets, in fingerprints, good examples are the databases NIST SD04 or NIST SD27 of the National Institute of Standards and Technology. We shall refer to this approach as "***a comparison with the state-of-the-art***", since the validation criteria can be deduced based on the performance of state-of-the-art algorithms. It should be noted here, that establishing the validation criteria as strictly equal to the performance of the state-of-the-art only makes sense in the case of either using the state of the art algorithm or being sure that the LR method proposed will be able to directly compete against and / or outperform the state-of-the-art, which might be rather challenging.

In case when such a special database does not exist and comparison with state-of-the-art methods is not an option; a baseline method can be developed, for example based on the score distributions (SS and DS) of the training dataset. We shall refer to this approach "***a comparison with the baseline***".

Alternatively, multiple LR methods can be developed at the same time on the training dataset, of which one will play the role of the baseline method from which the validation criteria will be defined. LR methods proposed, including the baseline, should be fit for purpose – a gamma function will not

be a good representation of clearly normal-like training set SS and DS score distributions – thus LR methods obviously not fit for purpose should be eliminated.

9. Validation decision

A validation procedure[9] should be concluded by a binary expression (e.g. pass / fail) regarding the LR method being fit / not-fit for forensic evaluation casework.
A set of recommendations can be issued alongside the validation decision, addressing mainly the shortcomings and limitations of the LR method under evaluation. These may contain applicability range of a LR method, clarity/distortion limits, description of sampling procedures, comparison algorithms used etc.

10. Validation report

The validation of non-standard methods is described in the ISO/IEC 17025 standard in section 5.4.4. *"When it is necessary to use methods not covered by standard methods, these shall be subject to agreement with the customer and shall include a clear specification of the customer's requirements and the purpose of the test and/or calibration. The method developed shall have been validated appropriately before use."* In the section 5.4.4 the ISO standard also lists the information recommended:

a) appropriate identification;
b) scope;
c) description of the type of item to be tested or calibrated;
d) parameters or quantities and ranges to be determined;
e) apparatus and equipment, including technical performance requirements;
f) reference standards and reference materials required;
g) environmental conditions required and any stabilization period needed;
h) description of the procedure, including

> *- affixing of identification marks, handling, transporting, storing and preparation of items,*
> *- checks to be made before the work is started,*
> *- checks that the equipment is working properly and, where required, calibration and adjustment of the equipment before each use,*

[9] The validation of a method should be understood as a procedure that uses the validation protocol.

- the method of recording the observations and results,
- any safety measures to be observed;

i) criteria and/or requirements for approval/rejection;
j) data to be recorded and method of analysis and presentation;
k) the uncertainty or the procedure for estimating uncertainty.

Prior to starting the validation of a LR method, a validation plan should be drawn by a forensic practitioner. It is mandatory for the reader to keep in mind, that the ISO/IEC 17025 standard was predominantly developed for the validation of analytical methods, therefore not all of the recommended information is applicable to the validation of LR methods. Especially the points e), f), g), h), j) and k) will be rather challenging to defend in the interpretation of forensic evidence. In compliance with the remaining recommendations from the ISO/IEC 17025 standard the validation plan should contain (but is not limited to) the following:

- Identification of LR method – point a)
- The intended use – point b)
- The performance characteristics – point d)
- The performance metrics – point d)
- The validation criteria – point i)
- The scope of the validation (Range of application of the LR method) – point b)
- Validation time span (applicable in cases in which the datasets used in the LR method development/validation stage are envisaged to get obsolete)

An example of the validation report is presented in chapter 7 and the readers not interested in all the different aspects regarding the stability of LR, robustness to the previously unseen data, LR method selection, presentation of coherence of the LR method or the Bayesian Networks developed for the evaluation of the first level detail fingermark evidence evaluation are advised to fast forward to the chapter 7.

Acknowledgements

Without going into too much of a detail it should be noted here, that the validation framework proposed reflects the authors point of view on ***"how to validate"*** LR methods used for forensic evidence evaluation, based on the inputs gained on the workshop. Even though a global consensus by all the validation workshop participants was not reached regarding certain aspects, the main objective of the framework proposed is to foster further discussions and to provide a decent starting point for validation of computer-based LR methods. Special credit belongs to all the participants in the validation workshop meeting, without whose inputs it would be impossible to draft this document.

The validation workshop, held in The Hague on 19th and 20th October 2011 was organized to define guidelines for the validation of computer-assisted LR methods developed for forensic evaluation. Topics discussed, namely ***"the LR as part of a decision process"*** and ***"dealing with uncertainty in the LR calculation"*** were deemed necessary to define a validation strategy based on validation criteria. Despite the fact that general consensus was not reached by all the participants on all the topics covered, especially on the topic of ***"Dealing with uncertainty in the LR calculation",*** the credit belongs to all the validation workshop participants, since the discussion with them served as an inspiration for writing the validation framework.

Names of the participants and their affiliation:

D. Ramos, J. Gonzalez-Rodriguez – *Universidad Autonoma de Madrid Spain*
Christophe Champod – *Université de Lausanne, Switzerland*
Niko Brümmer – *AGNITIO, Madrid, Spain*
R.N.J. Veldhuis – *University of Twente, The Netherlands*
D. Meuwly, M. Sjerps, A. Bolk, A. Ruifork, Ch. Berger, D. van Leeuwen, I. Alberink, H. Hanned, A. de Jongh, P. Vergeer, K. Slooten, L. Peschier, R. Haraksim, J, Leegwater, A. Lubach, J. Vermeulen, K. Herlaar – *Netherlands Forensic Institute, The Netherlands*

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) in cooperation with the Netherlands Forensic Institute, the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid.

References

[1] – Alberink, I., de Jongh, A. and Rodriguez, C. (2013), *Fingermark Evidence Evaluation Based on Automated Fingerprint Identification System Matching Scores: The Effect of Different Types of Conditioning on Likelihood Ratios,* Journal of Forensic Sciences. doi: 10.1111/1556-4029.12105
[2] – Champod C, Evett IW, Jackson G., *Establishing the most appropriate databases for addressing source level propositions.,* Sci Justice 2004;44:153–64.
[3] – D. Meuwly, *Forensic Individualization from Biometric Data*, Science & Justice 2006, 46, pp. 205-213
[4] – Hepler AB, Saunders CP, Davis LJ, Buscaglia J., *Score-based likelihood ratios for handwriting evidence.,* Forensic Sci Int 2012;219(1–3):129–40.
[5] – P. Gill, C.H. Brenner, S.J. Buckleton, A. Carracedo, M. Krawczak, W.R. mayr, N. Morling, M. Prinz, P.H. Schneider, B.S. Weir, *DNA Commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures*, Forensic Sci. Int., 160, 90-101, (2006)
[6] – Evett, I., "*Toward a uniform framework for reporting opinions in forensic science casework,* Science & Justice 38(3): 198 – 202, 1998
[7] – AFSP, A. o. F. S. P., *Standards for the formulation of evaluative forensic science expert opinion*, Science and Justice(49): 161-164, 2009
[8] – European Network of Forensic Science Institutes, *The development and implementation of an ENFSI standard for reporting evaluative forensic evidence*, 2a ENFSI annual report 2012
[9] – Cook, R., I. W. Evett, G. Jackson, P. J. Jones and J. A. Lambert, *A method for case assessment and interpretation.* Science & Justice **38**(3): 151-156, 1998
[10] – D. V. Lindley, *A problem in forensic science*, Biometrika, 62(2), 1977
[11] – Neumann C. et.al.: *Quantitative assessment of evidential weight for fingerprint comparison i. generalization to the comparison of a mark with a set of ten prints from a suspect*, Forensic Science International, 207(1:3):101-105, 2011
[12] – C. Champod, I.W. Evett, *A probabilistic approach to fingerprint evidence, Journal of Forensic Identification*, 51:101-122, 2001
[13] – C. Champod, I. Evett, B. Kuchler*, Earmarks as evidence: a critical review,* Journal of Forensic Sciecnes, 46:1275-1284, 2001. 10
[14] – C. Champod, D. Meuwly, *The inference of identity in forensic speaker recognition*, Speech Communication, 31:193-203, 2000
[15] – K. Hoffman, *Statistical evaluation of the evidential value of human hairs possibly coming from multiple sources*, Journal of Forensic Sciences, 36:1053-1058, 1991
[16] – C. Champod, D. Baldwin, F. Taroni, S.J. Buckleton, *Firearms and tool marks identification: the Bayesian approach*, AFTE (Association of Firearms and Toolmarks Examiners) Journal, 35:307-316, 2003
[17] – C.G.G. Aitken, F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004
[18] – C. Champod, F. Taroni, *Interpretation of evidence: the Bayesian approach*, pp. 379-398, Taylor and Francis, London 1999
[19] – D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzales-Rodriguez, J. Ortega-Garcia, *Speaker verification using speaker- and test-dependent fast score normalization,* Pattern Recognition Letters, 28(1):90-98, 2007
[20] – COUNCIL FRAMEWORK DECISION 2009/905/JHA of 30 November 2009 on Accreditation of forensic service providers carrying out laboratory activities, Official Journal of the European Union, 9.12.2009, online http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:322:0014:0016:EN:PDF
[21] – International Organization for Standardization EN ISO/IEC 17025, *General requirements for the competence of testing and calibration laboratories*, ICS: 03.120.20, stage 90/93 (2010-12-15)

[22] – Council of the European Union CEU DS 1459/11, *Council conclusions on the vision for European Forensic Science 2020 including the creation of a European Forensic Science Area and the development of forensic science infrastructure in Europe,* 3135th JUSTICE and HOME AFFAIRS Council meeting Brussels, 13 and 14 December 2011, online http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/jha/126875.pdf
[23] – International Laboratory Accreditation Cooperation, *ILAC-G19:2002 Guidelines for Forensic Science Laboratories*, 2002
[24] – Dutch Accreditation Council (RvA), *Explanation of ISO/IEC 17025:2005*, 30. November 2010
[25] – K. Popper, The Logic of Scientific Discovery (Taylor & Francis e-Library ed.), London and New York: Routledge / Taylor & Francis e-Library (2005)
[26] – Bureau International des Poids et Measures, *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, JCGM 200:2012
[27] – European Network of Forensic Science Institutes, *Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science*, working version 04-11-2013
[28] – K.J. Smalldon, A.C. Moffat, The Calulation of Discrimination Power for a Series of Correlated Attributes, J. Forens. Sci. Soc. (1973), 13,291
[29] – J. Quiñonero-Candela et al., *Dataset Shift in Machine Learning Shift in Machine Learning*. The MIT Press 2009 The MIT Press, 2009.
[30] – G.S. Morrison, *Measuring the validity and reliability of forensic likelihood-ratio systems*, Science & Justice 2011, 51, pp. 91-98
[31] – A. Martin et al., *The DET Curve in Assessment of Detection Task Performance*, National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940; 1997
[32] – D. Ramos, J. Gonzales-Rodriguez, *Reliable support: measuring calibration of likelihood ratios*, Forensic Sci. Int.: in press (2013)
[33] – D. Ramos, J. Gonzales-Rodriguez, G. Zadora, C. Aitken, *Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods*, J. Forensic Sci 2013
[34] – N. Brümmer and J. du Preez, *Application independent evaluation of speaker detection*, Computer Speech and Language, 20(2--3):230-275, 2006
[35] – D. van Leeuwen, N. Brummer, *An introduction to Application-Independent Evaluation of Speaker Recognition Systems*, in Speaker Classification I: Fundamentals, Features, Methods, Christian Müller (Ed.), Springer 2007.

Chapter 2

Influence of the datasets size on the stability of the LR in the lower region of the Within Source Distribution

BTFS 2013 : BBfor2 conference for the interaction of biometric technology research and forensic science

Rudolf Haraksim
Didier Meuwly

1. Abstract

This article focuses on the statistical evaluation of the fingermark evidence using the likelihood ratio (LR) approach. It studies the influence of the quantity of data used to model the within (WS) and between (BS) source variability. The LR system built for the experiment uses an Automated Fingerprint Identification System (AFIS) feature extraction and comparison algorithm, fingermark and fingerprint datasets coupled with a generative approach for modelling the WS and BS variability. This article concentrates on the computation of LRs of the same source in the lower region of the WS distribution. It analyzes the behaviour of the LR with an increasing number of entries in the WS datasets while maintaining the constant proportion of the BS dataset in an attempt to determine the amount of same source scores necessary to achieve consistent LR performance.

2. Introduction

While the question of the comparison of complete fingerprints seems to be an issue long solved in the biometric world with many commercial algorithms and applications available, quite some issues arise when analysing forensic fingermarks (traces). When a fingerprint and a fingermark are subjected to forensic evaluation, the fingermark is almost always partial; its quality severely degraded due to uncontrolled imposition (clarity, distortion) and due to the effects of the development methods.

While the AFIS matching and comparison algorithm is able to achieve great results in terms of performance and speed while producing shortlists of candidates, it is not used in the current practice for the statistical evaluation of fingermarks and fingerprint evidence. Forensic evidence (*E*) in this case is considered the similarity score resulting from the fingermark and fingerprint comparison. In order to quantify the weight of the forensic evidence we start off with a set of mutually exclusive propositions, the one of the prosecution $\boldsymbol{H_p}$ and the one of the defence $\boldsymbol{H_d}$:

- $\boldsymbol{H_p}$ – the fingermark originates from the individual that is also the source of the fingerprint
- $\boldsymbol{H_d}$ – the fingermark originates from an unknown individual, randomly selected

$$LR = \frac{\Pr(E \mid Hp)}{\Pr(E \mid Hd)} \qquad (1)$$

where *Pr* indicates the probability of observing the evidence *E* given one of the two propositions.

The calculation of the LR implies the modelling of the WS and BS scores distributions using a discriminative, generative or hybrid approach [1]). The main objective of this article is to study the influence of the size of the datasets on the stability of the LR. The influence will be studied using a generative approach[1] for the modelling of the within and between source variability.

An ideal situation would be to dispose of a quantity of score observations large enough to cover the whole range of the BS and WS distributions. However in the tails of these distributions a well-calculated LR value is difficult to obtain, due to the rarity of the scores. In the regions where the number of scores is sufficient to describe reliably the WS and BS the LR

[1] In the generative approach we “generate” the score distributions from the the discrete datasets (similarity scores).

value is generally low, and the stability of the LR can be considered as an indicator for the robustness[2] and of the reliability[3] of the method.

In this work we shall analyse the region of the lower tail of the WS score distribution - see figure 1 (similar issues addressed in [6]). We are interested in this region mainly due to the fact that similarity scores in this particular area can "shift" the scales in favour of either of the propositions. Ideally we would like to observe a stabile LR support to either of the propositions, however with the varying number of the WS scores we observe variation in the LRs as well.

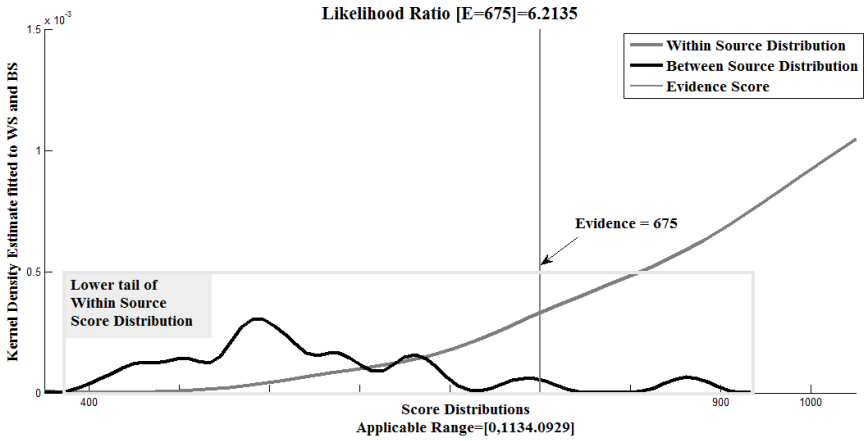


Figure 1. – Area of interest (lower tail of the WS score distribution)

In this initial study we will model the similarity scores produced by the AFIS algorithm using the Kernel Density Function (KDF). This choice is based on the fact that we are dealing with discrete datasets and because the comparison algorithm produces multimodal score distributions. Since we are interested in observing the influence of the different sizes of datasets on the LR stability, the over-fitting, which in most of the cases is considered a drawback of the KDF seems to be a desirable side-effect for this particular application.

Before any method developed can be used in a forensic casework, a validation step needs to provide insight about its robustness and reliability (LR > 1 if H_p true, LR < 1 if H_d is true). The aim of this article is to study the stability of the LR produced and in particular the variations due to data when

[2] Robustness is defined as the ability of a method to maintain the tendency of its performance when reducing the quality conditions of the data under examination.
[3] Reliability defined as the capability of the method of not degrading the trueness of the LR when used in all the possible cases for which it has been designed.

calculating probabilities for both the numerator and denominator of the LR. We will show the influence of lowering the quantity of data used for modelling the WS and BS scores on the stability of the LRs. Despite the fact that relatively small number of individuals is used in this study, it provides a valuable insight on the LR stability depending on the decreasing number of WS scores.

3. Datasets used

For modelling the BS scores, large quantities of reference fingerprints are available, for example ten-print cards originating from a police fingerprint databases. It is not necessarily the case for WS scores, where a limited number of fingermarks and corresponding fingerprints with the ground truth known is available. Different approaches have been proposed in the literature to handle the data sparsity under H_P [3, 4].

Both methods rely on the use of simulated fingermarks from the suspected individual. In [4] these simulated fingermarks are compared with a set of corresponding fingerprints (multiple fingerprints per finger), when in [3] large quantities of simulated fingermarks are compared with a single fingerprint in the fingermarks produced by this method are not completely equivalent to real crime-scene fingermarks but for the purpose of this article and based on the results published in [3], their similarity is considered as sufficient (see figure 2). The number of minutiae and the effect of distortion, present in the set of fingermaks used, represent the key elements of variability for the calculation of the evidential value.

Figure 2. – Simulated fingermark on the left vs. visualized real fingermark from a crime-scene on the right

Simulated fingermarks with 8 minutiae configurations were chosen for this article, as a majority of the fingermarks recovered as pieces of evidence contains less than 12 minutiae, which is the numerical standard in most countries using a numerical standard. In these countries fingermarks with less than 12 minutiae are currently not considered as evidence that can be presented at court and would primarily benefit from the approach described in this paper.

3.1 LR model and size of the datasets used

Figure 3 illustrates the LR model used in this article. The nomenclature used to describe the different datasets refers to the one used in [2].

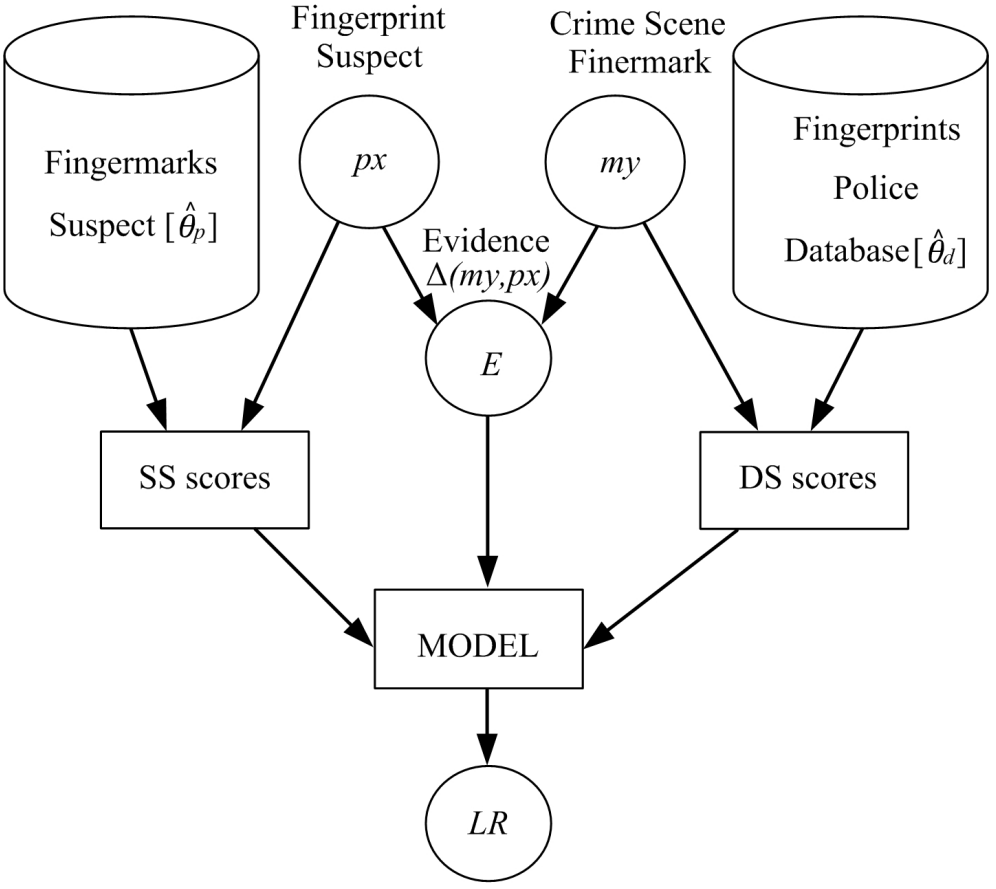


Figure 3. – The LR model

The fingerprint police database consists of electronic copy of ten-print cards. For the purpose of this article we have selected a population of 20.000 individuals (200.000 fingerprints) to represent the BS population.

Since we aim to establish the stability of the LRs in the lower region of the WS score distribution, we will use data from four individuals, for which we have large quantity of simulated fingermarks available – ranging from 2.179 to 8.455. In practice, collecting a WS dataset counting 1000s' of fingermarks for a suspected individual is a time consuming procedure which largely depends on the willingness of the suspect to cooperate (in many cases impossible).

In the following section a forensic evaluation will be described together with the calculation of a likelihood ratio.

4. Evidence Evaluation

As indicated in figure 3, we proceed with evidence evaluation in multiple stages:

- Establish the value of the evidence (E) – a similarity score between a fingermark or fingerprint
- Model the WS distribution based on the comparison of the marks and prints of the same individual (same finger)
- Model the BS distribution based on the comparison of the marks and prints of the different individual (different fingers)
- Calculate the Likelihood Ratio

According to [5] the LR is calculated in the following way:

$$LR = \frac{\Pr(E \mid H_p, \Delta_{SS}(m,p))}{\Pr(E \mid H_d, \Delta_{DS}(m,p))} \qquad (2)$$

where:

$\Delta_{SS}(m,p)$ is the similarity score of the marks and print of the same source

$\Delta_{DS}(m,p)$ is the similarity score of the marks and prints of the different source

In order to obtain calculate the evidence same source in the same dataset, one of the simulated fingermarks (on a leave-one-out basis) will play the

role of the crime scene mark and will be compared to the reference print of the same individual. If the total number of the simulated marks per individual is n, a total of n-1 fingermarks will be available to form the WS score distribution.

As indicated earlier, for WS and BS score distribution modelling we will use the KDF function.

For measuring the stability of the LRs we will vary the number of the WS and BS scores using random sub-sampling. Ideally, with increasing number of the WS scores we should observe more stable LR. More data is in general more informative, especially in the tails of the WS and BS distributions.

In the following section we shall study the influence of the size of the WS and BS datasets on the stability of the LR.

5. Method used

Since we aim to examine the lower tail of the WS score distribution, we will focus on the similarity score interval 375 – 900 (shown in figure 1). The similarity scores are dimensionless, which advocates for the use of the LR framework. Simulated fingermarks of 4 individuals are used in this study.

Table 1 – Proportion of simulated fingermarks

	No. of fingermarks
Individual 1	8455
Individual 2	4666
Individual 3	3179
Individual 4	3758

Individual 1 is used as a benchmark (largest number of simulated fingermarks available) to study the influence of the varying size of the simulated marks and police database datasets. We defined 5 experimental conditions:

1.Equal proportion of WS and BS scores (Symmetric)
2. WS[8455] and BS varying (WSmax)
3. WS[500] and BS varying (BSmin)
4. WS varying and BS[500] (BSmin)
5. WS varying and BS[200’000] (BSmax)

These conditions (where available) will be applied to all 4 individuals.

For all scenarios, the smallest number of WS scores tested counts 500 with 500 scores increments until the WSmax (where available). Similarly the smallest number of BS scores configuration counts 500 with 500 scores increments until BSmax. Since we have a lot more scores available for the BS, we will examine the influence of the amount of BS scores on the stability of the LR with 20.000, 50.000, 100.000 and 200.000 scores.

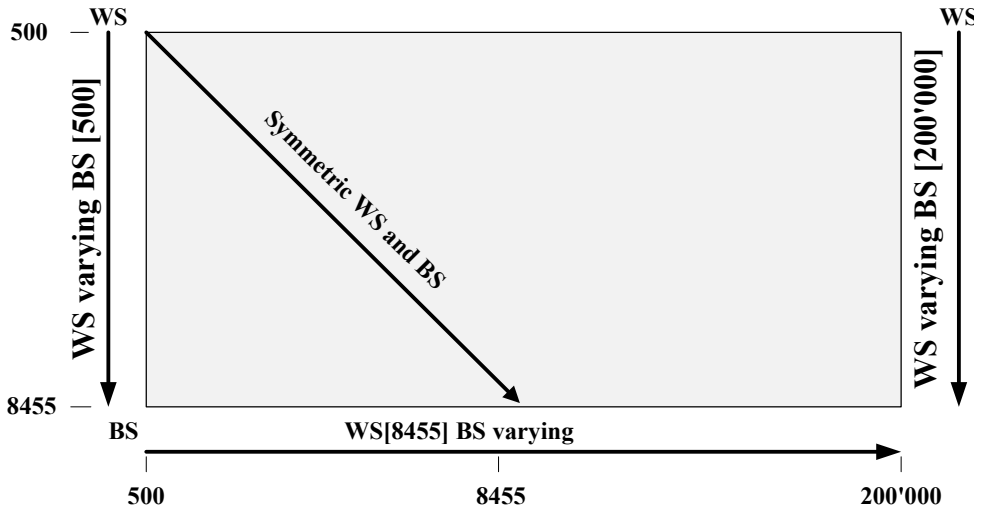


Figure 4 – Four scenarios for LR stability analysis

Please recall that we selected the similarity score interval range from 375 to 900 (see Figure 1). Based on the initial assumption that the LRs in this region are of low order of magnitude, we will place the LRs into 4 bins (10^{-2} < LR < 10^{-1}; 10^{-1} < LR < 10^{0}; 10^{0} < LR < 10^{1}, LR > 10^{1}) in order to analyse the LR behaviour. We are particularly interested in observing the varying proportions of the LRs crossing the value of the neutral evidence (LR_E = 1), changing the support of $\boldsymbol{H_p}$ to $\boldsymbol{H_d}$ and the actual value of the LRs (observation of the E at a fixed value with changing the experimental conditions). The influence of the varying sizes of the WS and BS datasets on the stability of the LR is presented in the following chapter.

6. Results

The experimental setup with the most similarity scores (BSmax, WSmax) was considered as ideal condition and best achievable results, which we aim to approach with increasing number of the similarity scores. In this sense, we want to get as close to the "ideal LR value"[4] with the minimum number of scores. Reader should also keep in mind that our aim here is to understand the data rather than draw conclusions of the rather erratic behaviour of the LRs produced.

Results are divided into two sections: firstly we will look at the stability of the LR for the individual 1 (counting the most WS scores), while in the second part we will attempt to replicate the results for the remaining individuals.

The sum of all the LRs in the 4 LR ranges is equal (126 – given by the total number of *E* scores for which the LRs have been calculated).

6.1 LR stability analysis

In figure 5 one of the populations (BS or WS) is fixed while other one varies from 500 to 8000 (however LRs have been analysed on the whole range of BS 500 - 200000).

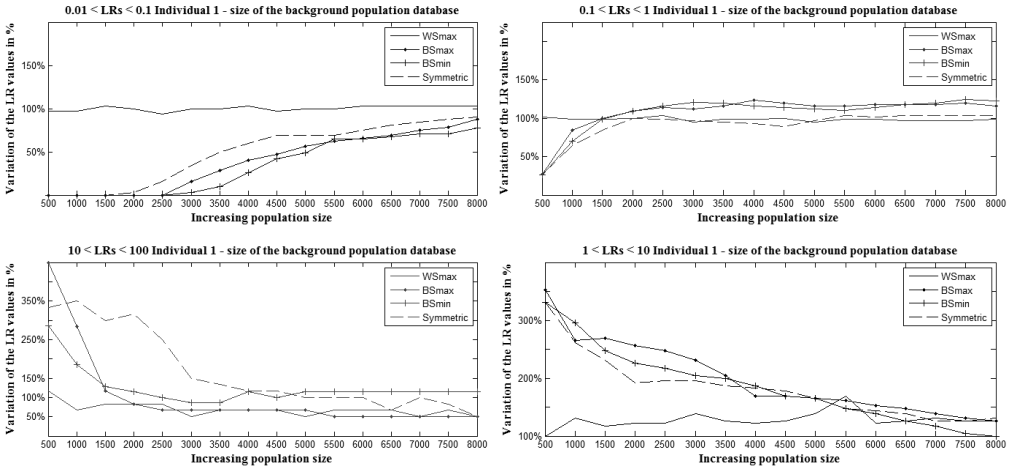


Figure 5 – Experimental setups results for individual 1

The stability of the LRs can be observed and compared with varying size of the BS population (BSmin, BSmax…) The experimental results for the

[4] Ideal LR value is the one obtained with the most ammount of data (BSmax and WSmax).

individual 1 show that about 4000 scores (WS) are needed to obtain a stabile behaviour of ± 10% of the LR values, for the selected LR bin ranges.

Calculated LR values for each piece of evidence E under different experimental conditions are presented in figure 6 on the log-scale. For the experimental condition 1 (symmetric WS and BS) [1000] 85% of LRs support H_p, on contrary in the symmetric set WS and BS [4000] only 46% supports H_p (horizontal line in figure 6 indicates LR = 1 and demonstrates the LR shift in support of different hypothesis).

The size of the BS population does not have a significant influence on the overall stability of the LR. The symmetric experimental condition converges the fastest to the ideal LR value; therefore this condition will be replicated for the remaining individuals.

