Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie

Universiteit Leiden

Assessing Performance of Score-Based Likelihood Ratio Methods for Forensic Data

by

René Neijmeijer

s1436643

A thesis

submitted in partial fulfilment of the

requirements for the degree of

Master of Science

Methodology and Statistics in Psychology

Thesis Supervisors:

Dr. W.D. (Wouter) Weeda (Internal)

Dr. A (Annabel) Bolck (External)

Leiden University

2016

Abstract

As part of my Master's Degree in Psychology and my internship at the Netherlands Forensic Institute I have structurally studied the performance of several score-based likelihood ratio methods. The method is a promising technique to establish the evidential value of forensic items, yet there is a need for research into the implementation of this method. A score function is part of the method and there are many to choose from. This thesis has attempted to expose a variability in performance between different score methods and to determine which scores are most appropriate for different forensic data. Several scores have been tested for a forensic dataset. In addition, this data was transformed four times for the purpose of testing the same scores under different conditions. These new four datasets separately included extreme values, negative values, a combination of extreme and negative values and binary values. Performance was measured by proportions of misleading evidence, the Cost Likelihood Ratio metric and the distributions of likelihood ratios. Results show that there is indeed variability in performance between scores. The Canberra and Clark distances appeared to perform best at establishing evidential value for continuous data and the Manhattan distance for binary data. Additionally, the Canberra and Clark distances also performed well for data with extreme values. Contrarily, none of the scores selected for this thesis managed to perform well on the two datasets with negative values. The results show that it is advisable to properly select a score for the likelihood ratio method.

Acknowledgements

Forensic statistics has been quite an unfamiliar topic to me before I started the internship and the writing of my thesis. This has been very inspiring and motivating to me and it has kept me going until the end. However, I could not have done this without the help of others. The Netherlands Forensic Institute has been a place that held many helpful and inspiring colleagues that introduced me to the subject. In particular, I would like to thank Annabel Bolck, my external internship supervisor. She has guided me throughout my exploration of the forensic science, without taking me by the hand too much. I would also like to thank my partner intern, Frederique Kool, who has helped me with a lot of mathematical issues, of which I had only little basic knowledge. Of course, I could not have written my thesis without the help of my internal supervisor, Wouter Weeda. His feedback has been very constructive and his response was promptly. Last, but not least, I would like to thank my family and friends, who have been interested in my studies and helped me to stay motivated. I would like to thank Leiden University and particularly the Psychology Master Specialisation Methodology and Statistics, coordinated by Tom Wilderjans. It has been an inspirational and challenging year and it has prepared me well for this internship/thesis. I am certain that I am now also prepared for what is awaiting me in the future.

Table of Contents

1. Introduction

1.1 Problem Statement

Forensic science is firmly intertwined with statistics, chiefly concerning the strength of evidence. Statistics have long been used to define the strength of a particular evidence and its application is occasionally present in criminal law. (Finkelstein & Levin, 1990; Aitken & Stoney, 1991; Gastwirth, 2000). Due to the diversity of forensic expertise (e.g. biology, psychology, chemistry) statistics are applied to a diverse collection of data. As a result, the forensic science had not known a uniform definition of the strength of evidence and its calculation. (Sjerps, 2004). This has changed as of the slow Bayesian revolution (Aitken, 1995; Robertson & Vignaux, 1995) which set on about 20 years ago and has developed after (Aitken & Taroni, 2004). The introduction of the likelihood ratio (LR) to the forensic field opened the door to a possibly uniform statistic. Forensic science often occupies itself with comparative research, i.e. two objects (e.g. glass pieces, DNA or fibres) are compared to each other in such a way that an expert can estimate how strong the evidence is. The estimation of the strength of evidence is traditionally based on the expert's evaluation and he or she concludes with a definite estimate of the origin. Contrarily, the Likelihood Ratio, as the name already discloses, is not a definite estimate of the origin, rather a probability. The probability is based on hypothetical statements on the origin of the evidence.

The LR being a probability is supposed to overcome the problem of case-specific conclusions. Many scientists believe that forensics should not be involved with conclusions on judicial ground. This conveys that when forensic scientists are to examine evidence, their findings should not explicitly point out any definite charges towards anyone or anything that is being accused. Instead, forensics should deal with evidence-specific hypotheses and they should report the likelihoods without drawing any conclusions who or what is responsible for these events. This leaves a jury, a judge or whatever legal body that is entitled to make juridical decisions to interpret the LR and other case-specific information, in order to form a conclusion.

Another problem that the LR addresses is that of the lack of objectification in forensic science. This is a concern that has been expressed by Evett (1998) and the US National Research Council (2009). The absence of a uniform and quantitative assessment of evidential value meant there were no possibilities to measure validity and reliability of the results that were provided by the forensic field. And there is little need to explain the importance of the measurements of validity and reliability in a scientific field, particularly in such which is concerned with matters as criminal law. The LR enables objectification of the methods that aim to calculate the involved probabilities. Besides, repetition of

the method enables validity and reliability even more. The more the method is applied, the more and the better the method can be reviewed. Fortunately, the ratio has been applied in many articles already, e.g. Davis et al. (2011) and Neumann (2007). Consider the list of references for more articles on the LR, which have been or will be discussed in this thesis. The LR can be both verbal and numerical. Some institutes, such as the Netherlands Forensic Institute and the Swedish National Forensic Centre, already report their findings through verbal LR's.

Despite its opportunities and growing popularity there are still many challenges to overcome. There is not just one LR method, as it is applied to a variety of data with many different characteristics. One method which implications can still be explored is the score-based approach. This approach is promising and overcomes some of the technical difficulties that other approaches sometimes fails to conquer. (Davis et al., 2012; Egli et al., 2007; Gonzalez-Rodriguez et al., 2007; Meuwly, 2006; Neumann et al., 2006; Neumann et al., 2009; Neumann et al., 2012; Nordgaard & Höglund, 2011). If the objects that are being compared have too many dimensions for modern-day computers, the score-based approach overcomes this problem by collecting most of the information in just one dimension. However, the score-based approach is still far from perfect and even this approach in turn knows several variations.

One of the variable aspects is the way of comparing forensic objects to each other. The score-based approach applies score formulas to realise this comparison (e.g. distances and similarities). There are many formulas to choose from (Cha, 2007; Choi et al., 2010) and they generally create different scores among each other. Different scores, which implies different comparison results, lead to different LR conclusions. Yet this is seldom acknowledged. There are plenty of examples of studies that either apply conventional scores or justify one choice of score. For example, in a study on comparisons of striated tool marks, Baiker et al. (2014) used a cross-correlation metric, in which this score is successfully proven to be capable of performing comparisons between striation marks on bullets. Baiker et al. concluded that the cross-correlation metric was found to be successful, validated by the results of training-test data. Although they justified their score selection by an article of De Kinder (1999) other potential scores may have outperformed the current results. More examples of literature in which one score is considered, may it be justified or not, are Chazal et al. (2005) and Pierrini et al. (2006), suggesting that this is not an incidental case. In addition, there are authors that explicitly affirmed the potential consequences of choosing just 1 score. Neumann (2007), for instance, was aware of the fact that he based his results solely on 1 score and noted that undoubtedly more research is needed on the score itself, which could lead to more convincing results. Other authors who explicitly state the limitation of their selective choice of score (and may

justify this) are Davis et al. (2011), Hepler et al. (2011), Horswell et al. (2002) and Inoue et al. (2003). Fortunately, at times effort is attributed to score selection and often it yields insight in the behaviour and fit of the chosen scores. For instance, the Bray-Curtis distance was found to be best suited compared to the Euclidean distance in a study by Quaak and Kuiper (2011) on the comparison of bacterial profiles in forensic soil comparisons. More examples of such score comparisons are Esseiva et al. (2003), Lociciro et al. (2007), Marquis et al. (2008), Neumann & Margot (2008) and Pervouchine & Leedham (2006).

All of these studies show that, first of all, it seems to be that often just 1 score is considered and, second of all, it may be beneficial to consider more than just one score. Since forensic data is a diverse collection of various data types, there is a need for research into the appropriateness of score selection for certain types of data. This concern has also been expressed by i.a. Hepler et al. (2012) and Neumann (2006), yet I have not found any author who has carried out such a study, at least not in a LR framework. In this thesis, I do not aim to come up with one universal supreme score, as the performance of a score will vary among types of data. Furthermore, I cannot cover all types of data, merely a selection of common ones in the forensic field. Nonetheless, my intentions are that at the end of this thesis I will have illustrated that score selection is a vital part of the score-based method construction phase and to provide insight into the appropriateness of scores for specific types of data. Based on methodological research, rather than on a theoretical debate, I aim to answer the question: To what degree is score selection a vital part of the Likelihood Ratio method and which score should be used for specific types of data?

The following paragraph of this chapter will elaborate more on the technicalities of the LR method, particularly the score-based approach. In chapter 2 the methods will be discussed, in chapter 3 the results and in chapter 4 the discussion on those findings.

1.2 Likelihood Ratio Specifics

Before we continue, the LR should will be explained in order to understand what is being researched. In the forensic field, the LR is a number that often represents the strength of evidence. More specifically, it often represents a likelihood value that some particular item originates either from a particular source or from any other random source. The LR is based on the Bayesian rule, which constitutes the conditional probability of event A given event B:

$$P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B]} \quad [1]$$

The Bayes rule in odd terms can be put into the forensic notation, which is:

$$\frac{P[H_p|E]}{P[H_d|E]} = \frac{P[H_p]}{P[H_d]} \cdot \frac{P[E|H_p]}{P[E|H_d]}$$

$$\begin{Bmatrix} post. \\ odds \end{Bmatrix} = \begin{Bmatrix} prior \\ odds \end{Bmatrix} \cdot \quad \{LR\} \quad [2]$$

E stands for evidence and is usually a combination of evidence features or a comparison between features of some recovered item from a crime scene (measurement Y) and a control item (measurement X). H_p is generally the hypothesis in which both X and Y originate from the same source and H_d is usually the hypothesis in which they do not originate from the same source, however depending on the case these can be stated otherwise. The formula signifies that the multiplication of the prior odds and the LR generates the posterior odds. For example, firstly a judge estimates the probability of a particular person committing a crime versus another random person committing that same crime, before the strength of evidence is evaluated by an expert, which in turn updates the odds from prior to post. The traditional forensic mode of operation focused on the posterior odds, which is the final product on which a juridical verdict is based. Now the focus of the forensic experts is shifted towards the LR, leaving the prior and the posterior odds for a legal entity to be interpreted, as it is believed that forensics should not be immersed with the events and/or circumstances related to the evidence. Consider, for example, two glass pieces: one that was found at a crime scene (Y) and one that is questioned whether or not it originates from the crime scene (X). H_p is the hypothesis that states that both glass pieces come from the same source. H_d is the hypothesis that stated that they do not come from the same source. The LR nominator gives the probability of finding E (the combination and/or comparison of the two glass pieces) given that H_p holds and the LR denominator gives the probability of finding E given that H_d holds. If the probability under H_p is bigger than the probability under H_d, the LR will be bigger than 1, indicating

that the evidence is pointing in the direction that it is more likely the two glass pieces come from the same source. The opposite is true for LR values that are below 1. The further the value from 1 (approaching infinity for H_p-supporting values and 0 for H_d-supporting values), the stronger the support for that hypothesis. The Netherlands Forensic Institute maintains a verbal scale for the values that the LR can take, which can be found in appendix I. For instance, a LR between 2 and 10 means that the forensic findings are slightly more probable given H_p relative to H_d. LR's between 100 and 10,000 mean that the forensic findings are much more probable given H_p relative to H_d.

The arrangement of E is what this thesis is most engaged with. There are two major approaches to the LR method: The feature-based approach and the score-based approach. The first one is the probability evaluation of the combination of characteristics (features) of the forensic items. Consider the two glass pieces again, and this time it is also made known that both are made of a specific type of sand that contains a specific amount of silicon. X contains .8 gram of silicon and Y contains .9 gram. In this case the nominator of the LR is the joint probability of finding this combination of silicon quantities, given that the pieces are indeed from the same source. Imagine that this type of glass with silicon quantities within the .5 - 1.0 gram range is extremely rare, then this probability would be high. The LR denominator is the probability of finding this combination under the hypothesis that they do not originate from the same source. In this 'rare-glass-case', this probability would be low. Eventually this leads to a high LR, whereas the LR would be lower if the sand were to be very common.

The score-based approach produces E in a different way. Instead of looking at the combination of features, a comparison between the two features is considered. Let us consider the glass pieces again, but now a different characteristic, namely the quantity of silicon that is measured in both pieces. The score-based approach does not combine these two quantities, rather compares them to each other. The comparison is often conducted by means of scores. If X would have .5 gram of silicon and Y would have .9 gram, then the Manhattan distance would simply be a score of .4 gram. See also Figure 1, in which X and Y are compared to each other by means of a score formula, which creates E. The LR numerator is the probability of finding this score under the hypothesis that the pieces do indeed originate from the same source, i.e. the likelihood of finding this score, given the population of scores that are comparisons between glass pieces that are known to be from the same source. The LR denominator is the probability of the same score under the hypothesis that they do ***not*** originate from the same source, i.e. the likelihood of finding this score, given the population of scores that are comparisons between glass pieces that are known ***not*** to be from the same source. The first population of scores are called the within scores (E_w) and the second are called the

between scores (E_b). These are available through a background database, which consists of a substantial sample of items from several sources. In our glass case this database would consist of several measurements of silicon quantities in several glass sources. The probabilities can be estimated through a standard statistical distribution model, such as Gaussian, Weibull or Gamma. However, as the distribution in Figure 2 shows, it is often hard to find an appropriate model. We see a distribution of between scores. This distribution might come close to a Gamma distribution, however, that model does not take the peak between .1 and .15 into account. Non-parametric methods, such as the Kernel Density Estimation [8], instead, use the data in order to smooth the distribution. (Silverman, 1986). Often distributions of scores even have more peaks than the example in Figure 2 and KDE is able to account for this.

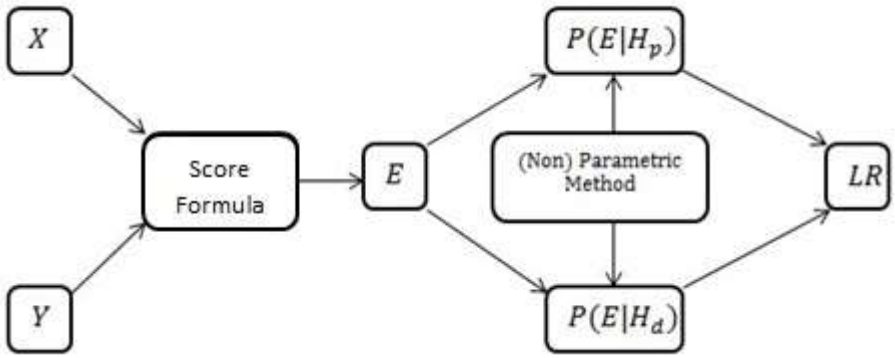


Fig.1. The score-based approach to the LR method. The quantitative measurements of X and Y are compared to each other by means of a particular score formula which produces E. Through a probability estimator method, the probabilities of finding that score under H_p ($P(E|H_p)$) and under H_d ($P(E|H_d)$) are estimated and the division between these two is the LR.