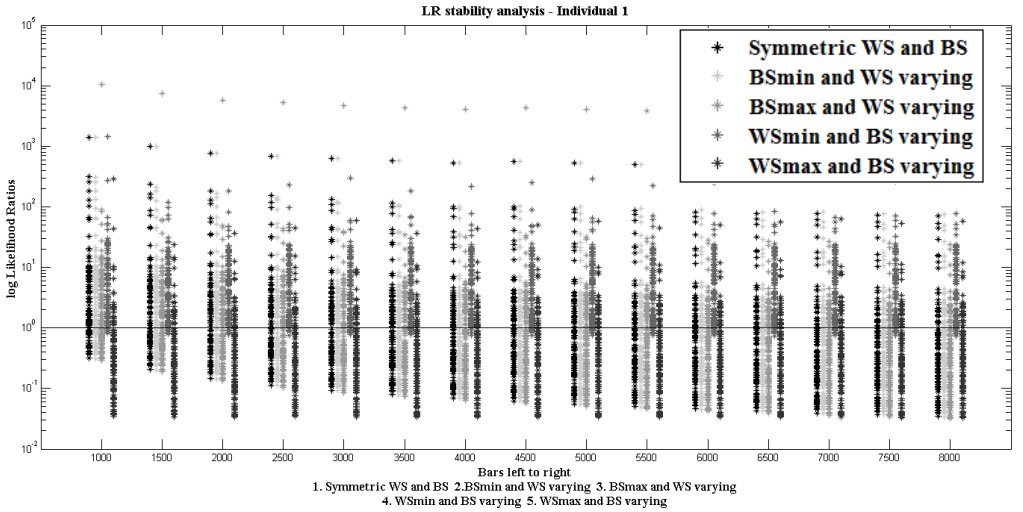


Figure 6 – log(LRs) presented with varying BS population

6.2 Replication for the remaining individuals

The stability of the LRs is analysed using the experimental condition 1 (symmetric WS and BS). Figure 7 illustrates the experimental results for the individuals 2, 3 and 4.

The ideal LR value was calculated from the LRs in the configuration (BSmax and WSmax) of each individual. No LR lower than 10^{-1} was recorded for individuals 2 – 4; hence this bin will remain empty.

The results observed advocates for using the LR calculation method as described in [5]. Despite of the different size of the within source dataset for the 4 individuals, the stabilizing effect of increasing the size of the datasets

on the LRs (as observed in the benchmark) was replicated with amongst all four individuals. Analysing the results separately, within source scores dataset counting 4500 seems sufficient to reach stability of ± 10% of the LR values for individual 2, 3000 for individual 3 and population size of 2000 for individual 4. More general conclusions cannot be drawn from such a limited number of individuals.

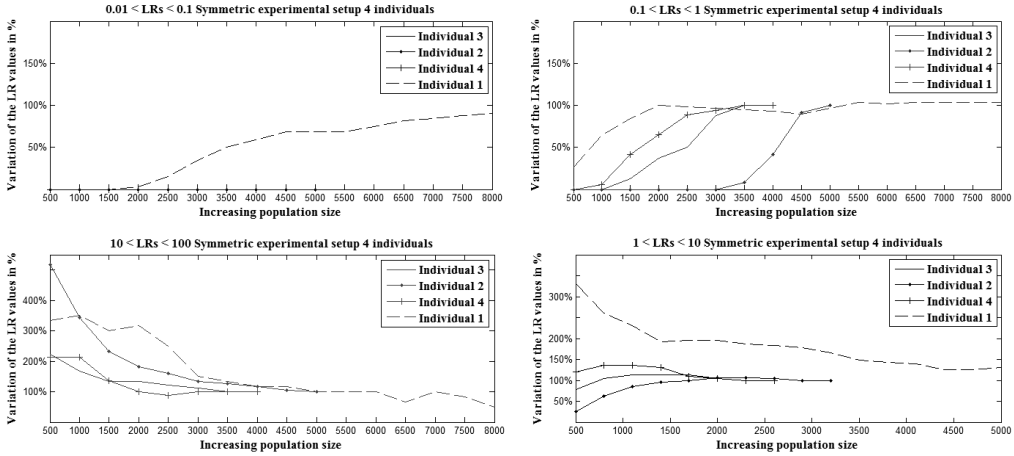


Figure 7 – Differences in stability of the LR amongst 4 individuals using symmetric experimental condition.

7. Discussion and conclusions

The aim of this article was to study the influence of the size of datasets on the stability of the LR. Judging from the experiments conducted, the increase in the between source population size does not seem to have much influence on the LR stability. The symmetric experimental setup has shown to produce the most stable LRs, while a significant variability was observed between the WSmin and WSmax experiments (see figure 6).

The stabilizing trend of the LR due to the increasing size of within source population was replicated for all four individuals, however the results show differences in the minimum number of the within source scores necessary to obtain a stable LR amongst the different individuals and call for further tests with datasets of comparable sizes before a generic threshold can be set.

The use of simulated fingermarks in the experiments show that they are a valuable evaluation tool, as they are relatively easy to produce in significant quantities and one can be “beyond any doubt” certain regarding their origin.

8. Future work

This article is intended as a preliminary study on the stability of the LRs and shows how the LRs behave with varying population sizes. The future work will focus on obtaining similarly large datasets of simulated fingermarks to individual 1 and extend the study for the *E* different source. Following research will be dedicated to non-parametric methods and model-based approaches.

Acknowledgements

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) in cooperation with the Netherlands Forensic Institute.

References

[1] D. Ramos, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Universidad Autonoma de Madrid, November 2007.
[2] C. Neumann, Quantifying the weight of evidence from a forensic comparison: a new paradigm, RSS 175, Part 2, (2011) pp 1 – 26.
[3] C. M. Rodriguez, A. de Jongh, D. Meuwly, Introducing a semi-automated method to simulate large numbers of forensic fingermarks for research on fingerprint identification, JFS 57(2), (2012) 334 – 342 .
[4] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability, Forensic Science International 167 (2007) 189 – 195.
[5] C. Neumann, C. Champod, R. PuchSolis, N. Egli, A. Anthonioz, D. Meuwly, A.Bromage-Griffiths, Computation of Likelihood Ratios in fingerprint identification for configurations of three minutiae. JFS 51(6), (2006) 1255 – 1266.
on, 55(4):480, 2005
[6] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, A review of calibration methods for biometric systems in forensic applications, In: 33rd WIC Symposium on Information Theory in Benelux, Boekelo, Netherlands, (May 2012), pp. 126-133, WIC. ISBN 978-90-365-3383-6

Chapter 3

Fingerprint Evidence Evaluation: Robustness to the Lack of Data

EAFS 2012 : 6th European Academy of Forensic Science Conference

Rudolf Haraksim
Didier Meuwly
Peter Vergeer

1. Abstract

Different approaches have been adopted throughout the scientific community for the fingerprint evidence evaluation using Likelihood Ratios (LR). Such approaches necessitate the use of fingerprint and fingermark data with a known ground truth of their origin. Depending on the approach, the type and quantity of data used to model the within and between source variability varies.

In this work we focus on evaluating the robustness to lack of data of two different approaches: the Non-Anchored and the Finger-Anchored. Robustness is defined as the ability of the method to maintain its performance when reducing the quantity of data. For a comparison of the two approaches we will limit the size of the training datasets used to produce the same source (SS) and different source (DS) distributions to 100, 500, 1000 and 2000 score samples, maintaining the quantity of the fingermarks with the known origin used for testing (8455 fingermarks).

2. The Approaches

In order to maintain a relative equivalence when comparing the two approaches, the Kernel Density Function has been used in both of them to model SS and DS score distributions. The use of other score to LR mapping functions would also be suitable.

2.1 The Finger-Anchored Approach

The finger-anchored is suspect-specific, meaning that in both the numerator and denominator of the LR are conditioned on suspect's fingerprint.

$$LR_{finger} = \frac{P(\Delta(my, px) \mid H_p, I, \Delta([mx], px)}{P(\Delta(my, px) \mid H_d, I, \Delta([mz], px)}$$

where:

my is the fingermark found on the crime scene
px is the fingerprint of the suspect
$\Delta([mx], px)$ is the distance between the marks and prints of the suspect
$\Delta([mz], px)$ is the distance between the marks of other individuals and the print of the suspect

2.2 The Non-Anchored Approach

In the non-anchored approach we aim to model the world population in both the numerator and denominator of the LR.

$$LR_{non} = \frac{P(\Delta(my, px) \mid H_p, I, \Delta([m\hat{z}, p\hat{z}])}{P(\Delta(my, px) \mid H_d, I, \Delta([m\hat{z}], [p\hat{z}])}$$

where:

my is the fingermark found on the crime scene
px is the fingerprint of the suspect
$\Delta([m\hat{z}, p\hat{z}])$ is the distance between marks and prints of the same source
$\Delta([m\hat{z}], [p\hat{z}])$ is the distance between marks and prints of the different source

3. Problem description

The question is, whether it is more suited for the forensic evidence evaluation to have limited quantity of a suspect-specific data modelled using a robust approach, or a large quantity of generic data modelled by an approach where robustness might change with increasing training datasets.

Our initial assumption, based on the fact that the pool of the training dataset is significantly larger in the non-anchored than in the finger-anchored approach, is that the robustness of the LRs produced should improve with increasing quantity of training data.

4. Datasets used

The main difference in the two approaches is in the different datasets used for training the models. For obtaining the SS and DS score distributions we sub-sample with replacement 1000 times the training datasets of both methods. Our target distributions contain 100, 500, 1000 and 2000 samples.

Table 1. Datasets used for training and testing in different approaches.

Approach	Training Data Set		Testing Data Set	
	Same Source	Different Source	Same Source	Different Source
Non-Anchored	16.560	31.200.000	8.455	8.455
Finger-Anchored	8.455	16.560	8.455	8.455

5. Results

The robustness to the lack of data of the two approaches will be demonstrated by calculating the distribution of the Cllr and $Cllr^{cal}$, which are a measure of performance and calibration loss, over the 1000 sub-sample iterations.

Figure 1 illustrates the performance of the two approaches, compared over 1000 sub-sample iterations and shows robustness of the finger anchored-approach when the quantity of the training data decreases compared to the non-anchored.

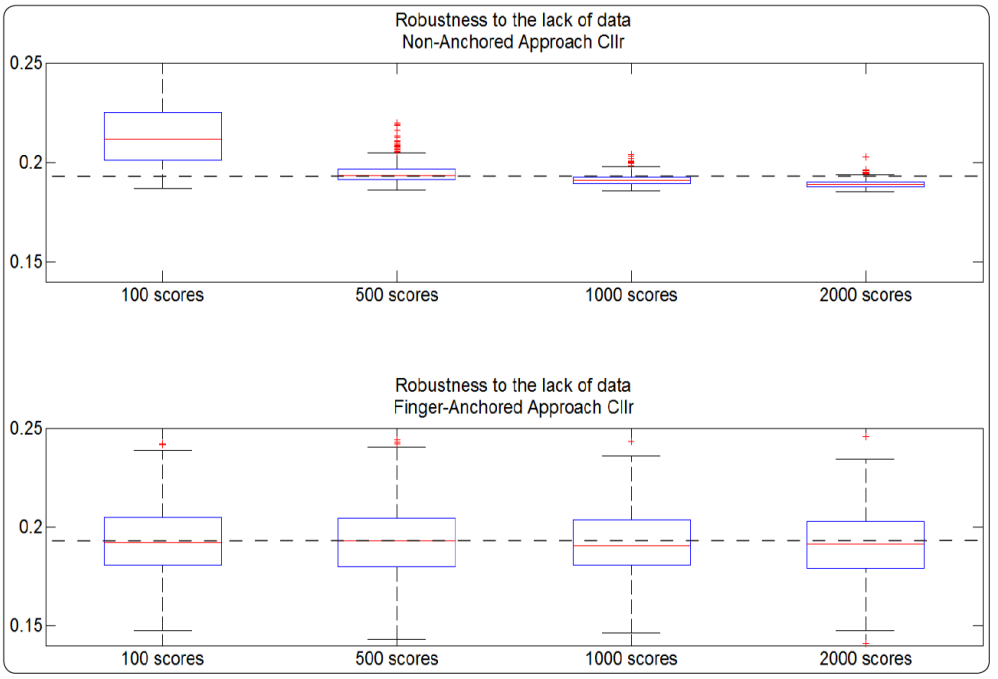


Figure 1. Cllr values across the different quantities of training data samples.

In Figure 2 we observe the variation in the calibration loss over the 1000 sub-sample iterations, when decreasing the quantity of training data. For small-sized training datasets the finger-anchored approach shows smaller calibration loss compared to the non-anchored approach.

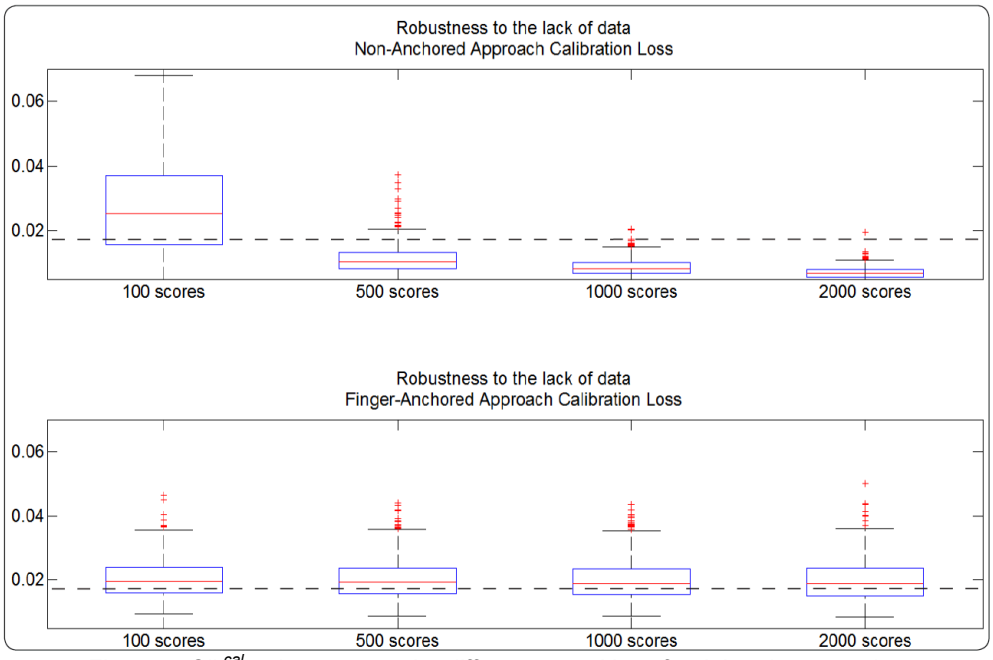


Figure 2. $Cllr^{cal}$ values across the different quantities of training data samples.

6. Discussion and conclusion

In the figures above we observed that the performance Cllr and calibration loss $Cllr^{cal}$ of the finger-anchored approach remain stable when the size of the training dataset decreases. For the non-anchored approach both the performance and calibration loss decrease when the size of the dataset decreases.

From the Cllr values obtained we can also conclude, that the finger-anchored approach outperforms the non-anchored approach for the training dataset containing 100 similarity scores as most of the time in the forensic evidence evaluation we are dealing with limited-sized datasets.

Therefore the finger-anchored approach is probably better suited for the task, together with similar approaches that are conditioned on the suspect.

Acknowledgements

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) in cooperation with the Netherlands Forensic Institute, the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid and the National Police Services Agency of the Netherlands.

Selected References

[He12] A. B. Hepler et al., Score-based likelihood ratios for handwriting evidence, Forensic Science International, doi:10.1016/j.forsciint.2011.12.009, 2012
[Me06] D. Meuwly, Forensic Individualization from Biometric Data, Science & Justice 46, 205-213, 2006
[ECM07] Egli N., Champod C., Margot P.: Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - Modelling within finger variability. Forensic Science International 167(2-3), pp. 189-195, 2007
[Ra07] D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, PhD thesis, Universidad Autonoma de Madrid, 2007

Chapter 4

Validation of Likelihood Ratio Methods for Forensic Fingermark Evaluation: Handling Multimodal Score Distributions

(to be) submitted : Science & Justice (May 2014)

Rudolf Haraksim
Daniel Ramos
Didier Meuwly

1. Abstract

This article deals with the problem of validation of Likelihood Ratios (LR) calculated from multimodal distributions of scores computed by an Automated Fingerprint Identification System (AFIS) feature extraction and comparison algorithm. This algorithm was primarily developed for forensic investigation rather than for forensic evaluation. The fingermark and fingerprint comparison is speed-optimized and performed on three different stages, each of which outputs discriminating scores of different magnitudes, together forming a multimodal score distribution. In this work we will highlight some of the problems related to modelling such distributions using standard methods, such as Kernel Density Estimate (KDE) and propose solutions to issues like data sparsity, dataset shift and over-fitting. Furthermore we will present a method robust to the above-mentioned issues. We should stress here, that the aim of this article is to present a global solution to a range of problems of a given AFIS algorithm, applicable in similar cases when a likelihood ratio needs to be calculated from a multimodal score distribution.

2. Introduction

In this work we propose an approach for handling multimodal score distributions produced by an "AFIS algorithm" (shorthand notation for AFIS feature extraction and comparison algorithm), with the aim of LR computation for forensic fingermark evaluation [1]. The problem of the LR calculation from similarity scores, such as those produced by an AFIS algorithm has been described before [2, 3]. Commercial AFIS algorithms are designed with computational efficiency in mind; therefore a multi-stage scoring process resulting in multimodal distribution may be common among other commercial AFIS algorithms. The fingermark and fingerprint comparison in our case is performed in three different stages, each of which outputs discriminating scores[1] of different magnitudes, together forming a multimodal score distribution. The basic functionality of our AFIS algorithm will be described in more detail in the following section.

The scores produced by any biometric system can be split according to the origin of the samples tested. For the fingerprint modality, the scores can be categorized into Same-Source (SS) scores when the mark(s) and print(s) originate from the same finger and Different-Source (DS) scores when the mark(s) and print(s) originate from two different fingers. The mark is typically a fingermark recovered from an object and the print is a rolled fingerprint captured as reference in a fingerprint ten-print card. Neumann et al. [4] have described a model to assess the evidential value for fingerprint comparison with the example given for the case of 12-minutiae configurations (a numerical standard followed by many countries). Other approaches based on AFIS algorithms have been proposed in [2], where the comparison between the mark and the print is not restricted to twelve minutiae, and is done automatically by an AFIS algorithm. The benefit of the use of an AFIS algorithm relies in its ability to perform large-scale comparisons with huge quantities of fingerprints, because the feature vector of the print can be extracted and compared automatically. When the number of the minutiae is sparse, modern AFIS algorithms are able to extract other type of information from the image (minutiae "handedness" [5], orientation field estimation [6] to name a few).

The performance of standard generative LR calculation methods generally diminishes with the decreasing quantity of scores in the training set used to build the models for LR calculation. As will be shown in subsequent sections,

[1] The reason for not using the term "similarity scores" is that some scores not only take into account similarity, but the tipicality as well. The aim of the score in either case is discrimination. We will not use the score to discriminate, but to compute a LR to assess the evidence value. For these reasons we prefer to use the term "discriminating score" wather than "similarity score".

even a model providing the best fit to the training data may produce to some extent misleading and unreliable LRs. This happens when the tails of SS and DS score distributions are rather poorly described, or when the AFIS scores are either completely missing or very sparse in the training set (in [7] some examples are provided). The main contribution of the approach proposed is the division of the score range in several regions and calculation of LRs in all regions separately following the rules of probability. It should not only be robust to the problem of data sparsity in the tails of the distributions, but also handle the problem of the multimodal distribution itself.

Despite the fact that different models are considered in different regions in which the score range is divided, the main contribution of the article does not focus on a particular model used in each of the regions, because the distribution of such scores may vary significantly from one AFIS algorithm to the other. Conversely, the contribution relies in handling the multimodal score distribution output of the scoring algorithm by combining the value of evidence in all regions in a formal way and producing well-calibrated LRs [8, 9]. This division of the score range and the subsequent combination proposed in this paper can be applied to any comparison algorithm outputting scores, which present multimodal distribution, not only an AFIS.

3. The three regions of the AFIS algorithm

The commercial “off-the-shelf” AFIS algorithms producing discriminative scores are primarily developed to support the process of selection of candidates for forensic investigation and not aimed for the process of description of the evidential value for forensic evaluation [1]. The algorithm selected was speed optimized to perform large number of comparisons in the shortest time possible. It fully uses the concept of “early-outs”, where in our case the database search is split into 3 consecutive stages.

As shown in figure 1, the scores that result from the AFIS algorithm are structured in three regions (R). In Region 1 (R1) the system finds a few minutiae in agreement (the algorithm assigns a score of “-1”). In Region 2 (R2) some similarities are observed, but not enough to warrant a full comparison (the algorithm outputs scores in the region of 0 - 300). Sores produced in R1 and R2 are referred to as “early outs”. Finally in Region 3 (R3) the full comparison of all the features is performed (the algorithm outputs scores bigger than 300).

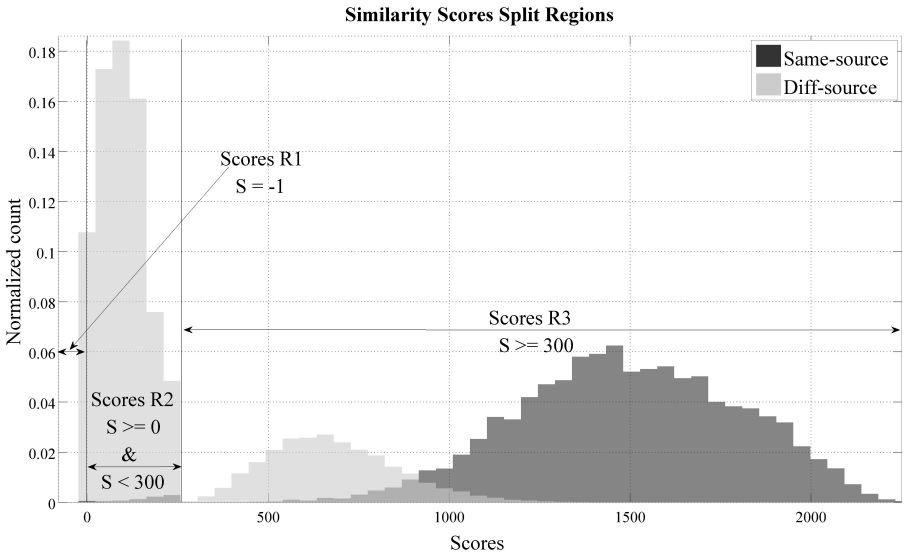


Figure 1 – Different regions of the scores produced by the AFIS algorithm

3.1 Probability of observing a score in a particular region

We can define three events E1, E2, E3 of observing a score in regions R1, R2 and R3 respectively. These regions do not overlap and cover the full range of possible scores. Therefore the events E1, E2 and E3 are mutually exclusive and exhaustive given the three regions.

In a typical AFIS score distribution, the score tends to be bigger as more support is given to the SS proposition, and lower as more support is given to the DS proposition. Thus, ideally in all regions higher scores should be SS and lower scores should be DS. This is illustrated in an example in figure 1, where the separation of SS and DS scores in each region is far from perfect and should be taken into account by the probabilistic model.

Additionally, it should be taken into account that the distribution of SS and DS scores observed in each region may not be uniform, mainly due to the early-out scoring process. As a matter of fact, the observation of the score in one of the three regions alone has “some” evidential value. In most of the cases the majority of the SS scores projects into the R3 region, because a comparison showing high degree of similarity tends to be a SS comparison. Conversely, the majority of the DS scores projects in the R2 region, because a comparison showing low degree of similarity tends to be a DS comparison.

As we will show later, classical models for multimodal distributions such as KDE present many problems related to dataset shift, data sparsity and in some cases complete absence of either SS or DS scores from a particular region. In this work, we will propose an approach robust to those problems for the whole range of scores. The approach is be based on the fact that we model the scores independently in all three regions in order to provide well-calibrated LR values that can be relied on, and combine the outcomes following the rules of probability.

4. LR calculation from AFIS scores

The question of forensic evaluation of fingermark at source level consists in evaluating the likelihood of the two following mutually exclusive propositions:

H_p – The fingermark and the fingerprint in the case both originate from the same finger
H_d – The crime-scene fingermark and the fingerprint in the case originate from different fingers
to form a likelihood ratio, following the formula from the Bayesian probability theory:

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} \qquad \text{(Eq. 1)}$$

where $P(E|H_p)$ is the conditional probability (or likelihood) of observing the evidence (E) under the prosecution proposition (H_p) and the $P(E|H_d)$ is the conditional probability (or likelihood) of observing the evidence (E) under the defence proposition (H_d).

In case of a system that outputs continuous scores S (e.g. AFIS), Equation 1 becomes [4]:

$$LR = \frac{f(S \mid H_p)}{f(S \mid H_d)} \qquad \text{(Eq. 2)}$$

where $f(S|H_p)$ is the probability density function for observing a score (S) under the prosecution proposition (H_p) and $f(S|H_d)$ is the probability density function for observing a score (S) under the defence proposition (H_d). The probability density considers SS scores (fingermarks of a given finger of the suspect and the reference print of the same finger of the suspect) for the numerator of the LR and DS scores (comparison of the questioned fingermark and fingerprints from a subset of the national police database – a collection of 10-print cards) for the denominator of the LR.

4.1 LR model for score-based biometric systems

In the forensic literature different strategies have been proposed for producing LRs from AFIS scores. In the field of score-based biometric recognition [3, 10, 11, 12, 13, 14], the following LR model has been defined:

$$LR = \frac{f(\Delta(my,px) \mid H_p, \Delta\hat{\theta}_p)}{f(\Delta(my,px) \mid H_d, \Delta\hat{\theta}_d)} \qquad \text{(Eq. 3)}$$

where for the fingerprint evidence evaluation datasets are defined in the following way:

$\Delta(my,px)$ – a score between the fingermark *my* found on the crime scene and the fingerprint *px* of the suspect
$\Delta\ \hat{\theta}_p$ – scores obtained from comparing a training set of simulated fingermarks of the suspect with the reference fingerprint of the suspect
$\Delta\hat{\theta}_d$ – scores obtained from comparing the crime scene fingermark and a subset of fingerprints from the population database used in the model (in this case a subset of operational 10-print card police database)

Furthermore, we will use below the following notation to refer to the parameters of the models:

θ – represents the parameters of the model (e.g. mean, variance) that need to be trained
$\hat{\theta}$ – represents a value given to the parameters of the model, obtained from the scores of training set

4.2 Datasets used

Since it is notoriously difficult to find forensically relevant, sufficiently large datasets with the ground truth about the origin of the samples known, we decided to use a set of simulated[1] [15] 8-minutiae fingermarks from 6 individuals paired with their corresponding fingerprints. The fingermarks were obtained by capturing an image sequence of the finger of each individual from an optical live scanner (Smiths Heimann Biometrics ACCO 1394S live scanner) and splitting the frames captured into 8 minutiae configurations.

For modelling the SS scores (numerator in Equation 3) we used the AFIS scores of simulated fingermarks and the corresponding reference fingerprint

[2] Simulated fingermarks in this case refer to series of image captions of a finger moving on a glass plate of the fingerprint scanner (the procedure is described in detail in [15]).

as training data, captured from the same individual under controlled conditions. For modelling the DS scores (denominator in Equation 3) we used the mark in the case compared against a 200,000 - fingerprint subset of the population database. The values assigned to the parameters of the distributions $\hat{\theta}$ are obtained from the data summarized in the Table 1.

Table 1: Same and different source scores.

Individual	$\Delta\ \hat{\theta}_p$ - same source scores	$\Delta\ \hat{\theta}_d$ - different source scores
Person 1	8,455 marks 1 print	8,455 marks 200,000 prints
Person 2	2,751 marks 1 print	2,751 marks 200,000 prints
Person 3	4,666 marks 1 print	4,666 marks 200,000 prints
Person 4	2,206 marks 1 print	2,206 marks 200,000 prints
Person 5	3,179 marks 1 print	3,179 marks 200,000 prints
Person 6	3,758 marks 1 print	3,758 marks 200,000 prints

For example scores for the Evidence Same Source (E_{SS}) are obtained on a "leave-one-out" basis from the SS score distribution (fingermarks of Person 1 and fingerprint of the Person 1) and scores for the Evidence Different Source (E_{DS}) are obtained from the AFIS scores of the fingermarks of Person 1 with the fingerprints of Persons 2-6. This process is repeated iteratively for each person. In the "leave-one-out" approach we iteratively sweep through the set of fingermarks. With every iteration we delegate one of the fingermarks to play the role of the crime-scene mark *my* and maintain the remaining fingermarks to form SS and DS score distributions. The concept of the LR and the method used will be discussed in length in the following sections.

5. Baseline Model

The multimodal character of the SS and DS score distributions and the non-overlap of the three regions suggests the use of flexible, and non-parametric score to LR transformation models (if all three regions are modelled together). A popular choice in the literature [7] has been the Kernel Density Estimation (KDE), which will be used as the baseline reference model, even

though over fitting of the score distribution is a known limitation of the KDE[3]. In the KDE baseline experiment we treat all the SS (and DS) scores in all three regions together to calculate LR's from the AFIS scores (model illustrated in figure 2). KDE (or any other parametric / non-parametric modelling method) will however not be of much use particularly in the R1 region, since all the scores in this region have the same discrete value *S = -1*. This is an excellent example of a limitation of the use of KDE for this kind of score distribution.

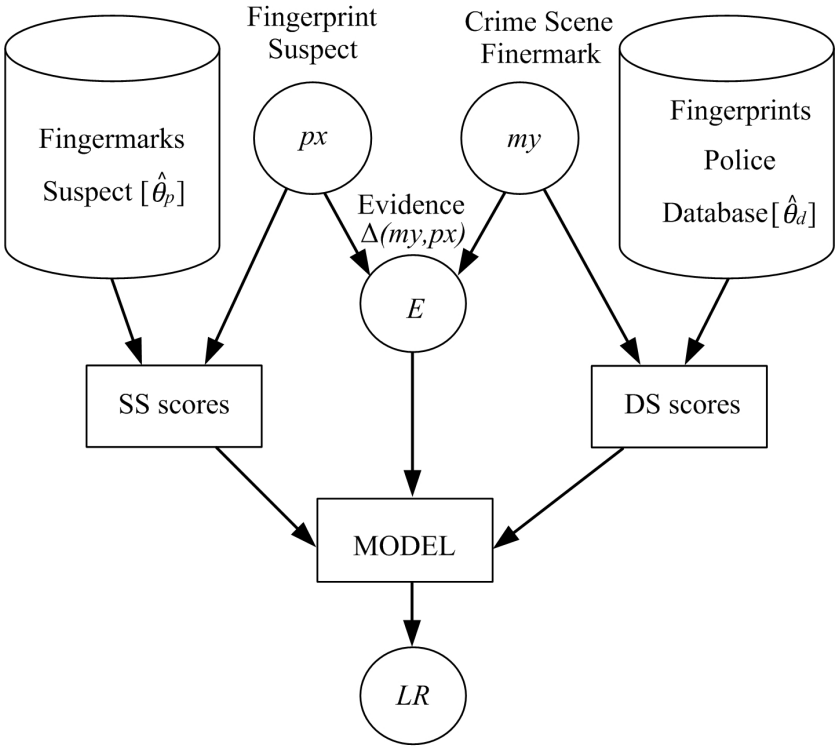


Figure 2 – The Baseline Model[4]

5.1 The Dataset Shift in the Different Source (DS) scores[5]

Traditional generative approaches like KDE used to treat similar score distributions show degraded performance – due to the lack of data,

[3]Although the KDE is a non-parametric model, we will use the notation in equation 3, where the $\hat{\theta}$ in this case are all the training score values and the mean and variance of the kernel [16, 17].

[4] SS and DS scores in this case can be modelled either using generative or discriminative approaches [10].