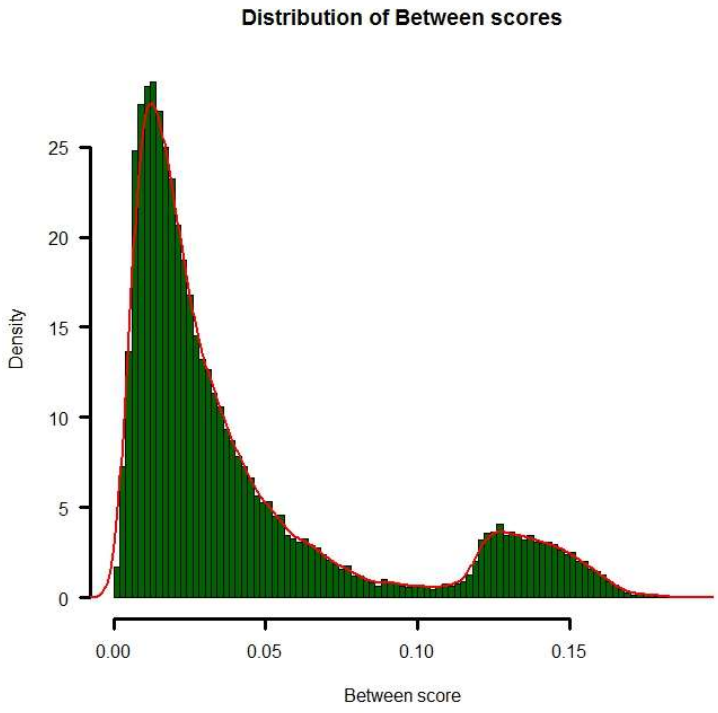


Fig. 2. Example of a distribution of between scores with the red line being the density curve (as estimated by the KDE).

2. Methods

In this chapter, two methods will be explained that aim to expose a difference in score behaviour and performance among scores. Hence the question which scores are most appropriate for which type of data can be answered. First of all, in 2.1 a series of modest experiments have been conducted, which have indicated differences in behaviour and performance of scores in simple settings. These simple settings are 3 different types of data that have characteristics which can be obstacles for scores: data with extreme values, data with negative values and binary data. This provided a reasonable footing for a more complex methodology. In 2.2 the data that was used for this thesis is explained and justified. In 2.3 it is explained how the distances and data were compared to each other and what computations needed to be done to realise an answer to the research question. Finally, the tools that helped to analyse the results and draw conclusions are explained in 2.4.

2.1 Score experiments

2.1.1 Euclidean Distance

$$s(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad [3]$$

One of the most well-known and used score measures is the Euclidean distance [3], based on the Pythagorean theorem. Here, $s(X,Y)$ means the score that is a result of a comparison between the (quantitative) features of two forensic items. It is probably also one of the most discussed scores. In general, this distance behaves and performs well and is therefore widely used. As soon as the data introduces difficulties, its application may become problematic. For example, if multivariate data is not equally scaled (i.e. different units), the Euclidean distance is not trustworthy. In particular, due to the fact that it squares over $X_i - Y_i$, which enhances the effect of distortion in case of different units. Ertöz et al. (2003) found that the Euclidean distance does not function well in case of binary/categorical data and data with many variables. The last instance is also described in the study by Aggarwal et al. (2001). Troyanskaya et al. (2001) claim that the Euclidean distance performs poorly for noisy data (lots of spikes/outliers).

2.1.2 Manhattan Distance

$$s(X,Y) = \sum_{i=1}^{n} |X_i - Y_i| \qquad [4]$$

The Manhattan distance is similar to the Euclidean distance, yet simpler. It takes the absolute value of the difference between X and Y. This property of absoluteness should make the score more robust to extreme values in comparison to the Euclidean distance. Consider point 1 (p1) and point 2 (p2) in a 10-dimensional-space (see Table 1). The Euclidean distance between these two points is 6.32 and the Manhattan distance is 16. (see Table 2 for the results). Now consider a point 3 (p3) which is equal to p2, except for variable B which now has a clear extreme value of 21. For p1 and p3 the Euclidean distance increases by a factor of 3.16 to 20 and the Manhattan distance increases by a factor of 2.13 to 34. While the Manhattan distance does not feature a particularly strong defence mechanism against the influence of an outlier, it does manage to control it better than the Euclidean distance. According to Aggarwal et al. (2010), the Manhattan distance works better in higher dimensional spaces than other scores, which could be attributed to its simplicity. One problem that the Manhattan distance, like the Euclidean distance, does not overcome well, is the problem of binary and categorical data.

Table 1

Experiment 1: Extreme Values

Point	A	B	C	D	E	F	G	H	I	J
p1	3	2	4	0	1	2	3	1	2	0
p2	0	3	3	2	2	2	5	3	2	4
p3	0	21	3	2	2	2	5	3	2	4

Note: p1 and p2 are two imaginary measurements with 10 variables and p3 only differs from p2 in variable B

Table 2

Results of Experiment 1

	Euclidean Distance	Manhattan Distance	Bray-Curtis Distance	Jaccard Distance
$s(p1,p2)$	6.23	16	.36	.47
$s(p1,p3)$	20	34	.55	.83
Factor	3.16	2.13	1.51	1.77

Note: This experiment applied the continuous versions of the scores.

2.1.3 Bray-Curtis Distance

$$s(X,Y) = \frac{\sum_{i=1}^{n} |X_i - Y_i|}{\sum_{i=1}^{n} (X_i + Y_i)} \quad [5]$$

This score is favourite among the ecologists (Looman & Campbell, 1960), which often work with count data. It might not be so useful for non-count data, given that this data may contain negative values which could result into a negative denominator. This could then again lead to a difficult interpretation of the metric value itself.

Table 3

Experiment 2: Negative Values

Point	A	B	C	D	E	F	G	H	I	J
p4	3	-2	4	0	-1	-2	-3	1	2	0
p5	0	-3	3	-2	-2	-2	5	3	-2	-4
p6	0	3	3	-2	2	-2	5	3	-2	-4
p7	0	-3	3	-2	-2	-2	-5	3	-2	-4

Note: These are 4 imaginary measurements of which p6 differs from p5 in variable B and E and p7 from p5 in variable G

Table 4

Results of Experiment 2

	Euclidean Distance	Manhattan Distance	Bray-Curtis Distance	Jaccard Distance
$s(p4, p5)$	10.77	26.00	-13.00	.94
$s(p4, p6)$	12.17	32.00	4.00	1.06
$s(p4, p7)$	7.48	20.00	-1.66	.60

Note: This experiment applied the continuous versions of the scores.

Consider 4 different points in a 10-dimensional-space (see Table 3). The Manhattan distance between p4 and p5 is 26 and the Bray-Curtis distance is -13. (See Table 4 for the results). The sum of p5 is a negative value, leading to the negative quality of the score. Point 6 differs from p5 in such a way that two negative values are transformed into their positive equivalents (variable B and E). The individual Manhattan distances of B and E between p4 and p6 are now bigger than the distances between p4 and p5. Depending on the context, the whole p4 vs. p6 distance should intuitively be bigger too. The results show that the Manhattan distance between p4 and p6 is 32 and the Bray-Curtis distance is 4. Both are higher numbers, so the assumption was correct. Now consider p7, which differs from p5 in such a way that variable G is turned into its negative equivalent. This time it feels as if p7 and p4 are closer to each other than p5 and p4. The Manhattan distance is 20, which

supports that assumption, but the Bray-Curtis distance is $-1\frac{2}{3}$. Suddenly, the Bray-Curtis distance becomes hard to interpret, when the intuitive sequence of dissimilarity from small to big is: p4 vs. p7, p4 vs. p5, p4 vs. p6 and the according Bray-Curtis distance sequence is: $-1\frac{2}{3}$, -13, 4. A bigger problem arises when the denominator sums up to zero, as this is mathematically impossible, yet, theoretically, this could occur with non-count data. The experiment of Table 1, in which p3 introduces an extreme value, can also be applied to the Bray-Curtis distance in comparison to the previous two. The distance between p1 and p3 is 1.51 times bigger than the distance between p1 and p2. This shows that the Bray-Curtis distance is more robust to extreme values than the Manhattan and the Euclidean distances.

2.1.4 Jaccard Distance

$$s(X,Y) = \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} X_i Y_i} \quad \text{(Continuous)} \quad [6]$$

$$s(X,Y) = \frac{b+c}{a+b+c} \quad \text{(Binary)} \quad [7]$$

a = Amount of variables that are both present (1)
b = Amount of variables which are present in X (1) and absent in Y (0)
c = Amount of variables which are absent in X (0) and present in Y (1)

The Jaccard distance, like the previous and more scores, is diverse as it knows different versions for different types of data. The experiment of Table 1 shows that the version for continuous data [7] provides a distance that is a 1.78 factor bigger due to the introduction of an extreme value. This is lower than the Euclidean and the Manhattan shifts and slightly higher than the Bray-Curtis factor. Consider the points in Table 5. The Euclidean and Manhattan distances[1] between p8 and p9, and between p10 and p11 are exactly the same, respectively 1.41 and 2. (See Table 6 for the results). This is odd in some cases where the mutual absence of a variable does not imply similarity. When two measurements both share a variable (presence), it logically implies a similarity. However, when they both lack a variable (absence), it does not necessarily mean that they are very similar to each other. According to this understanding, the distance between p8 and p9 should be bigger than the distance between p10 and p11. This phenomenon is tackled by Jaccard and Jaccard-like distances that consider the properties of absence and presence. For Table 5 the Jaccard distances are 1 and

[1] Note that the proper binary distances can be found in the appendix III. In this case the formulas 3, 4 and 5, however, do give the same results as their binary cousins.

0.22. The Bray-Curtis distance[1] demonstrates the same relationship (distances of 1 and 0.13 respectively). Of course, it depends on the context and the meaning of the binary values, whether shared absence or presence means similarity, which is the case for all previous intuitive assumptions made.

Table 5

Experiment 3: Binary Data

Point	A	B	C	D	E	F	G	H	I	J
p8	1	0	0	0	0	0	0	0	0	0
p9	0	0	0	0	0	0	0	0	0	1
p10	1	1	1	0	1	1	1	1	1	0
p11	0	1	1	0	1	1	1	1	1	1

Note: These are 4 imaginary measurements with only binary values of 0 and 1.

Table 6

Results of Experiment 3

	Euclidean Distance	Manhattan Distance	Bray-Curtis Distance	Jaccard Distance
$s(p8, p9)$	1.41	2	1	1
$s(p10, p11)$	1.41	2	.13	.22

Note: This experiment applied the binary versions of the scores.

2.1.5 Final Scores

The previous paragraph shows that scores behave differently with respect to the anticipated result in a rather uncomplicated setting. Whether or not this is also true for the likelihood ratio method is to be observed next. And if so, which score would be most appropriate? The studies by Cha (2007) and Choi et al. (2010) list several scores that have been carefully inspected. These are 56 potential scores for continuous data and 76 potential scores for binary data. Note that some of these are written as similarity measures instead of distance measures. This would not affect the interpretation of the final results and these similarity measures can be used instead. For time and computation power reasons the number of scores had to be reduced. Cha and Choi used hierarchical clustering to find similarities between the scores. Their results and the popularity of scores in the forensic field are the foundations on which the selection of 20 continuous and 20 binary scores have been made. The selected scores and their formulas can be found in appendices II and III.

2.2 Data and Transformation

For this thesis, essentially any type of data would have been suitable, as there is yet much to be discovered about the role of data with respect to the performance of LR methods. A dataset adapted for experimental usage, but extracted from real data, had been made available by Peter Zoon, a scientific researcher at the Netherlands Forensic Institute. The dataset contained multiple measurements of 7 chemical elements of certain knives, conducted with the use of a Scanning Electron Microscope (SEM). There are 15 knives and for each knife, 20 measurements had been carried out on 20 different locations on the knife. One knife is one source, indicating that there were 15 sources, each containing 20 measurements, which is a total of 300 measurements. For each measurement, the quantities of 7 chemical elements had been measured. Some summary statistics can be found in Table 7. There were not many curiosities in the data. At first glance the variables did not seem to follow any conventional or the same distribution. What they all did have in common is that the variables were continuous and the measurements did not take any negative values.

Table 7

Statistics of Element Concentrations (in %)

Chemical Element	A	B	C	D	E	F	G
Mean	.88	.03	14.60	.34	83.11	.78	.26
Median	.59	.00	14.19	.30	84.36	.39	.00
SD	1.12	.07	1.71	.36	3.94	1.69	.50
Min	.13	.00	12.21	.00	68.59	.00	.00
Max	5.38	.38	23.68	1.47	87.26	8.26	1.83
No. Outliers *	15	2	14	4	12	8	1
Presence in source **	15	2	15	8	0	12	4

Note: The mean, median, standard deviation, minimum value and maximum value refer to the complete dataset of 300 measurements.

* The number of outliers out of 300 measurements

** The amount of times the element was present in a source (out of 15 sources).

There were two qualities of the data that needed attention: Absence of some variables in some sources and the number of outliers. The knives mostly consisted of element E and C and those plus element A were always present in any knife, whereas the rest were sometimes absent. Especially element B (only present in 2 of the 15 sources) and G (only present in 4 of the 15 sources) were rarer than the other elements. This indicates that this data knew many zeroes, which represented absence of that variable. It was expected that this would affect the LR performance. Zeroes are mathematically quirky numbers which sometimes behave differently and/or unexpectedly compared to other numbers. This has also been illustrated by Ertöz et al. (2003), who performed similar

experiments as in paragraph 2.1. They found that the Euclidean and the Manhattan distance do not perform well in data with many zeroes, as opposed to the Jaccard distance. The scores as a result of this data may also have been affected and consequently the LR's too. The other data characteristic worth mentioning are the number of outliers detected. Outliers were detected per cluster (i.e. the measurements per source per element, e.g. all the 20 measurements of element A in knife 1) that were located outside the range of the corresponding quartile ± 1.5 × the interquartile range. The outliers may have distorted the distribution of scores and therefore the probabilities in the numerator and the denominator of the LR. Also, this may have had an impact on the performances of some scores and, consequently, on the LR performance.

In 2.1, three other data qualities had been disclosed, which seem to affect score performance. These three qualities were the inclusion of extreme values, the inclusion of negative values and binary data. These three qualities are not uncommon in the forensic field. The number of outliers in the knife dataset already suggest the presence of extreme values, although one may not necessarily classify these values as extreme. In a forensic context, extreme data can be a result of measurement error. Because the data conclusions may have a substantial judicial impact, the omission of outliers/extreme values is not always uncontroversial. Besides, there may be financial and/or technical restrictions. Data with negative values is most common when the data is scaled. One could argue to rescale the data so that it loses its negative property. However, some data is constituted in such a way that the negative property is meaningful and should not be lost in the process. Binary data does occur too and requires specialised score formulas. It was decided to transform the original dataset to additional datasets which incorporated these three qualities.

For the data with extreme values, in each cluster 2 random values have been replaced by a value that was the maximum value of that cluster plus the cluster's standard deviation multiplied by a factor that randomly lay between 2 and 4. If such a cluster happened to have only values of 0, the cluster has been left untouched, as it makes no sense to insert extreme values, which would indicate the presence of that chemical element, for a source knife that does not even contain that element at all. This dataset does not differ from the original one in terms of outlier presence, yet this time the outliers are significantly more extreme.

For the transformation into data with negative values, the cluster has been demeaned. Also, the previous generated data with extreme values has been demeaned in order to create a dataset for the interaction effect between the two qualities. Perhaps the effect of extreme values differs whether or not the data includes negative values.