[5] Even though the Dataset Shift affects the baseline model, is not a characteristic of the baseline model.

increased distortion in the fingermark, but also due to the dataset shift. The dataset shift is in the literature defined as a difference between the training and testing[6] score distributions [18], in our case the dataset shift is most obvious when comparing fingermarks of a particular individual to the Different Source fingerprints captured in our laboratory and to the population database supplied prints obtained from 10-print cards (see figure 3). The dataset shift in our case occurs due to different fingerprint capturing methods. Whereas the population database fingerprints are high-resolution scans from the 10-print cards (e.g. inked fingerprints are first rolled on a paper, scanned and post-processed), the laboratory-captured fingerprints used in this article are produced using a high-resolution fingerprint live scanner (e.g. direct caption of the fingerprint).

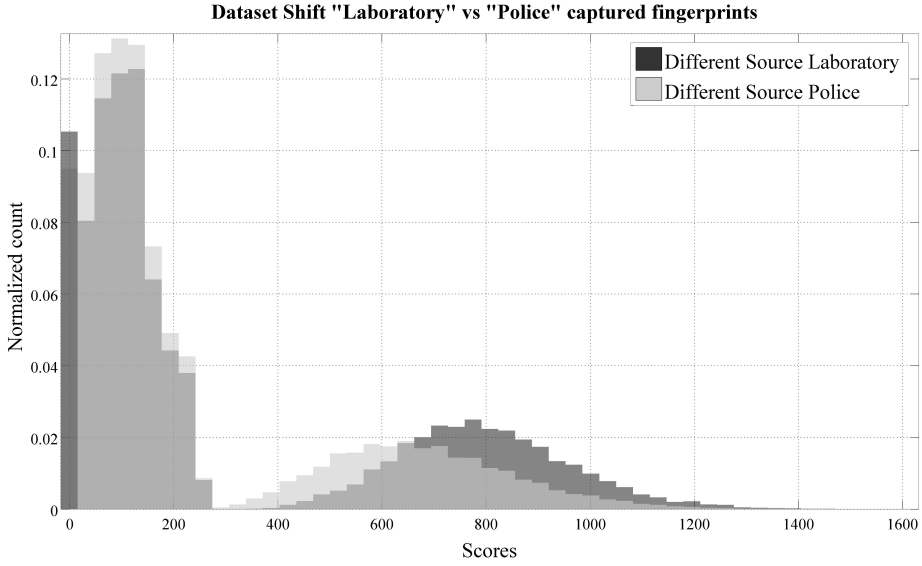


Figure 3 – Example of a dataset shift in the DS AFIS scores (Laboratory vs. Police captured fingerprints)

5.2 Data sparsity leading to extreme LR values

In some cases LR calculation from scores using the KDE for the E_{SS} results in the LRs of huge magnitudes (even infinity)[7] and for the E_{DS} in the LRs strongly supporting the wrong proposition (LR = 2.28^{91}) on the "log" scale as shown in Figures 4 and 5. The resulting performance of such model is seriously degraded [8], despite its visually excellent fit.

[6] Splitting the data available into the training and testing sets are one of the measures of the statistical validity commonly used across all biometric modalities for determining problems with the data itself or a model representation.

[7] Likelihood Ratio values of such magnitudes don't have a meaning in forensic evaluation.

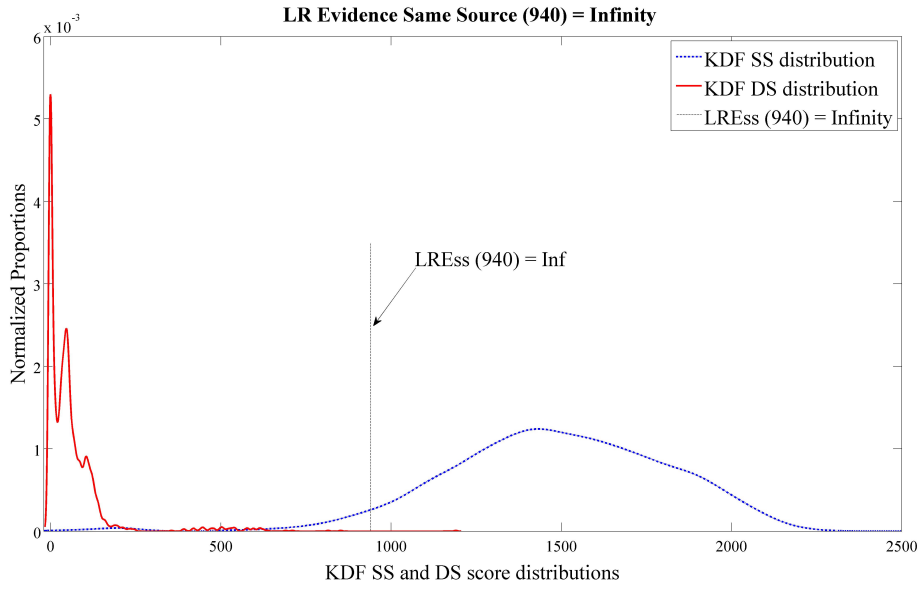


Figure 4 – KDE producing LR = ∞ for the E_{SS} = 940

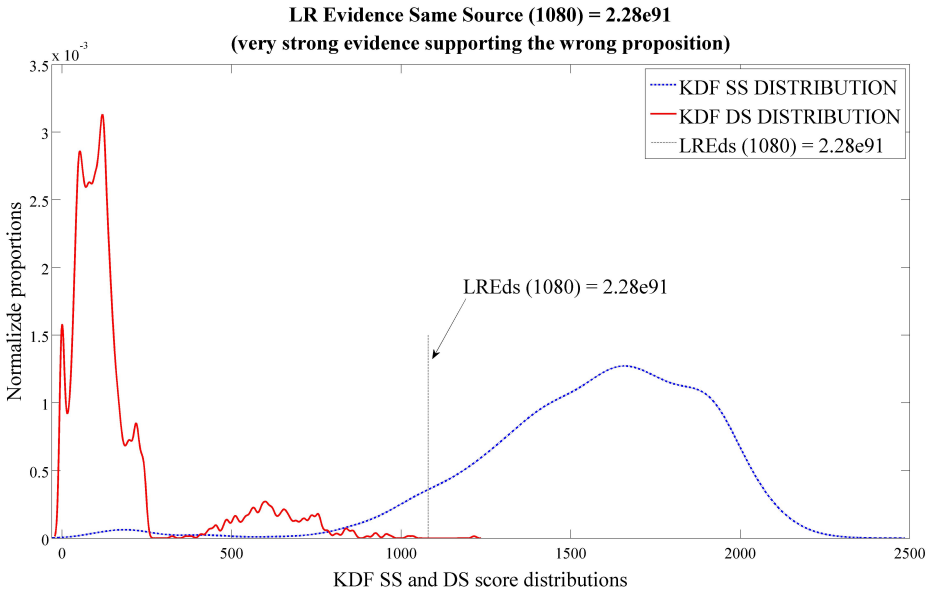


Figure 5 – KDE producing to LR = 2.28^{91} for the E_{DS} = 1080 (very strong support of the wrong proposition)

Figures 4 and 5 indicate that based on the training data, the baseline model in our case does not provide reliable LR values due to the poor description of the tails of the SS and DS score distributions. An LR value of ∞ in this case can be wrongly equated to a categorical decision towards the prosecution proposition, something that has to be definitely avoided [19]. On

the other hand, LR_{EDS} of 2.28[91] provides a very strong support to the wrong proposition. This effect is amplified by dataset shift that is significant as explained above.

From the above-mentioned we can specify two major requirements. The new approach should be robust to the data sparsity (lack of data in general), the dataset shift, but most of all – the LRs produced should be well calibrated and contained within "reasonable" boundaries.

6. Multimodal Approach

In the method proposed we will split the SS and DS score distributions into the three regions of interest, since the events of observing an AFIS score in different regions are mutually exclusive and exhaustive.

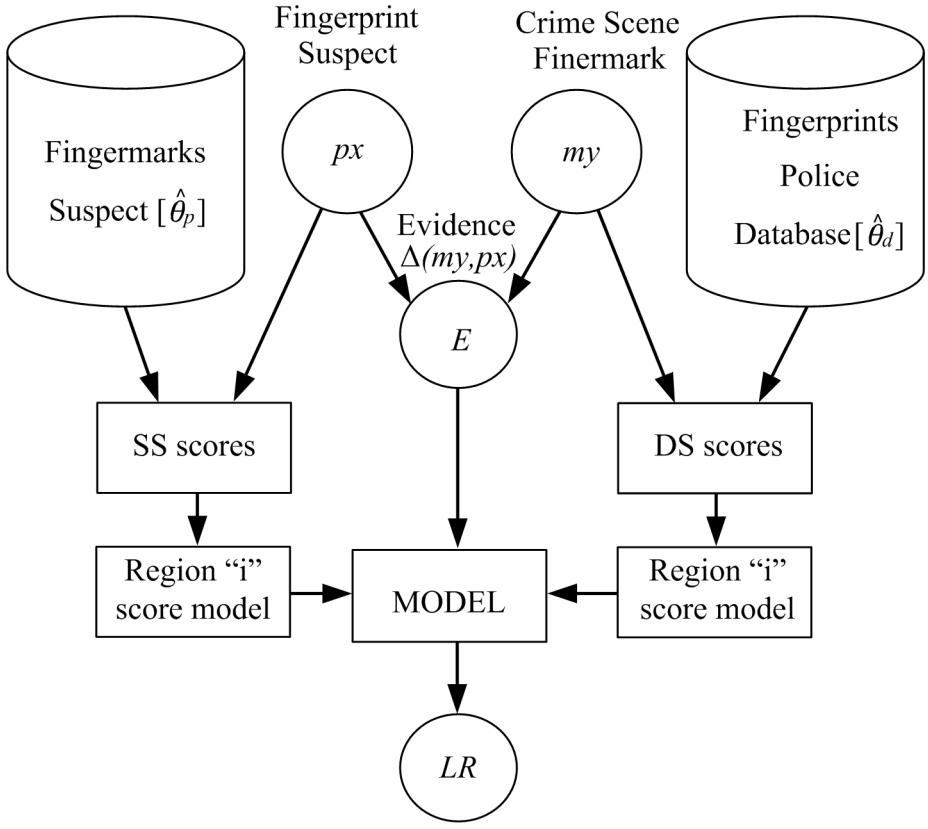


Figure 6 – Multimodal approach

In each region both the numerator and the denominator of the LR can be extended by the proportion of scores observed in a particular region under

both propositions $P(R_i \mid H_y)$ - where *i* = *1:3* and *y* = *p* or *d*. Transforming LR back to ratio of probabilities we obtain following:

$$LR = \frac{P(S \mid R_1, H_p) \times P(R_1 \mid H_p) + P(S \mid R_2, H_p) \times P(R_2 \mid H_p) + P(S \mid R_3, H_p) \times P(R_3 \mid H_p)}{P(S \mid R_1, H_d) \times P(R_1 \mid H_d) + P(S \mid R_2, H_d) \times P(R_2 \mid H_d) + P(S \mid R_3, H_d) \times P(R_3 \mid H_d)}$$

(eq. 4)

A score cannot be observed in multiple regions simultaneously, thus, assuming that the score is observed in region R_i, the equation 4 simplifies to:

$$LR = \frac{P(S \mid R_i, H_p) \times P(R_i \mid H_p)}{P(S \mid R_i, H_d) \times P(R_i \mid H_d)} \qquad \text{(eq. 5)}$$

Where the $\frac{P(R_i \mid H_p)}{P(R_i \mid H_d)}$ is the ratio of probabilities of observing R_i scores given that the fingermark and the fingerprint originates from the same finger over the probability of observing R_i scores given that the fingermark and the fingerprint originates from different fingers.

6.1 Scores in the Region 3

$$LR = \frac{P(S \mid R_3, H_p) \times P(R_3 \mid H_p)}{P(S \mid R_3, H_d) \times P(R_3 \mid H_d)} \qquad \text{(eq. 6)}$$

Score distributions in the R3 region for both SS and DS are smooth as shown in figure 7. From the histograms of the SS and DS score distributions on figure 7 we consider as a reasonable initial assumption that the scores in the R3 region are distributed following a Gaussian (Normal) distribution.

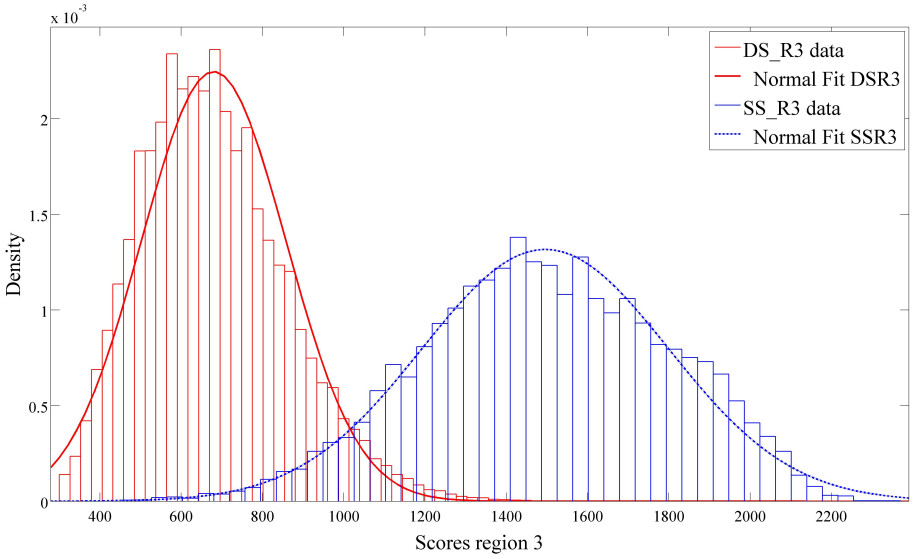


Figure 7 – Gaussian fit to the R3 region score distributions

As an alternative to the Gaussian distribution we propose to use a Linear Logistic Regression for modelling the scores in this region following the work in [20]. We have continuous sets of SS and DS scores (which fit into the R3 region) and the bigger the score the better support for the H_p proposition. If we expect our solution to be a monotonically rising function, we can approach to calculate LRs in a non-parametric way by using for example the Pool Adjacent Violators (PAV) algorithm, following the work in [21].

6.2 Scores in the Region 2

$$LR = \frac{P(S \mid R_2, H_p) \times P(R_2 \mid H_p)}{P(S \mid R_2, H_d) \times P(R_2 \mid H_d)} \qquad \text{(eq. 7)}$$

The DS score distribution in the R2 region appears to be skewed, and the SS score distribution seems to be monotonically rising in this region, therefore a Gaussian fit is not suitable for modelling the scores in this region (as seen in figure 8). A far better fit can be achieved using for example the Beta function (figure 9).

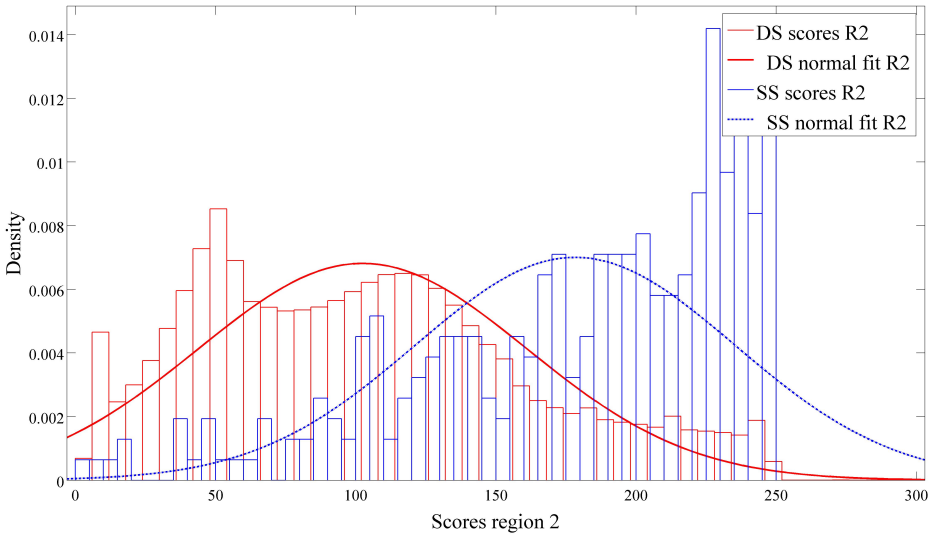


Figure 8 – Gaussian fit to the R2 region score distributions

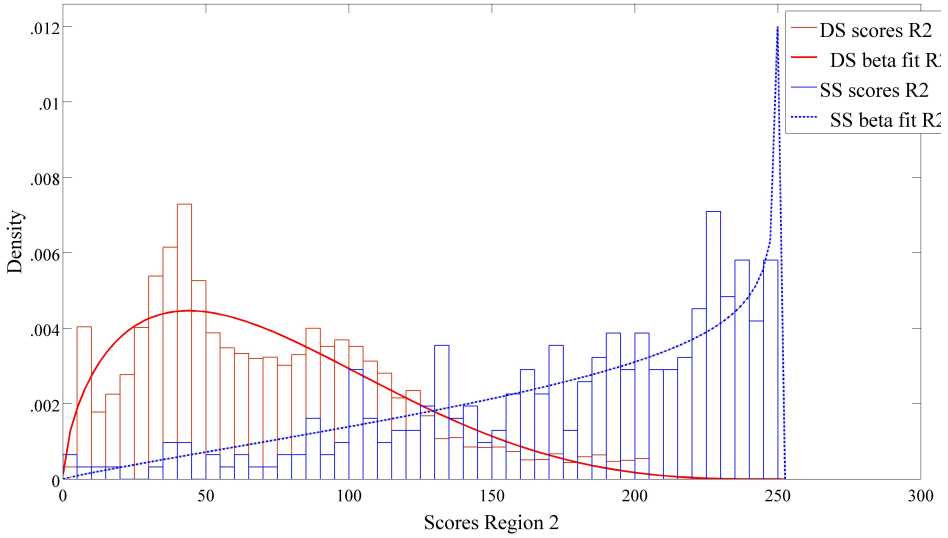


Figure 9 – Beta fit to the R2 region score distributions

Alongside the Beta function we will use the linear logistic regression and the non-parametric PAV algorithm in the R2 region as alternatives when calculating LRs from scores.

6.3 Scores in the Region 1

As mentioned in the beginning, all of the scores observed in the R1 region get one particular score (S_{R1} = -1) assigned by the AFIS algorithm. Equation 5 for the R1 region will have this form:

$$LR = \frac{P(S \mid R_1, H_p) \times P(R_1 \mid H_p)}{P(S \mid R_1, H_d) \times P(R_1 \mid H_d)} \quad \text{(eq. 8)}$$

where $P(S \mid R_1, H_p)$ is the probability of observing a -1 score (S_{R1} = -1) amongst all the scores observed in the R1 region under the H_p, an event, which is always true and we can write $P(S \mid R_1, H_p) = 1$. The same logic applies to the $P(S \mid R_1, H_d)$, which is a probability of observing a -1 score (S_{R1} = -1) amongst all the scores falling into the R1 region under H_d, an event which again is always true and we can write $P(S \mid R_1, H_d) = 1$.

If we apply above-mentioned conditions, eq. 4 further simplifies to a ratio of probabilities of observing score in the R_1 region under both propositions $\frac{P(R_1 \mid H_p)}{P(R_1 \mid H_d)}$. The scores in region R_1 possess certain evidential value, despite the fact that all of them share the same discrete value.

Let's assume that very few SS are observed in the R_1 region, and that they are mostly DS scores. If we observe a score of -1 (in the R_1 region), the LR should support the defence hypothesis. This happens if $LR = \frac{P(R_1 \mid H_p)}{P(R_1 \mid H_d)}$. Additionally, ignoring the scores in the R_1 region because all of them have the same value appears to be a waste of the discriminating information given by the fact that in R_1 there are mostly DS scores.

7. Robustness to the lack of data

The problem of assigning probabilities when no observations have been made in the training data has been studied for example in [22]. We still have to complete our model with a method to assign the following probability ratio for each region R_i:

$$\frac{P(R_i \mid H_p)}{P(R_i \mid H_d)}$$

which are probabilities of observing a score in i-th Region under both – prosecution and defence propositions.

6.4 Example with a simplified binary division

Assigning $P(R_i|H_p)$ and $P(R_i|H_d)$ to each of the different regions R_i, $i=1,2,3$; needs to consider some robustness about the sparsity of the scores in the training set. In order to illustrate this, we start with a simplified example, where we have divided the score axis in two regions R_1 and R_2 – a *binary* division. We consider for illustration the scores under the assumption that H_d is true, but this example can be analogously applied to the scores under the assumption that H_d is true. In order to assign a probability $P(R_i|H_d)$ that a given score will be observed in region R_i, we need some previous observations of regions where the scores have fallen – some training observations. Those observations are taken from the training scores $\Delta\hat{\theta}_d$, with N_p scores, in the following way. Let $R_d = \left\{R_d^1,...,R_d^{N_d}\right\}$ be a sample of random variables, where R_d^j represents the region in which the *j-th* different-source training score was observed. In this *binary* example, the possible outcomes of each R_d^j are R_1 and R_2. Then, the outcome of R_d^j will be the region in which the *j-th* score in $\Delta\hat{\theta}_d$ is observed. Thus, the training observations are the particular values of each of those random variables R_p^j. We assume that variables $R_d = \left\{R_d^1,...,R_d^{N_d}\right\}$ are identically distributed according to a Bernoulli distribution, where the probability that a score is observed in Region *i* is precisely $P(R_i|H_d)$, the parameter of the model. Moreover, we assume that the variables are *conditionally independent given the model*. Then, it can be shown that the *maximum likelihood* rule for the probability that a score will fall into Region *i* is as follows:

$$P\left(R_i\middle|H_d\right) = \frac{M_i}{N_d} \qquad \text{(eq. 9)}$$

where M_i is the number of scores in the training set observed in Region *i*. If the training scores under H_d for whatever reason contains zero score observations in Region *i*, i.e. $M_i = 0$, we get the following:

$$\frac{P\left(R_i\middle|H_p\right)}{P\left(R_i\middle|H_d\right)} = \frac{P\left(R_i\middle|H_p\right)}{\frac{M_i}{N_d}} = \frac{P\left(R_i\middle|H_p\right)}{\frac{0}{N_d}} = \infty \qquad \text{(eq. 10)}$$

In some cases this might result in a $LR = \infty$. An analogous derivation results in $LR = 0$ for same-source scores falling in a region where no same-source scores have been observed before.

An outcome of $LR = 0$ or ∞ is very likely to occur if a similarity score, either SS or DS, is not observed in one of the regions. The problem arises

particularly in the R1 region, where the SS scores are quite rare, but can also occur in the R2 or R3 region as well.

7.1 Bayesian solution

In order to avoid "zeroes" in either numerator or denominator of the LR and to assure a valid numerical input, we propose a Bayesian solution to $P(R_i|H_d)$. We start from the above binary example, where a maximum likelihood rule was considered. Under the same assumptions, if we instead consider that the probability $P(R_i|H_d)$, the parameter of the Bernoulli distribution, has a uniform prior distribution (in the [0,1] range), then it can be shown that the solution inferred is the *predictive distribution*, which takes the following form:

$$P\left(R_i \middle| H_d\right) = \frac{M_i + 1}{N_d + 2} \qquad \text{(eq. 11)}$$

A full derivation is tractable, and can be found in [23] (Equations (6.66) to (6.73)). This result is known as the *Laplace rule of succession* [24]. For simplicity the application of this rule on our dataset will be demonstrated on R1 region, where all the scores attain a discrete value S = -1. Recall the binary example, where in the R1 region we obtained LR = ∞ because there were no observed scores in that region in the training data. Suppose a number of DS training scores $N_d = 20$ and that none of these scores are observed in the Region 1, thus $M_i = 0$. Then, according to the previously proposed maximum-likelihood rule we would obtain $P(R_1 \mid H_d) = \frac{M}{N} = \frac{0}{20} = 0$ and the LR would be infinite. However, with the Bayesian uniform prior on the model's parameter (Laplace rule of succession) we get following $P(R_1 \mid H_d) = \frac{M+1}{N+2} = \frac{1}{22} \approx 0.05$, which with increasing number of scores will be approaching zero, but will still provide a non-zero numerical value. The interpretation of this result is, that additionally to the training data, a uniform prior for the model parameters forces to consider always at least the observation of one score in each of the regions. Therefore, if H_d is true, we have to consider $N_d + 2$ scores, and the scores observed in each region will be at least one. An analogous derivation provides equivalent interpretation for the case when H_p is true.

7.2 Generalization to more than 2 regions

The problem addressed in this paper requires a generalization with respect to the rule of succession for the binary example, because we are dividing the score range into more than 2 regions. That means that the variables

$\left\{R_d^1,...,R_d^{N_d}\right\}$ will now have more than 2 possible outcomes, and therefore their distribution cannot be a Bernoulli distribution. The generalization to more than 2 possible outcomes, say *Q* possible regions, involves the assumption that the variables $\left\{R_d^1,...,R_d^{N_d}\right\}$ follow a multinomial distribution. Moreover, since there are now *Q* parameters for this multinomial model, the prior uniform distribution of the model parameters will be a Dirichlet distribution, particularized for the case of uniform variables. Under these conditions, the derivation of the predictive distributions $P\left(R_i|H_d\right)$ for each of the regions can be found in [25], and therefore generalizes the rule for more than 2 regions. That generalization provides the following result for the predictive distribution:

$$P\left(R_i|H_d\right)=\frac{M_i+1}{N_d+Q} \qquad \text{(eq. 12)}$$

or, in the case of 3 regions as in the problem we address in this article, we have:

$$P\left(R_i|H_d\right)=\frac{M_i+1}{N_d+3} \qquad \text{(eq. 13)}$$

Again, the analogous derivation produces a similar result for the case where H_p is true.

In our model, equation 13 will be used in all three regions to assign all the probabilities $P\left(R_i|H_p\right)$ and $P\left(R_i|H_d\right)$. This is because in cases where there are both SS and DS scores present values, the probabilities do not change significantly with respect to the maximum likelihood solution. In cases where there are zero scores of either SS or DS it will give robustness to the model, avoiding results of $LR = 0$ or $LR = \infty$.

The motivations for the use of the Laplace rule of succession and its generalization are thoroughly justified in [23] and [24].

8. Experiment

We will measure the discrimination and calibration [8, 9] of the two approaches – baseline KDE and Multimodal in terms of Detection Error Trade-off (DET) curves [26] and Equal Error Rate (EER), Cost-Log-Likelihood-Ratio (C_{llr} and C_{llr}^{min}) and the Empirical-Cross-Entropy (ECE) [8].

The DET curve shows a trade-off between two types of errors – false acceptance and false rejection. The Equal Error Rate (EER) is the rate at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal if a threshold is used for the biometric system. More details regarding the EER can be seen in [26]; Cllr – the measure of calibration and $Cllr^{min}$ its discrimination component are in length described in [21]; and the ECE, which is closely related to Cllr, can be summarized as a measure of accuracy, as a sum of discriminating power and calibration with an information-theoretical interpretation [8, 9]. Cllr and the ECE in this work will use the LR values produced by proposed models and provide a quantitative measure of calibration.

In the baseline model the Kernel Density Estimate will be fitted to the SS and DS score distributions from which the LR's will be calculated. Since in the multimodal approach we treat all regions (R1, R2 and R3) separately, we will attempt to find the best performing model for each region. We will consider Normal and Beta functions in attempt to represent the score distributions parametrically and the linear logistic regression (LinLogReg) and the Pool Adjacent Violators (PAV) algorithm for non-parametric LR calculation in the way shown in Table 2:

Table 2. Different methods for LR calculation for the multimodal and baseline approach

Multi Modal Approach		
Region 1	**Region 2**	**Region 3**
$\propto (SS/DS)_{Bayesian}$	Beta	Normal
$\propto (SS/DS)_{Bayesian}$	Beta	PAV
$\propto (SS/DS)_{Bayesian}$	Beta	LinLogReg
$\propto (SS/DS)_{Bayesian}$	PAV	Normal
$\propto (SS/DS)_{Bayesian}$	PAV	PAV
$\propto (SS/DS)_{Bayesian}$	PAV	LinLogReg
$\propto (SS/DS)_{Bayesian}$	LinLogReg	Normal
$\propto (SS/DS)_{Bayesian}$	LinLogReg	PAV
$\propto (SS/DS)_{Bayesian}$	LinLogReg	LinLogReg
Baseline approach		
KDE baseline for joint regions		

Each line in Table 2 describes a different combination of models for each of the three regions, where the term $\propto (SS/DS)_{Bayesian}$ is short form notation for the $\frac{P(R_1 \mid H_p)}{P(R_1 \mid H_d)}$ – a Bayesian solution presented earlier, representing the ratio of probabilities of observing R_1 scores given that the fingermark and the fingerprint originates from the suspect (same source) over the probability of observing R_1 scores given that the fingermark originates form another person (different source). In the R1 region probabilities are assigned to the events of observing a score following the Laplace rule of succession (described earlier).

9. Results

The results will be presented in two sections. In the first we compare Tippett plots[8] and the discriminating power of the different approaches (by means of DET plots and EER values), while in the second we will focus on the accuracy of the LR decomposed as discriminating power and calibration (Cllr and ECE plots).

9.1 LRs produced

In the baseline KDE approach we see a somewhat sub-optimal performance of the KDE (Figure 10), as the inverse cumulative density function of the LR_{ESS} fails to converge in the bottom right region. This results in extremely large values of the LR for same-source comparisons. In this region the KDE model over-fits the training data, especially in the tail of the SS and DS distributions. In the examples earlier we saw LRs reaching undesirably high values. Also note, that the $\log_{10}$LR values in the baseline KDE have been limited at 30 for displaying purposes. We established earlier that the values produced in this approach go far beyond the LR = 10^{91}, which is an extreme result that does not have a forensic meaning. It does not reflect a reliable assignment for the evidential value and is the result of an artefact of the modelling approach.

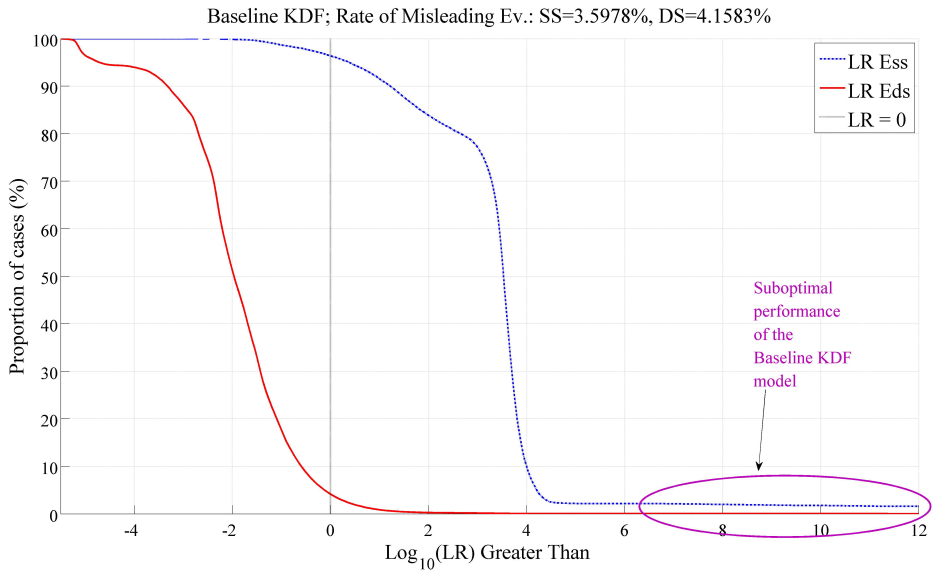


Figure 10 – Baseline KDE, Tippett Plot

[8] Tippett plots [7] are representations of the inverse cumulative density functions of LRs. Each of the curves represents LRs supporting one of the competing propositions.

Even though the same quantity of scores was used in both approaches, the resulting LR distributions do appear much more refined using the multimodal approach (Figure 11).

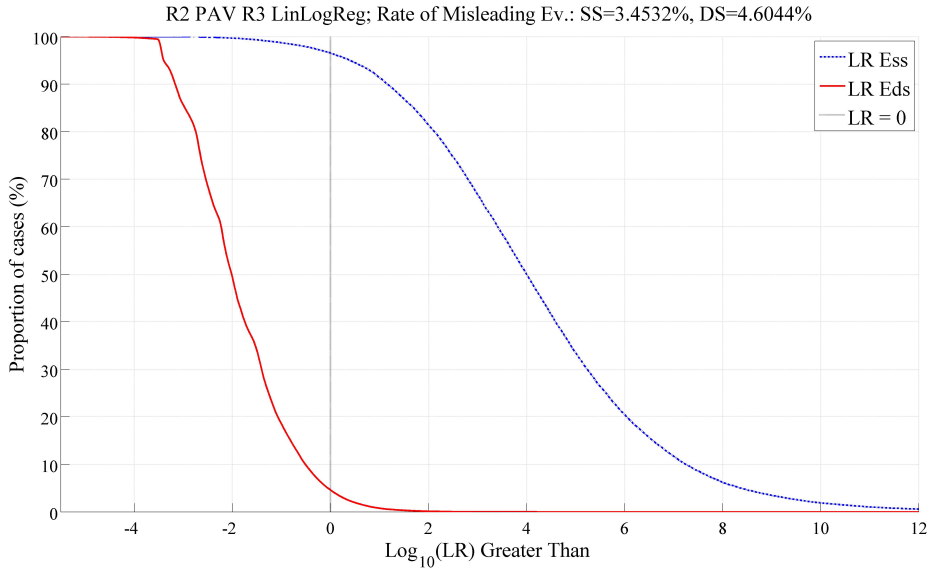


Figure 11 –Multimodal approach, Tippett plot

We can observe similar rates of misleading evidence in both cases (figures 10 and 11), however unlike the multimodal approach, in some cases (roughly 3% of the comparisons) the baseline KDE provides unjustifiably high LR values.

9.2 Discrimination

In this section we will have a look at the discriminating power of the two approaches, in the form of Detection Error Trade-off (DET[9]) curves [26] and Equal Error Rates (EER). On the DET curves below (figures 12 – 14) we see somewhat sub-optimal performance of the baseline KDE system (indicated by the deviation from the linear in the top left).

[9] The DET curve is a 2 dimensional plot of false acceptance and false rejection rates evenly handling both error types. The error rates are consecutively plotted on a Gaussian-warped scale. Thus, linearity of the DET curves is due to the assumed "normality" of the LRs. The closer the curve to the coordinate origin, the better the discrimination capabilities of the model. [26]

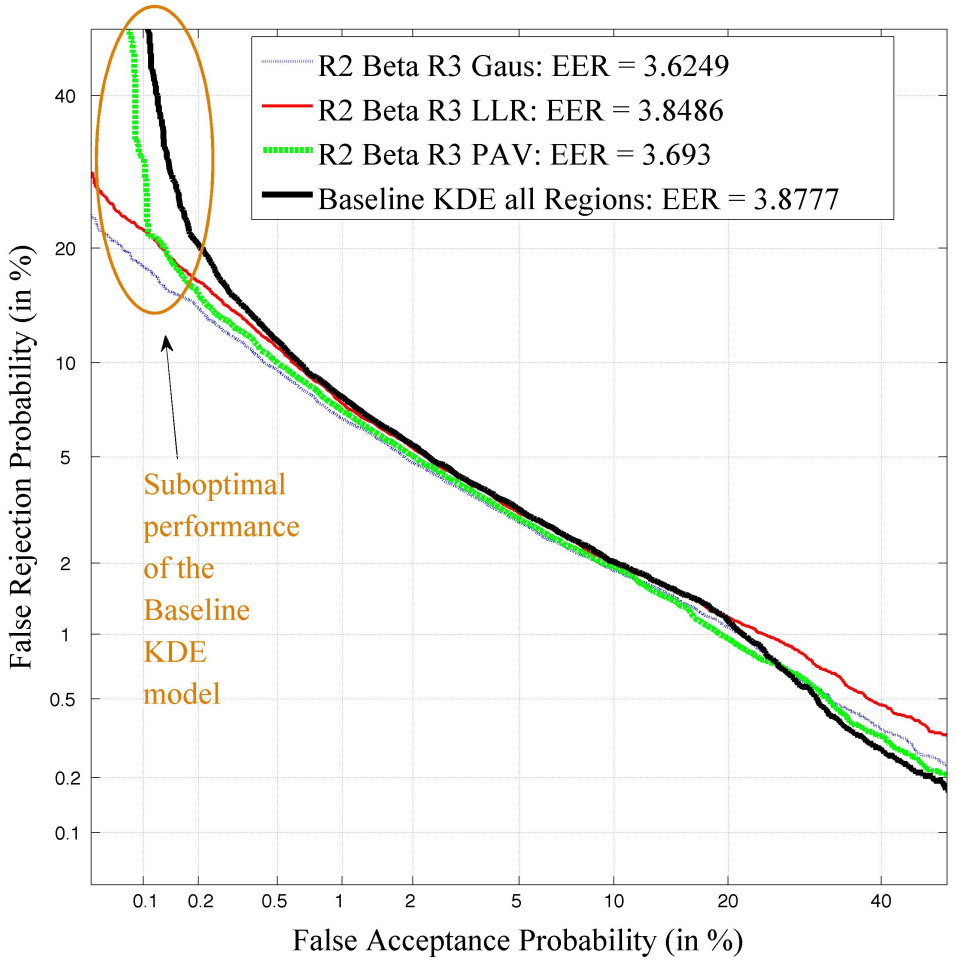


Figure 12 – Multimodal models compared to KDE baseline

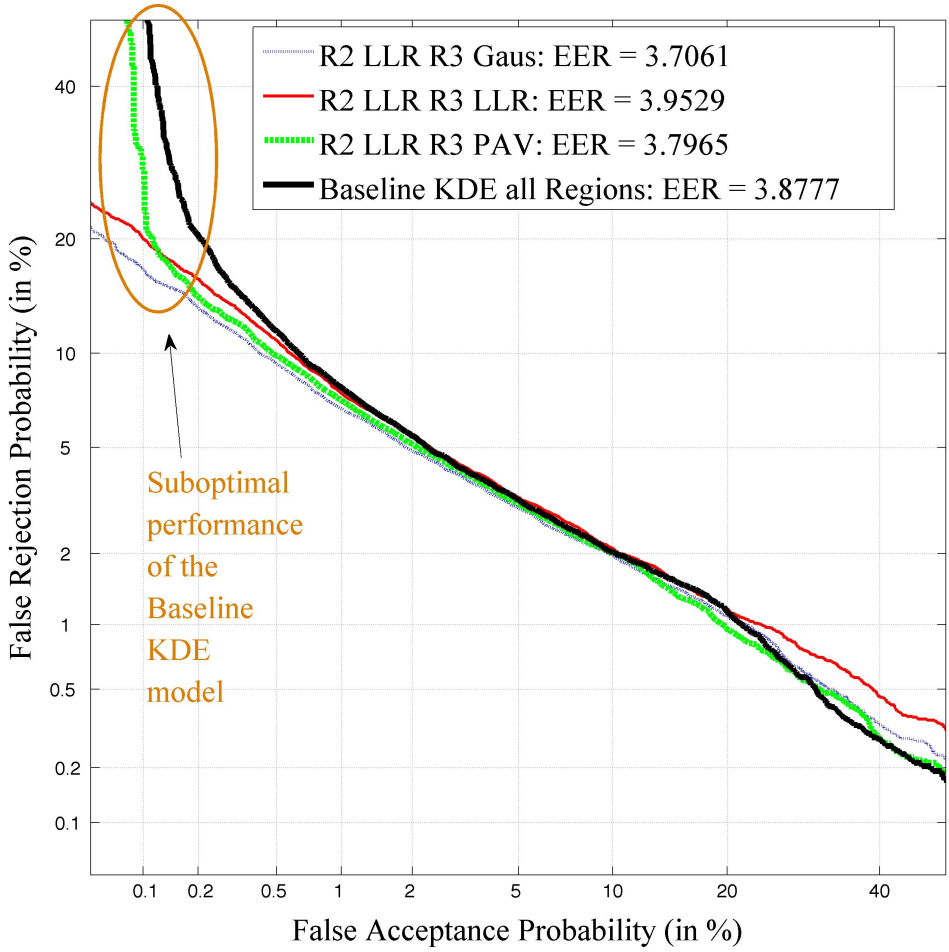


Figure 13 – Multimodal models compared to KDE baseline

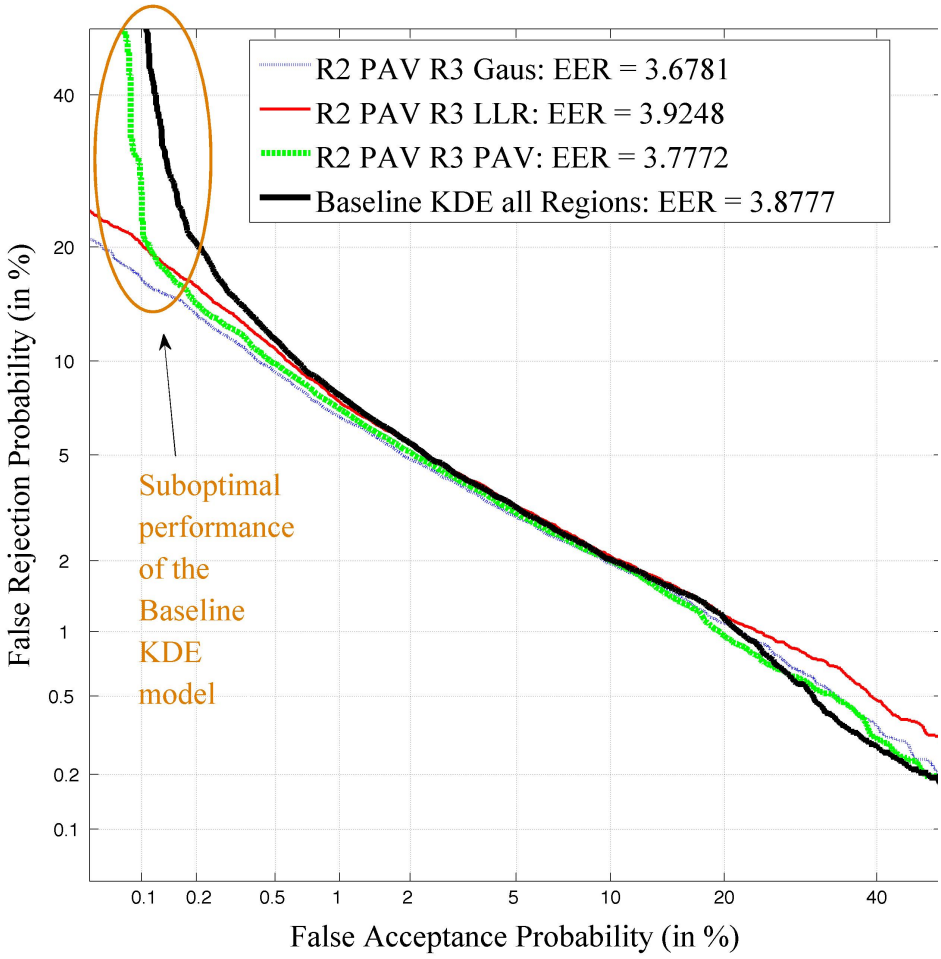


Figure 14 – Multimodal models compared to baseline KDE model

Similarly to the Tippett plots shown earlier, DET plots confirm the deviation from the optimal performance of the KDE model (thick solid black curve). They also show similar behaviour for the models where R3 scores were modelled using PAV (thick dashed green curve). Analysing Tippett plots we observed that in some cases the baseline model produces extreme LR values reflecting more an artefact of the modelling approach than expressing the real evidential value of the findings. In the top left corner of the DET curves (figures 12 -14) we clearly see a deviation from otherwise linear error rate distribution for both – the baseline KDE (mainly due to the over-fitting) and the multimodal approach in which the R3 region scores are modelled using the PAV algorithm (which appears not to be the best model

choice for almost Gaussian distributed scores). This happens because some of the DS evidence scores (roughly 0.1% of the total DS scores) yield an extremely big LR value, strongly supporting the wrong proposition. This is a highly undesirable effect, which will have consequences in the reliability of the LR, which are provided by the KDE model.

9.3 Calibration

Sub-optimal performance of the baseline KDE is further reflected in the ECE plots (ECE plot explained below).

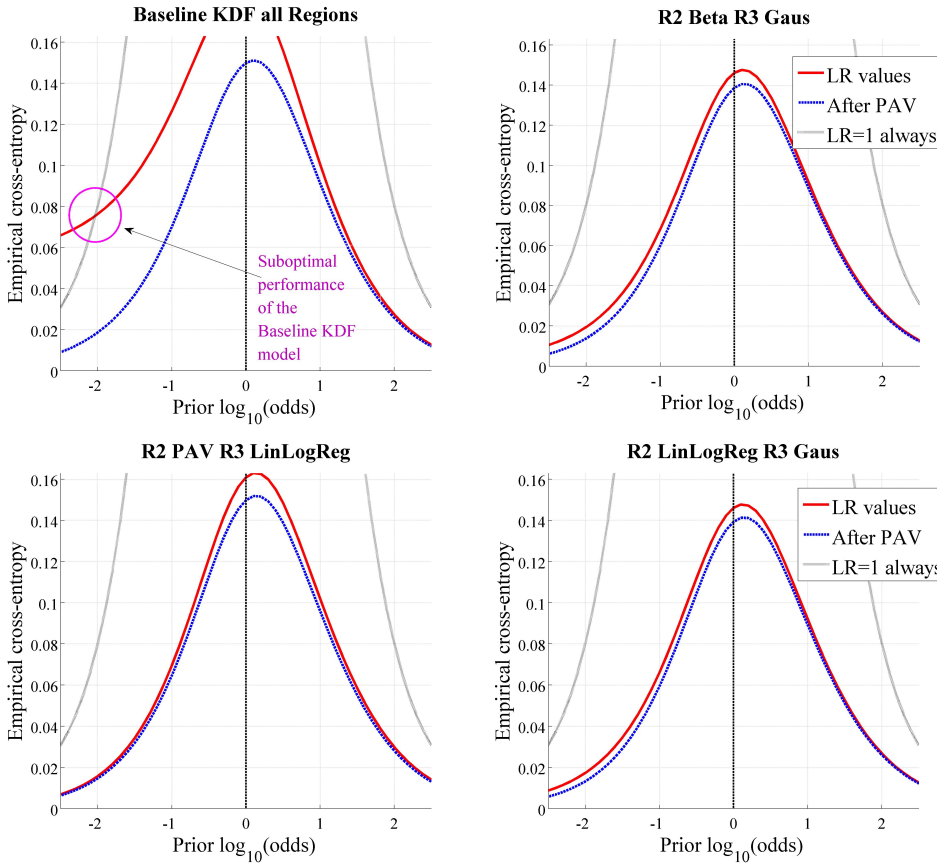


Figure 15 – Empirical Cross Entropy – selected multimodal systems compared to the KDE baseline

The red solid curve – ECE represents the accuracy of the LRs (the lower the red solid curve, the better the accuracy) and the blue dashed curve represents the discrimination of the LRs (the lower the blue dashed curve the better the discriminating power of the LR). The red solid minus the blue

dashed curve provides calibration of the LRs. The smaller the distance between the red solid and blue dashed curves, the better the calibration of the system. The lower the blue dashed curve, the better the discriminating capabilities of the system. Ideally, both red and blue line should be below the black dotted curve, which represents a reference system that continuously returns *LR = 1*. The ECE has information-theoretical interpretation and also there is a direct link between the Cllr and its ECE. The Cllr on the ECE plot lays on the intersection of the red solid curve and "zero" prior log odds and the $Cllr^{min}$ lays on the intersection of the blue dashed curve with the "zero" prior log odds line.

We can observe an undesired behaviour of the baseline KDE model at the prior log odds smaller than 10^{-2}, where its performance is even worse than that of the reference system which constantly returns LR = 1 (LR = 1 is equivalent to I don't know which of the H_p / H_d to support). The experiments and the visualisation tools used seek to present operational limits / constraints of any model developed and in this article are illustrated on the KDE model. This does not mean that the KDE should not be used in similar cases for modelling multimodal distributions; it simply shows the operational constraints of this model. It also warns about the low reliability of the LRs under certain situations, particularly when the prior odds in the case will be low. In other words, it appears to be "safe" to rely on the LR produced by the baseline KDE model only for the prior-log-odds bigger than 10^{-2}. Conversely, the rest of the multimodal models proposed in this work are reliable in all cases, because the calibration of those models is good for all the regions of the prior odds (example shown in Figure 15). Also all the multimodal models present better accuracy than the reference system outputting LR=1 always.

All of the multimodal systems provide more calibrated LRs than the baseline model. The overall results of the experimental section are summarized in table 3:

Table 3. Discriminating capabilities and calibration of the different approaches

Multi Modal Approach			Performance		
Region 1	**Region 2**	**Region 3**	**EER**	**$Cllr^{min}$**	**Cllr**
∝ (SS/DS)	Beta	Gauss	3.62	0.14	0.15
∝ (SS/DS)	Beta	PAV	3.69	0.14	0.15
∝ (SS/DS)	Beta	LinLogReg	3.84	0.15	0.16
∝ (SS/DS)	PAV	Gauss	3.67	0.14	0.15
∝ (SS/DS)	PAV	PAV	3.77	0.14	0.15
∝ (SS/DS)	PAV	LinLogReg	3.92	0.15	0.16
∝ (SS/DS)	LinLogReg	Gauss	3.70	0.14	0.15
∝ (SS/DS)	LinLogReg	PAV	3.79	0.14	0.15
∝ (SS/DS)	LinLogReg	LinLogReg	3.95	0.15	0.16
Reference system			**Performance**		
Baseline KDE all regions[10]			3.87	0.15	0.19

The best calibrated and best performing system was the multimodal approach where the LRs were modelled using the Beta function for the R2 and Gaussian function in the R3 region. The improvement of this system compared to the baseline KDE was approximately 21% for the Cllr and 6.5% for the EER. The discrimination and calibration of all remaining multimodal systems provided similar results, which prove the usefulness of the multimodal approach proposed.

Note, that we are not stating that the best model for any AFIS algorithm is given by the combination of Beta and Gaussian models. These models may be very different for different algorithms. For example in [2] the AFIS algorithm used outputs scores, which are shown to be better modelled by a log-normal distribution. However the main contribution of this article is the proposed division of the score range in order to model multimodal distributions, which can be viewed as general for any biometric system. Thus the aim of this experimental section was showing the usefulness and robustness of this model based on such a division, especially with respect to the other classical approaches, such as KDE.

10. Discussion

The main drawback of methods such as KDE when attempting to model the whole range of multimodal score distributions is a poor description of the tails of the training score distributions, together with a tendency to over-fitting. In extreme cases using a KDE we observed LRs of enormous magnitude supporting the correct proposition (e.g., $LR_{Ess} = 10^{130}$, LR = ∞), or even supporting the wrong proposition (e.g., $LR_{Eds} = 10^{91}$). This provides an

[10] The performance of the baseline KDE method was only possible to measure after removing the extreme outliers (LR = infinity) and setting a hard limit at log(LR) = 30. As such the reader is required to treat the KDE baseline method results with certain amount of moderation in mind.

illusion of certainty that transcends reality [18] and leads to a misleading interpretation of the evidence. In the ECE plots we observed bad calibration of the baseline KDE model in the low prior-odds region, a problem which is avoided using any of the multimodal models proposed in this article. This issue is hard to deduce from the Tippett plots alone and that is where the Cllr and ECE then prove themselves to be valuable validation performance evaluation tools.

In the multimodal models proposed here, we have split the SS and DS score distributions into three different regions (R1, R2 and R3) depending on the evidence score observed. We have modelled the score distributions in the R2 region using the PAV, linear logistic regression and the Beta function; the scores in the R3 region using the linear logistic regression, PAV and Gaussian function; the scores in the R1 region using the ratio of probabilities of observing the score under either proposition. In the multimodal approach the resulting LRs were structured in much more confined intervals. Using the multimodal approach we did not dramatically improve the discrimination capabilities of the system in terms of EER (6.5% relative improvement for the best performing multimodal system vs. the baseline KDE), however we significantly improved the calibration (25% relative improvement for the best performing multimodal system vs. the baseline KDE). We have shown that using a multimodal approach we can produce well-calibrated LRs for the whole range of the prior odds in a case (as shown on the ECE plots in figure 13).

In the multimodal approach using different modelling techniques we obtained almost identical *Cllr* and $Cllr^{min}$ rates. In the best performing multimodal system the score distributions were modelled using Beta distribution in the R2 region and Gaussian distribution in the R3 region, achieving the EER = 3.62%, Cllr = 0.15 and $Cllr^{min}$ = 0.14. This system will be used in the future work, due to its computational simplicity and relative ease of implementation.

11. Conclusion

In this work we have focused on the model selection for the LR computation from AFIS scores presenting multimodal distributions. The criteria addressed by the model were mainly focused on the robustness – to the lack of data, to the dataset shift and to the over-fitting. The best performing model resulting from the experiment (scores modelled using Beta in region 2 and Gaussian distribution in region 3) was selected for further work. The approach proposed takes into account all regions of the multimodal score distribution and provides a well-calibrated output. The benefits and functionality of the approach were shown on multimodal score distributions

produced by an AFIS fingerprint feature extraction and comparison algorithm. We think this approach might also be useful in other forensic fields, where the automatic approaches used for forensic evaluation at source level produce multimodal distributions of distances or scores. The models used in each of the regions can be adapted to other score distributions if necessary. The idea of splitting the score range in order to handle multimodality presented in this work can be generalized and used with other biometric systems.

With the model selected, in the future work we will proceed further with the definition of additional validation criteria, apply "real" forensic marks to the model selected (rather than simulated marks) and reproduce the results for 5 – 15 minutiae configurations based on the data from real forensic casework.

Acknowledgements

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) in cooperation with the Netherlands Forensic Institute, the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid and the National Police Services Agency of the Netherlands.

References

[1] – D. Meuwly and R.G.F. Veldhuis, *Forensic Biometrics: From two communities to One Discipline*, International Conference of the Biometrics Special Interest Group, Proceedings of the BIOSIG 2012, pp. 207 – 218

[2] – Nicole Egli, *Interpretation of partial fingermarks using an automated fingerprint identification system*, Doctoral Thesis, 2009

[3] – Amanda B. Hepler et al., *Score-based likelihood ratios for handwriting evidence*, Forensic Sci. Int. (2012), 219 (1-3): 129-40

[4] – C. Neumann, I. Evett et al., *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. (2011), 207(1-3):101-5

[5] – Cao et. al., *"Minutia handedness: A novel global feature for minutiae-based fingerprint matching"*, Pattern Recognition Letters 33, (2012), pp. 1411 – 1421

[6] – Feng et. al., *"Orientation Field Estimation for Latent Fingerprint Enhancement"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2013), v. 35, n. 4

[7] – D. Meuwly, *Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique*, PhD thesis, 2001

[8] – D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken, *Information-theoretical assessment of the performance of likelihood ratio methods*, Journal of Forensic Sciences (2012)

[9] – D. Ramos, J. Gonzales-Rodriguez, *Reliable support: measuring calibration of likelihood ratios*, Forensic Sci. Int.: in press (2013)

[10] – D. Ramos, *Forensic Evaluation of the Evidence using Automatic Speaker Identification System*, Doctoral Thesis, 2007
[11] – D. Meuwly, *Forensic Individualization from Biometric Data*, Science & Justice (2006), v. 46, pp. 205 – 213
[12] – N. Egli et al, *Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability*, Forensic Sci. Int. (2007), v. 167, pp. 189 – 195
[13] – J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro and J. Ortega-Garcia, *Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems*, Forensic Sci. Int. (2005), v. 155, n. 2-3, pp. 126-140
[14] – J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and J. Ortega-Garcia, *Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition*, IEEE Transactions on Audio, Speech and Language Processing (2007), v. 15, n. 7, pp. 2104 – 2115
[15] – C. M. Rodriguez et al., *Introducing a semi-automated method to simulate a large number of forensic fingermarks for research on fingerprint identification*, Journal of Forensic Sciences (2012), 57/2
[16] – R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd edition, ch. 1, WILEY,1997
[17] – D. Lucy, *Introduction to Statistics for Forensic Scientists,* Edition 1, WILEY, 2005
[18] – J.G. Moreno-Torres et al, *A Unifying view of Dataset Shift in Classification*. Pattern Recognition (2011)
[19]– C. Champod and I. Evett, *A probabilistic Approach to Fingerprint Evidence*, Journal of Forensic Identification (2001), 51(2): pp. 101-122
[20] – N. Brummer, L. Burget, et al., *Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006,* IEEE TASLP (2007), v.15, n.7
[21] – N. Brummer and J. du Preez, *Application Independent Evaluation of Speaker Detection*, Computer Speech and Language (2006)
[22] – Ch. Brenner, *Fundamental problem of forensic mathematics-The evidential value of a rare halotype*, Forensic Sci. Int.: Genetics 4 (2010), pp. 281-291
[23] – E. T. Jaynes, *Probability Theory: The Logic of Science*, ch. 18, WILEY, 1994
[24] S. L. Zabell, *The Rule of Succession*, Erkenntnis v. 31, n. 2/3, pp. 283-321, 1989
[25] W. E. Johnson, *The Logical Foundations of Science*, Cambridge University Press, Logic, Part III, 1924
[26] – A. Martin et al., *The DET Curve in Assessment of Detection Task Performance*, National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940; 1997

Chapter 5

Measuring Coherence of Computer-Assisted LR Methods: Experimental Example

submitted – Forensic Science International (April 2014)

Rudolf Haraksim
Daniel Ramos
Didier Meuwly
Charles Berger

Abstract

Measuring the performance of forensic evaluation methods that compute likelihood ratios (LRs) is relevant for both the development and the validation of such methods. A framework of performance characteristics categorized as primary and secondary is introduced in this study to help achieve such development and validation. Ground-truth labelled fingerprint data is used to assess the performance of an example likelihood ratio method in terms of those performance characteristics. Discrimination, calibration, and especially the coherence of this LR method are assessed as a function of the quantity and quality of the trace fingerprint data. Assessment of the coherence revealed a weakness of the comparison algorithm in the computer-assisted likelihood ratio method used.

1. Introduction

Forensic research makes progress in the field of evaluation of forensic findings. An increasingly adopted approach [1] uses a logical framework based on Bayes' Theorem to report forensic evidence in terms of likelihood ratios [1,2]. Computer-assisted LR methods (also referred to simply as *LR methods*), have been developed to assist the forensic practitioner in his role of forensic evaluator [3,4,5,6,7,8,9]. In these methods pattern recognition algorithms are often used for the feature extraction (analysis), the feature comparison, and statistical models are used for the evaluation of the forensic findings.

In this article the term validation refers to a series of experiments, and the application of a set of performance metrics and validation criteria to demonstrate validity. This is different from Ref. [10], where the term validity was defined as a single metric and equated to accuracy. The specific performance characteristics, performance metrics and validation criteria are used to describe the performance of methods computing LRs and to assess the limits of their validity when used for casework. The LR describes the strength of the evidence, and does not imply a decision by itself. Therefore, the validation of LRs is not the validation of a decision process, but of a description process. We define *coherence* as a performance characteristic, understood as the ability of a LR method to perform better and to maintain low rates of misleading evidence as some measured parameters influencing quality in the features studied improve, and vice versa. A concrete example is provided when studying and assessing the coherence of a forensic fingermark evaluation method, based on a comparison algorithm of an AFIS (Automated Fingerprint Identification System). When analysing the coherence of the method we hope to observe a LR value increasing with the intrinsic quantity and quality of the information present in the trace data (such as the length of a speech fragment or the number of minutiae in a fingermark).

Forensic service delivery makes progress in the field of quality assurance. Initiatives in the European Network of Forensic Science Institutes (ENFSI) focus on best practices, method validation and service accreditation [11,12]. But because LR methods for forensic evaluation are still very new, the question of their validation has not been addressed yet in the context of quality assurance. Currently, performance characteristics, performance measures, and validation criteria exist to assess analytical forensic methods [13] and human-based methods used for forensic evaluation [14,15]. These approaches are however not suitable for the validation of LR methods developed for forensic evaluation. Such a validation requires specific performance characteristics, performance measures and validation criteria related to the nature of the LRs and the computation methods involved.

Studying the coherence contributes to describing the performance of the LR method using datasets in which some measurable parameters influencing the strength of the evidence vary. The variation of the length of utterances in forensic automatic speaker recognition and the variation of the number of minutiae in fingermarks are examples of such parameters. Coherence is a highly desirable property of a LR method.

The remainder of this article is structured as follows. The definition of coherence in a set of performance characteristics is presented in Section 2. Section 3 introduces the experimental example for assessment of the coherence of LRs assigned using computer-assisted methods. The different datasets used to measure the performance characteristics are described in Section 4, while the relevance of the use of the datasets and their specificity is described in Section 5. The performance metrics related to the performance characteristics used are introduced in Section 6. Results in terms of coherence of the LR method are presented in Section 7, followed by general discussion and conclusions in Section 8.

Throughout this article we frequently use the terms performance characteristic – a measurable property (or a set of measurable properties) of LRs; and performance metrics – a quantitative description of the performance characteristic. These definitions are ours and the terms may have different meanings in other related works.

2. Performance Characteristics

Several performance characteristics have been defined to assess the performance of computer-assisted LR methods developed for forensic evaluation. We propose to structure them into primary and secondary performance characteristics. Primary performance characteristics directly measure desirable properties of the LRs. The secondary performance characteristics measure how sensitive primary performance characteristics are to factors like the quantity of information in the data, and to the forensic casework circumstances, such as degraded quality, different technical and temporal conditions related for example to the acquisition of trace and test[1] specimens, representativeness of the data, etc.