For the binary data, the transformation was conducted by variable. The variable values had to be split into two halves, of which one was attributed a 0 and the other was attributed a 1. The split criterion needed to be the same across the variables, but it also needed to gain the same effect. Because some variables were sparse, the median of each variable should be the split criterion, rather than the mean. For some variables, this was identical to the logical argument of the presence or absence of that particular chemical element. No interaction effects between the binary data and the other two have been explored, as it made no sense to transform negative and extreme data into binary data, since the properties would get lost in the transformation.

Table 8

Overview of datasets

A_1	Adapted dataset, extracted from real data, including 15 sources with 20 measurements each.
A_2	Transformed dataset of A_1, to which extreme values are added
A_3	Transformed dataset of A_1, which has been demeaned to create negative values.
A_4	Transformed dataset of A_2, which has been demeaned to create in interaction dataset.
A_5	Transformed dataset of A_1 to binary values with the cluster median as the split criterion.

2.3 Computations

Each score is applied in combination with each dataset, except for score 6 till 9 in combination with A_3 and A_4, for these formulas contain natural logs and roots that cannot process negative values.

For both scenarios, the following was considered: a piece of metal is found on a victim's bone, which is assumed to come from a knife. This is synonymous to 1 measurement of 1 source from a data set. This measurement was called the Y (recovered). The X (control) was a measurement that was carried out on a certain knife. This is also 1 measurement of 1 source from a dataset. Please take note that although 1 measurement may seem slim, in real practice, although not necessarily this dataset, it is often not feasible to carry out more measurements, due to technical and/or financial limitations.

In order to measure performance, it was necessary to establish the origin of X and Y. Fortunately, an experimental dataset as described in 2.2 gives the opportunity to keep track of this. There were two propositions to be taken into account. Proposition 1 is the scenario for which H_p is true, which indicates the hypothesis that X and Y come from the same source, and proposition 2 is the scenario for which H_d is true, which indicates the hypothesis that X and Y do not come from the same souce. Under H_p, E is a score between X and Y that must come from the same source, a within score (E_w).

Under H_d, E is a score between X and Y that must come from different sources, a between score (E_b). The following procedure have been executed for every score and dataset combination. Firstly, all the possible E_w scores were computed, out of all the possible X and Y comparisons under H_p. The first measurement was compared to all other 19 measurements from that same knife, the second measurement was compared to all the remaining 18 measurements and not 19, so that no double scores were computed. Then the third measurement was compared to all other 17 measurements and eventually this leads to the summation of $19 + 18 + 17 + 16 \ldots + 1$. Because this was done for every knife, this summation was carried out 15 times. This has led to 2850 E_w scores ($(19 + 18 + 17 + 16 \ldots + 1) \times 15 = 2850$). Secondly, all the E_b scores were computed out of all the possible X and Y comparisons under H_d. The first measurement was compared to all measurements from other knives, which are 280. The first measurement of the second knife was compared to all other measurements from other knives, except for the first knife, in order to avert double scores, which are 260 scores. Eventually this leads to the summation of $280 + 260 + 240 + 220 \ldots + 20$. Since this counts for each measurement within a knife, this summation was carried out 20 times. This has led to 42,000 E_b scores ($(280 + 260 + 240 + 220 \ldots + 20) \times 20 = 42{,}000$).

Because there are 2850 E_w's and 42000 E_b's, it was possible to compute 2850 LR's under H_p (LR_p) and 42000 LR's under H_d (LR_d). In other words, there were 2850 LR's of which was established that X and Y came from one knife and 42000 LR's of which was established that X and Y came from two different knives. In order to get from E to LR_p and LR_d, the numerator and the denominator of the LR needed to be calculated. The LR_p numerator ($P(E_w|H_p)$) is the probability of finding an E_w score in an E_w distribution and the LR_p denominator ($P(E_w|H_d)$) is the probability of finding that same E_w score in an E_b distribution. The two probabilities have been estimated through Kernel Density Estimation with a Gaussian Kernel [8]. For $P(E_w|H_p)$ it means that the E_w is s and it is evaluated against all other E_w's, which are the s_i's. All of these evaluations add up to the probability of finding that within score under H_p. For $P(E_w|H_d)$ the E_w is evaluated against all other E_b's. These add up to the probability of finding that within score under H_d. Inversely, the LR_d numerator ($P(E_b|H_p)$) is the probability of finding an E_b score in an E_w distribution and the LR_d denominator ($P(E_b|H_d)$) is the probability of finding that same E_b score in an E_b distribution.

$$f(s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{1}{\sqrt{(2\pi)}h\hat{o}_s} \exp(-\frac{1}{2h^2\hat{o}_s^2} \cdot (s - s_{i)}{}^2) \qquad [8]$$

$$\hat{o}_s = \sqrt{\frac{\sum_{i=1}^{n_s}(s_i - \bar{s})^2}{n_s - 1}}$$

$$h = \frac{4}{(2+p)n_s}^{\frac{1}{p+4}}$$

$$p = \ No. of\ dimensions\ of\ s\ (which\ is\ always\ 1)$$

$$n_s = Length\ of\ s\ (either\ 2850\ or\ 42{,}000)$$

2.4 Analysis

After all the calculations, there was a pair of two LR distributions (LR_p and LR_d) for each of the 20 scores and for each dataset (A_1 till A_5). Intuitively, one could say, the more support for the corresponding hypothesis (e.g. the higher the score is under LR_p), the better the LR-method. A good starting metric for this would be one that represents misleading evidence. Misleading evidence is when the LR supports a hypothesis which is not true. For this to evaluate, a threshold is required, which theoretically could be anywhere on the LR scale. From a mathematical point of view, LR = 1 would be the most appropriate threshold, which means that there is no support for either of the two hypotheses. Therefore, in this thesis, a correct LR would be when the LR is higher than 1 if H_p holds and lower than 1 if H_d holds. A more conventional way of reporting this is the opposite one: The False Positives (FP) and the False Negatives (FN), which represent misleading evidence. A convenient graphical representation of the FP and the FN is the Tippett plot. Please consider Figure 3, which is an example. The Tippett plot presents two cumulative lines, one for the LR_p (red solid) and one for the LR_d (green dashed). On the x-axis are the log-transformed LR's and the y-axis tells what proportion of an LR distribution is higher than the log-LR that it corresponds to. The threshold is now log(1), which is 0. The FP is the green line right from the 0 line, the FN is the red line left from the 0 line. An ideal Tippett plot would present the red line pushed to the upper right corner, whereas the green line would be in the lower left corner.

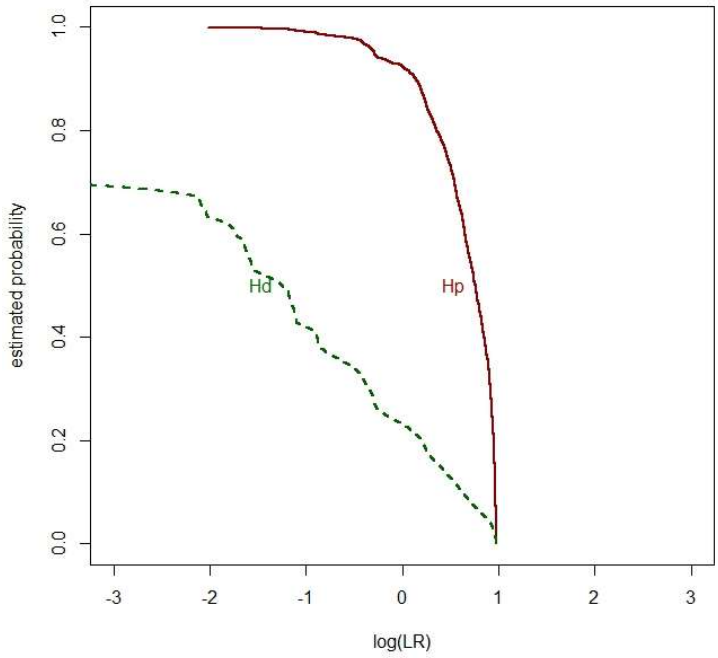


Fig. 3. A Tippett Plot example

But this only captures 1 element of the method's performance. An even better way of ranking and comparing the methods is by applying the log-likelihood-ratio cost ($Cllr$), adapted for validation purposes by Morrison (2011), based on a more general introduction by Van Leeuwen and Brümmer (2007). The $Cllr$ is computed using the following formula:

$$Cllr = \frac{1}{2}\left(\frac{1}{N_{LR_p}}\sum_{i=1}^{N_{LR_p}} 2_{log}\left(1+\frac{1}{LR_{p_i}}\right) + \frac{1}{N_{LR_d}}\sum_{i=1}^{N_{LR_d}} 2_{log}\left(1+LR_{d_i}\right)\right) \qquad [9]$$

The metric can be interpreted as a metric that expresses the cost of faulty decisions that could be made, based on the threshold of $LR = 1$. Anything below a Cllr value of 1 is acceptable and the lower, the better. If, for instance, LR_p is below 1, which is faulty because it is calculated in the scenario that H_p holds, then it can be deduced from the formula that $Cllr$ will increase. The lower LR_p is, the higher $Cllr$. The opposite counts for LR_d. In this way, not only the quantity of instances a faulty decision could be made, but also the cumulative probability of actually making that decision is evaluated.

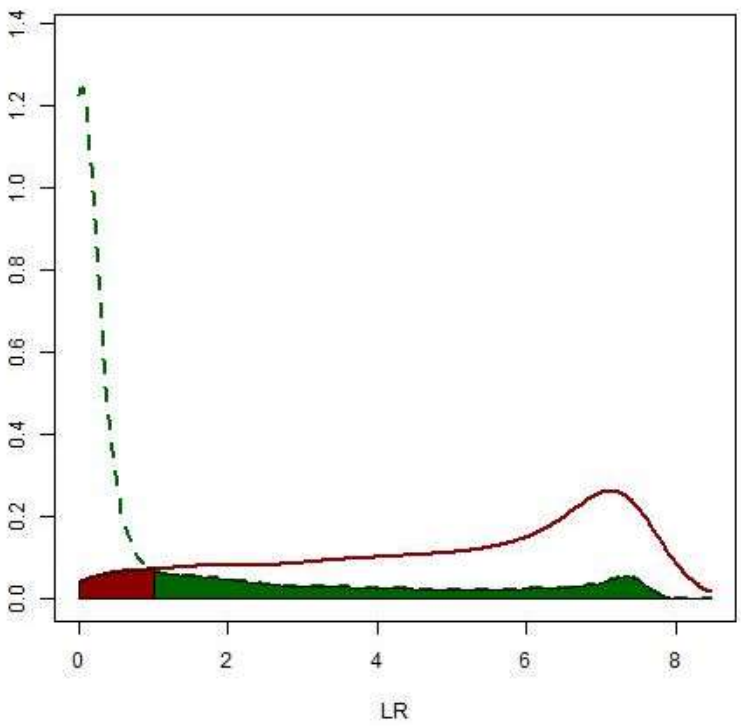


Fig. 4. Example of a PDF plot of a LR_p (red) and LR_d (green).

Apart from Tippett plots, the distributions of LR_p and LR_d can be visualised through probability density functions (PDF's, also created through Kernel Density Estimation). These are very informative as they illustrate the coherence between the two hypotheses and depict what consequences the choice of score has on the distributions. An example of a PDF and the coherence between the LR_d and the LR_p distribution can be found in Figure 4. This graph portrays the PDF's of the LR_d's (the green dashed line) and the LR_p's (the red solid line). On the x-axis are the LR's and on the y-axis is the density. The green area illustrates the FP and the red area illustrates the FN. The bigger the area, the more misleading evidence. Please bear in mind that a steep descent after the LR=1 threshold, which would still form a considerable green area right after it, is a result of KDE smoothing, which does not take the threshold into account. Line peaks indicate a higher density of those LR's in the corresponding LR distribution. The location of the peaks of the two lines are the most important. In Figure 4 there are three peaks. The first one is the green peak below the threshold of 1. This indicates that a lot of LR_d's are between 0 and 1, exactly what they should be. However, there is also a smaller green peak around LR = 7, indicating another concentration of LR_d's, which is on the wrong side of the threshold. Please bear in mind that it is hard to compare peak heights between the two threshold areas (below and above LR = 1). The higher peak of the green line on the left side of the threshold compared to the lower peak on the right side of the

threshold does not directly imply a bigger concentration of LR_d's below the threshold. Due to LR being a ratio, the density in range 0 < x < 1 is naturally higher than the density in range x > 1. This could theoretically be solved by a log transformation. Unfortunately, the LR_d and LR_p data know many zeroes and even when the same rules are applied as during the score calculation process (See Appendix II and III), the figures are often less illegible than the original figures. Peaks within a threshold area are easier to compare, with the red peak on the right indicating a higher concentration of LR_p's compared to LR_d's in that same area.

3. Results

The results for the misleading evidence can be found in Table 9 and 10 and the results for the Cost Likelihood Ratio can be found in Table 11. There are too many results to discuss all of them, however, I detected a pattern among the probability density curves and scores. Therefore I will discuss the results by score group and I will show one or two PDF's per group. The rest of the PDF's can be found in Appendices IV till VIII. An overview of the score groups can be found in Table 12.

3.1 Dataset A_1

Let us start with the LR distributions of A_1, the original dataset (see Table 8). The Euclidean LR_d curve (see Figure 5), the green curve which displays the distribution of LR's under H_d, peeks below the $LR = 1$ threshold and slowly descends above it, after which it approaches the x-axis. This comes close to what an ideal LR_d curve should be. It could improve on the False Positive area as there is still a considerable green area, but it is doing not too bad, compared to other scores. In Table 9 the FP proportions are shown for all datasets and distances, including the respective rankings. (Lowest proportion means best ranking). From this table one can conclude that the Euclidean distance ($sc13$) ranks 10th. Much worse are the Euclidean False Negatives, which is ranked last. Taking a look back at the PDF, one can see that the red area is considerably large, which holds one of the two peaks the red LR_p curve knows, the other one located around x = 4. This is obviously not a preferred characteristic of an LR_p curve, as there should only be a peak above the threshold. Other distributions that show similar patterns are those of the following scores, which I will call d_A: the Chebyshev ($sc2$), the Minkowsky L3 ($sc11$), the Tanimoto ($sc14$), the Avg(L1, Loo) ($sc15$), the Kulczynski ($sc16$), the Manhattan ($sc17$) and the Bray-Curtis ($sc18$). (All PDF's can be found in Appendices IV-VIII). All of these have relatively high FN proportions and relatively average FP proportions. Take a look at Figure 7, which illustrates the Tippett lines for the Euclidean distance. The H_p line descends already very quickly, before the $\log(LR) = 0$ threshold, which indicates high FN. The s_A's are also doing average on the $Cllr$.