To assess the performance of computer-assisted LR methods, several performance characteristics have been defined recently in forensic evaluation [16]. A very important one is accuracy, defined as the

[1] In the fingerprint modality the trace usually refers to the fingermark recovered from the crime scene and the test specimen usually refers to the rolled, inked fingerprint of a suspected individual.

combination of discrimination (discriminating power) and calibration [16,17,18].

- **Accuracy** is defined as the closeness of agreement between the decision – driven by a LR computed by a given method – and the ground truth. The LR is accurate if it helps to lead to a decision that is correct[2]. In case of source level inference, the ground truth relates to the following pair of propositions:

  - H_p: The pair of specimens compared come from the same source (SS)
  - H_d: The pair of specimens compared come from different sources (DS)

  Ground-truth labels are defined as SS (same source) when the LR was calculated for specimens originating from the same source, and as DS (different source) when the LR was calculated for specimens originating from the different sources. If an experimental set of LR values is to be evaluated, and the corresponding ground-truth label of each of the LR values is known, then a given LR value is evaluated as more accurate if it supports the true (known) proposition to a higher degree, and vice-versa.

- **Discrimination** (or discriminating power) is a property of a set of LRs that allows distinguishing between the propositions involved. See [16,17] for details.

- **Calibration** is another property of a set of LRs. Perfect calibration of a set of LRs means that those LRs can probabilistically be interpreted as the evidential value of the comparison result for either proposition in a Bayesian evaluation framework. Finding a LR = x will be x times more probable under H_p than under H_d (in other words, the LR of the LR is the LR [19,20]). Under those conditions the LR is exactly as big or small as is warranted by the data. Well-calibrated LRs tend to increase with the discrimination of a given method [16].

[2] The LR does not imply a decision, but the accuracy measurement is inserted in a decision-theoretical process as explained in [16,17].

2.2 Example factors influencing the primary performance characteristics

- **Quality** of the data is a measurable parameter that has no information about the proposition, but can predict the performance of that comparison. In other words, specimens of high quality to be compared in a forensic case predict good performance of that comparison while low quality samples predict bad performance of a LR method. Examples are the quantity of minutiae in fingerprint comparisons or the signal-to-noise ratio in speaker recognition.

- **Quantity**[3] or amount of data, e.g. the length of a speech fragment, the number of minutiae in a fingermark, etc.

- **Representativeness** of the data used to train the LR method for the data used in operational conditions. The smaller the dataset shift [22] between the two, the more representative the training data is for those in operational conditions. In the next section we present an experimental example to illustrate the measurement of *coherence* in this framework, the datasets used together with the LR method and the performance measures used to establish the *coherence* LRs produced by the model tested.

2.3 Secondary performance characteristics

- **Coherence** is defined as the ability of the method to yield LRs with better performance with an increase of the quantity and quality of the information present in the data.
- **Generalization** is defined as the property of a given method to maintain its performance under dataset shift. LR method 1 generalizes better than LR method 2 if, under similar conditions of dataset shift in both methods, the performance of method 1 decreases less than the performance of method 2.

  Robustness is the ability of the method to maintain performance when the quantity or quality of the data decreases. For instance, method 1 is more robust to data sparsity than method 2 if, with decreasing amount of data, the performance of method 1 decreases less than the performance of method 2.

[3] Quality is not an intrinsic property, but depends e.g. on the ability of a system to extract features from the specimens, and to compare and evaluate this information.

In the next section we present an experimental example to illustrate the measurement of coherence, discuss the datasets used in the LR method development and the performance measures used to establish the coherence of LRs produced by the method.

3. Measuring Coherence: Experimental example with fingerprint LR's

The comparison of the minutiae of a fingermark and fingerprint using an AFIS comparison algorithm results in a comparison score. The strength of evidence of this score can be assessed in terms of a LR. Since the LR method in our case consists of modelling the SS and DS score distributions, it is referred to as a LR model from here on. A detailed description of the LR model used – derived from [6] – is beyond the scope of this article, since the aim is to present the validation methodology with the focus on the analysis of coherence.

Recall the set of propositions from the Section 2.1. Without loss of generality we can rephrase them to fit our fingerprint example:

- H_p: The fingermark and fingerprint come from the same source (SS)
- H_d: The fingermark and fingerprint come from different sources (DS)

Having defined the set of propositions with respect to which the comparison scores are evaluated, we proceed to build the LR model [6]:

- Use the minutiae comparison algorithm to compare the fingermarks of a suspect with the fingerprint of a suspect to produce a same source score distribution (SS)
- Use the minutiae comparison algorithm to compare the crime scene fingermark to the fingerprint of a suspect to produce the evidence score (E)
- Use the minutiae comparison algorithm to compare the crime scene fingermark to a database of fingerprints of individuals other than the suspect to produce a different source score distribution (DS)
- Model the SS and DS score distributions using probability density functions or a discriminative approach e.g. using logistic regression [18]
  - Compute the strength of the evidence given by the likelihood ratio:

$$LR = \frac{p(E \mid H_p)}{p(E \mid H_d)} \qquad \text{(eq. 1)}$$

The comparison algorithm applied in this work to generate scores is a commercial product Motorola bis 9.1, used as a black-box. The minutiae extraction and comparison technology remains outside the scope of this work, but we still present some of its functionality. The algorithm used is speed-optimized and outputs comparison scores in three separate score ranges. The comparison algorithm considers two different comparison methods depending on the number of minutiae in the mark: one for 5 to 10 minutiae configurations and one for configurations of 11 and more minutiae. The maximum score is directly proportional to the number of features in agreement. We get back to the two methods of the comparison algorithm in section 7.

4. Datasets used

We use two different datasets – one with simulated fingermarks to obtain the values of the parameters of the model and a relatively small one with forensic fingermarks to determine validity of the LR model for forensic casework. In the following sections we present the two datasets used in more detail. We justify their degree of similarity both numerically using the Kullback-Leiber (KL) divergence, a measure commonly used in probability and information theory [21], and visually by comparing the histograms of selected score distributions.

4.1 Forensic Dataset

The forensic dataset consists of data from real forensic cases: 58 identified fingermarks in 12-minutiae configuration and their corresponding fingerprints. The ground-truth labels of the dataset, indicating whether a fingermark / fingerprint pair originates from the same source is denoted as "ground-truth by proxy" because of the nature of the pairing between fingermarks and fingerprints: they have been assigned after examination by human examiners, taking into account not only the 12 minutiae, but also other minutiae, ridge pattern, etc. The minutiae feature vectors[4] of the fingermarks have been manually extracted by examiners while the minutiae feature vectors of the fingerprints have been automatically extracted using a feature extraction algorithm and manually checked by examiners.

In order to obtain multiple minutiae configurations for the LR method validation, the minutiae extracted from the fingermarks have been clustered into configurations of 5 to 12 minutiae, according to the method described in [23]. Following the clustering procedure we obtain 481 minutiae clusters in a

[4] Minutiae feature vectors of a fingermark or fingerprint in our case consist of feature type, position, and orientation (parallel to the ridge flow).

5-minutiae configuration from the 58 fingermarks with 12 minutiae. For each cluster in the marks, a same-source (SS) score is obtained by comparing each minutiae cluster from a fingermark with the corresponding reference print. Similarly, a different-source (DS) score distribution is obtained by comparing a fingermark to a subset of a police fingerprint database. This subset consists of roughly 10 million 10-print cards captured in 500 dpi. The higher the number of minutiae in each cluster, the lower the number of clusters, as can be seen in Table 1. An example of a forensic fingermark is presented in Figure 1.

Table 1: Forensic dataset sizes for SS and DS scores. Note that the number of SS scores is the same as the number of clusters for a given minutiae number.

	SS scores	DS scores
5 minutiae	481	10,283,780
6 minutiae	432	9,236,160
7 minutiae	426	9,107,880
8 minutiae	387	8,274,060
9 minutiae	342	7,311,960
10 minutiae	286	6,114,680
11 minutiae	190	4,062,200
12 minutiae	58	1,240,040

4.2 Simulated marks Dataset

Simulated fingermarks were obtained by capturing a video sequence of a finger of a known individual moving on a glass plate in different directions in order to capture as much distortion as possible. Reference print(s) of the same finger of the same individual were recorded on a 10-print card. This dataset consists of 200 individuals (100 male and 100 female) times 10 video sequences (1 per finger). The process of obtaining the simulated marks dataset is described in detail in [23].

The simulated dataset consists of 25,000 fingermarks of known origin, from which we produce the SS and DS score distributions (the number of simulated fingermarks differs per configuration[5] as shown in Table 2). An example of a simulated fingermark on a forensic background is presented in Figure 1.

[5] The difference in the number of simulated fingermarks per configuration is caused by the sub-sampling of the original fingerprint captured from a video sequence of a finger moving on the glass surface of a fingerprint sensor [21].

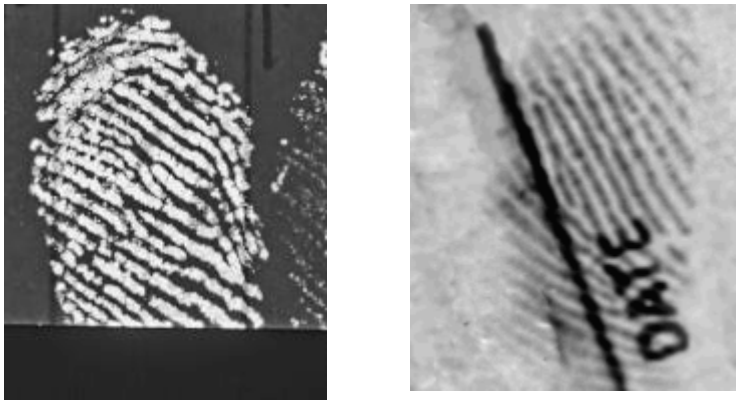


Figure 1 - Forensic (left) vs. simulated (right) fingermark.

There are several advantages in using a simulated fingermarks dataset:

1.) The contrast and the clarity of the images captured from the video sequences are high which allows for automatic minutiae extraction.
2.) It is relatively easy and cost-efficient to scale up the experiment and produce more simulated marks.

Table 2: Simulated dataset SS and DS dimensions

	SS scores	DS scores
5 minutiae	16,653	33,306,000
6 minutiae	25,058	50,116,000
7 minutiae	24,876	49,752,000
8 minutiae	25,015	50,030,000
9 minutiae	25,036	50,072,000
10 minutiae	24,994	49,988,000
11 minutiae	24,658	49,316,000
12 minutiae	24,443	48,886,000

5. Measuring similarity between the datasets

Since the two datasets (forensic and simulated) were acquired under different conditions, it is appropriate to establish the degree of similarity between the distributions of the scores generated by them. We use the KL (Kullback-Leiber) divergence to quantitatively express the similarity between the DS score distributions of the two datasets. We convert the score distributions into normalized histograms representing relative frequencies of observations of comparison scores in each of the two datasets – forensic (*F*) and simulated (*S*) – and compute the KL divergence as follows:

$$KL = \sum_i F(i) \cdot \ln\left(\frac{F(i)}{S(i)}\right) \quad \text{(eq. 2)}$$

where the index *i* in Equation 2 refers to the *i*-th bin in the histogram. Note that if the two distributions *F* and *S* are identical the KL divergence is equal to zero, and the more similar the histograms are, the smaller is the divergence.

Since the KL divergence is a non-commutative distance between the two distributions *F* and *S*, we propose to calculate the distance between *F* and *S* and *S* and *F*. The final, symmetric KL divergence is represented as the average of those two distances:

$$KL_{sym} = \frac{\sum_i F(i) \cdot \ln\left(\frac{F(i)}{S(i)}\right) + \sum_i S(i) \cdot \ln\left(\frac{S(i)}{F(i)}\right)}{2} \quad \text{(eq. 3)}$$

where index *i*, as in Equation 2 refers to *i*-th bin in the histogram.

The KL divergence of the two datasets, calculated using Equation 3, is presented in Table 3. Recall from Equation 2 that the more similar the two score distributions are, the closer to zero is the resulting KL_{sym}. The highest degree of similarity between the simulated and the forensic dataset is found for the fingermarks clustered in 6-minutiae configuration, while the lowest degree of similarity is found for the fingermarks in 5-minutiae configuration.

Table 3: KL_{sym} divergence of the DS comparison scores (simulated and forensic dataset)

Configuration	$\boldsymbol{KL_{sym}}$
5 minutiae	0.0336
6 minutiae	0.00725
7 minutiae	0.0105
8 minutiae	0.01915
9 minutiae	0.01295
10 minutiae	0.01025
11 minutiae	0.01375
12 minutiae	0.0107

For better understanding the KL divergence, the similarity of the two score distributions can also be visually assessed in Figures 2 and 3. We compare the normalized histograms of the scores for the simulated and the forensic datasets, presenting as an example the results for the 5-minutiae configurations (lowest degree of similarity KL_{sym} = 0.033) and the 6-minutiae configurations (highest degree of similarity KL_{sym} = 0.007). The difference

between these most similar and least similar score distributions appears negligible in Figures 2 and 3.

Establishing a degree of similarity between the two datasets acquired under different conditions is a very important step in LR method development, especially when using probability density functions to produce LRs. We conclude that the simulated dataset is a representative approximation of the forensic dataset.

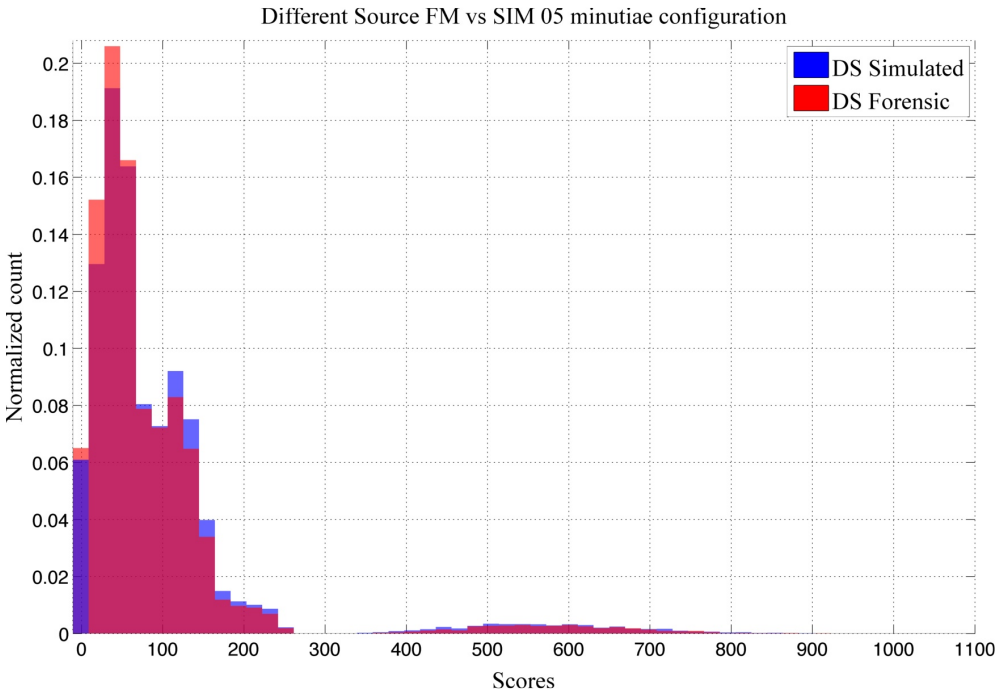


Figure 2 - Normalized score distribution for 5-minutiae configurations of forensic versus simulated datasets (lowest degree of similarity).

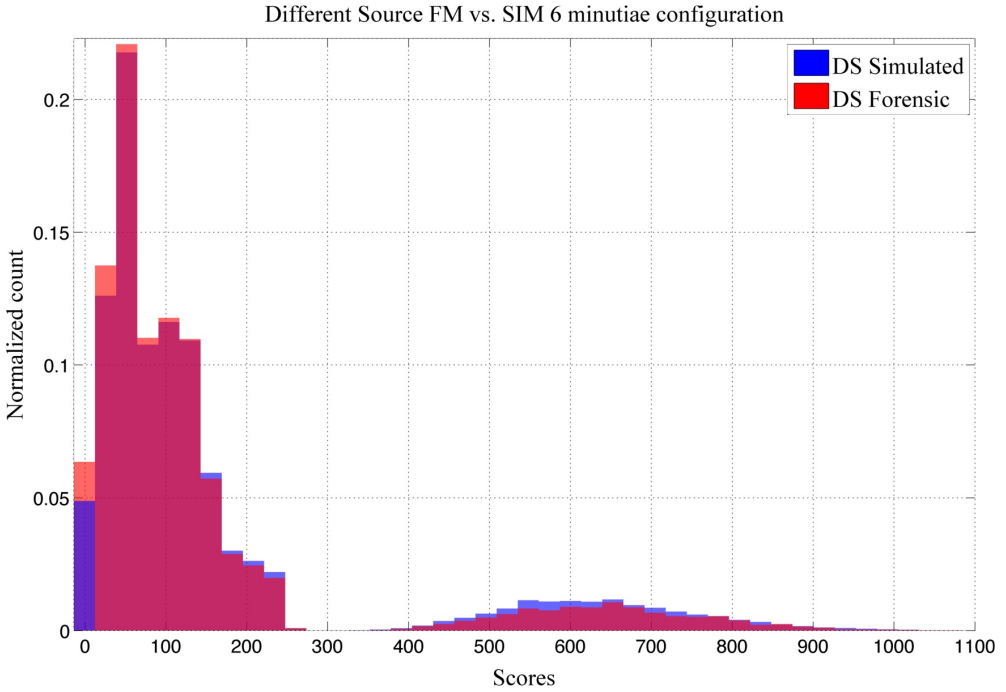


Figure 3 - Normalized score distribution for 6-minutiae configurations of forensic versus simulated datasets (highest degree of similarity).

6. Performance metrics used

In this part we introduce a set of plots and performance measures used to evaluate the performance of the model for different minutiae configurations. Although alternative measures can be used to illustrate the coherence of the LR method, we think that visual representations and measures proposed are sufficient.

6.1 Detection Error Trade-off (DET) plot and Equal Error Rate (EER)

The DET plot [25] presents the false acceptance rate (FAR) as a function of the false rejection rate (FRR). The error rates are plotted on a Gaussian-warped scale. This makes the DET curves linear when the log(LR) values are normally distributed. The closer the curve is to the origin, the better the discrimination of the method. The intersection of a DET curve with the diagonal of the DET plot marks the Equal Error Rate (EER). The EER is used as a performance measure to show the coherent behaviour of the LR method. For example, when comparing forensic fingermarks in different minutiae configurations the EER should be larger for configurations with fewer minutiae (see Figure 4). Even if a DET plot is meant to characterize a

system that makes decisions, it is informative about the coherence of the LR method when evaluating datasets with different quantities of information.

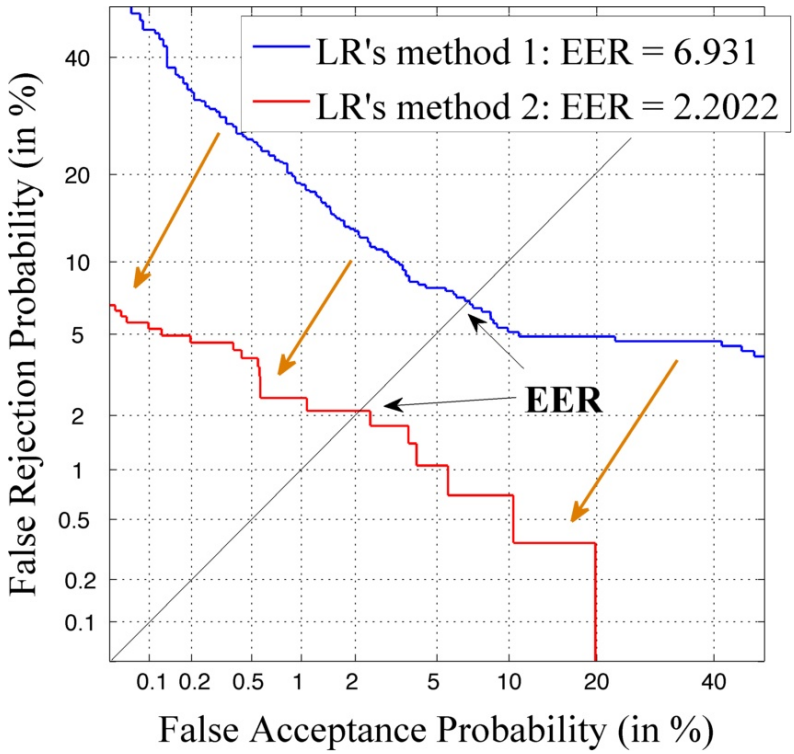


Figure 4 - DET curves showing the performance of the same LR method with different quantities of information. The dashed curve shows worse discrimination in the LRs of comparisons for 6-minutiae configurations, while the solid line shows better discrimination in the LRs of comparisons for 10-minutiae configurations. The equal error rates are given by the intersection of the curves with the diagonal of the plot, and are 6.9% and 2.2%, respectively.

6.2 Tippett plots

Tippett plots [26] are representations of cumulative distributions of LRs. The curves in it represent the proportion of comparisons resulting in a log(LR) greater than t versus that value t, when either proposition H_p or H_d is true. In a Tippett plot, the rates of misleading evidence for either proposition can be observed at the intersection of each of the curves and the vertical at $t = 0$. The log(LR) value zero corresponds to a LR value of 1. Using Tippett plots it is relatively easy to distinguish the performance of an LR method when presented with different quantities of evidential information.

Examples of Tippett plots are shown in Figure 5 for the 5 and 10-minutiae configurations. The decrease in misleading evidence due to the 5 additional minutiae can clearly be seen.

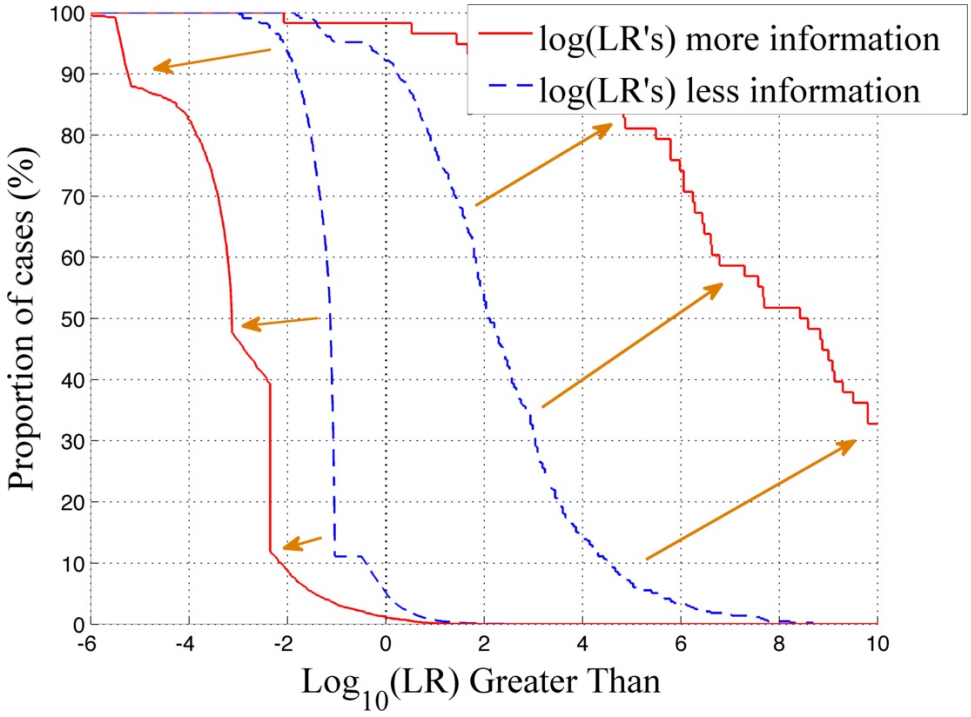


Figure 5 – Tippett plots showing the performance of the same LR method with different quantities of information. Dashed lines show less evidential information captured in the LRs of comparisons for 5-minutiae configurations, while solid lines show more evidential information captured in the LRs of comparisons for 10-minutiae configurations.

Using the Tippett plots it is relatively easy to distinguish the quantity of the evidential information within the LR values captured by the LR method presented with datasets under different conditions. Tippett plots for a LR method evaluating the strength of evidence in fingermarks with 5-minutiae configuration (dashed line) and 10-minutiae configuration are presented in Figure 5. Arrows indicate the change of the Tippett plots when LR method is presented with additional information (in this case 10 additional minutiae).

6.3 Empirical Cross-Entropy (ECE) plot and the Log likelihood ratio cost (CIlr)

The Empirical Cross-Entropy or ECE plot [16,17] is a representation of the performance and calibration of the LR values and complements other already established methods such as those discussed above [17]. The C_{llr} is a closely related cost function of the log(LR) defined in Ref. [18]. ECE and C_{llr} are both lower when the likelihood ratio correctly supports the ground-truth proposition. The difference between them lies in the interpretation of both measures. The C_{llr} is interpreted as an average decision cost for all prior probabilities. On the other hand, the ECE has an information-theoretical interpretation as the amount of information lacking compared to full knowledge of the ground-truth, on average in a given set of LR values. The C_{llr} is an average over costs and priors, and therefore is not giving the performance for a given value of the prior, but for an average of all possible priors. An ECE-plot shows the ECE for a certain range of priors [16,17]. It can be easily shown that the C_{llr} is the ECE at prior log-odds of 0 (i.e. a prior probability of 0.5). In this sense, the ECE is a more general and interpretable performance metric than the C_{llr} in a forensic context, where no decision is to be made by the forensic examiner and where the value of the prior changes very much from one case to another. It also appears to be more suitable to show the validity of a method over a relevant set of priors that are generally unknown. On the other hand, the C_{llr} is a summary of the ECE in a single number, useful for comparing and ranking methods.

We use the C_{llr} as a measure of accuracy, consisting of two components: discrimination $C_{llr}^{\min}$ and calibration C_{llr}^{cal} [18]. The solid curve in the ECE plot also represents accuracy: the lower it is, the better the accuracy of the method. The dashed curve represents the discrimination, and is sometimes referred to as “accuracy after PAV”, because it is the ECE after applying the Pool Adjacent Violators algorithm (PAV). It is an algorithm that improves the calibration of a set of LRs while not affecting their discrimination, see [18] for details. The difference between these two curves represents calibration losses: the smaller the distance, the better the LR method’s calibration.

Besides the information-theoretical aspect, the ECE provides the “range of application” of the LR method under evaluation. A LR method should perform better than a reference method producing LR = 1 for the whole range of prior probabilities. In a range of prior probabilities where this is not the case, using the LR method would be worse than not using any method at all.

Figure 6 presents an example for the sake of illustration, showing the ECE plots of the LR method evaluating the fingermarks in 5-minutiae configuration in two different settings: uncalibrated and calibrated with PAV. Calibrating the LR method not only improves the accuracy of the LR method (here measured by the C_{llr}), it also extends the applicable range of this method. The uncalibrated LR method presents an ECE larger than that of the reference method for prior log-odds above 0.5, which does not happen for the calibrated method. Note that the LRs used for the right hand plot were calibrated using the data from the left hand plot, which explains why applying PAV using the right hand plot's own data still reduces the ECE somewhat.

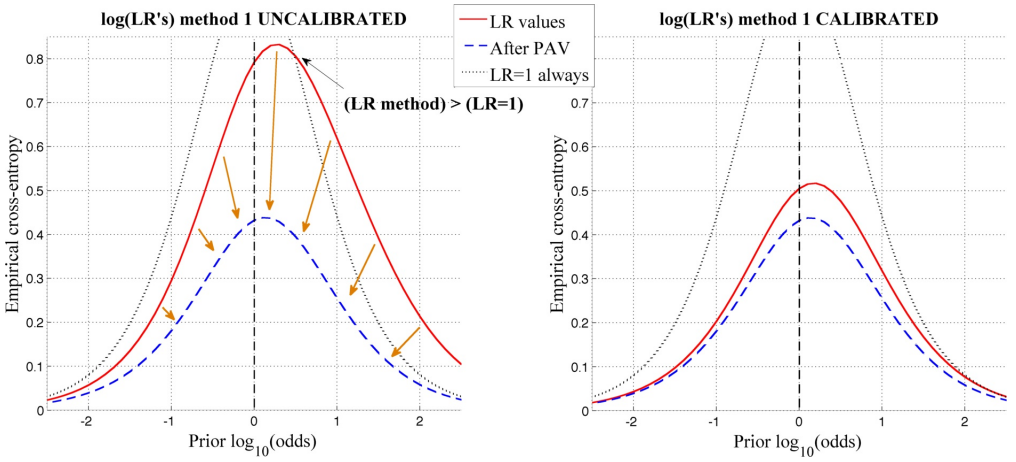


Figure 6 - ECE plots for the same LR method (same set of LR values) before and after calibration (leave one out cross-validation used for calibration). On the left-hand-side the solid curve represents uncalibrated LRs, and the dashed curve gives the ECE after PAV. The LRs on the right-hand-side are calibrated using the PAV transform resulting from the data used for the left ECE plot. The dramatic lack of calibration is visible in the left plot by the fact that above prior-log(odds) = 0.5 the ECE exceeds that of the reference method which always gives LR = 1). For that range of prior odds the uncalibrated method performs worse than a method that always returns the "I don't know" answer (i.e., always yielding LR = 1).

7. Results

We use the same LR method to produce LR values for 5 to 12-minutiae configuration comparisons. To describe the performance of the LR method for each forensic *n*-minutiae configuration dataset, the LR method is trained with the corresponding *n*-minutiae simulated fingermark dataset.

In order to establish the coherence of the LRs produced by the LR method selected, we measure the primary performance characteristics: accuracy (using C_{llr} and ECE as a measure), discrimination (using $C_{llr}^{\min}$ and ECE-

after-PAV as a measure) and calibration (using C_{llr}^{cal} and the difference between ECE and ECE-after-PAV as a measure). Recall that the coherence is not a primary but a secondary performance measure: it describes the variation of the performance of the LR method when varying quality or quantity of the information (in our case the number of minutiae).

The performance as a function of the number of minutiae is presented using ECE, Tippett and DET plots. The C_{llr}, $C_{llr}^{\min}$, and EER are determined for all minutiae configurations and presented in Table 4.

Table 4: Relative increase in performance of the LR model when introducing additional minutiae

Configuration	DET-EER	Discriminating power $Cllr^{min}$	Accuracy Cllr
5 minutiae	15.69	0.43	0.5
6 minutiae	6.91	0.26	0.28
7 minutiae	3.95	0.14	0.16
8 minutiae	2.42	0.11	0.13
9 minutiae	1.56	0.063	0.075
10 minutiae	2.19	0.063	0.074
11 minutiae	2.73	0.081	0.1
12 minutiae	1.82	0.057	0.084

The ECE plots in Figure 7 show a decreasing trend (solid curves), which corresponds to increased accuracy and discrimination (dashed curves) when increasing the number of minutiae from 5 to 10. The values for the accuracy and discrimination show the same trend and are summarized in Table 4. The sudden increase of these plots and values for the 11-minutiae configurations are related to the comparison algorithm, which changes its method from 11 minutiae onwards.