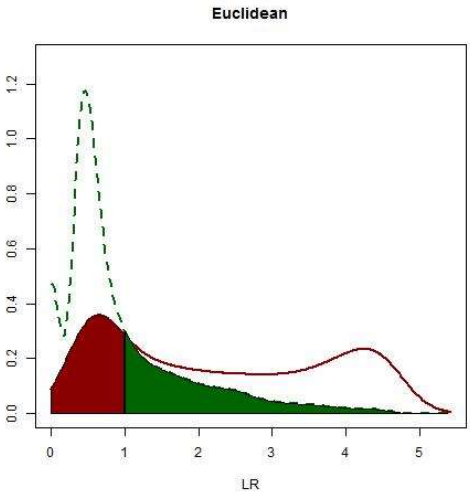


Fig.5. PDF for Euclidean distance (A_1)

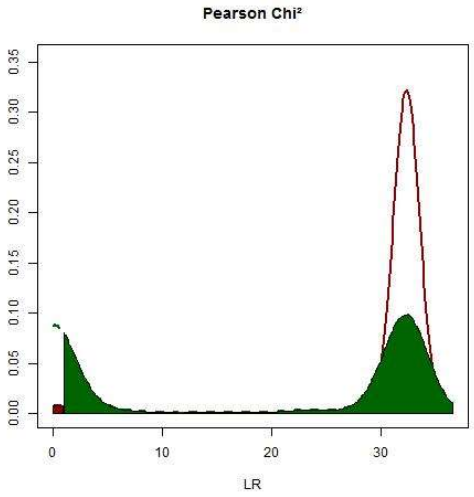


Fig. 6. PDF for Pearson χ² (A_1)

Figure 6 holds the curves for the Pearson χ² ($sc1$). The LR_d curve peaks below the threshold and in the 27 < x < 37 range. There is an evident distinction between these two peaks, with almost no density in between. The relatively major extent of the green area cannot be missed. The Pearson χ² is also ranked last for the FP propotions. The LR_p curve peaks at the same areas as the LR_d curve, yet to a lesser extent below the threshold. The Pearson χ² ranks first in the FN list. The imbalance in the FN and FP ranks is also clear in Figure 7, along with the somewhat healthier Euclidean Tippett plot. The LR_d almost vertically descends from a proportion of more than a half to 0 around $\log(LR) = 1.5$, which indicates a very high FP. Other scores with similar distributions, albeit around different LR values, are those of the following scores: The Neyman χ² ($sc3$), the K divergence ($sc6$), the Taneja ($sc7$), the Kullback-Leibler ($sc8$), the Kumar Johnson ($sc9$), the Additive Symmetric χ² ($sc10$) and the Jaccard ($sc12$). These are referred to as s_B.

Table 9

FP

I*	A_1		A_2		A_3		A_4		II**	A_5	
$sc1$	0.5736	(20)	0.7397	(19)	0.8476	(14)	0.9298	(15)	$sb1$	0.1759	(11)
$sc2$	0.3029	(13)	0.3171	(11)	0.1916	(7)	0.1411	(5)	$sb2$	0.0267	(1)
$sc3$	0.4644	(16)	0.5433	(16)	0.9305	(15)	0.913	(12)	$sb3$	0.2034	(14)
$sc4$	0.161	(2)	0.1535	(3)	0.0644	(1)	5e-04	(1)	$sb4$	0.1924	(13)
$sc5$	0.1598	(1)	0.1611	(4)	0.0645	(2)	5e-04	(1)	$sb5$	0.0883	(7)
$sc6$	0.1673	(3)	0.2088	(6)	NA		NA		$sb6$	0.0267	(1)
$sc7$	0.1676	(4)	0.2072	(5)	NA		NA		$sb7$	0.0405	(5)
$sc8$	0.348	(15)	0.4662	(15)	NA		NA		$sb8$	0.2933	(18)
$sc9$	0.3225	(14)	0.7889	(20)	NA		NA		$sb9$	0.2052	(16)
$sc10$	0.2169	(5)	0.6181	(17)	0.826	(13)	0.9144	(14)	$sb10$	0.3826	(19)
$sc11$	0.2988	(12)	0.3193	(13)	0.1797	(5)	0.187	(10)	$sb11$	0.0849	(6)
$sc12$	0.5477	(17)	0.6883	(18)	0.47	(12)	0.513	(12)	$sb12$	0.3826	(19)
$sc13$	0.2912	(10)	0.3186	(12)	0.1832	(6)	0.173	(9)	$sb13$	0.0267	(1)
$sc14$	0.2592	(6)	0.2785	(7)	0.4068	(11)	0.0002	(1)	$sb14$	0.0267	(1)
$sc15$	0.2934	(11)	0.322	(14)	0.2096	(8)	0.1593	(7)	$sb15$	0.2034	(14)
$sc16$	0.2815	(9)	0.3029	(10)	0.9997	(16)	0.9956	(16)	$sb16$	0.1759	(11)
$sc17$	0.2688	(7)	0.2942	(9)	0.2701	(10)	0.1694	(8)	$sb17$	0.1693	(10)
$sc18$	0.2688	(7)	0.2888	(8)	0.249	(9)	0.3196	(11)	$sb18$	0.1153	(8)
$sc19$	0.5515	(19)	0.0899	(1)	0.1445	(3)	0.1455	(6)	$sb19$	0.2655	(17)
$sc20$	0.5495	(18)	0.136	(2)	0.1445	(3)	0.0675	(4)	$sb20$	0.1399	(9)

Note: Ranking between brackets, 1st means best in terms of performance.

* For score names, please see appendix II (for A_1 - A_4).

** For score names, please see appendix III (for A_5).

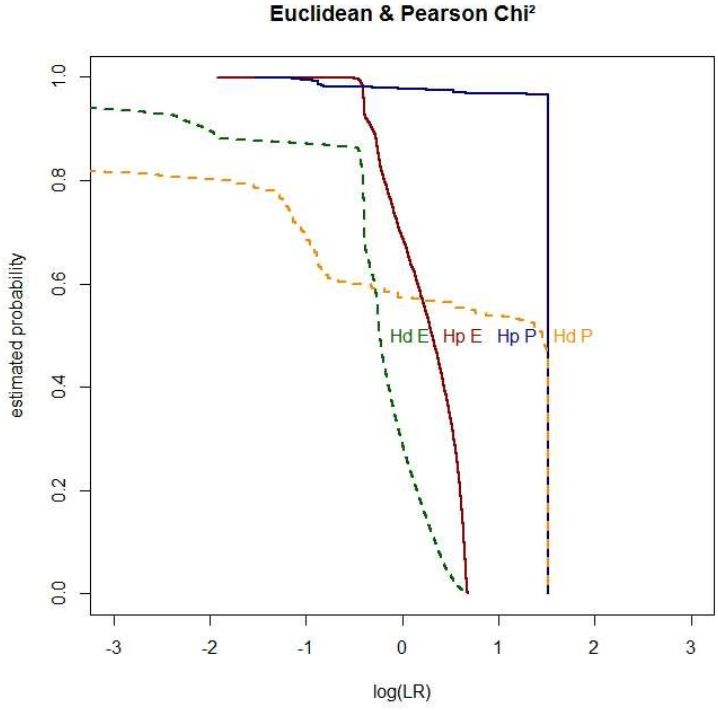


Fig. 7. Tippett plot with solid red (LR_p) and dashed green (LR_d) lines for the Euclidean distance (A_1) and solid blue (LR_p) and dashed orange (LR_d) lines for the Pearson χ^2 (A_1).

Table 10

FN

I*	A_1		A_2		A_3		A_4		II**	A_5	
$sc1$	0.0239	(1)	0.0077	(2)	0.0063	(2)	0.0067	(2)	$sb1$	0.1351	(3)
$sc2$	0.287	(14)	0.4407	(17)	0.6909	(10)	0.8084	(10)	$sb2$	0.4481	(16)
$sc3$	0.027	(2)	0.0151	(4)	0.0098	(3)	0.0154	(4)	$sb3$	0.1618	(6)
$sc4$	0.073	(5)	0.0747	(6)	0.8768	(15)	0.9986	(13)	$sb4$	0.1618	(6)
$sc5$	0.0772	(6)	0.0996	(7)	0.8768	(15)	0.9986	(13)	$sb5$	0.3418	(14)
$sc6$	0.1316	(9)	0.3821	(10)	NA		NA		$sb6$	0.4481	(16)
$sc7$	0.0839	(7)	0.0509	(5)	NA		NA		$sb7$	0.4214	(15)
$sc8$	0.0856	(8)	0.2284	(9)	NA		NA		$sb8$	0.2772	(12)
$sc9$	0.0344	(3)	0.0039	(1)	NA		NA		$sb9$	0.1351	(3)
$sc10$	0.0442	(4)	0.013	(3)	0.0102	(4)	0.0105	(3)	$sb10$	0.0877	(1)
$sc11$	0.3088	(19)	0.4407	(17)	0.7095	(12)	0.7561	(7)	$sb11$	0.3046	(13)
$sc12$	0.1453	(10)	0.1579	(8)	0.3561	(5)	0.3709	(4)	$sb12$	0.0877	(1)
$sc13$	0.3105	(20)	0.4333	(16)	0.7014	(11)	0.7674	(8)	$sb13$	0.4481	(16)
$sc14$	0.2947	(18)	0.4189	(14)	0.5221	(6)	0.9996	(15)	$sb14$	0.4481	(16)
$sc15$	0.2888	(17)	0.427	(15)	0.6691	(9)	0.7842	(9)	$sb15$	0.1618	(6)
$sc16$	0.2772	(13)	0.3961	(11)	0	(1)	0.0004	(1)	$sb16$	0.1351	(3)
$sc17$	0.287	(14)	0.4053	(12)	0.5902	(7)	0.7404	(6)	$sb17$	0.1618	(6)
$sc18$	0.287	(14)	0.4088	(13)	0.6204	(8)	0.574	(5)	$sb18$	0.6667	(20)
$sc19$	0.1586	(12)	0.9235	(20)	0.8702	(13)	0.8481	(11)	$sb19$	0.1719	(10)
$sc20$	0.154	(11)	0.8895	(19)	0.8705	(14)	0.9305	(12)	$sb20$	0.2456	(11)

Note: Ranking between brackets, 1st means best in terms of performance.

* For score names, please see appendix II (for A_1 - A_4).

** For score names, please see appendix III (for A_5).

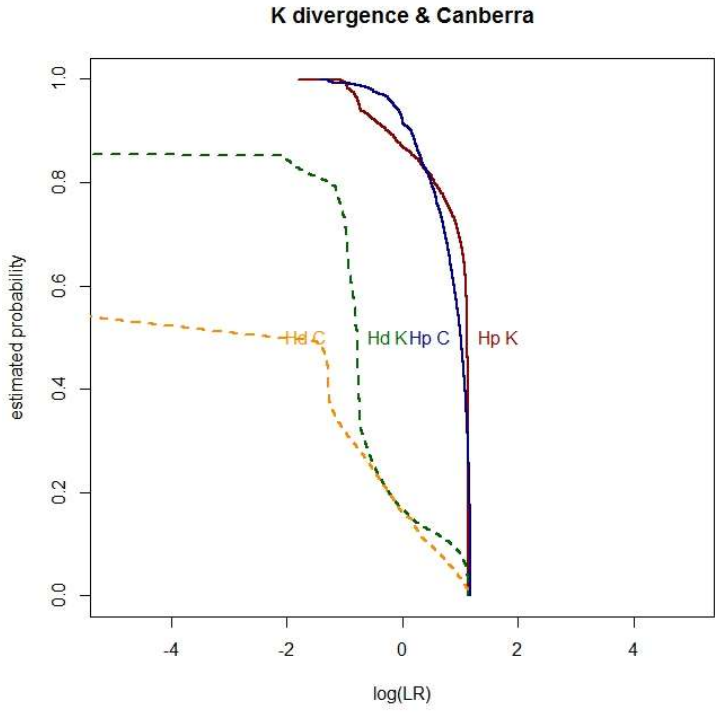


Fig. 8. Tippett plot with solid red (LR_p) and dashed green (LR_d) lines for the K divergence (A_1) and solid blue (LR_p) and dashed orange (LR_d) lines for the Canberra distance (A_1).

Table 11

$Cllr$

I*	A_1		A_2		A_3		A_4		II**	A_5	
$sc1$	1.4695	(20)	0.992	(18)	0.9543	(1)	1.0292	(11)	$sb1$	0.552	(1)
$sc2$	0.8004	(13)	0.9367	(13)	0.9869	(6)	0.9964	(7)	$sb2$	0.6019	(7)
$sc3$	1.1288	(19)	1.0103	(19)	1.2827	(15)	1.8284	(16)	$sb3$	0.5945	(5)
$sc4$	0.4235	(2)	0.4496	(1)	1.0893	(14)	1.0655	(12)	$sb4$	0.6095	(8)
$sc5$	0.4134	(1)	0.4729	(2)	1.0887	(13)	1.0655	(12)	$sb5$	0.6813	(12)
$sc6$	0.5489	(3)	0.9397	(14)	NA		NA		$sb6$	0.9957	(19)
$sc7$	0.5583	(4)	0.5146	(3)	NA		NA		$sb7$	0.6181	(11)
$sc8$	0.6599	(6)	0.8963	(9)	NA		NA		$sb8$	0.8153	(17)
$sc9$	0.9187	(18)	1.0809	(20)	NA		NA		$sb9$	0.5819	(4)
$sc10$	0.6312	(5)	0.8412	(4)	0.9905	(9)	1.1059	(14)	$sb10$	0.7475	(15)
$sc11$	0.8075	(14)	0.9316	(12)	0.9876	(8)	0.9957	(6)	$sb11$	0.7271	(14)
$sc12$	0.9095	(17)	0.9582	(15)	0.9719	(2)	0.9842	(1)	$sb12$	0.7508	(16)
$sc13$	0.8003	(12)	0.9205	(10)	0.987	(7)	0.9951	(4)	$sb13$	0.6179	(10)
$sc14$	0.7623	(7)	0.8749	(5)	1.031	(12)	1.1166	(15)	$sb14$	1.0059	(20)
$sc15$	0.7911	(11)	0.9251	(11)	0.9857	(4)	0.9952	(5)	$sb15$	0.6017	(6)
$sc16$	0.768	(10)	0.8798	(7)	22.2835	(16)	0.9967	(8)	$sb16$	0.5596	(2)
$sc17$	0.7648	(8)	0.8839	(8)	0.9816	(3)	0.9902	(3)	$sb17$	0.573	(3)
$sc18$	0.7648	(8)	0.8771	(6)	0.986	(5)	0.9888	(2)	$sb18$	0.8537	(18)
$sc19$	0.8921	(15)	0.9851	(17)	1.0126	(10)	1.0031	(10)	$sb19$	0.6922	(13)
$sc20$	0.8978	(16)	0.9695	(16)	1.0126	(10)	1.0025	(9)	$sb20$	0.6177	(9)

Note: Ranking between brackets, 1st means best in terms of performance.

* For score names, please see appendix II (for A_1 - A_4).