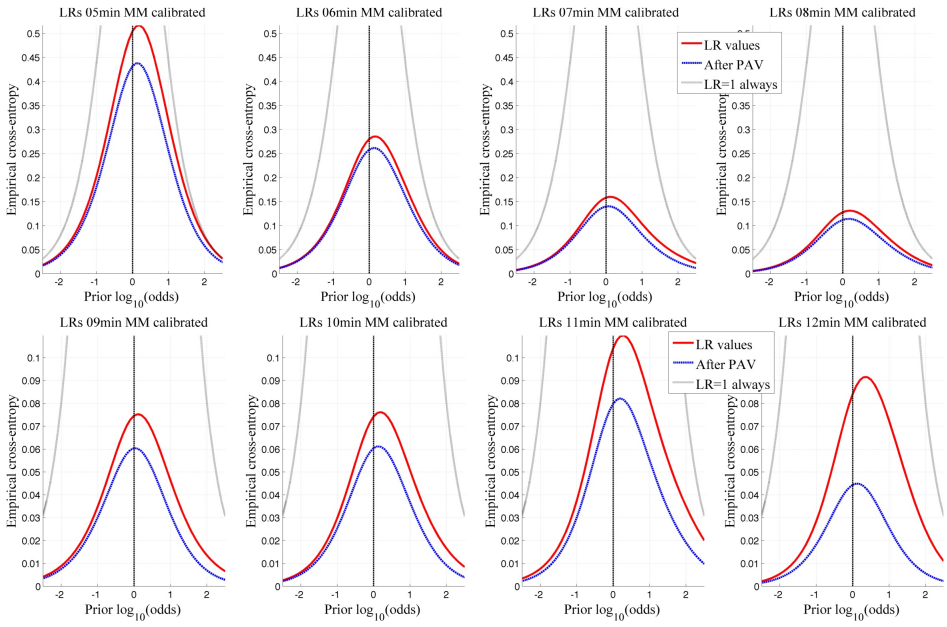


Figure 7 - ECE plots for LRs generated for forensic marks with 5 to12-minutiae configurations. Note the different scaling of the *y*-axis in the upper and lower row of plots.

The Tippett plots in Figure 8 also show coherence of the method with the increasing distance between the curves based on LRs supporting either proposition as the number of minutiae increases. In an ideal system the rates of misleading evidence would be equal to zero, and both curves in the Tippett plots would be maximally separated. The coherence is observed in the Tippett plots when with the increasing number of minutiae there is a decreasing trend in the rates of misleading evidence and an increase in the separation of the curves. The rate of misleading evidence in favour of H_d (RMED [24, 26]) decreases from 31% for 5-minutiae configurations to 3.5% for 12-minutiae configurations, while the rate of misleading evidence in favour of H_p (RMEP [24, 26]) decreases from 1.2% for 5-minutiae configurations to 0.06% for 12-minutiae configurations.

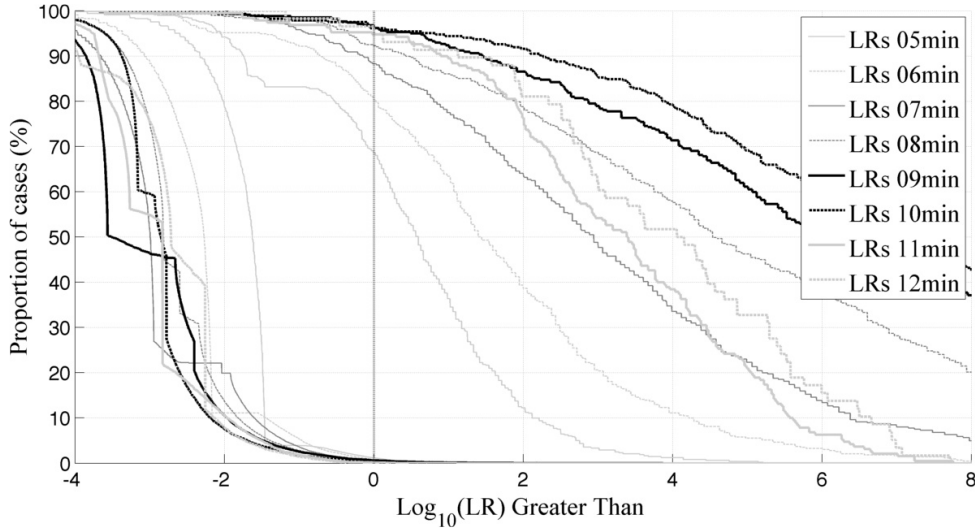


Figure 8 - Tippett plots for LRs generated for forensic marks with 5 to12-minutiae configurations.

The DET curves in Figure 9 capture the discrimination in a lot more detail, complementing the Tippett plots. Coherent behaviour of the LR method used can be observed in the decreasing values of the EER for an increasing number of minutiae. The best performance in terms of EER was achieved for the 9-minutiae configuration dataset (EER = 1.6%). The worst performance of the LR method was observed for the 5-minutiae configuration dataset (EER = 15.7%). Table 4 lists the EER values and apart from the overall decreasing trend shows increases for 10 and 11 minutiae. Not too much meaning can be attached to this because of the overlap and irregular behaviour of the DET curves for the highest number of minutiae.

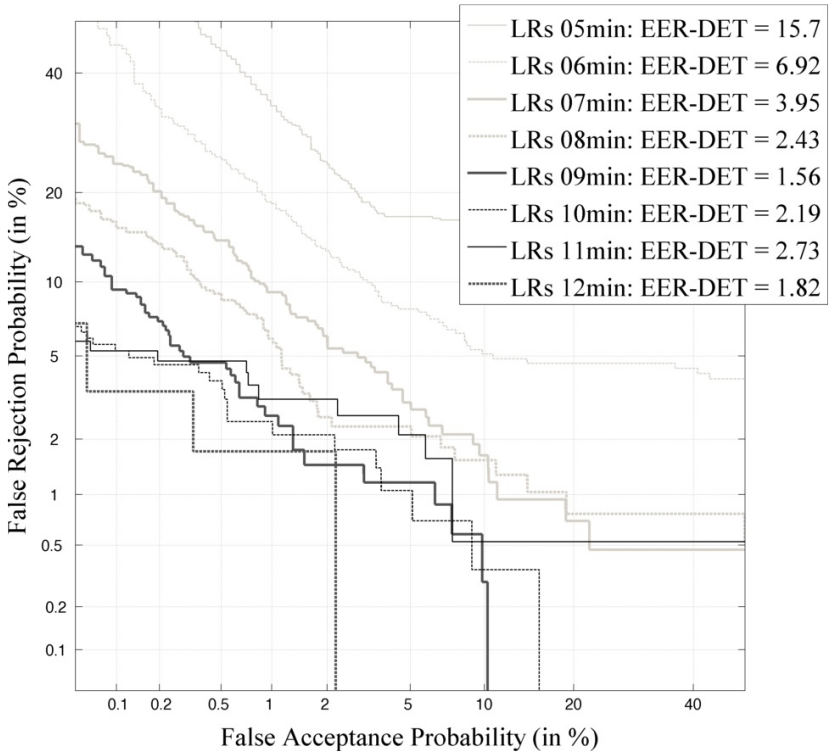


Figure 9 - DET plots for LRs generated for forensic marks with 5 to12-minutiae configurations.

8. Discussion & Conclusions

The purpose of this article is to introduce coherence as a secondary performance characteristic for LR methods developed for forensic evaluation, and to demonstrate its use with an experimental example. In Section 2 we have split various performance characteristics into primary and secondary ones with examples of factors influencing the primary performance characteristics. We then focused on one performance characteristic in particular – the coherence – by giving an experimental example from the area of forensic fingerprint examination. Coherence has been defined as the property of a given method to perform better when the quality or quantity of information increases, which in our experimental example has been simulated by varying the number of minutiae present in fingermarks from 5 to 12.

The performance of the LR method was evaluated using different performance measures (Rates of Misleading Evidence, C_{llr} and EER) and their corresponding graphical representations: Tippett, ECE, and DET plots. The LR method used showed coherent behaviour: performance increased

with the number of minutiae increasing from 5 to 10. It also showed somewhat incoherent behaviour and a small decrease in performance when moving from 10 to 11 minutiae.

This incoherent feature of the comparison algorithm's performance is believed to be caused by a switch of the method it uses when more than 10 minutiae are present. The experimental example therefore reveals the importance of coherence in order to detect points of improvement in computer-assisted LR methods.

Acknowledgements

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) at the Netherlands Forensic Institute, and in collaboration with the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid and the National Police Services Agency of the Netherlands.

References

[1] – G.S. Morrison, *Measuring the validity and reliability of forensic likelihood-ratio systems*, Science & Justice 2011, 51, pp. 91-98
[2] – I.W. Evett, *Towards a Uniform Framework for Reporting opinions in Forensic Science Casework*, Science & Justice 1998, 38, pp.198-202
[3] – D. V. Lindley, *A problem in forensic science*, Biometrika 1977, 64, 2, pp. 207-13
[4] – C. Neumann, I. Evett et al., *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. 2011, 207(1-3), pp. 101-5
[5] – Amanda B. Hepler et al., *Score-based likelihood ratios for handwriting evidence*, Forensic Sci. Int. 2012; 219 (1-3): 129-40
[6] – D. Meuwly, *Forensic Individualization from Biometric Data*, Science & Justice 2006, 46, pp. 205-213
[7] – N. Egli et al, *Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability*, Forensic Sci. Int. 2007, 167, pp. 189-195
[8] – J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro and J. Ortega-Garcia, *Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems*, Forensic Sci. Int. 2005, 155 (2-3), pp. 126-140
[9] – G. Zadora, A. Martyna, D. Ramos, C. Aitken. "Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data". John Wiley and Sons, 2014
[10] – J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, n. 7, pp. 2104-2115, September 2007.
[11] – European Network of Forensic Science Institutes, *Annual Report 2012*: http://www.enfsi.eu/news/enfsi-annual-report-2012
[12] – R. Gill, *FSS Report on Study on Obstacles to Cooperation and Information-sharing among Forensic Science Laboratories and other Relevant Bodies of Different Member States and between these and Counterparts in Third Countries*, 2008 http://www.enfsi.eu/sites/default/files/documents/report_project_terrorism_0.pdf
[13] – ILAC-G19:2002, *Guidelines for Forensic Science Laboratories*
[14] – RvA-T015, 2010, *Explanation of NRN-RN ISO/IEC 17025:2005*
[15] – ISO/IEC 17025:2005, *General requirements for the competence of testing and calibration laboratories*
[16] – D. Ramos, J. Gonzalez-Rodriguez, *Reliable support: Measuring calibration of likelihood ratios,* Forensic Sci. Int., Vol. 230, pp. 156-169, May 2013
[17] – D. Ramos, J. Gonzales-Rodriguez, G. Zadora, C. Aitken, *Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods*, J. Forensic Sci Vol. 58, n. 6, pp. 1503-1518, November 2013
[18] – N. Brümmer, J. du Preez, *Application independent evaluation of speaker detection*, Computer Speech Lang 2006, 20(2-3):230-75
[19] – I.J. Good, *Weight of Evidence: A Brief Survey*, Bayesian Statistics 2, pp. 249-270, 1985
[20] – D.A. van Leeuwen, N. Brümmer, *The distribution of calibrated likelihood-ratios in speaker recognition*, in proceedings Interspeech 2013
[21] – T. Cover and J. Thomas, Elements of Information Theory 2nd Ed., John Wiley and Sons, 2006.
[22] – J. Quiñonero-Candela et al., *Dataset Shift in Machine Learning Shift in Machine Learning*. The MIT Press 2009 The MIT Press, 2009.

[23] – C. M. Rodriguez et al., *Introducing a semi-automated method to simulate a large number of forensic fingermarks for research on fingerprint identification*, J. Forensic Sci 2012, Mar;57(2):334-42.
[24] – C. Nemuann et al., *Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Three Minutiae*, J Forensic Sci, 2006 Nov;51(6):1255-66
[25] – A. Martin et al., *The DET Curve in Assessment of Detection Task Performance*, Proc. EuroSpeech (1997) p. 1895–1898
[26] – D. Meuwly, *Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique*, PhD thesis, 2001

Chapter 6

Assignment of the Evidential Value of a Fingermark General Pattern (GP) using a Bayesian Network (BN)

BIOSIG 2013 : IEEE International Conference of the Biometrics and Special Interest Group

Rudolf Haraksim
Didier Meuwly
Gina Doekhie
Peter Vergeer
Marjan Sjerps

1. Abstract

When visible on a fingermark, the general pattern maintains its importance in the fingerprint examination procedure, since the difference between the general pattern of a fingermark and a fingerprint is sufficient for exclusion. In the current work, the importance of the general pattern is extended by evaluating the strength of evidence of a match given corresponding general pattern. In current practice (due to the lack of statistical support for the general pattern evidence) the fingerprint examiners assign personal probabilities to the general pattern evidence based on their knowledge and experience, while in this work the probabilities are calculated using a Bayesian Network, which is fed by empirical data.

2. Introduction

In this article, we aim to assign a value to the correspondence of the general patterns (GP) in terms of descriptive and inferential statistics. We have developed two Bayesian Networks (BN) – one at the level of finger and one at the level of person – to assist the fingerprint examiners in statistical quantification of probabilities they assign to the general pattern evidence. The main motivation for using BNs is their ability to model the dependencies between different types of evidence in a logically correct framework.

When a fingerprint examiner compares a fingermark retrieved from a crime-scene to a reference fingerprint of a suspected person, (s)he exploits all the available information to assign its evidential value: properties of the ridge flow (level 1), of the minutiae (level 2) and of the ridges themselves (level 3). Recently tools producing Likelihood Ratios (LR) have been developed, allowing the fingerprint examiners to quantify the evidential value of spatial configurations of minutiae [Ne11, ECM07, AJR13, NCJ12, FSS07]. According to [Ne11], the evidential value assigned to the spatial configuration of the minutiae present in a fingermark can be expressed using a likelihood ratio (LR) and a set of propositions at the level of the finger[1]:

Hp: the fingermark was left by a specific finger
Hd: the fingermark was left by an unknown finger

or at the level of the person:

Hp: the fingermark was made by the person who made the set of fingerprints
Hd: the fingermark was made by some unknown person

In absence of realistic data, the numerator of the LR has been reduced by a factor of 10 in [Ne11] when the propositions are considered at the level of the person, to account for the uncertainty in relation to which of the ten fingers of a donor the fingermark originates. The aim of this article is to complement these approaches, using real forensic fingermark and fingerprint data as well as a BN to account for the probability from which of the 10 fingers of a donor the fingermarks retrieved from crime-scene originated and to quantify the evidential value of the shape of the ridge flow classified as a GP.

In the following sections of this article we will provide firstly an insight in the datasets used for constructing the networks, and secondly present each of

[1] The factfinders phrase their questions at the level of the person, which is then investigated at the level of the finger.

the BNs proposed paired with a case example. Finally we will assign the evidential value in form of likelihood ratios. Such likelihood ratios can be combined with the evidential value assigned to other corresponding features of the fingermark and fingerprint, for example the minutiae configuration.

3. Data used and descriptive statistics

3.1 Data labeling

By convention, the fingers are numbered from 1 to 10, starting from the right thumb (labelled finger 1) and ending to the left little finger (labelled finger 10). Numerous systems exist to assign GP to the shape of the ridge flow. In this work, the data are labelled according to the GP classification codes of the ANSI/NIST-ITL 1-2000 format [NIST11]: plain arch, tented arch, left loop, right loop and whorl. A 6th class labelled "unknown" merges the ANSI-NIST codes "unable to print" and "unable to classify".

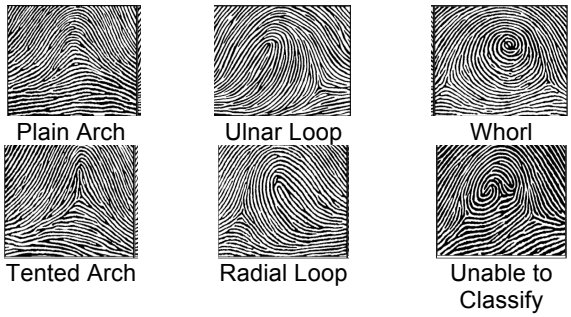


Figure 1 – General pattern classification

In 1975, A. J. Brooks conducted a study on the fingermarks identified in Chicago during the period from 1969 to 1973, to determine from which of the 10 fingers of a donor the fingermarks retrieved from crime-scenes originated [Br75]. Since this time, too little attention has been paid to the study of datasets of identified fingermarks [RJM12]. More attention has been dedicated recently to the study of the distribution of the GP on the 10 fingers [Sw05, NBMM09, GARG08]. These studies use various GP classification codes and only the results presented in [GARG08] classify the shape of the ridge flows, with codes similar enough to the ANSI-NIST codes to be compared to the results of the present study.

Due to their age, rarity, diversity or origin, we have replicated these studies independently in 2012 in our country using the most recent operational data, ensuring the applicability of the results in this country and at the present time.

3.2 Identified fingermarks – finger number

A total of 11555 identified fingermarks [2] from the years 2010 (4032 identifications) and 2011 (7523 identifications) was used to determine from which of the 10 fingers of a donor the fingermarks retrieved from crime-scenes originate. These data reflect the operational activity as processed by the national police force in the field of fingermark examination in these two years. For each identified fingermark, the finger number, the GP and the gender of the donor of the (corresponding) reference fingerprint general were provided. The results summarizing the distribution of fingers identified in the police investigations are presented in the table 1.

Table 1 – Proportions of identified fingermarks (Brooks vs. Police identified fingermarks)

Finger Number	Brooks Identified Fingermarks	Police Identified Fingermarks
1	15.06	15.59
2	11.69	16.97
3	13.57	10.64
4	10	6.9
5	2.22	2.1
6	14.05	15.26
7	10.17	9.62
8	13.2	11.67
9	7.81	8.07
10	2.22	3.18

The proportions of identified fingermarks will be integrated into the "FingerNumber" node of both BNs (described in the following section). The results have also been compared to the results of the Brooks study.

Despite the 35 years separating the two studies, the diversity of the populations studied and the fact that the quantity of data of the present study supersedes almost 4 times the dataset of Brooks, we observe similar results. The descriptive statistics presented indicate that differences smaller than 2% are observed between the two datasets. Our interpretation is that inferences made using these results are valid on the long term and are not sensitive to the diversity of the populations. We also observe fact that both hands are similarly represented in the criminal activity (47% left hand vs. 53% right hand), despite the fact that the majority of the human population

[2] We are aware that no ground truth exists for a decision regarding identification of a crime-scene fingermark and a corresponding reference fingerprint of a suspect. Due to the fact that 12-minutiae numerical standard is addopted in many countries (including ours) we consider the identifications carried out by fingerprint examiners based on this standard as an acceptable ground truth by proxy.

is right-handed.

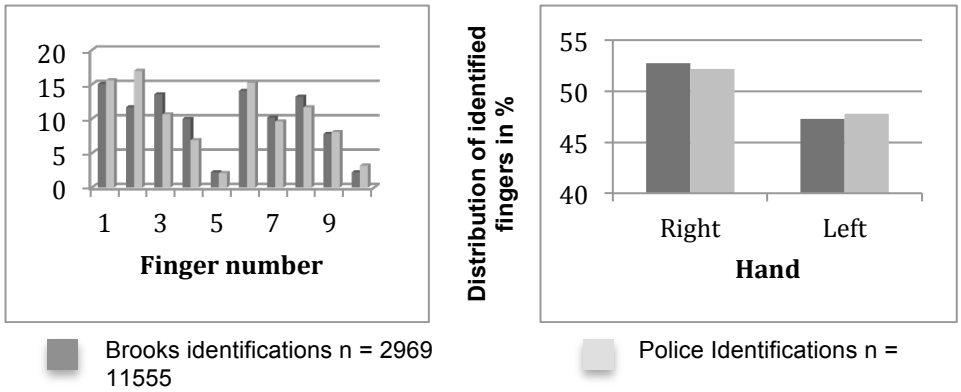


Figure 2 – Results of the comparison of the results of the present study with the results of the Brooks study [Br75]

3.3 Reference fingerprints – General pattern

The dataset consists of inked, digitized and encoded 10-print cards of the police fingerprint database. The GP of these prints has been assigned manually by fingerprint examiners. For each print, additional information regarding the finger number and the gender of the donor is available. 10-print cards from 312.484 individuals have been randomly selected from the original dataset to study the distribution of the GP over the 10 fingers. 72.5% of the data originates from male donors and 26.8% from female donors. For 0.7% of the data the gender was unknown.

Table 2 – GP distribution (%) on different fingers of the right hand (females and males)

Finger No.	**1**		**2**		**3**		**4**		**5**	
GP \ Gender	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
Plain arch	2.2	4.1	5.3	7.8	3.7	5	1	1.7	0.6	1.3
Tented arch	1.1	1.5	11.8	10.3	7.3	6.6	3	3.1	3	3.6
Right Loop	47.3	54.9	29.6	36.6	65.6	72.5	43.9	53.1	77.6	83.1
Left Loop	0.4	0.4	16.4	12	1.4	0.8	1	0.9	0.2	0.2
Whorl	48.7	39	36.3	32.9	21.5	14.9	50.8	41.1	18	11.4
Unknown	0.3	0.2	0.7	0.4	0.5	0.3	0.4	0.3	0.5	0.4

The information related to the GP, to the finger number, and to the gender have been exploited in combination in order to study the distribution of the GP on the 10 fingers. The results for the female and male donors are presented in the Tables 2 and 3. They will be integrated into the node variable “GeneralPattern” of the two BNs described in the next section.

Table 3 – GP distribution (%) on different fingers of the left hand (females and males)

Finger No.	6		7		8		9		10	
GP \ Gender	M	F	M	F	M	F	M	F	M	F
Plain arch	3.9	6.5	5.3	8.1	4.6	7.4	1.3	2.3	0.8	1.9
Tented arch	1.7	2	12	11.5	7.7	8.3	3.1	3.7	3.1	4.1
Right Loop	0.5	0.9	14.3	15.8	1.1	1.5	0.4	0.8	0.1	0.2
Left Loop	55.3	55	34.1	32.9	64.8	64.2	55.2	57.5	82	84.3
Whorl	38.3	35.4	33.8	31.3	21.4	18.3	39.6	35.4	13.5	11.2
Unknown	0.3	0.2	0.6	0.4	0.5	0.3	0.4	0.3	0.5	0.5

In the Figure 3 the results are compared to the results of the Gutierrez [GARG08] study. As the entries in the Tables 2 and 3 indicate minor differences of the order of 2% between the relative frequencies of GPs for females and males. The prints labelled as *Plain* and *Tented Arch* of our study have been merged into one class labelled *Arch* to fit the classification codes used in [GARG08].

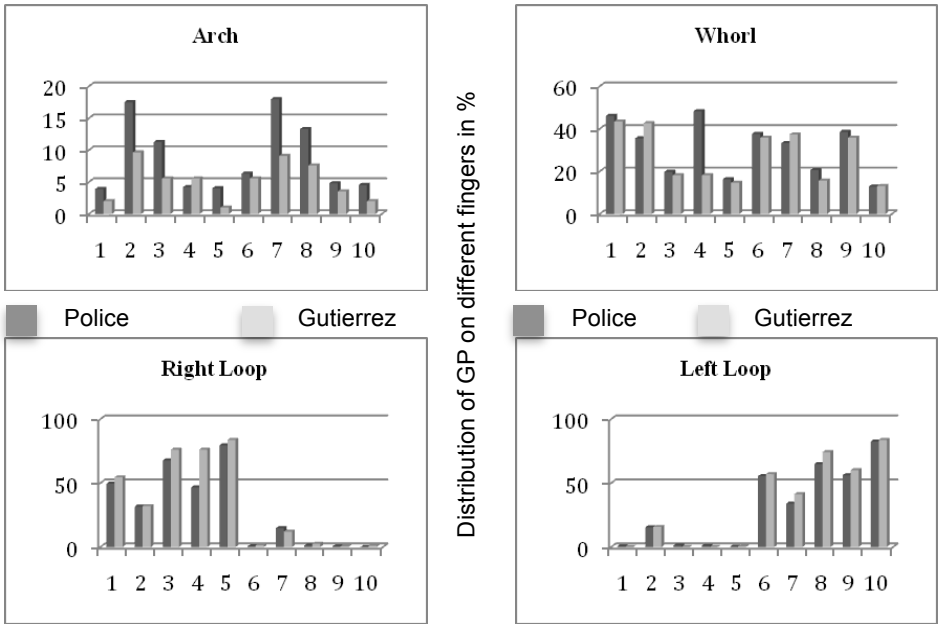


Figure 3 – Distribution of the GPs on the 10 fingers (Police vs Gutierrez datasets)

Despite the fact that 312.484 individuals were used in our study and only 200 individuals in the study of Gutierrez, we observe similar proportions of right and left loops distributed over different fingers. However, proportions of arches and whorls appears to be quite different for some fingers. Unfortunately, the difference in encoding standards used by Nithin [Sw05] and Swofford [NBMM09] prevent a direct comparison.

4. Inferential statistics using Bayesian Networks

From practice and experience the fingerprint examiners acquire an implicit knowledge of the distribution of the GP over the 10 fingers and of the relative contribution of the 10 fingers to the fingermarks retrieved from crime scenes. They make use of this knowledge when assigning evidential value to the correspondences and differences observed between a fingermark and a fingerprint.

Two BNs integrating the descriptive statistics described in Section 2 have been built to quantify the evidential value resulting from the combination of the GP and the relative contribution of the 10 fingers. The utility of these networks is to assist the fingerprint examiners to refine the numerator of the LR when they consider propositions at the finger and person level. In other words, the use of BNs allows the examiners to support their personal probabilities with statistical data. Concretely, we propose two BNs to assist the examiner, the first one for the finger level (3.1) and the second one for the person level (3.2). The BN models are „built for purpose“ and their implicit validation and justification is subject to further research.

4.1 Finger level (Distinctivness of the GP)

At the finger level the BN informs about the rarity of a GP observed on each finger number of a random person (based on the population). The node “Finger Number” contains the distribution from which of the 10 fingers of a random donor the fingermark originated; the node “Hand” encapsulates the proportion of right / left handed in the identified fingermarks; the node general “General Pattern” contains the distribution of the GP over the 10 fingers and the node “Gender” contains the proportions of male / female / unknown donors of identified fingermarks. We express the dependency of the GP node on the finger number and the gender[3] by *P(GP|FN,G)*.

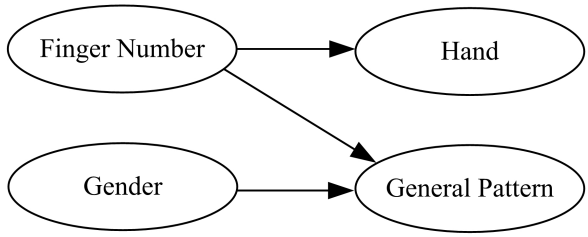
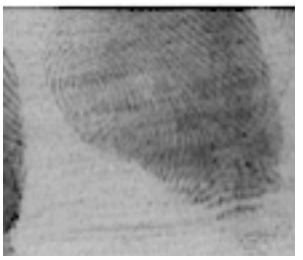


Figure 4 – the finger level BN

[3] The gender dependency has been made explicit, despite the minor differences in the GP distribution between the male and female population (see Identified Fingermarks section of this article).

Case example



A fingermark containing a GP labelled as a whorl is recovered from the surface of a ceramic mug. The BN calculates the probability (posterior odds) for this whorl to have been left by each finger of a randomly selected donor. In this case the BN indicates that this mark has the highest probability to have been left by the finger number 1 and the lowest probability to have been left by the finger number 5. This result is useful for 2 purposes. Firstly, it allows for searching the database per finger number, starting from the most common finger. Secondly, the posterior odds indicate that the evidential value expected strongly depends on which finger of a donor it can be paired to.

The propositions to be tested are: the mark originates from donor's finger 1 vs. the mark originates from any other finger (2-10) of the same donor. The posterior odds provided by the BN allow to calculate the posterior odds ratio P (F1|GP) / P ($F_{2\sim10}$|GP). The uninformed prior odds of 1/10 in absence of data are updated using the descriptive statistics of the Table 4. The evidential value for a whorl observed on a fingermark paired to the finger number 1 (vs. on any other finger) as calculated in the table 4 is 1.46. In other words, it is 1.46 times more likely to observe a whorl if it originates from the finger number 1 than if it originates from any other finger number of a donor randomly selected. The calculation for the highest and lowest evidential value has been added for illustration purposes.

Table 4 – LR values for the most rare, case example and most common GP

Evidence	Evidential Value	Prior Odds (in %)	Posterior Odds (in %)	LR
Whorl on Finger 1	Example given	15.59/84.51	21.26/78.74	1.46
Right loop on Finger 10	Highest	3.18/96.82	0.02/99.98	0.000609
Right loop on Finger 5	Lowest	2.10/97.90	6.22/93.78	3.09

$$\frac{P(F_1 \mid GP)}{P(F_{2-10} \mid GP)} = LR \times \frac{P(F_1)}{P(F_{2-10})}$$

$$\frac{21.26}{78.74} = LR \times \frac{15.89}{84.51}$$

Equation 1 – LR calculation from the prior and posterior odds

4.2 Person level

A few extra nodes need to be introduced in the previous BN to address the propositions at the person level. The node “Suspect Print GPs” contains the classification code of the 10 GPs for the donor of the 10-print card paired to the fingermark (GP code). The node “Source of the Mark” contains the pair of alternative propositions to be tested: the mark originates from the donor of the 10-print card vs. the mark originates from a donor randomly selected. For a practical reason the prior odds ratio for these 2 propositions is set to ½ (prior odds = 1). The choice for the prior of 1 is a conscious choice to force the posterior odds to be equal to LR. We do not mean to imply that equal prior odds are a good choice for any other purpose than extracting the LR from the BN.

The probabilities of the GP of the mark (“Mark General Pattern” node) directly depend on the finger number, the gender and GP code of the donor of the 10-print card. In the case of correspondence between the GP code and finger number of the fingermark and fingerprint of the donor of the 10-print card, the numerator of the likelihood ratio is equal to 1; it is equal to 0 in the case of a difference. For the denominator of the LR, the probability of correspondence between the GP code and finger number of the fingermark and the fingerprint of another person is determined by the data of the Tables 1, 2 and 3.

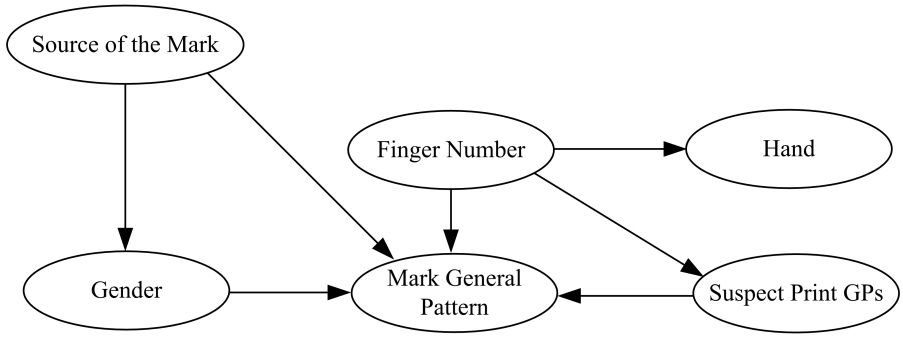


Figure 5 – the person level BN

Case example

At the person level we use the same fingermark as in the previous example: a whorl found on a ceramic mug on the crime-scene. Based on eyewitness testimony the police arrest a person, from which a 10-print card is produced. The Table 5 summarizes the GP classification codes of this donor.

Table 5 – Description of the general pattern code the donor of the 10-print card (A – Arch, W – Whorl, R – Right loop, L – Left loop)

Finger Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
General Pattern	R	A	R	W	R	L	L	W	L	L

In the BN shown in Figure 5, the GP of the fingermark is given in the node "Mark General Pattern" and the GP code of the donor is given in the node "Suspect Print GPs". The probability that the fingermark originates from the donor of the GP code given the finger number and the correspondence of the GP divided by the probability that the fingermark originates from another person given the same evidence: *P(Hp|GP) / P(Hd|GP)* is calculated in the node "Source of the Mark". This posterior odds ratio is equivalent to the LR , since the prior odds for the 2 propositions have been set to 0.5 (odds 1/1).

The BN at the level of the person uses the general pattern code of the donor (distribution of the GPs over all 10 fingers) together with the whorl found on the crime-scene information to calculate the LR at the person level directly. This information is quantified for the two sets of propositions in the node "Source of the Mark": *P(Hp|GP) = 30.73* and *P(Hd|GP) = 69.27* (as shown in Figure 5). The value calculated as presented in Table 6 is 0.44 *(30.73/69.27).* It means that it is slightly less probable to observe a whorl if it originates from the donor of the 10-print card than if originates from a donor randomly selected. For illustration purposes, the LR has also been calculated for the other available GPs: arch, right loop and left loop.

Table 6 – LR values for the most rare, case example and most common GP

Data from the BN	P(GP|Hp) in %	P(GP|Hd) in %	**LR**
Arch (given suspects GP)	80.73	19.27	4.19
Whorl given suspects GP (case example)	30.73	69.27	0.44
Right loop (given suspects GP)	51.43	48.57	1.06
Left Loop (given suspects GP)	57.64	42.36	1.36

Tables 2 and 3 show that the arch is the most rare classification code for a GP. Similarly, as in the previous example, we can attempt to evaluate the smallest and largest LR. Unlike in previous example however, we now operate at the level of the person, hence the LR depends not only on the GP found on the crime scene but also on distribution of the GPs in the population. It is directly dependent on the general pattern of the suspect as well. Given the general patterns of the suspect in this case the smallest LR corresponds to the whorl found on the crime-scene and the biggest LR corresponds to the arch. These LR values remain modest, but the strategy

consists in measuring and combining the evidential value of each characteristic available in the fingermark that can be paired with a reference fingerprint. Clearly, the LRs obtained for the first level information are calculated with the intention to combine them with the LRs calculated for the second level of information, based on the spatial arrangement of the minutiae.

5. Conclusions

When no prior information is available it is reasonable to assume the refinement of 1/10 when moving from the finger to person level as proposed by Neumann et al [Ne11].

The two BNs developed in this article combine the statistical information regarding the GP distribution over different fingers contained within the fingermarks (police identifications) and fingerprints (police database). The main motivation for using the BNs is their ability to model the dependencies between different types of evidence. They also provide a practical solution when quantifying the rarity of the GP found on the crime-scene fingermark and a finger of a random donor (level of the finger) or when quantifying the weight of the GP evidence found on the crime-scene fingermark and GP code of a donor (level of the person).

The choice between the two BNs proposed depends on the needs of the fingerprint examiner and/or operational conditions of the systems used to assign the evidential value of the second level details. Our aim in both cases was to quantify the evidential value contained in the first level detail fingermark/fingerprint comparison in meaningful LR values, which could be further combined with LR values obtained from the second level detail fingerprint evidence evaluation process or any other case related evidence.

6. Future work

Future work will include validation of the BN models developed in terms of enhanced application scenarios, sensitivity analysis and further improvement of the BN to support any finger combination. Also, further investigation is needed when utilizing the tools developed in combination with other fingerprint/case related evidence.

Acknowledgements

This research was motivated by the work carried out at the University of Lausanne presented by prof. Christophe Champod at the IPES 2010 in Florida entitled *"The Use of Probabilistic Networks in the Area of Fingerprints"*. The work was carried out in cooperation with the National Police Services Agency of the Netherlands (KLPD), the University of Amsterdam, department of interdisciplinary research and statistics (WISK) of the Netherlands Forensic Institute, and the European Union MC-ITN FP7 BBfor2 project (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803). More detailed summary of the work presented can be found in [Do13].

References

[AJR13] Alberink I., Jongh A., Rodriguez C.: Fingermark evidence evaluation based on AFIS matching scores: the effect of different types of conditioning on Likelihood Ratios. Journal of Forensic Sciences, 2013, In Press

[Br75] Brooks A.J.: Frequency of Distribution of Crime Scene Latent Prints, Journal of Police Science and Administration, pp. 292-293, 1975

[Do12] Doekhie G.: A Bayesian Network for Assigning Probabilities on which Finger left a Mark, MSc Thesis, University of Amsterdam, 2012

[ECM07] Egli N., Champod C., Margot P.: Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - Modelling within finger variability. Forensic Science International 167(2-3), pp. 189-195, 2007

[FSS07] Fang G., Srihari S.N., Srinivasan H.: Generative fingerprint individuality using ridge types, In Proc. 3rd International Symposium on Information Assurance and Security, pp. 423-428, Aug. 2007

[GARG08] Gutiérrez-Redomero E., Alonso C., Romero E., and Galera V.: Variability of fingerprint ridge density ina sample of Spanish Caucasians and its application to sex determination. Forensic science internatonal,180(1):17—22, 2008

[NBMM09] Nithin M.D., Balaraj B.M., Manjunatha B., and Mestri S.C.: Study of fingerprint classification and their gender distribution among south Indian population. Journal of Forensic and Legal Medicine, 16(8):460 – 463, 2009

[NCJ12] Nagar A., Choi H.S., Jain A.K.: Eveidential Value of Automated Latent Fingerprint Comparison: An Empirical Approach, IEEE Transactions onf Information Forensics and Security, Vol. 7, no. 6, pp. 1752 – 1765, Dec. 2012

[Ne11] Neumann C. et.al.: Quantitative assessment of evidential weight for fingerprint comparison i. generalization to the comparison of a mark with a set of ten prints from a suspect, Forensic Science International, 207(1:3):101-105, 2011

[NIST11] American National Standard for Information Systems – Data Format for the Interchange of Fingerprint, Facial& Other Biometric Information, ANSI/NIST-ITL 1-2011

[RJM12] Rodriguez C.M., Jongh A., and Meuwly D.: Introducing a semi-automatic method to simulate large numbers of forensic fingermarks for research on fingerprint identification. Journal of Forensic Sciences, 2012

[Sw05] Swofford H.: Fingerprint patterns: a study on the finger and ethnicity prioritized order of occurrence. Journal of Forensic Identification, 55(4):480, 2005

Chapter 7

Multimodal LR Method for Fingerprint Evidence Evaluation: Validation Report

Rudolf Haraksim

Introduction

Validation of non-standard methods is described in the ISO/IEC 17025 standard in section 5.4.4. *"When it is necessary to use methods not covered by standard methods, these shall be subject to agreement with the customer and shall include a clear specification of the customer's requirements and the purpose of the test and/or calibration. The method developed shall have been validated appropriately before use."* In the section 5.4.4 the ISO standard also lists the information recommended:

a) appropriate identification;
b) scope;
c) description of the type of item to be tested or calibrated;
d) parameters or quantities and ranges to be determined;
e) apparatus and equipment, including technical performance requirements;
f) reference standards and reference materials required;
g) environmental conditions required and any stabilization period needed;
h) description of the procedure, including

- affixing of identification marks, handling, transporting, storing and preparation of items,
- checks to be made before the work is started,
- checks that the equipment is working properly and, where required, calibration and adjustment of the equipment before each use,
- the method of recording the observations and results,
- any safety measures to be observed;

i) criteria and/or requirements for approval/rejection;
j) data to be recorded and method of analysis and presentation;
k) the uncertainty or the procedure for estimating uncertainty.

Prior to starting the validation of a LR method, a validation plan should be drawn by a forensic examiner. It is mandatory for the reader to keep in mind, that the ISO/IEC 17025 standard was predominantly developed for the validation of analytical methods, therefore not all of the recommended information is applicable to the validation of LR methods. Especially the points e), f), g), h), j) and k) will be rather challenging to defend in the interpretation of forensic evidence. In compliance with the remaining recommendations from the ISO/IEC 17025 standard the validation plan should contain (but is not limited to) the following:

- Identification of LR method – point a)
- The intended use – point b)
- The performance characteristics – point d)
- The performance metrics – point d)
- The validation criteria – point i)
- The scope of the validation (Range of application of the LR method) – point b)
- Validation time span (applicable in cases in which the datasets used in the LR method development/validation stage are envisaged to get obsolete)

1. Intended use

This validation report presents the empirical validation example of the ***multimodal LR method***, developed for use in forensic fingerprint evidence evaluation as presented in chapter 4. It follows the validation protocol drafted in chapter 1 of this thesis, addressing the validation requirements specified. Validation requirements will be specified and addressed using the validation criteria, performance metrics, graphical representations, datasets used, experiment description and validation decision as presented in chapter 1. Where applicable, a reference to the chapter / section in the thesis dedicated to the specific requirement will be provided. The report will be followed by a summary section, in which a set of recommendations and a final validation verdict will be presented.

2. Method Description

The validation criteria of the LR method developed in chapter 4 were set as a comparison with the "baseline". Both, the Kernel Density Estimate (KDE) baseline LR method and the multimodal LR method proposed in chapter 4 use the discriminating scores produced by AFIS comparison algorithm for the evidence Same Source (SS) and for the evidence Different Source (DS).

Two different datasets have been used:

- Simulated dataset (see chapter 4 for more details)
- Forensic dataset (see chapter 5 for more details)

Main motivation for the use of simulated data was the fact that the fingermarks were available in large quantities and the source of origin was a-priori known due to the fact that the simulated fingermarks we re obtained

in controlled conditions. Due to the fact that the quantity of the simulated fingermarks supersedes the quantity of forensic fingermarks by a factor of 100 (there is 100times more simulated data than the real forensic), the simulated data was used in the training stage while the real forensic data was used in the validation stage.

Baseline KDE LR method featured several undesirable aspects (see chapter 4 for more details):

- Over-fit on the training dataset
- Not robust to the previously unseen data
- Not robust to the dataset shift
- Produced unconstrained LR values of irrational magnitudes (LR = infinity)

The multimodal LR method presented solutions to the above-mentioned issues while maintaining (and improving on) the performance of the baseline KDE method.

3. Validation Matrix

Validation matrix including all different aspects of the validation report is presented in table 1 below:

Table 1: Aspects of empirical validation

Performance Characteristic	Performance Metric	Graphical Representation	Validation criteria	Experiment	Data	Analytical result	Validation Decision
Accuracy	Cllr, EER	ECE plot DET plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail
Discriminating power	$Cllr^{min}$	ECE^{min} plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail
Calibration	$Cllr^{cal}$	Tippett plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail
Robustness	Cllr, EER LR range	ECE plot DET plot Tippett plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail
Coherence	Cllr, EER	ECE plot DET plot Tippett plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail
Generalization	Cllr, EER	ECE plot DET plot	According to the definition	Description	Data used	+/- [%] compared to the baseline	Pass / fail

It is to be understood, that due to the definition of the likelihood ratio as being the result of a ***probabilistic inference*** and not a ***measurement***, ***no quantitative ground truth exist for the likelihood ratio because*** of the ***"Bayesian interpretation of probabilities as a degree of belief"*** [Lindley 1976]. ***Therefore it is not possible to establish unique relation between a pair of samples and a likelihood ratio value.***

For this reason we omit the precision from the performance characteristics all together and re-instantiate the term accuracy using a new definition (see chapter 1 for more details).

4. Performance characteristics and the metrics associated

The performance characteristics have been structured in the primary and secondary ones, presented in table 2. The primary characteristics of the LR method under evaluation are related directly to performance metrics and focus on desirable properties (e.g. goodness of a set of LR values, in which we are assessing whether a set of LR values is good or bad, adequate or non-adequate, whether it has desirable properties or not). The secondary characteristics describe how the primary metrics behave in different situations, in some cases simulating the typical casework conditions (e.g., degraded quality of samples, varying conditions in training data and evidence, etc.).

Table 2 – Performance characteristics definitions

Performance Characteristic	Performance Metric	Definition
Accuracy	Cllr, EER	Closeness of agreement between a LR computed by a given method and the ground truth status of the proposition in a decision-theoretical framework. The LR is accurate if it helps to lead to a decision that is correct according to the ground truth of the propositions. In case of source level inference, the ground truth relates to the following pair of propositions: • H_p: the pair of samples tested originate from the same source (SS) • H_d: the pair of samples tested originate from different sources (DS) If an experimental set of LR values is to be evaluated, and the corresponding ground-truth labels of each of the LR values are known, then a given LR value is evaluated as more accurate if it supports the true

Performance Characteristic	Performance Metric	Definition
		(known) proposition to a higher degree, and vice-versa
Discriminating power	$Cllr^{min}$	Performance property representing the capability of a given method to distinguish amongst forensic comparisons under each of the propositions involved
Calibration	$Cllr^{cal}$	In probabilistic terms can be defined as the property of a set of LRs. Perfect calibration of a set of LR's means that those LR's can probabilistically be interpreted as the evidential value of the comparison result for either proposition. Finding an LR=x will be x times more probable under H_p than under proposition H_d. Under those conditions the LR is exactly as big or small as is warranted by the data. The strength of evidence of well-calibrated LRs tends to increase with the discrimination power for a given method
Secondary Performance Characteristics	Performance Metric	Definition
Robustness	Cllr, EER	Stability of the performance measure to the variation of a given factors, and as the improvement of the performance measure with the increase of that factor. For instance, method A is more robust to data sparsity than method B if, as the data gets sparser, the performance of method A degrades less than the performance of method B.
Coherence	Cllr, EER	Focuses on the variation of some measurable parameters[1] in the features[2] studied, perceived as influencing the strength of evidence, like the quantity of minutiae in the fingerprint field or the signal to noise ratio in speaker recognition field.
Generalization	Cllr, EER	Property of a given method to maintain its performance under dataset shift. *"A dataset shift occurs when the joint distributions of inputs and outputs differs between the training data (used to build the LR methods) and the testing data(previously unseen)"* [18] used to compute LRs in operational conditions.

[1] Parameter can be seen as a measurable value of the degradation of the extracted features due to forensic conditions (signal to noise ratio, distortion, clarity). LR method can be the robust to these parameters.

[2] Feature is to be understood as a carrier of information extracted from raw data.Coherence is related to the information carried by the features.

5. Validation criteria

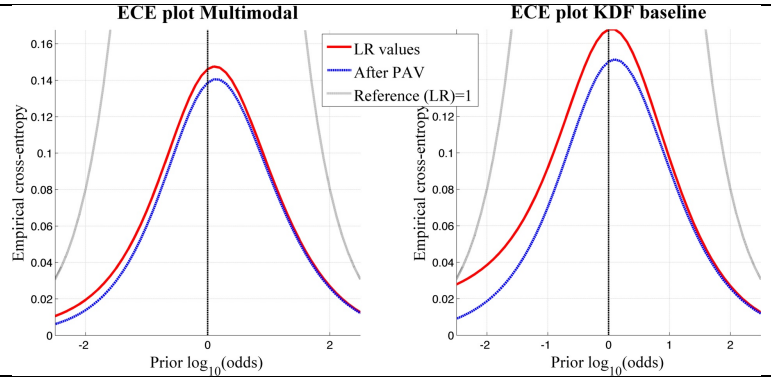
As presented in chapter 1, validation criteria for the LR methods can come either from the comparison with the state-of-the-art or from a comparison with a baseline LR method. The latter case will be used in this report.

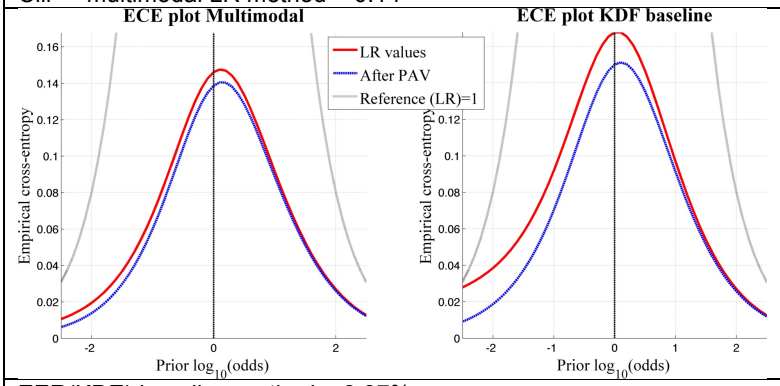
Following validation criteria, based on the performance of the baseline KDE method were extracted, presented in table 3 below:

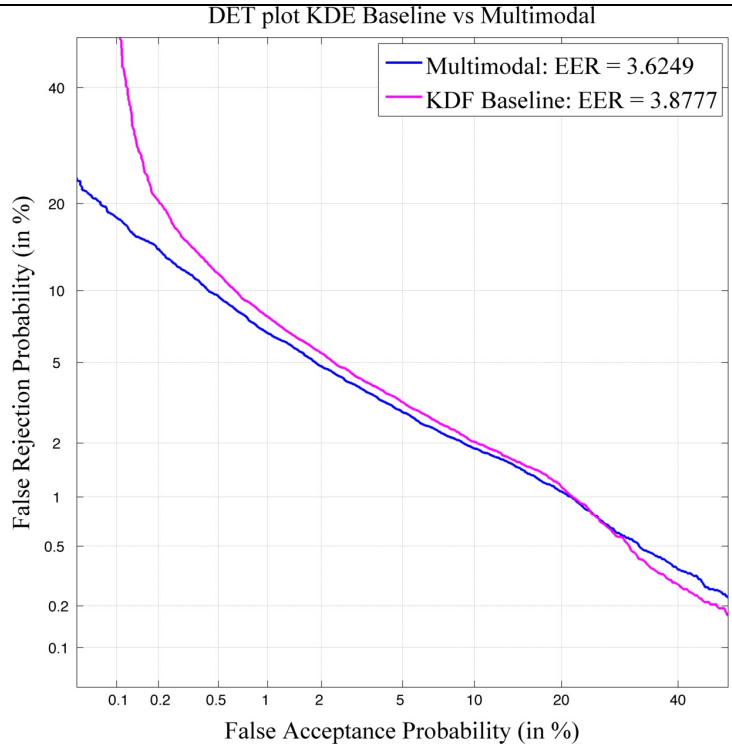
Table 3 – validation criteria

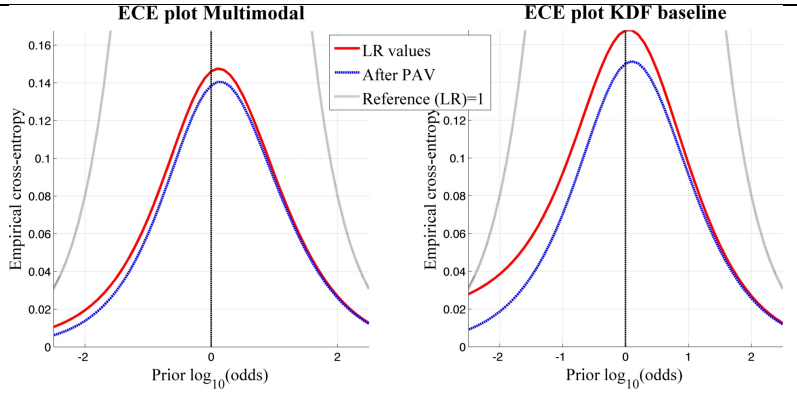
Performance Characteristic	Performance Metric	Validation criteria
Accuracy	Cllr, EER	Cllr <= 0.19
Discriminating power	$Cllr^{min}$	$Cllr^{min}$<= 0.15 EER <= 3.87%
Calibration	$Cllr^{cal}$	$Cllr^{cal}$ =< 0.04
Robustness to the lack of data	Cllr, EER	LR values of meaningful and interpretable magnitudes
Coherence	Cllr, EER	$Cllr_{12min}$< $Cllr_{11min}$ … $Cllr_{6min}$< $Cllr_{5min}$ EER_{12min}< EER_{11min} … EER_{6min}< EER_{5min}
Generalization	Cllr, EER	$Cllr_{Simulated}$ = $Cllr_{Forensic}$ +/- 5% $EER_{Simulated}$ = $EER_{Forenscis}$ +/- 5%

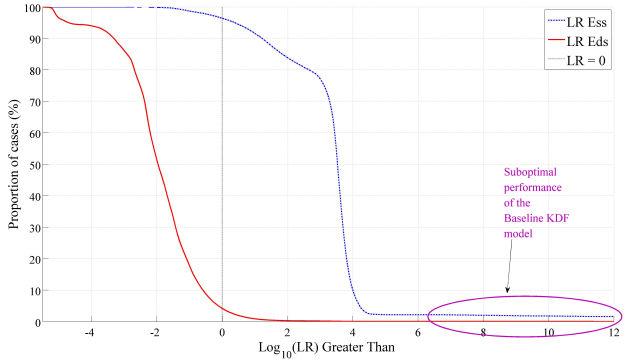
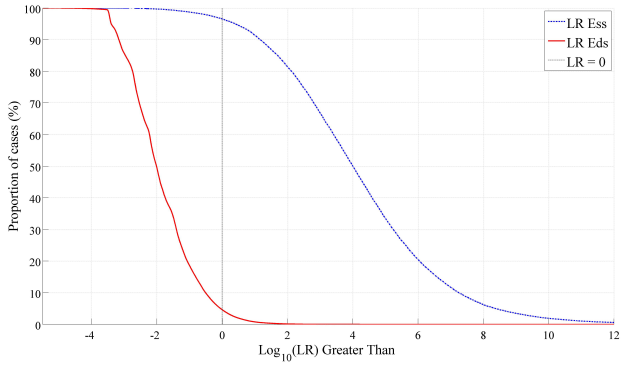
6. The Validation Report

Requirement	**1. Accuracy** Closeness of agreement between an assigned LR and the ground truth status of the proposition in a decision-theoretical framework.
Metric	Cllr
Representation	ECE plot
Validation Criterion	Validation criterion for accuracy is based on the Kernel Density Function (KDE) baseline LR method. Using the simulated dataset in 8 minutiae configuration Cllr(KDE) = 0.19 Better or comparable multimodal LR method Cllr value on the simulated dataset in 8minutae configuration is expected than the KDE baseline. Cllr <= 0.19
Experiment	The Cllr will be measured for both methods – KDE baseline and the multimodal LR – on the simulated dataset
Data	Simulated dataset: fingermarks in 8 minutiae configuration, corresponding fingerprints, reference subset of operational police database
Analytical result(s)	Cllr KDE baseline method = 0.19 Cllr multimodal LR method = 0.15 
Validation Decision	**Pass**

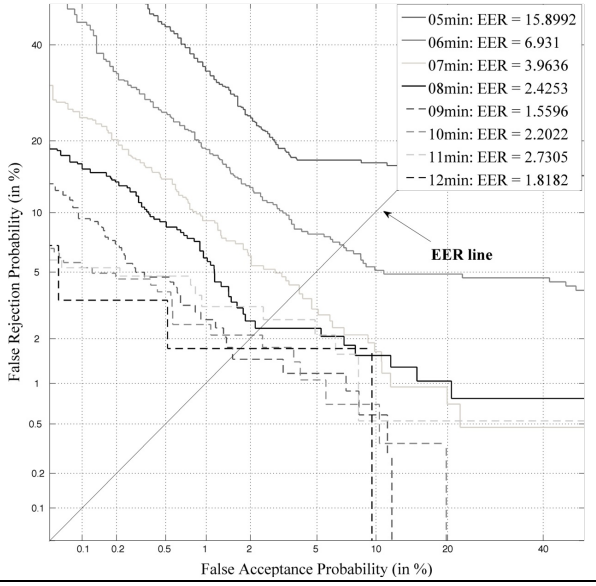
Requirement	**2. Discriminating power** Performance property representing the capability of a given method to distinguish amongst forensic comparisons under each of the propositions involved
Metric	EER, $Cllr^{min}$
Representation	ECE plot, DET plot
Validation Criterion	Validation criterion for accuracy is based on the Kernel Density Function (KDE) baseline LR method. Using the simulated dataset in 8 minutiae configuration $Cllr^{min}$(KDE)= 0.15 and EER(KDE) = 3.87% Better or comparable multimodal LR method $Cllr^{min}$ value on the simulated dataset in 8minutae configuration is expected than the KDE baseline. $Cllr^{min}$<= 0.15 EER <= 3.87%
Experiment	The Cllr will be measured for both methods – KDE baseline and the multimodal LR – on the simulated dataset
Data	Simulated dataset: fingermarks in 8 minutiae configuration, corresponding fingerprints, reference subset of operational police database
Analytical result(s)	$Cllr^{min}$ KDE baseline method = 0.15 $Cllr^{min}$ multimodal LR method = 0.14
	
	EER(KDE) baseline method = 3.87% EER multimodal LR method = 3.62%

	
Decision	**Pass**

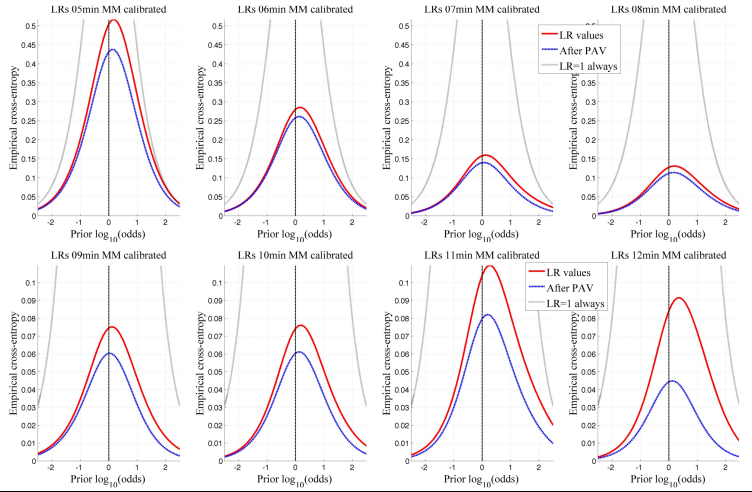
Requirement	**3. Calibration** Defined as the property of a given set of LR values for yielding the same set of LR values when computing the LR trained from the same data (in other words, the LR of the LR is the LR for a given set of LR values)
Metric	$Cllr^{cal}$
Representation	ECE plot
Validation Criterion	Validation criterion for accuracy is based on the Kernel Density Function (KDE) baseline LR method. Using the simulated dataset in 8 minutiae configuration $Cllr^{cal}$(KDE)= 0.04 Better or comparable multimodal LR method $Cllr^{cal}$ value on the simulated dataset in 8minutae configuration is expected than the KDE baseline. $Cllr^{cal}$ =< 0.04
Experiment	The Cllr will be measured for both methods – KDE baseline and the multimodal LR – on the simulated dataset
Data	Simulated dataset: fingermarks in 8 minutiae configuration, corresponding fingerprints, reference subset of operational police database
Analytical result(s)	$Cllr^{cal}$ KDE baseline method = 0.04 $Cllr^{cal}$ multimodal LR method = 0.01 
Decision	**Pass**

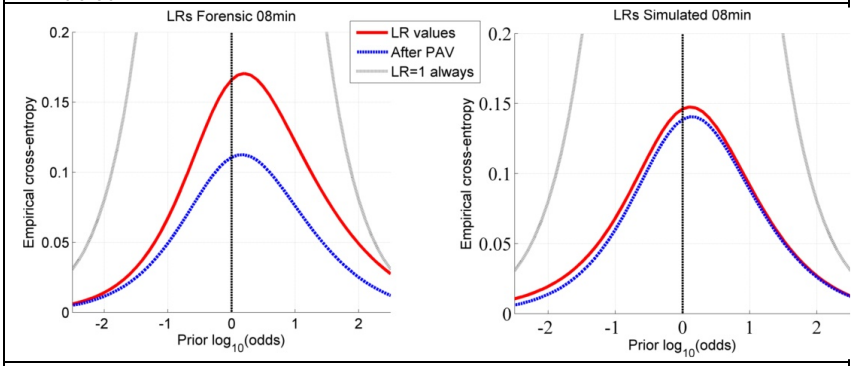
Requirement	**4. Robustness to the lack of data** Data driven LR methods do have a tendency to provide LR values of different magnitudes. Inappropriate (not suitable) LR methods may result in LR values of huge magnitudes (LR = +/- infinity, which given the limited amount of data can not resemble reality). See chapter 4, section 5 for more details.
Metric	N/A
Representation	Tippett plot
Validation Criterion	Reasonably constrained LR values assigned by a LR method.
Experiment	A post-processing
Data	Simulated dataset used for both – KDE baseline and the multimodal LR method
Analytical result(s)	  From the Tippett plots above it is obvious that the KDE baseline method is not a sufficient representation of the simulated dataset. The multimodal LR method developed is more robust to the lack of data than the KDE baseline method. See chapter 4, section 9 for full details.
Decision	**Pass**

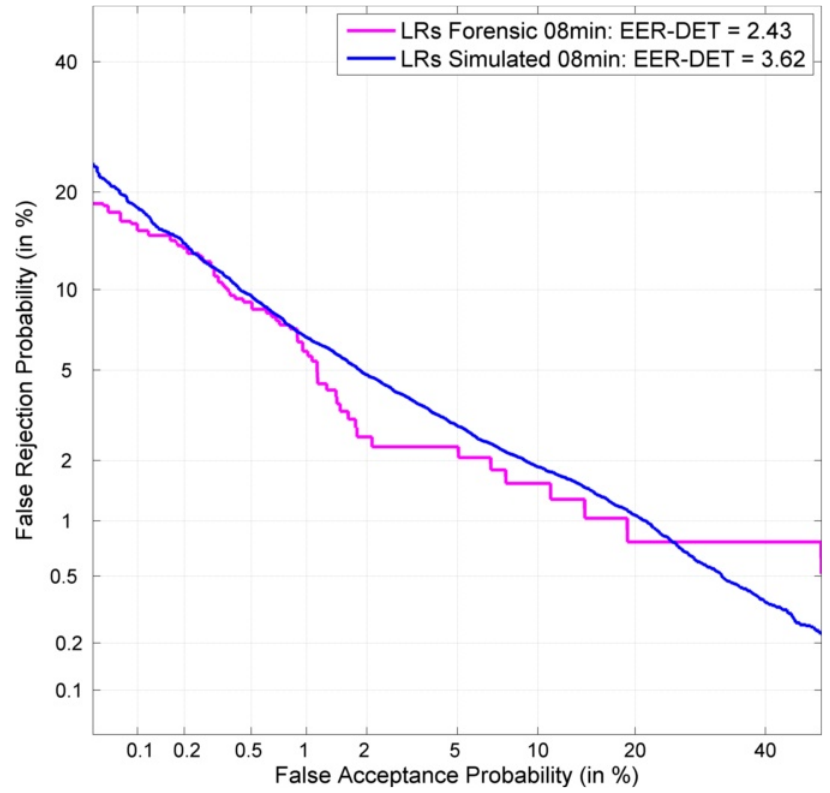
<table>
<tr><td>Requirement</td><td>5. Coherence

Focuses on the variation of some measurable parameters (Cllr, EER) in the features (minutiae) studied, perceived as influencing the strength of evidence, like the quantity of minutiae in the fingerprint field or the signal to noise ratio in speaker recognition field.</td></tr>
<tr><td>Metric</td><td>Cllr, Cllr^{min}, EER</td></tr>
<tr><td>Representation</td><td>ECE plot, DET plot</td></tr>
<tr><td>Validation Criterion</td><td>Observe improvement in the performance metrics with the increasing number of minutiae.</td></tr>
<tr><td>Experiment</td><td>Vary the number of minutiae from 5 to 12 minutiae and observe improvement in Cllr, Cllr^{min} and EER</td></tr>
<tr><td>Data</td><td>Multimodal LR method trained using the simulated dataset
Multimodal LR method validated using the forensic dataset</td></tr>
<tr><td>Analytical result(s)</td><td>
<table>
<tr><th>Configuration</th><th>DET-EER</th><th>Discrimination Calibrated Cllr^{min}</th><th>Accuracy Calibrated Cllr</th></tr>
<tr><td>5 minutiae</td><td>15.9</td><td>0.43</td><td>0.5</td></tr>
<tr><td>6 minutiae</td><td>6.9</td><td>0.26</td><td>0.28</td></tr>
<tr><td>7 minutiae</td><td>3.9</td><td>0.14</td><td>0.16</td></tr>
<tr><td>8 minutiae</td><td>2.4</td><td>0.11</td><td>0.13</td></tr>
<tr><td>9 minutiae</td><td>1.5</td><td>0.063</td><td>0.075</td></tr>
<tr><td>10 minutiae</td><td>2.2</td><td>0.063</td><td>0.074</td></tr>
<tr><td>11 minutiae</td><td>2.7</td><td>0.081</td><td>0.1</td></tr>
<tr><td>12 minutiae</td><td>1.8</td><td>0.057</td><td>0.084</td></tr>
</table>


</td></tr>
</table>

	
Decision	**Pass –** with remark * There are two cores at the minutiae comparison algorithm. One core of the algorithm is used for comparing fingermarks in 5 to 9 minutiae configuration; another core is used for comparing fingermarks in 10+ minutiae configuration. In order to fully demonstrate the coherence effects it would be beneficiary to replace the twin-cored comparison algorithm by a dedicated minutiae comparison algorithm that would work across the whole range of minutiae configurations.