** For score names, please see appendix III (for A_5).

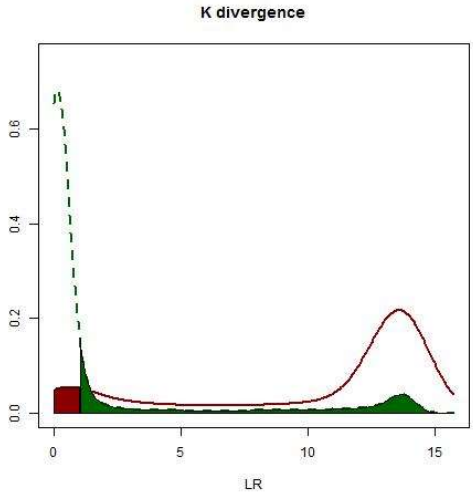


Fig. 9. PDF for K Divergence (A_1)

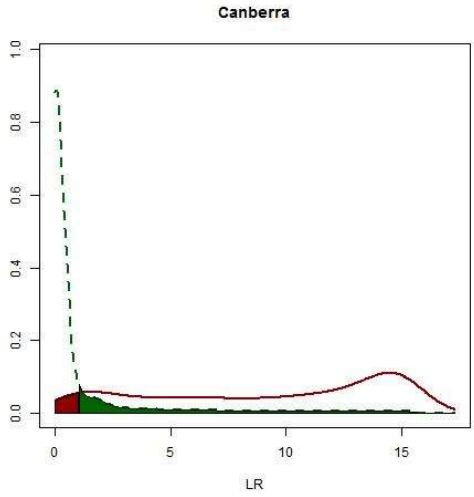


Fig. 10. PDF for Canberra distance (A_1)

Table 12

Overview of score groups

	Description	A_1
s_A	PDF similar to Figure 5. Performance is average.	$sc2, sc11,$ $sc13 - sc18$
s_{B1}	PDF similar to Figure 6. Performance is poor.	$sc1, sc3,$ $sc9, sc12$
s_{B2}	PDF similar to Figure 9. Performance is good.	$sc6 - sc8$ $sc10$
s_C	PDF similar to Figure 10. Performance is best.	$sc4$ & $sc5$
s_D	PDF similar to Figure 11. Performance is poor.	$sc19$ & $sc20$

Although the s_B score PDF's seem to be very much alike, they differ substantially in terms of performance. For example, the Pearson χ^2 is doing worst with a $Cllr$ of 1.4695, whereas the K divergence is ranked third with a $Cllr$ of .5489 and the PDF looks much healthier (see Figure 9). The s_B scores can be subdivided into s_{B1} and s_{B2}. The Pearson χ^2, the Neyman χ^2, the Kumar Johnson and the Jaccard belong to the former and the K divergence, the Taneja and the Kullback-Leibler belong to the latter. The s_{B1} score have in common that they all use powers in their equations, yet they do not consider roots (which for example the Euclidean distance does). This results in an asymmetry, which does not favour the performance. The FP proportions are scoring worst and the FN proportions are scoring best and dominates the top 3. The low FN proportions, however, could not prevent these scores having the 4 highest $Cllr$ scores. The s_{B2} scores are doing much better, with both low FP scores and FN scores. The $Cllr$ scores make the s_{B2} scores the second most successful. These scores do not use powers and instead they use natural logs in their respective equations. The Additive Symmetric χ^2 does not belong to s_{B1}, as the asymmetry is solved. Despite lacking in the use of a natural log, in terms of performance the score should be assigned to s_{B2}.

The next group, the s_C scores, seem to perform the best. The Canberra curves ($sc5$) can be found in Figure 10 and only the Clark distance ($sc4$) generates similar patterns. The LR_d curve peaks below the threshold and little significant green area is visible. The LR_p curve is quite flat and almost uniform, still favouring the correct side of the threshold. They rank best in FP and $Cllr$ and still very well in FN and if it weren't for the unfair asymmetry of the s_{B1} scores, they would have done even better. One common feature, which they share with the s_B's (except for the Jaccard), is the level of normalisation, which is different from the s_A's. The division in the s_C's takes place on the individual level, implying that the division is done before the summation, whereas for the s_A's the division is

done after the summation(s). This is, for instance, the only distinction between the Canberra and the Bray-Curtis distance.

The two remaining scores, the d_D scores, are the Pearson's ρ ($sc19$) and the Cosine similarity ($sc20$). The equations appear complicated, but by the use of some math tricks, one can easily prove that these scores are fairly similar. Hence the similar curves (see Figure 11 for the Cosine similarity). The LR_d curve has three peaks, of which one is above the threshold, located at the same spot as one of the LR_p peaks. The other LR_p peak is below the threshold. Seemingly, these scores are not of great value.

The results for the first dataset show a considerable diversity in score behaviour and performance. They also suggest that the d_C scores, the Canberra and the Clark distances, have the best performances. Whether or not this holds for the remaining datasets as well, will be discussed next.

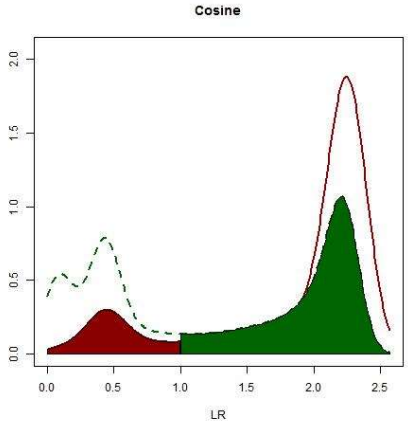


Fig. 11. PDF for Cosine similarity (A_1)

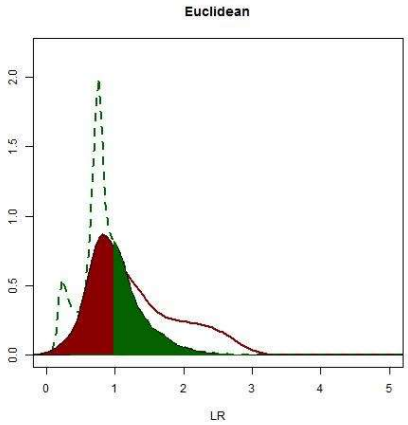


Fig. 12. PDF for Euclidean distance (A_2)

3.2 Dataset A_2

The addition of more extreme values to the A_2 data, has changed the distributions of the LR. Figure 12 shows the PDF's for the LR_d and the LR_p for the Euclidean distance. Compared to Figure 5 the density appears to be shifted more towards the threshold, suggesting that the LR weakens, which might not come as a surprise, considering that the original data is not as coherent anymore. All of the s_A scores for dataset A_2 have reacted similarly as for dataset A_1. The LR_p curve above the threshold is flatter and both curves below the threshold are higher compared to the pdfs under A_1. Apart from the expected rise in FP and FN proportions, the PDF's have not remarkably changed. In fact, the $Cllr$ of these scores have improved from an average ranking of 10.375 in A_1 to 9 in A_2. The s_{B1} score PDF's preserve their shapes, except for the Jaccard, which looks slightly more like a d_D.

Just like the s_A scores, the d_{B1} peaks have become higher and the LRs higher than 1 have decreased. The $Cllr$ scores are still as poor, ranking once again last (Jaccard is ranking 15th). Also, the s_{B2} scores appear to have been affected differently among each other. The K divergence now looks more like a s_A , with which the LR_p peak is centred around 1 and it is now dropping eminently from 3rd to 14th place in the $Cllr$ ranking. The Kullback-Leibler turns out to be more like a s_D and has undergone the same shift towards the threshold as was seen before. Only the Taneja and the Additive Symmetric χ^2 come out well. Despite the lower LR's above the threshold, the shapes remain relatively similar, keeping the peaks on the right sides, except for that LR_d bump right under the LR_p peak. The Taneja ranks 3rd in the $Cllr$ and the Additive Symmetric χ^2 ranks 4th.

All data considered, the Canberra and the Clark distance appear to be most suited again. The curves are a bit jerkier above the threshold, yet the peaks have not become as sharp as for the previously discussed scores. The FP and FN proportions are still relatively low.

3.3 Dataset $A_3 - A_5$

The conclusion for the negative data, A_3 and A_4, can be very short: All of the scores are performing relatively bad. Many of the $Cllr$ scores, 15 out of 32, are above the acceptable value of 1 and the misleading evidence proportions are relatively high. The lowest $Cllr$ is .9543 for A_3 and .9842 for A_4. (See Table 11). The PDF's often depict overlapping curves or sometimes behave fairly strange, e.g. Kulczynski distance for A_3 and the Clark distance for A_4. Despite the evasive results, the conclusion that either a more appropriate score measure should be found and negative data should be treated differently and more carefully is valuable.

The typical s_{B1} curves are also present in the PDF's for A_5, the binary data. In Figure 13 the Simpson similarity curves ($sb12$) look like the s_{B1} curves; highest peaks on the right side of the threshold with almost no density in between and a smaller faulty located peak beneath them. The other three similarities that are comparable are the Sokal&Sneath-III ($sb6$), the Peirce ($sb8$) and the Yule W ($sb10$). Together with the Kulczynski-I ($sb14$) and the Forbes-I similarity ($sb18$), these are the six worst ranked in the $Cllr$ list, having a score of more than .7475. There is another group of scores that show similar PDF's, with one major difference. (See Figure 14 for the Manhattan PDF). The overall FP area is usually smaller, yet, there is an additional LR_d peak just after the threshold. These scores include the Manhattan ($sb1$), the Euclidean ($sb2$), the Bray-Curtis ($sb3$), the Stiles ($sb4$), the Pearson-I ($sb7$), the Pearson-III ($sb9$), the Braun&Banquet ($sb13$), the Jaccard ($sb15$), the Driver&Kroeber ($sb17$) and the Sokal&Sneath-I ($sb20$). The Anderberg ($sb5$), Tarantula ($sb11$) and Eyraud ($sb16$) similarities are similar, yet the area above the threshold is more jerky. Together these

dominate the top 14 in the $Cllr$ ranking. Most of the successful ones seem to lack the d in their respective equations, which is when i and j are both 0. If the average results for scores that use d (average rank = 11.33, average $Cllr$ = .704) are compared to those that do not (average rank = 9.81, average $Cllr$ = .6783) with a simple independent t-test, no significant results are found (t = -.411, df = 17.3, p = .686). Apparently, the inclusion or exclusion make no difference and do not necessarily determine the shape of the curves.

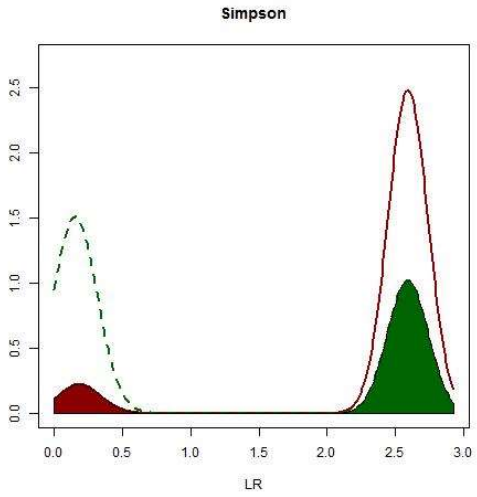


Fig.13. PDF Simpson similarity (A_5)

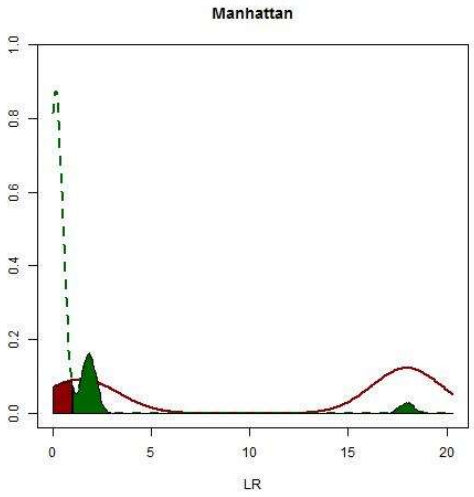


Fig.14. PDF Manhattan distance (A_5)

4. Discussion

My intentions were that at the end of this thesis I would have illustrated that score selection is a vital part of the score-based method construction phase and I would have provided insight into the appropriateness of scores for specific types of data. I have accomplished this by systematically testing different score methods on a real dataset and transformed versions of it and by analysing the performances based on relative statistics and graphical representations. The first intention has evidently been proven. There is much variation between the scores in terms of performance and anticipated results. The Canberra and Clark distances are outperforming the other groups of scores in terms of discriminating power ($Cllr$) and LR distributions that they produce. The variation between scores is compatible with the score experiments in 2.1. In just one experiment, which applies 4 scores in a very simple dataset, there is already difference in performance with respect to the anticipated result. There is also much variation between the datasets, which indicates that special attention should be devoted to the appropriateness of the chosen score in relation to the dataset that is worked with. The variation between the datasets is shown in the experiments of 2.1, in which the same variation between scores in different experiments explains that not every score will perform well in every situation. Relatively, the Euclidean distance, which is considered to be the most conventional one, did not appear to be very successful overall. Ertöz et al. (2003), Aggarwal et al. (2001) and Troyanskaya et al. (2001), as discussed in chapter 1, already told that the Euclidean distance is not robust against data difficulties. The original data contained many zeroes, which has been an obstacle for the Euclidean distance, and the transformations with added difficulties have not been tackled well by the Euclidean distance either. There is much to gain if proper score selection were not to be neglected and more conventional and unconventional scores would be considered more. A more suitable score leads to a more valid and reliable method, which is undoubtedly crucial in the forensic field and part of the objectification of strength of evidence. Moreover, as the LR is to be interpreted and put into context by other juridical parties, the LR should be handed to them without too many complicated caveats that may confuse them even more. The score-based approach method is by many authors acclaimed to be very promising and devoting more effort to the improvement of the method is worth considering.

The second part of the question was: Which score should be used for specific types of data? The Clark and Canberra distances seem to be most appropriate for this thesis' data, which was continuous data, and the distances are relatively robust against extreme values. For binary data, the Manhattan distance is most appropriate. The results show that powers easily enhance distortions and are best accounted for by roots. The Clark distance, for example, is doing better than the

Pearson and the Neyman χ^2. In addition, normalisation should take place on the individual level. The Canberra distance is doing better than the Bray-Curtis distance. Both apply the same formula, yet the former normalises on the individual level and the latter on all variables at once. Therefore, it is wise to reconsider opting for a score that takes these two remarks about powers and normalisation into account. This, however, may be true for the types of data that have been used for this thesis, whereas other, easier types may even be beneficial for distances as the Pearson and the Neyman χ^2 and the Bray-Curtis. The Bray-Curtis, for instance, may flourish in count data as what it was originally designed for. (Looman & Campbell, 1960).

Although these are some conclusive remarks, contrarily to the first part of my intentions, the second part does not have a conclusive answer overall. The research question encompasses many facets of the score-based likelihood method, i.e. many scores and data to choose from. There is not one universal score that can be classified as the best score for score-based LR methods which could be appropriate for all kinds of data, although the Clark and Canberra for continuous data and the Manhattan for binary data are advisable choices.

There is yet a score to be found that is applicable for negative data. The results show that LR distributions become very corrupt and misleading evidence proportions are very high. None of the scores applied in this thesis could accurately cope with the complications that negative data introduces. Because A_4 was a combination of extreme values and negative data, little could be said about the interaction effect, as these results were even more distorted. It is not unexpected that this type of data is difficult to manage. However, the selection of 20 scores was not sufficient to find at least one acceptable score. Research into the mathematics of other scores should help to identify the right score. Another surprising result was that the LR's remained relatively low. Hardly could they ever get over a LR of 50, which is only in the second lowest verbal scale that is maintained by the NFI. There are two main reasons that account for this. Firstly, it is most probably a result of the quality of the data, which forms score distributions (E_w and E_b) that are rather similar. Consequently, the numerator and the denominator of the LR differ to a lesser extent, resulting in low LR's. Secondly, in general score-based approached to the LR method tend to provide lower LR's. Since all information is reduced to 1 score, there is a loss of information which leads to a decrease in the strength of evidence.

As the results lead to the conclusion that score performance relies on the type of data, this is in turn a caution for the rest of the conclusions that could be drawn. The results may differ in other types of data and more research is necessary. During the final stage of this thesis, the procedure was also run for another type of real data that has just recently been made accessible by the NFI, used in drugs

comparisons. Some results were similar, for example the Canberra and the Clark distance doing relatively well and the s_C scores doing relatively poor. Other results were very different, e.g. the Pearson's ρ succeeded remarkably, which is in contrast with this thesis' results. It indicates that much is left to be discovered on the behaviour of scores in an environment as the forensic fields that holds a considerable diversity of data types. This research could be repeated with other types of real data, such as categorical data, non-sparse data, high-dimensional data and oddly distributed data. Additionally, simulated data enables the regeneration of results. A systematic and repetitive process will provide the forensic world with more valuable information on the behaviour of scores in the score-based LR method.