Requirement	**6. Generalization: Robustness to previously unseen data and the dataset shift** Capability of a method to keep its performance under dataset shift, which is here defined as the difference in the conditions between the training or population data (used to train the LR methods) and the data that will be used as evidence in operational conditions.
Metric	Cllr, EER
Representation	ECE plot, DET plot
Validation Criterion	$Cllr_{Forensic}$ not worse than $Cllr_{Simulated}$ + 5% $EER_{Forenscis}$ not worse than $EER_{Simulated}$ + 5%
Experiment	Multimodal LR method trained using the simulated dataset and tested using the previously unseen forensic dataset. An example using fingermarks in 8 minutiae configuration will be used.
Data	Multimodal LR method trained using the simulated dataset Multimodal LR method validated using the forensic dataset
Analytical result(s)	$Cllr_{Simulated}$ = 0.15 $Cllr_{Forensic}$ = 0.17
	
	$EER_{Simulated}$ = 3.62% $EER_{Forensic}$ = 2.43%

	
Decision	*ΔCllr = +2.55% (worse performance on the Forensic dataset)* *ΔEER = -32.87% (better performance on the Forensic dataset)* **Pass**

7. Summary

The multimodal LR method developed for the forensic fingerprint evidence evaluation appears to be satisfying the validation requirements specified in chapter 1. Summary across different performance characteristics is presented in the table 4 below.

Table 4 – Validation decisions across different performance characteristics

Performance Characteristic	Validation Decision
Accuracy	Pass
Discrimination	Pass
Calibration	Pass
Robustness	Pass
Coherence	Pass **with constraint*
Generalization	Pass

Multimodal LR method constraints

Coherent behaviour in terms of improvement of LR values produced for the range of 5-12 can be observed for 5-9 minutiae configurations and again for 10-12 minutiae configurations (being judged independently due to the fact that two different versions of AFIS comparison algorithm is used – one for 5-9 minutiae and another one for 10-12 minutiae).

Multimodal LR method range of application

Besides the information-theoretical aspect, the ECE provides another interesting insight – that is the range of application of the LR method under evaluation. We can safely assume that one of the most desirable properties of a LR method should be to obtain better performance for the whole range of priors than the one of a reference method producing LR = 1 all the time (equivalent to I don't know). Such a reference method has an interesting property – in long term it is perfectly calibrated, it is however as well completely useless for making predictions. Since the accuracy of a LR method in terms of Cllr represents "goodness" of predictions of the LR method under evaluation, a LR method can be deemed "good" if the Cllr values produced by a LR method don't exceed the ones of the reference method.

This gives a rise to the definition of the range of application. The range of priors, in which the LR method under evaluation performs better than the reference method, represents the range of application of a LR method judged using the Cllr metric and shown in the ECE plots. Ranges for different minutiae configurations are presented in table 5 below:

Table 5– Range of application of the multimodal LR method.

Minutiae configuration	$Prior_{log}odds$ uncalibrated	$Prior_{log}odds$ calibrated
5 minutiae	<-2.5,0.5>	<-2.5,2.4>
6 minutiae	<-2.5,1.4>	<-2.5,2.5>
7 minutiae	<-2.5,2.4>	<-2.5-2.5>
8 minutiae	<-2.5,2.7>	<-2.5,2.5>
9 minutiae	<-2.5,2.5>	<-2.5,2.5>
10 minutiae	<-2.5,2.5>	<-2.5,2.5>
11 minutiae	<-2.5,2.8>	<-2.5,2.5>
12 minutiae	<-2.5,2.5>	<-2.5,2.5>

Conclusion

The multimodal LR method has been validated for the 5-12 minutiae configurations. Even though LR values produced for different minutiae configurations show coherent behaviour and the method performance "outside" of these configurations could be predicted, minutiae configurations outside of the validated minutiae configurations should be subject to further experiments and validation.

Validation Decision

The multimodal LR method can be deemed fit for forensic casework within the constraints indicated.

Epilogue

Summary

There are three more or less important, if not equally important steps that lead from the detection of the trace material on a crime scene to the presentation of its strength of evidence in court: the recovery of the trace material (ideally done by a certified crime-scene investigator), the analysis of the trace material (usually done using an accredited method) and the forensic evaluation of the analytical results. The interpretation of the strength of evidence is ideally done using the logically correct approach. Currently it is at best presented as a ratio of probabilities ideally based on personal probabilities of forensic practitioners and expressed on a verbal scale, also due to the fact there are no criteria of validation and acceptance in the practice of the automatic methods for interpretation. A guideline exists for the validation of human-based methods for forensic evaluation though they differ from laboratory to laboratory[1]. Accreditation standards are described for forensic laboratories[2] (EN ISO/IEC 17025), however no guideline or framework exists for the validation of (semi)-automatic evaluation methods. The necessity for a framework and guidelines for validation of semi-automatic likelihood ratio (LR) methods arises for these methods to reach the necessary acceptance from the forensic practitioners community to be used in practice. A EU requirement for the use of accredited methods for the forensic evaluation of DNA and fingerprint evidence states that: *"Member States shall take the necessary steps to comply with the provisions of this Framework Decision in relation to dactyloscopic data by 30 November 2015."*

As the title of the thesis "Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation: *Application in Forensic Fingerprints*" suggests, this thesis mainly dealt with the forensic interpretation of discriminating scores produced by Automated Fingerprint Identification System (AFIS) and despite the fact that the validation framework for LR methods used for forensic evidence evaluation was in theory developed for application across the whole range of biometric modalities, its applicability was presented in the area of forensic fingerprints.

Answering the first research question: ***"Which criteria should be used to validate a LR-based inference model?"*** several literature surveys were conducted, addressing issues of guidelines and standards for validation of LR methods used for forensic evidence evaluation (chapter 1); a theoretical framework was proposed for the validation of LR methods used for forensic evidence evaluation (chapter 1); the theoretical framework developed was

[1] European Network of Forensic Science Institutes, *Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science*, working version 04-11-2013
[2] International Organization for Standardization EN ISO/IEC 17025, *General requirements for the competence of testing and calibration laboratories*, ICS: 03.120.20, stage 90/93 (2010-12-15)

applied to validate fingerprint LR method based on the AFIS scores (chapter 7); several issues were addressed in the course of the LR method development, namely robustness to the dataset shift (generalization – chapter 3,4) and robustness to the lack of data (data sparsity – chapter 4).

Answering the second research question: ***"What performance characteristics and metrics should be used to report the findings?"*** the measures of accuracy (chapter 1,4,5), discriminating power and calibration in (forensic) biometrics (chapter 4,5); use of Bayesian Networks (chapter 6) for fingerprint evidence evaluation and evidential value of the first level detail fingerprint evidence have been considered; different methods were used to calculate the LR's from the fingerprint AFIS scores and their performance evaluated using the performance metrics proposed in the theoretical framework (chapter 2,4); the primary performance characteristics used for the validation of the LR's presented in chapter 1 were used as an indicator of performance of the human-in-the-loop (appendix A).

Somewhat remotely stands the development of the Bayesian Network for the first level detail (General Pattern) fingerprint evidence evaluation (chapter 6).

A framework for validation of semi-automatic LR methods was presented in the first chapter of this thesis, motivated by the validation workshop organised in The Hague on 21 – 22 October 2011. This chapter addressed the questions of ***"what is the role of the LR as a part of the decision process"***, ***"how to deal with the uncertainty in the LR calculation"***, ***"what to validate"*** and ***"how to validate the LR methods"***. Performance metrics deemed suitable for measuring primary and secondary performance characteristics were introduced and the non-applicability of performance measures for the validation of analytical methods, borrowed from analytical methodology, was highlighted.

The issue of stability of the LR in the LR ≈ 1 region (LR > 1 supporting the prosecution and LR < 1 supporting the defence proposition) was presented in **Chapter 2**. The graphical representation chosen in this chapter (bar charts) has the tendency to encourage readers do draw "confidence intervals" around the LR. Although confidence intervals possess certain merit when measuring physical characteristics in analytical measurements (for which true values exist), these measures do not seem appropriate when dealing with the LR – argument presented in Chapter 1 (paragraph 5.1). On the other hand stability in the region where the LR ≈ 1 region is a desirable property of a well performing LR method. The LR method used in this chapter – the Kernel Density Function – showed unstable behaviour, when with the varying amount of data the LR values oscillated between the

support of either Hp or Hd propositions. The KDF baseline method was replaced by multimodal LR method in Chapter 4.

Chapter 3 was dedicated to the study of the robustness of the LR-based method to a lack of training and test data. The dependence on the quantity of data in the modelling stage was highlighted when using two different modelling approaches – suspect dependent (suspect anchored) and suspect independent (non-anchored). The suspect dependent approach showed higher robustness to the lack of training data than the suspect independent approach, for which the performance of the LR method degraded severely. Although quite illustrative and suitable for the purpose, the choice of box-plots might misleadingly encourage readers to draw confidence intervals around the LRs, analogous to the bar charts in the previous chapter. Currently, two major schools of thought influence the forensic community: the "Bayesian" and the "Frequentist", which causes heated discussions regarding how the uncertainty in the forensic evidence evaluation should be addressed. The Frequentists petition for the use of confidence intervals around the LR values, while the Bayesian like to consider the uncertainty encapsulated within the LR method and integrated out in the inference process.

The selection of an appropriate LR method, when dealing with multimodal score distributions was addressed in **Chapter 4**. Kernel Density Function (KDF) was used as a baseline LR method, against which the performance of a proposed multimodal LR method was evaluated. Despite having relatively good discriminating power, the KDF LR method showed sub-optimal performance in robustness to the lack of data (producing instable LRs, sometimes of irrational magnitudes) and the generalization due to dataset shift. These issues in particular were addressed in the development of the multimodal LR method. Primary performance characteristics were used to evaluate the performance of both methods in the modelling stage, while the secondary performance characteristics were used to determine appropriateness / usability of both the KDF and multimodal LR method.

Chapter 5 dealt with the coherence of LR values based on the quality of the data described by certain measurable parameters (number of minutiae) and the quality of the results (LR values) measured by the tools described in chapter 1. Chapter 5 firstly provided a definition of coherence and set it into the validation framework and subsequently showed an experimental example from fingerprint evidence evaluation using the multimodal LR method developed in Chapter 4. The comparison algorithm applied to computing discriminating scores uses 2 matching methods; hence the coherent behaviour of the multimodal LR method was observed separately for 5 – 10 minutiae configurations and for 11 and 12 minutiae configurations.

While Chapters 2 to 5 focused on the fingerprint minutiae, **Chapter 6** was dedicated to the first level detail fingerprint evidence – the General Pattern (GP) Graphical models – Bayesian Networks (BN) – were used to quantify the evidential value of GPs. Although the magnitudes of the resulting LRs is relatively low compared to LRs for second level detail fingerprint evidence, evidential values obtained from the GP using the BNs were deemed helpful in supporting the correct proposition in cases in which extreme distortion is present in the crime-scene fingermark and the only information visible feature is the GP. Furthermore, assuming independence, the first level detail and second level detail fingerprint evidence can be combined.

Finally, **Chapter 7** presented the validation of the multimodal LR model developed in Chapter 4 in the form of a validation report. Following the framework developed in Chapter 1 the report presented the scope of validation, method description, validation criteria and the empirical results. The validation report plays a significant role in the process of accreditation of such a method.

Appendix A showed an example of the use of the primary performance characteristics when evaluating the LR values based on the features extracted from the fingermarks by individual fingerprint examiners and automatically compared to the reference fingerprints. Although the initial assumption was that given the same training the LR's produced would show similar performance, the results were marginally different. It is important to note, that the human – the fingerprint examiner – plays a significant role in the forensic process and that the features extracted manually can vary from one examiner to another.

Research applications

1. The validation framework presented in chapter 1 will serve as a baseline for guideline for validation of LR methods used for forensic evidence evaluation. Despite the fact that the validation framework was empirically applied to the forensic fingerprints (presented in a form of a validation report in chapter 7), it can be universally applied to the range of forensic/biometric LR methods.

2. The multimodal LR method developed in chapter 4 can be applied to a range of forensic/biometric score-based applications, which similarly to the NFI AFIS system output multimodal score distributions.

3. As shown in the Appendix A, the performance characteristics/metrics developed within the scope of the validation framework can be used (given certain assumptions) to evaluate the performance of human practitioners.

Future work

Several research topics arise from the work conducted in this thesis. Following ones are particularly worthy addressing:

1. Open the validation framework (chapter 1) to the critiques in order to foster discussions within the scientific community and develop a guideline for the validation of LR methods used for forensic evidence evaluation.

2. Further exploration of additional secondary performance metrics and performance characteristics and their importance in the validation framework (chapter1). By introducing additional performance characteristics the validation procedure would become more robust and reliable.

3. Explore the usefulness of the multimodal LR model developed for the fingerprint evidence evaluation in realistic/operational conditions (operational validation) by the forensic practitioners (chapter 5).

4. Application of the validation framework to measure the performance of Bayesian Networks and explore the possibilities of combining the first and second level detail fingerprint evidence (Appendix A).

Biography

The author was born in Košice, Slovak republic on the 31st December 1981. He obtained the undergraduate degree from the faculty of Electrical Engineering and Informatics at the Technical University in Košice, Slovak Republic; graduating in Computer Science and Information Technologies in 2005.

In 2007 he completed his graduate studies by obtaining an MSc degree in Computer Science and Networking Technology at the Faculty of Engineering and Technology at the Manchester Metropolitan University, United Kingdom.

Following the work experience in the Institut français de recherche pour l'exploitation de la mer (IFREMER) in La Seyne-sur-Mer (France) in underwater robotics; Pildo Labs in Barcelona, Spain and Septentrio in Leuven, Belgium in satellite navigation he was offered a PhD fellowship at the Netherlands Forensic Institute in collaboration with the University in Twente focusing on the validation of LR methods used for forensic evidence evaluation.

List of publications

With forensic relevance:

R. Haraksim, D. Meuwly, *Fingerprint Evidence Evaluation – Robustness to the Lack of Data*, in proceedings EAFS 2012, Den Haag

R. Haraksim, D. Meuwly, G. Doekhie, M. Sjerps, P. Vergeer, *Assignment of evidential value of a fingermark general pattern using a Bayesian Network*, in proceedings BIOSIG 2013, Darmstadt

R. Haraksim, D. Meuwly, *Influence of the datasets size on the stability of the LR in the lower region of the within source distribution*, in proceedings BTFS 2013, Nijmegen

R. Haraksim, D. Ramos, D. Meuwly, *Validation of Likelihood Ratio Methods for Forensic Evaluation: Handling Multimodal Score Distributions*, submitted Science and Justice 2013

R. Haraksim, D. Ramos, D. Meuwly, Ch. Berger, *Measuring Coherence of Computer-Assisted LR methods: Experimental example*, submitted Science and Justice, 2014

R. Wang, D. Meuwly, R. Veldhuis, D. Ramos, J. Fierrez, and R. Haraksim, *Weighted complex spectral minutiae representation for forensic fingerprint recognition.* Pattern Recognition, 2013a. to be submitted by October, 2013

R. Wang, D. Ramos, D. Meuwly, R. Veldhuis, J. Fierrez, and R. Haraksim, *Assessing latent fingerprint distortion using forensic databases and minutiae paring by human experts,* In Proc. BBfor2 Conference on Biometric Technologies in Forensic Science (BTFS), Nijmegen, the Netherlands, Oct. 2013e. 49

Other:

R. Haraksim, L. Brignone, J. Opderbecke, *Multiple AUV control in an operational context: a leader – follower approach,* in proceedings Oceans 2008, Bremen

F. Maurelli, Y. Petillot, A. Mallios, S. Krupinski, R. Haraksim , P. Sotiropoulos, *Investigation of portability of space docking techniques for autonomous underwater docking,* in proceedings Oceans 2008, Bremen

F. Maurelli, F. Aklilu, R. Haraksim, Y. Petillot, *Robust localization of an autonomous underwater vehicle using sonar data and bayesian techniques*, in proceedings Oceans 2008, Bremen

Appendix A

Semi-automatic LR method:

Measuring performance of the human-in-the-loop

Rudolf Haraksim

1. Introduction

A very important aspect worthy consideration is the human factor in the process of fingerprint evidence evaluation using a validated semi-automatic LR method. Three roles of human involvement in the process have been identified: at the level of the methodology - choice of the technology and inference (LR) method, at the level of the development - implementation of the technology and inference (LR) method and at the level of the practice - use of the inference (LR) method in forensic practice. Another important aspect of the human factor is the acceptance of the validation process and the decision of using the LR method in practice, which involves setting up a hybrid approach for merging the evaluated strength of evidence of the human-based and semi-automatic method.

While the analytical results produced by computational methods are straightforward to interpret, the performance of human-based methods for interpretation and the human factor in the decision process should also be critically addressed. Given the same inputs, a computer-based LR method will always output a LR of the same magnitude (human factor in the LR method development resides in the choices made by the forensic developer); the same piece of evidence evaluated independently by two human examiners on the other hand might yield a different answer. This being said, the repeatability and reproducibility to the unitary precision of the LR is hardly the biggest issue[1]. Measuring performance of humans perhaps better fits within the domain of psychometrics, but using the primary performance characteristics described in Chapter 1, the validation framework developed can be extended to evaluate the performance of human examiners the same way it is used for evaluation of different LR methods.

Using an example from fingerprints we will attempt to evaluate the difference in the likelihood ratios produced by a LR method from the similarity scores, based on the feature vectors[2] extracted by the fingerprint examiners from the fingermarks and start with these assumptions[3]:

1. Treat each human (fingerprint examiner) as an independent "feature extraction system". The fingerprint examiners were trained in the

[1] Several approaches exploiting the same set of features should provide answers within a certain range and the LR value should be supported empirically with relevant data with the scientifically sound inferences made.
[2] Feature vectors considered here consist of spatial configuration of minutiae poitns and their orientation.
[3] We measure the performance of a subtask performed by the practitioner, it is not really the performance of the assignment of the strength of evidence that is directly measured.

same way, they should extract the features in the same way and the resulting LR values are expected be similar.

2. Ground truth is known for each fingermark and fingerprint pair.
3. Each examiner extracts the features from the constant number of fingermarks.
4. The feature comparison algorithm and the LR method do not change.

In the following example we aim to assess the performance of the LR values based on the similarity scores produced from the features vectors extracted by different fingerprint examiners.

2. Experimental example

Due to the limited number of participants – three – in this study, the following results are to be taken as an example of how the performance of "human examiners" can be evaluated, rather than as a prescriptive framework. The performance characteristics used for fingerprint examiner performance evaluation are in length described in chapter 1 paragraph 5 and related performance metrics in paragraph 6. Measuring the accuracy (Cllr), discriminating power (EER and $Cllr^{min}$) and calibration ($Cllr^{cal}$) we obtained results summarized in the table 1 below.

Table 1: Comparison of performance of three different fingerprint examiners

Performance characteristic	Performance measure	***Examiner 1***	***Examiner 2***	***Examiner 3***
Accuracy	**Cllr**	0.17	0.18	0.7
Discriminating power	**EER**	2.42	3.01	6.9
	$Cllr^{min}$	0.11	0.1	0.27
Calibration	**$Cllr^{cal}$**	0.06	0.08	0.43

Accuracy, discriminating power and calibration of the LR values of the features extracted by examiners 1 and 2 show similar results, while the LR values of the features extracted by examiner 3 appears to show lower performance. It is however rather difficult to draw any kind of conclusions based on such a small sample. The difference observed between the examiner 3 and the examiners 1 and 2 in this case depends on the amount of experience and will be subjected to further evaluation. Having used identical datasets, the same feature comparison algorithm and the same LR method for evaluation of fingermark evidence, it gives an indication that the differences observed are at the fingermark feature extraction level, which

was the only task of the fingerprint practitioner (the rest of the procedure was fully automated). It is important to realize, that in the evidence evaluation process presented only one parameter varies, namely the feature extraction from fingermarks – which is in any case human task.

Figures 1 to 3 below provide a graphical representation of the results in table 1. The plots of accuracy (ECE), discriminating power (DET and ECE^{min} plot) and calibration (Tippett plot) as specified in chapter 1 (paragraph 6.3) of this thesis will be used.

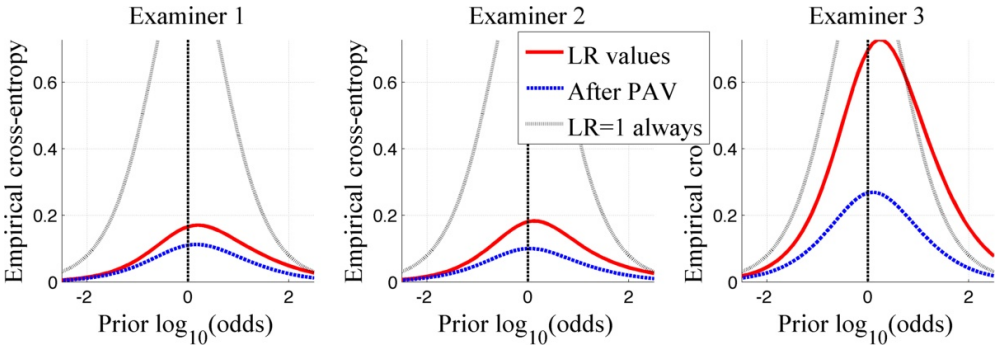


Figure 1: ECE plots (Examiner1 & Examiner2 & Examiner3 – 8 minutiae configuration dataset)

Figure 1 presents the *ECE* plot, capturing the accuracy and discriminating power of the LRs achieved by the comparison algorithm based on the features provided by individual practitioners. Recall from chapter 1 paragraph 6.3, that the blue dashed line (ECE^{min}) represents the discriminating power for each practitioner, red line (*ECE*) represents the accuracy for each practitioner and black dashed line represents a neutral system that always outputs LR = 1. The intersection of *ECE* and ECE^{min} with prior log odds = 0 we find values of *Cllr* and $Cllr^{min}$. From the *ECE* plots we can also conclude similar accuracy and discriminating power of the examiners 1 and 2 and somewhat lower performance of examiner 3.

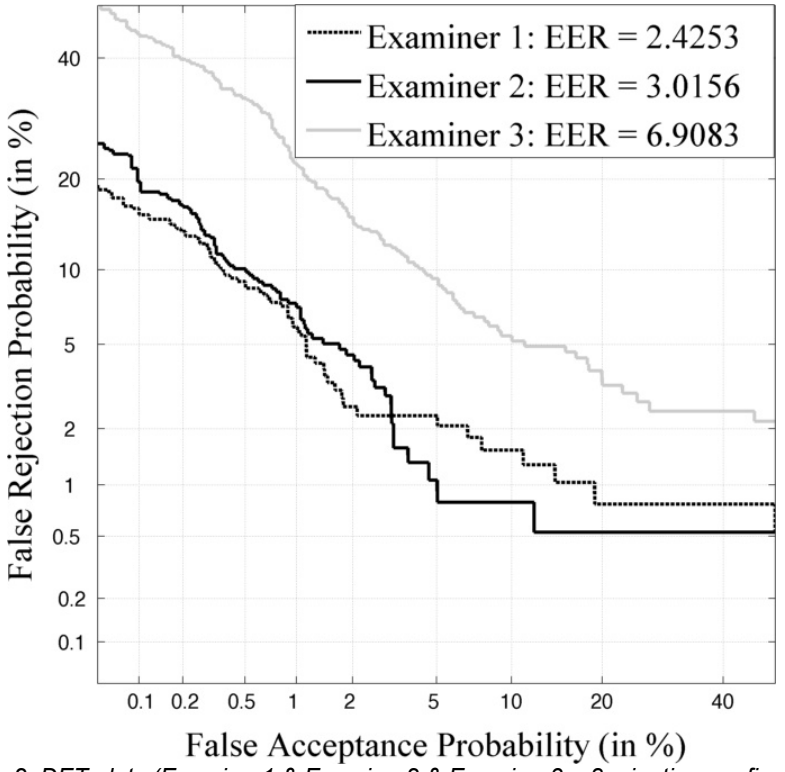


Figure 2: DET plots (Examiner1 & Examiner2 & Examiner3 – 8 minutiae configuration dataset)

The lower performance of the LR's based on the features extracted by examiner 3 is also observed in DET plots presented in figure 2. While the discriminating power of the LR values of examiner 1 and 2 is almost similar (black dashed and black solid lines), the accuracy of the LR values of examiner 3 (solid grey line) is more than twice as bad, as confirmed by the equal error rate (please note that the equal error rate represents one single operating point on the DET curve and there may be other operating points worthy considering).

Tippett plots presented in figure 3 conclude the graphical representation of the performance evaluation. Similar performance of LR values based on the features of the examiners 1 and 2 is clearly visible (black solid and black dashed line), together with the sub-optimal performance of the LR's of the examiner 3 (solid grey line). Further visible in the *Tippett* plots is the miscalibration of the third examiner when compared to the first two examiners. This is visible by marginally disproportional rates of misleading evidence in the upper curve of the *Tippett* plot supporting the prosecution proposition.

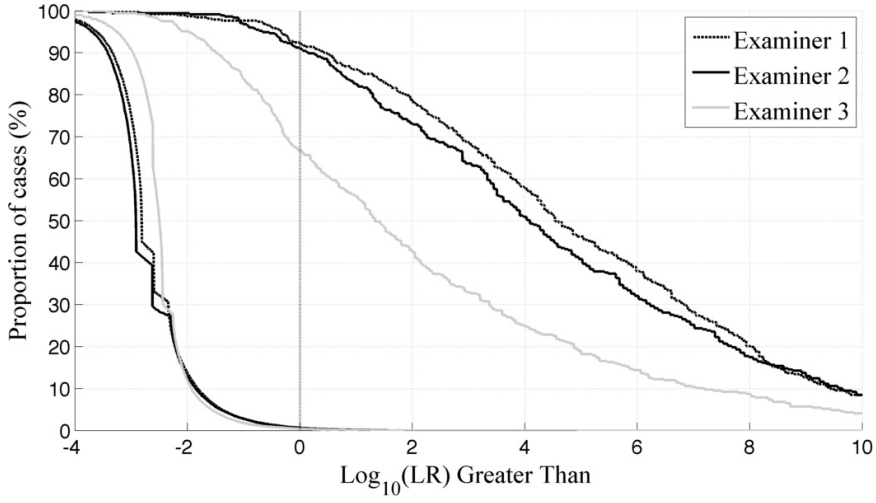


Figure 3: Tippett plots (Examiner1 & Examiner2 & Examiner3 – 8 minutiae configuration dataset)

Even though the three examiners in our example received the same training and should have extracted the fingermark features in the same way, the performance measures of accuracy, discriminating power and calibration vary, especially in the case of the LR's based on the features extracted by the examiner 3 and of the LR values based on the features extracted by the examiners 1 and 2. The differences observed are closely related to the years of practice and the amount of experience of each of the participating examiners.

The example above highlights the usefulness of validation tools presented and their potential use in performance assessment of forensic examiners. They partially allow monitoring the degree of expertise and the development of a practitioner.

3. Conclusion

Although the aim of this thesis was development of validation framework for semi-automatic LR methods, it is very important to realize, that it is the practitioner, who is in charge of the forensic process. The tools developed in for the valuation of the performance of LR methods provided to the practitioner help him in his evidence evaluation task. It is the human – forensic examiner, who chooses the features that feed the automated comparison algorithm and it is also the human – forensic evaluator, who is responsible for choosing a "well performing" LR method. It is again the human – the fact finder, who combines the LR values with relevant prior information supplied in case to make appropriate decision. Using the tools developed, the practitioner should be able to put the results in a meaningful perspective and to detect erroneous behaviour or unexpected results.

The aim of this work was not to eliminate the human practitioner from the process, since the forensic traces remain challenging to be adequately perceived by off-the-shelf biometric feature extraction algorithms and the human forensic practitioners remain a solid chain in the interpretation of the strength of evidence. The tools presented, applied to the human-in-the-loop performance evaluation, can be used to get better information about the degree of proficiency of each practitioner in the development, rather than in a competitive perspective.

Printed and bound by: www.ipskampdrukkers.nl
Cover designed by: Rudolf Haraksim
ISSN: 1381-3617
ISBN: 978-90-365-3648-6
http://dx.doi.org/10.3990/1.9789036536486