This thesis is one of the first works that has systematically tested several score-based likelihood ratio methods and the results are rewarding. It has shown that there is variability in performance between a handful of score-based likelihood ratio methods and between a handful of data settings. Hopefully, more researchers will continue to bring in more knowledge of score behaviour and performance in various likelihood ratio method environments.

List of References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. Database Theory — ICDT 2001 Lecture Notes in Computer Science, 420-434.

Aitken, C. G. (1995). Statistics and the evaluation of evidence for forensic scientists. Chichester: J. Wiley.

Aitken, C. G., & Stoney, D. A. (1991). The use of statistics in forensic science. New York: E. Horwood.

Aitken, C. G., & Taroni, F. (2004). Statistics and the evaluation of evidence for forensic scientists. Chichester, England: Wiley.

Baiker, M., Keereweer, I., Pieterman, R., Vermeij, E., Weerd, J. V., & Zoon, P. (2014). Quantitative comparison of striated toolmarks. Forensic Science International, 242, 186-199.

Cha, S. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences, 1(4), 300-307.

Chazal, P. D., Flynn, J., & Reilly, R. (2005). Automated processing of shoeprint images based on the Fourier transform for use in forensic science. IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE Trans. Pattern Anal. Machine Intell., 27(3), 341-350.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics, 8(1), 43-48.

Davis, L. J., Saunders, C. P., Hepler, A., & Buscaglia, J. (2012). Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. Forensic Science International, 216(1-3), 146-157.

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

Esseiva, P., Dujourdy, L., Anglada, F., Taroni, F., & Margot, P. (2003). A methodology for illicit heroin seizures comparison in a drug intelligence perspective using large databases. Forensic Science International, 132(2), 139-152.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. Proceedings of the 2003 SIAM International Conference on Data Mining, 47-58.

Evett, I. (1998). Toward a uniform framework for reporting opinions in forensic science casework. Sci. Just, 38(3), 198-202.

Finkelstein, M. O., & Levin, B. (1990). Statistics for lawyers. New York: Springer-Verlag.

Gastwirth, J. L. (2000). Statistical science in the courtroom. New York: Springer.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2007). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech & Language, 20(2-3), 331-355.

Hepler, A. B., Saunders, C. P., Davis, L. J., & Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. Forensic Science International, 219(1-3), 129-140.

Horswell, J., Cordiner, S. J., Maas, E. W., Martin, T. M., Sutherland, K. B., Speir, T. W., . . . Osborn, A. M. (2002). Forensic Comparison of Soils by Bacterial Community DNA Profiling. Journal of Forensic Sciences J. Forensic Sci., 47(2).

Inoue, H., Kanamori, T., Iwata, Y. T., Ohmae, Y., Tsujikawa, K., Saitoh, S., & Kishi, T. (2003). Methamphetamine impurity profiling using a 0.32 mm i.d. nonpolar capillary column. Forensic Science International, 135(1), 42-47.

Kinder, J. D., & Bonfanti, M. (1999). Automated comparisons of bullet striations based on 3D topography. Forensic Science International, 101(2), 85-93.

Leeuwen, D. A., & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. Lecture Notes in Computer Science Speaker Classification I, 330-353.

Lociciro, S., Hayoz, P., Esseiva, P., Dujourdy, L., Besacier, F., & Margot, P. (2007). Cocaine profiling for strategic intelligence purposes, a cross-border project between France and Switzerland. Forensic Science International, 167(2-3), 220-228.

Looman, J., & Campbell, J. B. (1960). Adaptation of Sorensen's K (1948) for Estimating Unit Affinities in Prairie Vegetation. Ecology, 41(3), 409-416.

Marquis, R., Weyermann, C., Delaporte, C., Esseiva, P., Aalberg, L., Besacier, F., . . . Zrcek, F. (2008). Drug intelligence based on MDMA tablets data. Forensic Science International, 178(1), 34-39.

Meuwly, D. (2006). Forensic Individualisation from Biometric Data. Science & Justice, 46(4), 205-213.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice, 51(3), 91-98.

National Research Council of the National Academies, Strengthening Forensic Science in the United States: A Path Forward. The National Academies Press, Washington. 2009.

Neumann, C., Evett, I. W., & Skerrett, J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm. Journal of the Royal Statistical Society: Series A (Statistics in Society), 175(2), 371-415.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D., & Bromage-Griffiths, A. (2006). Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Three Minutiae. Journal of Forensic Sciences, 51(6), 1255-1266.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2006). Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae. Journal of Forensic Sciences, 52(1), 54-64.

Neumann, C., & Margot, P. (2009). New perspectives in the use of ink evidence in forensic science. Forensic Science International, 185(1-3), 38-50.

Nordgaard, A., & Höglund, T. (2011). Assessment of Approximate Likelihood Ratios from Continuous Distributions: A Case Study of Digital Camera Identification*. Journal of Forensic Sciences, 56(2), 390-402.

Pervouchine, V., & Leedham, G. (2007). Extraction and analysis of forensic document examiner features used for writer identification. Pattern Recognition, 40(3), 1004-1013.

Pierrini, G., Doyle, S., Champod, C., Taroni, F., Wakelin, D., & Lock, C. (2007). Evaluation of preliminary isotopic analysis (13C and 15N) of explosives A likelihood ratio approach to assess the links between semtex samples. Forensic Science International, 167(1), 43-48.

Quaak, F. C., & Kuiper, I. (2011). Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons. Forensic Science International, 210(1-3), 96-101.

Robertson, B., & Vignaux, G. A. (1995). Interpreting evidence: Evaluating forensic science in the courtroom. Chichester: J. Wiley.

Sjerps, M. (2004). Onderzoek Forensische Statistiek. NAW, 5(2), 4-9.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., . . . Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.

Appendix I: Verbal Scale of LR

Verbal Scale of Likelihood Ratios maintained by NFI (2016)

Values* of LR	**Verbal equivalent (two options of phrasing are suggested)**
1-2	The forensic findings provide no assistance in addressing the issue.
2 – 10	The forensic findings are slightly more probable given one proposition relative to the other.
10 – 100	…are more probable given…proposition...than proposition...
100 – 10,000	…are much more probable given… proposition...than proposition...
10,000 – 1,000,000	…are very much more probable given… proposition...than proposition...
1,000,000 and above	…are extremely more probable given… proposition...than proposition...

* Likelihood ratios corresponding to the inverse (1/X) of these values (X) will express the degree of support for the specified alternative compared to the first proposition.

Appendix II: Continuous Scores

In-text the continuous scores are referred to as sc's. All formulas in this list are distances, except for the Cosine similarity ($sc20$). X and Y are the two measurements being compared to each other and n is the number of variables a measurement has, which is 7 in this thesis.

1. Pearson χ^2 $\sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{Y_i}$
2. Chebyshev $max_i|X_i - Y_i|$
3. Neyman χ^2 $\sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{X_i}$
4. Clark $\sqrt{\sum_{i=1}^{n} \left(\frac{|X_i - Y_i|}{X_i + Y_i}\right)^2}$
5. Canberra $\sum_{i=1}^{n} \frac{|X_i - Y_i|}{X_i + Y_i}$
6. K divergence $\sum_{i=1}^{n} X_i \ln \frac{2X_i}{X_i + Y_i}$
7. Taneja $\sum_{i=1}^{n} \left(\frac{X_i + Y_i}{2}\right) \ln \left(\frac{X_i + Y_i}{2\sqrt{X_i Y_i}}\right)$
8. Kullback-Leibler $\sum_{i=1}^{n} X_i \ln \frac{X_i}{Y_i}$
9. Kumar Johnson $\sum_{i=1}^{n} \left(\frac{\left(X_i^2 - Y_i^2\right)^2}{2(X_i Y_i)^{\frac{3}{2}}}\right)$
10. Additive Symmetric χ^2 $\sum_{i=1}^{n} \frac{(X_i - Y_i)^2 (X_i + Y_i)}{X_i Y_i}$
11. Minkowski L_3 $\sqrt[3]{\sum_{i=1}^{n} |X_i - Y_i|^3}$
12. Jaccard $\frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} X_i Y_i}$
13. Euclidean $\sqrt{\sum_{i=1}^{n} |X_i - Y_i|^2}$

14. Tanimoto $\frac{\sum_{i=1}^{n} X_i+\sum_{i=1}^{n} Y_i-2\sum_{i=1}^{n}\min(X_i,Y_i)}{\sum_{i=1}^{n} X_i+\sum_{i=1}^{n} Y_i-\sum_{i=1}^{n}\min(X_i,Y_i)}$

15. Avg(L_1, L_∞) $\frac{\sum_{i=1}^{n}|X_i-Y_i|+ma\ _i|X_i-Y_i|}{2}$

16. Kulczynski $\frac{\sum_{i=1}^{n}|X_i-Y_i|}{\sum_{i=1}^{n}\min(X_i,Y_i)}$

17. Manhattan $\sum_{i=1}^{n}|X_i - Y_i|$

18. Bray-Curtis (Sørenson) $\frac{\sum_{i=1}^{n}|X_i-Y_i|}{\sum_{i=1}^{n}(X_i+Y_i)}$

19. Pearson ρ $\frac{1-\frac{\sum_{i=1}^{n}\left(X_i-\frac{\sum X}{N}\right)\left(Y_i-\frac{\sum Y}{N}\right)}{\sqrt{\left(\left(X_i-\frac{\sum X}{N}\right)^2\right)\left(\left(Y_i-\frac{\sum Y}{N}\right)^2\right)}}}{2}\cdot 100$

$\frac{1-\frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}}{2}\cdot 100$

20. Cosine $\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$

Please note that sometimes data causes technical problems for some scores. The following issues were found and the corresponding treatments were implemented:

1) 0 divided by 0; This is treated as 0
2) Anything else divided by 0; The 0 is replaced by $1e^{-04}$
3) 0 log0; This is treated as 0
4) Log of 0; The 0 is replaced by $1e^{-04}$

Appendix III: Binary Scores

In-text the continuous scores are referred to as sc's. The first 3 formulas in this list are distances and the rest are similarity measures. The two measurements X and Y are first transformed into a, b, c and d. Here a is the number of variables that are both present in X and Y, b is the number of variables that is only present in X, c is the amount of variables that is only present in Y and d is the amount of variables that are absent in both X and Y. Because there are 7 variables, $a + b + c + d = 7$, which is in turn n.

1. Manhattan $b + c$
2. Euclidean $\sqrt{b + c}$
3. Bray-Curtis $\frac{b+c}{2a+b+c}$
4. Stiles $\log_{10} \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$
5. Anderberg $\frac{\sigma-\sigma'}{2n}$

   $\sigma = \max(a,b) + \max(c,d) + \max(a,c) + \max(b,d)$

   $\sigma' = \max(a+c, b+d) + \max(a+b, c+d)$
6. Sokal & Sneath-III $\frac{a+d}{b+c}$
7. Pearson-I $\frac{n(ad-b\ )^2}{(a+b)(a+c)(b+d)(c+d)} = \chi^2$
8. Peirce $\frac{ab+bc}{ab+2bc+cd}$
9. Pearson-II $\sqrt{\frac{\chi^2}{n+\chi^2}}$
10. Yule w $\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
11. Tarantula $\frac{a(c+d)}{c(a+b)}$

12. Simpson $\frac{a}{\min(a+b,a+c)}$

13. Braun & Banquet $\frac{a}{\max(a+b,a+c)}$

14. Kulczynski-I $\frac{a}{b+c}$

15. Jaccard $\frac{a}{a+b+c}$

16. Eyraud $\frac{n^2(na \quad (a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$

17. Driver & Kroeber $\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$

18. Forbes-I $\frac{na}{(a+b)(a+c)}$

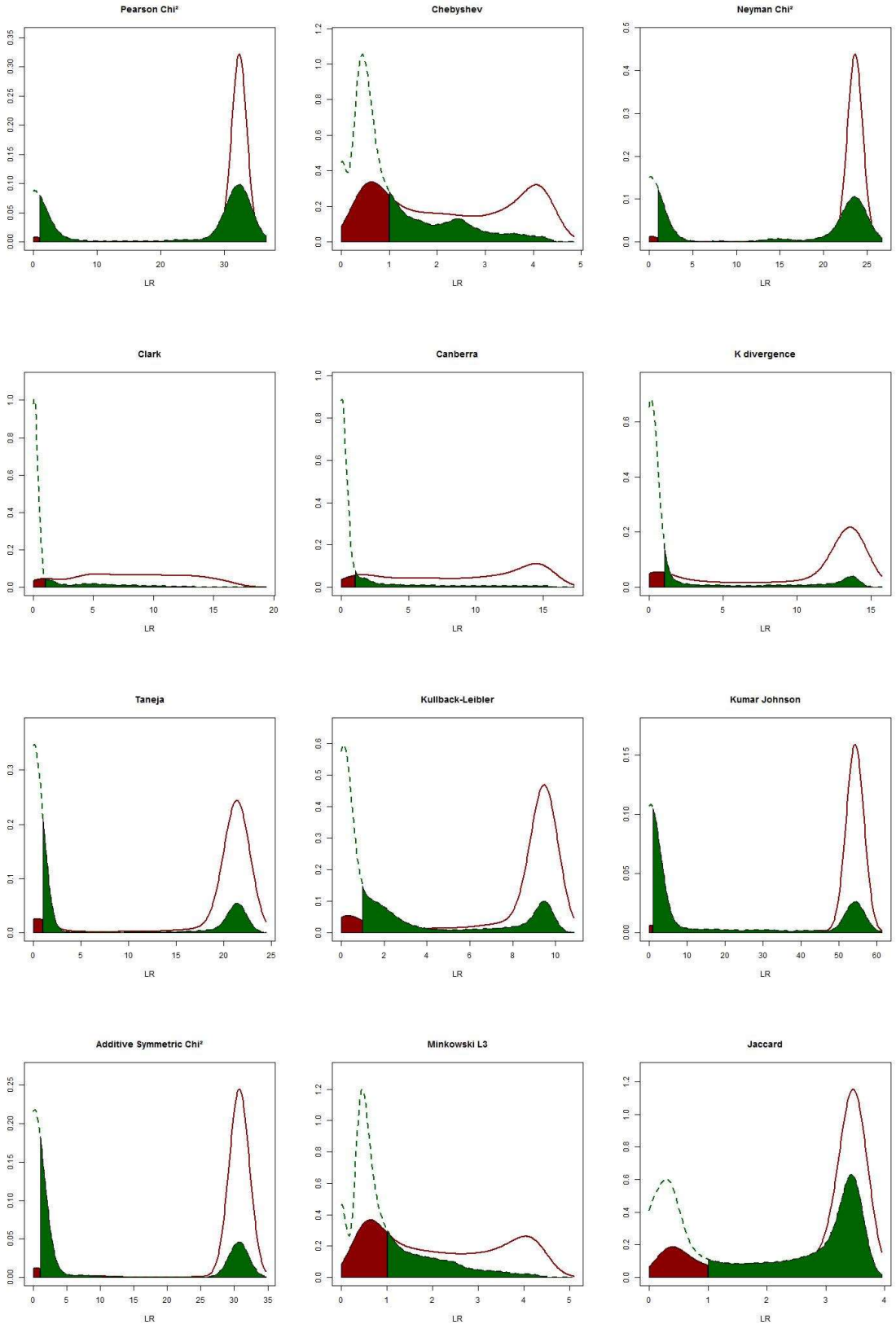
19. Fager & McGowan $\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{\max(a+b,a+c)}{2}$

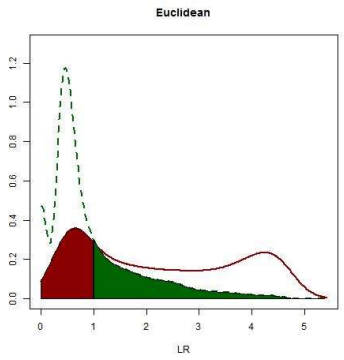
20. Sokal & Sneath-I $\frac{a}{a+2b+2c}$

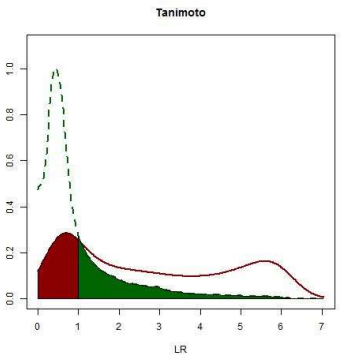
Please note that sometimes data causes technical problems for some scores. The following issues were found and the corresponding treatments were implemented:

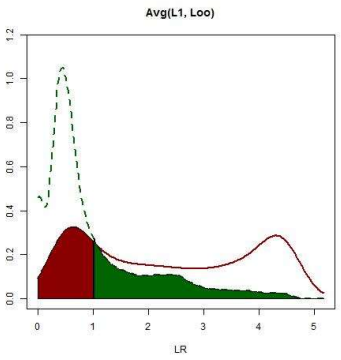
1)	0 divided by 0;	This is treated as 0
2)	Anything else divided by 0;	The 0 is replaced by $1e^{-04}$
3)	0 log0;	This is treated as 0
4)	Log of 0;	The 0 is replaced by $1e^{-04}$

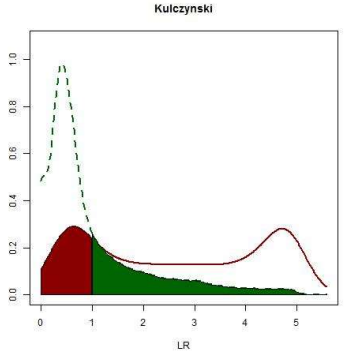
Appendix IV: PDF's of LR's for A_1

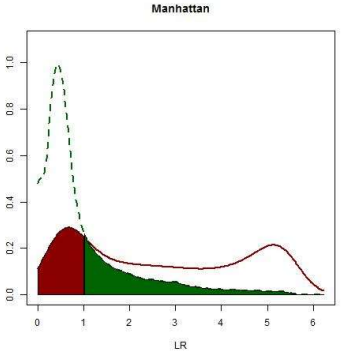


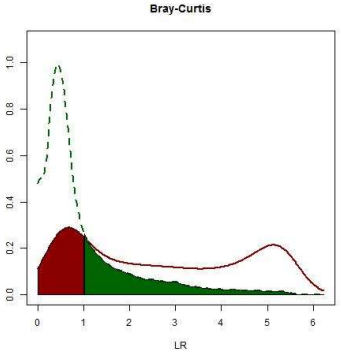

Euclidean
LR

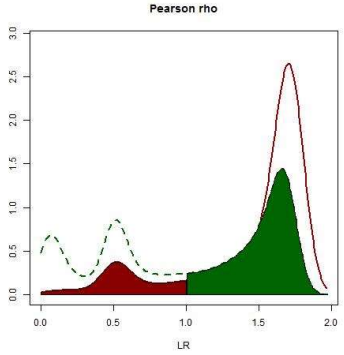

Tanimoto
LR

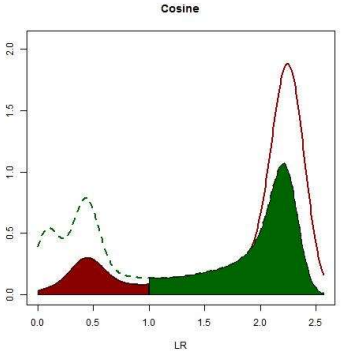

Avg(L1, Loo)
LR


Kulczynski
LR

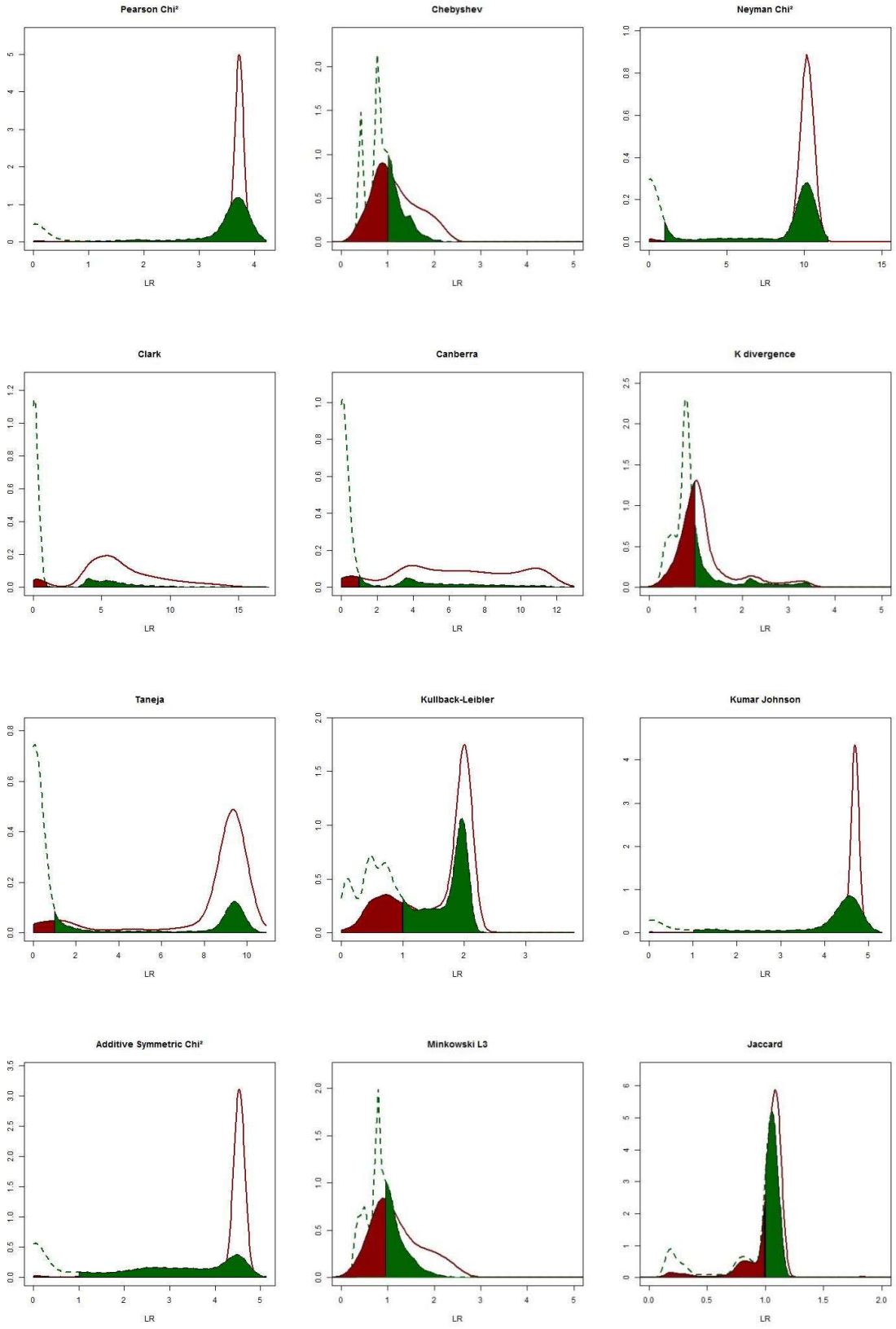

Manhattan
LR

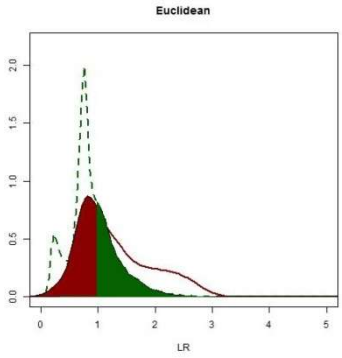

Bray-Curtis
LR

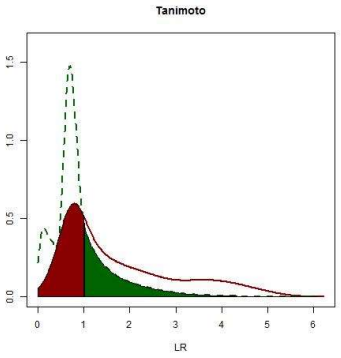

Pearson rho
LR

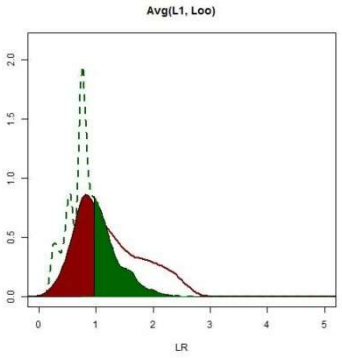

Cosine
LR

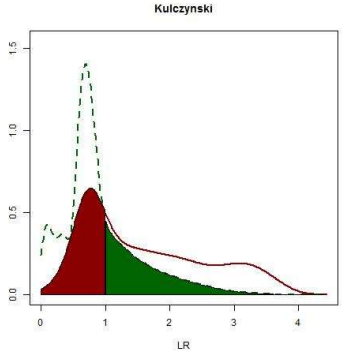
Appendix V: PDF's of LR's for A_2

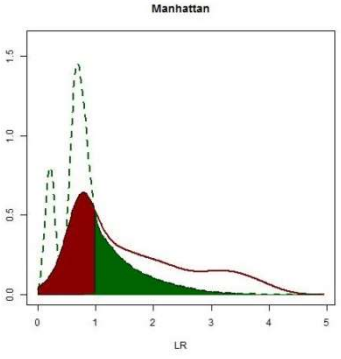


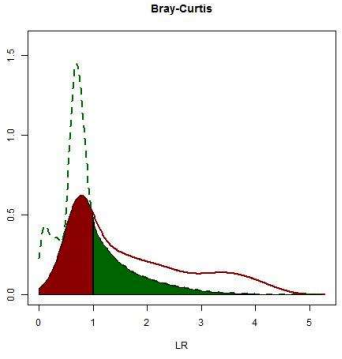

Euclidean
LR

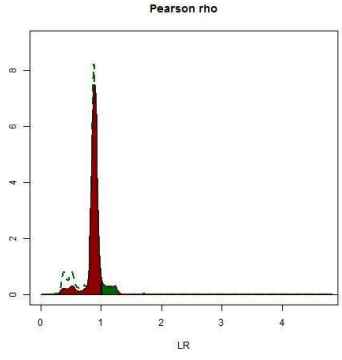

Tanimoto
LR

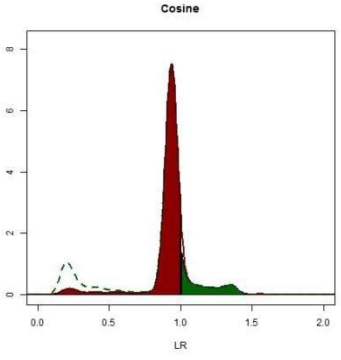

Avg(L1, Loo)
LR


Kulczynski
LR

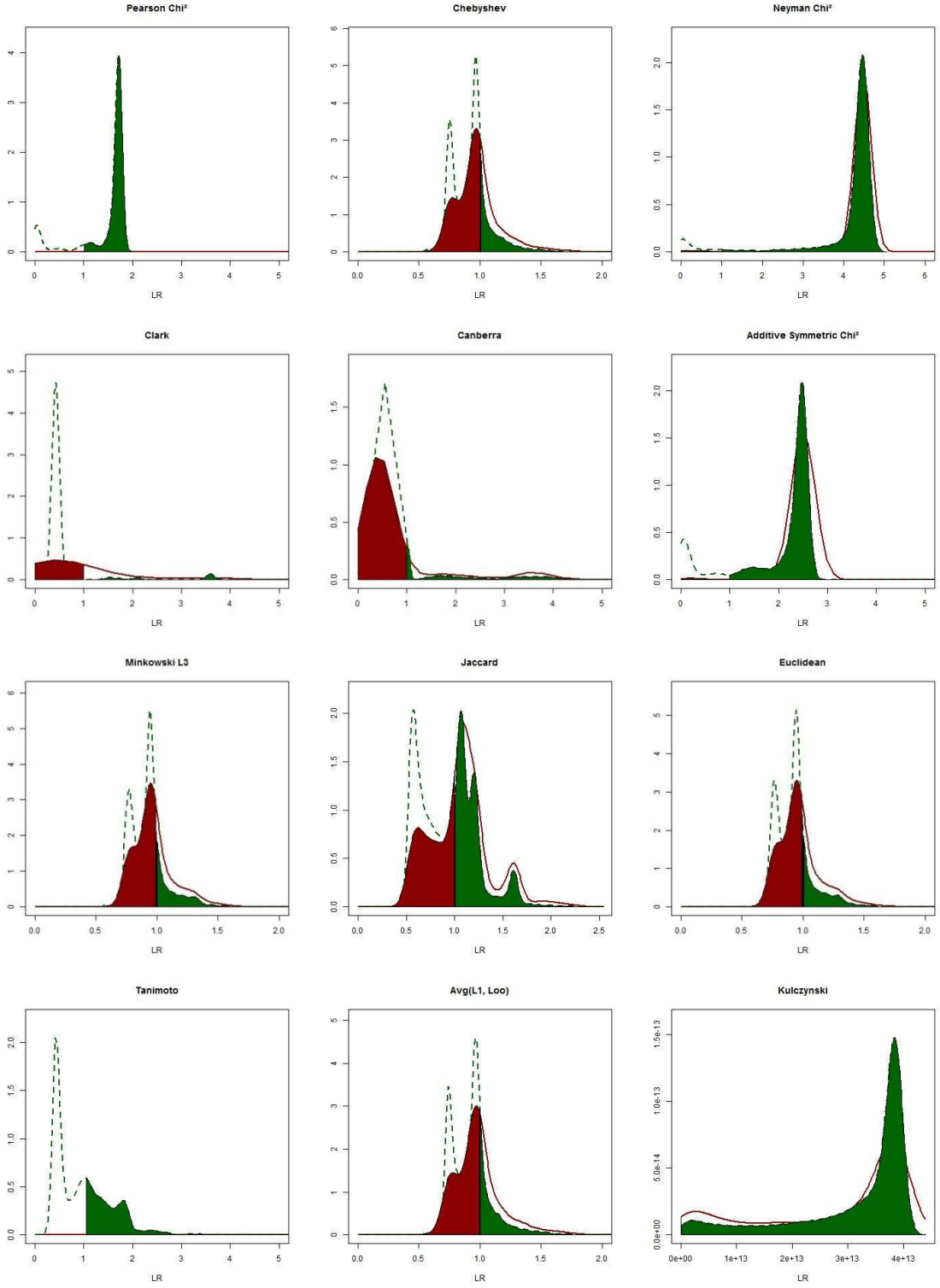

Manhattan
LR

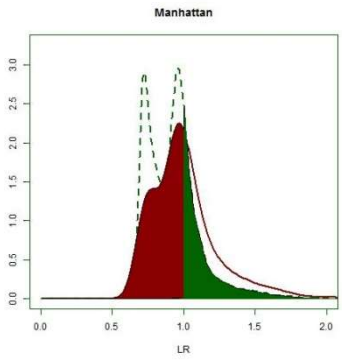

Bray-Curtis
LR

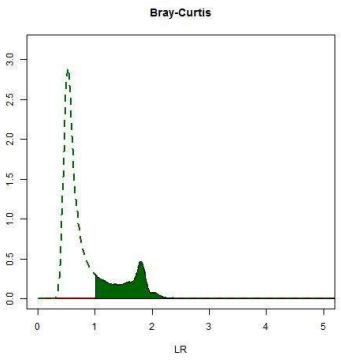

Pearson rho
LR

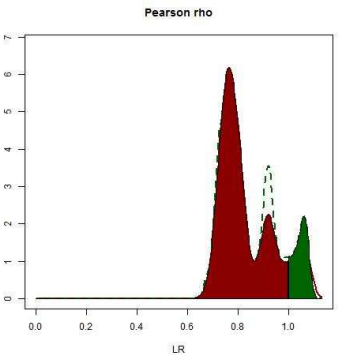

Cosine
LR

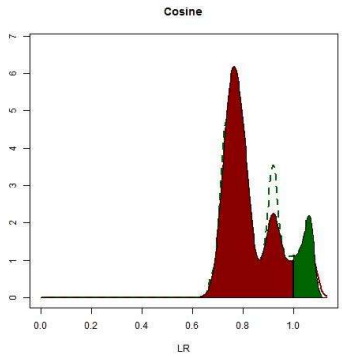
Appendix VI: PDF's of LR's for A_3



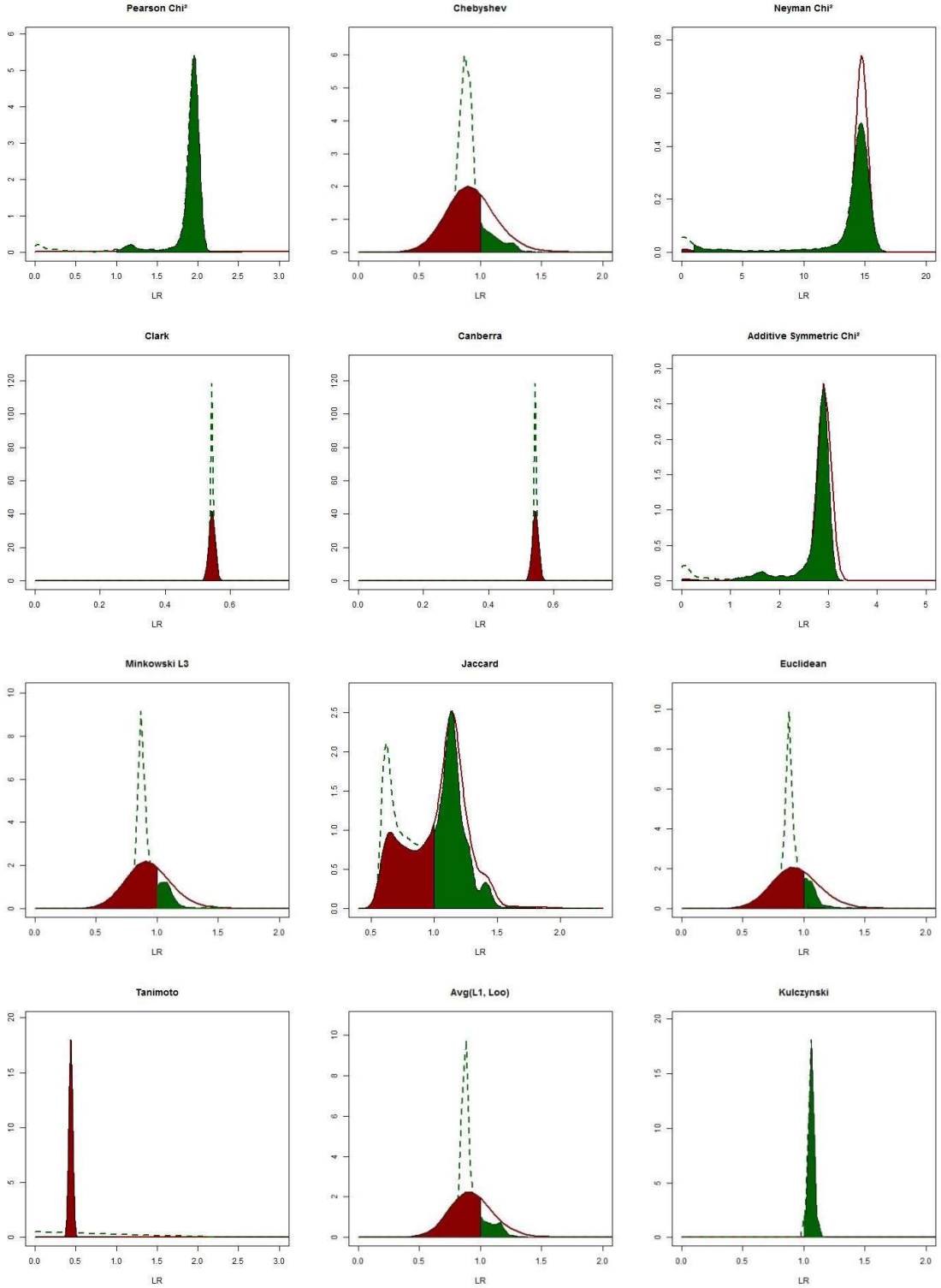

Manhattan
LR

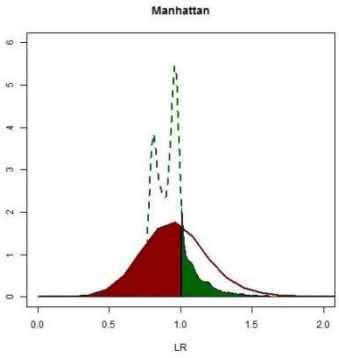

Bray-Curtis
LR

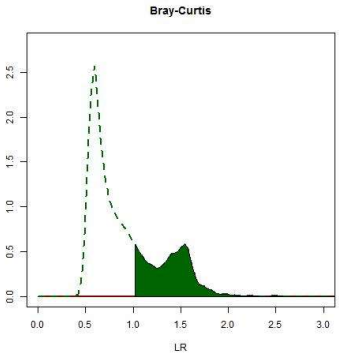

Pearson rho
LR

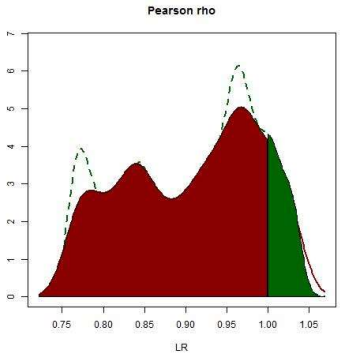

Cosine
LR

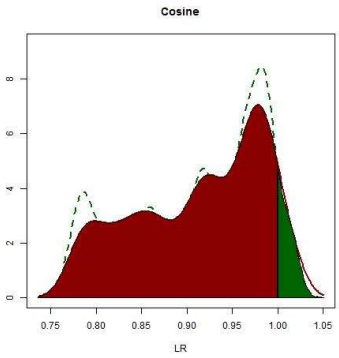
Appendix VII: PDF's of LR's for A_4



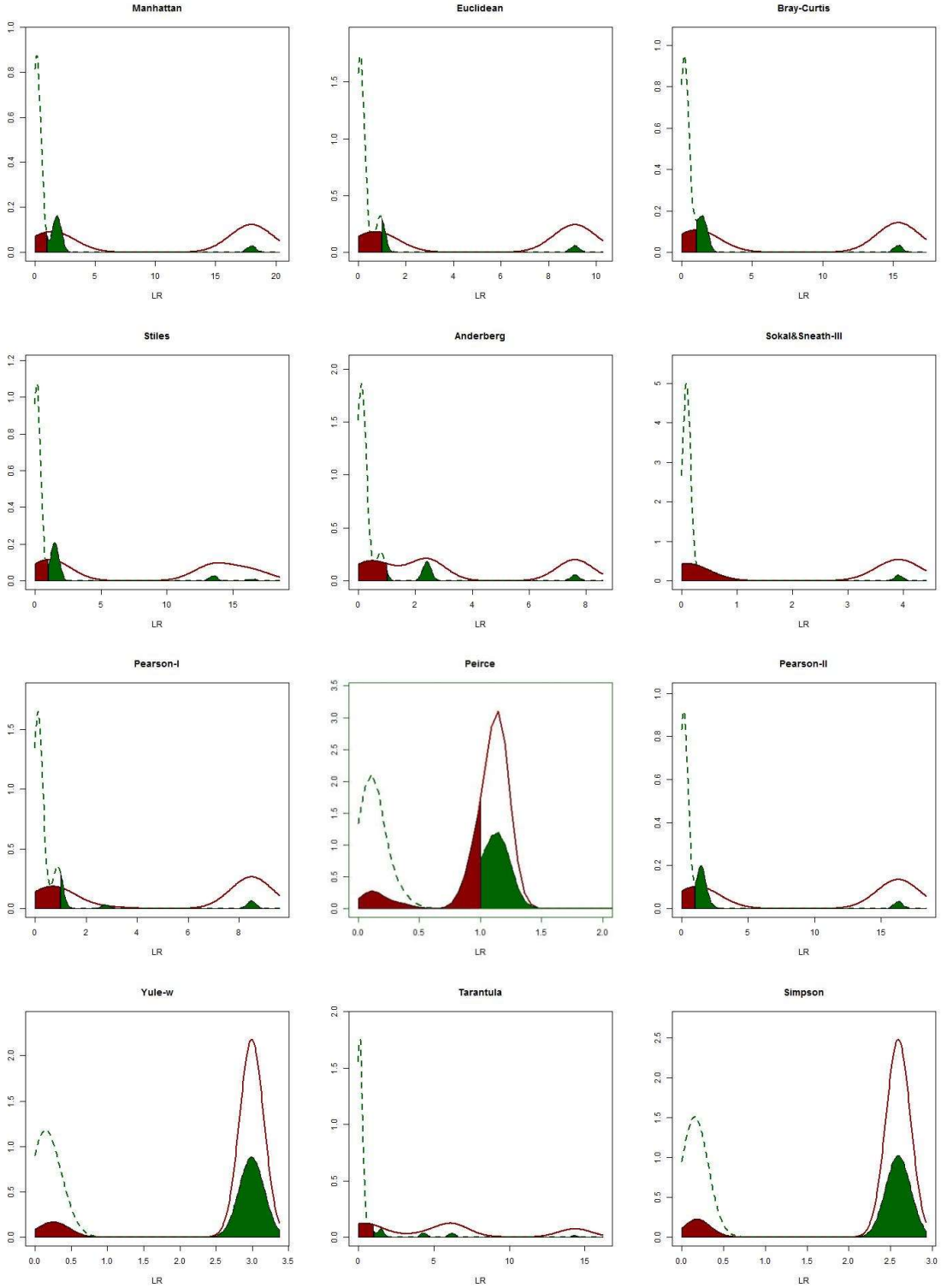

Manhattan
0.0
0.5
1.0
1.5
2.0
LR

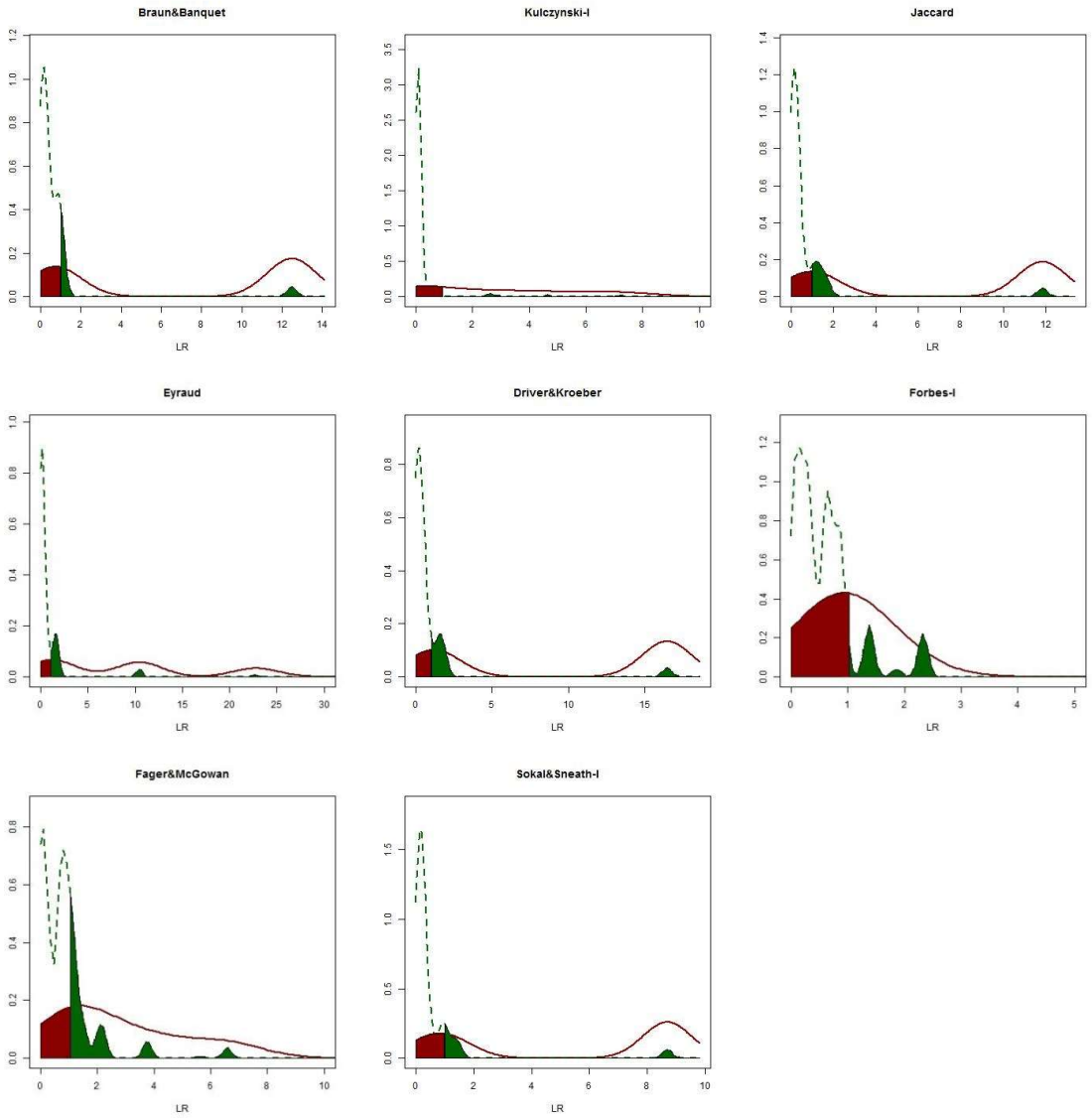

Bray-Curtis
0.0
0.5
1.0
1.5
2.0
2.5
3.0
LR


Pearson rho
0.75
0.80
0.85
0.90
0.95
1.00
1.05
LR


Cosine
0.75
0.80
0.85
0.90
0.95
1.00
1.05
LR

Appendix VIII: PDF's of LR's for A_5




Braun&Banquet
Kulczynski-I
Jaccard
Eyraud
Driver&Kroeber
Forbes-I
Fager&McGowan
Sokal&Sneath-I
LR